Efficient inverse graphics in biological face processing

Ilker Yildirim,[1,3*] Winrich Freiwald,[2,3*] Josh Tenenbaum[1,3*]

[1]**Department of Brain & Cognitive Sciences, MIT, Cambridge, MA**

[2]**Laboratory of Neural Systems, The Rockefeller University, New York, NY**

[3]**The Center for Brains, Minds, and Machines, MIT, Cambridge, MA**

[*] **Correspondence: ilkery@mit.edu, wfreiwald@rockefeller.edu, jbt@mit.edu.**

Abstract

Vision must not only recognize and localize objects, but perform richer inferences about the underlying causes in the world that give rise to sensory data. How the brain performs these inferences remains unknown: Theoretical proposals based on inverting generative models (or "analysis-by-synthesis") have a long history but their mechanistic implementations have typically been too slow to support online perception, and their mapping to neural circuits is unclear. Here we present a neurally plausible model for efficiently inverting generative models of images and test it as an account of one high-level visual capacity, the perception of faces. The model is based on a deep neural network that learns to invert a three-dimensional (3D) face graphics program in a single fast feedforward pass. It explains both human behavioral data and multiple levels of neural processing in non-human primates, as well as a classic illusion, the "hollow face" effect. The model fits qualitatively better than state-of-the-art computer vision models, and suggests an interpretable reverse-engineering account of how images are transformed into percepts in the ventral stream.

Introduction

Perception confronts us with a basic puzzle: how can our experiences be so rich in content, so robust to environmental variation, and yet so fast to compute, all at the same time? Vision theorists have long argued that the brain must not only recognize and localize objects, but make inferences about the underlying causal structure of scenes (Olshausen; Yuille and Kersten, 2006; Barrow and Tenenbaum, 1978). When we see a chair or a tree, we perceive it not only as a member of one of those classes, but also as an individual instance with many fine-grained three-dimensional (3D) shape and surface details, which persist in long-term memory (Brady et al., 2008) and are crucial for planning our actions – sitting in that

chair or climbing that tree. Similarly, when seeing a face, we not only identify a person, but also perceive details of momentary shape, texture, and subtleties of expression. Various frameworks for scene analysis by inverting causal generative models, also known as "analysis-by-synthesis", have been proposed in computational vision (Barrow and Tenenbaum, 1978; Lee and Mumford, 2003; Blanz and Vetter, 1999; Barron and Malik, 2013; Kulkarni et al., 2015a), and these models have some behavioral support (Yildirim and Jacobs, 2013; Erdogan and Jacobs, 2017). However, inference in these models is typically based on top-down stochastic search, which is highly iterative and implausibly slow: a single scene percept may take hundreds of iterations to compute (which could be seconds or minutes on conventional hardware), in contrast to the nearly instantaneous processing in the visual system which is mostly feedforward (DiCarlo et al., 2012). There is also no direct empirical evidence about whether or how analysis-by-synthesis models are implemented in stages of actual neural processing.

In part for these reasons, recent work in computational vision has focused on a different class of architectures, deep convolutional neural networks (DCNNs), which are both more directly relatable to neural circuits and more consistent with the fast bottom-up processing dynamics of the brain (DiCarlo et al., 2012; Serre et al., 2007). DCNNs consist of many layers of features arranged in a feedforward hierarchy, discriminatively trained to optimize detection of objects and object categories from labeled data. They have been instrumental both in leading engineering applications (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015) and in predicting neural responses at the level of single units in macaque cortex as well as fMRI in humans (Yamins et al., 2014; Eickenberg et al., 2017; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015). Despite their impressive successes, however, DCNNs do not attempt to address the question of how vision infers the causal structure underlying images. How we see so much so quickly, how our brains compute rich descriptions of scenes with detailed 3D shapes and surface appearances, in a few hundred milliseconds or less, remains a challenge for all existing approaches.

Here, we present a computational model that combines the best features of analysis-by-synthesis and neural network approaches in order to answer that challenge. Our efficient inverse graphics (EIG) network recovers 3D object shape and texture from a single image with a performance similar to the best inverse graphics approaches, but does so quickly

using only feedforward computations, without the need for iterative algorithms. The model consists of two parts: a probabilistic generative model based on a multistage 3D graphics program for image synthesis (Fig. 1A), and an approximate inverse function of this generative model based on a DCNN that inverts (explicitly or implicitly) each successive stage of the graphics program (Fig. 1B), layer by layer. The inverse model, also known as an inference network or recognition model, is different from conventional DCNNs for vision in two critical ways: (1) It is trained to produce the inputs to a graphics engine, the latent variables of the generative model, rather than to predict class labels such as object categories or face identities; (2) it is trained in a completely self-supervised way, with inputs and outputs internally synthesized by the generative model component, rather than requiring externally supervised training on large sets of labeled images. In this way, the EIG network embodies principles similar to the Helmholtz machine originally proposed by Hinton and colleagues in the 1990's (Hinton et al., 1995; Dayan et al., 1995), but with a generative model that is based on a graphics program (instead of a generic function approximator) and thus more faithfully captures the causal structure of how real-world scenes give rise to images. We return to this contrast in the discussion at the end of the paper.

As a test case, we apply our EIG model in the domain of face perception where, in a rare co-occurrence, data from brain imaging, single-cell recordings, quantitative psychophysics and classic visual illusions all come together to strongly constrain possible models. Although the Helmholtz machine and other analysis-by-synthesis models have existed for decades as theoretical proposals, only now do we have the empirical data needed to test these theories in a strong way. EIG implements the hypothesis that the targets of ventral stream processing, a series of interconnected cortical areas, are 3D scene properties analogous to the latent variables in a causal generative model of image formation (referred to as the "latent variables" hypothesis). We compare EIG against a broad range of alternatives, including both lesions of EIG (leaving out components of the model) and multiple variants of state-of-the-art networks for face recognition in computer vision. These variants implement an alternative hypothesis that the targets of ventral stream processing are points in an embedding space optimized for discriminating across facial identities (referred to as the "discriminative" hypothesis), without necessarily any goal of reconstructing the structure of the 3D scene. The EIG model, and therefore the latent variables hypothesis, but not

other models, accounts for the full set of neural and behavioral data, at the same time as it matches the most challenging perceptual function of the system: computing a rich, accurate percept of the intrinsic 3D shape and texture of a novel face from an observed image in a single, fast, feedforward circuit.

Efficient Inverse Graphics (EIG) Network

The model consists of two components, a probabilistic generative model and a deep network-based recognition model. The generative model takes the form of a hierarchy of latent variables and causal relations between them, representing multiple stages in a probabilistic graphics program for sampling face images (Fig. 1A). The top level random variable specifies an abstract person identity, F, drawn from a prior $Pr(F)$ over a finite set of familiar individuals but allowing for the possibility of encountering a new, unfamiliar individual. The second level random variables specify scene properties: an intrinsic space of 3D face shape S and texture T descriptors drawn from the distribution $\mathrm{Pr}(S, T|F)$, as well as extrinsic scene attributes controlling the lighting direction, L, and viewing direction (or equivalently, the head pose), P, from the distribution $\mathrm{Pr}(L, P)$. We implement this stage using the Basel Face Model (a probabilistic 3D Morphable Model) (Blanz and Vetter, 1999; IEE, 2009), although other implementations are possible. These 3D scene parameters provide inputs to a z-buffer algorithm $\Psi(\cdot)$ that outputs the third level of random variables, corresponding to intermediate-stage graphics representations (or 2.5D components) for viewpoint-specific surface geometry (normal map, N) and color (albedo or reflectance map, R), $\{N, R\} = \Psi(S, T, P)$. These view-based representations and the lighting direction then provide inputs to a renderer, $\Phi(\cdot)$, that outputs an idealized face image, $I = \Phi(N, R, L)$. Finally, the ideal face image is subject to a set of image-level operations including translation, scaling, and background addition, $\Theta(\cdot)$, that outputs an observable raw image, $O = \Theta(I)$ [Fig. 1A; Supporting Online Material (SOM) Section 1].

In principle, perception in this generative model can be formulated as MAP (Maximum A Posterior) Bayesian inference as follows. We seek to infer the individual face F, as well as intrinsic and extrinsic scene properties S, T, L, P that maximize the posterior probability

$$\Pr(F, S, T, L, P|O) \propto$$
$$\int_{I,N,R} dI\ dN\ dR\ \Pr(O|I) \cdot \Pr(I|N, R, L) \cdot \Pr(N, R|S, T, P) \cdot \Pr(L, P) \cdot \Pr(S, T|F) \cdot \Pr(F), \quad (1)$$

where $\Pr(N, R|S, T, P)$, $\Pr(I|N, R, L)$ and $\Pr(O|I)$ express likelihood terms induced by the mappings Ψ, Φ, and Θ respectively, and we have integrated out the intermediate representations of surface geometry and reflectance N and R, which perceivers do not normally have conscious access to, as well as the ideal face image I. Traditional analysis-by-synthesis methods seek to maximize Eq. 1 by stochastic local search, or to sample from the posterior by top-down Monte Carlo inference methods; all of these computations can be very slow. Instead we consider a bottom-up feedforward recognition model that is trained to directly estimate MAP values for the latent variables, F^*, S^*, T^*, L^*, P^*.

This recognition model comprises a bottom-up hierarchy of functional mappings that parallels (in reverse) the top-down hierarchy of the generative model, and exploits the conditional independence structure inherent in the generative model for efficient modular inference. In general, if a random variable (or set of variables) Z renders two (sets of) variables A and B conditionally independent in the generative model, and if our goal in recognition is to infer A from observations of B, then an optimal (maximally accurate and efficient) feedforward recognition model can be constructed in two stages that map B to Z and Z to A respectively(Stuhlmüller et al., 2013; Lin et al., 2017). Here our recognition model exploits two such crucial independence relations: (i) The observable raw image is conditionally independent of the 2.5D face components, given the ideal face image and (ii) The 2.5D components are conditionally independent of person identity, given the 3D scene parameters that describe the individual's face. This conditional independence structure suggests a recognition network with three main stages, which can be implemented in a sequence of deep neural networks where the output of each stage's network is the input to the next stage's network.

The first stage segments and normalizes the input image to compute the attended face image, i.e., the most probable value for the ideal image I^* given the observed image O, by maximizing $\Pr(I|O)$ using a DCNN module trained for three-dimensional face segmentation (Jackson et al., 2017) and adapted to compute the face region given images of faces with background clutter (f_1 in Fig. 1B; SOM Section 2.1).

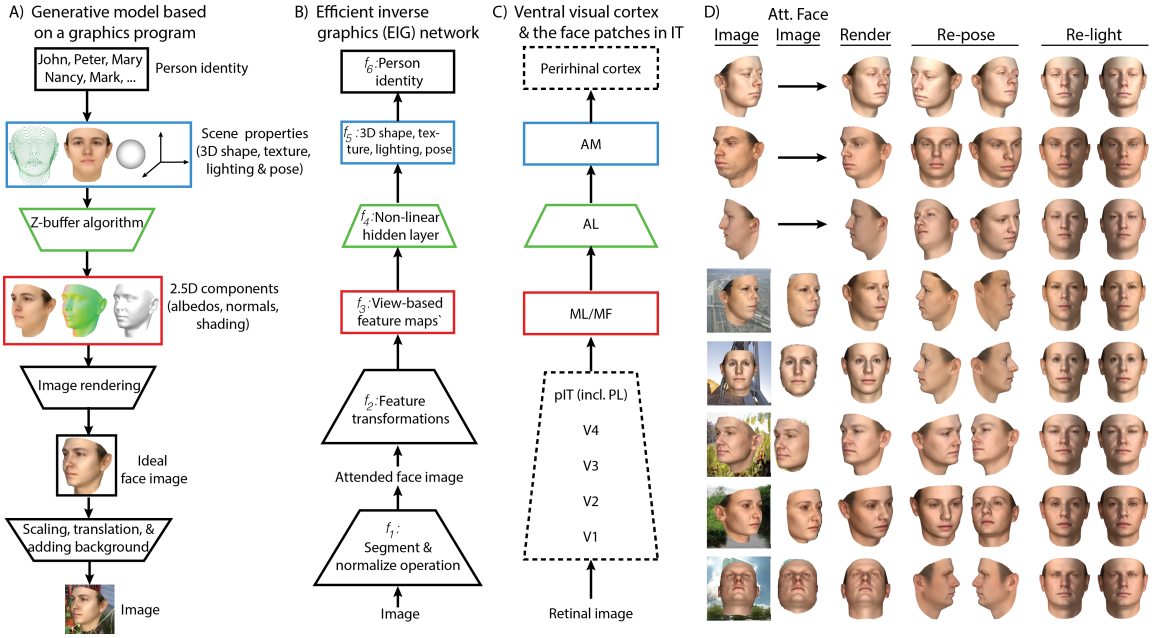

Figure 1: Overview of the modeling framework. (A) Schematic of the probabilistic generative that consists of a distribution over familiar person identities, detailed view of a graphics program with its key stages exposed, and an image-level transformations module. (B) A recognition model for efficient inference in the generative model that is based on DCNNs. f_1 is for face segmentation and zooming, f_2 to f_5 for 3D scene parameter inference, and finally f_6 for person identity recognition. (C) Schematic of the ventral stream with the three face patches indicated. Colored boxes in (A) to (C) show the hypothesized explanations of the neural sites based on the generative and recognition models. Rectangles indicate representations, trapezoids indicate transformations or algorithms using representations. (D) Example results of inference using the recognition model. Inferred scene parameters are rendered, re-posed, and re-lit using the generative model. Results are shown for images without backgrounds skipping the face segmentation step and for images with backgrounds where the output of f_1 applied to the input is also shown.

