Walk a mile in my shoes! 3D visual perspective taking in humans and machines.

Peisen Zhou*, Drew Linsley*, Alekh K. Ashok, Gaurav Gaonkar, Akash Nagaraj, Francis E Lewis, Thomas Serre
Carney Institute for Brain Science, Brown University
Department of Cognitive Linguistic & Psychological Sciences, Brown University
{peisen_zhou,drew_linsley}@brown.edu

Abstract

Visual perspective taking (VPT), and the ability to analyze scenes from a different viewpoint, is an essential feature of human intelligence. We systematically evaluated if a large zoo of over 300 deep neural networks (DNNs) could solve this task like humans can. While DNNs rival human performance on 3D tasks like depth perception, they are significantly worse than humans at VPT. Our findings indicate that despite the incredible progress of DNNs over recent years to rival or exceed human performance on many different visual tasks, significant progress is still needed for them to perceive and function like humans in complex 3D environments.

Keywords: Visual Perspective Taking; Computer Vision;

Introduction

Piaget posited in his theory of cognitive development that human children gain an ability to predict which objects are visible from another viewpoint before the age of 10 (Piaget et al., 1956; Frick et al., 2014). This ability of "Visual Perspective Taking" is a marker for the theory of mind (Aichhorn et al., 2006) and a fundamental feature of human intelligence (VPT, Fig. 1A). In contrast, deep neural networks (DNNs) have been reported to rival or surpass human performance on a variety of visual tasks (Geirhos et al., 2021; Linsley* et al., 2021; Lee et al., 2017) despite taking little — if any — inspiration from how biological brains develop or work. Can today's state-of-the-art DNNs learn strategies for VPT, or are additional insights from Neuroscience and Cognitive Science needed to induce these capabilities into models?

Here, we performed the first large-scale comparison of VPT abilities in humans and machines. While prior VPT experiments tested participants on a handful of real-world scenes or a slightly larger number of generated scenes, we devised an approach to generate an unbounded number of real-world VPT stimuli by leveraging a state-of-the-art 3D computer graphics method known as "3D Gaussian Splatting" (Kerbl et al., 2023) (Fig. 1B). Each generated image shows a single object, a green camera, and a red ball. We then posed two tasks on each image to assess an observer's ability to perceive complementary properties of the depicted 3D scene (Fig. 2C). (1) Depth perception: Is the green camera closer than the red ball? (2) Visual perspective taking: Can the green camera see the red ball? After evaluating 327 different DNNs representing many of the leading approaches, from Visual Transformers (ViT) (Dosovitskiy et al., 2020) trained on ImageNet-21k (Ridnik et al., 2021) to ChatGPT4 (Achiam et al., 2023) and Stable Diffusion 2 (Rombach et al., 2021), we found that while some DNNs rival human accuracy at depth perception accuracy, *there is no model in existence today with a VPT ability comparable to humans*.

Approach and Method

Data generation. We systematically evaluated 3D perception of humans and machines using data generated by Gaussian Splatting (Kerbl et al., 2023) models trained on the Common Objects in 3D (Co3D) (Reizenstein et al., 2021) dataset. Co3D contains videos of unique objects spanning 50 different categories, and we trained Gaussian Splatting models on objects from 30 of the categories. We then added a green camera and red ball into each model (Fig. 1B) in 3D. Next, we generated a large number of images from a camera placed in the 3D scene, and derived ground-truth answers for a depth perception task and visual perspective taking task at each position (Fig. 1C). Thus, we were able to measure an observer's ability to solve two distinct 3D scene perception tasks while holding visual statistics of the stimuli constant. We generated 7,408 unique images in total.

