nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01759-7

Perception of 3D shape integrates intuitive physics and analysis-by-synthesis

Received: 9 January 2023

Accepted: 12 October 2023

Published online: 23 November 2023

Check for updates

Ilker Yildirim[1,2,3,7] ✉, Max H. Siegel[4,5,7] ✉, Amir A. Soltani[4,5,7], Shraman Ray Chaudhuri[6] & Joshua B. Tenenbaum[4,5] ✉

Many surface cues support three-dimensional shape perception, but humans can sometimes still see shape when these features are missing—such as when an object is covered with a draped cloth. Here we propose a framework for three-dimensional shape perception that explains perception in both typical and atypical cases as analysis-by-synthesis, or inference in a generative model of image formation. The model integrates intuitive physics to explain how shape can be inferred from the deformations it causes to other objects, as in cloth draping. Behavioural and computational studies comparing this account with several alternatives show that it best matches human observers (total $n = 174$) in both accuracy and response times, and is the only model that correlates significantly with human performance on difficult discriminations. We suggest that bottom-up deep neural network models are not fully adequate accounts of human shape perception, and point to how machine vision systems might achieve more human-like robustness.

For more than a century, vision scientists have studied the many cues that humans or machines use to recover shape. Edges or bounding contours, gradients of shading or texture, stereo disparity and motion parallax are just a few of the cues that can be computed from the visible surface of an object and that can reliably indicate an object's three-dimensional (3D) shape across different views[1]. When available, surface cues effectively support shape perception in humans and machines. However, a set of recent studies[2–5] present a challenge to the classical cue-based theory of shape perception: even when a surface is obscured, humans can sometimes perceive shape, without directly seeing the object at all. Consider the synthesized images of cloth-covered objects in Fig. 1a,b, in which each object is completely occluded by a thin, cotton-like fabric draping over it. Although the draped shapes look very different from the comparison objects (shown in randomly chosen orientations), observers can nonetheless pick out which unoccluded airplane or chair matches the 3D shape of the appropriate occluded object (answers: Fig. 1b, top left matches bottom left; Fig. 1b, top left matches bottom right).

In this Article, we ask: how can people perceive object shape (and pose, size, category and so on) in these images, when all the classic visual cues are mostly or entirely absent? Even those image cues that are present may be highly misleading, as they belong not to the underlying object's shape but to the surface of the occluding cloth. Somehow we are able to interpret the shape of the cloth as an interaction between the underlying rigid object's geometry and the way the cloth drapes and deforms upon contact. While sculptors have long exploited this capacity of the visual system to depict human faces and figures, only recently have detailed behavioural studies provided convincing evidence that humans somehow 'undo'[3,5] the effect of the cloth to access the hidden object. Thus far, a computational account of vision that can explain shape perception, even in the absence of surface cues, remains absent.

There are at least two ways one might explain how people perceive the shape of draped objects, corresponding to two contemporary frameworks which each seek to advance beyond classical cue-based approaches to 3D shape. One possibility is that, instead of a relatively constrained set of interpretable, meaningful cues, often derived from

[1]Department of Psychology, Yale University, New Haven, CT, USA. [2]Department of Statistics & Data Science, Yale University, New Haven, CT, USA. [3]Wu-Tsai Institute, Yale University, New Haven, CT, USA. [4]Department of Brain & Cognitive Sciences, MIT, Cambridge, MA, USA. [5]The Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA. [6]Department of Electrical Engineering & Computer Science, MIT, Cambridge, MA, USA. [7]These authors contributed equally: Ilker Yildirim, Max H. Siegel, Amir A. Soltani. ✉e-mail: ilker.yildirim@yale.edu; maxs@mit.edu; jbt@mit.edu

https://doi.org/10.1038/s41562-023-01759-7

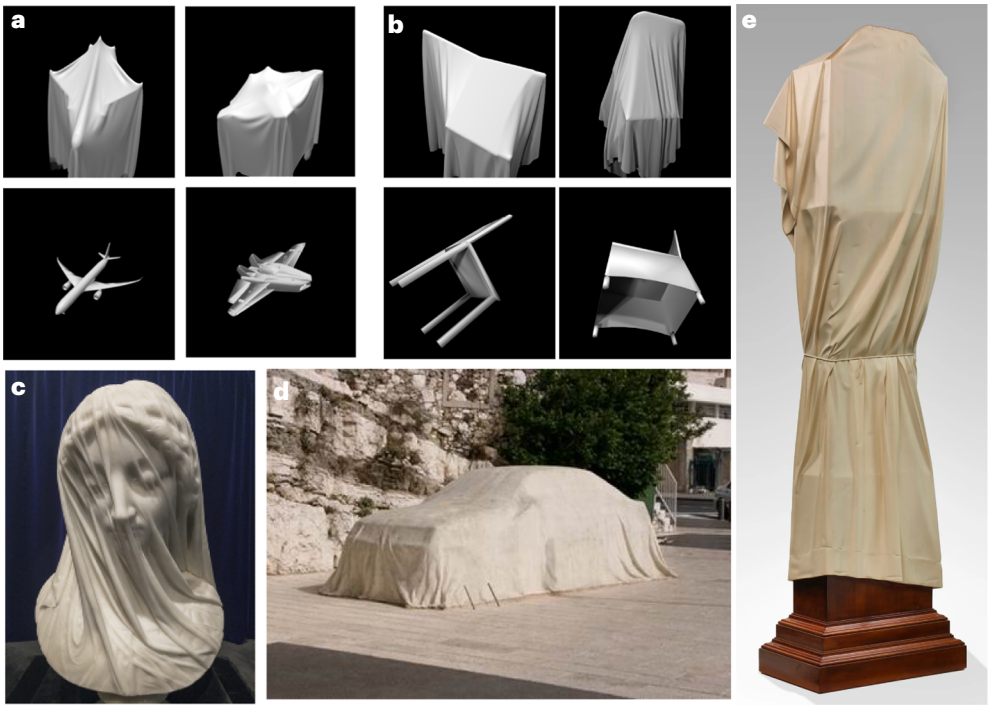

Fig. 1 | Seeing 3D shape through a cloth. a, Bottom row: two different airplanes. Top row: the same airplanes, draped with cloth and presented in random order and in a random pose. **b**, Same as **a** but for two different chairs. Despite the variation in viewpoint and complete occlusion of the cloth-draped objects, human observers can still match the cloth-draped and unoccluded object pairs (see note in main text for correct answer). **c**–**e**, Sculptors have long displayed their skill in works that are crafted from a single rigid material but convey both an illusory effect of cloth draping and rich 3D shape for the object under the illusory cloth, as in Giovanni Strazza's (c. 1850s) 'Veiled Virgin' (**c**, marble), Gabriel Klasmer's (2000) 'Car in the Sun' (**d**, fibreglass), or Wendell Castle's (1985) 'Ghost Clock' (**e**, mahogany, partially bleached).

an analysis of the geometry of objects and of image formation processes, the human visual system might engage a much larger set of cues that are obtained through some black-box learning mechanism (and are therefore difficult to write down or interpret). Recent computer vision models based on deep convolutional neural networks (DCNNs) have demonstrated learned feature hierarchies that facilitate impressive object recognition capabilities[6,7] and that are relatively robust to variation in appearance and pose even though the model training objective does not explicitly include these goals. Moreover, these same features have been shown to enable many other seemingly disparate visual tasks, including shape perception, with only minor adaptation (for example, fine-tuning, or adding one or a small number of additional output layers)[8]. Perhaps these features are sufficiently robust to generalize across even more extreme image transformations, such as cloth occlusion.

A second possibility is that we see 3D shape via 'analysis by synthesis', or inference in a physics-based generative model of how scenes form and give rise to images[9,10]. On this view, shape perception is not driven solely or primarily by a fixed, universal set of image cues, computed bottom-up from any image and sufficient for any downstream task. Instead, we infer 3D shape through a top-down interpretation process based on an internal model of how images are formed and the role that shape plays in that model. The generative model approach sees cloth draping as just one exemplar of a potentially unbounded space of atypical presentations of objects, in which some aspects of the physics of scenes and images grossly alter an object's appearance from its typical form while remaining easily interpretable by humans: consider as other examples viewing an object such as the chair or airplane in Fig. 1 outside in a rainstorm, or under 10 feet of cloudy water, or through coloured plastic wrap, or in the light of a full moon at night. The open-ended compositionality of the visual world may imply that it is difficult or impossible to specify or learn a single set of bottom-up image cues or features that reliably and robustly encode an object's 3D shape even in such atypical, rare conditions. Instead, a system should model individual, scene-level causes—the physical objects and processes that generate images—and how they combine and yield visual input. Then, by reversing their effects, it may recover the original physical scene. Thus a visual system could still identify a draped object and even perceive its fine-grained 3D shape if it were able to model and somehow invert cloth physics.

Our goal in this paper is to use the perception of objects under cloth as a case study to evaluate concrete versions of each of these accounts of shape perception. The theoretical merits of the pure bottom-up and top-down approaches have been extensively debated in the literature, but it has been difficult to find strong evidence distinguishing the bottom-up cue-based and top-down model-based approaches; until recently, neither discriminative classifiers nor Bayesian generative models performed well in realistic visual tasks, so comparisons with biological vision were limited to controlled scenarios with simplified, non-naturalistic stimuli[11]. Advances in algorithms and computing hardware, however, have led to DCNN and analysis-by-synthesis models that achieve good performance with complex natural images and can now be rigorously evaluated as models of how we perceive 3D shape in challenging cases with rich naturalistic stimuli. They also let us explore various hybrid accounts that so far have received very little direct evaluation in human psychophysics: in particular, we compare human judgements with top-down analysis-by-synthesis models attempting to match images at either the level of raw pixels or intermediate-level representations based on DCNN features.

To our knowledge, only two empirical studies[12,13] have compared modern neural networks with Bayesian models as accounts of human perception (although a number of papers evaluate either one or the other type of model; see for example refs. 14,15). For example, ref. 13 defined a compositional generative model of 3D shape and

compared human judgements of shape similarity with those derived from bottom-up classifiers and from top-down inference in a 3D generative model, concluding that the latter might underlie shape perception because it correlated better with human responses. But this study, while pioneering, provides only limited evidence for top-down analysis-by-synthesis in perception. The best bottom-up model was also quantitatively predictive of human judgements and performed almost as well as the top-down model. With more training and improved DCNNs, the quantitative gap between these models might be expected to narrow even further. In addition, neither model improved dramatically over a simple baseline matching test to target images at the pixel level. Here we demonstrate a stronger, qualitative distinction between model classes enabled by our completely occluded cloth-draped stimuli: standard DCNNs and pixel-based observers, unlike people and our generative models, perform no better than chance on harder instances and even with extensive specialized training show little improvement.

Our design choices offer several other advantages. Because we chose uncommon stimuli with variable difficulty, we find meaningful variance in human performance and response time, which allows for finer-grained model evaluations and comparisons of humans and models on trial-by-trial speed as well as accuracy. The generative model that we consider performs iterative inference with substantial stimulus-driven variation in the number of computational steps and therefore can be directly compared with subjects' reaction times, potentially revealing signatures and roles for feedback or recurrence in biological shape perception.

Results

The object-under-cloth task

While in some cases humans can recognize draped objects from a single image (for example, Fig. 1c,d), we chose as our experimental setting a visual match-to-sample task that allows us to directly address the above two possibilities. We choose this setting as we are primarily interested in how generative models can support online, detailed 3D shape perception, rather than object categorization or any kind of memory-based process. The essence of our proposal is that observers may perceive the 3D shape of cloth-covered objects in arbitrary orientations by approximately simulating in their minds the process of how cloth drapes over the object in three dimensions, and imagining what the resulting 2D image would look like. So we constructed an experimental task that should be directly solvable via this mechanism: we show observers an unoccluded matching object along with a target draped shape (that is, the initial matching object rendered in computer graphics under a simulation of cloth draping) and an unoccluded distractor object, in a two-alternative forced choice task. We call this the 'occluded' condition to contrast with a control 'unoccluded' condition (see below).