The second stage is the heart of our EIG model, and consists of a second DCNN module trained to estimate intrinsic and extrinsic scene properties $\{S^*, T^*, L^*, P^*\}$ maximizing $\Pr(S, T, L, P|I^*)$ from the attended face image. This network is adapted from the architecture of a standard DCNN (Krizhevsky et al., 2012) for object recognition, which consists of four convolutional layers (f_2 in Fig. 1B) ending in a fifth, top convolutional feature space (TCL, f_3 in Fig. 1B), followed by two fully connected layers (FCLs, f_4 and f_5 respectively). The second FCL f_5 is the key difference from the conventional object recognition pipeline: instead of being trained to predict class labels, f_5 is trained to predict scene properties, $\{S, T, L, P\}$. Training begins from a pre-trained version of the basic architecture, fixing or fine-tuning weights up to layer f_4, with only weights in the new scene property layer f_5 being learned from random initial values. Training images for stage two were generated by forward-simulating (or "dreaming") images drawn from the generative model (in the spirit of the Helmholtz machine (Hinton et al., 1995)), each with a different randomly drawn value for the scene parameters $\{S, T, L, P\}$, and using the generative model to produce the corresponding ideal face image I conditioned on those scene parameters (SOM Section 2.2).

Finally, a third recognition stage estimates the most likely face identity label F^* given the scene properties, maximizing $\Pr(F|S^*, T^*, L^*, P^*)$. This module comprises a single new FCL f_6 for person identity classification, and like the previous module is trained (with fine-tuning of f_5) on another self-generated set of simulated faces drawn randomly from the generative model but starting from the prior over individuals $P(F)$, which can be specific to a particular set of faces encountered in an individual experiment (see SOM Section 2.3).

Together these three modules form a complete recognition model for the generative model of face images, which satisfies the crucial characteristics of face perception and perceptual systems more generally: The recognition model (i) infers both rich 3D scene structure and the identities or class labels of individuals present in the scene, in a way that is robust to many dimensions of image variation and clutter, and (ii) computes these inferences in a fast, almost instantaneous manner given observed images.

The model's inferences are both qualitatively reasonable (Fig. 1D) and quantitatively accurate (SOM Section 1.2; Fig. S1), suggesting it is a good functional solution to the problem of face perception (but see SOM Section 2.4 for a discussion of potential weaknesses as well). In the remainder of the paper, we ask how well the model captures the mechanisms of

face perception in the mind and brain, by comparing its internal representations (especially f_3, f_4, f_5) to neural representations of faces in the primate ventral stream, and its estimates of intrinsic and extrinsic face properties with the judgments of human observers in several hard perceptual tasks.

Efficient inverse graphics stages explain macaque face-processing hierarchy

The best-understood neural architecture on which we can evaluate EIG as an account of perception in the brain is the macaque face-processing network (Freiwald and Tsao, 2010) (Fig. 2A-i; see SOM Section 4.1 and 4.2 for experiment and neural data analysis details). This three-level hierarchy exhibits a systematic progression of tuning properties: neurons in the bottom-level face patches ML/MF are tuned to specific poses; those in the mid-level patch AL exhibit mirror-symmetry for pose; and those in the top-level patch AM exhibit view-robust identity coding (Fig. 2A-ii). It has also been argued that these neural populations encode a multidimensional space for face, based on controlled sets of synthetically generated images. (Leopold et al., 2006; Freiwald et al., 2009; Chang and Tsao, 2017). However, it remains unclear how the full range of three-dimensional shapes and appearances for natural faces viewed under widely varying natural lighting and view conditions might be encoded, and how high-level face space representations are computed from observed images through the multiple stages of the face-processing hierarchy.

We address these questions by first quantifying the population-level tuning properties for three successive levels of face patches, ML/MF, AL and AM, using linear combinations of three idealized similarity templates representing the abstract properties of view specificity, mirror symmetry, and view-invariant identity selectivity (Kietzmann et al., 2012; Guntupalli et al., 2016) (Fig. 2A-iv, SOM Section 4.5) to fit the empirical similarity matrices for neural populations in each of these patches. The weights of these different matrices measure, in objective terms, how view-specificity decreases from ML/MF to AM, how mirror-symmetry peaks in AL, and how view-invariant identity coding increases from ML/MF to AL and further to AM (Fig. 2A-iii), complementing the qualitative features shown in the population-level similarity matrices (Fig. 2A-ii).

We then evaluated the ability of the EIG network and other models to explain both these qualitative and quantitative tuning properties of ML/MF, AL and AM. In particular we contrast EIG with several variants of the VGG network, a state-of-the-art DCNN for machine face recognition built via supervised training with millions of labeled face images from thousands of individual identities (Parkhi et al., 2015) (SOM Section 3.2). These comparisons allow us to tell apart the latent variable hypothesis and the discriminative hypothesis of the neural representations.

We first test all models using the FIV image set, the set of 175 face images (25 individuals in 7 poses) shown to monkeys during neural recording (Fig. 2B). Given these stimuli, the EIG network (Fig. 2C-i) faithfully reproduces all patterns in the neural data, both qualitatively (Fig. 2C-ii) and quantitatively using the idealized similarity matrix analysis (Fig. 2C-iii). The EIG model also closely tracks the functional compartmentalization observed in the face-processing hierarchy: layer f_3 best correlates with the ML/MF, layer f_4 best correlates with AL, and layer f_5 best correlates with AM ($p < 0.05$, Fig. 2C-iv).

Comparing EIG to VGG (Fig. 2E-i) allows us to evaluate the discriminative hypothesis, and comparing EIG to a lesion ("EIG$^-$", Fig. 2D-i) that omits the initial object segmentation (f_1) lets us test whether this stage of the recognition model (which has not been a component of previous ventral stream models (DiCarlo et al., 2012; Serre et al., 2007; Yamins et al., 2014)) is needed. Across the top three levels of each network (f_3, f_4 and f_5 for EIG and EIG$^-$; TCL, FFCL and SFCL for VGG), only the full EIG network gave the best fit at each model-level to the corresponding level of neural processing (ML/MF, AL, and AM, respectively; panels (iv) of Figs. 2C, 2D, 2E); the full EIG model also correlated more highly than either alternative model with the corresponding level of neural data ($p < 0.05$ in all cases except for AM, where fits were not significantly different). Both VGG and EIG$^-$ gave rise to patterns of selectivity with some qualitative similarity to those of the neural data (Fig. 2D-ii, 2E-ii), but with pronounced quantitative differences. Both alternative models were substantially more view-invariant in their first and second stages (f_3, f_4 for EIG$^-$ and TCL, FFCL for VGG) when compared to either the neural data ($p < 0.05$; Fig. 2D-iii, 2E-iii) or the full EIG model ($p < 0.05$). Most dramatically, for both alternative models, the two highest layers (f_4, f_5 and FFCL, SFCL) were almost indistinguishable from each other, which fails to reflect the clear distinction of function in both the corresponding

neural sites (AL and AM) and the corresponding layers of EIG. In short, only EIG captured the full progression of three functionally distinct stages from ML/MF through AM, suggesting that the face processing network begins with an initial face segmentation stage, and culminates in targets that encode 3D scene properties rather than features optimized for identity discrimination.

To better understand the differences between models, we tested EIG, EIG$^-$, VGG, and three variants of the VGG architecture using a synthetic analog of the FIV image set (FIV-S; SOM Section 1.1), in which only faces were rendered (without clothing or backgrounds) in different viewing and lighting positions. Using these synthetic faces gives us full control over how each network is trained, and lets us unconfound the influences of network architecture, training set, and loss function (or training objective; see SOM Sections 3.2, 3.3, 3.4, and 3.5 for these controls). We find additional support for the inverse-graphics account of the primate face patch network (the "latent variables" hypothesis): The classic neural selectivity patterns across all three levels of ML/MF, AL and AM appear pristinely in the EIG network tested on synthetic faces, and arise uniquely when a recognition model is trained with targets that are 3D scene properties – that is, when the network is trained to infer the 3D shape and texture inputs to a causal generative model of observed face images (see Fig. S2 and SOM Section 3.1).

Finally we ask whether intermediate stages of the face-processing hierarchy, ML/MF and AL in the primate brain or f_3 and f_4 in the EIG network, can be given an interpretable representational account as we did for AM and f_5, or whether instead these patches are best understood simply as a hierarchy of "black box" function approximators whose responses arise just as the locally optimal parameterization of a deep recognition architecture that has been trained to infer the 3D shape and texture properties of faces at the level of AM/f_5. Fig. 1 suggests one possible interpretation based on correspondences between the graphics and inverse-graphics pathways: ML/MF could be understood as computing a reconstruction of an intermediate stage of the generative model, the 2.5D components of a face (e.g., albedos and surface normals) analogous to the "intrinsic images" or "2.5D sketch" of classic computer vision systems (Marr, 1982; Barrow and Tenenbaum, 1978). It is also possible that these patches compute a reconstruction of an earlier stage in the generative model such as the attended face image (corresponding to the output of f_1), or that they are just

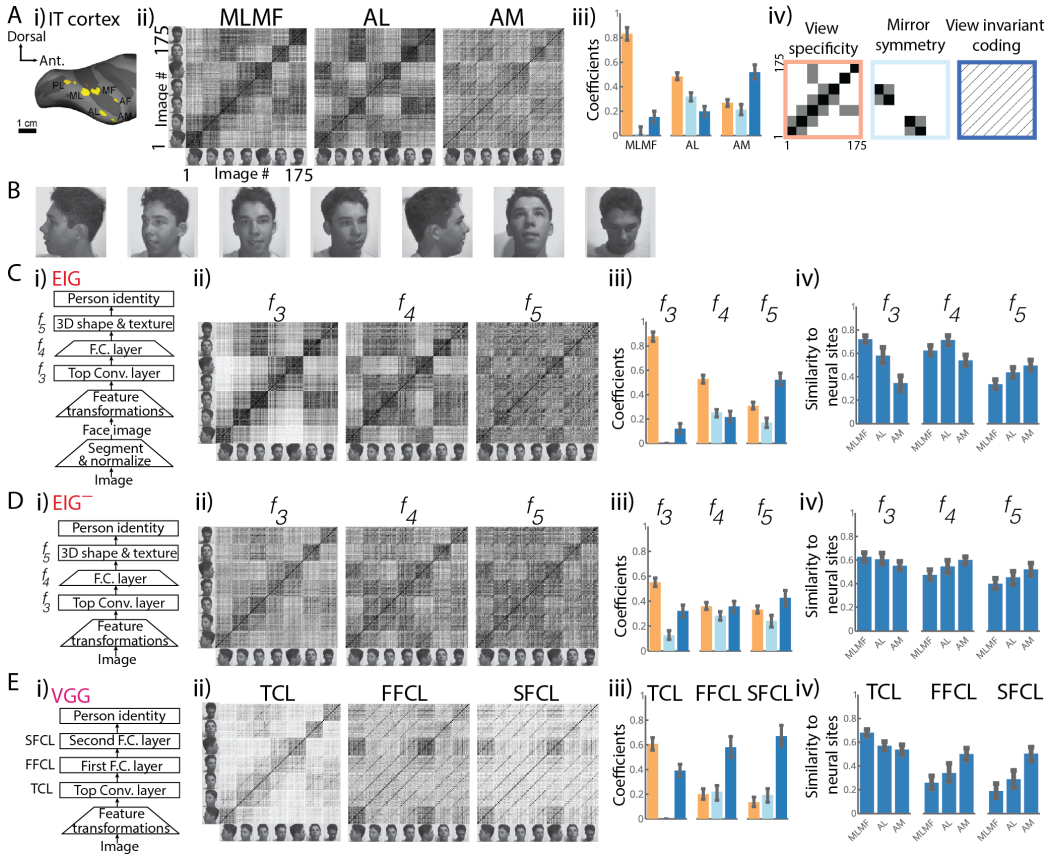

Figure 2: Inverse graphics in the brain. (A) (i) Inflated macaque right hemisphere showing six temporal pole face patches, including middle lateral and middle fundus areas ML/MF, anterior lateral area, AL, and anteriormedial area, AM. (ii) Population similarity matrices from face patches, from left to right: ML/MF, AL, and AM. Each matrix shows 175 × 175 population response correlation coefficients to images of 25 individual identities shown in seven different poses (FIV image set) (Freiwald and Tsao, 2010). (iii) Normalized coefficients resulting from a linear decomposition of the population similarity matrices in terms of idealized similarity matrices for view-specificity, mirror-symmetry, and view-invariance shown in (iv), in addition to a constant background factor to account for overall mean similarity. (B) Sample images from the FIV image set illustrating one of the 25 identities. (C) Full EIG network tested with FIV image set. (i) Schematic of the EIG network with its key layers indicated, f_3, f_4, and f_5. (ii) Layer-wise similarity matrices. On a given layer, each entry is the Pearson's r value between two vectors, where each vector is the activation of all units given an image. (iii) Normalized linear regression coefficients for the idealized similarity matrices. (iv) Correlation coefficients between each of the population similarity matrices and each of the layer-wise similarity matrices. (D) EIG$^-$ network tested with FIV image set. (E) VGG network tested with FIV image set. Sub-figures follow the same convention. Error bars show 95% bootstrap confidence intervals (CIs; SOM Section 4.6).

stepping stones to higher-level representations without distinct functional interpretations in terms of the generative graphics model. We computed similarity matrices for each of these candidate interpretations (each generative model stage), as well as for the raw pixel images as a control (Fig. 3A; see SOM Section 3.6 for how 2.5D components of the FIV images are approximated). We then correlated these similarity matrices with those for ML/MF and AL. We find that the 2.5D components best explain ML/MF ($p < 0.001$), and closely resemble their overall similarity structure (Fig. 3B). Attended images also provide a better account of ML/MF than the raw pixel images ($p < 0.001$) but significantly worse than the 2.5D components ($p < 0.001$ for each component; Fig. 3B). We also find that the 2.5D components explain f_3 layer responses in the EIG model better than the raw pixel images, and better than the attended face image when these can be discriminated (SOM Section 3.6, Fig. S3).