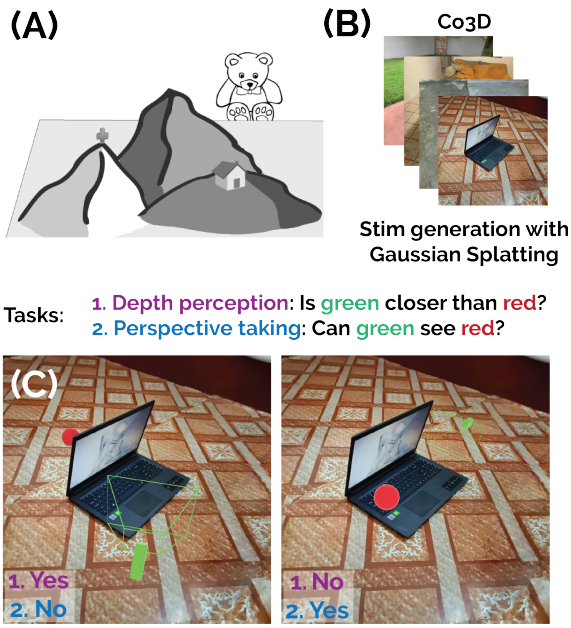

Figure 1: **(A)** Visual Perspective Taking (VPT) is the ability to predict which objects are visible from someone else's point of view, and it is essential for daily interaction with objects. Here, Piaget's "Three Mountains Task" tests VPT by having the viewer describe the scene from the perspective of the bear (Piaget & Inhelder, 1967). **(B)** We generate real-world VPT stimuli using 3D Gaussian Splatting Kerbl et al. (2023) trained on Co3D Reizenstein et al. (2021) **(C)** Each stimulus shows a red ball and a green camera. In the VPT task, observers decide whether the green camera can see the red ball. In the depth perception task, they decide if the green camera is closer than the red ball.

Human psychophysics. We measured the accuracy of 20 human participants on both the depth perception and VPT tasks. To do this we split our dataset into training (6568 images), validation (730 images), and test sets (110 images). Human participants were trained on a small subset of the total training dataset (20 images), and evaluated on the test set (110 images). The trials consisted of a fixation cross pre-

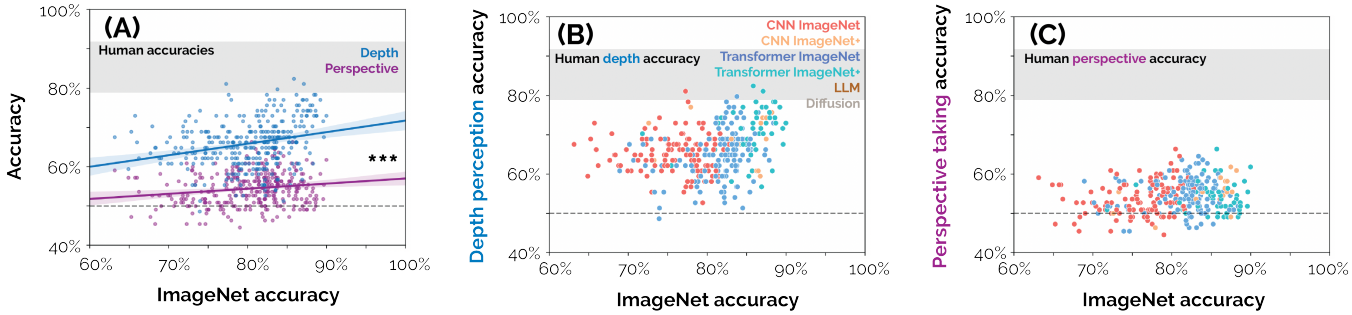

Figure 2: **Humans are significantly better than DNNs at visual perspective taking**. (**A**) While some DNNs rival human performance on depth perception, the average human accuracy (86%) was 20 percentage points better than the best DNN (66%) at VPT. DNNs were also significantly better at depth perception than VPT ($p < 0.001$). (**B**) DNN accuracy on depth perception correlated ($\rho = 0.26$, $p < 0.001$) with accuracy on ImageNet object classification — a standard pretraining task for vision DNNs. Some model classes, like visual transformers trained on datasets much larger than ImageNet (ImageNet+; $\rho = 0.28$, $p < 0.01$), had even stronger correlations between their object classification and depth perception accuracy. (**C**) There was a weak but significant correlation between ImageNet object classification and VPT performance, but no model type exhibited an advantage over any other.

sented for 500ms followed by the stimulus for 3s. Participants had to respond using left/right keys on their keyboard. Participants were recruited from `https://www.prolific.com/` and had to answer either a depth or VPT task posed on each image. The experiment lasted around 15 minutes and participants were paid $\$5$ for their time.

Model zoo and training. We tested 318 DNNs from PyTorch Image Models (TIMM) (Wightman, 2019), foundational vision models like MAE (He et al., 2022), DINO v2 (Oquab et al., 2023), iBOT (Zhou et al., 2021), SAM (Kirillov et al., 2023), midas (Ranftl et al., 2020) and Depth Anything (Yang et al., 2024) that have recently been reported to have surprising capabilities for 3D scene analysis (El Banani et al., 2024), the stable diffusion image generation model, and the state-of-the-art large vision language models (VLMs) ChatGPT4 (Achiam et al., 2023), Gemini (Team et al., 2023), and Claude 3 (Anthropic, 2024). We extracted decisions from the TIMM and foundational vision models by training linear probes, we used a zero-shot evaluation procedure to test the diffusion model (Li et al., 2023), and we tested the VLMs using the same procedure as humans.

