THE 3D-PC: A BENCHMARK FOR VISUAL PERSPECTIVE TAKING IN HUMANS AND MACHINES

Drew Linsley[†1], **Peisen Zhou**[†1], **Alekh Karkada**[†1], **Akash Nagaraj**[1], **Gaurav Gaonkar**[1], **Francis E Lewis**[1], **Zygmunt Pizlo**[2], **Thomas Serre**[1]

drew_linsley@brown.edu

ABSTRACT

Visual perspective taking (VPT) is the ability to perceive and reason about the perspectives of others. It is an essential feature of human intelligence, which develops over the first decade of life and requires an ability to process the 3D structure of visual scenes. A growing number of reports have indicated that deep neural networks (DNNs) become capable of analyzing 3D scenes after training on large image datasets. We investigated if this emergent ability for 3D analysis in DNNs is sufficient for VPT with the *3D perception challenge* (3D-PC): a novel benchmark for 3D perception in humans and DNNs. The 3D-PC is comprised of three 3D-analysis tasks posed within natural scene images: **1.** a simple test of object *depth order*, **2.** a basic VPT task (*VPT-basic*), and **3.** another version of VPT (*VPT-Strategy*) designed to limit the effectiveness of "shortcut" visual strategies. We tested human participants (N=33) and linearly probed or text-prompted over 300 DNNs on the challenge and found that nearly all of the DNNs approached or exceeded human accuracy in analyzing object *depth order*. Surprisingly, DNN accuracy on this task correlated with their object recognition performance. In contrast, there was an extraordinary gap between DNNs and humans on *VPT-basic*. Humans were nearly perfect, whereas most DNNs were near chance. Fine-tuning DNNs on *VPT-basic* brought them close to human performance, but they, unlike humans, dropped back to chance when tested on *VPT-Strategy*. Our challenge demonstrates that the training routines and architectures of today's DNNs are well-suited for learning basic 3D properties of scenes and objects but are ill-suited for reasoning about these properties as humans do. We release our 3D-PC datasets and code to help bridge this gap in 3D perception between humans and machines.

1 INTRODUCTION

In his theory of cognitive development, Piaget posited that human children gain the ability to predict which objects are visible from another viewpoint before the age of 10 (Piaget et al., 1956; Frick et al., 2014). This "Visual Perspective Taking" (VPT) ability is a foundational feature of human intelligence and a behavioral marker for the theory of mind (Aichhorn et al., 2006). VPT is also critical for safely navigating through the world and socializing with others (Fig. 1A). While VPT has been a focus of developmental psychology research since its initial description (Piaget et al., 1956; Bukowski, 2018; Martin et al., 2019) (Fig. 1B), it has not yet been studied in machines.

One of the more surprising results in deep learning has been the number of concomitant similarities to human perception exhibited by deep neural networks (DNNs), trained on large-scale static image datasets (Yamins & DiCarlo, 2016; Yamins et al., 2014). For example, DNNs now rival or surpass human recognition performance on object recognition and segmentation tasks (Geirhos et al., 2021; Linsley* et al., 2021; Lee et al., 2017), and are the state-of-the-art approach for predicting human neural and behavioral responses to images (Serre, 2019). There is also a growing number of reports indicating that DNNs trained with self-supervision or for object classification learn to encode 3D

[†]These authors contributed equally to this work.
[1]Carney Institute for Brain Science, Brown University, Providence, RI.
[2]Department of Cognitive Sciences, University of California-Irvine, Irvine, CA.

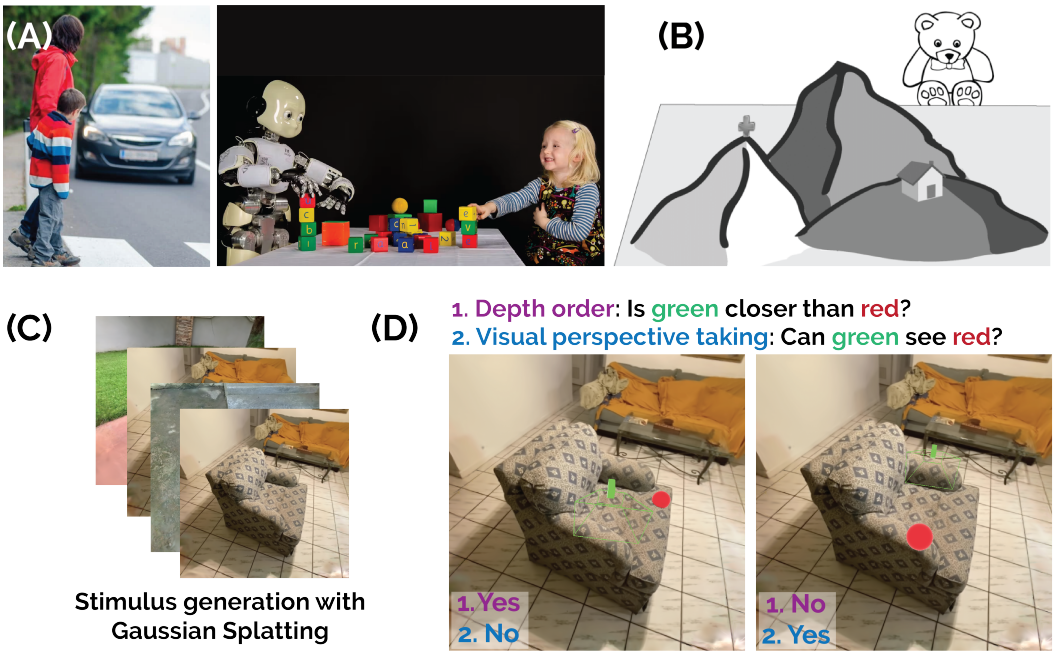

Figure 1: **Visual Perspective Taking (VPT) is the ability to analyze scenes from different viewpoints.** **(A)** Humans rely on VPT to anticipate the behavior of others. We expect that this ability will be essential for creating the next generation of AI assistants that can accurately anticipate human behavior (images are CC BY-NC). **(B)** VPT has been studied in developmental psychology since the mid-20th century using cartoon or highly synthetic stimuli. For example, Piaget's "Three Mountains Task" asks observers to describe the scene from the perspective of a bear (image from Bruce et al. (2017)). **(C)** Here, we use Gaussian Splatting (Kerbl et al., 2023) to develop a 3D scene generation pipeline for the *3D perception challenge* (`3D-PC`), to systematically compare 3D perception capabilities of human and machine vision systems. **(D)** The `3D-PC` tests **1.** Object depth perception, and **2.** VPT.

properties of objects and scenes that humans are also sensitive to, such as the depth and structure of surfaces (Li et al., 2023; Ranftl et al., 2022; Saxena et al., 2023; Liu et al., 2023; Goodwin et al., 2022; Tang et al., 2023; Amir et al., 2021; Chen et al., 2023; Bhattad et al., 2023; El Banani et al., 2024). Are the emergent capabilities of DNNs for 3D vision sufficient for solving VPT tasks?

Here, we introduce the *3D perception challenge* (`3D-PC`) to address this question and systematically compare 3D perceptual capabilities of humans and DNNs. The `3D-PC` evaluates observers on (Fig 1): **1.** identifying the order of two objects in depth (*depth order*), **2.** predicting if one of two objects can "see" the other (*VPT-basic*), and **3.** another version of VPT that limits the effectiveness of "shortcut" solutions (Geirhos et al., 2020a) (*VPT-Strategy*). The `3D-PC` is distinct from existing psychological paradigms for evaluating VPT (Piaget et al., 1956; Bukowski, 2018; Martin et al., 2019) and computer vision challenges for 3D perception (El Banani et al., 2024; Amir et al., 2021) in two ways. First, unlike small-scale psychology studies of VPT, the `3D-PC` uses a novel "3D Gaussian Splatting" (Kerbl et al., 2023) approach which permits the generation of endless real-world stimuli. Second, unlike existing computer vision challenges, our approach for data generation means that the `3D-PC` tests and counterbalances labels for multiple 3D tasks on the exact same images, which controls for potential confounds in analysis and interpretation. We expect that DNNs which rival humans on the `3D-PC` will become ideal models for a variety of real-world applications where machines must anticipate human behavior in real-time, as well as for enriching our understanding of how brains work (Fig. 1A).

Contributions. We built the `3D-PC` and used it to evaluate 3D perception for human participants and 327 different DNNs. The DNNs we tested represented each of today's leading approaches, from

Visual Transformers (ViT) (Dosovitskiy et al., 2021a) trained on ImageNet-21k (Ridnik et al., 2021) to ChatGPT4 (Achiam et al., 2023) and Stable Diffusion 2.0 (Rombach et al., 2021).