We chose ten different everyday object categories (airplane, bicycle, bus, car, chair, guitar, motorcycle, pistol, rifle and table) and sampled object meshes for each category from a large repository of 3D shapes[16]. We used 24 unique exemplars from each category, yielding 120 visual-matching trials; each trial used two shapes, and each unique shape appeared in one trial. Each trial presented an unoccluded target shape, a distractor shape and the target shape after cloth draping. We varied the similarity between the distractor and matching items (Fig. 2a, right) to generate visual-matching triplets ranging in difficulty. In half of the trials, the target and distractor objects were drawn from the same category; these we term 'harder' trials because same-category shapes are generally more difficult to distinguish than different-category objects, which we call 'easier'.

To create the cloth-occluded stimuli, we simulated cloth draping via a particle-based physics engine[17]; we chose simulation parameters (for example, number of iterations) and the mechanical and material properties of the simulated cloth (for example, stiffness and mass) to enable efficient, stable simulation of natural-looking cotton-like cloth (Methods).

In the unoccluded condition, we use the same objects but show the target shape without cloth (Fig. 2b). In this version of the task, viewpoint variability and the shape similarity between the matching and distractor test items are the only confounding variables.

Physics-based analysis-by-synthesis

We formalize the problem of matching a cloth covered object with its unoccluded counterpart as approximate Bayesian inference in a causal generative model. Our physics-based analysis-by-synthesis (PbAS) method combines physics and graphics knowledge with statistical inference and optimization. The model consists of three components: a generative model for scenes and images, feature extraction for approximately Bayesian inference (using a pseudo-likelihood approach) and a simulator-in-the-loop inference engine based on Bayesian optimization. As an account of how people can perceive the shapes of objects under cloth (or other challenging viewing conditions), we posit that each of these three components has some analogue in the mind and brain, and that they operate and interact in something like the ways we specify here—not precisely as we have implemented them, but close enough that the speed and accuracy characteristics of the PbAS model can be quantitatively compared with human behaviour, along with different model variants and alternative accounts.

The generative model in PbAS captures the physical scene variables, including object shape and pose, cloth properties and the mechanics of how they interact, which together produce the geometry of the occluding cloth surface. It further includes a model of graphics—how surface geometry, material and light interact to generate an image (some factors, such as optics, are handled implicitly; see, for example, ref. 18 for an explicit treatment, including modelling, of human representation of visual scene geometry). Given a hypothesized 3D object shape in a hypothesized pose, the model produces a synthesized or hypothetical image which may be compared with the image actually observed. In the analysis-by-synthesis framework, perception requires inverting this process to recover the object shape and pose likely to have given rise to the observed image (Fig. 3a,b).

Like most generative models, PbAS is too complex to invert exactly. A ubiquitous approximation algorithm, Markov chain Monte Carlo (MCMC), iteratively constructs samples from a target distribution like the posterior, but in our case requires far too many iterations to work because each step includes costly physics simulation. We sought instead to maximize the posterior using Bayesian optimization[19], or BayesOpt, which relative to MCMC provides a guided inference scheme where the next scene hypothesis to evaluate is informed by all (instead of only the current) evaluations of the posterior function[20]. BayesOpt simultaneously estimates and optimizes the posterior, providing an algorithm that efficiently samples increasingly more probable hypotheses for object shape and rotation given an input occluded image (Fig. 3b). The likelihood of a scene hypothesis is computed by comparing its corresponding rendered hypothesis image with the input, using a feedforward feature hierarchy f_{enc} implemented as the first fully connected layer of a pre-trained DCNN[7] (see the next section for a discussion of this choice). While the goal of inference in our model is posterior probability maximization, the optimization trajectory is also of interest for comparison with human behaviour.

Psychologically, BayesOpt can be seen as implementing a kind of goal-conditioned mental imagery (also see ref. 21 for an application in the context of mental rotation), in which the model interprets a target object (for example, a cloth-draped image) in the context of test objects (Fig. 3a,b), to determine which test object results in a mental image that better matches the target. PbAS can arrive at a reasonable percept rapidly (Fig. 3c,d) compared with sampling-based methods like standard MCMC. It therefore provides a more plausible quantitative standard for understanding average human accuracy, how accuracy improves with longer viewing time, and stimulus-driven variability in response time.

https://doi.org/10.1038/s41562-023-01759-7

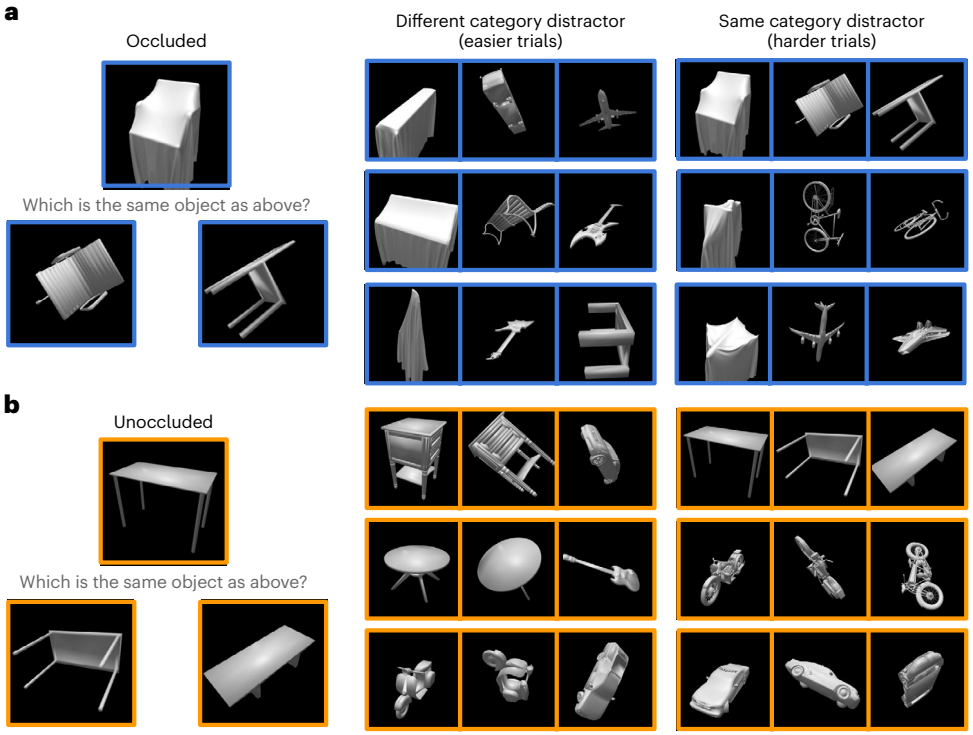

Fig. 2 | Matching a target shape to one of two unoccluded test objects. **a**, Left: a trial in the occluded task condition. The top image shows a 'target item', the bottom-left image shows a 'matching test item' and the bottom-right image shows a 'distractor test item'. Right: trials from the occluded task. Each triplet displays, from left to right, target item, matching test item and distractor test item for one trial. We show 'easier' trials, with different-category distractor and matching test items, and 'harder' trials, where both test items are of the same category. **b**, Left: a trial from the unoccluded task condition, spatial configuration as in **a**. Right: each triplet shows a trial from the unoccluded task, showing instances of easier and harder trials.

Synthesis: generative model. The generative model consists of (1) latent variables describing the scene: a 3D object shape S and its rotation R; (2) a forward physics simulator along with cloth parameters: cloth size, position, stiffness, mass and friction, denoted f_ψ; and (3) a rendering function and lighting parameters, together denoted f_Γ. We set the physics simulation parameters f_ψ and renderer parameters f_Γ to the same values as used for stimuli generation (see 'Solving the object-under-cloth task using the model' section in Methods). While the model is designed to perceive cloth-covered objects, it applies to unoccluded objects, as in the unoccluded task condition, as a special case by setting f_ψ to the identity function.

Given an occluded input observation (indicated as 'Input' in red frame, Fig. 3b) and an unoccluded 'context object' (in blue frame, Fig. 3b), we wish to estimate the object shape S and rotation R that best explains the input image. More formally, we wish to invert the generative model to find scene hypotheses that explain perceptual input using Bayesian inference, which amounts to finding the posterior

$$\begin{aligned}\Pr(S,R|I_{\text{obs}}) &\propto \Pr(I_{\text{obs}}|S,R,f_\psi,f_\Gamma)\Pr_u(S)\Pr(R)\delta_{f_\psi}\delta_{f_\Gamma}\\ &= \Pr(I_{\text{obs}}|I_{\text{hyp}})\Pr_u(S)\Pr(R),\end{aligned} \quad (1)$$

where $\Pr(I_{\text{obs}}|S,R,f_\psi,f_\Gamma)$ is a likelihood term induced by the physics engine f_ψ and rendering function f_Γ, and the delta functions select fixed physics f_ψ and rendering f_Γ parameters. For brevity, in the equality in equation (1) and below we write $I_{\text{hyp}} = f_\psi(f_\Gamma(S,R))$ for the hypothesis image given latent scene parameters and suppress the delta notation. We next explain each term.

The context object allows observers to form a distribution $\Pr_u(S)$ over the possible shape of the draped object; because the context object is presented as a 2D rendering, its shape is uncertain. Even though human observers do not need auxiliary shape information to process cloth-occluded images, this accompanying context object provides a computationally tractable shape hypothesis space for generative modelling. In this way, PbAS can be thought of as searching over a limited hypothesis space of shapes; for future work towards relaxing this constraint, see Discussion. We represent this shape uncertainty $\Pr_u(S)$ using a categorical distribution over the K nearest neighbours of the actual context object (excluding the ground-truth context object from the hypothesis space) in a large repository of shapes (the ShapeNet dataset[16]; Fig. 3b). In our simulations we take $K = 4$ and each neighbour is assigned a probability as a function of its distance rank (Methods). We place a uniform prior over rotations $\Pr(R)$ covering the half-sphere centred at canonical pose.

Executing the physics simulator f_ψ with a scene hypothesis (a sampled shape S and its rotation R) results in a draped cloth geometry G (Fig. 3b). Passing the resulting scene to the rendering function f_Γ in turn yields an image $I_{\text{hyp}} = f_\Gamma(G)$ of the cloth-draped object (Fig. 3b) that is a hypothesis image that may be compared with an input observed image I_{obs} to evaluate its likelihood under the scene hypothesis.

The detailed geometry resulting from cloth simulation (for example, the particular pattern of wrinkles) can vary substantially with even small changes in the values of random variables[17]; therefore, calculating an accurate likelihood (through marginalization) for any scene hypothesis is computationally intractable[20]. As a result, we define a pseudo-likelihood function $\Pr(I_{\text{obs}}|I_{\text{hyp}})$ based on the distance $D(I_{\text{obs}}, I_{\text{hyp}})$ between the input and hypothesis images in a suitable feature space arising from an encoder $f_{enc}(\cdot)$; here, we set $D = \ell_1$ and adopt the features computed by the first fully connected layer of AlexNet[7] as the encoder f_{enc}. The PbAS (pseudo-)likelihood for an input image, given a hypothesis image rendered from a scene proposal, is then

$$\Pr(I_{\text{obs}}|I_{\text{hyp}}) \propto \exp\left(-\left\|(f_{\text{enc}}(I_{\text{obs}}) - f_{\text{enc}}(I_{\text{hyp}})\right\|_1\right)$$

