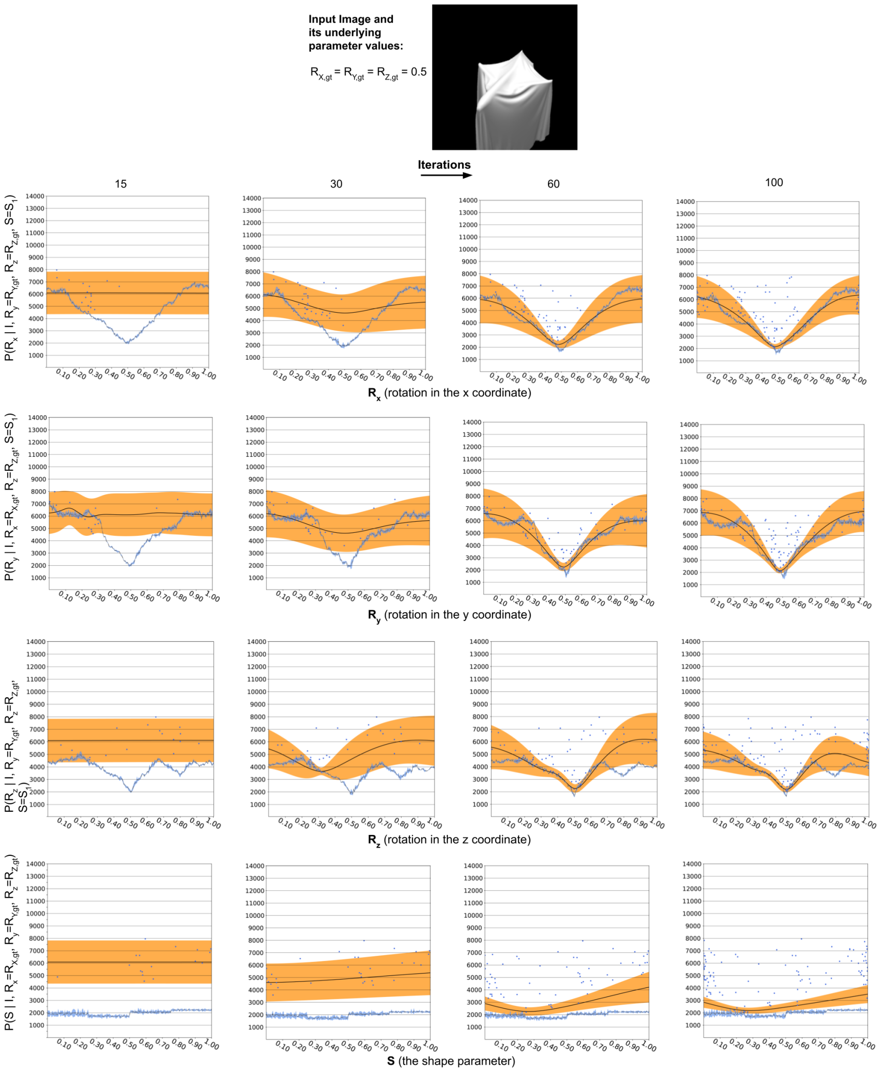
nature human behaviour

Article

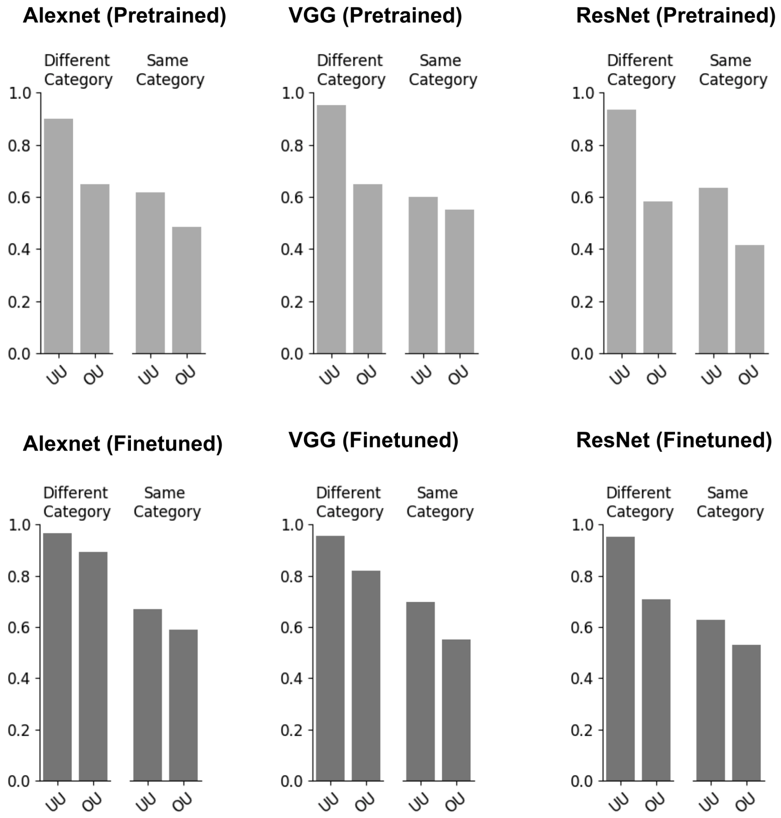
https://doi.org/10.1038/s41562-023-01759-7