- We found that DNNs were very accurate at determining the *depth order* of objects after linear probing or text-prompting. DNNs that are state-of-the-art on object classification matched or exceeded human accuracy on this task.
- However, DNNs dropped close to chance accuracy on *VPT-basic*, whereas humans were nearly flawless at this task.
- Fine-tuning the zoo of DNNs on *VPT-basic* boosted their performance to near human level. However, the performance of the DNNs — but not humans — dropped back to chance on *VPT-Strategy*. DNNs overly rely on the size and location features of objects instead of estimating their line-of-sight to solve VPT like humans do.
- Our findings demonstrate that the visual strategies necessary for solving VPT do not emerge in DNNs from large-scale static image training or after directly fine-tuning on the task.We release the `3D-PC` data, 3D Gaussian Splatting models, code, and human psychophysics to support the development of models that can perceive and reason about the 3D world like humans.

2 Related Work

3D perception in humans. The visual perception of 3D properties is a fundamentally ill-posed problem (Todd, 2004; Pizlo, 2010), which forces biological visual systems to rely on a variety of assumptions to decode the structure of objects and scenes. For example, variations in the lighting, texture gradients, retinal image disparity, and motion of an object all contribute to the perception of its 3D shape. 3D perception is further modulated by top-down beliefs about the structure of the world, which are either innate or shaped by prior sensory experiences, especially visual and haptic ones. In other words, humans learn about the 3D structure of the world in an embodied manner that is fundamentally different than how DNNs learn. In light of this difference, it would be remarkable if DNNs could accurately model how humans perceive their 3D world.

Visual perspective taking in humans. VPT was devised to understand how capabilities for reasoning about objects in the world develop throughout the course of one's life. At least two versions of VPT have been introduced over the years (Michelon & Zacks, 2006; Pizlo, 2022). The version of VPT that we study here — known in the developmental literature as "VPT-1" — is the more basic form, which is thought to rely on automatic feedforward processing in the visual system (Michelon & Zacks, 2006). In light of the well-documented similarities between feedforward processing in humans and DNNs (Serre, 2019; Kreiman & Serre, 2020), we reasoned that this version of VPT would maximize the chances of success for today's DNNs.

3D perception in DNNs. As deep neural networks (DNNs) have increased in scale and training dataset size over the past decade, their performance on essentially all visual challenges has improved. For example, there has also been significant progress in training DNNs to solve 3D computer vision tasks. For example, it is possible to train models that can precisely decode object and scene depth Yang et al. (2024), generate a high-resolution 3D model of an environment Kerbl et al. (2023), or even predict new camera views of an object given a single video of it Zhang et al. (2024). DNNs represent the state-of-the-art approach to nearly every 3D vision task today.

Surprisingly, there is also growing evidence that 3D perception emerges in DNNs even when they are trained on static image tasks. For example, DNNs trained with a variety of self-supervised learning techniques on static image datasets learn to represent the depth, surface normals, and 3D correspondence of features in scenes (Ranftl et al., 2022; Saxena et al., 2023; Liu et al., 2023; Goodwin et al., 2022; Tang et al., 2023; Amir et al., 2021; Chen et al., 2023; Bhattad et al., 2023; El Banani et al., 2024). While similarities between DNNs and human 3D perception have yet to be evaluated systematically, it has been shown that there are differences in how the two reason about the 3D shape of objects (Qian & Ullman, 2024). The `3D-PC` complements prior work by systematically evaluating which aspects of human 3D perception today's DNNs can and cannot accurately represent.

Limitations of DNNs as models of human visual perception. Over recent years, DNNs have grown progressively more accurate as models of human vision for object recognition tasks (Geirhos

et al., 2021; 2020a). At the same time, these models which succeed as models of human object recognition struggle to capture other aspects of visual perception (Bowers et al., 2022) including contextual illusions (Linsley et al., 2020), perceptual grouping (Kim* et al., 2020; Linsley et al., 2021), and categorical prototypes (Golan et al., 2020). There are also multiple reports showing that DNNs are growing less aligned with the visual strategies of humans and non-human primates as they improve on computer vision benchmarks (Linsley et al., 2023b; Fel* et al., 2022; Linsley et al., 2023a). The 3D-PC provides another axis upon which the field can evaluate DNNs as models of human vision.

3 METHODS

The 3D-PC. To enable a fair comparison between human observers' and DNNs' 3D perceptual capabilities, we designed the 3D-PC framework with two goals: **1.** posing different 3D tasks on the same set of stimuli, and **2.** generating a large number of stimuli to properly train DNNs on these tasks. We achieved these goals by combining 3D Gaussian Splatting (Kerbl et al., 2023), videos from the Common Objects in 3D (Co3D) (Reizenstein et al., 2021) dataset, and Unity (Juliani et al., 2018; Pranckevicius, 2023) into a flexible data-generating framework.

Figure 2: **3D-PC examples.** We tested 3D perception in images generated by Gaussian Splatting. Each image depicts a green camera and a red ball. These objects are placed in the scene in a way that counterbalances labels for depth order task and VPT-basic tasks.

Our procedure for building the 3D-PC involved the following three steps. **1.** We trained Gaussian Splatting models on videos in Co3D (Fig. 1C). **2.** We imported these trained models into Unity, where we added green camera and red ball objects into each 3D scene, which were used to pose visual tasks (Fig. 1D). **3.** We then generated random viewpoint trajectories within each 3D scene,

rendered images at each position along the trajectory, and derived ground-truth answers for *depth order* and VPT tasks for the green camera at every position from Unity.

Our approach makes it possible to generate an unlimited number of visual stimuli that test an observer's ability to solve complementary 3D perception tasks (*depth order* and VPT) while keeping visual statistics constant and ground truth labels counterbalanced across tasks. For the version of `3D-PC` used in our evaluation and released at `https://huggingface.co/datasets/3D-PC/3D-PC`, the *depth order* and *VPT-basic* tasks are posed on the same set of 7,480 training images of 20 objects and scenes, and a set of 94 test images of 10 separate objects and scenes (Fig. 2). We held out a randomly selected 10% of the training images for validation and model checkpoint selection.

To build the *VPT-Strategy* task, we rendered images where we fixed the scene camera while we moved the green camera and red ball objects to precisely change the line-of-sight between them from unobstructed to obstructed and back. We reasoned that this experiment would reveal if an observer adopts the visual strategy of taking the perspective of the green camera, which is thought to be used by humans (Michelon & Zacks, 2006), from other strategies that relied on less robust feature-based shortcuts. This dataset consisted of a test set of 100 images for 10 objects and scenes that were not included in *depth order* or *VPT-basic*.

Psychophysics experiment. We tested 10 participants on *depth order*, 20 on *VPT-basic*, and 3 on *VPT-Strategy*. 33 participants were recruited online from Prolific. All provided informed consent before completing the experiment and received $15.00/hr compensation for their time (this amounted to $5.00 for the 15–20 minutes the experiment lasted). These data were de-identified.

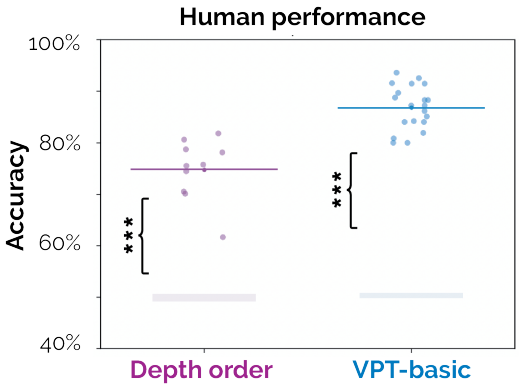

Figure 3: **Human accuracy for object depth order and VPT-basic tasks.** Bars near 50% are label-permuted noise floors; lines are group means. The difference is significant, *** $= p < 0.001$.

Participants were shown instructions for one of the `3D-PC` tasks, then provided 20 examples to ensure that they properly understood it (Appendix Fig 8). These examples were drawn from the DNN training set. Each experimental trial consisted of the following sequence of events overlaid onto a white background: **1.** a fixation cross displayed for 1000*ms*; **2.** an image displayed for 3000*ms*, during which time the participants were asked to render a decision. Participants pressed one of the left or right arrow keys on their keyboards to provide decisions.

Images were displayed at 256×256 pixel resolution, which is equivalent to a stimulus between $5^{o} - 11^{o}$ of visual angle across the range of display and seating setups we expected our online participants used for the experiment.