https://doi.org/10.1038/s41562-023-01759-7

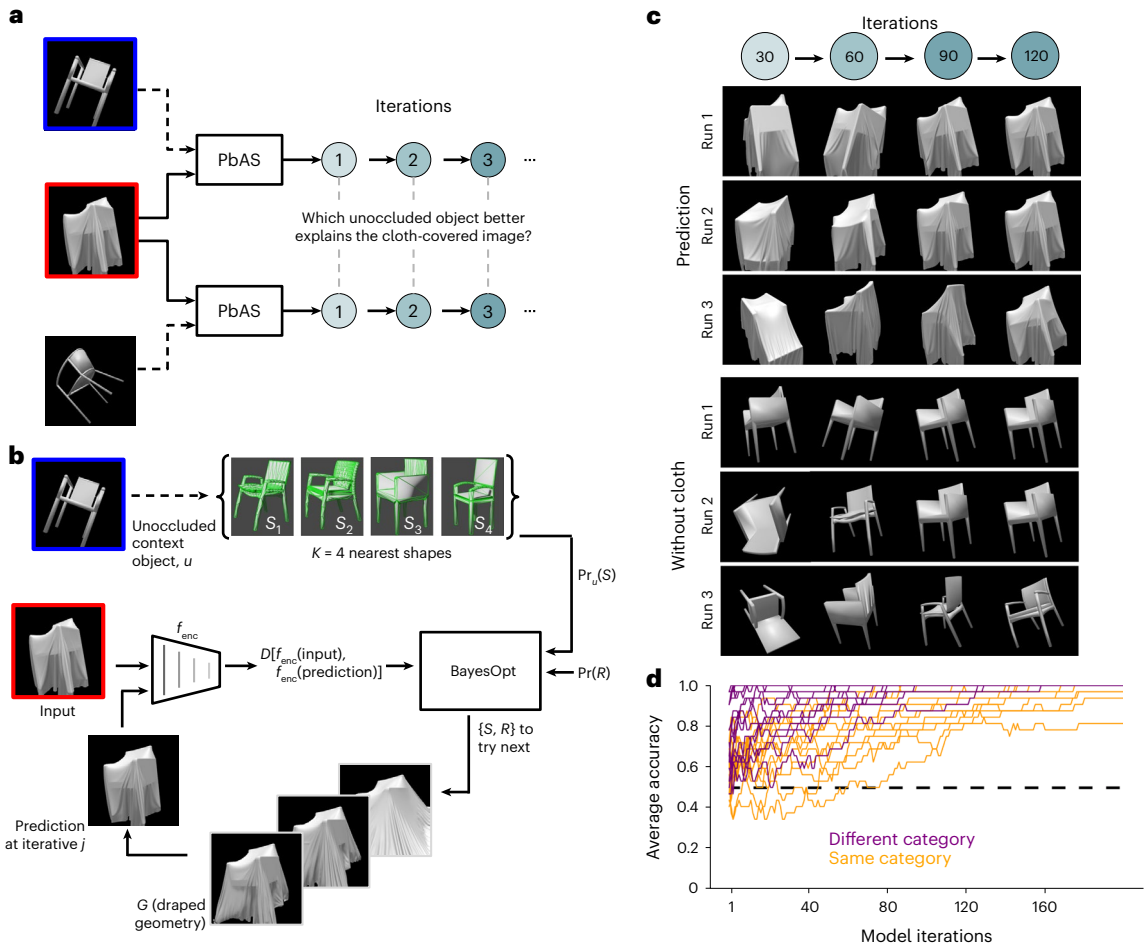

Fig. 3 | Overview of PbAS. a, Application of the PbAS model to solve the object-under-cloth task. Given an image triplet, two PbAS models are run in parallel; each execution takes as input a test item and the target item. On each iteration, the two executions of the model are compared to determine how well the target item is explained by each test candidate. **b**, Interpreting an input cloth-covered image (red frame), with a context unoccluded object (u; blue frame) supplying a prior over object shape $\Pr_u(S)$. Bayesian optimization (BayesOpt) efficiently guides inference, improving shape (S) and rotation (R) hypotheses across iterations. S and R proposals initialize the cloth draping simulation, then are evaluated by computing the distance D between the current scene hypothesis (rendered to a hypothesis image) and the input in a suitable feature encoding space f_{enc}. **c**, Visualization of three inference (**b**) trajectories over time. Rows are independent runs of PbAS each with input as in **a** and **b** (red frame) and show the cumulative best scene hypothesis at each iteration. Blocks show hypotheses visualized with (top; 'Prediction') and without (bottom; 'Without cloth') cloth occlusion. Model estimation accuracy improves with increasing iteration number, but some uncertainty remains as the model (like people) cannot in general perfectly identify the shape or pose of a draped object. **d**, Evolution of model accuracy averaged across multiple runs in the occluded task condition. Model predictions by iteration for 15 same-category 'harder' and 15 different-category 'easier' trials.

With these choices for the prior and likelihood, the posterior $\Pr(S, R|I_{obs})$ of equation (1) depends only on two terms: the discrepancy $\Pr(I_{obs}|I_{hyp})$ between the observed image and the rendered latent parameters, and the uncertainty $\Pr_u(S)$ over the shape of the context image.

By measuring the discrepancy between rendered scene hypotheses and observed images in terms of DCNN encoder-based features, the PbAS model as described is an instance of a hybrid top-down/bottom-up (or model-based/cue-based) approach to 3D shape recovery (see also ref. 22). We also consider a purely top-down analysis-by-synthesis approach that is identical except that image discrepancies are computed in terms of raw pixel deviations. The likelihood is then simply

$$\Pr(I_{obs}|I_{hyp}) \propto \exp\left(-\left\|I_{obs} - I_{hyp}\right\|_1\right)$$

We refer to this alternative as the 'Pixel Likelihood PbAS' model, or 'Pixel-PbAS' for short.

Analysis: inference using Bayesian optimization. The posterior $\Pr(S, R|I_{obs})$ of equation (1) contains all information that our model extracts from an observed image I_{obs}, but computing this distribution is intractable. Standard simulation-based inference methods based on MCMC ensure eventual convergence to the full posterior but in practice spend too many iterations in low-probability regions[20]. We focus instead on the maximum a posterior setting: finding the best single scene interpretation rather than the full posterior over all possible latent variable settings. Following previous work in simulation-based inference[23–25], we employ Bayesian optimization (or BayesOpt[19]); unlike gradient-based algorithms, BayesOpt allows us to optimize functions that include procedures, such as our scene renderer, which do not expose or do not support gradients. See Methods for an overview and details of BayesOpt applied in PbAS and Supplementary Fig. 1 for an illustration of the internals of this process.

Solving the object-under-cloth task using the model. Human participants see two unoccluded context objects (that is, test items) and one

target object on each trial. Recall that, by its design, the PbAS model interprets a target (that is, cloth-draped or unoccluded depending on experiment condition) object in the context of an unoccluded object. Thus, to model a given trial, we form two pairs, each consisting of a context object (either the matching item or the distractor item) and the target object, and apply PbAS to each pair (Fig. 3a). Each PbAS run aims to explain the same input image, but with different shape hypotheses derived from either the matching object or the distractor object. At every iteration, we save the current best parameter estimates (that is, shape and rotation) and the log posterior score for that scene hypothesis. Using the odds ratio decision rule, we obtain the model's best estimate of the underlying shape for each inference step.

We ran the PbAS model 32 times on each trial, for 200 iterations each, and treated each of these runs as a simulated participant (although with finer temporal resolution). At each of the 200 iterations, we averaged the binary decisions across runs to obtain mean accuracy predictions—that is, simulating the accuracy of participants' average shape choices. In our analysis we compare the dynamics of model choice with human decisions sampled at three different time intervals, corresponding to three different presentation durations that varied across experimental conditions (see the 'Inference dynamics explain human accuracy and response times' section for comparisons of models and human behaviour). Figure 3d shows how the average model performance changes as a function of iteration for a subset of our stimuli.

Bottom-up models based on DCNNs. To help evaluate the PbAS model and its correspondence with human perception, we considered several well-studied bottom-up models as comparisons for human and model performance. Recent computer vision models based on DCNNs learn powerful visual feature hierarchies achieving state-of-the-art object recognition performance. These feature hierarchies are relatively robust to variation in pose and lighting, can predict certain aspects of variance in neural and behavioural data, and are considered the 'current best models of the primate visual stream'[26]. Moreover they are useful for visual tasks beyond object recognition; these features have been used for a number of other vision problems, such as object localization and pose estimation[27], among others, with minor or no modification. In testing these pre-trained models, our goal is not to establish whether DCNNs, considered as a model class, can perform the object-under-cloth task. DCNNs are universal function approximators; with enough data, enough compute, and the right architecture and optimization procedure, they are probably able to learn to perform our visual-matching task. Instead, our goal is to assess whether the features learned from categorizing objects in natural scenes can suffice to perceive cloth-occluded shapes as well.

Because our synthesized stimuli and task design differ from those used for the pre-trained DCNNs, we also test the same networks after fine-tuning them using images similar to our experimental stimuli. We tested the following architectures, each pre-trained using ImageNet[28]: AlexNet[7], ResNet-50 (ref. 29) and VGG16 (ref. 30). Each DCNN was fine-tuned separately for the cloth-occluded and unoccluded conditions. The task was the same visual matching problem presented to humans: given an image containing two unoccluded test shapes and one target object (a 'triplet'; objects sampled from a total of 50 shapes), determine which test shape corresponds to the target. We repeated this process 32 times; thus we fine-tuned 32 copies (to match the number of PbAS runs per trial) of each architecture for each occlusion condition. We report the average accuracy of these 32 fine-tuned networks. For dataset generation, fine-tuning and evaluation procedures, see Methods.

For both the pre-trained and fine-tuned conditions, we found that no architecture was more accurate than AlexNet (Supplementary Fig. 2). Therefore, we use both the pre-trained AlexNet and our fine-tuned variant in our comparisons of bottom-up models with behaviour.

Inference dynamics explain human accuracy and response times

To evaluate PbAS as a candidate model for human perception, we compared its predictions on the object-under-cloth task with two key behavioural measures: average accuracy and response time. We recruited human subjects and assigned them to either the occluded or unoccluded condition (Fig. 2a,b, left). Participants were also divided into three presentation time conditions: the two fixed (1 or 2 s) time conditions and the unlimited time condition, which presents stimuli until subjects respond. In total, the experiment consisted of 2 occlusion × 3 presentation time = 6 conditions in a between-subjects design. We chose a between-subject design, rather than a within-subject design, primarily because an experiment that tested multiple conditions would require more time and focus from participants than is feasible, in our experience, for an online experiment. Nevertheless, we note that our design offered sufficient power for our statistical analyses; as shown in Supplementary Fig. 3, nearly all of our pairwise behavioural comparisons were statistically significant.

As is typical in modelling studies, we compared the average accuracy of PbAS and alternative models with that of humans. Because accuracy measures alone might simply favour models that are more performant, we also examined how PbAS 'response times'—the number of inference iterations used per trial—might explain human response times on the same trials.

Explaining human accuracy across presentation times. We first established that behavioural performance is significantly affected by task setting. While participants performed well above chance across all occlusion and presentation time conditions, their performance varied with respect to these design parameters. Most obviously, human performance was better in the unoccluded setting, $t(72) = 6.868$, $P < 0.001$. All statistical tests reported in this article are two-tailed, including both the parametric and nonparametric variants. With longer presentation time, average performance significantly improved (1 s versus 2 s $t(51) = -3.187$, $P = 0.002$; 2 s versus Unlimited $t(52) = -3.049$, $P = 0.003$, Supplementary Fig. 3a; for results broken down by occlusion condition, see Supplementary Fig. 3b) and response times increased (1 s versus 2 s $t(51) = -5.616$, $P < 0.001$; 2 s versus Unlimited $t(52) = -4.121$, $P < 0.001$, Supplementary Fig. 3c; for results broken down by occlusion condition, see Supplementary Fig. 3d). We also note that we did not observe any learning effects throughout the experiment, with participants' average performance remaining fairly constant across trials (Supplementary Fig. 6).

The design of our behavioural experiment offers a multifaceted view of human performance in terms of presentation time, trial difficulty (defined as whether test items are of same or different category; Fig. 2) and occlusion condition. In Fig. 4a, we show average human accuracy levels for each presentation time, pooled with respect to the two difficulty types ('different category' versus 'same category') and two occlusion conditions (occluded versus unoccluded). Observers performed significantly above chance even in the most challenging setting with cloth occlusion, same-category distractors and the briefest presentation time (1 s). Note also that performance improved with longer presentation time in the same-category distractor trials where, unlike the easier different-category distractor case, performance does not reach ceiling even with unlimited presentation time. We now ask whether PbAS and bottom-up models can explain these nuanced results.