Results While human participants were significantly above chance at the depth perception and VPT tasks, the capabilities of DNNs in our zoo were mixed (Fig. 2A). Nearly all DNNs were significantly above chance ($p < 0.001$) at solving the depth task, and their accuracy on it correlated with object classification performance on ImageNet (Deng et al., 2009) (a standard proxy for a model's overall effectiveness FEL et al. 2022). Some of the DNNs we tested rivaled human performance on the depth perception task (*e.g.*, a Visual Transformer trained on ImageNet-21K). We also observed strong and significant correlations between certain model families, like Transformers, and performance on the depth task (Fig. 2B). However, DNNs were significantly worse than humans at the VPT task (Fig. 2A, $p < 0.001$). Their performance on VPT was also significantly worse than on the depth task ($p < 0.001$). The DNN with the highest performance on VPT — the MixNet XL with a DeiT3 base — was 66% accurate, which paled in comparison to the average human accuracy of 86% (Fig 2C). Our results indicate that no class of model or scale of training is sufficient to help DNNs achieve human-level performance at VPT.

Conclusion

Progress in deep learning over the past several years has largely been driven by a scale-up of existing methods. This approach has been undeniably effective for nearly every domain of intelligence it has been applied to, from vision to language, and from biomedicine to physics. However, our work demonstrates that this DNN scale-up is insufficient for inducing models with at least one fundamental aspect of biological intelligence — our ability to switch visual perspectives and analyze the world from a new viewpoint. This ability for VPT has been studied for over a half-century in developmental psychology, and its development has been well-characterized in humans. Thus, our work suggests that some facet of biological development that supports VPT is missing from today's leading DNNs. Characterizing and developing algorithmic abstractions of these missing principles for DNNs may help form the foundation for machines that can perceive the world and others like humans do.

Acknowledgments

This work was funded by ONR grant N00014-24-1-2026. We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program and computing hardware supported by NIH Office of the Director grant S10OD025181 through the Center for Computation and Visualization at Brown University.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind? *NeuroImage*, *30*(3), 1059-1068. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811905024079 doi: https://doi.org/10.1016/j.neuroimage.2005.10.026

Anthropic. (2024). *Claude*. Retrieved from https://www.anthropic.com/claude

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, *abs/2010.11929*. Retrieved from https://arxiv.org/abs/2010.11929

El Banani, M., Raj, A., Maninis, K.-K., Kar, A., Li, Y., Rubinstein, M., ... Jampani, V. (2024). Probing the 3D Awareness of Visual Foundation Models. In *Cvpr.*

FEL, T., Rodriguez, I. F. R., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems.* Retrieved from https://openreview.net/forum?id=ZYKWi6Ylfg

Frick, A., Möhring, W., & Newcombe, N. S. (2014, April). Picturing perspectives: development of perspective-taking abilities in 4- to 8-year-olds. *Front. Psychol.*, *5*, 386.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021, June). Partial success in closing the gap between human and machine vision.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16000–16009).

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023, July). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, *42*(4).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... others (2023). Segment anything. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4015–4026).

Lee, K., Zung, J., Li, P., Jain, V., & Sebastian Seung, H. (2017, May). Superhuman accuracy on the SNEMI3D connectomics challenge.

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., & Pathak, D. (2023). Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 2206–2217).

Linsley*, J. W., Linsley*, D. A., Lamstein, J., Ryan, G., Shah, K., Castello, N. A., ... Finkbeiner, S. (2021, December). Superhuman cell death detection with biomarker-optimized neural networks. *Sci Adv*, *7*(50), eabf8142.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... others (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Piaget, J., & Inhelder, B. (1967). *The child's conception of space*. London: Routledge & K. Paul.

Piaget, J., Inhelder, B., Langdon, F. J., & Lunzer, J. L. (1956). *La représentation de l'espace chez l'enfant. the child's conception of space... translated... by FJ langdon & JL lunzer. with illustrations*. New York; Routledge & Kegan Paul: London; printed in Great Britain.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, *44*(3), 1623–1637.

Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., & Novotny, D. (2021). Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International conference on computer vision.*

Ridnik, T., Baruch, E. B., Noy, A., & Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *CoRR*, *abs/2104.10972*. Retrieved from https://arxiv.org/abs/2104.10972

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. *CoRR*, *abs/2112.10752*. Retrieved from https://arxiv.org/abs/2112.10752

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., ... others (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wightman, R. (2019). *Pytorch image models.* https://github.com/rwightman/pytorch-image-models. GitHub. doi: 10.5281/zenodo.4414861

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.