Model zoo. We evaluated a wide range of DNNs on the `3D-PC`, which represented the leading approaches for object classification, self-supervised pretraining, image generation, depth prediction, and vision language modeling (VLM). Our zoo includes 317 DNNs from PyTorch Image Models (TIMM) (Wightman, 2019), ranging from classic models like `AlexNet` (Krizhevsky et al., 2012a) to state-of-the-art models like `EVA-02` (Fang et al., 2023) (see Appendix 1 for the complete list). We added 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), and `Midas` (Ranftl et al., 2022) (obtained from the GitHub repo of El Banani et al. (2024)). We also included `Depth Anything` (Yang et al., 2024), a foundational model 3D scene analysis and depth prediction (El Banani et al., 2024), as well as the `Stable Diffusion 2.0` (Rombach et al., 2021) image generation model. Finally, we added 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 evaluated a total of 327 models on the `3D-PC`.

Model evaluation. We evaluated all models except for the VLMs on the *depth order* and *VPT-basic* tasks in this challenge by training linear probes on image embeddings from their penultimate layers.

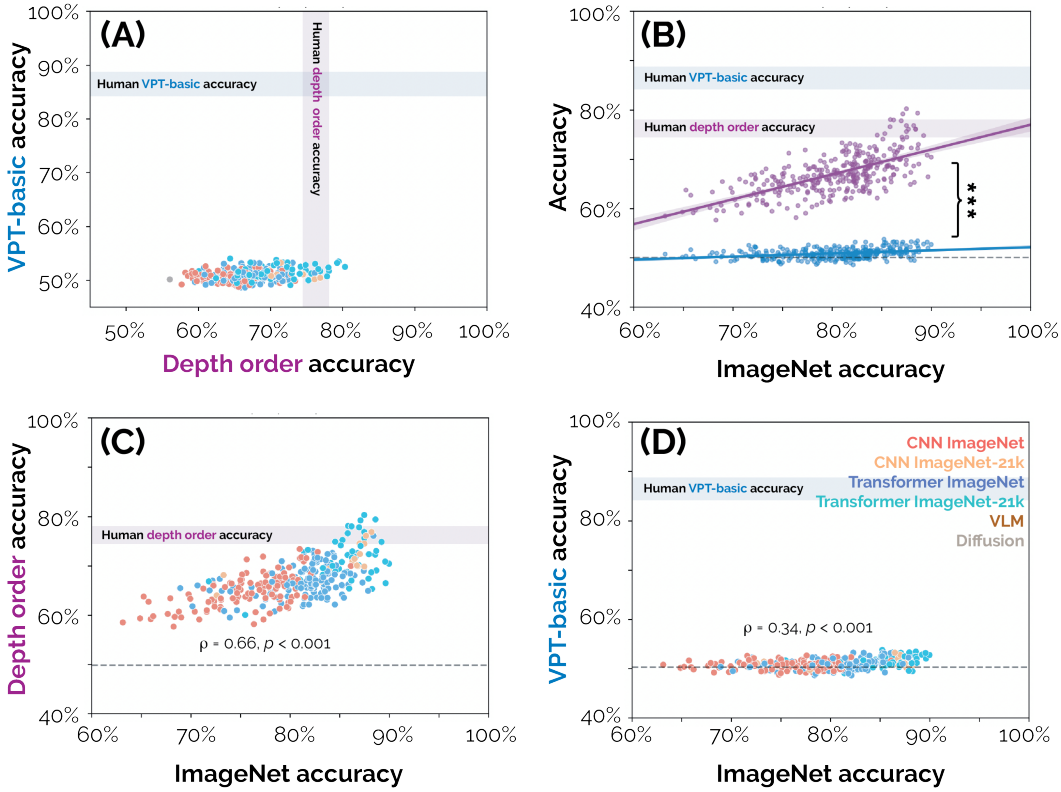

Figure 4: **DNN performance on the depth order and VPT-basic tasks in the 3D-PC after *linear probing or prompting*. (A, B)** DNNs are significantly more accurate at depth order than *VPT-basic*. Human confidence intervals are S.E.M. and ***: $p < 0.001$. **(C, D)** DNN accuracy for *depth order* and *VPT-basic* strongly correlates with object classification accuracy on ImageNet. Dashed lines are the mean of label-permuted human noise floors.

Linear probes were trained using PyTorch (Paszke et al., 2019) for 50 epochs, a 5*e*-4 learning rate, and early stopping (see Appendix A.5 for details). Training took approximately 20 minutes per model using NVIDIA-RTX 3090s. We tested the `Stable Diffusion 2.0` model by adopting the evaluation method used in Li et al. (2023) (see Appendix A.8 for details). We evaluated the VLMs by providing them the same instructions and training images (along with ground truth labels) given to humans, then recording their responses to images from each task via model APIs.

To test the learnability of the `3D-PC`, we also fine-tuned each of the TIMM models in our zoo to solve the tasks. To do this, we trained each of these models for 30 epochs, a 5*e*-5 learning rate, and early stopping (see Appendix A.5 for details). Fine-tuning took between 3 hours and 24 hours per model using NVIDIA-RTX 3090s.

4 Results

Humans find VPT easier than determining the depth ordering of objects. Human participants were on average 74.73% accurate at determining the *depth order* of objects, and 86.82% accurate at solving the *VPT-basic* task (Fig. 3; $p < 0.001$ for both; statistical testing done through randomization tests (Edgington, 1964)). Humans were also significantly more accurate at solving *VPT-basic* than they were at the *depth order* task.

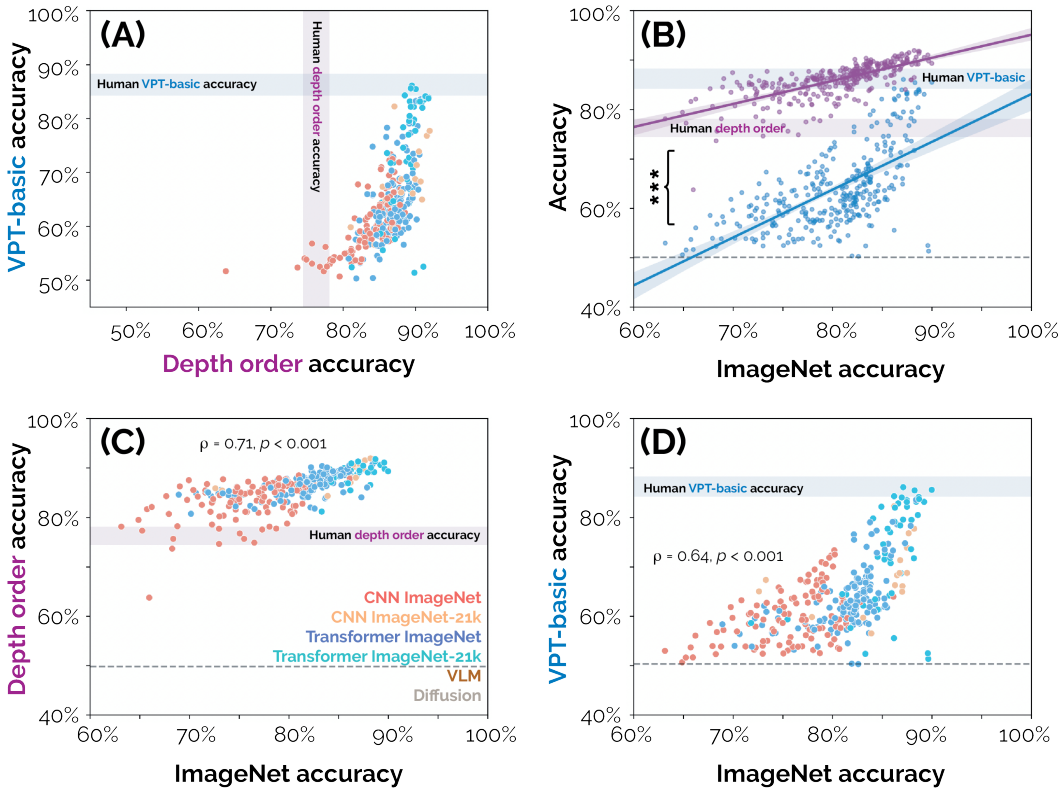

Figure 5: **DNN performance on the depth order and VPT-basic tasks in the 3D-PC after *fine-tuning*. (A)** Fine-tuning makes DNNs far better than humans at the *depth order* task and improves the performance of several DNNs to be at or beyond human accuracy on *VPT-basic*. **(B)** Even after fine-tuning, there is still a significant difference in model performance on *depth order* and *VPT-basic* tasks, $p < 0.001$. **(C, D)** DNN accuracy on both tasks after fine-tuning correlates with ImageNet object classification accuracy. Human confidence intervals are S.E.M. and ***: p < 0.001. Dashed lines are the mean of label-permuted human noise floors.