We compared average human accuracy levels for each presentation time condition (collapsing over occlusion and difficulty) with PbAS accuracy at each model iteration. The comparison used the ℓ_2 distance. We found that the longer the presentation time, the more model iterations are needed to best match behaviour: the fit for 1-s data requires fewer iterations (48[41, 54], where $[l, u]$ indicates lower/upper 95% confidence intervals based on bootstrap resampling of participants) than

https://doi.org/10.1038/s41562-023-01759-7

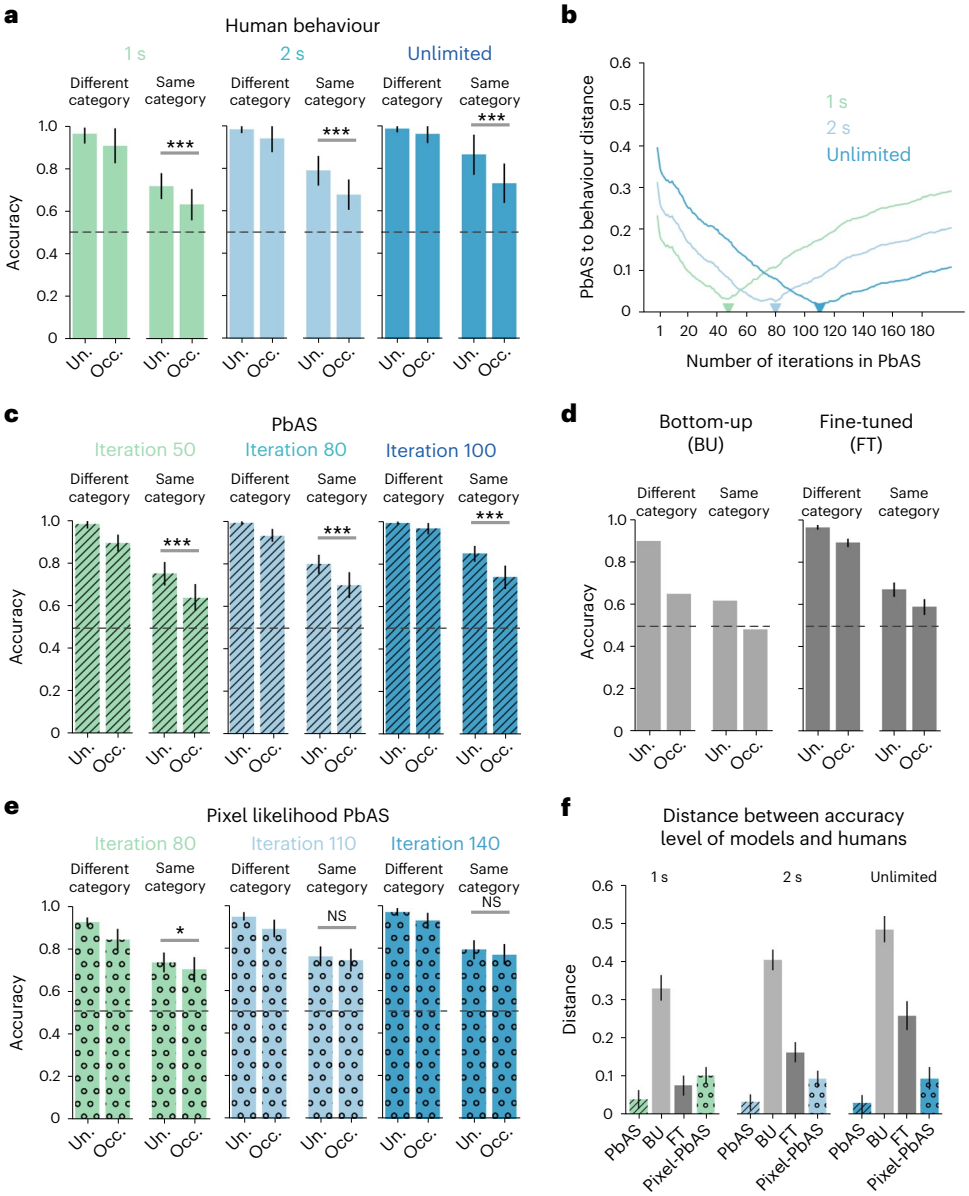

Fig. 4 | PbAS explains how human accuracy increases with longer stimulus presentation time. a, Average behavioural accuracy for each presentation time, occlusion condition and difficulty (independent samples of participants for each combination of presentation time and occlusion condition: 1-s Unoccluded $n = 29$; 1-s Occluded $n = 24$; 2-s Unoccluded $n = 30$; 2-s Occluded $n = 25$; Unlimited Unoccluded $n = 28$; Unlimited Occluded $n = 25$). A trial is said to be hard if the distractor test item is of the same category as the target item, and easy otherwise. **b**, Divergence between model and human performance at each model iteration. Coloured lines show ℓ_2 distance between PbAS model and human accuracy in indicated presentation time condition. Human accuracy at each increasing presentation time is best matched by model at correspondingly greater iteration (coloured triangles). **c**, Average accuracy of the PbAS model at the three iteration numbers chosen to be close to the best-matching iterations marked by coloured triangles in **b**. We show results for 50 rather than 48 iterations (see text). Evolution of PbAS accuracy levels over these snapshots closely matches human accuracy levels at the corresponding presentation times (compare **a**). **d**, Average accuracy of the bottom-up network (BU) and the fine-tuned (FT) variants; FT model reports ensemble average. Unlike humans and the PbAS model, in harder cloth-occluded trials with same-category distractors, the BU and FT models remain close to chance (dashed lines). **e**, Average accuracy of the 'Pixel-PbAS' model, a variant operating on pixels rather than intermediate perceptual features. Relative to the PbAS model, this model requires more iterations to reach human-level performance; more critically, it qualitatively misses a key aspect of behaviour by performing equally well across occlusion conditions, specifically in the harder same category trials (for the details of statistical comparisons, see text). Error bars in **a**,**c**,**d** (right plot) and **e** show standard deviation; significance in these panels is determined using two-tailed independent-sample t-tests. Samples in **c**,**d** (right plot) and **e** are independent runs of respective models ($n = 32$ in each case). **f**, Average of the bootstrapped ℓ_2 distance between human accuracy (**a**) and models: PbAS, bottom-up network pre-trained (BU) and after fine-tuning (FT), and PbAS without image encoding (using pixels for likelihood computation; 'Pixel-PbAS'). For the PbAS and Pixel-PbAS models, for each presentation time, we present the distances based on their corresponding best-matching iteration number. Error bars show 95% bootstrapped confidence intervals ($n = 5{,}000$ bootstrap samples); $^{***}P < 0.001$; $^{*}P < 0.05$; NS, not significant ($P > 0.05$).

are needed for the 2-s condition (80[63, 83]), and even more iterations (110[102, 131]) are needed to match the unlimited time data (Fig. 4b). The performance of the PbAS model at the best-fitting iteration numbers for each presentation time closely matches their corresponding behavioural accuracies (compare Fig. 4a,c, which shows model performance at iterations 50, 80 and 110 for simplicity; model accuracy levels at 48 and 50 iterations are essentially identical). In particular, the correspondence between PbAS and behaviour (measured as the ℓ_2 distance

https://doi.org/10.1038/s41562-023-01759-7

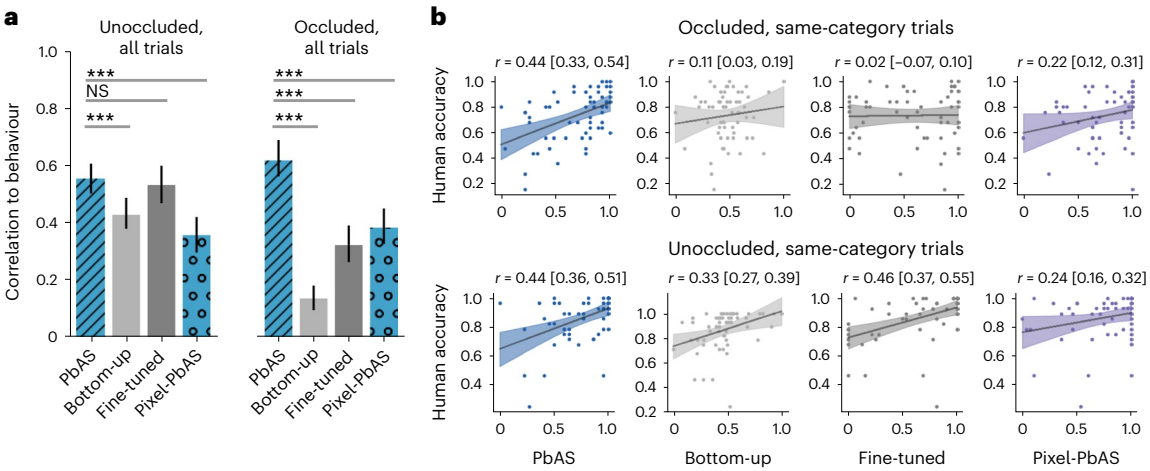

Fig. 5 | Fine-grained analysis of human accuracy at the level of individual trials in the unlimited time condition. PbAS explains behaviour better than alternative models. **a**, Average of the bootstrapped trial-level accuracy correlations between models and humans in the unlimited time condition. PbAS, bottom-up network pre-trained (BU) and after fine-tuning (FT), and the PbAS model without image encoding ('Pixel-PbAS'). The fine-tuned model reports ensemble average of multiple fine-tuned networks. Error bars show bootstrapped 95% confidence intervals (n = 5,000 bootstrap samples). Statistical comparisons are made using direct bootstrap hypothesis testing based on a two-tailed test threshold (***$P < 0.001$; NS, $P = 0.632$; all tests reported are two-tailed). **b**, The hardest, same-category trials reveal that only the PbAS model consistently correlates with behaviour in both the unoccluded and occluded conditions. The x-axis values are normalized to range between 0 and 1. Correlation coefficients are indicated on each scatter plot; bootstrapped 95% confidence intervals in brackets (n = 5,000 bootstrap samples). Shaded regions show 95% confidence intervals of the standard error in linear regressions (solid lines).

between behavioural and model accuracy levels) is stronger than it is for any other model (Fig. 4f, $P < 0.001$ using direct bootstrap hypothesis testing, except PbAS versus FT in the 1-s condition; see below).

Unlike the PbAS model, the bottom-up features derived from pre-trained DCNNs failed to explain human accuracy levels, nor did they after fine-tuning these networks separately for each occlusion condition (Fig. 4d,f). As expected, the performance of the pre-trained bottom-up network declined substantially under occlusion, but it did so even for the easier different-category distractor trials (Fig. 4d, Bottom-up (BU)). For the harder cloth-draped, same-category trials, the performance of the bottom-up model reduced to chance (Fig. 4d, Bottom-up (BU)). Fine-tuning this network improved its overall performance, but most of this improvement manifested in the different-category trials and indeed its performance remained near chance in the harder cloth-occluded trials with same category distractors (Fig. 4d, Fine-tuned (FT)). These results are reflected in the correspondence between human and network accuracy. In all but one condition, the discrepancy between bottom-up and fine-tuned models, and human behaviour, is higher than it is for PbAS (Fig. 4f; $P < 0.001$, direct bootstrap hypothesis testing, for each pairwise comparison of PbAS and other models). In the 1-s condition, the fine-tuned model is statistically inseparable from PbAS ($P = 0.055$), but it decouples from behaviour in finer-grained trial-by-trial analysis, as we explain in the next section (see also Fig. 5 and Supplementary Figs. 4 and 5).

Overall, unlike PbAS, the discrepancy between human and network accuracy levels increased with presentation time, suggesting the need for additional computations beyond the bottom-up processing implemented in these DCNN models (Fig. 4f; $P < 0.001$ for each pairwise comparisons of BU-1 s versus BU-2 s, BU-2 s versus BU-Unlimited, FT-1 s versus FT-2 s, and FT-2 s versus FT-Unlimited).