AL has no such straightforward representational account, but it may be understood as implementing a densely connected hidden-layer mapping the estimated 2.5D face components (in ML/MF and f_3) to estimated 3D face properties (in AM and f_5). Because this transformation is highly nonlinear, some kind of hidden layer is required in any feedforward recognition network (Rumelhart et al., 1985; Leibo et al., 2017), and this could be the role of AL in the primate brain and the corresponding layer f_4 in EIG. Note that such an intermediate layer appears to be functionally missing from VGG and its variants trained to discriminatively predict identity rather than 3D object properties. These models always show very similar responses in all their fully connected layers (compare Fig. 2C and 2E and also see Fig. S2D-G). We conjecture that this AL-like intermediate stage nonlinearity is not necessary because the fully connected layers of VGG are solving a different task than EIG or the brain: VGG appears to be mapping high-level image features (computed at the top of the convolutional layers) to person identities which are almost linearly decodable from these features, without ever having to explicitly represent the 3D properties of a face (SOM Section 3.7, Fig. S4).

Efficient inverse graphics scene parameters predict human behavior

We also tested EIG and alternative models' ability to explain human face perception, by comparing their responses to people's judgments in a suite of challenging unfamiliar face

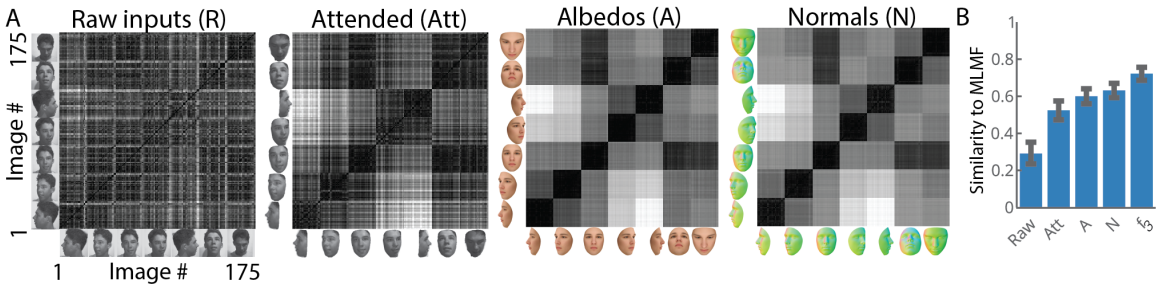

Figure 3: Understanding ML/MF computations using the generative model and the 2.5D (or intrinsic image) components. (A) Similarity matrices based on raw input images, attended images, albedos, and normals. Colors indicate the direction of the normal of the underlying 3D surface at each pixel location. (B) Correlation coefficients between ML/MF and the similarity matrices of each image representation in (A) and f_3. Error bars indicate 95% bootstrap CIs.

recognition tasks(Hancock et al., 2000). In three experiments (inspired by the passport photo verification task), subjects were asked to judge whether two sequentially presented face images showed the same or different identity (Fig. 4A). In Experiment 1 ("Regular"), both study and test images were presented with pose and lighting directions chosen randomly over the full range covered by the generative model. Experiments 2 and 3 probed generalization abilities, using the same study items from Experiment 1 but test items that extended qualitatively the range of training stimuli. In Experiment 2 ("Sculpture"), the test items were images of face sculptures (i.e., texture-less face shapes in frontal pose) eliminating all of the texture information in the input. In Experiment 3, the test items were flat frontal facial textures and distorted using a fish-eye lens effect to reduce shape information in the input (SOM Sections 5.1-3). We hypothesized that if face perception is based on inverting a generative model with independent shape and texture latents, as in EIG but not the VGG or VGG-FT models, participants might be able to selectively attend to shape or texture latents in order to optimize performance.

We compared all three of these models' predictions with human judgments, but the comparison between EIG and VGG-FT is especially revealing. These two models are both trained using an equal number of images synthesized from the same graphics program used to generate the stimuli (although VGG-FT is fine-tuned on top of the VGG network which itself is trained with millions of other face images), but their training targets are

different: EIG's target is a set of scene parameters, the latent variables of the generative face model (latent variables hypothesis), while VGG-FT's target is an embedding space for discriminating person identities (discriminative vectors hypothesis). This allows our behavioral results, like the neural data, to test between these two different hypotheses about the functional goal of face perception.

We compared average human responses – i.e., Pr("Same"), frequency of the "same" response – to the models' predicted similarity across trials. A model's predicted similarity for a given trial was computed as the similarity between the model's outputs (i.e., its top layer) for the study and test items (SOM Section 5.4). The VGG and VGG-FT networks' outputs for an image is their identity-embedding spaces, the layer SFCL. (We found that no other layer in the VGG network provided a better account of the human behavior than its SSFL layer.) EIG's output is its shape and texture parameters, which unlike other models supports selective attention to these different aspects of a face. For each experiment we fit a single weight for the shape parameters in EIG's computation of face similarity (constant across all trials and participants); the weight of the texture component is 1 minus that value.

Overall, participants performed significantly better than chance (average percent correct performance across experiments were 66% in Experiment 1, 64% in Experiment 2, and 61% in Experiment 3; see SOM Section 5.1-3 for further behavioral analysis). In trial-by-trial comparisons to behavior, EIG consistently predicted human error patterns across all three experiments, with r values $0.70[0.65, 0.76]$, $0.64[0.58, 0.69]$, and $0.54[0.47, 0.61]$ (where $[l, u]$ indicates lower/upper 95% confidence intervals; Fig. 4B). EIG also exhibited better generalization when the test images were qualitatively different from the natural texture or shape of the study images, fitting human judgments significantly better than both alternative models in Exps. 2 and 3 ($p < 0.001$ for all comparisons based on direct bootstrap hypothesis tests; SOM Section 5.5). We found that EIG's inferred shape weights were not significantly different from 0.50 (uniform weighting of shape and texture parameters) for Experiments 1 and 3, but it attended to shape parameters almost exclusively for Experiment 2 (mean value of $0.95[0.84, 1]$; Fig. 4C). These results show that EIG predicts the patterns of human face perception more accurately than other models, especially under atypical

stimulus conditions, and lends further support to the latent variables hypothesis over the discriminative hypothesis.

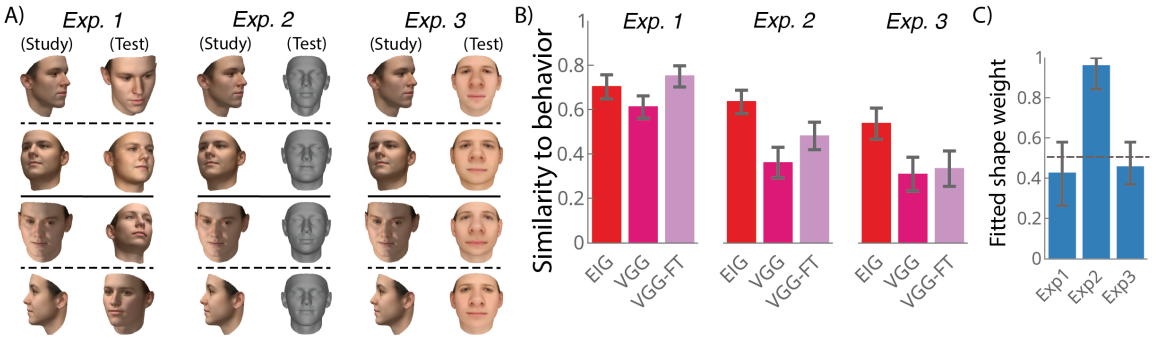

Figure 4: Across three behavioral experiments, EIG consistently predicts human face recognition performance. (A) Example stimuli testing same-different judgments ("same" trials rows 1-2, "different" trials rows 3-4) with normal test faces (Exp. 1), "sculpture" (texture-less) test faces (Exp. 2), and fish-eye lens distorted shade-less facial textures as test faces (Exp. 3). (B) Correlations between model similarity judgments and human judges' probability of responding "same". (C) Best-fitting weights of the shape latents (relative to texture latents) in the EIG model predictions for Exps. 1-3. Note in Exp. 3, in addition to shape distortion, texture is also distorted due to a different lighting and rendering mode used for generating the test items. Error bars indicate 95% bootstrap CIs (SOM Section 5.5).

Human face perception is susceptible to illusions, and our model naturally captures one of the most famous. In the hollow face illusion, a face mask reversed in depth (so the nose points away from the viewer) and lit from the top or side appears to be a normally shaped face, but with the lighting coming from the bottom or alternate side, respectively(Gregory, 1970). The illusion is often explained by saying that strong top-down priors on the shapes of faces bias how we interpret otherwise ambiguous cues to depth in images, such as shading patterns which result from the interaction of surface normals and lighting direction (Gregory, 1997). Our model provides a related but distinct account, in which the prior on convexity for face shapes is implicitly encoded in the bottom-up weights of the EIG network which has learned to jointly extract both intrinsic face properties such as shape as well as extrinsic scene properties such as lighting direction. We quantitatively compared our model's inferences about lighting direction with people's judgements, in both graded versions of the hollow face illusion and normal lighting direction variation, as a control (Fig. 5). We found

that the EIG network, like humans, perceived the light source direction to covary illusorily with graded reversal of the face depth map, in a highly nonlinear pattern inflecting just when depth values turned negative; in contrast, varying lighting direction in a normal way while keeping face shape constant (the control condition) was perceived linearly and largely veridically by both people and the model. That the EIG model captures the nonlinear interaction of depth and lighting percepts in the hollow face illusion, as well as the fact that these percepts are formed nearly instantenously upon seeing the stimulus, provides further evidence that scene perception is implementing some form of analysis-by-synthesis via an efficient bottom-up recognition network rather than slow top-down hypothesis testing mechanisms.

Discussion

Our results suggest that the primate ventral stream approaches face perception – and perhaps object perception more generally – with an "inverse graphics" strategy implemented approximately but efficiently in a feedforward hierarchical network: Observed images are mapped via a segmentation and normalization mechanism to a 2.5D-like map of intrinsic surface properties (view-centered geometry and albedo) represented in ML/MF, which is then mapped via a nonlinear transform through AL to a largely viewpoint-independent representation of 3D object properties (shape and texture) in AM. The EIG network simulates this process and captures the key qualitative and quantitative features of neural responses across the face-patch system, as well as human perception for both typical and atypical face stimuli. The EIG model thus suggests how the structure of the visual system is optimized for its function: computing a rich and accurate representation of the shape and texture of a novel object from an input image in a single, fast, feedforward pass.

Our results are consistent with strong evidence that neurons in areas ML/MF and AM code faces in terms of a continuous "shape-appearance" space Chang and Tsao (2017), not simply discrete identities. However, the EIG model goes beyond this finding to address core, long-standing questions of neural computation: How is the ultimate percept of an object (or face) derived from an image via a hierarchy of intermediate processing stages, and why does this hierarchy have the structure it does? EIG is an image-computable model that faithfully reproduces representations in all three face patches of ML/MF, AL and AM, and explains

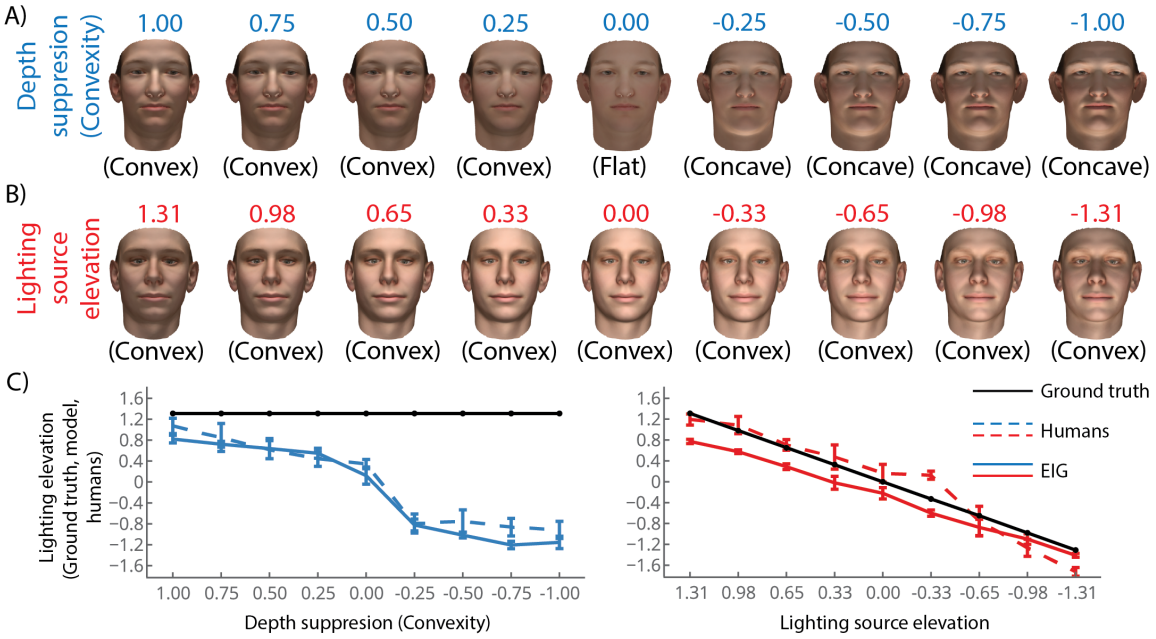

Figure 5: Light source localization task. On a given trial, participants saw an image of a face lit by a single light source and judged the elevation of the light source using a scale from 1 (bottom of the face) to 7 (top of the face; also see SOM Section 5.6). (A) One group of participants were presented with images of faces that were always lit from the top, but where the shape of the face was gradually reversed from a normally shaped face (convexity=1) to a flat surface (convexity=0) to an inverted hollow face (convexity=-1). (B) Another group of participants were presented with images of normally shaped faces (convexity=1) lit from one of the 9 possible elevations ranging from the top of the face to the bottom. (C) Normalized average ratings of the depth-suppression participants, EIG's predicted light source elevation ($r = 0.95, p < 0.01$), and the ground truth light source location. (D) Normalized average ratings of the lighting source elevation participants, EIG's predicted light source elevation ($r = 0.95, p < 0.01$), and the ground truth light source location. Error bars indicate one standard deviation.

mechanistically how each stage is computed. It also suggests why these representations would be computed in the sequence observed, in terms of a network for moving from 2D images to 2.5D surface components to 3D object properties, to efficiently invert a generative model of how face images are observed in 3D scenes. The model thus gives a systems-level functional understanding of perhaps the best characterized circuitry in the higher ventral stream.