DNNs learn depth but not VPT from static image training. DNNs showed the opposite pattern of results on *depth order* and *VPT-basic* tasks as humans after linear probing or prompting (Fig. 4): 15 of the DNNs we tested fell within the human accuracy confidence interval on the *depth order* task, and three even outperformed humans (Fig. 4A). In contrast, while humans were on average 86.82% accurate at *VPT-basic*, the DNN which performed the best on this task, the ImageNet 21K-trained `beit` (Bao et al., 2021a), was 53.82% accurate. Even commercial VLMs struggled on *VPT-basic* and were around chance accuracy (`ChatGPT4`: 52%, `Gemini`: 52%, and `Claude 3`: 50%). The *depth order* task was significantly easier for DNNs than *VPT-basic* ($p < 0.001$), which is the opposite of humans (Fig. 4B).

ImageNet accuracy correlates with the 3D capabilities of DNNs. What drives the development of 3D perception in DNNs trained on static images? We hypothesized that as DNNs scale up, they learn ancillary strategies for processing natural images, including the ability to analyze the 3D structure of scenes. To investigate this possibility, we focused on the TIMM models in our DNN zoo. These models have previously been evaluated for object classification accuracy on ImageNet, which we used as a stand-in for DNN scale (Fel* et al., 2022; Linsley et al., 2023a;b). Consistent with our hypothesis, we found a strong and significant correlation between DNN performance on ImageNet and *depth order* task accuracy ($\rho = 0.66, p < 0.001$, Fig. 4C). Despite the very low accuracy of DNNs on *VPT-basic*, there was also a weaker but still significant correlation between performance

on this task and ImageNet ($\rho = 0.34, p < 0.001$, the difference in correlations between the tasks is $\rho = 0.32, p < 0.001$; Fig. 4D). These results suggest that monocular depth cues develop in DNNs alongside their capabilities for object classification[1]. However, the depth cues that DNNs learn are poorly suited for VPT.

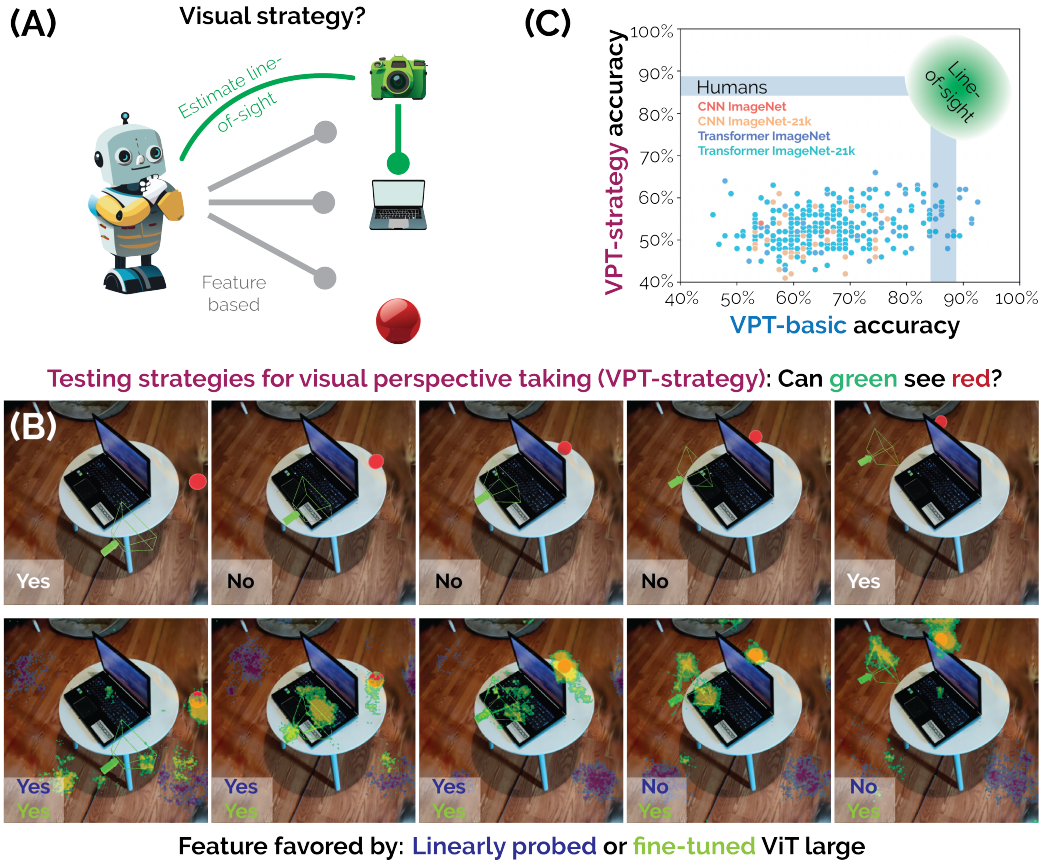

Figure 6: **Even DNNs fine-tuned on VPT-basic fail on VPT-Strategy. (A)** To better characterize the strategy used by humans and DNNs to solve VPT, we devised a new test, *VPT-Strategy*, in which the green camera and red ball are moved through a scene while holding the scene camera and a centrally-positioned object still. This task is easily solvable if an observer *estimates the line-of-sight* of the green camera; other strategies, such as those that rely on specific image features like the relative sizes and positions of objects (*feature based*), may be less effective. **(B)** Examples of *VPT-Strategy* stimuli along with the ground-truth label (top-row) and predictions by a `ViT large` after linearly probing or fine-tuning for *VPT-basic* (bottom-row). Decision attribution maps from each version of the `ViT large`, derived from "smooth gradients" (Smilkov et al., 2017), are overlaid onto bottom-row images (purple/blue=linearly probed, yellow/green=fine-tuned). The fine-tuned ViT locates the green camera and red ball but renders incorrect decisions. **(C)** DNNs fine-tuned on *VPT-basic* fail to solve *VPT-Strategy*; they rely on a brittle feature-based strategy. Humans, on the other hand, are 87% accurate; they likely estimate line-of-sight.

DNNs can solve *VPT-basic* after fine-tuning. One possible explanation for the failure of today's DNNs on *VPT-basic* is that the task requires additional cues for 3D vision that cannot be easily learned from static images. To explore this possibility, we fine-tuned each of the TIMM models in our DNN zoo to solve *depth order* and *VPT-basic* (Fig. 5A). There was still a significant difference between DNN performance on the two tasks (Fig. 5B, $p < 0.001$), but fine-tuning caused 97% of the

[1]More work is needed to identify a causal relationship between the development of monocular depth cues and object recognition accuracy.

DNNs to exceed human accuracy on *depth order*, and four of the DNNs to reach human accuracy on *VPT-basic*. DNN performance on the tasks more strongly correlated with ImageNet accuracy after fine-tuning than linear probing (compare Fig. 5C/D and Fig. 4C/D). We also compared the errors these DNNs made on both tasks to humans. We found nearly all of the fine-tuned DNNs were aligned with humans on *depth order*, and a handful were aligned with humans on *VPT-basic* (Fig. 9).

DNNs learn different strategies than humans to solve VPT. Prior work has found that Deep Neural Networks (DNNs) can achieve human-level performance on visual tasks while using fundamentally different strategies than humans (Linsley et al., 2023b; Fel* et al., 2022; Linsley et al., 2023a). To investigate whether this issue applies to VPT, we designed a new experiment which we call *VPT-Strategy*.

VPT-Strategy was inspired by past work in developmental psychology, where it has been proposed that humans solve VPT by estimating line-of-sight (Fig. 6A) because humans respond in predictable ways when objects in a scene are slightly repositioned (Pizlo, 2022; Michelon & Zacks, 2006). In *VPT-Strategy*, observers solve the VPT task on a series of images rendered from a fixed camera viewpoint which show the green camera and red ball incrementally repositioned from one side of the screen to the other, passing by an occluding object in the process. This makes it possible to precisely map out the moments at which the green camera's view of the red ball is unoccluded, then occluded, then unoccluded once more. DNNs that have been fine-tuned on *VPT-basic* behaved differently than humans on this task: humans were 87% accurate, but the highest performing DNN, the `Swin Transformer` (Liu et al., 2021) trained on ImageNet-21k, was only 66% accurate (Fig. 6C). Since DNNs cannot generalize to *VPT-Strategy*, it means that they do not learn to estimate line-of-sight to solve VPT.

To better understand the source of DNN failures, we derived "smooth gradients" (Smilkov et al., 2017) decision attribution maps of every DNN that we tested on *VPT-Strategy*, and then measured the overlap of these maps with the green camera and red ball objects using Dice similarity coefficient (Dice, 1945). These attribution maps revealed that the DNNs we evaluated did indeed learn to attend to the locations of the green camera and red ball objects when trying to solve VPT (Fig. 6B and Fig. 10), and the precision of their localization increased as a function of accuracy on *VPT-basic* (Fig. 10). When considered alongside the poor DNN performance on *VPT-strategy*, these results imply that DNNs learn a brittle feature-based strategy that is dependent on object properties like size and location instead of estimating the line-of-sight of the green camera like humans do (Fig. 6C).

5 Discussion

Deep neural networks (DNNs) have rapidly advanced over recent years to the point where they match or surpass human-level performance on numerous visual tasks. However, our `3D-PC` reveals there is still a significant gap between the abilities of humans and DNNs to reason about 3D scenes. While DNNs match or exceed human accuracy on the basic object *depth order* task after linear probing or prompting, they struggle remarkably on even the basic form of VPT that we test in the `3D-PC`. Fine-tuning DNNs on *VPT-basic* allows them to approach human-level performance, but unlike humans, their strategies do not generalize to the *VPT-Strategy* task.

A striking finding from our study is the strong correlation between DNNs' object classification accuracy on ImageNet and their performance on *depth order* and *VPT-basic*. This correlation suggests that monocular depth cues emerge in DNNs as a byproduct of learning to recognize objects, potentially because these cues are useful for segmenting objects from their backgrounds. The difference in DNN effectiveness for *depth order* versus *VPT-basic*, however, indicates that these cues are not sufficient for reasoning about the 3D structure of scenes in the way that VPT demands.

Thus, today's approaches for developing DNNs, which primarily focus on static image datasets, may be poorly suited for enabling robust 3D perception and reasoning abilities akin to those of humans. Incorporating insights from human cognition and neuroscience into DNNs, particularly in ways biological visual systems develop 3D perception, could help evolve more faithful models of human intelligence.

Limitations One key limitation of our study is that we were not able to isolate the precise cues that humans versus DNNs use to solve *depth order*. While both were very capable of solving the

task, it's very possible that both relied on different monocular depth cues. For example, the relative size of objects is a very salient cue for solving the task. However, object occlusion and interposition is another strong cue in this dataset, and it's unclear if DNNs and humans rely on one or the other, or both, or other cues like the lighting and shadows of objects. Future work that can drill down on the strategies of humans and DNNs for perceiving 3D spatial properties will make significant contributions for understanding the alignment of biological and machine vision.

Another key limitation of our study is that our version of VPT represents the most basic form studied in the developmental psychology literature. While solving this task is evidently an extraordinary challenge for DNNs, it is only one small step towards human-level capabilities for reasoning about 3D worlds in general. Far more research is needed to identify additional challenges, architectures, and training routines that can help DNNs perceive and reason about the world like humans do. We release our `3D-PC` data, 3D Gaussian Splatting models, and code anonymously at `https://huggingface.co/datasets/3D-PC/3D-PC` to support this goal.

Finally, we validated developmental psychology work that humans rely on a line-of-sight strategy to solve VPT, and demonstrated in our *VPT-Strategy* task that DNNs learn to use a different strategy after fine-tuning. We provided evidence through our attribution map experiment (Fig. 10) that DNN strategies are highly focused on the target objects. However, we were not able to characterize exactly how DNNs use object features for VPT, which may be critical for developing the next-generation of vision models that can solve VPT like humans do.

ACKNOWLEDGMENTS

Our work is supported by ONR (N00014-24-1-2026, NSF (IIS-2402875), the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). Additional support was provided by the Carney Institute for Brain Science and the Center for Computation and Visualization (CCV). We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program as well as computing hardware supported by NIH Office of the Director grant S10OD025181.

REFERENCES

Openai Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, S Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, B Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, L Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, C Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, S Gray, Ryan Greene, Joshua Gross, S Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, I Kanitscheider, N Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, J Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, A Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, J Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, A Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, S McKinney, C McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P Mossing, Tong Mu, Mira Murati, O Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang

Long, Cullen O'Keefe, J Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, M Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, T Sherbakov, Jessica Shieh, S Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D Sokolowsky, Yang Song, Natalie Staudacher, F Such, Natalie Summers, I Sutskever, Jie Tang, N Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, P Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. *arXiv preprint arXiv:2303. 08774*, March 2023.

Markus Aichhorn, Josef Perner, Martin Kronbichler, Wolfgang Staffen, and Gunther Ladurner. Do visual perspective tasks need theory of mind? *Neuroimage*, 30(3):1059–1068, April 2006.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *arXiv preprint arXiv:2112. 05814*, 2021.

Anthropic. Claude, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of image transformers. June 2021a.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021b.

Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. StyleGAN knows normal, depth, albedo, and more. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 73082–73103. Curran Associates, Inc., 2023.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, Benjamin D Evans, Jeffrey Mitchell, and Ryan Blything. Deep problems with neural network models of human vision. *Behav. Brain Sci.*, pp. 1–74, December 2022.

Catherine D Bruce, Brent Davis, Nathalie Sinclair, Lynn McGarvey, David Hallowell, Michelle Drefs, Krista Francis, Zachary Hawes, Joan Moss, Joanne Mulligan, Yukari Okamoto, Walter Whiteley, and Geoff Woolcott. Understanding gaps in research networks: using "spatial reasoning" as a window into the importance of networked educational research. *Educational Studies in Mathematics*, 95(2):143–161, June 2017.

Henryk Bukowski. The neural correlates of visual perspective taking: a critical review. *Current Behavioral Neuroscience Reports*, 5(3):189–197, September 2018.

Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *International Conference on Computer Vision (ICCV)*, 2021a.

Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. June 2023.

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks, 2017.

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 589–598, 2021b.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS 2021*, 2021. URL https://openreview.net/forum?id=5kTlVBkzSRx.

Cheng Cui, Tingquan Gao, Shengyu Wei, Yuning Du, Ruoyu Guo, Shuilong Dong, Bin Lu, Ying Zhou, Xueying Lv, Qiwen Liu, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-lcnet: A lightweight cpu convolutional neural network, 2021.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

St'ephane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, July 1945.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformer. In *ECCV*, 2022.

A Dosovitskiy, L Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M Dehghani, Matthias Minderer, G Heigold, S Gelly, Jakob Uszkoreit, and N Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021a.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021b.

E S Edgington. RANDOMIZATION TESTS. *J. Psychol.*, 57:445–449, April 1964.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. In *CVPR*, 2024.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *arXiv preprint arXiv:2106.09681*, 2021.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis, 2023.

Thomas Fel*, Ivan Felipe*, Drew Linsley*, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Adv. Neural Inf. Process. Syst.*, 2022.

Andrea Frick, Wenke Möhring, and Nora S Newcombe. Picturing perspectives: development of perspective-taking abilities in 4- to 8-year-olds. *Front. Psychol.*, 5:386, April 2014.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652âĂŞ662, Feb 2021. ISSN 1939-3539. doi: 10.1109/tpami.2019.2938758. URL http://dx.doi.org/10.1109/TPAMI.2019.2938758.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020a.

Robert Geirhos, K Meding, and Felix Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *NeurIPS*, 2020b.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. June 2021.

Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci. U. S. A.*, 117 (47):29330–29337, November 2020.

Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-Shot Category-Level object pose estimation. In *Computer Vision – ECCV 2022*, pp. 516–532. Springer Nature Switzerland, 2022.

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve Jegou, and Matthijs Douze. Levit: A vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12259–12269, October 2021.

Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rexnet: Diminishing representational bottleneck on convolutional neural network, 2020a.

Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations, 2020b.

Kai Han, Yunhe Wang, Qiulin Zhang, Wei Zhang, Chunjing Xu, and Tong Zhang. Model rubikâĂŹs cube: Twisting resolution, depth and width for tinynets. *Advances in Neural Information Processing Systems*, 33:19353–19364, 2020c.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019. URL `http://arxiv.org/abs/1905.02244`.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A general platform for intelligent agents. September 2018.

B Kerbl, Georgios Kopanas, Thomas Leimkuehler, and G Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42:1–14, July 2023.

Junkyung Kim*, Drew Linsley*, Kalpit Thakkar, and Thomas Serre. Disentangling neural mechanisms for perceptual grouping. *International Conference on Representation Learning*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and Others. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Gabriel Kreiman and Thomas Serre. Beyond the feedforward sweep: feedback computations in the visual cortex. *Ann. N. Y. Acad. Sci.*, 1464(1):222–241, March 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012a.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012b.

Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the SNEMI3D connectomics challenge. May 2017.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection, 2022a.

Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. *arXiv preprint arXiv:2212.08059*, 2022b.

Drew Linsley, Junkyung Kim, Alekh Ashok, and Thomas Serre. Recurrent neural circuits for contour detection. *International Conference on Learning Representations*, 2020.

Drew Linsley, Girik Malik, Junkyung Kim, Lakshmi Narasimhan Govindarajan, Ennio Mingolla, and Thomas Serre. Tracking without re-recognition in humans and machines. In M Ranzato, A Beygelzimer, Y Dauphin, P S Liang, and J Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19473–19486. Curran Associates, Inc., 2021.

Drew Linsley, Pinyuan Feng, Thibaut Boissin, Alekh Karkada Ashok, Thomas Fel, Stephanie Olaiya, and Thomas Serre. Adversarial alignment: Breaking the trade-off between the strength of an attack and its relevance to human perception. June 2023a.

Drew Linsley, Ivan F Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Adv. Neural Inf. Process. Syst.*, 2023b.

Jeremy W Linsley*, Drew A Linsley*, Josh Lamstein, Gennadi Ryan, Kevan Shah, Nicholas A Castello, Viral Oza, Jaslin Kalra, Shijie Wang, Zachary Tokuno, Ashkan Javaherian, Thomas Serre, and Steven Finkbeiner. Superhuman cell death detection with biomarker-optimized neural networks. *Sci Adv*, 7(50):eabf8142, December 2021.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *International Workshop on Computational Aspects of Deep Learning at 17th European Conference on Computer Vision (CADL2022)*. Springer, 2022.

Andrew K Martin, Garon Perceval, Islay Davies, Peter Su, Jasmine Huang, and Marcus Meinzer. Visual perspective taking in young and older adults. *J. Exp. Psychol. Gen.*, 148(11):2006–2026, November 2019.

Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022.

Pascale Michelon and Jeffrey M Zacks. Two kinds of visual perspective taking. *Percept. Psychophys.*, 68(2):327–337, February 2006.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and Others. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304. 07193*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, High-Performance deep learning library. December 2019.

Jean Piaget, Bärbel Inhelder, Frederick John Langdon, and J L Lunzer. *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, 1956.

Zygmunt Pizlo. *3D Shape: Its Unique Place in Visual Perception*. Mit Pr, January 2010.

Zygmunt Pizlo. *Problem Solving: Cognitive Mechanisms and Formal Models*. Cambridge University Press, new edition edition, July 2022.

Aras Pranckevicius. Unity gaussian splatting. `https://github.com/aras-p/UnityGaussianSplatting`, 2023.

Peng Qian and Tomer D Ullman. Shape guides visual pretense. January 2024.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr DollÃąr. Designing network design spaces, 2020.

Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637, March 2022.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3D category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. April 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution image synthesis with latent diffusion models. December 2021.

Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J Fleet. Zero-Shot metric depth with a Field-of-View conditioned diffusion model. December 2023.

Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annu Rev Vis Sci*, 5:399–426, September 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. June 2017.

Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. Single-path nas: Designing hardware-efficient convnets in less than 4 hours, 2019.

Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions, 2019.

Mingxing Tan and Quoc V. Le. Mixconv: Mixed depthwise convolutional kernels, 2019.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile, 2019.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1363–1389. Curran Associates, Inc., 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and Others. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312. 11805*, 2023.

James T Todd. The visual perception of 3D shape. *Trends Cogn. Sci.*, 8(3):115–121, March 2004.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, July 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herv'e J'egou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 32–42, October 2021b.

Asher Trockman and J. Zico Kolter. Patches are all you need?, 2022.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *ECCV*, 2022.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):1–10, 2022.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. URL `http://arxiv.org/abs/1611.05431`.

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9981–9990, October 2021.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, March 2016.

Daniel L K Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23):8619–8624, June 2014.

Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401. 10891*, 2024.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation, 2019.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022a.

Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *arXiv preprint arXiv:2210.13452*, 2022b.

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv [cs.CV]*, November 2024.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks, 2020.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, A Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with online tokenizer. *ArXiv*, abs/2111.07832, November 2021.

A Appendix

A.1 Author Statement

As authors of this dataset, we bear all responsibility for the information collected and in case of violation of rights and other ethical standards. We affirm that our dataset is shared under a Creative Commons CC-BY license.

A.2 Data Access

We release benchmarking code, data, and 3D Gaussian Splatting models anonymously at `https://huggingface.co/datasets/3D-PC/3D-PC`.

A.3 Potential negative societal impacts of this work

The most obvious potential negative impact of our work is that advancing visual perspective taking (VPT) capabilities in artificial agents could potentially enable militaristic applications or surveillance overreach. However, we hope that our benchmark will aid in the development of AI-based assistants that can better anticipate and react to human needs and social cues for safer navigation and interaction. We also believe that our benchmark will guide the development of better computational models of human 3D perception as well as the neural underpinnings of these abilities.

A.4 Data Generation

To generate data for the `3D-PC`, we first trained 3D Gaussian Splatting (Kerbl et al., 2023) models on videos from the Common Objects in 3D (Co3D) (Reizenstein et al., 2021), which yielded 3D representations of each scene. We then imported trained models into Unity (Juliani et al., 2018) using Unity Gaussian Splatting (Pranckevicius, 2023) and added 3D models of the green camera and red ball to each. Finally, we rendered 50 images along a smooth viewpoint camera trajectory sampled near the original trajectory used for training the Gaussian Splatting model. For each 3D scene, we created 5 positive and 5 negative settings for VPT.

To generate *VPT-basic*, the generation process was repeated for 30 Co3D videos from 10 different categories. We removed any images where the green camera and red ball were not visible. We then split the images into a training set of 7480 images from 20 scenes and a testing set of 94 images from 10 other scenes. For the *depth order* task, we used the same data splits but removed any ambiguous samples where the objects were similarly close to the camera. The resulting dataset for the *depth*

order task contains 4787 training images and 94 testing images. The same set of testing images is used for both model and human benchmarks.

For *VPT-Strategy*, we used the same process to generate data from 10 additional Co3D scenes not included in *VPT-basic* and additionally controlled the positions of the green camera and the red ball. The angle between these two objects was held constant while we moved them so that their line of sight was unobstructed, obstructed, and then unobstructed once again. For each Co3D scene, we rendered 10 settings from a fixed viewpoint camera position, resulting in 100 images in total for *VPT-Strategy*.

A.5 MODEL ZOO

We linearly probed 317 DNNs from Pytorch Image Models (TIMM) (Wightman, 2019) (Table 1) along with foundational vision models following the procedures in (El Banani et al., 2024). All DNNs were trained and evaluated with NVIDIA-RTX 3090 GPUs. All linear probes were trained for 50 epochs, with a $5e-4$ learning rate, a $1e-4$ weight decay, a 0.3 dropout rate, and a batch size of 128. We fine-tuned each of the TIMM models for 30 epochs, a $5e-5$ learning rate, $1e-4$ weight decay, 0.7 dropout rate, and a batch size of 16. Linear probing took approximately 20 minutes per model, and fine-tuning varied from 3 to 24 hours on a NVIDIA-RTX 3090 GPU.

A.6 VLM EVALUATION

We evaluated the following proprietary VLMs on the *VPT-basic* and *depth order* tasks: GPT-4 (`gpt-4-turbo`), Claude (`claude-3-opus-20240229`), and Gemini (`gemini-pro-vision`). To evaluate these VLMs, we used their APIs to send queries containing 20 training images, with ground truth answers as context, as well as a test image. The prepended 20 training images meant that for every example in the challenge, VLMs were given the opportunity to learn, "in-context", how to solve the given task.

The prompt we used for the depth task was "`In this image, is the red ball closer to the observer or is the green arrow closer to the observer? Answer only BALL if the red ball is closer, or ARROW if the green arrow is closer, nothing else.`" and the prompt for the *VPT-basic* task was "`In this image, if viewed from the perspective of the green 3D arrow in the direction the arrow is pointing, can a human see the red ball? Answer only YES or NO, nothing else`". We evaluated each model's generated responses across multiple temperatures, ranging from 0.0 to 0.7 in increments of 0.1, and we report the average of the best 3 runs. Note that while this evaluation approach gives the VLMs more opportunities to perform well on our benchmark than other models, they still struggled immensely (see main text).

A.7 VLM CHAIN-OF-THOUGHT EVALUATION

Chain-of-Thought (COT) prompting is a technique used to improve VLM and LLM performance by showing models the conceptual steps that might be helpful for solving a problem. COT has proven very effective on reasoning tasks. To understand if COT would also help DNNs solve VPT, we devised a prompting approach described below (see Fig. 7 for an example image used when prompting for COT reasoning). However, COT did not improve VLM accuracy on VPT (it achieved chance accuracy).

Depth order: Here is a sample training image, from the perspective shown, is the red ball closer to the observer or is the green arrow closer to the observer? Answer only 'BALL' if the red ball is closer, or 'ARROW' if the green arrow is closer, nothing else. To answer this, let's think step by step: The red ball is above the suitcase, which is a small distance away from the observer, approximately 3 feet. The green arrow is in front of the suitcase, which is closer to the observer, approximately 1 foot away. Thus the answer is: ARROW.

Figure 7: **An example image for Chain-of-Thought prompting of vision language models to solve VPT.**

VPT: Here is a sample training image, if viewed from the perspective of the green 3D arrow, in the direction the arrow is pointing, can a human see the red ball? Answer only 'YES' or 'NO', nothing else. Let's think step by step: If a human were looking in the direction the arrow was pointing, they would face a few possible occlusions, such as the suitcase. The red ball is above the suitcase, so the suitcase wouldn't block the view of the red ball. Thus, the answer is: YES

A.8 Stable Diffusion Evaluation

We followed the method of Li et al. (2023) to evaluate `Stable Diffusion 2.0` on the `3D-PC`. This involved trying multiple prompts to optimize the zero-shot classification performance of the `Stable Diffusion 2.0` model, on *VPT-basic* and *depth order* tasks. For *VPT-basic* we found that the prompt `"A photo with red ball is visible from the green arrow's perspective"` for positive class and `"A photo with red ball not visible from the green arrow's perspective"` for the negative class led to the best performance. For the *depth order* task, the prompt with the highest performance was `"A photo with green arrow closer to the camera as compared to red ball"` and `"A photo with red ball closer to the camera as compared to green arrow"` for positive and negative classes respectively.

A.9 Human Benchmark

We recruited 30 participants through Prolific, compensating each with $5 upon successful completion of all test trials. Participants confirmed their completion by pasting a unique system-generated code into their Prolific accounts. The compensation was prorated based on the minimum wage. We also incurred a 30% overhead fee per participant paid to Prolific. In total, we spent $195 on these benchmark experiments.

A.9.1 Experiment design

At the outset of the experiment, we acquired participant consent through a form approved by the Institutional Review Board (IRB). The experiment was performed on a computer using the Chrome browser. Following consent, we presented a demonstration with instructions and an example video. Participants had the option to revisit the instructions at any time during the experiment by clicking a link in the top right corner of the navigation bar.

In the *depth order* task, the participants were asked to classify the image as "positive" (the green arrow in closer to the viewer) or "negative" (the red ball is closer) using the right and left arrow keys respectively. The choice for keys and their corresponding instances were mentioned below the image on every screen (See Appendix Fig. A1. Participants were given feedback on their response

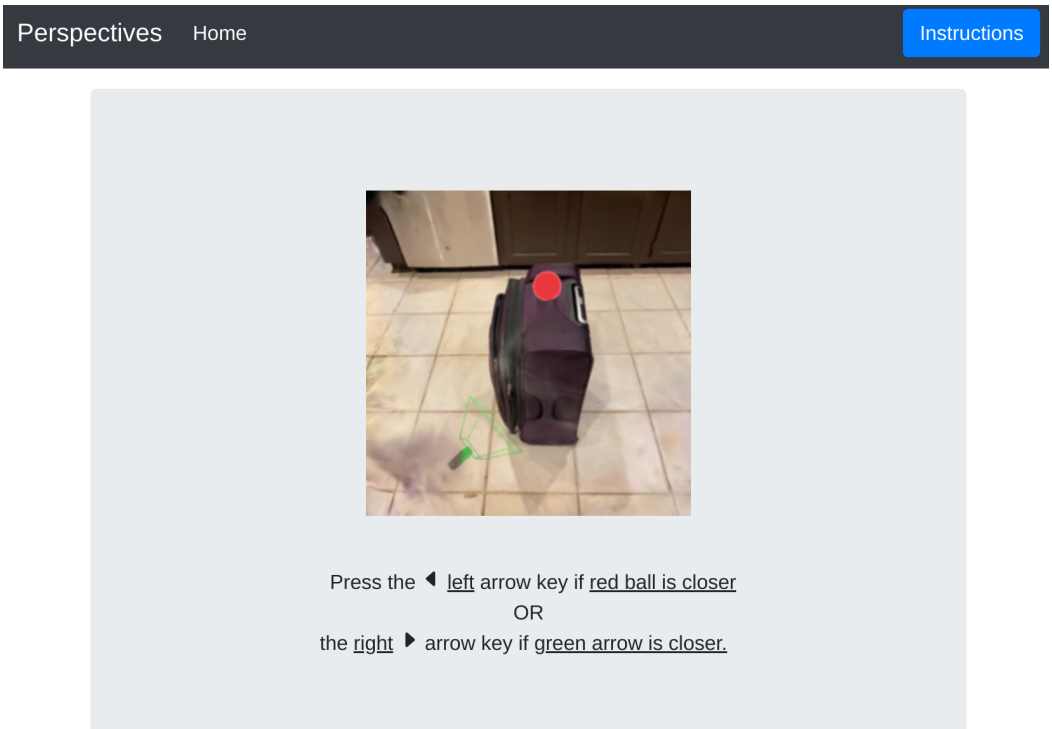

Figure 8: **An experiment trial.**

(correct/incorrect) during every practice trial, but not during the test trials. In the VPT tasks, the choices were "the green arrow/camera see the red ball" or "the green arrow/camera can not see the red ball".

The experiment was not time-bound, allowing participants to complete it at their own pace. Participants typically took around 20 minutes. After each trial, participants were redirected to a screen confirming the successful submission of their responses. They could start the next trial by clicking the "Continue" button or pressing the spacebar. If they did not take any action, they were automatically redirected to the next trial after 1000 milliseconds. Additionally, participants were shown a "rest screen" with a progress bar after every 40 trials, where they could take additional and longer breaks if needed. The timer was turned off during the rest screen.

A.10 Human vs. DNN decision making on *VPT-basic*

We compared the decision strategies of humans and DNNs on *VPT-basic* by measuring the correlations between their error patterns with Cohen's κ (Geirhos et al., 2020b). Model κ scores were mostly correlated with accuracy on *VPT-basic* after linear probes and fine-tuning (Fig. 9). However, while nearly all DNNs were highly correlated with human error patterns after fine-tuning, the correlation between κ scores and task accuracy disappeared (Fig. 9B, purple dots).

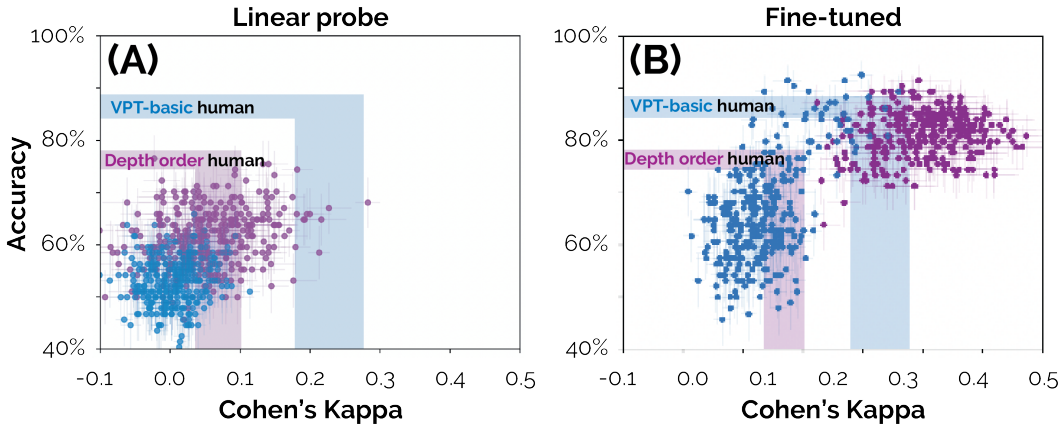

Figure 9: **Error pattern correlations (Cohen's κ) between humans and DNNs on *VPT-basic* and *Depth order*.**

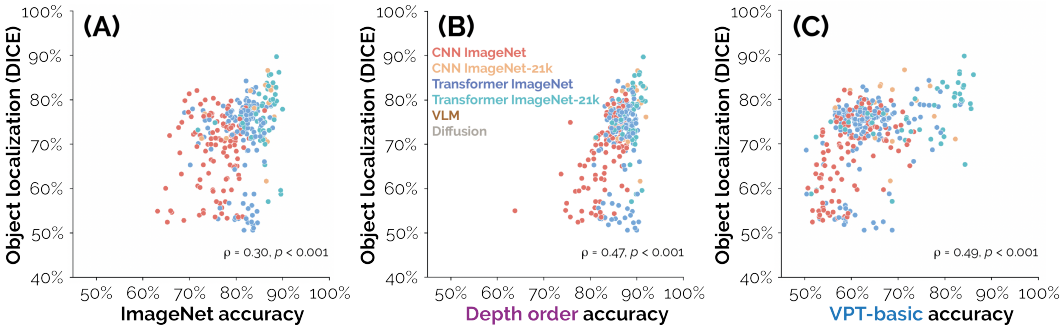

Figure 10: **DNNs fine-tuned on VPT-basic learn to solve the task by overfitting to the positions of the target objects. (A, B, C)** There are significant correlations between DNN accuracy on ImageNet, fine-tuned accuracy on Depth/VPT-basic tasks, and the extent to which their attribution maps localize the target objects in VPT. Given the dramatic failure of DNNs to solve *VPT-Strategy*, this finding implies that DNNs rely on a and brittle strategy of memorizing object features like size and position to solve VPT.

Architecture	Model	Versions
CNN	AlexNet (Krizhevsky et al., 2012b)	1
	ConvMixer (Trockman & Kolter, 2022)	3
	ConvNeXT (Liu et al., 2022)	10
	DenseNet (Huang et al., 2018)	4
	DLA (Yu et al., 2019)	5
	DPN (Chen et al., 2017)	6
	EfficientNet (Tan & Le, 2020)	4
	GhostNet (Han et al., 2020b)	1
	HRNet (Sun et al., 2019)	8
	LCNet (Cui et al., 2021)	3
	MixNet (Tan & Le, 2019)	4
	MnasNet (Tan et al., 2019)	3
	MobileNet (Howard et al., 2019)	14
	RegNet (Radosavovic et al., 2020)	6
	Res2Net (Gao et al., 2021)	5
	ResNet (He et al., 2015)	26
	ResNeSt (Zhang et al., 2020)	3
	RexNet (Han et al., 2020a)	5
	ResNext (Xie et al., 2016)	2
	SPNASNet (Stamoulis et al., 2019)	1
	TinyNet (Han et al., 2020c)	2
	VGG (Simonyan & Zisserman, 2014)	14
Transformer	BEiT (Bao et al., 2021b)	9
	CAFormer (Yu et al., 2022b)	6
	CaiT (Touvron et al., 2021b)	3
	ConViT (d'Ascoli et al., 2021)	3
	CrossViT (Chen et al., 2021a)	2
	DaViT (Ding et al., 2022)	3
	DeiT (Touvron et al., 2021a)	12
	EfficientFormer (Li et al., 2022b)	7
	EVA (Fang et al., 2023)	9
	FocalNet (Yang et al., 2022)	6
	LeViT (Graham et al., 2021)	5
	MaxViT (Tu et al., 2022)	6
	MobileViT (Mehta & Rastegari, 2022)	3
	MViT (Li et al., 2022a)	3
	PiT (Heo et al., 2021)	8
	PVT (Wang et al., 2022)	7
	Swin (Liu et al., 2021)	16
	Twins-SVT (Chu et al., 2021)	5
	ViT (Dosovitskiy et al., 2021b)	36
	Volo (Yuan et al., 2022)	7
	XCiT (El-Nouby et al., 2021)	6
	PoolFormer (Yu et al., 2022a)	8
Hybrid	CoaT (Xu et al., 2021)	7
	CoAtNet (Dai et al., 2021)	8
	EdgeNeXt (Maaz et al., 2022)	1
	Visformer (Chen et al., 2021b)	2
Foundation	Depth Anything (Yang et al., 2024)	1
	DINOv2 (Oquab et al., 2023)	1
	iBoT (Zhou et al., 2021)	1
	MAE (He et al., 2022)	1
	MiDas (Ranftl et al., 2022)	1
	SAM (Kirillov et al., 2023)	1
VLM	ChatGPT4 (Achiam et al., 2023)	1
	Gemini (Team et al., 2023)	1
	Claude 3 (Anthropic, 2024)	1
Diffusion	Stable Diffusion 2.0 (Rombach et al., 2021)	1

Table 1: **The 327 DNN models used in our study.**

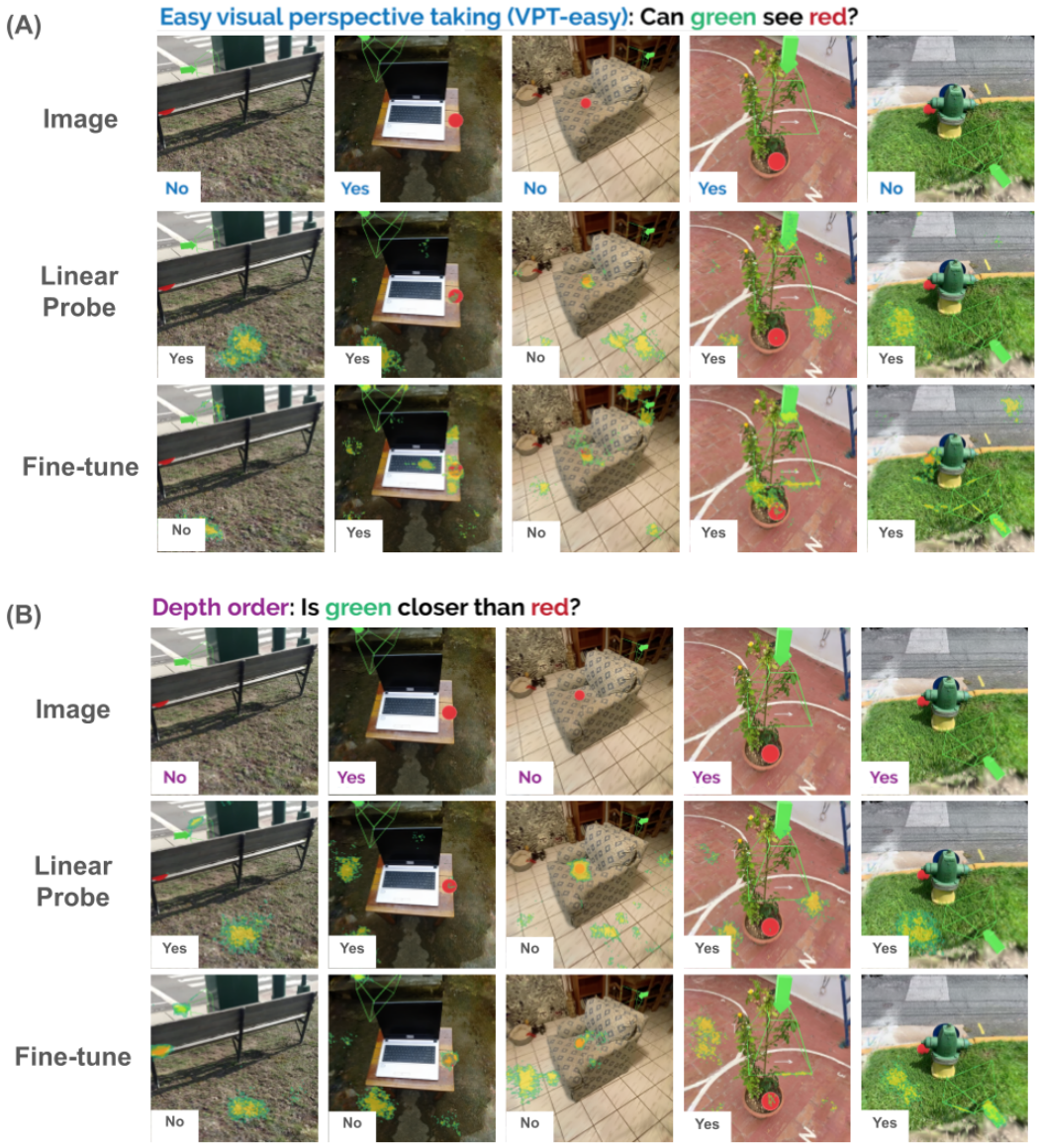

Figure 11: **Examples of decision attribution maps from a `ViT large`.** **(A)** Attribution maps from linear probed and fine-tuned DNNs on VPT-basic. **(B)** Attribution maps from linear probed and fine-tuned DNNs on depth order.