These results provide support for the role of top-down computations (the generative model) in the hybrid architecture embodied in PbAS: the DCNN feature hierarchies that alone cannot explain behaviour are useful when they guide inference (by defining the likelihood) in the generative model. Is this bottom-up component necessary to explain behaviour? We evaluated a model that removed the image encoding module. This ablation—referred to as the Pixel Likelihood PbAS (or 'Pixel-PbAS' for short)—computes likelihood in the pixel space, keeping everything else unchanged from PbAS. We found that this ablation fails to reproduce an important aspect of behaviour: unlike the PbAS model and human judgements, the Pixel-PbAS model performs equally well in the harder (that is, same category) occluded and unoccluded trials (Fig. 4e; humans: 1 s $t(22) = 4.467$, $P < 0.001$; 2 s $t(23) = 5.324$, $P < 0.001$; Unlimited $t(23) = 4.712$, $P < 0.001$; PbAS: 1 s $t(30) = 7.45$, $P < 0.001$; 2 s $t(30) = 7.24$, $P < 0.001$; Unlimited $t(30) = 9.14$, $P < 0.001$; Pixel-PbAS: 1 s $t(30) = 2.47$, $P = 0.016$; 2 s $t(30) = 1.42$, $P = 0.161$; Unlimited $t(30) = 1.93$, $P = 0.058$). Moreover, it takes longer to reach human level accuracy relative to PbAS, requiring about 30 more iterations for each presentation time condition (Fig. 4e and Supplementary Fig. 7). Finally, this model does not match behaviour as well as PbAS; using its best-fitting iteration numbers, the distance to behaviour is greater than that of PbAS at each presentation time condition (Fig. 4f; $P < 0.001$ using direct bootstrap hypothesis testing). However, we note that, unlike the bottom-up models, the distance from the Pixel-PbAS model to behaviour is constant or decreases slightly across presentation time conditions, indicating that the iterative refinement of scene hypotheses is still crucial to explain how behavioural performance improves with longer exposure times. These results establish that both top-down and bottom-up components of the PbAS architecture are needed to account for behaviour. Relative to the bottom-up models, PbAS's superior account of behaviour is not merely a result of its better task performance, but is instead due to its making similar perceptual judgements, and errors, as humans. The next two sections provide further evidence for these conclusions using fine-grained error and response time analyses.

Explaining trial-level human accuracy. Next, we evaluated the ability of the models to explain average human accuracy at the level of individual trials. In the unlimited time condition, we found that the trial-by-trial accuracy of the PbAS model at the best-fitting iteration (iteration 110, marked by the dark-blue triangle in Fig. 4b) correlated well with behaviour, and did so consistently in both occlusion conditions ($r = 0.55$, $P < 0.001$ and $r = 0.62$, $P < 0.001$ in unoccluded and cloth-occluded conditions; Fig. 5a). In the unoccluded condition, PbAS better correlated with behaviour relative to the pre-trained bottom-up network features ($P < 0.001$, using bootstrap direct hypothesis testing),

https://doi.org/10.1038/s41562-023-01759-7

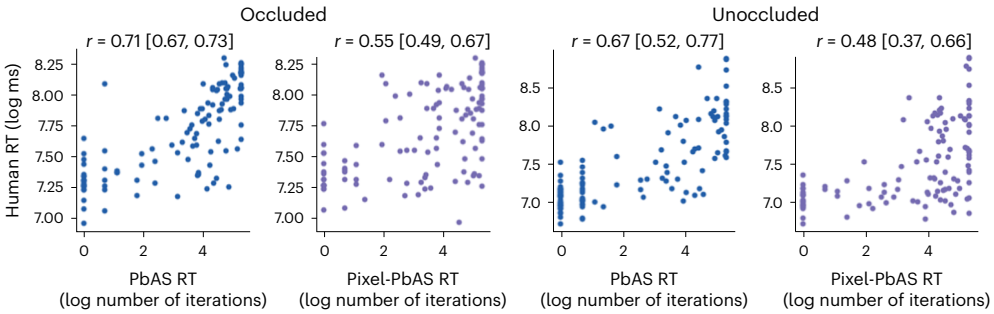

Fig. 6 | Trial-level response time comparisons. Trial-by-trial average human response times (RTs) (log-transformed milliseconds) are explained by the PbAS model (log-transformed number of iterations) based on a simple decision threshold (for details, see text). The PbAS model captures significantly more variance than the ablated Pixel-PbAS model in each occlusion condition ($P < 0.001$, using direct bootstrap hypothesis testing). For each comparison, the mean correlation and bootstrapped 95% confidence intervals (in brackets) are shown.

but fine-tuning was effective in closing the gap; PbAS and the fine-tuned model showed no difference ($P = 0.31$). However, in the occluded condition, the PbAS model better explained behaviour relative to both the pre-trained and fine-tuned alternatives ($P < 0.001$ for each comparison; Fig. 5a). PbAS also correlated with behaviour better than the Pixel-PbAS model in both the unoccluded and occluded conditions ($P < 0.001$; Fig. 5a; for qualitatively similar results in the other two presentation time conditions, see Supplementary Fig. 4) Despite the superior quantitative account of PbAS, we note that none of the models considered could explain all of the reproducible variance in the behavioural data. Split-half correlations across participants (Methods) in the unlimited presentation time condition were around $r = 0.80$, $P < 0.001$ for both occlusion conditions, significantly higher ($P < 0.001$) than the correlation achieved by PbAS.

What underlies the PbAS model's ability to consistently account for behavioural accuracy at the trial level across both occlusion conditions? We hypothesize that both top-down generative knowledge and bottom-up feature encoding are crucial. To address this, we first notice that in the easier, different-category trials humans performance is at ceiling, especially in the unlimited time condition (see the 'different category' bars in Fig. 4a). There is therefore little variance to explain in these easier trials. Thus, we focus on the difficult same-category trials where there is appreciable variance in behavioural accuracy across trials. We find that in these difficult trials, when compared with the bottom-up models, only PbAS can account for behaviour in both occlusion conditions (Unoccluded: $r = 0.44$[0.36, 0.51]; Occluded: $r = 0.44$[0.33, 0.54]). In the regular, unoccluded condition, the fine-tuned model (and to some extent the pre-trained model) can explain some of these fine-grained behavioural patterns, however, these models, especially the fine-tuned model, decouple from behaviour under cloth occlusion (Fig. 5b; FT: Unoccluded: $r = 0.46$[0.37, 0.55]; Occluded: $r = 0.02$[−0.07, 0.10]; BU: Unoccluded: $r = 0.33$[0.27, 0.39]; Occluded: $r = 0.11$[0.03, 0.19]). The Pixel-PbAS model also falls short of the performance of the full PbAS model in both occlusion conditions ($P < 0.001$ using direct bootstrap hypothesis testing; Fig. 5b; for qualitatively similar results in the other two presentation time conditions, see Supplementary Fig. 5), further demonstrating the necessity of both top-down generative knowledge and the bottom-up image embedding for successful prediction of behaviour. However, we note in these same-category trials, too, PbAS falls short of explaining the full extent of the reproducible variance in the data (split-half correlations in behaviour: $r = 0.79$, $P < 0.001$ and $r = 0.71$, $P < 0.001$ in the unoccluded and occluded conditions).

Explaining trial-level response times as iterative inference. Our analyses have so far focused on accuracy. Here we analyse human response times to ask whether the time course of inference in PbAS can explain the evolution of observers' perceptual decision-making at the level of individual trials—how long they decide to view a stimulus before making their choice. Thus, in the unlimited time condition, we compare the number of iterations required for the model to arrive at a decision on a given trial (in a given experimental condition) with the average human response time for that trial. To do so, we devised a simple decision rule in the model that applies to individual trials. At each model iteration, this decision rule compares the average model accuracy to a criterion set to the average participant accuracy within the trial's condition. We record the earliest iteration that PbAS performance exceeds that criterion (or the maximum iteration number, 200, otherwise) and take it as a predictor for that trial's average response time. This is akin to a drift-diffusion model[31] where evidence accumulation naturally arises from the iterative refinement of scene hypotheses in the PbAS model (notice that, unlike standard drift diffusion models, the drift rate and other parameters arise from model inference; no parameters are fit save the criterion). The results are response time predictions for each trial of each condition in the experiment.

Despite the simplicity of this decision rule, we found a remarkable correspondence between the number of iterations needed to solve a trial in PbAS and the time humans took to respond on that trial (Fig. 6); the relationship holds for both occlusion conditions (Unoccluded: $r = 0.67$[0.52, 0.77]; Occluded: $r = 0.71$[0.67, 0.73]). No parameters (beyond taking human performance as criterion for each condition) were fit to explain response times. Because the Pixel-PbAS model also performs iterative inference, we can test its ability to explain response time data as we did with PbAS. We found that PbAS gave a better account of response time data than the ablated model in each occlusion condition ($P < 0.001$ using bootstrap hypothesis testing; Fig. 6). We also explored measures of 'amount of processing' in the bottom-up networks (the pre-trained and fine-tuned models) as predictors of human response times, finding that PbAS better explains behaviour both qualitatively and quantitatively, in each occlusion condition (Supplementary Fig. 9).

Discussion

We presented evidence for the use of generative model computations in visual perception, in the form of physics-based mental simulations. Our behavioural results as well as recent related literature[2–4] raise a fundamental question: How is it possible to perceive the shape of an object when none of the classic visual cues to shape are visible? We proposed that the mind and brain exploit internal representations of the physical processes which form scenes and images. Our PbAS model incorporates knowledge of scene structure and dynamics to explain, through online optimization and physics simulation, why a cloth-covered object appears the way it does—as the result of dropping a cloth on an inferred shape in an inferred pose. We tested PbAS in a

https://doi.org/10.1038/s41562-023-01759-7

shape matching task that required subjects to match a cloth-draped object with its unoccluded (and randomly rotated) counterpart, in the presence of a distractor. The PbAS model predicts not only overall human accuracy in this visual matching task, but also how performance improves with longer stimulus presentation times. Crucially, the number of inference steps needed to reach a behaviourally determined performance threshold predicts, on a trial-by-trial basis, average participant response times.

Our work adds to the growing literature showing that perception in the brain can be understood as efficient approximate inference in generative models, or analysis-by-synthesis[9,13,32,33]. Past studies have examined some predictions of this theory, but have not provided quantitative evidence that such rich generative models—incorporating shape, object interaction dynamics and sensory features—are used online during perception. PbAS also differs from previously considered generative models in its focus on scene elements and causal processes, which when composed allow it to interpret images which are outside typical perceptual experience. In this way, our work identifies the flexible use of ad hoc dynamic scene properties in perception, such as cloth mechanics, that only indirectly influence image formation and are not usually seen as cues to 3D shape. Perceiving shape through cloth occlusion highlights how such 'nuisance' variables can play a central role in 3D object perception. Our work argues that the compositional use of generative models provides the best way of understanding how these factors influence perception.

Bottom-up models based on DCNNs performed poorly both in the object-under-cloth task and in mimicking human behaviour. A DCNN that has been fine-tuned on thousands of images of cloth-occluded objects produces behaviour with roughly similar average accuracy as humans in our briefest presentation conditions (1 s), but unlike the PbAS model fails to explain how performance improves with time and does not correlate at all with trial-by-trial accuracy in the most challenging conditions (occluded with cloth and same-category distractors, for all presentation conditions tested). DCNNs, as a model class, should in principle be able to learn any mapping from inputs to outputs, but our fine-tuning results show that, in practice, the data requirements can be substantial (and probably exceed human experience) and the best results far from human-like. Given the broader context of the many atypical, challenging viewing conditions that the visual system may encounter, these findings underscore the importance of generalization and robustness, ongoing challenges for DCNNs, and illustrate how top-down knowledge can enable perception in difficult novel contexts. Bottom-up models do, however, play an important role in our framework; relative to the ablated Pixel-PbAS model, the hybrid architecture implemented in PbAS demonstrates that powerful feature hierarchies can usefully facilitate or guide inference in generative models. This perspective is compatible with much research on 'core' object recognition showing the explanatory power of bottom-up models[34,35]. Future work should also evaluate continuing developments in DCNNs, trained using alternative loss functions, architectures or datasets, including, for example, contrastive objectives[36], vision-language models[37] and non-convolutional architectures[38], which may show improved generalization to difficult perceptual tasks (as explored in refs. 39,40). This is especially relevant for brief exposure times (for example, 1 s, Fig. 4d,f), where the fine-tuned model shows some initial alignment with behaviour in terms of average accuracy levels.