Anatomical connectivity and temporal dynamics of responses in the face patches suggest extensive feedback and other non-hierarchical connectivity that our current model does not capture (Grimaldi et al., 2016). Following earlier models of primate face and object processing (Serre et al., 2007; Yamins et al., 2014; Leibo et al., 2017), however, we see a feedforward hierarchical network such as EIG as only a first approximation of the system's functional architecture, reflecting the first feedforward pass of spiking activity through the ventral stream. The EIG model naturally extends to architectures involving feedback or skip connections, which can support more powerful perceptual inferences as we discuss below.

The EIG model also has broader implications for neuroscience, perception, and cognition. Beyond the specific domain of faces, the finding that IT cortex supports decoding of category-orthogonal shape information in addition to object category identity (Hong et al., 2016) suggests that an extension of EIG could account for how the brain perceives the three-dimensional structure of objects and scenes more generally. This perspective also suggests a resolution to the problem of interpretability in systems neuroscience (Yamins and DiCarlo, 2016): Today's best performing models are remarkable for their ability to fit stimulus-dependent variance in neural firing rates, but often without an interpretable explanation of what those neurons are computing. Our work suggests that in addition to maximizing variance explained, computational neuroscientists could aim for "semi-interpretable" models of perception, in which some neural populations (such as ML/MF and AM) can be understood as representing stages in the inverse of a generative model (such as 2.5D components and 3D shape and texture properties), while other populations (such as AL) might be better explained as implementing necessary hidden-layer (nonlinear) transforms between interpretable stages.

In offering a solution to the problem of how scene percepts can be so rich in content, yet so fast to compute from observed images, EIG builds on a more general class of approaches to efficient analysis-by-synthesis including Helmholtz machines (Dayan et al., 1995; Nair et al., 2008) and the more recent variational autoencoders (VAEs) (Kingma and Welling, 2015; Kulkarni et al., 2015b). In these approaches, as in EIG, patterns of inference in a top-down generative model are learned by a complementary bottom-up recognition network, which can then approximate the generative model's inferences on new inputs without going through costly iterative computations. These earlier approaches, however, used generative

models learned implicitly with generic function approximators rather than instantiated in an explicit graphics model as we do; this has the advantage of being very general, but the weaknesses of producing only rough approximations to natural-looking images for faces or any given real-world class of objects, and only approximately untangling the underlying factors of variation or the independent degrees of freedom for objects in a scene. Our work shows that inverse graphics networks can be implemented for much richer, structured generative models based on a graphics model capable of generating realistic images for a wide range of objects in natural scenes. We showed this specifically for one important class of natural objects, faces, but our network is general and could be applied to other object classes thought to have have functionally specific brain representations (bodies, hands, word forms), as well as objects more generally.

Our approach can be extended in a number of directions important for human and machine perception. EIG networks can be augmented with multiple scene layers in order to parse faces (or other objects) under occlusion (Yildirim et al., 2017; Moreno et al., 2016). They can be deployed in parallel or in series (using attention) to parse out multiple objects in a scene (Romaszko et al., 2017; Eslami et al., 2017; Wu et al., 2017a). They can even be extended to other modalities through which we perceive physical objects, such as touch, and can support flexible crossmodal transfer, allowing objects that have only been experienced in one modality (e.g., by sight) to be recognized in another (touch). We have already developed several of these extensions in the domain of faces (Yildirim et al., 2017), but much more work remains to be done to explore the full potential of the EIG approach.

Most intriguingly, our work suggests a clear role for causal models of image formation in the visual system, and in perception more generally. Unlike many leading accounts of visual recognition based on deep networks for classification (DiCarlo et al., 2012; Serre et al., 2007; Krizhevsky et al., 2012), our approach naturally supports explicit representations of physical objects in terms of their 3D shape and other properties (e.g, substances they are made of); these basic components of generative models based on graphics engines become the targets for training inverse-graphics recognition networks. Generative models in the brain could also support other functional roles: They could be used during online perception to refine a percept – particularly in hard cases such as under dim light or under heavy occlusion– by enforcing re-projection consistency with intrinsic image based surface representations (Wu

et al., 2017b; Yildirim et al., 2015; Kulkarni et al., 2015a; Wu et al., 2017a), and they could also support higher functions in cognition such as mental imagery, planning, and problem solving (Battaglia et al., 2013; Wu et al., 2015). It remains to be determined which of these functions of generative models are actually operative in the brain, as well as where and how generative models might be implemented in neural circuit. VAEs, and their close cousins GANs (Goodfellow et al., 2014) and capsules (Sabour et al., 2017), as well as RCNs (George et al., 2017), are recent developments in deep neural networks that suggest at least partial hypotheses for how graphics models might be implemented neurally, but none of these suggestions are yet well grounded in experimental work. We hope that the success of the EIG approach here will inspire future work to explore potential neural correlates of these architectures, as well as the other roles that generative models could play in perception, cognition, and learning.

Acknowledgments

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216; the National Eye Institute of NIH (R01 EY021594 to W.A.F.); The New York Stem Cell Foundation (to W.A.F.); 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.). W.A.F. is a New York Stem Cell FoundationRobertson Investigator. A high performance clustering environment for computations (Openmind) was provided by the McGovern Institute for Brain Research. We thank Michael Janner and Tejas Kulkarni for many helpful discussions. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

Jonathan Barron and Jagannath Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

HG Barrow and JM Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, page 2, 1978.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110 (45):18327–18332, 2013.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

Le Chang and Doris Y Tsao. The Code for Facial Identity in the Primate Brain. *Cell*, 169 (6):1013–1028, 2017.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The Helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Goker Erdogan and Robert A Jacobs. Visual shape perception as bayesian inference of 3d object-centered shape representations. *Psychological Review*, 2017.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. 2017.

Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

Winrich A Freiwald, Doris Y Tsao, and Margaret S Livingstone. A face feature space in the macaque temporal lobe. *Nature neuroscience*, 12(9):1187, 2009.

D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, A. Lavin, and D. S. Phoenix. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 2017. ISSN 0036-8075. doi: 10.1126/science.aag2612. URL `http://science.sciencemag.org/content/early/2017/10/26/science.aag2612`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Richard L Gregory. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 352(1358):1121–1127, 1997.

Richard Langton Gregory. The intelligent eye. 1970.

Piercesare Grimaldi, Kadharbatcha S Saleem, and Doris Tsao. Anatomical connections of the functionally defined face patches in the macaque monkey. *Neuron*, 90(6):1325–1342, 2016.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

J Swaroop Guntupalli, Kelsey G Wheeler, and M Ida Gobbini. Disentangling the representation of identity from head view along the human face processing pathway. *Cerebral Cortex*, 27(1):46–53, 2016.

Peter JB Hancock, Vicki Bruce, and A Mike Burton. Recognition of unfamiliar faces. *Trends in cognitive sciences*, 4(9):330–337, 2000.

Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158, 1995.

Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613, 2016.

A 3D Face Model for Pose and Illumination Invariant Face Recognition, Genova, Italy, 2009. IEEE.

Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *Proceedings of the International Conference on Computer Vision*, 2017.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10 (11):e1003915, 2014.

Tim C Kietzmann, Jascha D Swisher, Peter König, and Frank Tong. Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. *The Journal of Neuroscience*, 32(34):11763–11772, 2012.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Advances in Neural Information Processing Systems*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Tejas D Kulkarni, Pushmeet Kohli, Joshua B Tenenbaum, and Vikash Mansinghka. Picture: A Probabilistic Programming Language for Scene Perception. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4390–4399, 2015a.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015b.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

Joel Z Leibo, Qianli Liao, Fabio Anselmi, Winrich A Freiwald, and Tomaso Poggio. View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation. *Current Biology*, 27(1):62–67, 2017.

David A Leopold, Igor V Bondar, and Martin A Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572, 2006.

Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*, volume 2. 1982.

Pol Moreno, Christopher KI Williams, Charlie Nash, and Pushmeet Kohli. Overcoming occlusion with inverse graphics. In *European Conference on Computer Vision (ECCV) Workshops*, pages 170–185. Springer, 2016.

Vinod Nair, Josh Susskind, and Geoffrey E Hinton. Analysis-by-synthesis by learning to invert generative black boxes. In *International Conference on Artificial Neural Networks (ICANN)*, pages 971–981. Springer, 2008.

Bruno A. Olshausen. Perception as an inference problem. In M. Gazzaniga and R. Mangun, editors, *The Cognitive Neurosciences*.

O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *British Machine Vision Conference (BMVC)*, 2015.

Lukasz Romaszko, Christopher KI Williams, Pol Moreno, Pushmeet Kohli, Jan Czarnowski, Stefan Leutenegger, Andrew J Davison, Renata Khasanova, Pascal Frossard, Iaroslav Melekhov, et al. Vision-as-inverse-graphics: Obtaining a rich 3d explanation of a scene from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 851–859, 2017.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego Institute for Cognitive Science, 1985.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, 2017.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

Jiajun Wu, Ilker Yildirim, Joseph J Lim, William T Freeman, and Joshua B Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*, pages 127–135, 2015.

Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *Advances In Neural Information Processing Systems*, 2017b.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Ilker Yildirim and Robert A Jacobs. Transfer of Object Category Knowledge Across Visual and Haptic Modalities: Experimental and Computational Studies. *Cognition*, 126:135–148, 2013.

Ilker Yildirim, Tejas D Kulkarni, Winrich A Freiwald, and Joshua B Tenenbaum. Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual Conference of the Cognitive Science Society*, 2015.

Ilker Yildirim, Michael Janner, Mario Belledonne, Christian Wallraven, Winrich Freiwald, and Tenenbaum Joshua B. Causal and compositional generative models in online perception. In *Annual Conference of the Cognitive Science Society*, 2017.

Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Supplementary Online Material

Ilker Yildirim,[1,3*] Winrich Freiwald,[2,3*] Josh Tenenbaum[1,3*]

[1]**Department of Brain & Cognitive Sciences, MIT, Cambridge, MA**
[2]**Laboratory of Neural Systems, The Rockefeller University, New York, NY**
[3]**The Center for Brains, Minds, and Machines, MIT, Cambridge, MA**

[*] **Correspondence: ilkery@mit.edu, wfreiwald@rockefeller.edu, jbt@mit.edu.**

1. Probabilistic graphics program

Our generative model builds on and extends the Basel Face Model (BFM) (Paysan et al., 2009), a statistical shape and texture model obtained by applying probabilistic principal component analysis (Tipping and Bishop, 1999) on a data set of 200 laser-scanned human heads. BFM is publicly available and consists of a mean (or norm) face shape, a mean texture, two sets of principal components of variance, one for shape and the other for texture, and their corresponding eigenvectors that projects these principal components to 3D meshes.

The principal components of shape S and texture T accept a standard normal distribution such that $\Pr(S)$ and $\Pr(T)$ are each multivariate standard normal distributions with $S \in \mathbb{R}^{D_S}, T \in \mathbb{R}^{D_T}$. Each sample from $\Pr(S)$ (or $\Pr(T)$) is a vector in a $D = D_S$ (or $D = D_T$) dimensional space specifying a direction and a magnitude to perturb the mean face shape (or the mean texture) to obtain a new unique shape (or texture). Mean shape and texture correspond to $s = \{0, 0, ..., 0\}$ and $t = \{0, 0, ..., 0\}$. (Uppercase letters are used for random variables and lowercase letters are used for assignments of these random variables to draws from their respective distributions. Non-random model parameters, such as D are also uppercase.) We set $D_S, D_T = 200$ in our analysis. We found that the exact

values of D_S and D_T did not matter as long as they were not too small, which leads to very little variation across the samples.

We used the part-based version of BFM where the principal components of shape and texture were partitioned across four canonical face parts: (i) outline of the face, (ii) eyes area, (iii) nose area, and (iv) mouth area. Each face-part (e.g., shape of the nose area or texture of the eyes area, etc.) was represented using $200/4 = 50$ principal components. There are four advantages of using BFM: it (i) allows a separable representation of shape and texture, (ii) provides a probability distribution over both of these properties, (iii) allows us to work with lower dimensional continuous vectors (400 dimensions in this case) as opposed to very high dimensional meshes (meshes consisting of about 1 million vertices), and (iv) consists of dimensions that are often but not always perceptually interpretable (e.g., a dimension controlling the inter-eye distance).

The full scene description in the model also requires choosing extrinsic scene parameters including the lighting direction and viewing direction or head pose. In our simulations, we used Lambertian lighting where the lighting direction L can vary along azimuth L_a and elevation L_e. $\Pr(L_a)$ and $\Pr(L_e)$ are uniform distributions in the range $\{-1.4^{rad}\}$ to $\{1.4^{rad}\}$. The head pose P can vary along the z-axis P_z with $\Pr(P_z)$ a uniform distribution in the range -1.5^{rad} to 1.5^{rad}, and the x-axis P_x with $\Pr(P_x)$ a uniform distribution in the range -0.5^{rad} to 0.5^{rad}. In practice, rotation in the x-axis was conditionally dependent on rotation in the z-axis in order to avoid rendering the back and hollow side of the face mesh. In particular, we truncate $\Pr(P_x)$ whenever P_z takes values greater than 0.75^{rad} or less than $-0.75°$. Finally, we rendered each scene to a 227×227 pixels color image, unless otherwise mentioned.

The behavioral stimuli were generated using the same $\Pr(L)$ and $\Pr(P)$ as described above unless otherwise mentioned (see Section 5).

1.1 Synthetic FIV image sets

The FIV-S stimuli underlying Figs. 2F-I used the pose distributions in Table S1. Each of the 25 identities (i.e., unique pairs of shape and texture properties) were rendered at 7 different poses and with frontal lighting.

Pose Category	Azimuth (P_z)	Elevation (P_x)
Frontal	$N(0, 0.05)$	$N(0, 0.05)$
Right-half profile	$0.75 + N(0, 0.05)$	$N(0, 0.05)$
Right profile	$1.50 + -1 * abs(N(0, 0.05))$	$N(0, 0.05)$
Left-half profile	$-0.75 + N(0, 0.05)$	$N(0, 0.05)$
Left profile	$-1.50 + abs(N(0, 0.05))$	$N(0, 0.05)$
Up	$N(0, 0.05)$	$0.5 + N(0, 0.05)$
Down	$N(0, 0.05)$	$-0.5 + N(0, 0.05)$

Table S1: Pose distributions for the FIV-S image set (in radians).

The image set underlying Fig. S3B (referred to as FIV-S-2) used the same prior over lighting and pose as the generative model, $\Pr(L)$ and $\Pr(P)$. It used the same 25 identities as FIV-S image set and as is the case with FIV-S image set we rendered 7 images (each with its own randomly drawn pose and lighting parameters) per identity making 175 images in total. Additionally, to increase the variability at the levels of raw and attended images, we converted half of these images to gray-scale.

1.2 Conventional top-down inference with MCMC

Given a single image of a face as observation, I, and an approximate rendering engine, $G(\cdot)$ – a combination of the z-buffer $\Psi(\cdot)$ and image rendering $\Phi(\cdot)$ stages introduced in the main text – face processing in this probabilistic graphics program can be defined as inverting the graphics pipeline using Bayes' rule:

$$\Pr(S, T, L, P|I) \propto \Pr(I|I_S) \cdot \Pr(I|S, T, L, P) \cdot \Pr(S, T, L, P) \cdot \delta_{G(\cdot)}$$

where I_S is a top-down sample generated using the probabilistic graphics program, and $\delta(\cdot)$ is a Dirac delta function. (We dropped the corresponding Dirac delta functions in Equation 1 in the main text in order to avoid cluttered notation.) We assume that the image likelihood is an isotropic standard Gaussian distribution, $P(I|I_S) = N(I; I_S, \Sigma)$. Note that the posterior space is of high-dimensionality consisting of more than 400 (404, to be exact) highly coupled shape, texture, lighting direction, and head pose variables, making inference a significant challenge for conventional methods.

Markov chain Monte Carlo (MCMC) methods provide a general framework for inference in generative models which have a long history of application to inverse graphics problems (Yuille and Kersten, 2006). For this specific face model, we have explored both

traditional single-site MCMC and a more efficient multi-site elliptical slice sampler (Murray et al., 2009) to infer the 3D shape and texture vectors given an image, I_D (Kulkarni et al., 2014). Proposals in elliptical slice sampling are based on defining an ellipse using an auxiliary random variable $X \sim N(0, \Sigma)$ around the current state of the latent variables (shape and texture properties), and sampling from an adaptive bracket on this ellipse based on the log-likelihood function. For the lighting direction and pose parameters, single-site Metropolis-Hastings steps are used. At each MCMC sweep, the algorithm iterates a proposal-and-acceptance loop over twelve groups of random variables: four shape vectors (each of length 50), four texture vectors (each of length 50), and four scalars for lighting direction and pose parameters. The detailed form of the proposal and acceptance functions can be found in (Murray et al., 2009). This method often converges to reasonable inferences within a few hundred iterations, although with substantial variance across multiple runs of the algorithm (Fig. S1). In contrast, the EIG algorithm which we describe below and in the main text reliably produces inferences that are as accurate as the best of these MCMC runs (Fig. S1), far more quickly. EIG avoids the need for iterative computation by estimating 3D shape and texture latents via a single feedforward pass through a deep recognition network. Further comparisons between MCMC and efficient recognition networks for inverse graphics (using an earlier version of EIG, without the initial face detection stage and using a more limited training regime and loss function) can be found in (Yildirim et al., 2015).

2. EIG model

The EIG model is a multistage neural network that attempts to estimate the MAP (Maximum A Posteriori) 3D scene properties and identity of an observed face image (approximately maximizing the posterior in Equation 1 of the main text). EIG comprises three recognition modules arranged in sequence to take advantage of the conditional independence structure in the generative (graphics) model. These three modules compute (1) a segmentation and normalization of the face image; (2) an estimate of the 3D face shape and texture; and (3) a classification of the individual whose face is observed.

Below we describe how each of these modules is constructed. The EIG network can also be seen as a multitask network that is designed to solve several tasks at once, including segmentation, 3D scene reconstruction, and identification, where the generative model de-

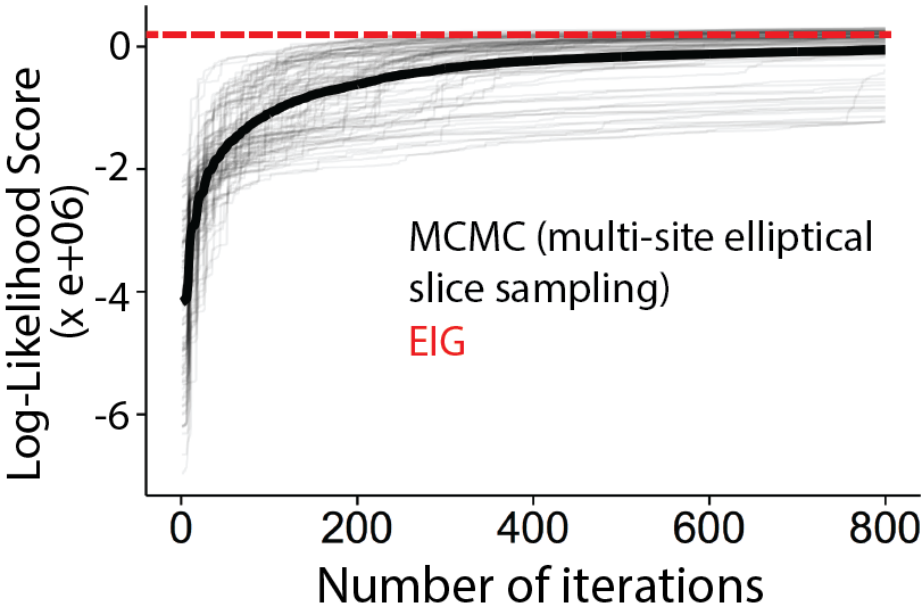

Figure S1: We evaluated the log-likelihood scores $P(I|S,T,L,P)$ of 100 randomly generated images based on their reconstructions using the EIG network's inferred scene parameters (outputs at its layer f_5), and compared them to the evolution of the log-likelihood scores in MCMC. The EIG estimates are computed almost instantaneously, with no iterations, yet achieve a higher score and lower variance (mean score $\sim 2.5 \times 10^5$; standard deviation $\sim 1 \times 10^5$; dashed red line shows the mean) than the MCMC algorithm. In addition, the MCMC algorithm requires a great deal more time because it must perform hundreds of iterations to achieve a similar level of inference quality (mean score $\sim -5 \times 10^5$; standard deviation $\sim 8 \times 10^5$; thick black line shows the mean, thinner black curves show 100 individual runs of the algorithm).

termines which tasks should be solved and the conditional independence structure of the generative model determines the order in which they should be solved.

2.1 Estimating face image given a transformed image, $\mathbf{Pr}(I|O)$

Given an observation consisting of a face image with background, O, MAP inference involves estimating I^* that maximizes $\Pr(I|O)$. This can be achieved by a segmentation of the observed image that only consists of the face-proper region and excludes the rest.

We implemented this inference problem using a convolutional neural network (Fukushima, 1988; LeCun and Bengio, 1995), referred to as f_1 in the main text. We took a recent convolutional neural network with an hour-glass architecture that is trained for volumetric

3D segmentation of faces from images (Jackson et al., 2017). This model takes as input an image and outputs a 3D voxel map where a value of 1 indicates inside the face region and a value of 0 indicates outside the face region. The output of this network is a rough and noisy estimation of the face shape in the form of a voxel grid, V_{xyz}, of dimensions 192 (width) $\times$192 (height) $\times$200 (depth), which in addition to the face-region, also includes filled disconnected regions that are outside the face-proper region.

We adapted this output for accurate 2D segmentation in the following way. We first sum over the depth dimension of V_{xyz} to obtain a 2D map, V_{xy}, of dimensions 192×192. We then binarize V_{xy} (i.e., replace all non-zero entries with 1) and compute its connected components in Matlab. We segment O using the largest connected region of V_{xy} as the mask. Finally, we normalize it by zooming in on the segmented image using bicubic interpolation such that the resulting image's longer dimension is 227. We minimally translate this resized image such that it doesn't exceed the boundaries of the image. In practice, this procedure yields good estimates for I^*.

2.2 Scene parameters given face image, $\mathbf{Pr}(S, T, L, P|I)$

Given a face image as input, MAP inference involves estimating the scene properties (latent variables in the graphics program), $\{S^*, T^*, L^*, P^*\}$ maximizing $\Pr(S, T, L, P|I)$. We accomplished this using a recognition model by learning to map inputs to their underlying latent variables in the graphics program.

Our recognition model is a convolutional neural network obtained by modifying AlexNet's network architecture in the following way (Krizhevsky et al., 2012): we removed its top two fully-connected layers and replaced them with a single new fully-connected layer. Each layer in the network implements a cascade of functions including convolution, rectified linear activation, pooling, and normalization. The details of the resulting network architecture is given in Table S2.

We initialized the parameters of f_2, f_3, and f_4 using the corresponding weights of AlexNet that was pre-trained on a large corpus of images, namely the Places data set (Zhou et al., 2016). The pre-trained network weights are provided by its authors and can be downloaded at
http://places2.csail.mit.edu/models_places365/alexnet_places365.caffemodel. This data set

Type	Patch size/stride	Output size
Convolution (f_{2_1})	11x11/4	96x55x55
Max Pooling (f_{2_2})	3x3/2	96x27x27
Convolution (f_{2_3})	5x5/1	256x27x27
Max Pooling (f_{2_4})	3x3/2	256x13x13
Convolution (f_{2_5})	3x3/1	384x13x13
Convolution (f_{2_6})	3x3/1	384x13x13
Convolution (f_3)	3x3/1	256x13x13
Max Pooling	3x3/2	256x6x6
Full-connectivity (f_4)		1x4096
Full-connectivity (f_5)		1x404

Table S2: Recognition model architecture

consists of about 2.5 million images and their corresponding place labels such as "beach," "classroom," "landscape," etc. (365−way categorization in total). The parameters of the new fully-connected layer (also referred to as scene properties layer or latents layer) were initialized randomly. We chose to use Places data set pre-trained weights to ensure that the network started with already trained but generic visual feature extractors not specifically related to faces. We also avoided using a face corpus pre-trained weights as this would require access to a large labeled data set of weights, which EIG doesn't require.

To learn the mapping from images to their latent variable representations, we drew $200,000$ random samples from the generative model. A random sample is as a tuple of $\{s_i, t_i, l_{a_i}, l_{e_i}, p_{z_i}, p_{x_i}, i_i\}$, each random variable drawn from their respective distributions. Each image was a 227×227 color image with its corresponding target a concatenation of all the latent variables making a continuous vector of length 404 (200 shape properties, 200 texture properties, and 4 extrinsic scene parameters). Using stochastic gradient descent over minibatches of size 20 examples, we finetuned the parameters of f_3, f_4 starting from their pre-trained weights and trained the parameters of f_5 starting from random initialization. The network learns a regression from images to their latent variable vectors based on the mean squared error (MSE) loss function. In our simulations we used a learning rate of 10^{-4}. In order to ensure that gradients were large enough throughout training, we also multiplied the target latent variable vectors by 10. We accounted for this pre-processing step by dividing the outputs of the network by 10 at test time. We trained the model until

the error stopped decreasing on a held-out validation set, which was achieved by 64 epochs of training.

2.3 Person identity given scene parameters, $\mathbf{Pr}(F|S,T,L,P)$

We provide the details of $\Pr(F)$ before describing this final component of the recognition model. In principle, this distribution is over a finite set of familiar individuals but allowing for possibility of encountering a new, unfamiliar individual (Allen et al., 2016). Here, we approximated $\Pr(F)$ as a uniform distribution over a set of familiar individuals. Specifically, we treated $\Pr(F)$ as a multinomial categorical distribution with K outcomes (i.e., K unique person identities) with each outcome equally probable. Each person identity is chosen as a pair of shape and texture properties and denoted as $\Pr(S,T|F)$.

Given scene properties, MAP inference involves estimating the person identity, F^*, maximizing $\Pr(F|S,T,L,P)$. To estimate F^* given scene properties, we extended the recognition model with a new fully-connected layer, f_6, of length K. To learn this mapping from scene properties to identities, we generated a new data set of $K * M$ images where M is the number of times the shape and texture properties associated with each of the K identities were rendered. For each image, we randomly draw the lighting direction and pose properties from their respective prior distributions, $\Pr(L)$ and $\Pr(P)$. In our simulations we set K to 25 and M to 400.

In our FIV experiments, we do not have access to the ground truth shapes or textures of the 25 person identities in that image set and so cannot use the graphics program for generating a training image set. Instead, for a given identity, we obtained the $M = 400$ images by a bootstrapping procedure applied to the whole set of 7 attended images for that identity. Given the image bounding box of the face proper region, we randomly and independently stretched or shrank each side of the bounding box by 15%. We resized the resulting bounding boxes by a randomly chosen scale between $75 - 99\%$. Finally, we translated the resulting bounding boxes in the image randomly but ensuring that the entire face proper region remains in the image. We refer to the resulting image set as bootstrapped FIV image set.

The training procedure was identical for the FIV and FIV-S experiments. We train the new identity classification layer f_6 and finetune the scene properties layer f_5 using

$M * K = 10,000$ images and their underlying person identity labels minimizing the cross-entropy loss (Kevin, 2012). We used a learning rate of 0.005 for training f_6 and half of that value for finetuning f_5. We performed stochastic gradient descent with minibatches of size 20 until the training performance exceeded 99%. Across multiple training sessions, we found that this took two epochs of training for FIV-S image set and between 13 and 14 epochs for FIV image set.

All of our models are implemented in Torch, and will be made publicly available at https://github.com/iyildirim/efficient_inverse_graphics.

2.4 Weaknesses of EIG

We note two potential weaknesses of the recognition model. First, it may not perform as well when the segmentation step f_1 fails (e.g., too much of the background is left in the attended face image). This is an issue only if the face doesn't cover a spatially significant portion of the input image. Second, the model's reconstruction accuracy may degrade when the observed faces have shapes and textures far from the regions of high prior probability in the generative model, $\Pr(S, T)$. We see these weaknesses mostly as challenges for the model as currently implemented, with a rather limited set of face experiences for training compared to what an individual encounters over the course of their lifetime – let alone what is effectively a much broader base of experience over evolutionary time that also shapes the brain's representations.

The training procedure underlying the third component of our recognition model, $\Pr(F|S, T, L, P)$, helps alleviate the second issue by allowing finetuning of f_5, thereby adjusting $\Pr(S, T, L, P|I)$ to the given training set (e.g., the bootstrapped FIV image set).

3. Alternative models and their evaluations

In this section, we provide the details of the VGG network and its variants and their further evaluations using the FIV-S image set.

3.1 Overall results based on the FIV-S image set

On the FIV-S images, EIG$^-$ (Fig. S2C) fit just as well as the full EIG model (Fig. S2B), in both qualitative and quantitative terms. This confirms our expectation that the face

segmentation stage is needed only to handle background clutter in the image, or hair or clothing that might occlude or distract from face shape and appearance, but that could also provide spurious cues to a familiar person's identity. When tested on the FIV-S images, VGG again showed almost no difference between its top two layers (Fig. S2D), unlike both the neural responses (Fig. 2A, main text) and the EIG networks. VGG also showed less view-invariance overall (Fig. S2D ii, iii; $p < 0.05$) in comparison to its performance on the FIV images. This relative lack of generalization across viewpoint suggests a form of dataset bias. Consistent with this interpretation, a version of VGG fine-tuned to images from the graphics programs (i.e., images similar to the FIV-S faces; referred to as VGG-FT and described further in detail below) gave rise to patterns of results very similar to the regular VGG's results on the more natural FIV image set, with strong view-invariance in the top fully connected layers (Fig. S2E). Crucially, the training data used to fine tune VGG-FT matches EIG's training data, which suggests that the superior fit of EIG to the neural data on the natural FIV faces, relative to VGG, is more likely a consequence of their respective targets as opposed to differences in training or test set distributions. Two other VGG variants that further interpolated towards the EIG$^-$ model were also tested to rule out possible alternative explanations for its lower fit, due to differences in architecture (ID network, described further in detail below; Fig. S2F) or training loss functions (Regress-ID, described further in detail below; Fig. S2G). Taken together, these results on the FIV-S image set positively support the inverse-graphics hypothesis for how the multistage recognition network of primate face perception is organized: classic neural selectivity patterns across all three levels of ML/MF, AL and AM arise uniquely when a recognition model is trained with targets that are 3D scene properties – that is, when the network is trained to infer the inputs to a causal generative model of observed face images.

We now provide the details for each of the VGG variants.

3.2 VGG network

We used the VGG face network (here referred to as VGG network for brevity) that is publicly available,
http://www.robots.ox.ac.uk/ṽgg/software/vgg_face/. This network consists of 13 convolutional layers (8 more layers than AlexNet) and 3 fully-connected layers. The data set used

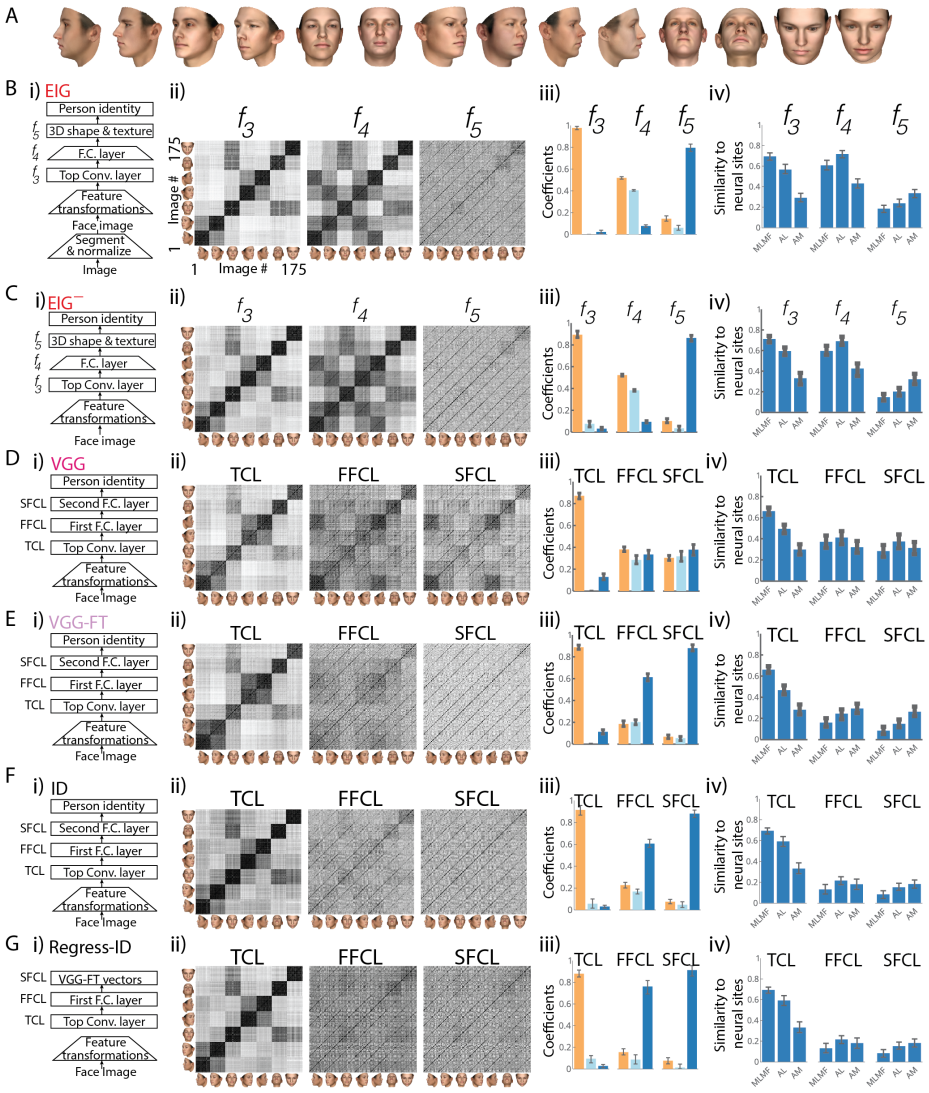

Figure S2: Extending Fig. 2 from main text. (A) Sample images from the FIV-S image set. (B) Full EIG network tested using the FIV-S image set. (C) EIG^{-} network tested with FIV-S image set, which replicates the results with the segmentation step. (D) VGG network, (E) VGG-FT network, (F) ID network, and (G) Regress-ID network tested with the FIV-S image set. Sub-figures follow the same conventions as Fig. 2, main text. Error bars show 95% bootstrap confidence intervals (CIs).

for training this network consisted of more than 2.5 million images where each image is labeled with one of the 2622 person identities. The details of the network architecture, its training data set, and training procedure can be found in (Parkhi et al., 2015).

3.3 VGG-FT network

VGG-FT network was identical to VGG in architecture except for its very last fully-connected layer. We replaced the 2622-dimensional final classification layer (a fully-connected layer of length 2622) in VGG with a 500-dimensional classification layer (a fully-connected layer of length 500).

To train this network, we obtained a new data set where the person identities and training images were generated using the graphics program (Section S1). We first randomly sampled 500 identities as pairs of shapes and textures from $Pr(S, T|F)$. We then rendered each identity using 400 viewing conditions randomly drawn from $Pr(L, P)$, identical to EIG's training data set. This procedure gave us a total of $200,000$ images and their corresponding identity labels (from 1 to 500).

We initialized the weights of VGG-FT using the weights of VGG except for its classification layer, which was initialized using random weights. We then finetuned the weights associated with its TCL, FFCL, and SFCL and trained its classification layer (i.e., its third fully-connected layer; TFCL) using stochastic gradient descent to minimize a cross-entropy loss.

3.4 ID network

We tested two other VGG variants to rule out possible alternative explanations for its lower fit, due to differences in architecture or training loss functions. The first of these, the ID network, is trained using the same data set as VGG-FT, but unlike VGG-FT it shared the same architecture and pre-training weights as EIG. Specifically, the architecture of the ID network was based on AlexNet similar to EIG except for its top layer (i.e., its TFCL), which was a 500-dimensional classification layer. We provide the details of this network's architecture in Table S2.

In the same way as was done for EIG, we initialized the weights of this network using the weights of AlexNet pre-trained on the Places data set, except for its new classification layer. The training data set ($200,000$ images obtained using the graphics program) and training procedure (fine-tuning TCL, FFCL, and SFCL and training TFCL to minimize a cross-entropy loss) were identical to that of VGG-FT.

This model gave rise to patterns similar to that of the VGG-FT network (Fig. S2F).

Type	Patch size/stride	Output size
Convolution	11x11/4	96x55x55
Max Pooling	3x3/2	96x27x27
Convolution	5x5/1	256x27x27
Max Pooling	3x3/2	256x13x13
Convolution	3x3/1	384x13x13
Convolution	3x3/1	384x13x13
Convolution (TCL)	3x3/1	256x13x13
Max Pooling	3x3/2	256x6x6
Full-connectivity (FFCL)		1x4096
Full-connectivity (SFCL)		1x4096
Full-connectivity (TFCL)		1x500

Table S3: ID network architecture

3.5 Regress-ID network

Although the ID network matches EIG in training data and architecture, the loss function it optimizes is different from EIG. We built Regress-ID to further evaluate the discriminative hypothesis using the identical loss function as EIG, the MSE loss. Moreover Regress-ID's training set was the identical set of $200,000$ images we used for training EIG.

The Regress-ID network's architecture is identical to the ID network except for it doesn't have a classification layer (i.e., removing the TFCL layer from the ID network gives the Regress-ID). We paired each image in the training set with a discriminative vector representation as its target. The discriminative vector representations were obtained using the VGG-FT network: for each image, we recorded the SFCL activations of the VGG-FT network. We trained the Regress-ID network to map images to their discriminative vector representations using stochastic gradient descent and a MSE loss. This variant of VGG also gave rise to very similar patterns as its two other variants (Fig. S2G).

3.6 Functionally interpreting ML/MF and f_3 using the generative model

Albedos and normals for each of the 25 person identities in the FIV image set are approximated using EIG and the generative model. The 3D shape and texture properties for each frontal-pose FIV image are inferred using EIG (outputs at f_5). Given the resulting 3D meshes, we obtained the face proper regions by masking out the neck, ears, and hair from the resulting 3D meshes for each identity. Using the generative model, we rendered the 2.5D components of each of the masked meshes at the 7 mean pose values underlying

the extrinsic scene parameter distribution for the FIV-S image set (Table S1). Finally, we adjusted the size and location of faces in the images using the same normalization procedure as the attended images (SOM Section 2.1).

We hypothesized that the random variables in the generative model (Fig. 1A) that are hierarchically below the 3D scene properties each provide a conditional independence stage that could be exploited by ML/MF or AL. We tested this hypothesis using the similarities arising from each of the potential conditional independence stages (Fig 3A, main text): raw input images, attended images, and 2.5D components including albedos and normals. The attended images and 2.5D components are both better accounts of ML/MF than the raw images ($p < 0.001$; raw images $r = 0.29[0.24, 0.35]$, attended images $r = 0.52[0.47, 0.57]$, albedos $0.60[0.56, 0.64]$, and normals $0.63[0.59, 0.67]$) with the 2.5D components providing a significantly better account than the attended images ($p < 0.001$ for each 2.5D component). However, f_3 continued to provide a better account of ML/MF than the 2.5D components ($p < 0.001$).

Similar to our ML/MF results (Fig. 3B), we found that f_3 itself was highly correlated with the 2.5D components to a much better degree than the raw input images (albedos $0.83[0.81, 0.85]$, normals $0.84[0.82, 0.86]$, raw images $0.36[0.28, 0.44]$; $p < 0.001$ for each comparison), with the exception that attended images correlated with f_3 as highly as the 2.5D components ($0.86[0.85, 0.87]$; Fig. S3A). To better understand f_3 and its relationship to attended images, we used the FIV-S-2 image set (SOM Section 1.1) which consists of higher image-level variability in a way that allowed us to tell apart attended images from the 2.5D components. Unlike the attended images, albedos and normals continued to correlate consistently well with f_3 (raw images $0.25[0.23, 0.29]$, attended images $0.46[0.44, 0.50]$, albedos $0.85[0.82, 0.86]$, normals $0.87[0.85, 0.88]$; Fig. S3B). These results collectively suggest that both ML/MF and f_3 can be understood as 2.5D-like surface representations and also suggests use of image sets with broader image-level variability in future experiments for better understanding ML/MF computations.

3.7 Linear decodability of the 2.5D-like representations

We consistently find that VGG and its variants discriminatively trained to estimate face identity (Fig. 2E and Fig. S2D-G) do not produce an AL-like mirror symmetric represen-

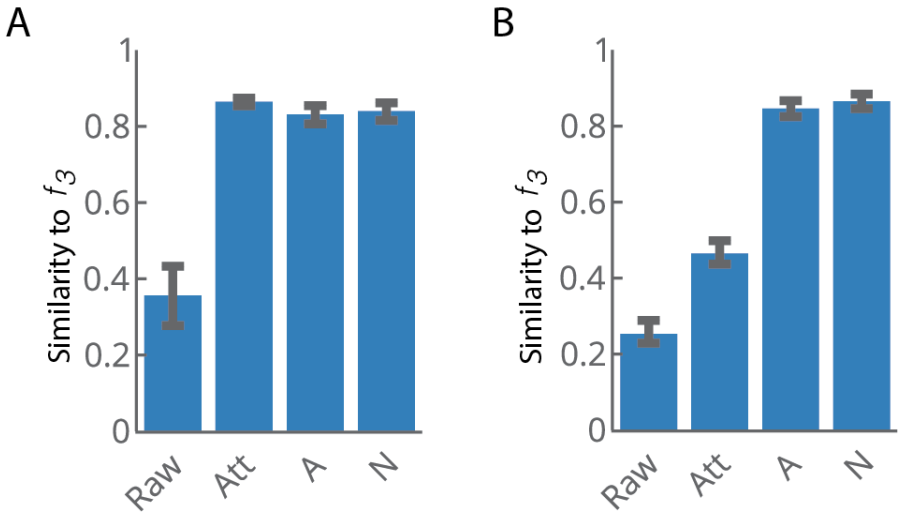

Figure S3: Extending Fig. 3 from main text. (A) FIV image set based comparisons of f_3 similarity patterns to that of raw images, attended images, and the 2.5D components. (B) FIV-S-2 image set based comparisons of f_3 similarity patterns to that of raw images, attended images, and the 2.5D components.

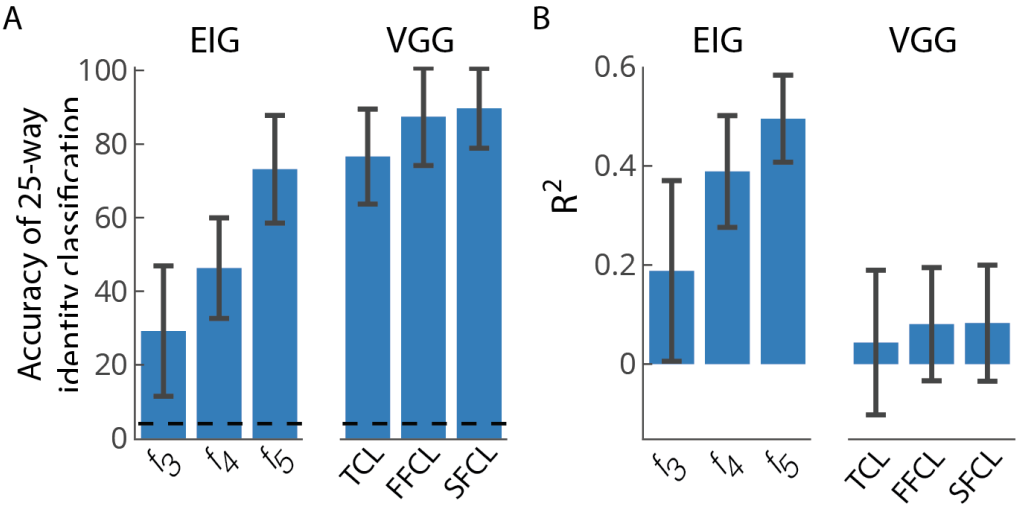

Figure S4: (A) Average accuracy of a 25-way linear classifier decoding FIV identities from the VGG network and the EIG network. Dashed line shows chance performance (4%). (B) Average goodness-of-fit R^2 values resulting from linearly decoding approximate shape and texture properties of the FIV images from the VGG network and the EIG network. Error bars indicate standard deviation. All results are based on held-out test sets (see text for further details).

tation distinct from both the 2.5D-like representation in ML/MF and EIG-f_3 and the 3D scene property representation in AM and EIG-f_5; instead, all fully connected layers of these networks have similar responses with strong viewpoint-invariant identity coding from the first fully connected layer (FFCL) upwards. To explain this, we hypothesized that these discriminatively trained networks are performing a fundamentally different computation in their hidden layers than EIG and the face-patch circuitry: While EIG and the ventral stream appear to need a distinct hidden-layer transformation to solve the nonlinear mapping from 2.5D surface components to 3D object properties – our interpretation for the function of AL and EIG-f_4 – the identity classification task that VGG and its variants are trained for might be linearly solvable from the high-level image features computed in these networks' top convolutional layer (TCL), with no need for further nonlinear transformations.

To test this hypothesis, we attempted to linearly decode identity on the FIV faces from each of the models: layers TCL, FFCL, and SFCL in VGG and layers f_3, f_4, f_5 in EIG. Specifically, we trained a one-layer linear-softmax classification network on ML/MF, and on the max-pooling outputs of EIG-f_3 and VGG-TCL, to decode all 25 FIV identities. We split the 175 FIV images to 6 poses ($6 \times 25 = 150$ images) for training and 1 pose (25 images) for testing with averaging results across all 7 possible splits. Fig. S4A shows the held-out test performance of the linear classifier. All layers in both networks gave rise to above chance (4%) decoding performance, but we found far better decodability of identity in the VGG network, with its TCL representation already achieving near FFCL and SFCL performance. In contrast, in the EIG network identity wasn't nearly as linearly decodable initially at its f_3 layer but increased to a comparable level of performance as the VGG network by layer f_5 (Fig. S4A). These results support our conjecture that the face identity might be linearly solvable from the TCL representations of the VGG network and its variants, without a need for further nonlinear transformations.

We also tested whether a nonlinear transformation on top of the 2.5D-like representations –e.g., the layer f_4 in the EIG network– are required for mapping these representations to 3D object properties. We attempted to linearly decode the shape and texture properties of the FIV images –approximated using the EIG network as its layer f_5 outputs given the 175 FIV images– based on both models, layers TCL, FFCL, and SFCL in VGG and layers f_3, f_4, and f_5 in EIG. We performed linear regression using the partial least squares (PLS)

method with 33 retained components (Helland, 2006; Pedregosa et al., 2011). We split the 175 FIV images to 6 poses ($6 \times 25 = 150$ images) for training and 1 pose (25 images) for testing with averaging results across all 7 possible splits. Fig. S4A shows the goodness-of-fit R^2 values on the held-out test sets. We found that these shape and texture vectors were not linearly decodable from any of the VGG layers (Fig. S4B), whereas it became increasingly more decodable in the EIG network from layer f_3 to f_4 (Fig. S4B). Notably, the intrinsic scene properties (i.e., the shape and texture properties) were much less linearly decodable at layer f_3 when compared to layer f_5 indicating that indeed the transformation from 2.5D-like representations to 3D scenes requires some nonlinear transformation.

4. Neural data analysis

The neural experiments and the data presented in the main text were originally reported in (Freiwald and Tsao, 2010).

4.1 Stimulus and experimental procedure

The neural experiments used the FIV image set. FIV included images of 25 person identities with each identity viewed at 7 different head orientations: left-profile, left-half-profile, straight, right-half-profile, right-profile, upwards, downwards. (The original recordings also used an 8th viewing condition,the back of the head, but we didn't analyze the corresponding data in this study).

Images were shown in a rapid serial presentation mode with 200 msec on-time followed by 200 msec blank screen with gray background. Images were presented centrally and subtended an angle of 7°. Monkeys were given a juice reward for maintaining fixation at the center of the screen for 3 seconds.

4.2 Neural recordings

Single-unit recordings were made from three male rhesus macaque monkeys (*Macaca mulatta*). Before the recordings, face-selective regions in each subject were localized using functional magnetic resonance imaging (fMRI). The face-selective regions were determined as the regions that were activated more to faces in comparison to bodies, objects, fruits, hands, and scrambled patterns. Single-unit recordings were performed at four of the fMRI-

identified face-selective patches, all in the inferior temporal cortex: middle lateral and middle fundus areas ML/MF, anterior lateral area, AL, and anteriormedial area, AM. Following the original study, we combined the responses from the regions ML and MF in our analysis due to their general similarity (referred to as ML/MF).

A single neuron was targeted at each recording session, in which each image was presented 1 to 10 times in a random order. Following (Freiwald and Tsao, 2010), we only analyze responses of the well isolated units.

4.3 Representational similarity matrices: Neurons

To compute the neural similarity matrices for a given neural site, each image was represented as a vector of the average spiking rates of all neurons recorded at that site. Following (Meyers et al., 2015), we obtained the average number of spikes for each neuron across the repetitions of a given image using the time-binned spike counts centered at 200 msec after stimulus onset with a time window of 50 msec in each direction. Following (Freiwald and Tsao, 2010), for each site, we min-max (range $[0, 1]$) normalized the average spiking rate of each neuron. For a given neural site, similarity of a pair of images was computed as the Pearson's correlation coefficient of the corresponding pair of the average spiking vectors. All spiking data was processed using the Neural Decoding Toolbox (Meyers, 2013).

4.4 Representational similarity matrices: Models

For a given image set, model, and the model's layer, images were represented as a vector of activations of all units in that layer. The model similarity of a pair of images (i.e., each entry in the similarity matrices in Figs. 2C-E and G-I, main text) is the Pearson's correlation coefficient of their corresponding activations vectors.

4.5 Linear regression analysis using the idealized similarity templates

For a given representational similarity matrix M, we solved the following linear equation.

$$M = c_1 * I_1 + c_2 * I_2 + c_3 * I_3 + c_4 * B \tag{1}$$

where $\{c_1, c_2, c_3, c_4\}$ are non-negative coefficients, I_1 is the idealized view-specificity matrix, I_2 is the idealized mirror-symmetry matrix, I_3 is the idealized view-invariant identity

coding matrix, and B is the background matrix. These matrices are shown in Fig. 2A-iv in the main text. All black entries have a value of 1, all gray entries have a value of 0.5 and all white entries have a value of 0. We solve this equation above using a non-negative least squares solver as implemented in the Python package `scipy`'s `nnls` method.

4.6 Bootstrap procedure

Due to the small number of subjects (N=3), we performed bootstrap analysis at the image-level. Following the procedure in (Nili et al., 2014), a bootstrap sample was obtained by sampling the 175 images in the FIV image set with replacement. Based on this sample, we computed the neural and the model similarity matrices. To avoid spurious positive correlations, we excluded all non-diagonal identity-pairs that could arise due to sampling-with-replacement. Based on the discussion in (Diedrichsen and Kriegeskorte, 2017) and following (Ejaz et al., 2015), we computed the Pearson correlation coefficient between pairs of representational similarity matrices. We repeated this procedure for 1000 bootstrap samples. Significance was measured using a direct bootstrap hypothesis testing procedure with a significance level of 0.05.

For the linear regression analysis with idealized similarity matrices, we again bootstrap sampled the 175 images with replacement and performed the linear regression using the resulting similarity matrix. We repeated this procedure for 1000 times.

5. Psychophysics methods

5.1 Experiment 1

A total of 48 participants were recruited over Amazon's crowdsourcing platform, Mechanical Turk (one additional participant was eliminated due to performing at or worse than the chance performance, 50%). The task took about 10 minutes to complete. Each participant was paid $1.50 ($9.00/hour). All participants provided their informed consent and were at the age of 18 or older according to their self-report.

The experimental procedure consisted of a simple "same"/"different" judgment task as the following. A study item was presented for 150 msecs, which was followed by a masking stimuli in the form of a scrambled image of a face for 500 msecs. Finally a test item appeared and stayed on until a response was entered (the participants were instructed to press "f" for

"same" and press "j" for "different"). They performed 10 practice trials before performing 96 experimental trials. Participants did not receive any feedback at all during the practice trial, which aimed to have participants get used to the experiment parameters (e.g., its interface). During the experimental trials, participants were shown their current average performance at every fifth trial.

The stimuli were 200×200 color images of faces photo-realistically rendered using the generative model. None of the stimuli across the experiments were used during training of the models. The viewing conditions for both the study and test items were drawn randomly from their respective prior distribution, $Pr(L, P)$. All participants saw the same image set (i.e., the viewing conditions were sampled once for all participants before the experiment began). There were 48 "same" trials and 48 "different" trials.

No study identity (i.e., a pair of shape and texture properties) was presented twice across trials. For the "different" trials, we chose the distractor face (the test item) by running a Metropolis-Hasting based search until 50 accepted steps. The search started from a random face but with matching lighting and pose parameters as that of the study item and increasingly moved closer to the study face w.r.t. likelihood $P(I|S, T, L, P)$ by generating proposals from the prior distribution over shape and texture properties, $Pr(S, T)$. This procedure aimed to ensure that the test facial identities in "different" trials were not arbitrarily different from the study item in obvious ways. Our data suggested that this procedure was effective: across the "different" trials average Pr("Same") was 0.35 with a standard deviation of 0.15, min value of 0.10 and max value of 0.71. All stimuli were rendered using Matlab's OpenGL-based rendering pipeline.

The average performance of participants was 66% with a standard deviation of 7%, a min value of 53%, and a max value of 78%. Two tailed t-tests revealed that the performance of the Experiment 1 participants were not statistically distinguishable from that of the Experiment 2 participants ($p = 0.23$) but both Experiment 1 and 2 participants performed more accurately than the Experiment 3 participants ($p < 0.001$ and $p < 0.05$). See the corresponding subsections below for further details of the accuracy distributions of Experiment 2 and 3 participants.

The average reaction time of participants was 1479 msecs, with a standard deviation of 626 msecs, a min value of 426 msecs, and a max value of 3754 msecs. Two-tailed t-tests

revealed that the reaction time distributions were not statistically distinguishable from each other across all pairs of the three experiments ($p > 0.2$ in all pair-wise comparisons). See the corresponding subsections for the reaction time statistics of Experiments 2 and 3.

To examine effects of learning in each experiment, we considered a moving-window performance of the participants as the trials progressed from the 1st trial to the 96th trial (Fig. S5). We used a window size of 10 trials and stride of 1. For each participant and moving-window index, we found that participant's average performance in the next 10 trials including the current trial. Even though the data suggested some learning in the early trials in Experiment 1 (e.g., significant difference between the 1st and 10th windows' performance, $p < 0.05$), there was no indication of learning in Experiments 2 and 3 (e.g., no significant difference between the 1st and 10th windows, $p > 0.35$ in both experiments).

5.2 Experiment 2

A total of 48 participants were recruited over Amazon's crowdsourcing platform, Mechanical Turk (seven additional participants were eliminated due to performing at or worse than the chance performance, 50% and two other participants were eliminated because their average reaction times were very short, 19 msecs and 30 msecs, much less than even the perception-to-action cycle of the expert video game players, 100 msecs). The task took about 10 minutes to complete. Each participant was paid \$1.50 (\$9.00/hour). All participants provided their informed consent and were at the age of 18 or older according to their self-report.

The stimuli and procedure were identical to Experiment 1 with the following exceptions. The test item was always presented frontal (i.e., frontal lighting and frontal pose) and without texture. This was achieved by assuming a uniform gray color for all vertices of the face mesh before rendering.

Participants' average accuracy was 64% with a standard deviation of 7%, a min value of 52%, and a max value of 80%. Their average reaction time was 1542 msecs, with a standard deviation of 564 msecs, a min value of 151 msecs, and a max value of 3716 msecs.

5.3 Experiment 3

A total of 44 participants were recruited over Amazon's crowdsourcing platform, Mechanical Turk (12 additional participants were eliminated due to performing at or worse than the

chance performance, 50%, and one other participant was eliminated because their average reaction time was very short, 16 msecs, much less than even the perception-to-action cycle of the expert video game players, 100 msecs). The task took about 10 minutes to complete. Each participant was paid $1.50 ($9.00/hour). All participants provided their informed consent and were at the age of 18 or older according to their self-report.

The stimuli and procedure were identical to Experiment 1 with the following exception. The test item was always presented frontal (i.e., frontal lighting and frontal pose), however, the texture was rendered on a flat surface in order to eliminate shape information from shading. In an attempt to further eliminate the shape information, we post-processed the resulting images by applying a fish-eye lens effect. All of the code used to generate stimuli as well as the actual images used in the experiments will be released at the time of publication.

Participants' average accuracy was 61% with a standard deviation of 6%, a min value of 51%, and a max value of 73%. Their average reaction time was 1403 msecs, with a standard deviation of 472 msecs, a min value of 508 msecs, and a max value of 2792 msecs.

5.4 Calculating Similarity(study,test)

For a given pair of study and test images, their predicted similarity by a model was computed as the similarity of their respective representations under the model. For the EIG network, this representation was its f_5 consisting of the shape and texture properties (a 400 dimensional vector), excluding the lighting and pose parameters. The model's similarity prediction was the Pearson's correlation coefficient of these two vectors.

For the other networks, the images were represented by their resulting SFCL activations. The model's prediction is the correlation coefficient of these two vectors. We found that no other layer in the VGG face network resulted in a better account of the human behavior than the layer we used. We also considered using other similarity metrics in addition to Pearson's correlation coefficient such as the cosine of the angle between two vectors and Euclidean distance. We found no significant difference in fits for any of the models.

5.5 Bootstrap analysis

In order to quantify the correlations between the models' predictions and the data, we sampled whole subject responses with replacement. We generated $10,000$ such bootstrap

samples. All p-values were estimated using direct bootstrap hypothesis testing (Efron and Tibshirani, 1994).

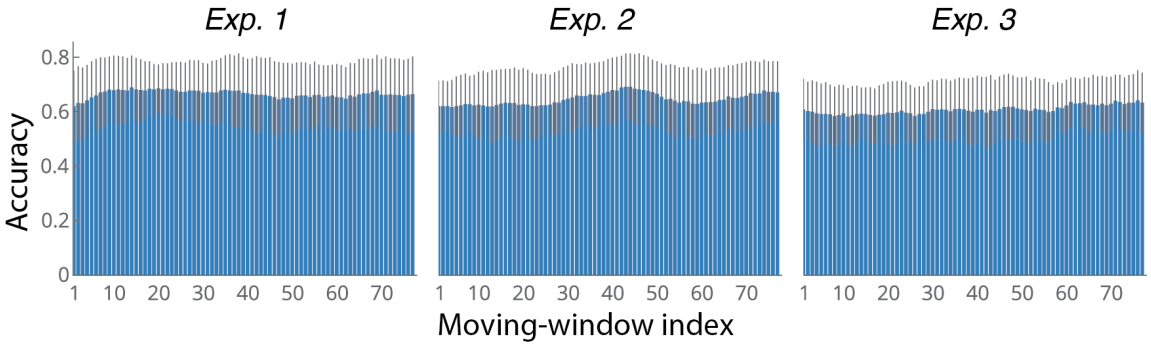

Figure S5: Learning curve analysis. The moving-window average performance of the participants in each experiment. We don't observe any pronounced effects of learning, especially in Exps. 2 and 3. Error bars indicate one standard deviation.

5.6 Hollow face illusion experiment

A total of 60 participants were recruited over Amazons crowdsourcing platform, Mechanical Turk. The task took about 10 minutes to complete. Each participant was paid $1.50 ($9.00/hour). Half of these participants participated in the light source elevation condition, and the other half participated in the depth-suppression condition. The experimental procedure was identical between the two groups.

Before the beginning of the experimental trials, both groups of participants were instructed that they would see images of faces that could be lit anywhere from the top of the face to the bottom of the face using an illustration of the range of possible scene lighting conditions (Fig. S6A). An example trial from the lighting source elevation condition is shown in Fig. S6B.

Both groups of participants had to complete 5 training trials before they moved onto 45 test trials. We only used the test trials in our analysis. Each of the 45 trials featured a different facial identity. In the depth-suppression group, each of the 9 levels of depth suppression (from 1, regular faces, to 0, flat face, to -1, fully inverted faces with nose pointing away from the observer; see also the main text) appeared 5 times throughout the experiment. In the lighting source elevation experiment, each of the 9 levels of elevation

appeared 5 times (from the top of the face, 1.31 radians of elevation, to the front of the face, 0 radians of elevation, to the bottom of the face, -1.31 radians of elevation; see also the main text).

For each condition, we z-scored each participant's responses (a total of 45 ratings each in the range of 1 to 7) before averaging all responses across participants and across the 9 levels. The error bars were obtained for each of the 9 levels as the standard deviation of the average values of the 5 stimuli items corresponding to that level.

Obtaining the EIG network's predictions was straightforward. For each condition, we ran the EIG model on the same set of 45 images as the human subjects, recording its outputs for the lighting elevation, L_e. We averaged the values for the 5 images of each of the 9 levels. The error bars in the main text (Fig. 5B, C) show the standard deviation across these five images. For each condition, the main text reports the linear correlation between the EIG model and the human behavior using their average predictions and responses across the 9 stimulus levels.

References

Kelsey R Allen, Ilker Yildirim, and Joshua B Tenenbaum. Integrating identification and perception: A case study of familiar and unfamiliar face processing. 2016.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, 13(4):e1005508, 2017.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC press, New York and London, 1994.

Naveed Ejaz, Masashi Hamada, and Jörn Diedrichsen. Hand use predicts the structure of representations in sensorimotor cortex. *Nature Neuroscience*, 18(7):1034–1040, 2015.

Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

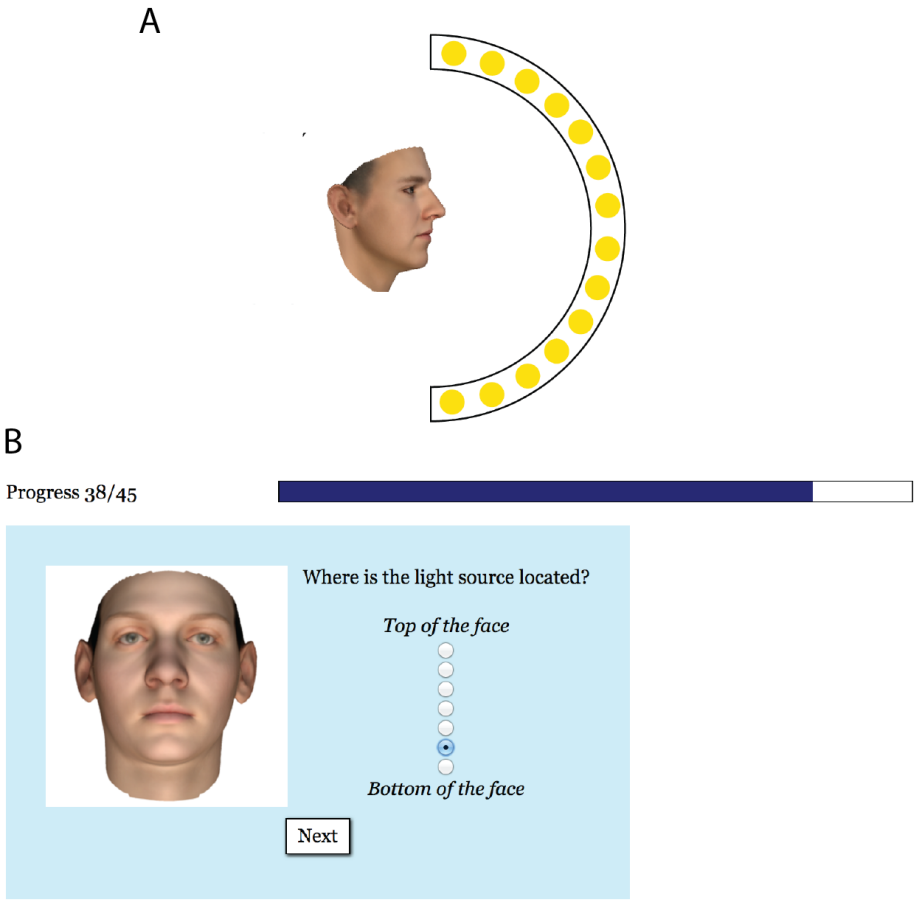

Figure S6: Lighting direction judgment experiment. (A) The lighting source could be located at one of the 9 locations frontal to the center of the face, as illustrated they covered the full range from above the face (1.31 rads) to below the face (-1.31 rads). (B) An example trial.

Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, 1988.

Inge Helland. Partial least squares regression. *Encyclopedia of statistical sciences*, 2006.

Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. 2017.

M Kevin. *Machine Learning: a Probabilistic Perspective*. The MIT press, 2012.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Tejas D Kulkarni, Ilker Yildirim, Pushmeet Kohli, Winrich A Freiwald, and Joshua B Tenenbaum. Deep Generative Vision as Approximate Bayesian Computation. In *Neural Information Processing Systems Workshop on Approximate Bayesian Computation*, 2014.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361, 1995.

Ethan Meyers. The neural decoding toolbox. *Frontiers in neuroinformatics*, 7:8, 2013.

Ethan M Meyers, Mia Borzello, Winrich A Freiwald, and Doris Tsao. Intelligent information loss: The coding of facial identity, head pose, and non-face information in the macaque face patch system. *The Journal of Neuroscience*, 35(18):7069–7081, 2015.

Iain Murray, Ryan Prescott Adams, and David JC MacKay. Elliptical slice sampling. *arXiv preprint arXiv:1001.0175*, 2009.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014.

O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *British Machine Vision Conference (BMVC)*, 2015.

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. Genova, Italy, 2009. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Ilker Yildirim, Tejas D Kulkarni, Winrich A Freiwald, and Joshua B Tenenbaum. Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual Conference of the Cognitive Science Society*, 2015.

Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.