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

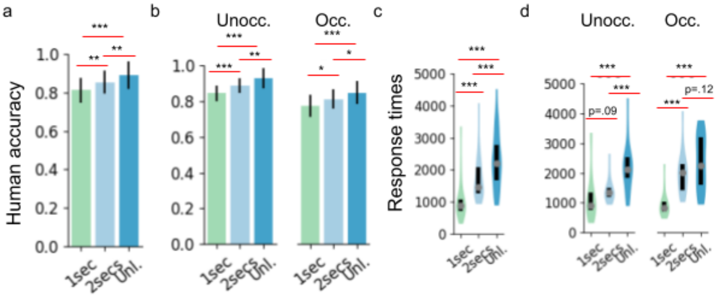
Supplementary Information



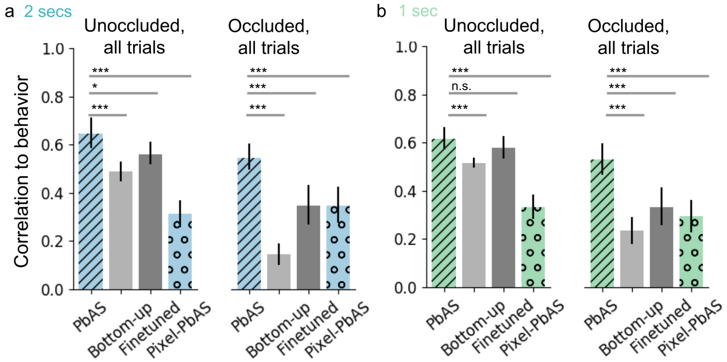
Supplementary Figure 1 The inputs, internals, and outputs of the "BayesOpt" box from Fig. 3 in the main text. We show the Gaussian Process (GP) approximation of the posterior distribution $Pr(S, R \mid I)$: snapshots at selected iteration numbers given the observed image I on the left. Rows show the GP estimation for each of three angular rotation parameters specifying object pose, as well as the shape parameter (see Materials & Methods; the x-axes show the full range of values each parameter can take, normalized to lie between 0 and 1). Each plot displays the mean (black line) and the uncertainty of the GP approximation of the posterior (orange shading showing 2 standard deviations) and the true posterior score (black line with blue shading), as a function of the indicated parameter, with all other latent variables set to ground truth (except for the shape, which is set to the nearest neighbor, S_1). (Notice that due to the stochasticity of the physics simulator, the true posterior score is a random variable; the shaded blue regions show standard deviation taken across multiple evaluations of the synthesis module). The GP approximation is initially poor (iteration #15), but rapidly improves. At each iteration, the Expected Improvement acquisition function chooses a new point to sample and evaluate according to this posterior estimate, by making a trade-off between the GP mean and covariance (see Materials & Methods). The blue dots on each panel show posterior evaluations for selected parameter values. Notice that the number of evaluated samples grows with the number of iterations.



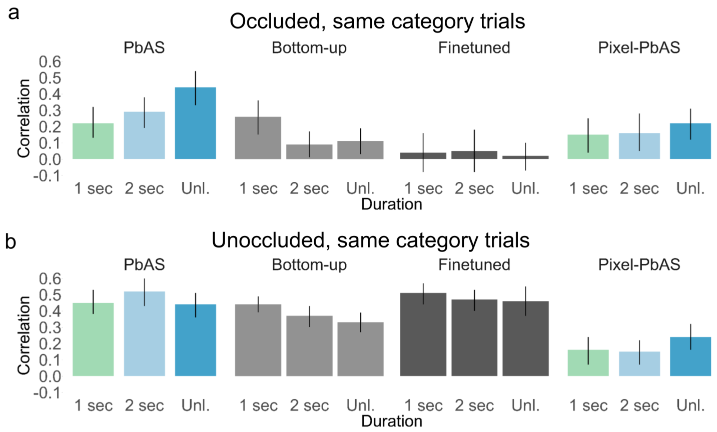
Supplementary Figure 2 Accuracy levels of the three models we considered including pretrained versions (top row) and after finetuning (bottom row). AlexNet results are presented in the main text.