The PbAS model suggests that perception of cloth-covered objects in the brain relies on a combination of feedforward, feedback and recurrent computation. We believe that this is valuable as, relative to the case of feedforward processing, there is little evidence to constrain or generate hypotheses regarding the role of feedback and recurrent computation in visual scene analysis[41]. PbAS suggests a new computational goal for feedback and recurrence in the brain, which is in some ways related to pattern theory as expressed in ref. 10: such processing might implement the progressive unfolding of one or a number of physical simulations. It is likely that these forms of neural computation

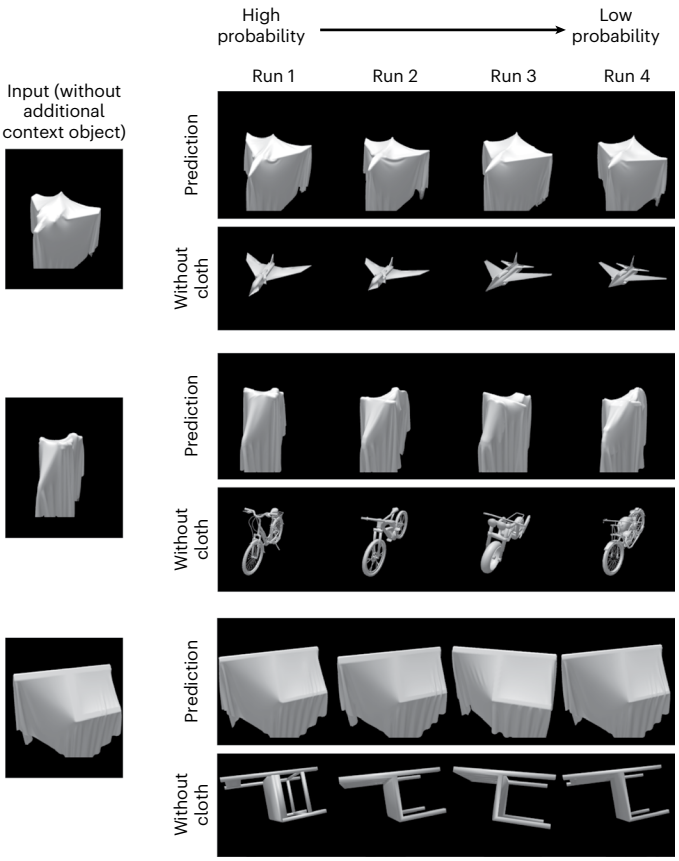

Fig. 7 | Seeing the shape of a single cloth-draped object, without the aid of unoccluded candidates (cf. Fig. 2a,b). By expanding its shape hypothesis space to contain a large set of category-specific objects (as opposed to the four nearest neighbours of the available context object as in our main model) and removing the unoccluded inference module, PbAS can obtain plausible estimates of 3D pose and geometry. Rows show (from left) input image containing target cloth-covered object and four inferred shape/pose hypotheses under this modified PbAS model, ordered from high to low posterior probability.

implement multiple computational goals needed for such diverse functions as attention, learning and perception[41]. The hypothesis suggested by PbAS—internal simulations of physical processes—is not exclusive of the others and future work should explore their combination.

The present implementation of PbAS accounts for behaviour in the specific matching task we studied here (Figs. 1a,b and 2). Future work should exploit its modular architecture to address other experimental paradigms and perceptual problems. For example, from an image of a single draped object (without a comparison unoccluded object), humans can often infer its category, approximate pose and partial shape (Fig. 1c–e). While evaluating PbAS in this more difficult scenario is beyond the scope of the present paper, the framework readily extends to this setting; for a demonstration from a proof-of-concept implementation, see Fig. 7. Another limitation of the current study is our treatment of scene parameters such as illumination and cloth material. For simplicity, we held these constant across all experimental stimuli, and the same parameters were incorporated into PbAS and Pixel-PbAS and used to generate training data for the neural network models. A more elaborated PbAS would attempt to estimate these scene properties as well (our use of relatively invariant feature embeddings, however, probably attenuates the effects of changes in environmental lighting). Varying these parameters when generating stimuli could also reveal how cloth material affects shape perception, an interesting direction for future work.

There is also room for improvement in matching human behaviour in the present matching task; even our best model, PbAS, falls

short of explaining all of the reproducible variance in people's judgements. Considering the models presented here, we find that the more performant a model, the better it correlates with human accuracy (Supplementary Fig. 8), suggesting that building performance-equated alternatives (for example, Pixel-PbAS) will help further refine the algorithmic basis of how physics-based mental simulations and shape inference are implemented in the mind and brain. PbAS relies on image features for approximate inference; because it is unlikely that a generative model can match sensory inputs exactly, good performance requires image representations that ignore irrelevant fine-scale variation but do covary with the visible 3D shape. To this end, we considered bottom-up features based on DCNNs, which are useful (as manifested in the comparisons of the PbAS and Pixel-PbAS models) but nevertheless biased towards local image patterns[42]. Future research should consider geometrically (for example, curvature regions[43]), topologically (for example, critical contours[44,45]) and physically informed feature layers (for example, physical contact[3]). In principle, models embedding such feature layers would not require an unrealistically accurate model of physics or graphics because these features, learned or designed to be robust to irrelevant image variation, might tolerate model mismatch.

Our results suggest that shape perception under cloth draping involves mental operations beyond the rapid, bottom-up processing believed characteristic of traditional object recognition[46]. To what extent are the computations hypothesized by PbAS—3D shape inference, mental rotation or mental simulations of physical and image formation processes—also engaged in rapid, automatic visual processing? How do they relate to other cognitive mechanisms supporting dynamic processing such as visual routines[47] and mental imagery[48]?

Recent psychophysical work suggests that these computations might be implemented in the visual system as part of spontaneous processing of sensory data. Ref. 5 studied cloth-covered object perception across a battery of visual tasks, finding evidence that scenes are rapidly and automatically parsed as the appropriate physical causes. In addition to behavioural probes, cognitive neuroscience can address where in the brain the computations specified by the PbAS model might be implemented (for a recent review of Bayesian causal models and inference in the brain, see ref. 49). For example, functional magnetic resonance imaging studies[50] have identified brain regions supporting intuitive physical judgements in a dorsal frontoparietal network; it is of significant interest to answer whether the same or similar brain regions are also recruited during the perception of cloth-draped objects.

Our work also raises questions about the origins of the generative models implemented in PbAS, including the prior over 3D shapes and the simulation of physical processes. A plausible hypothesis is that such generative models derive from a combination of innate capacities and experience-driven learning. Attempts to learn expressive distributions of shapes[51] and complex object dynamics[52,53], starting from certain helpful inductive biases implemented in neural network architectures or training objectives, suggest that such a hypothesis can be concretely and productively implemented and empirically evaluated (for examples, see refs. 54,55).

The PbAS framework discussed and supported here may play a broader role in visual processing beyond our cloth-draped object setting, unifying competencies beyond traditional shape and object perception. A common computational engine may therefore support perception of the dynamical properties of objects, such as the relative masses of colliding rigid bodies or single objects reacting to the application of external forces[56–58]; the stiffness of deformable objects undergoing natural transformations[59–61]; viscosity and flow of liquids[62–64]; and in general the perception of the physical (that is, non-intentional) causal history of an object[65–67]. In each of these cases, it is at least plausible that the brain uses generative models to simulate the physical processes that could have produced the observed scene, and compare the results of these simulations to the sensory input. A better understanding of how the brain supports these abilities could also lead to more robust, and more human-like, machine vision systems.

Methods

This study complies with all relevant ethical regulations and was approved by the Massachusetts Institute of Technology Institutional Review Board (the Committee on the Use of Humans as Experimental Subjects).

Generative model

Cloth simulations. We used the FLeX engine, a particle-based physics engine, for cloth physics simulation[17]. Simulation parameters as well as the mechanical-material properties of the cloth were chosen so as to achieve fast, stable simulation of natural-looking, cotton-like cloth. Simulation parameters were as follows: iterations, 4; subiterations, 19; particle radius, 0.0078; collision distance, 0.0078; shape collision margin, 0.00078; particle collision margin, 0.0; relaxation mode, default; relaxation factor, 1.3; drag, 0.09; damping, 0.0; dissipation, 0.0; restitution, 0.0. The mechanical-material properties of the cloth were as follows: strength stiffness, 0.8; bend stiffness, 0.64; shear stiffness, 0.4; particle mass, 1.0; static friction, 0.18; dynamic friction, 1.1.

To increase simulation efficiency, we simplified the geometry of the ShapeNet meshes using Blender[68]. First, we corrected the surface normals on each mesh by ensuring that they were consistent and pointed outwards. Second, we used Blender's 'Solidify' mesh modifier with the thickness parameter set to −0.0001. Finally, we merged faces that were adjacent and approximately coplanar (with surface normals differing by less than 0.02 rad ≈ 1.15°).

We initialized simulations by placing a square cloth (represented computationally with 210 × 210 particles) just above the geometric centre of the rotated object to be draped. We then ran the simulation for 150 steps, sufficient to fully drape all objects we tested. Each cloth simulation took between 3 s and 40 s on a NVIDIA 2080TI GPU, on the order of 1,000 times faster than alternative implementations using central processing unit-based cloth simulation and unsimplified meshes.

Image rendering. The scene was lit to minimize shadows. We placed 14 point lights with energy 0.5 on a sphere with radius 1.22 (object radius normalized to 1), with lights distributed approximately equidistant using the Fibonacci sphere algorithm. We rendered these scenes to 224 × 224 images using Blender's internal renderer.

To equate the texture appearance of the draped and unoccluded images, we replaced the optical materials associated with the original ShapeNet meshes with a diffuse material (diffuse colour 0.75 in each RGB channel, diffuse intensity 0.75 and specular intensity 0.07). We used a very similar material to render draped cloths (diffuse colour 0.8 in each RGB channel, diffuse intensity 0.8 and specular intensity 0.05). We reasoned that equating the texture appearance in this way would aid the bottom-up neural network models in emphasizing shape over texture[69].

Approximating shape distance. Given two shapes from ShapeNet S_i and S_j, we define a shape distance metric by (1) rendering each object in a standard canonical pose, (2) passing each image through a pre-trained AlexNet[7] and extracting feature activations at the first fully connected layer (that is, applying f_{enc} as for pseudo-likelihood evaluation during inference) and (3) evaluating the ℓ_1 distance between the feature activations for each shape. The resulting measure is similar to that used when calculating the pseudo-likelihood.

Shape prior. Given an unoccluded context object s_0, we modelled the observer's shape uncertainty $\Pr_u(S)$ as a categorical distribution over the $K = 4$ shapes nearest to s_0. Let d_{s_k} be the weight of the kth closest shape s_k to s_0; then $\Pr_u(S = s_k) \propto \exp(-d_{s_k})$ with $1 \le k \le K$. The Shapenet database forms a sparse approximation to the space of all object

shapes, and we found that the distance between an object and its closest neighbours could vary wildly; one reason is that some object classes have many more exemplars than others. Therefore, a prior defined solely using shape distance showed high variance across trials and was unsuitable for our purposes (for example, it induces arbitrary bias towards either the distracting or matching object from trial to trial). The unnormalized weights for each nearest shape were instead assigned on the basis of the rank order of their distance to the context object, starting at $d_{s_1} = 750$ and increasing at increments of 75 so that $d_{s_k} = 750 + (k-1)75$. The scale of these weights was chosen so that the relative contributions of the prior and likelihood were comparable.

Inference using Bayesian optimization

In comparison with traditional inference schemes based on random-walk MCMC[20], Bayesian optimization provides a more guided or 'active' approach to inference, where the next scene hypothesis to evaluate the posterior on is informed by all of the previous evaluations of the posterior. In adopting Bayesian optimization, we forego full posterior estimation (which MCMC can provide in principle) in favour of a good point maximum a posteriori estimate. This choice is further motivated by the computational cost of cloth simulation, which is responsible for nearly all of the work our model must do. BayesOpt requires many fewer iterations, and therefore cloth simulations, than random-walk MCMC. It trades expensive overhead (compared with other methods) in choosing search candidates for greater search efficiency[19].

Following ref. 24, we sought to learn a function from latent scene variables (that is, shape and rotation) to their (unnormalized) log posterior scores. By specifying a tractable Gaussian process (GP) prior over functions and conditioning on all available data, BayesOpt yields an online strategy for adaptively choosing parameter settings to evaluate and prescribes how the results update the GP posterior. The uncertainty in the GP approximation of the log posterior score decreases as the number of inference iterations increases (that is, as more evaluations of the posterior are observed). This probabilistic approximation is computationally cheap to evaluate and has support over the entire range of scene hypotheses (that is, can be evaluated for any scene hypothesis including those that are previously not evaluated).

BayesOpt requires specification of the GP kernel, which encodes prior assumptions about, for example, the smoothness of functions[70], and an acquisition function that selects the next hypothesis given the results of all previous evaluations. In our work, we used a Matérn kernel with $\nu = 1.5$ (the Matérn 3/2 kernel) and automatic relevance determination[70] to learn a probabilistic mapping from latent scene hypothesis onto posterior scores; and we use the expected improvement (EI) as our acquisition function, which favours scene hypotheses that are expected to most improve the posterior score. Each iteration of BayesOpt consists of (1) updating the estimated regression function (from scene hypotheses to posterior scores) and (2) optimizing the acquisition function to determine which scene hypothesis to evaluate in the next iteration. We next describe each of these two components.

Scene hypotheses are coded specially for BayesOpt. We represent rotations using normalized Euler angles $R = \{R_x, R_y, R_z\}$, with each orthogonal axis taking values in [0, 1] and together spanning the half-sphere of rotations. The shape variable is discrete, which we transform to a continuous encoding using vector quantization. We map the interval [0, 0.25) to shape hypothesis (that is, nearest neighbour) 1; [0.25, 0.50) maps to the next nearest neighbour, shape 2, and so on up to shape 4. The GP therefore learns a regression function from a four-dimensional input (three numbers for rotation, one for shape) to a scalar, the log posterior score.

With this GP approximation at hand, we define an 'acquisition function' that uses the current GP state to select the most promising scene hypothesis to try in the next iteration $j+1$, $(S^{(j+1)}, R^{(j+1)})$. Various active sampling (or learning) heuristics are proposed in the literature (see, for example, refs. 19,23,24). We adopt the EI acquisition function[19], which chooses the scene hypothesis that is expected to most improve the current posterior score, given all of the previous posterior evaluations. At each iteration j of our model, the inference procedure evaluates a scene hypothesis chosen to optimize the EI acquisition function. EI uses a parameter, denoted ϵ (set to 330 in our simulations), to trade off between how much to weigh the predicted posterior score versus the uncertainty around that prediction (notice that the GP-based probabilistic regression provides both the predicted mean posterior score and variance around that prediction for the entire range of scene hypotheses). To find the scene hypothesis that optimizes EI, we generate 100,000 random scene hypotheses and use the highest scoring to initialize further local search (using L-BFGS-B). This procedure yields the scene hypothesis (S_j, R_j) to be evaluated in the next iteration of the model. We evaluate the posterior at this scene hypothesis using equation (1).

We implemented our inference scheme using the BayesOpt[71] and GPy[72] packages.

Bottom-up models

We tested three DCNN architectures: AlexNet[7], ResNet50 (ref. 29) and VGG16 (ref. 30). These models provide powerful feature hierarchies that are learned as a result of training to classify images from the large-scale real-world ImageNet[28] dataset.

Imagesets for fine-tuning. Imagesets for fine-tuning were derived from the 5 shapes/category × 10 categories = 50 object shapes. These are the identical set of objects as those underlying the experimental training trials used to familiarize human participants with the task. We note that in our behavioural experiments we did not provide feedback during the training phase and indeed did not find any evidence of learning in our behavioural data (Supplementary Fig. 6). We used 8 imagesets per occlusion condition, with 500 unique trials in each set giving 500 × 8 = 4,000 image triplets. We evaluated how the amount of data used for fine-tuning influenced performance, finding that performance plateaued at 8 imagesets, compared with alternative groups of 1, 2, 8, 18, 28 and 38 imagesets. For each triplet, we sampled two objects and randomly rotated, draped and rendered them using our stimulus generation pipeline. We reserved two imagesets for test and the remaining were used for training. To minimize bias, a set of 8 imagesets was sampled from a larger pool of 54 at the beginning of each fine-tuning procedure. We fine-tuned each model 32 times for each occlusion condition and report accuracy averaged over the condition-specific replicas.

Modifying network architectures for fine-tuning. To fine-tune AlexNet and VGGG16, we removed their top classification layer and replaced it with a linear fully connected layer of size 120. We trained the added linear layer of size 120 from scratch and also fine-tuned the weights of the layer preceding it with the same learning rate (this fine-tuned layer would be the first fully connected layer in Alexnet).

Unlike AlexNet and VGG16, the ResNet-50 model does not contain multiple final fully connected layers; thus, we used a modified approach to fine-tune it. We replaced both its top classification layer as well as the preceding average pooling layer with a convolutional layer with kernel size 2, stride 2, and dilation 2, without zero-padding. This convolutional layer takes as input 2,048 feature maps (the number of output feature maps in the fourth residual block of the ResNet-50 model) each with dimensionality 7 × 7 and outputs 300 feature maps (each with dimensionality 3 × 3). The ReLU activation function is applied to the flattened outputs of this convolutional layer, which is followed by a single linear fully connected layer of size 120. We trained the weights of the new convolutional layer as well as the fully connected layer from scratch, while keeping all other weights in the network unchanged.

Details of the training procedure. To adapt the networks to our visual matching task, we used metric learning with a triplet margin loss[73]. The goal is to adapt the network's representational space so that distance in that space reflects the similarity structure of our stimuli. Concretely, the distance between an 'anchor' image and a 'positive example' should be smaller than the distance between the anchor and 'negative example'. A training triplet has the same structure as our behavioural match-to-sample task setup: anchor corresponds to the target item; positive example corresponds to the ground-truth matching test item; and the negative example corresponds to the distractor test item (remember that the training datasets are crafted differently for each occlusion condition and different networks are trained for each of these occlusion conditions). We fine-tuned each architecture for a total of 200 epochs and used a held-out test set to make sure the models did not overfit over the course of training.

We set batch size to 8 and set the triplet loss margin to 2.0. We used the ADAM optimizer[74] with ASMGrad[75] using the following optimization parameters. We set β_1 to 0.9, β_2 to 0.999, learning rate to 1.2×10^{-6} and ℓ_2 weight decay to 1.8×10^{-3} at the beginning of training. In an attempt to optimize the performance of the bottom-up networks, we explored a range of custom learning rate schedules as well as regularization methods. During training, we scheduled the learning rate as the following. The learning rate is multiplied by 1.2 from epoch 1 to epoch 12. From epoch 13 to epoch 161, the learning rate is annealed by multiplying it with 0.985, after which it was kept constant until epoch 200. In addition, to avoid overfitting, we employ regularization using a weight decay strategy and data augmentation. From epoch 13 to 161, we multiply the weight decay parameter by 1.04 and 1.03 in fine-tuning the occluded and unoccluded task conditions, respectively. We observed that, without this scheduled weight decay, models essentially memorized the training image set, giving rise to a substantial discrepancy between training and test performance. As a form of data augmentation, during training, we randomly perturb each image by adding white noise (variance set to 8.3×10^{-3}) with probability 0.3 (the added noise was restricted to the foreground pixels). All pixel values were truncated to ensure that their values lie between 0 and 1.

Evaluation of bottom-up models on the object-under-cloth task. The accuracy of the pre-trained bottom-up model on a given trial was calculated using the following procedure. Recall that each trial in the object-under-cloth task consists of three images: the target item, the matching test item, and the distractor test item. We compute a feature embedding of each of these three images from the first fully connected layer of the network. We define a correct answer (accuracy of 1 for this trial) from the network if the correlation between the embeddings of the target item and the matching test item (denoted corr_m) is greater than the correlation between the embeddings of the target item and the distractor test item (denoted corr_d). Otherwise, the network got the trial wrong (accuracy 0). The accuracy levels of the pre-trained bottom-up model underlying Fig. 4d,f are calculated in this way.

In Fig. 5 where we require a continuous covariate per trial from each model (as opposed to a binary accuracy label), we use Luce's choice rule (that is, softmax) to transform the above mentioned correlation values to a continuous score: $\text{corr}_m/(\text{corr}_m + \text{corr}_d)$. Notice that the model predictions in Fig. 5b are normalized to the range of [0, 1] for all models.

The trial-level accuracy of the fine-tuned model is calculated in a manner similar to the PbAS model. For a given trial and a fine-tuned network, we select the test item that is closer to the target item as the network's guess and report the fraction of correct guesses (that is, the closest test item was the matching test item) across the ensemble of 32 independently fine-tuned networks.

Behavioural methods

Participants. A total of 174 participants were recruited from Amazon's crowdsourcing platform Mechanical Turk. All participants self-confirmed to be at least at the age of 18 years old or older, and all provided informed consent before the beginning of the study. The experiment took about 20 min to complete. Each participant was paid US$1.50. A total of 12 subjects were excluded due to performing at or below chance performance (1 in Unoccluded-1 s; 4 in Occluded-1 s; 3 in Occluded-2 s; and 4 in Occluded-Unlimited). Approval for our behavioural study was obtained from the Massachusetts Institute of Technology Institutional Review Board (the Committee on the Use of Humans as Experimental Subjects), and we obtained each participant's informed consent before any experimental session.

Stimuli and procedure. We used 240 unique ShapeNet meshes from 10 object categories to create the 120 match-to-sample shape pairs in our task. We selected 24 objects from each category and allocated them evenly between the same-category (target and distractor from same object category) and different-category conditions, pairing each shape with another from the same category or a different category as appropriate. Pairings were sampled randomly without replacement. We thus obtained six same-category and six different-category pairs for each object category, with no duplicate shapes across trials.

We designed a visual matching experiment based on the object-under-cloth task. The experiment assigned participants to either the occluded or unoccluded conditions as well as one of three conditions varying presentation time lengths, for a between-subject design with 2 occlusion × 3 presentation time = 6 conditions. In the 1- and 2-s conditions, the target and test items were displayed for the indicated period of time and the unlimited time condition let participants view the items for as long as they wished, that is until their response. Images appeared and disappeared simultaneously.

The spatial organization of the display differed slightly by occlusion condition. In the unoccluded condition, the two test images were placed side by side, below the target item; for the occluded condition, the test images were placed side by side but above the target.

Participants completed 10 practice trials before moving on to the 120 experimental trials. Participants were provided with running feedback, seeing their average task performance at every fifth trial throughout the experiment except during the practice block (the performance feedback calculation excluded practice trial accuracy).

Split-half correlations. To estimate the data noise ceiling, we used bootstrapped split-half correlations. We sampled 1,000 random splits of our participants in each occlusion condition (only considering the unlimited presentation time condition), each split dividing the participants into two groups of equal size (participants were sampled without replacement for each partition). For a single division (one random split) of participants, we computed the average accuracy of each split-half on each trial, then correlated the group accuracies across all trials (in essence, we used the responses of one split of participants to model the responses of the other half). We did the same for each of the 1,000 random splits, yielding 1,000 bootstrap estimates of the behavioural noise ceiling and allowing us to assess their average value and spread. But because this procedure effectively halved our participant number, our split-half correlations probably underestimate the true noise ceiling.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Our behavioural data are publicly available at https://github.com/CNCLgithub/PbAS-model-human-comparisons. The experimental stimuli underlying the object-under-cloth task are publicly available at https://github.com/CNCLgithub/intuitive-physics-3d-shape-perception-stimuli.

Code availability

Code implementing the PbAS model, scripts for replicating model simulations, and a container for full reproducibility are publicly available at https://github.com/CNCLgithub/PbAS. Our custom Python scripts for data analysis are publicly available at https://github.com/CNCLgithub/PbAS-model-human-comparisons.