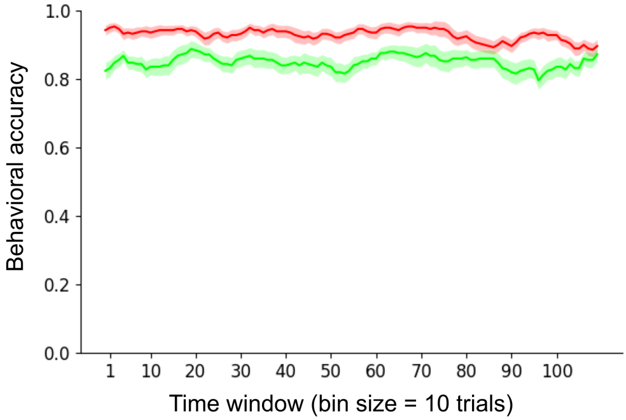
Supplementary Figure 3 Behavioral results. **(a)** Average human accuracy in the 3 presentation time conditions pooling data across the occlusion conditions. Overall, participants performed well above chance under all presentation time conditions. Behavioral accuracy improved with longer presentation times: 1 sec vs. 2 secs $t(51) = -3.187, p = .002$; 2 secs vs. Unlimited $t(52) = -3.049, p = 0.003$, 1 sec vs. Unlimited $t(51) = -6.404, p < .001$, two-tailed independent samples t-tests. (Using independent samples of participants for each bar: 1-sec $n = 53$; 2-secs $n = 55$; Unlimited $n = 54$.) **(b)** Average human accuracy shown separately for each occlusion and presentation time condition. Participants' average performance ranged from 73% in the cloth-occluded condition under 1 sec presentation time to 93% in the unoccluded condition under unlimited time. The gain in performance was significant within each occlusion condition for all pairwise comparisons of presentation times: Unocc 1 sec vs. 2secs $t(27) = -3.921, p < 0.001$; Unocc 2 secs vs. Unlimited $t(27) = -3.142, p = 0.003$; Unocc 1 sec vs. Unlimited $t(27) = -6.378, p < .001$; Occ 1 sec vs. 2secs $t(22) = -2.379, p = .022$; Occ 2 secs vs. Unlimited $t(23) = -2.246, p = .029$; Occ 1 sec vs. Unlimited $t(22) = -4.459, p < .001$, two-tailed independent samples t-tests. (Using independent samples of participants for each combination of presentation time and occlusion condition: 1-sec Unoccluded $n = 29$; 1-sec Occluded $n = 24$; 2-secs Unoccluded $n = 30$; 2-secs Occluded $n = 25$; Unlimited Unoccluded $n = 29$; Unlimited Occluded $n = 25$) **(c)** Average response times (in milliseconds; pooling data across the occlusion conditions) lengthen with longer presentation times, $p < .001$ for all pairwise comparisons of presentation time conditions: 1 sec vs. 2 secs $t(51) = -5.616, p < .001$; 2 secs vs. Unlimited $t(52) = -4.121, p < 0.001$; 1 sec vs. Unlimited $t(51) = -8.121, p < .001$, using two-tailed independent samples t-tests. (Using identical samples of participants as panel a.) **(d)** Average response times shown separately for each occlusion and presentation time condition. Lengthening of response times is still evident for each occlusion condition for all pairwise comparisons of presentation times (except in the 1 sec vs. 2 secs comparison in the unoccluded condition and 2 secs vs. Unl. comparisons in the cloth-occluded condition): Unocc 1 sec vs. 2 secs $t(27) = -1.749, p = 0.086$; Unocc 2 secs vs. Unlimited $t(27) = -4.410, p < .001$; Unocc 1 sec vs. Unlimited $t(27) = -4.874, p < .001$; Occ 1 sec vs. 2secs $t(22) = -6.732, p < .001$; Occ 2 secs vs. Unlimited $t(23) = -1.585, p = .120$; Occ 1 sec vs. Unlimited $t(22) = -7.723, p < .001$, two-tailed independent samples t-tests. (Using identical samples of participants as panel b.) Error bars show standard deviation. In panels (c) and (d), gray dots show the median, and thick black lines extend between the 25 and 75% percentile of the distribution.



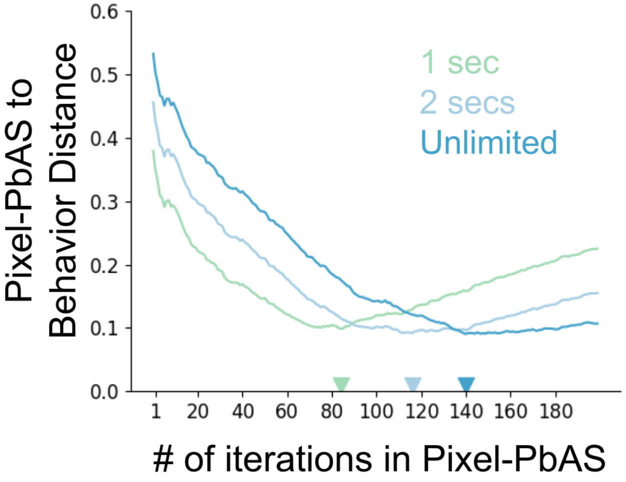
Supplementary Figure 4 Average of the bootstrapped trial-level accuracy correlations in the **(a)** 2 secs presentation time condition and **(b)** 1 sec presentation time conditions. The physics-based analysis-by-synthesis (PbAS) model correlates well with behavior across all presentation time and occlusion condition time conditions, relative to the alternatives based on bottom-up features optimized for image classification (BU: bottom-up network with pretrained weights from ImageNet dataset; FT: fine-tuned networks, separately fine-tuned for each occlusion conditions) and Pixel-PbAS, an ablation of PbAS without the bottom-up image encoding modules (using pixels directly for likelihood computation). "***": $p < .001$; "*": p = 046; "n.s." = .242. Error bars indicate bootstrapped 95% confidence intervals. Statistical comparisons are performed using two-tailed direct bootstrap hypothesis testing ($n = 5000$ bootstrap samples by sampling participants with replacement).



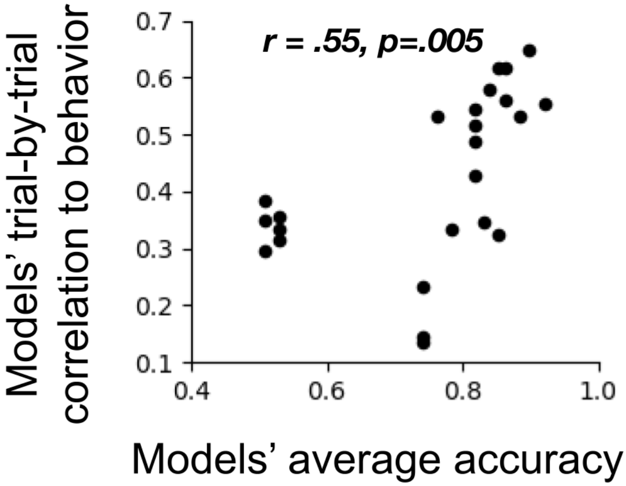
Supplementary Figure 5 Average of the bootstrapped trial-level accuracy correlations for the difficult, same category trials in the **(a)** Occluded and **(b)** Unoccluded conditions. Results are arranged by model type and stimulus presentation time. Error bars show bootstrapped 95% confidence intervals (n = 5000 bootstrap samples by sampling participants with replacement). In the easier unoccluded, shape-category conditions, all three models that use DCNN features to match images (PbAS, BU, and FT) perform similarly across all presentation times; Pixel-PbAS performs significantly worse across all presentation times. In the more difficult occluded, same-category conditions, PbAS clearly outperforms all other models, except for BU which performs similarly in the shortest (1 sec) presentation time. Notably both pure DCNN models, BU and FT, consistently correlate less well with human trial-level accuracies as presentation times increase, while PbAS correlations tend to increase, and FT correlations are not significantly different from zero in the challenging occluded same-category conditions (with BU correlations being only barely higher than zero in the 2 sec and unlimited conditions). This overall pattern is consistent with the success of DCNNs at capturing the rapid feedforward contributions to human object recognition for familiar stimuli viewed under standard conditions, and strengthens our proposal that more challenging viewing conditions and longer processing times engage top-down, iterative, generative model based computations of the form instantiated in PbAS. The combination of physics-based analysis by synthesis with DCNN features for matching generative model simulations to images, as instantiated in the full PbAS model but not Pixel-PbAS, is the only model that accounts well (and better than or equal to any other model) for all stimulus conditions and all presentation times.



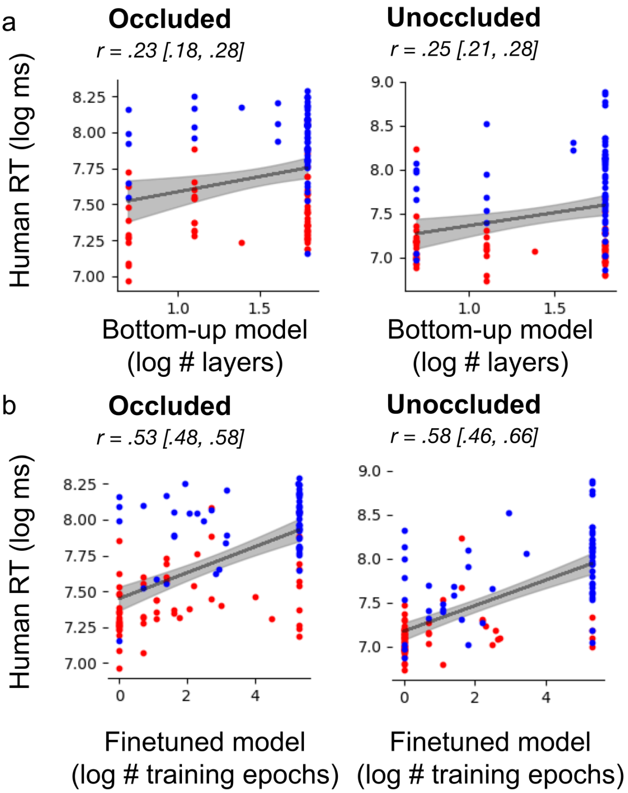
Supplementary Figure 6 Behavioral learning curves in the two occlusion conditions. We show moving window averages (window size=10) of human accuracy levels in the two occlusion conditions (solid lines; red=Unoccluded, green=Occluded) under the unlimited presentation time condition. We find no evidence of learning throughout the experiment. Shaded region shows standard error.



Supplementary Figure 7 Divergence between the Pixel-PbAS model (i.e., using pixels for likelihood without bottom-up image encoding) and human performance at each model iteration. Colored lines show ℓ_2 distance between this model and human accuracy for all trials in indicated presentation time condition. Colored triangles indicate the best matching iterations for each presentation time condition. This model asymptotes at a larger distance to behavior than the PbAS model.



Supplementary Figure 8 Correlation between a model's average accuracy (x-axis) and its trial-by-trial correlation to behavior (y-axis) – shown for all models (4 models [PbAS, Bottom-up, Fine-tuned, Pixel-PbAS]) and experimental condition (6 conditions), resulting in 24 data points. We see that overall, there is a medium-strength positive correlation between the accuracy of a model and its trial-by-trial consistency with behavior (p = .005). This observation motivates future work to develop more performance-matched models (similar to the Pixel-PbAS model) and use behavior to adjudicate among them.



Supplementary Figure 9 Evaluating the pretrained and fine-tuned models on response time data. **(a)** Comparing the number of network layers (log #) needed to reach a decision threshold in the pretrained bottom-up network vs. average human response times (log ms). We model response times using the network layer at which a decision threshold has been reached. To implement this proposal, we go through the network layers conv1, conv2, conv3, conv4, conv5, and fc1 in the pretrained bottom-up network, and make a decision as soon as the ratio $corr_m/(corr_m + corr_d)$ exceeds a certain threshold. The resulting correlations (unoccluded: r = .25; occluded: r= .23) are significantly lower than the correlations due to PbAS ($p < .001$; using direct bootstrap hypothesis testing), even though PbAS has no free parameters. To give the most favorable setting possible for the pretrained model, these results are based on the decision threshold that maximized these correlations: for these results, we make a decision when $corr_m/(corr_m+corr_d) >= .57$. **(b)** Comparing the number of training epochs (log #) required for the average accuracy of the 32 fine-tuned models to reach a decision threshold, vs. average human response times (log ms). This method focuses on the fine-tuned models and exploits the fact that we train the networks for 200 epochs during fine-tuning. We treat the training epochs much like the number of iterations in PbAS and make a decision at a given training epoch as soon as the average of the 32 fine-tuned models reaches average human accuracy in a given condition. This resulting correlations (unoccluded: r = .58; occluded: r = .53) are lower than that of the PbAS model ($p < .001$; using direct bootstrap hypothesis testing). Moreover, this relationship is largely categorical: The fine-tuned models require fewer training epochs in the easier different-category trials (red dots) and often take until the final training epoch in the harder same-category trials (blue dots). In contrast, the number of inference steps in PbAS correlates with behavioral response times in a much more graded manner (see Fig. 6 in the main text). Shaded regions show 95% CIs of the standard error in linear regressions (solid lines); the confidence intervals around the correlation values on top of each plot are bootstrapped 95% CIs based on resampling participants with replacement ($n = 5000$ bootstrap samples).