References

1. Bulthoff, H. Shape from X: psychophysics and computation. *Comput. Models Vis. Process.* 305–330 (1991).
2. Yildirim, I., Siegel, M. H. & Tenenbaum, J. B. Perceiving fully occluded objects via physical simulation. In *Proc. 38th Annual Conference of the Cognitive Science Society* 1265–1271 (Cognitive Science Society, 2016).
3. Phillips, F. & Fleming, R. W. The Veiled Virgin illustrates visual segmentation of shape by cause. *Proc. Natl Acad. Sci. USA* **117**, 11735–11743 (2020).
4. Little, P. C. & Firestone, C. Physically implied surfaces. *Psychol. Sci.* **32**, 799–808 (2021).
5. Wong, K. W., Bi, W., Soltani, A. A., Yildirim, I. & Scholl, B. J. Seeing soft materials draped over objects: a case study of intuitive physics in perception, attention, and memory. *Psychol. Sci.* **34**, 111–119 (2022).
6. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
8. Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613 (2016).
9. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
10. Mumford, D. in *Large-Scale Neuronal Theories of the Brain* (eds Koch, C. & and Davis, J.) 125–152 (MIT Press, 1994).
11. Liu, Z., Knill, D. C. & Kersten, D. Object classification for human and ideal observers. *Vis. Res.* **35**, 549–568 (1995).
12. Destler, N., Singh, M. & Feldman, J. Skeleton-based shape similarity. *Psychol. Rev.* https://doi.org/10.1037/rev0000412 (2023).
13. Erdogan, G. & Jacobs, R. A. Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychol. Rev.* **124**, 740 (2017).
14. Lee, M. J. & DiCarlo, J. J. An empirical assay of view-invariant object learning in humans and comparison with baseline image-computable models. Preprint at *bioRxiv* (2023).
15. Chandra, K., Li, T.-M., Tenenbaum, J. & Ragan-Kelley, J. Designing perceptual puzzles by differentiating probabilistic programs. In *ACM SIGGRAPH 2022 Conference Proceedings* 1–9 (ACM, 2022).
16. Chang, A. X. et al. ShapeNet: an information-rich 3D model repository. Preprint at https://doi.org/10.48550/arXiv.1512.03012 (2015).
17. Macklin, M., Müller, M., Chentanez, N. & Kim, T.-Y. Unified particle physics for real-time applications. *ACM Trans. Graph.* **33**, 1–12 (2014).
18. Koch, E., Baig, F. & Zaidi, Q. Picture perception reveals mental geometry of 3D scene inferences. *Proc. Natl Acad. Sci. USA* **115**, 7807–7812 (2018).
19. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **25**, 2951–2959 (2012).
20. Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proc. Natl Acad. Sci. USA* **117**, 30055–30062 (2020).
21. Hamrick, J. B. & Griffiths, T. L. Mental rotation as Bayesian quadrature. In *NIPS 2013 Workshop on Bayesian Optimization in Theory and Practice* (2013).
22. Wang, A., Mei, S., Yuille, A. L. & Kortylewski, A. Neural view synthesis and matching for semi-supervised few-shot learning of 3D pose. *Adv. Neural Inf. Process. Syst.* **34**, 7207–7219 (2021).
23. Järvenpää, M., Gutmann, M. U., Pleska, A., Vehtari, A. & Marttinen, P. Efficient acquisition rules for model-based approximate Bayesian computation. *Bayesian Anal.* **14**, 595–622 (2019).
24. Kandasamy, K., Schneider, J. & Póczos, B. Bayesian active learning for posterior estimation. In *24th International Joint Conference on Artificial Intelligence* 3605–3611 (PMLR, 2015).
25. Tamura, R. & Hukushima, K. Bayesian optimization for computationally extensive probability distributions. *PLoS ONE* **13**, e0193785 (2018).
26. Schrimpf, M. et al. Brain-score: which artificial neural network for object recognition is most brain-like? Preprint at *bioRxiv* https://doi.org/10.1101/407007 (2018).
27. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356 (2016).
28. Deng, J. et al. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
30. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556(2014).
31. Usher, M. & McClelland, J. L. The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.* **108**, 550 (2001).
32. Echeveste, R., Aitchison, L., Hennequin, G. & Lengyel, M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nat. Neurosci.* **23**, 1138–1149 (2020).
33. Yildirim, I., Belledonne, M., Freiwald, W. & Tenenbaum, J. Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979 (2020).
34. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
35. Yamins, D. L. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
36. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
37. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
38. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2020).
39. Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 1–12 (2022).
40. Geirhos, R. Partial success in closing the gap between human and machine vision. *Adv. Neural Inf. Process. Syst.* **34**, 23885–23899 (2021).
41. Gilbert, C. D. In *Principles of Neural Science* 5th edn (eds. Kandel, E. R. et al.) 556–576 (McGraw-Hill, 2013).

42. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
43. Feldman, J. & Singh, M. Information along contours and object boundaries. *Psychol. Rev.* **112**, 243 (2005).
44. Zucker, S. W. On qualitative shape inferences: a journey from geometry to topology. Preprint at https://doi.org/10.48550/arXiv.2008.08622 (2020).
45. Kunsberg, B. & Zucker, S. W. Critical contours: an invariant linking image flow with salient surface organization. *SIAM J. Imaging Sci.* **11**, 1849–1877 (2018).
46. Grill-Spector, K. & Kanwisher, N. Visual recognition: as soon as you know it is there, you know what it is. *Psychol. Sci.* **16**, 152–160 (2005).
47. Ullman, S. in *Readings in Computer Vision* (eds Fischler, M. A. & Firschein, O.) 298–328 (Elsevier, 1987).
48. Shepard, R. N. & Metzler, J. Mental rotation of three-dimensional objects. *Science* **171**, 701–703 (1971).
49. Shams, L. & Beierholm, U. Bayesian causal inference: a unifying neuroscience theory. *Neurosci. Biobehav. Rev.* **137**, 104619 (2022).
50. Fischer, J., Mikhael, J. G., Tenenbaum, J. B. & Kanwisher, N. Functional neuroanatomy of intuitive physical inference. *Proc. Natl Acad. Sci. USA* **113**, E5072–E5081 (2016).
51. Nash, C., Ganin, Y., Eslami, S. A. & Battaglia, P. Polygen: an autoregressive generative model of 3d meshes. In *International Conference on Machine Learning* (7220–7229) (2020).
52. Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. & Battaglia, P. W. Learning mesh-based simulation with graph networks. Preprint at https://doi.org/10.48550/arXiv.2010.03409 (2021).
53. Mrowca, D. et al. Flexible neural representation for physics prediction. In *Proc. 32nd International Conference on Information Processing Systems* 8813–8824 (2018).
54. Smith, K. et al. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Adv. Neural Inf. Process. Syst.* **32**, 8983–8993 (2019).
55. Piloto, L. S., Weinstein, A., Battaglia, P. & Botvinick, M. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nat. Hum. Behav.* **6**, 1257–1267 (2022).
56. Sanborn, A. N., Mansinghka, V. K. & Griffiths, T. L. Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychol. Rev.* **120**, 411 (2013).
57. Wu, J., Yildirim, I., Lim, J. J., Freeman, B. & Tenenbaum, J. Galileo: perceiving physical object properties by integrating a physics engine with deep learning. *Adv. Neural Inf. Process. Syst.* **28**, 127–135 (2015).
58. Schwettmann, S., Tenenbaum, J. B. & Kanwisher, N. Invariant representations of mass in the human brain. *eLife* **8**, e46619 (2019).
59. Bi, W., Shah, A. D., Wong, K. W., Scholl, B. & Yildirim, I. Perception of soft materials relies on physics-based object representations: Behavioral and computational evidence. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.12.443806 (2021).
60. Paulun, V. C., Schmidt, F., van Assen, J. J. R. & Fleming, R. W. Shape, motion, and optical cues to stiffness of elastic objects. *J. Vis.* **17**, 20–20 (2017).
61. Paulun, V. C. & Fleming, R. W. Visually inferring elasticity from the motion trajectory of bouncing cubes. *J. Vis.* **20**, 6–6 (2020).
62. Bates, C. J., Yildirim, I., Tenenbaum, J. B. & Battaglia, P. Modeling human intuitions about liquid flow with particle-based simulation. *PLoS Comput. Biol.* **15**, e1007210 (2019).
63. Kubricht, J., Zhu, Y., Jiang, C., Terzopoulos, D., Zhu, S. C. & Lu, H. Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Proc. 39th Annual Conference of the Cognitive Science Society* 3426–3431 (Cognitive Science Society, 2017).
64. Van Assen, J. J. R., Barla, P. & Fleming, R. W. Visual features in the perception of liquids. *Curr. Biol.* **28**, 452–458 (2018).
65. Chen, Y.-C. & Scholl, B. J. The perception of history: seeing causal history in static shapes induces illusory motion perception. *Psychol. Sci.* **27**, 923–930 (2016).
66. Fleming, R. W. & Schmidt, F. Getting "fumpered": classifying objects by what has been done to them. *J. Vis.* **19**, 15–15 (2019).
67. Schmidt, F., Phillips, F. & Fleming, R. W. Visual perception of shape-transforming processes: 'shape scission'. *Cognition* **189**, 167–180 (2019).
68. Blender Online Community Blender—a 3D modelling and rendering package. *Blender Institute* http://www.blender.org (2015).
69. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. Preprint at https://doi.org/10.48550/arXiv.1811.12231 (2018).
70. Rasmussen, C. E. & Williams, C. K. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
71. Nogueira, F. Bayesian Optimization: open source constrained global optimization tool for Python. *GitHub* https://github.com/fmfn/BayesianOptimization (2014).
72. GPy: a Gaussian process framework in Python. *GitHub* http://github.com/SheffieldML/GPy (2012).
73. Schultz, M. & Joachims, T. Learning a distance metric from relative comparisons. *Adv. Neural Inf. Process. Syst.* **16**, 41–48 (2003).
74. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2015).
75. Reddi, S. J., Kale, S. & Kumar, S. On the convergence of Adam and beyond. Preprint at https://doi.org/10.48550/arXiv.1904.09237 (2018).

Acknowledgements

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216; ONR MURI N00014-13-1-0333 (to J.B.T.); a grant from Toyota Research Institute (to J.B.T.); and a grant from Mitsubishi MELCO (to J.B.T.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. A high performance computing cluster (OpenMind) was provided by the McGovern Institute for Brain Research. We thank K. Smith, B. Egger, K. Allen, G. Erdogan, M. Tenenbaum, N. Kanwisher and V. Paulun for their comments on a previous version of this manuscript.

Author contributions

I.Y., M.H.S. and J.B.T. conceived and designed the study. I.Y. analysed the data. I.Y., M.H.S., A.A.S. and J.B.T. designed stimuli and experiments, and wrote and edited the manuscript. All authors contributed to the models.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01759-7.

Correspondence and requests for materials should be addressed to Ilker Yildirim, Max H. Siegel or Joshua B. Tenenbaum.

Peer review information *Nature Human Behaviour* thanks Yaniv Morgenstern, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Ilker Yildirim

Last updated by author(s): 2023-9-8

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used custom javascript code and the psiturk package to collect behavioral data.
Data analysis	We performed all analysis using custom scripts in Python v3.6. Our analysis scripts are publicly available at https://github.com/CNCLgithub/PbAS-model-human-comparisons.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Our behavioral data is publicly available at https://github.com/CNCLgithub/PbAS-model-human-comparisons. The experimental stimuli (images) underlying the object-under-cloth task are publicly available at https://github.com/CNCLgithub/intuitive-physics-3d-shape-perception-stimuli. The stimuli are generated using custom Python scripts and Blender computer graphics package.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We did not collect information on sex and gender as we did not believe our hypothesis pertained to any sex- and gender-based differences according to available literature.
Reporting on race, ethnicity, or other socially relevant groupings	We did not collect information on race, ethnicity, or other socially relevant grouping as we did not believe our hypothesis pertained to these variables according to available literature.
Population characteristics	Participants, who self-confirmed being at least at the age of 18, were randomly drawn from the user base of a crowdsourcing platform.
Recruitment	Participants were recruited using an online crowdsourcing platform in a randomized manner (as offered by the platform). We acknowledge that this introduces a selection bias, including the portion of the population with access to a computer and internet connection.
Ethics oversight	Institutional Review Board at Massachusetts Institute of Technology

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A behavioral experiment with quantitative psychophysics and computational modeling.
Research sample	Participants, who are self-confirmed to be at least 18 years or older, are sampled from the user base of Mechanical Turk platform. Information on sex was not collected. This sample is fairly representative given the goals of our study (basic mechanisms of visual perception), but we acknowledge that it is biased toward western cultures and selects people with access to computers and internet.
Sampling strategy	Participants were randomly assigned from the Mechanical Turk platform. We chose our sample size based on previous studies using similar methodology (computational vision/psychophysics).
Data collection	All data collection occurred online using a custom javascript-based web interface. Participants, all of whom were outside of the lab, interacted with that web interface using their browsers and responded to visual stimuli by pressing keys on their keyboards. We recorded their key presses and response times. The researchers did not accompany participants. Researchers were unaware of the conditions the participants were assigned to, and had no influence on their behavior or performance because all instructions and stimuli were presented automatically and remotely.
Timing	3/2019-6/2019
Data exclusions	A total of 12 subjects were excluded due to performing at or below chance performance (1 in Unoccluded-1sec; 4 in Occluded 1-sec; 3 in Occluded-2secs; and 4 in Occluded-Unlimited conditions). This criterion was pre-established.
Non-participation	A total of 202 participants visited or started the task but did not finish it. Typical reasons include generic connectivity issues such as poor internet connection.
Randomization	Participants were randomly assigned to conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging