NEUROSCIENCE

Efficient inverse graphics in biological face processing

Ilker Yildirim[1,2,3,4]*, Mario Belledonne[1,2,4], Winrich Freiwald[4,5]*, Josh Tenenbaum[1,4]*

Vision not only detects and recognizes objects, but performs rich inferences about the underlying scene structure that causes the patterns of light we see. Inverting generative models, or "analysis-by-synthesis", presents a possible solution, but its mechanistic implementations have typically been too slow for online perception, and their mapping to neural circuits remains unclear. Here we present a neurally plausible efficient inverse graphics model and test it in the domain of face recognition. The model is based on a deep neural network that learns to invert a three-dimensional face graphics program in a single fast feedforward pass. It explains human behavior qualitatively and quantitatively, including the classic "hollow face" illusion, and it maps directly onto a specialized face-processing circuit in the primate brain. The model fits both behavioral and neural data better than state-of-the-art computer vision models, and suggests an interpretable reverse-engineering account of how the brain transforms images into percepts.

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 also make inferences about the underlying causal structure of scenes (*1*–*3*). 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. These details can persist in long-term memory (*4*) and are crucial for planning our actions—sitting in that chair or climbing that tree. Similarly, when seeing a face, we can not only identify a person if they are familiar but also perceive so many details of shape, texture, and subtleties of expression even in people we have never met before.

To explain these inferences, early vision scientists proposed that scene analysis proceeds by inverting causal generative models, also known as "analysis-by-synthesis" or "inverse graphics." Approaches to inverse graphics have been considered for decades in computational vision (*3*, *5*–*8*), and these models have some behavioral support (*9*). However, inference in these models has traditionally been based on top-down stochastic search algorithms, such as Markov chain Monte Carlo (MCMC), which are highly iterative and implausibly slow. A single scene percept may take many iterations to compute via MCMC (which could be seconds or minutes on conventional hardware), in contrast to processing in the visual system, which is nearly instantaneous. While top-down processing likely plays a role in some of the brain's visual computations, such as surface segregation in complex scenes (*10*), both humans and nonhuman primates can extract rich high-level information about objects, faces, and scene gists in a time window (150 ms or less) that requires much (if not all) processing to be driven by a single feedforward pass (*11*, *12*).

In part for these reasons, modern work in computational vision and neuroscience has focused on a different class of architectures, deep convolutional neural networks (DCNNs), which are more consistent with the fast, mostly feedforward dynamics of vision in the brain and which benefit from simple, direct hypotheses about how their computations map onto neural circuits (*11*, *12*). DCNNs consist of many layers of features arranged in a feedforward hierarchy, typically trained discriminatively to optimize recognition of objects or object classes from labeled data. They have been instrumental both in leading engineering applications (*13*, *14*) and in predicting neural responses in the primate visual system, both at the level of single units in macaque cortex as well as functional magnetic resonance imaging (fMRI) in humans (*15*, *16*). Despite their impressive successes, however, conventional 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.

Most recently, a new class of computational architectures has been developed that can potentially answer this challenge, by combining the best features of DCNNs and analysis-by-synthesis approaches. Several artificial intelligence (AI) research groups, including ours, have shown how neural network "inference models" can be built from a feedforward or recurrent network architecture trained to infer the underlying scene structure, rather than to recognize objects or classify object categories as in conventional DCNNs. In contrast to early analysis-by-synthesis algorithms, inference is fast, following a single bottom-up pass from the image or a small number of bottom-up–top-down cycles, without the need for extensive iterative processing (*17*–*22*). These models have been developed in an engineering setting and are just beginning to be tested in machine vision problems; their correspondence with human perception or neural mechanisms is unexplored. Here, we introduce a specific model in this class, which we call the efficient inverse graphics (EIG) network, and evaluate it as an account of face perception, arguably the best studied domain of high-level vision. The EIG model not only makes a number of fine-grained, quantitatively testable predictions but also lets us evaluate the more general hypothesis that face perception in the brain is best understood in terms of an inference network that inverts a causal generative model (Fig. 1A), as opposed to the more conventional view in both AI and neuroscience that perception is best approached using neural networks optimized for classification, trained to recognize or distinguish object or face identities (*11*, *12*, *15*).

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA. [2]Department of Psychology, Yale University, New Haven, CT, USA. [3]Department of Statistics and Data Science, Yale University, New Haven, CT, USA. [4]The Center for Brains, Minds and Machines, MIT, Cambridge, MA, USA. [5]Laboratory of Neural Systems, The Rockefeller University, New York, NY, USA.
*Corresponding author. Email: ilker.yildirim@yale.edu (I.Y.); wfreiwald@rockefeller.edu (W.F.); jbt@mit.edu (J.T.)

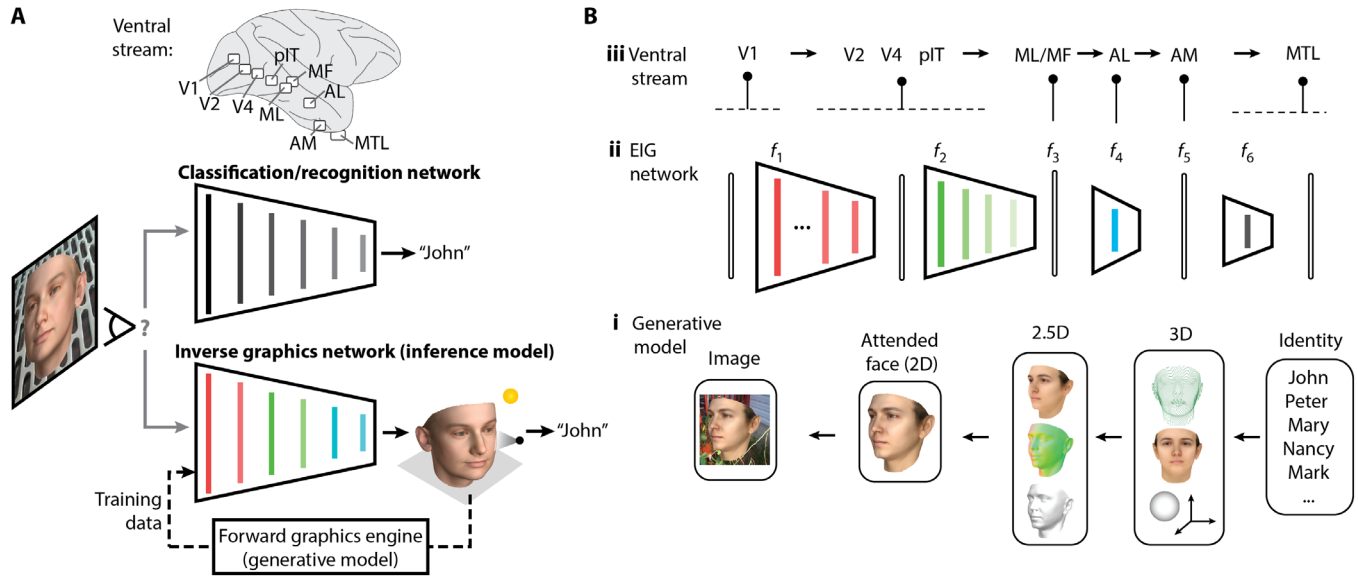

Fig. 1. Overview of the modeling framework. (**A**) Schematic illustration of two alternative hypotheses about the function of ventral stream processing: the recognition or classification hypothesis (top) and the inverse graphics or inference network hypothesis (bottom). (**B**) Schematic of the EIG model. Rounded rectangles indicate representations; arrows or trapezoids indicate causal transformations or inferential mappings between representations. (i) The probabilistic generative model (right to left) draws an identity from a distribution over familiar and unfamiliar individuals and then, through a series of graphics stages, generates 3D shape, texture, and viewing parameters, renders a 2D image via 2.5D image-based surface representations, and places the face image on an arbitrary background. (ii) The EIG inference network efficiently inverts this generative model using a cascade of DNNs, with intermediate steps corresponding to intermediate stages in the graphics pipeline, including face segmentation and normalization (f_1), inference of 3D scene properties via increasingly abstract image-based representations (convolution and pooling, f_2 to f_3), followed by two FCLs (f_4 to f_5), and finally a person identification network (f_6). (iii) Schematic of ventral-stream face perception in the macaque brain, from V1 up to inferotemporal cortex (IT), including three major IT face-selective sites (ML/MF, AL, and AM), and onto downstream medial temporal lobe (MTL) areas where person identity information is likely computed. Pins indicate empirically established or suggested functional explanations for different neural stages, based on the generative and inference models of EIG. Pins attached to horizontal dashed lines indicate untested but possible correspondences.

We find that the EIG model is uniquely compatible with known data on human and nonhuman primate face processing, and provides the first quantitatively accurate and functionally explanatory account of both neural population responses in macaques and a range of challenging perceptual judgments in humans.

The EIG model consists of two parts: a probabilistic generative model based on a multistage 3D graphics program for image synthesis (Fig. 1Bi) 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. 1Bii), layer by layer. The inverse model, also known as an inference network or inference model, is the heart of EIG and the component that we can most easily test with available neural and behavioral data. But the generative model is essential as well: It produces the training targets and training data for building the inference network, which is trained to infer the latent inputs or causes in the generative model conditioned on its outputs, rather than to predict class labels such as object categories or face identities as in conventional machine vision systems. The generative model, as we will see, also provides the basis for a functional interpretation of the representations the inference network learns. In this way, the EIG network embodies principles similar to earlier analysis-by-synthesis proposals that learn to approximate the underlying distribution of objects in an appropriate latent feature space (*8*, *23*) as well as to the Helmholtz machine originally proposed by Hinton and colleagues (*24*) in the 1990s and its modern cousins based on variational autoencoders (VAEs) (*17*, *18*). However, EIG differs from these approaches in that the generative model is based on an explicit graphics program (rather than a second deep neural network learned generically from data), and the EIG inference network is designed to parallel, in reverse, the graphics program's structure. This allows EIG to more faithfully capture the causal processes of how real-world scenes give rise to images and to exploit this structure for efficient learning and inference.

As a test case, we apply EIG 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. EIG implements the hypothesis that the downstream targets of the ventral visual pathway, a series of interconnected cortical areas in inferotemporal (IT) cortex, are 3D scene properties analogous to the latent variables in a causal generative model of image formation (referred to as the "latent variables" or inverse graphics hypothesis; Fig. 1A); moreover, EIG specifies a precise circuit mechanism by which these properties are plausibly computed in the ventral stream (Fig. 1Biii). 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, implementing versions of the 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 "classification" or "recognition" hypothesis; Fig. 1A). We also consider alternative instantiations of the latent variables hypothesis, based on VAEs, which replace the structured

generative graphics program in EIG with an unstructured generic deep neural network trained to reconstruct images. Only the EIG model, and therefore its more structured version of the latent variables hypothesis, accounts for the full set of neural and behavioral data, at the same time as it matches one of the most challenging perceptual functions of the ventral pathway: computing a rich, accurate percept of the intrinsic 3D shape and texture of a novel face from an observed image in a mostly feedforward pass.

RESULTS

EIG network

The core of EIG is the DCNN-based inference network, but we begin by describing the probabilistic generative model component, which determines the training objectives and produces the training data for the inference network. 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. 1Bi). 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 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 $\Pr(S, T \mid F)$, as well as extrinsic scene attributes controlling the lighting direction, L, and viewing direction (or equivalently, the head pose), P, from the distribution $\Pr(L, P)$. We implement this stage using the Basel Face Model (BFM; a probabilistic 3D morphable model) (*6*), 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, $I = \Phi(N, R, L)$. Last, the idealized 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. 1Bi; see Materials and Methods).

In principle, perception in this generative model can be formulated as MAP (maximum a posteriori) 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

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

where $\Pr(N, R \mid S, T, P)$, $\Pr(I \mid N, R, L)$, and $\Pr(O \mid I)$ 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 MCMC inference methods; all of these computations can be very slow. Instead, we consider a bottom-up feedforward inference model that is trained to directly estimate MAP values for the latent variables, F^*, S^*, T^*, L^*, P^*.

This inference network (Fig. 1Bii) 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 is to infer A from observations of B, then an optimal (maximally accurate and efficient) feedforward inference network can be constructed in two stages that map B to Z and Z to A, respectively (*25*, *26*). Here, our inference 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 an inference 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 $\Pr(I \mid O)$ using a DCNN module trained for face volume segmentation (*27*) and adapted to compute the face region given images of faces with background clutter (f_1 in Fig. 1Bii).

The second stage is the core of our EIG model and consists of a DCNN module trained to estimate intrinsic and extrinsic scene properties $\{S^*, T^*, L^*, P^*\}$ maximizing $\Pr(S, T, L, P \mid I^*)$ from the attended face image. This network is adapted from the architecture of a standard "AlexNet" DCNN (*13*) for object recognition, which consists of four convolutional layers (f_2 in Fig. 1Bii) ending in a fifth, top convolutional feature space (TCL; f_3 in Fig. 1Bii), followed by two fully connected layers (FCLs; f_4 and f_5, respectively). The training target for the second and final FCL f_5 is the key difference from the conventional object recognition or face recognition pipeline: Instead of being trained to predict class labels or identities, f_5 is trained to predict scene properties, $\{S, T, L, P\}$. Training begins from a pretrained 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 2 are generated by forward-simulating images drawn from the generative model [in the spirit of the Helmholtz machine (*24*)], 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.

Last, a third inference stage estimates the most likely face identity label F^*, given the scene properties, maximizing $\Pr(F \mid S^*, T^*, L^*, P^*)$. These identity labels are only introduced for familiar faces, with sufficient experience associating an individual's identity to that face. This module comprises a single new FCL f_6 for person identity classification and is trained on labeled image-identity pairs. We generate these pairs from real-world experience if available, or by simulating real-world experience drawing faces randomly from the generative model and its prior over individuals $P(F)$. In modeling particular experimental data, we tune training to the distribution of faces presented and introduce classification nodes for specific individuals if participants have sufficient opportunity to become familiar with them. See Materials and Methods for further details of each of the three stages of the EIG network.

Together, these three modules form a complete inference pipeline (approximately) inverting the generative model of face images, which satisfies the crucial characteristics of face perception and perceptual systems more generally: The inverse 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.

We have tested the EIG inference network on both synthetic and held-out real face images, both isolated and superimposed on random backgrounds, and compared its performance with classic top-down analysis-by-synthesis algorithms based on MCMC (*7*). EIG inferences are at least as accurate, assessed both quantitatively (Fig. 2A; see also Materials and Methods) and qualitatively (Fig. 2B), while being far faster (Fig. 2A). EIG also generalizes to real-world faces of different genders and complexions, at least for images that are reasonably close to the distribution of faces captured in the generative model, with neutral expressions and little or no occlusion (Fig. 2C). The model's reconstructions of real-world faces "in the wild" are not perfect but capture many identity-specific shape and texture details for each input image. Thus, we see EIG as a viable functional solution to the problem of face perception while recognizing that it also has limitations and can be improved in a number of ways (see Discussion as well as Materials and Methods for potential weaknesses and points of ongoing model development). In the remainder of the paper, we ask how well our current version of the model captures the mechanisms of face perception in the mind and brain, by comparing its internal representations (especially, f_3, f_4, and 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.

EIG stages explain the 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 (Fig. 3A; see Materials and Methods for experimental procedure and neural recording details) (*28*). Freiwald and Tsao (*28*) presented macaques with images of different individuals in different viewing poses (Fig. 3B) and found that this three-level hierarchy exhibits a systematic progression of tuning properties (Fig. 3Ci). Neurons in the bottom-level face patches ML/MF (middle lateral and middle fundus) have responses driven largely by the pose of a face, independent of the face's identity. Those in the midlevel patch AL (anterior lateral) also exhibit pose-specific tuning, but with a strong mirror symmetry effect: Faces in poses mirror reflected about the frontal view axes exhibit similar responses. Neurons in the top-level patch AM (anterior medial) exhibit view-robust identity coding. It has also been argued that these neural populations encode a multidimensional space for face, based on controlled sets of synthetically generated images (*29*, *30*). However, it remains unclear how the full range of 3D shapes and appearances for natural faces viewed under widely varying natural viewing 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 the 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 (Fig. 3Ciii) to fit the empirical similarity matrices for neural populations in each of these patches (see Materials and Methods). The coefficients of these different matrices (Fig. 3Cii) measure, in objective terms, how view specificity decreases from ML/MF to AM (yellow bars), how mirror symmetry peaks in AL (light blue bars), and how view-invariant identity coding increases from ML/MF to AL and further to AM (dark blue bars), complementing the qualitative features shown in the population-level similarity matrices (Fig. 3Ci).

We then evaluated the ability of the EIG network and other models to explain these qualitative and quantitative tuning properties of

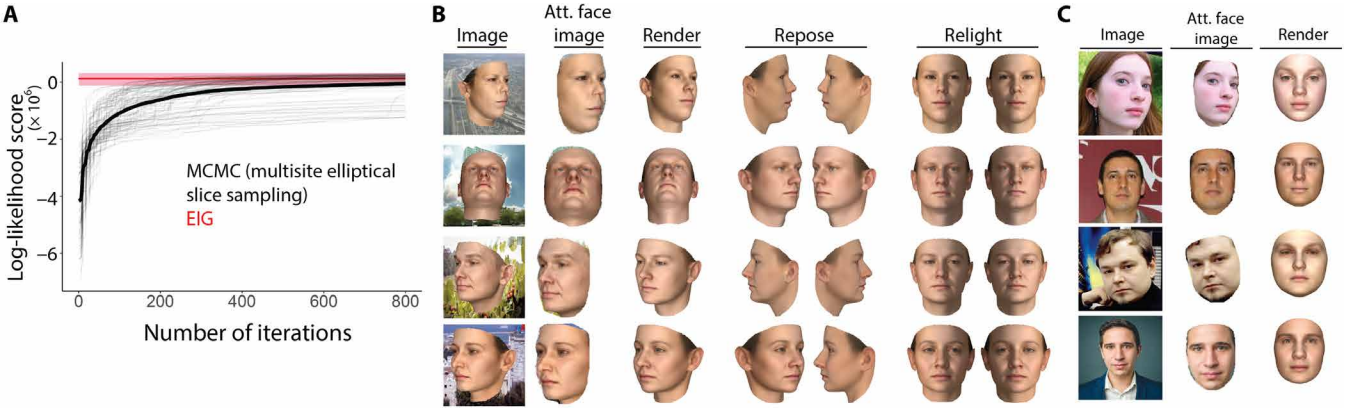

Fig. 2. Overview of the modeling framework. (**A**) Image-based log-likelihood scores for a random sample of observations using the EIG network's inferred scene parameters (layer f_5) compared to a conventional MCMC-based analysis-by-synthesis method. EIG estimates are computed with no iterations (red line; pink shows min-max interval), yet achieve a higher score and lower variance than MCMC, which requires hundreds of iterations to achieve a similar mean level of inference quality (thick line; thin lines show individual runs; see also Materials and Methods). (**B**) Example inference results from EIG, on held-out real face scans rendered against cluttered backgrounds. Inferred scene parameters are rendered, re-posed, and re-lit using the generative model. (**C**) Example inference results from the EIG network applied to real-world face images. Faces have been re-rendered in a frontal pose using the generative model applied to the latent scene parameters inferred by EIG. Although the EIG recognition network is trained only on samples from the generative model, it can still generalize reasonably well to real-world faces of different genders and complexions. Re-rendered results are not perfect, but they are recognizably more similar to the corresponding input face image than to other faces. All images are public domain and fetched from the following sources (from top to bottom): http://tinyurl.com/whtumjy, http://tinyurl.com/te5vzps, http://tinyurl.com/rcof3zj, and http://tinyurl.com/u8nxz7w.

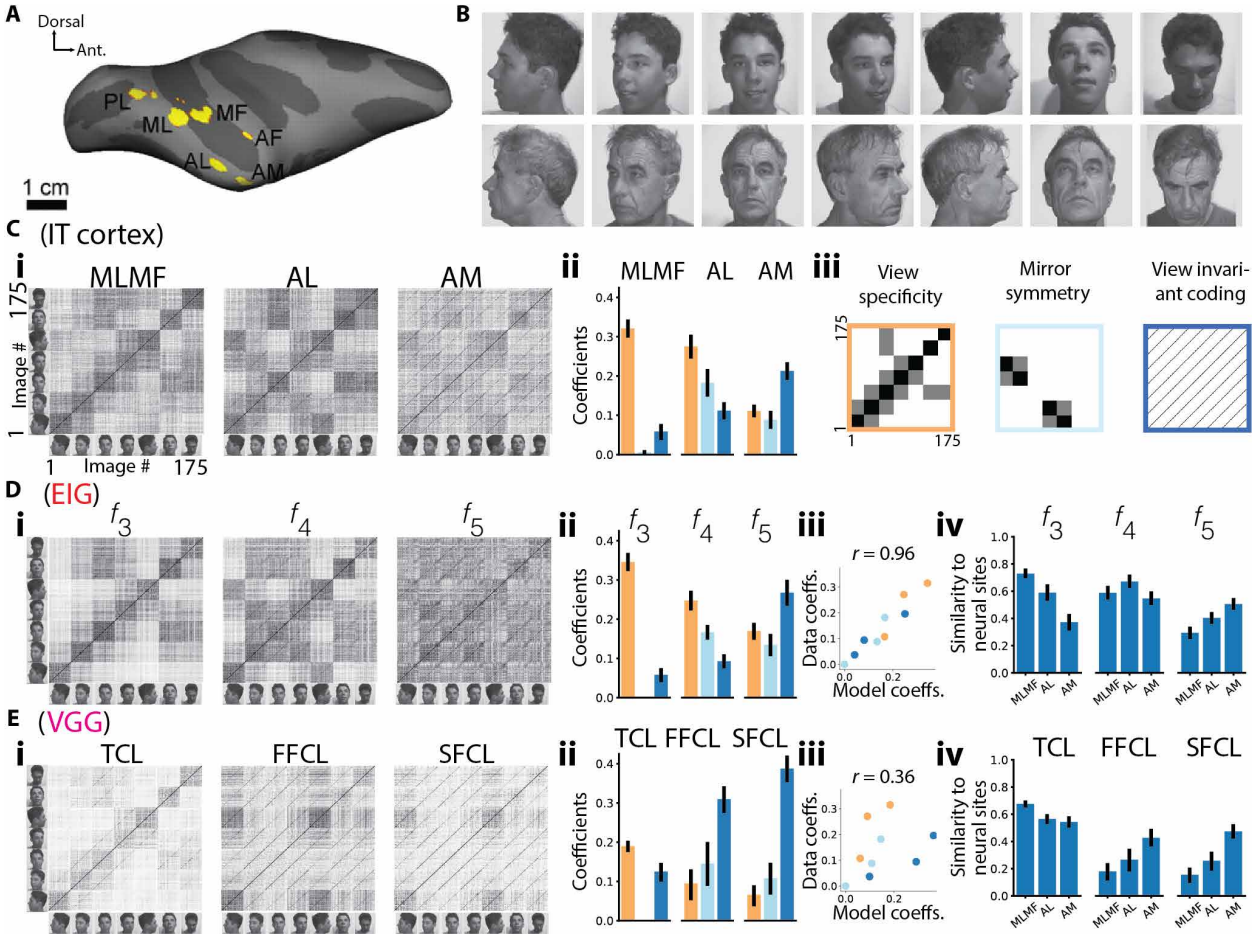

Fig. 3. Inverse graphics in the brain. (**A**) Inflated macaque right hemisphere showing six temporal pole face patches, including ML/MF, AL, and AM. (**B**) Sample FIV images consisting of 25 individuals each shown in seven poses, making a total of 175 images. These images were used in (*28*). Photo credit: Margaret Livingstone. (**C**) (i) Population-level similarity matrices for each face patch. Each matrix shows correlation coefficients of population-level responses for each image pair from the FIV image set (*28*). (ii) 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 (iii), in addition to a constant background factor to account for overall mean similarity. (**D**) (i) Similarity matrices for each key layer of the EIG network—f_3, f_4, and f_5—tested with FIV image set. Each image is represented as a vector of activations in the corresponding layer. (ii) Linear regression coefficients showing contribution of each idealized similarity matrix for each layer. (iii) Comparing full set of neural transformations to model transformations using these coefficients. (iv) Pearson's *r* between similarity matrices arising from each of the neural populations and model layers. (**E**) VGG network tested using FIV image set. Subpanels follow the same convention as the EIG results. Error bars show 95% bootstrap confidence intervals (CIs; see Materials and Methods).

ML/MF, AL, and AM. In particular, we contrast EIG with several variants of the VGG (Visual Geometry Group) network, based on 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 (see Materials and Methods) (*31*). These comparisons allow us to tell apart the inverse graphics hypothesis and the classification hypothesis at the level of neural representation.

We first test models using the face-identities-view (FIV) set of natural face images, with 175 images of 25 individuals in seven poses, shown to monkeys during neural recording of the face patches (Fig. 3B). The EIG network faithfully reproduces all patterns in the neural data, both qualitatively (Fig. 3Di) and quantitatively in terms of the idealized similarity matrix analysis (Fig. 3Dii). The coefficients of all three idealized similarity templates (view specificity, mirror symmetry, and view invariant coding) across all three levels of representation (f_3/ML/MF, f_4/AL, and f_5/AM) correlate almost perfectly between EIG and cortical face circuitry ($r = 0.96$; Fig. 3Diii). EIG also tracks the functional compartmentalization observed in the cortical hierarchy, as measured by raw correlations between similarities in corresponding layers: Similarity in layer f_3 best correlates with ML/MF, layer f_4 best correlates with AL, and layer f_5 best correlates with AM ($P < 0.05$; Fig. 3Div). By all these measures, EIG appears to capture the full progression of three functionally distinct stages in face processing, from ML/MF through AL up to AM.

We evaluate VGG based on its three layers architecturally analogous to EIG: VGG's TCL, analogous to f_3, which is the TCL of EIG; VGG's first FCL (FFCL), analogous to f_4, which is the FFCL of EIG; and VGG's second FCL (SFCL) analogous to f_5, the SFCL of EIG. These are also the layers of VGG most similar in response to the three levels of the face patch system. While VGG representations bear some similarity to analogous layers of representation in the neural data, VGG—in contrast to EIG—also showed profound qualitative

and quantitative differences from the brain in its patterns of selectivity (Fig. 3Ei). Representations in VGG were substantially more view invariant than either cortex or EIG across all three layers ($P < 0.05$, compare all yellow bars and dark blue bars in Fig. 3Eii versus Figs. 3Cii and 2Dii), with the biggest disparities occurring at the intermediate level (compare FFCL to f_4 and AL in Fig. 3, Eii, Dii, and Cii). There is no layer of VGG that shows the characteristic tuning of the intermediate patch AL, as f_4 of EIG does, nor does any layer of VGG correlate maximally with AL relative to other neural sites; each layer is either a better fit to ML/MF or AM (Fig. 3Eiv). Across all three levels in VGG, coefficients of the three idealized similarity matrices correlated much less strongly with analogous coefficients for neural data ($r = 0.36$; Fig. 3Eiii), suggesting a failure to capture how face processing progresses through the cortical hierarchy. Most dramatically, the two highest layers of VGG (FFCL and SFCL) were almost indistinguishable from each other (Fig. 3Ei), which fails to reflect the clear progression in function from mirror-symmetric tuning to view invariant coding that is seen in both the corresponding layers of EIG (f_3 and f_4) and the corresponding neural sites (AL and AM).

Other analyses show that VGG performance does not depend on whether it is fine-tuned to these specific face identities (as in Fig. 3E; see fig. S2 for VGG in its raw pretrained state) and that the initial face segmentation and normalization stage of EIG, which has not been a component of previous ventral stream models (*11*, *12*, *15*), is necessary for its strong performance (but has little effect on VGG; see section S1.1 and fig. S2). Together, these results strongly support the hypothesis that ventral stream face processing begins with an initial segmenting operation and culminates in targets that encode the latent variables of a face generative model, rather than mapping raw images to features optimized for face identity recognition or discrimination, as in conventional machine vision approaches.

To better understand the reasons why a fully brain-like pattern of responses arises in EIG, and the conditions under which it might arise in other neural network models, we studied a large number of model alternatives, varying in network architecture, training set and objective, and standard aspects of training procedure (see sections S1 and S2 and figs. S4 to S7). We used a controlled synthetic analog of the FIV image set, in which only faces were rendered (without clothing or backgrounds; FIV-S; see Materials and Methods). In particular, we tested several VAE variants that shared EIG's feedforward inference-network architecture but used a different training objective (image reconstruction loss; fig. S5) and used deep neural networks to parametrize a learned generative model (as opposed to EIG's structured graphics engine). We also tested several variants of the VGG architecture (fig. S4) to unconfound effects of the VGG architecture, training set, and training objective. We say that a model produces a "fully brain-like pattern of responses" to the extent that it has three progressive layers with idealized similarity coefficients matching those in ML/MF, AL, and AM (i.e., the bar plots shown in Fig. 3Cii), correlating highly across layers (as in Fig. 3Diii), and with raw similarities in each of these model layers, correlating maximally and distinctively with raw similarities in the corresponding neural sites (as in Fig. 3Div). Two aspects of the EIG network, its training targets and architecture, proved necessary to obtain fully brain-like representations: (i) The targets of inference should be the latent variables of the causal generative model (3D face shape and face texture descriptors), and (ii) there should be a stack of convolutional layers processing the attended face image, followed by at least one fully connected hidden layer between the TCL and the final layer trained to estimate the latent variables. Other aspects of the EIG training procedure, such as the magnitude of dropout and initialization with pretrained network weights, were not essential for producing fully brain-like responses but do make training much more efficient (figs. S6 and S7).

Last, 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 functional 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. Figure 1B (i and ii) suggests one possible functional 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 or surface depths) analogous to the "intrinsic images" or "2.5D sketch" of classic computer vision systems (*3*, *32*). 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 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. 4A; see section S3 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. 4B). Attended images also provide a better account of ML/MF than the raw pixel

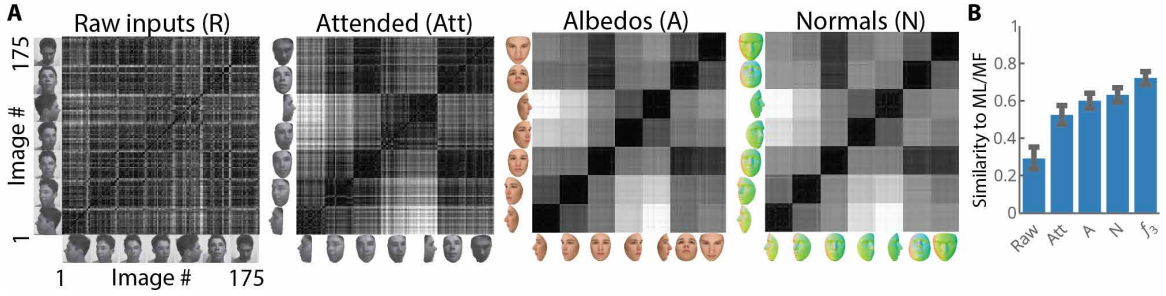

Fig. 4. Understanding ML/MF computations using the generative model and the 2.5D (or intrinsic image) components. (**A**) Similarity matrices based on raw input (R) images, attended images (Att), albedos (A), and normals (N). Colors indicate the direction of the normal of the underlying 3D surface at 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.

images ($P < 0.001$) but significantly worse than the 2.5D components ($P < 0.001$ for each component; Fig. 4B). 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 (see section S3 and fig. S8).

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). This highly nonlinear transformation can be facilitated using some kind of hidden layer and 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 predict identity rather than 3D object properties. These models always show very similar responses in all their FCLs (Fig. 3E and see also fig. S4). We conjecture that this AL-like intermediate stage nonlinearity is not necessary, because the FCLs 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 (see section S3.1 and fig. S9). The VGG network design may be a reasonable, perhaps even a superior, way to build a system for face perception if the goal is merely to classify or recognize individuals through their facial appearance, as in most of today's computer vision system. But the brain needs to compute much richer information about the 3D shape and texture of faces to analyze expressions, emotions, mood, and health or to use face perception as a cue in spoken language understanding. The inverse graphics design of the EIG network offers a possible route to those richer percepts, and our analyses suggest that the ML/MF-AL-AM circuit may be the locus of these computations in the brain.

EIG scene parameters predict human behavior

We also tested EIG and alternative models' ability to explain the behavioral aspects of face perception by comparing their responses to people's judgments in a suite of challenging unfamiliar face identity matching tasks (*33*). 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. 5A). 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., textureless face shapes rendered with a stone-like uniform gray albedo in frontal pose), eliminating all cues from skin coloration or texture normally present in face inputs. In experiment 3, the test items were flat frontal facial textures, produced by distorting normal images using a fish-eye lens effect to reduce shape information in the input (see Materials and Methods and sections S4.1, S4.2, and S4.3).

We hypothesized that if face perception is based on inverting a generative model with independent 3D shape and texture latents, as in EIG but not VGG, VGG-Raw, or other classification/recognition alternatives, then participants might be able to selectively attend to shape or texture estimates in their internal representations to optimize performance on these different challenge tasks. Crucially, EIG and VGG models are both trained using an equal number of images synthesized from the same graphics program used to generate the stimuli (although VGG is fine-tuned on top of the VGG network, which itself is trained with millions of other face images); only their training targets are different: latent variables of the generative face model for EIG versus an embedding space for discriminating person identities for VGG. This allows our behavioral analyses, like our neural analyses, to test between the two different hypotheses about the functional goal of face perception, inference in a generative model versus classification, or recognition of individuals' identities.

For each experiment, we compared average human responses—i.e., Pr("Same"), the proportion of participants responding "same" to a given trial—to the models' predicted similarity of the given pair of face images on that trial, across all 96 trials of the experiment. 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 (see Materials and Methods). The VGG and VGG-Raw networks' outputs for an image are their identity-embedding spaces, or SFCL. (No other layer in the VGG network provided a better account of the human behavior than its SFCL layer.) EIG's output is its shape and texture parameters, represented in f_5, which, unlike other models, supports selective attention to these different

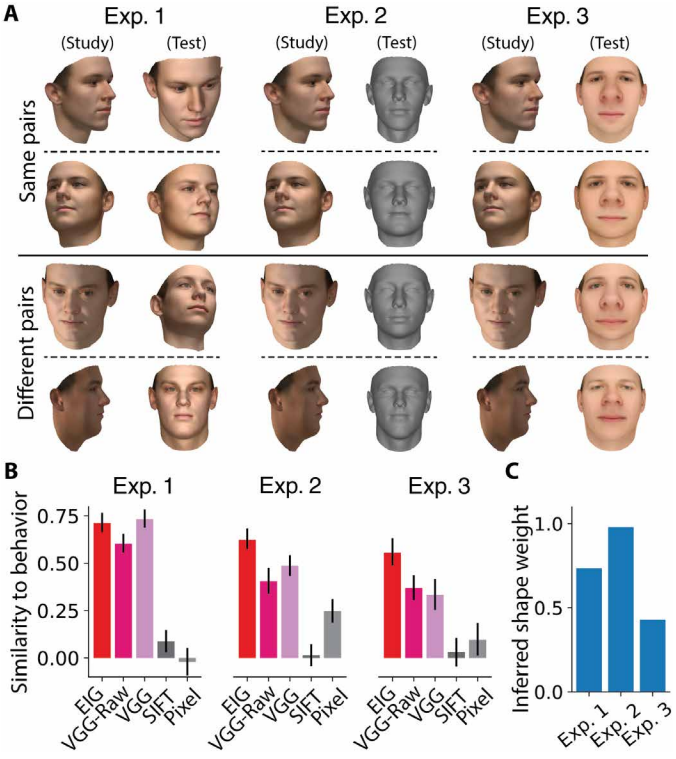

Fig. 5. Across three behavioral experiments, EIG consistently predicts human face identity matching performance. (**A**) Example stimuli testing same-different judgments (same trials, rows 1 and 2; different trials, rows 3 and 4) with normal test faces (experiment 1), "sculpture" (textureless) test faces (experiment 2), and fish-eye lens distorted shadeless facial textures as test faces (experiment 3). (**B**) Correlations between model similarity judgments and humans' probability of responding same. (**C**) Inferred weights (a value between 0 and 1 that maximized model's recognition accuracy) of the shape properties (relative to texture properties) in the EIG model predictions for experiments 1 to 3. Error bars indicate 95% bootstrap CIs (see Materials and Methods).

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 (see Materials and Methods).

Overall, participants performed significantly better than chance (50% correct): Average performance was 66% correct in experiment 1, 64% in experiment 2, and 61% in experiment 3 (section S4 for model-free behavioral analysis.) In trial-by-trial comparisons to behavior, we first evaluated simple image matching methods by either using pixels or more sophisticated scale or rotation invariant transforms of these pixels (scale-invariant feature transform or "SIFT" features) to determine similarity between study and test images (see Materials and Methods). We found that these methods correlated weakly or not at all with human trial-by-trial judgments (Fig. 5B), showing that simple image-matching strategies cannot explain behavior. EIG, on the other hand, consistently predicted human responses across all three experiments, with r values 0.71[0.66,0.76], 0.62[0.56,0.67], and 0.55[0.48,0.62] (where [l, u] indicates lower/upper 95% confidence intervals; Fig. 5B). VGG (though not VGG-Raw) performed comparably on experiment 1, but EIG fit human judgments significantly better than both alternative models in experiments 2 and 3 ($P < 0.001$ for all comparisons based on direct bootstrap hypothesis tests; see Materials and Methods). EIG's ability to selectively attend to shape and texture plays a critical role here. In experiment 1, EIG's inferred attention weight showed a baseline bias toward shape properties (shape weight =0.73), but this weight shifted in the predicted directions in both experiments 2 and 3 (Fig. 5C). In experiment 2, which completely eliminated texture cues, EIG's inferred attention weight focused almost exclusively on shape (shape weight =0.98). In experiment 3, which distorted shape while preserving texture, EIG's inferred attention weight focused slightly more on texture properties (shape weight =0.43), which represents an even larger shift toward texture relative to the baseline value in experiment 1. These results suggest that EIG captures human face perception abilities more accurately than other models, especially under less familiar stimulus conditions and tasks requiring extreme generalization between study and test faces. They also lend further support to the inverse graphics hypothesis over the classification hypothesis for ventral stream face processing.

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) appears to be a normally shaped face with two distinctions: (i) hollow faces lit from the top or side appear to be lit from the bottom or alternate side, and (ii) hollow faces appear flatter than normal faces (*34*, *35*). It has been suggested that this illusion could be a result of Bayesian inference, arising from the integration of top-down priors for natural face geometry, appearance, and lighting with ambiguous bottom-up cues to depth such as shading patterns (*34*, *35*). To our knowledge, this proposal has not previously been tested quantitatively or implemented in a working computational model. Here, we psychophysically study the hollow face effect in greater detail using graded levels of depth reversals and test EIG quantitatively as a computational account of human illusory percepts at a trial-level granularity.

We compared our model's inferences about lighting direction and face depth with people's judgments, in both graded versions of the hollow face illusion and normal lighting direction variation, as a control (Fig. 6, A and B). 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 (Fig. 6, C and D). We also found that the EIG network, like humans, perceived depth-inverted faces as more flat when compared to their control counterparts, with the lighting source elevation matched to its illusorily perceived location in the depth-inverted condition; the EIG network closely matched the magnitude of flattening in depth judgments as a function of the level of depth reversal for hollow faces, as well as a subtle effect of lighting elevation on judged depth in the control condition (Fig. 6, E and F). We also attempted to decode these same lighting and profile depth parameters from the VGG network, and found significantly worse fits to human judgments in all cases, but especially in depth judgments where VGG fits were barely better than chance (see section S4 and fig. S13). The fact that the EIG network captures the nonlinear interaction of depth and lighting percepts in the hollow face illusion does not uniquely support EIG as an account of the ventral face pathway; a "vanilla" network could be trained to estimate either lighting or profile depth from face images and might predict the same judgments. Rather, EIG's success here relative to VGG, without EIG having to be trained specially on these atypical images or ever being trained explicitly to estimate profile depth, provides further evidence that ventral stream face perception as modeled by EIG is implementing some form of fast approximate analysis-by-synthesis or inverse graphics computation, as opposed to being optimized for recognition of face identity.

DISCUSSION

Our results suggest that the primate ventral stream approaches at least the first feedforward pass in face perception—and perhaps object perception more generally—with an inverse graphics strategy implemented via an efficient hierarchical inference network: Observed images are mapped via a segmentation and normalization mechanism to a view-centered, image-like representation of surface shape and appearance in ML/MF, which is then mapped via a nonlinear transform through AL to a largely viewpoint-independent representation of 3D properties (3D shape and texture) in the most anterior stage of AM. More speculatively, the middle stage of representation (ML/MF) could correspond to something like the classic computer vision proposals for a 2.5D sketch (*32*) or intrinsic image (*3*) maps of intrinsic surface properties (surface normals and albedo), represented in a viewer-centric coordinate frame. 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 might be optimized for its function: computing a rich representation of behaviorally relevant causal properties underlying the appearance of a novel object or scene, as quickly and as accurately as possible.

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 (*30*), 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

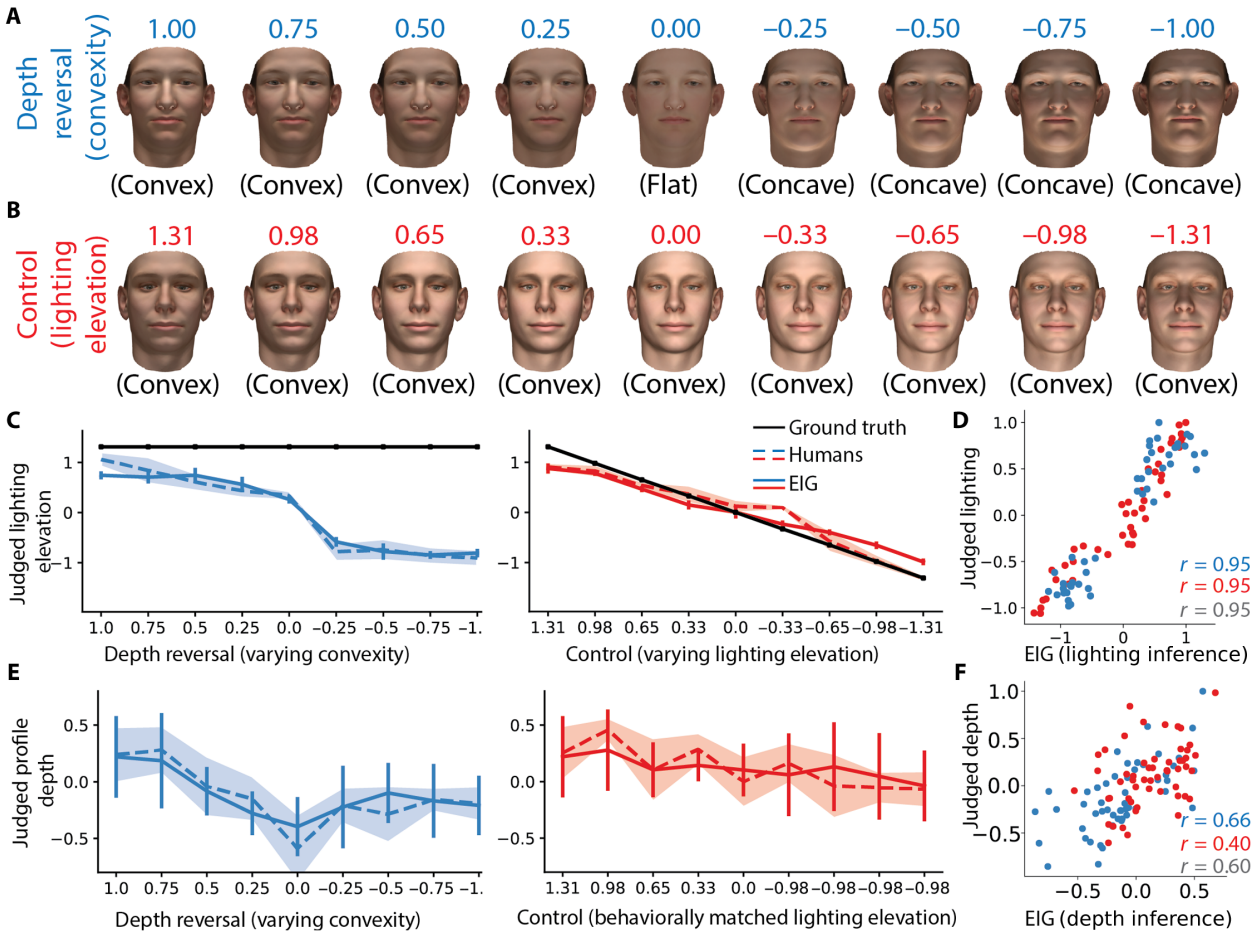

Fig. 6. Psychophysics of the "hollow face" effect. On a given trial, participants saw an image of a face lit by a single light source and judged either the elevation of the light source (C and D) or the profile depth of the presented face (E and F) using a scale between 1 and 7 (see also Materials and Methods and sections S4.4 and S4.5). (**A**) One group of participants (depth-suppression group) was 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 (control group) was presented with images of normally shaped faces (convexity = 1) lit from one of the nine possible elevations ranging from the top of the face to the bottom. (**C**) Normalized average light source elevation judgments of the depth-suppression group (left), the control group (right), EIG's lighting elevation inferences, and the ground truth light source location. (**D**) Average human judgments versus EIG's lighting source elevation inferences across all 90 trials without pooling to nine bins. Pearson's *r* values are shown for all trials (gray), control trials (red), and depth-suppression trials (blue). (**E**) Normalized average profile depth judgments of the depth-suppression group (left), control group (right), and EIG's inferred profile depth. (**F**) Average human judgments versus EIG's inferred profile depths across all 108 trials without pooling to nine bins. Pearson's *r* values are shown as in (D).

does? EIG is an image-computable model that faithfully reproduces representations in all three face patches of ML/MF, AL, and AM and explains 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, which exploits the conditional independence properties of a generative model for how face scenes produce images to efficiently invert that process. 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 the existence of feedback and other nonhierarchical connectivity that our current model does not capture (*36*). Following earlier models of primate face and object processing (*12*, *15*, *37*), we see a feedforward hierarchical network such as EIG as only a first approximation of the system's functional architecture—a natural starting point, as so much rich information about faces (and objects and scenes) is already computed in the first 150 ms of feedforward inference but clearly just a first step that future work should go beyond. More generally, there are important functions of vision that can be understood in terms of inverting generative models, such as segregating multiple objects or surfaces in complex or cluttered scenes, which appear to depend on feedback or recurrent connections, especially to early visual areas (V1/V2) (*10*). Explaining these neural computations could benefit greatly from the study of EIG architectures that integrate bottom-up and top-down processing (*20*, *38*). It is also possible that such feedback architectures could provide a fuller account of the mechanisms by which the computations in our EIG network are implemented in the brain.

The EIG network also likely deviates from biology in the mechanisms by which its weights are optimized to learn the functional

mapping from images to latent scene properties. We used stochastic gradient descent (SGD) for optimization, the standard learning procedure typically used in deep neural networks including all of the alternative models that we considered in this study and most state-of-the-art hierarchical models of cortical responses (*16*). Several aspects of SGD are incompatible with our current understanding of plasticity and development in neural circuits, but there is active ongoing work in exploring more biologically plausible (though functionally similar) learning mechanisms for deep neural networks, and it would be valuable to explore these learning mechanisms for efficiently inverting generative models in future work. It is also possible that the brain constructs something like an EIG network through a multiplicity of different learning mechanisms at the circuit level, e.g., including reinforcement learning or evolutionary strategies (*39*), and we hope to consider these possibilities in future work as well.

A potential limitation of both our behavioral and physiological studies is that all our stimuli used isolated images of faces, and we only tested our model applied to those images, while in the real world, people typically encounter faces in the context of much more complicated scenes. The architecture and training of our model explicitly tackles this complexity: The EIG inference network is trained from imagined images that overlay faces on complex backgrounds, and its first stage of processing in inference corresponds to segmenting and isolating a face from the rest of the scene (Fig. 1B and fig. S1), which can be done reliably even in complex scenes. As we show in Fig. 2 (B and C), the model generalizes to faces on complex cluttered backgrounds and, to a limited extent, also to real-world photos of faces, at least with neutral expressions and near frontal poses. Future psychophysical and physiological experiments could test human and nonhuman primate face perception with these more complex natural scenes, and we would expect the model to continue to predict the same basic phenomena that we study here. We and other groups are also working on extending the EIG approach to capture a wider range of face percepts in natural scenes. One mechanism that we are currently exploring is to grow the model's support through a bootstrap-like procedure, starting with the basic model presented here and iteratively improving both the generative model and the inference network by exposure to increasingly diverse face images. Such a bootstrap learning procedure might also be a more biologically plausible mechanism for how human face perception develops over the lifetime (*40*).

Comparing computational models to human judgments and using them to characterize internal representations has a rich history in face perception. Previous work by O'Toole and colleagues (*41*) compared the performance of machine vision systems to human performance, looking at both recent DCNNs (including the VGG face networks that we consider here) as well as many earlier face recognition systems. These comparisons focused on overall accuracy and relative difficulty of different naturally occurring faces, but did not attempt to test whether the internal representations used in different models correspond to those in the brain. They also did not consider alternative hypotheses about the function of face perception networks, realized in different network architectures, training regimes, or loss functions as we considered here, especially in the context of comparing inverse graphics versus classification hypotheses, which is the focus of our work. Also related to our findings are recent results from Zhan *et al.* (*42*), who used a generative model much like ours (but not an efficient inverse network) to make inferences about the representations underlying memory for familiar faces. They asked participants to judge the degree of similarity between random samples from the generative model and familiar faces, and identified via reverse-correlation methods certain aspects of 3D shape that were most diagnostic—and more diagnostic than 2D texture maps. This is similar to the greater weight on 3D shape over texture features we observed in face identity matching (experiment 1; Fig. 5C), suggesting at least some representational commonalities between online perception of unfamiliar faces and familiar faces stored in memory. More generally, our work shows how these 3D shape and texture map features can be computed from images efficiently, with network mechanisms that bear close resemblances to the ventral stream face patch circuitry, and how these representations can used to support flexible behavior, as in matching faces under conditions when normal texture or shape cues are unavailable or distorted (experiments 2 and 3; Fig. 5C).

That our model simultaneously explains the full macaque face patch system and the outputs of human psychophysical judgments provides further support that human and nonhuman primate face systems share at least broadly similar organization (*43*). Future work should characterize correspondences (and discrepancies) between our model and neural activity in three different face areas more closely. Recent work comparing VGG face network representations with neural representations in humans using intracranial electroencephalography (iEEG) data (*44*) does suggest a consistent picture with our results presented here. Grossman *et al.* (*44*) find evidence that VGG only matches human face representations up to the model network's TCL, in areas of human IT thought to best correspond to the middle face patches we study here (ML/MF). They take this as evidence that human face circuitry is performing a more "pictorial" form of processing than VGG's recognition computations, but they do not specify an alternative network architecture or concrete computational hypothesis for what that pictorial processing might be. Our model suggests one such hypothesis, in the form of 2.5D or intrinsic image components, which capture facial appearance and shape in a view-based, image-centric frame, and correspond well to middle face patch representations in macaques. Our model also suggests how those pictorial 2.5D representations can lead downstream to a full 3D description of face shape and appearance, which would correspond to more anterior face regions that [as noted by Grossman *et al.* (*44*)] have yet to be studied intracranially in humans and then further downstream to representations of familiar individuals' identities [e.g., medial temporal lobe (MTL) and perirhinal cortex], which have been characterized in both humans and macaques (*45*).

Our approach also has broader implications for neuroscience, perception, and cognition. The finding that IT supports decoding of category-orthogonal shape information for a wide range of objects, in addition to object category identity (*46*), suggests that an extension of EIG could account for how the brain perceives the 3D structure of objects beyond the domain of faces. With other collaborators, we have recently shown in an AI context that EIG-like networks for efficient inference of 3D shapes from 2D images via 2.5D sketches can work for arbitrary object classes (e.g., chairs and cars) (*21*) and can even generalize to a range of novel, unseen classes (*47*). In future work, we hope to explore these models of how the ventral visual pathway processes other object classes with functionally specific, localized representations (bodies, hands, and word forms), as well as objects more generally.

If this larger program is successful, it may offer a resolution to the problem of interpretability in visual neuroscience (*48*): 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.

The EIG approach can also be extended to richer perceptual inferences where there is currently no consensus on how these computations are implemented in the brain. EIG networks can be augmented with multiple scene layers to parse faces or other objects under occlusion (*49*). They can be deployed in parallel or in series (using attention) to parse out multiple objects in a scene (*17*). They can also be extended with anatomically aware generative models to imagine or process facial expressions (*50*). 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) (*49*). All of these extensions suggest testable hypothesis for neural computations and representations, in ways that could also point to crucial functional roles for feedback or recurrent processing, which our work here does not address.

Last, while our work suggests a functional role for causal generative models in the visual system, it leaves open many questions about their nature, use, and origins. Interpreted most literally, EIG implies that the brain uses feedforward inference networks as the workhorse of object perception, but uses generative models to provide the targets for training those networks, and as a source of internally generated training data (possibly at multiple stages, in a recognition pipeline that inverts a multistage generative process). Generative models in the brain could also support other functional roles; however, 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 (*7*, *21*). They could also support higher functions in cognition such as mental imagery, planning, and problem solving (*50*, *51*). It remains to be determined which of these functions are actually operative in the brain, as well as where and how generative models might be implemented in neural circuits, and how they might be built over development, from some combination of genetically programmed mechanisms and early perceptual experience. VAEs, and their close cousins GANs (*50*, *52*), capsules (*38*), and GQNs (*17*), as well as RCNs (*20*), are recent developments in artificial network architectures that suggest at least partial hypotheses for how graphics models might be implemented neurally or constructed through learning, 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.

MATERIALS AND METHODS

Generative model

Our generative model builds on and extends the BFM (*53*), a statistical shape and texture model obtained by applying probabilistic principal components analysis on a dataset of 200 laser-scanned human heads (100 female). 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 project 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 $S \in R^{D_S}$, $T \in R^{D_T}$. 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 $s = \{0,0, \ldots,0\}$ and $t = \{0,0, \ldots,0\}$. (Uppercase letters are used for random variables, and lowercase letters are used for assignments of these random variables to a sample from their respective distributions. Nonrandom 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 are 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) 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 (e.g., 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 equivalently, 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 $\{-1.4^{\text{rad}}\}$ to $\{1.4^{\text{rad}}\}$. 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 -0.75^{rad} to 0.75^{rad}. Last, we rendered each scene to a 227×227–pixel color image, unless otherwise mentioned, with backface culling.

Synthetic FIV image sets

The FIV-S stimuli underlying figs. S4 to S8 used the pose distributions in Table 1. Each of the 25 identities (i.e., unique pairs of shape and texture properties) were rendered at seven different poses and with frontal lighting.

The image set underlying fig. S8B (referred to as FIV-S-2), instead of using the pose distributions in Table 1, 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 each rendered seven times (each with its own randomly drawn pose and lighting parameters), making 175 images in total. In addition, to increase the variability at the level of raw and attended images, we converted half of these images to grayscale.

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

Table 1. Pose distributions for the FIV-S image set (in radians).

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	$1.50 + -1 * \text{abs}(N(0,0.05))$	$N(0,0.05)$
Left-half profile	$-0.75 + N(0,0.05)$	$N(0,0.05)$
Left profile	$-1.50 + \text{abs}(N(0,0.05))$	$N(0,0.05)$
Up	$N(0,0.05)$	$0.5 + N(0,0.05)$
Down	$N(0,0.05)$	$-0.5 + N(0,0.05)$

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's 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 δ function. (We dropped the corresponding Dirac δ functions in Eq. 1 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.

MCMC methods provide a general framework for inference in generative models and have a long history of application to inverse graphics problems (*2*). For this specific face model, we explored both traditional single-site MCMC and a more advanced and efficient multisite elliptical slice sampler (*54*) to infer the shape and texture properties given an image, I_D. 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 12 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 (*54*). This method often converges to reasonable inferences within a few hundred iterations, although with substantial variance across multiple runs of the algorithm as shown in Fig. 2A. The *y* axis values in that figure are the log-likelihood scores $P(I \mid S, T, L, P)$ of 100 individual chains each given as input a different face image (with clean background). The log-likelihood score for each iteration of each chain is calculated by rendering and comparing the current MCMC estimate with the input image. The log-likelihood scores for the EIG network on Fig. 2A are computed in the same way except that its estimates are outputs at its layer f_5.

The EIG estimates are computed almost instantaneously, with no iterations, yet achieve a higher score and lower variance (mean score, red line, $\sim 2.5 \times 10^5$; SD $\sim 1 \times 10^5$; pink region shows worst to best scores) than the MCMC algorithm. 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 $\sim -5 \times 10^5$; SD $\sim 8 \times 10^5$; thick black line shows the mean, and thinner black curves show 100 individual runs of the algorithm).

In summary, the EIG network that we describe below and in the main text reliably produces inferences that are as accurate as the best of these MCMC runs but far more quickly. EIG avoids the need for iterative computation by estimating 3D shape and texture properties via a single feedforward pass through a deep inference network. Further comparisons between MCMC and efficient inference 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 (*19*).

EIG model

The EIG model is a multistage neural network that attempts to estimate the MAP 3D scene properties and identity of an observed face image (approximately maximizing the posterior in Eq. 1). EIG comprises three inference modules arranged in sequence to take advantage of the conditional independence structure in the generative (graphics) model. These three modules compute (i) a segmentation and normalization of the face image, (ii) an estimate of the 3D face shape and texture, and (iii) 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 determines which tasks should be solved and the conditional independence structure of the generative model determines the order in which they should be solved.

Estimating face image given a transformed image, Pr(I | O)

Given an observation consisting of a face image with cluttered 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, referred to as f_1 in the main text. We took a recent convolutional neural network with an hourglass architecture that is trained for volumetric 3D segmentation of faces from images (*27*). 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) ×192 (height) ×200 (depth), which we found in practice often includes filled but disconnected regions that are outside the face-proper region.

We adapted this output for accurate 2D segmentation of the face-proper region 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 nonzero entries with 1) and compute its connected components. We produce a segmentation of *O* using the largest connected region of V_{xy} as the mask. Last, we normalize this region by zooming in on the segmented image using bicubic interpolation such that the resulting image's longer dimension is 227. In practice, this procedure yields good estimates for I^*. We also applied a small amount of translation (25 pixels) away from the left or right border for the normalized FIV images, which better aligned them with the samples from the generative model.

Scene parameters given face image, 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 accomplish this using an inference model by learning to map inputs to their underlying latent variables in the graphics program.

Our inference model is a convolutional neural network, with each layer implementing a cascade of functions including convolution, rectified linear activation, pooling, and normalization. We obtained this model by modifying AlexNet network architecture in the following way (*13*): We removed its top two FCLs and replaced them with a single new FCL. The details of the resulting network architecture are given in Table 2.

We initialized the parameters of f_2, f_3, and f_4 in the inference model using the corresponding weights of AlexNet that was pretrained on a large corpus of images, namely, the Places dataset (*55*). The pretrained network weights are provided by its authors and can be downloaded at http://places2.csail.mit.edu/models_places365/alexnet_places365.caffemodel. This dataset consists of about 2.5 million images and their corresponding place labels such as "beach," "classroom," and "landscape" (365-way categorization). The parameters of the new FCL (also referred to as scene properties layer or latents layer) were initialized randomly. Using these pretrained weights ensured that the earlier layers of the inference model provided a good generic visual feature extractor not specifically related to faces. We also avoided using a face corpus pretrained weights as this would require access to a large labeled dataset of weights, which EIG does not require.

To learn the mapping from images to their latent variable representations, we drew 200,000 random samples from the generative model. Each resulting image was a 227 × 227 color image, and each target was a concatenation of all the latent variables making a vector of length 404 (200 shape properties, 200 texture properties, and 4 extrinsic scene parameters). Half of the images were added background and were first segmented and normalized using f_1, whereas the other half of the images were not added background and were directly used during training. We fine-tuned the parameters of f_3, f_4 starting from their pretrained weights and trained the parameters of f_5 starting from random initialization. The network learns a mapping from these images to their latent variable representations, which we accomplish minimizing a mean squared error loss function using SGD with minibatches of 20 examples. In our simulations, we used a learning rate of 10^{-4}. To ensure that gradients were large enough throughout training, we multiplied the target latent variable vectors by 10. We accounted for this preprocessing step by dividing the outputs of the network by 10 at test time. We trained the model for 75 epochs.

Table 2. Inference model architecture.

Type	Patch size/stride	Output size
Convolution (f_{21})	11 × 11/4	96 × 55 × 55
Max pooling (f_{22})	3 × 3/2	96 × 27 × 27
Convolution (f_{23})	5 × 5/1	256 × 27 × 27
Max pooling (f_{24})	3 × 3/2	256 × 13 × 13
Convolution (f_{25})	3 × 3/1	384 × 13 × 13
Convolution (f_{26})	3 × 3/1	384 × 13 × 13
Convolution (f_3)	3 × 3/1	256 × 13 × 13
Max pooling	3 × 3/2	256 × 6 × 6
Full connectivity (f_4)		1 × 4096
Full connectivity (f_5)		1 × 404

Person identity given scene parameters, Pr(F | S,T,L,P)

We provide the details of Pr(F) before describing this final component of the inference model. In principle, this distribution is over a finite set of familiar individuals but allowing the possibility of encountering a new, unfamiliar individual. 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 inference model with a new FCL, f_6, of length K. To learn this mapping from scene properties to identities, we generated a new dataset 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.

For our FIV experiments, we do not have access to the ground truth shapes or textures of the 25 person identities, and therefore, we cannot use the graphics program for generating a training image set. Instead, for a given identity, we obtained M = 400 images by a bootstrapping procedure applied to the whole set of seven attended face 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 and 99%. Last, we translated the resulting bounding boxes in the image randomly but ensuring that the entire face-proper region remained in the image. We refer to the resulting image set as the 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 fine-tune the scene properties layer f_5 using $M * K$ = 10,000 images and their underlying person identity labels minimizing cross-entropy loss. We used a learning rate of 0.0005. We performed SGD with minibatches of 20 examples until the training performance was high (e.g., >95%). In practice, it took two additional epochs of training for the FIV-S image set and 20 additional epochs of training for the FIV image set.

A detailed diagram of our generative model, the EIG network, and a schematic of the ventral visual cortex hierarchy are shown in fig. S1 (complementing Fig. 1B). All of our models are implemented using the PyTorch machine learning library (*56*) and are available at https://github.com/CNCLgithub/EIG-faces.

Weaknesses of EIG

We note two potential weaknesses of the inference 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). We observed that this is an issue only if the face does not cover a spatially significant portion of the input image. Second, as we consider in

Discussion, 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 inference model, Pr(*F* | *S*, *T*, *L*, *P*), helps alleviate the second issue by allowing fine-tuning of f_5, thereby adjusting Pr(*S*, *T*, *L*, *P* | *I*) to the given training set (e.g., the bootstrapped FIV image set).

VGG network

The VGG network is based on the raw pretrained VGG face network (referred to as VGG-Raw) that is publicly available, http://www.robots.ox.ac.uk/~vgg/software/vgg_face/. This network consists of 13 convolutional layers (eight more layers than AlexNet) and three FCLs (same as AlexNet). The dataset used for training this network consisted of more than 2.5 million face images, where each image is labeled with one of 2622 person identities. The details of the network architecture, its training dataset, and training procedure can be found in (*31*).

Similar to the EIG network, the VGG network is obtained by fine-tuning this pretrained VGG-Raw network on the relevant image sets. For our FIV experiments, we used the same bootstrapped training dataset of FIV images as described above. We replaced VGG-Raw's top 2622-way fully connected classification layer [i.e., its third FCL (TFCL)] with a 25-way classification layer for the FIV identities. Training of VGG started from their pretrained values in VGG-Raw, except this final layer, which was initialized with random weights. We trained that new classification layer (TFCL) and fine-tuned the weights in TCL, FFCL, and SFCL using SGD to minimize a cross-entropy loss.

For our FIV-S experiments, we replaced the final classification layer in the pretrained VGG-Raw network with a 500-way classification layer. To train this network, we obtained a new dataset with the person identities and training images coming from the generative model. 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 dataset. This procedure gave us a total of 200,000 images and their corresponding identity labels (from 1 to 500). In line with the training of the VGG-Raw network, the VGG network as well as the EIG network used two standard data augmentation methods including making an image grayscale with a low probability (0.1) and mirror reflecting an image with probability 0.5. As for our FIV experiments, we initialized the weights of the VGG network using the weights of the pretrained VGG-Raw network except for its classification layer, which was initialized using random weights. We then fine-tuned the weights associated with its TCL, FFCL, and SFCL and trained its classification layer using SGD to minimize a cross-entropy loss. We used a learning rate of 0.0001 with minibatches of size 20 images.

Neural data analysis

The neural experiments and the neural data presented in the main text were originally reported in (*28*).

Stimulus and experimental procedure

The neural experiments used the FIV image set. FIV included images of 25 person identities, with each identity viewed at seven different head orientations: left-profile, left-half profile, straight, right-half profile, right-profile, upward, and downward. (The original recordings also used an eighth viewing condition, the back of the head, which we did not analyze in this study.)

Images were shown in a rapid serial presentation paradigm with 200 ms on-time followed by 200 ms off-time with a 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 s.

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 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: ML/MF, AL, and 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 (*28*), we only analyze responses of the well-isolated units.

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 (*57*), 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 ms after stimulus onset with a time window of 50 ms in each direction. Following (*28*), 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 were processed using the Neural Decoding Toolbox (*58*).

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 (e.g., each entry in the similarity matrix shown in Fig. 3Di) is the Pearson's correlation coefficient of their corresponding activations vectors.

Linear regression analysis using the idealized similarity templates

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

$$M = c_1 * I_1 + c_2 * I_2 + c_3 * I_3 + c_4 * B \quad (2)$$

where $\{c_1, c_2, c_3, c_4\}$ are 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. 3Ciii. 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 using a nonnegative least squares solver as implemented in the Python package scipy's nnls method.

Bootstrap procedure

Because of the small number of subjects ($N = 3$), we performed bootstrap analysis at the image level. Following the procedure in (*59*), a bootstrap sample was obtained by sampling the 175 images in the FIV image set with replacement. On the basis of this sample, we computed the neural and the model similarity matrices. To avoid spurious positive correlations, we excluded all nondiagonal identity pairs that could arise due to sampling with replacement. We computed the Pearson's correlation coefficient between pairs of representational similarity matrices [see the discussion in (*60*)]. We repeated this procedure for 10,000 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 each time. We repeated this procedure for 10,000 times. All *P* values were estimated using direct bootstrap hypothesis testing.

Psychophysics methods

Experiment 1

The experimental procedure consisted of a simple "same"/"different" judgment task as the following. A study item was presented for 150 ms, which was followed by a masking stimuli in the form of a scrambled image of a face for 500 ms. Last, 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). Participants performed 10 practice trials before performing 96 experimental trials. Participants did not receive any feedback 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 photorealistically 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 with respect to the likelihood $P(I \mid S, T, L, P)$ 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 an SD of 0.15, minimum value of 0.10, and maximum value of 0.71. All stimuli were rendered using Matlab's OpenGL-based rendering pipeline.

Experiment 2

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.

Experiment 3

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 to eliminate shape information from shading. In an attempt to further eliminate the shape information, we postprocessed the resulting images by applying a fish-eye lens effect.

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, we used its layer f_5 consisting of the shape and texture properties (a 200-dimensional vector for each component), excluding the lighting and pose parameters. The model's similarity prediction was a weighted sum of the Pearson's correlation coefficient between each pair of these two vectors—the correlation coefficient between the shape parameters and the correlation coefficient between the texture parameters. For each experiment, we estimated a fixed weight for the shape parameters, w_s, with a simple grid search using 50 linearly spaced values from 0 to 1. w_s was assigned to the value that maximized the model's performance on that experiment, and the weight of the texture parameters was given as $w_t = 1 - w_s$. (To calculate performance, similarities were transformed to binary same/different judgments using the median similarity across trials as the threshold.) We also considered fitting w_s as a free parameter to maximize the correlation between the behavioral responses and the model predictions per experiment and found that this variant was consistent with the performance-based matching method described above.

For the VGG 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 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.

To evaluate the pixel model, we flattened the pixel values for each of the study and test images and took the correlation between these two flattened vectors as their similarity. To evaluate the SIFT model, we first extracted their SIFT features, which are in the form of histograms, for each of the study and test images and determined their similarity using chi-square distance between the histograms of the study and test items. We also considered a number of other distance metrics appropriate for histograms including Bhattacharyya distance, intersection between two histograms, linear correlations, and Kullback-Leibler divergence but found that none of these alternatives improved over our results based on the chi-square distance.

Bootstrap procedure

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.

Lighting elevation judgment task

Both groups of participants—light source elevation group or the control condition and the depth-suppression group or the illusory condition—had to complete five 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 nine 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 five times throughout the experiment. In the lighting source elevation experiment, each of the nine levels of elevation appeared five 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 nine levels. The error bars were obtained for each of the nine levels as the SD of the average values of the five 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 five images of each of the nine levels. The error bars in Fig. 6C show the SD across these five images. The main text also reports trial-level correspondence between the model and the behavior as the correlation of model's predicted lighting elevations and the average human response per each of the 90 test trials (Fig. 6D).

Face depth judgment task

Before the beginning of the experimental trials, participants were instructed that they would see frontal images of faces and some faces would be flatter than others. They were shown several examples of fairly flat and fairly deep faces, which were samples chosen from either tail of the flatness distribution of 3000 randomly generated heads. On a given trial, participants were presented frontal image of a face (excluding neck and the ear) and were asked to judge the profile depth of it using a scale of 1 to 7. Next to the flat end (wide end) of the continuum, participants were presented with the profile view of an altered mean-BFM-face with its depth scaled to −3 (+3) SDs away from the mean depth of the abovementioned 3000 faces. An example trial in this experiment is shown in fig. S12.

Participants had to complete 10 training trials before they moved onto 108 test trials. We only used the test trials in our analysis. The 108 trials featured 54 different facial identities, with each identity rendered once as a regular face and once with depth suppression. These identities were uniformly assigned to the nine depth-suppression levels (six identities per level). When rendering an identity as a regular face, we set the lighting elevation location to match where it would be perceived given its depth-suppression level according to the results in Fig. 6C. The actual values used are indicated in the x axis of Fig. 6E (right panel). When rendering an identity with depth suppression, we always place the lighting elevation at the top, at 1.31 radians. Following previous work (*34*), we rendered only the face proper region excluding ears and neck. This procedure resulted in six images per each of the nine depth-suppression levels and six images per each of the nine control levels.

For each condition, we z-scored each participant's responses (a total of 108 ratings each in the range of 1 to 7) before averaging all responses across participants. The error bars were obtained for each of the nine levels as the SD of the average values of the five stimuli items corresponding to that level.

The EIG network can be readily used to estimate depth of a given face image. We ran the EIG model on the same set of 108 images as the human subjects, recording its outputs for the shape parameters, S. We then assigned a depth for each input image as the average displacement of the three key points (nose, left cheek, and right cheek) of the face shape with respect to the underlying aligned coordinate system of MFM. This coordinate system is in arbitrary units, so we z-scored model's predictions to bring it to the same scale as the behavioral data. The error bars in Fig. 6E show the SD across six images falling under the same pair of depth-suppressed or control and one of the nine levels. The main text also reports trial-level correspondence between the model and the behavior as the correlation of the model's predicted depth and the average human response per each of the 108 test trials (Fig. 6F).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/10/eaax5979/DC1

Section S1. Alternative architectures and loss functions
Section S2. Impact of pretrained weights and dropout rate
Section S3. Functionally interpreting ML/MF and f_3 using the generative model
Section S4. Psychophysics methods and model-free analysis
Fig. S1. A more detailed diagram of the modeling framework.
Fig. S2. Evaluation of VGG-Raw, VGG$^+$, and EIG$^-$ networks based on the FIV image set (extending Fig. 3).
Fig. S3. Scatter plots of data and model similarity matrices and analysis of earlier network layers (extending Fig. 3).
Fig. S4. Evaluation of alternative models using the FIV-S image set.
Fig. S5. Evaluation of the VAE models using the FIV-S image set.
Fig. S6. Trade-off arising from the choice of training targets and the use of pretrained weights.
Fig. S7. Variants of the EIG network architecture each trained from scratch without pretraining.
Fig. S8. Comparison of intermediate stages of the generative model to f_3.
Fig. S9. Decoding analysis.
Fig. S10. Learning curve analysis.
Fig. S11. Lighting direction judgment experiment.
Fig. S12. Snapshot of a trial from the depth judgment experiment.
Fig. S13. Decoding lighting elevation and profile depth from the VGG network.
Table S1. ID network architecture their architectures, loss functions, and training procedures.
Table S2. VAE decoder architecture.
Table S3. VAE-QN pose architecture dimensional vector.
References (*61*–*65*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. B. A. Olshausen, Perception as an inference problem, in *The Cognitive Neurosciences*, M. Gazzaniga, R. Mangun, Eds. (MIT Press, 2013).
2. A. Yuille, D. Kersten, Vision as Bayesian inference: Analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
3. H. Barrow, J. Tenenbaum, Recovering intrinsic scene characteristics from images, in *Computer Vision Systems* (Elsevier, 1978), p. 2.
4. T. F. Brady, T. Konkle, G. A. Alvarez, A. Oliva, Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14325–14329 (2008).
5. T. S. Lee, D. Mumford, Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* **20**, 1434–1448 (2003).
6. V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in *Annual Conference on Computer Graphics and Interactive Techniques* (ACM Press/Addison-Wesley Publishing Co., 1999), pp. 187–194.
7. T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, V. Mansinghka, Picture: A probabilistic programming language for scene perception, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), pp. 4390–4399.
8. A. M. Martinez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 748–763 (2002).
9. G. Erdogan, R. A. Jacobs, Visual shape perception as bayesian inference of 3d objectcentered shape representations. *Psychol. Rev.* **124**, 740–761 (2017).
10. V. A. Lamme, P. R. Roelfsema, The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
11. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

12. T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6424–6429 (2007).
13. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2012), pp. 1097–1105.
14. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), pp. 1–9.
15. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
16. S.-M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Comput. Biol.* **10**, e1003915 (2014).
17. S. M. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, D. P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. Rabinowitz, H. King, C. Hillier, M. Botvinick, D. Wierstra, K. Kavukcuoglu, D. Hassabis, Neural scene representation and rendering. *Science* **360**, 1204–1210 (2018).
18. T. D. Kulkarni, W. F. Whitney, P. kohli, J. Tenenbaum, Deep convolutional inverse graphics network, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2015), pp. 2539–2547.
19. I. Yildirim, T. D. Kulkarni, W. A. Freiwald, J. B. Tenenbaum, Efficient and robust analysisby-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations, in *Annual Conference of the Cognitive Science Society* (2015).
20. D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, A. Lavin, D. S. Phoenix, A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science* **358**, eaag2612 (2017).
21. J. Wu, Y. Wang, T. Xue, X. Sun, W. T. Freeman, J. B. Tenenbaum, MarrNet: 3D shape reconstruction via 2.5D sketches, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2017).
22. R. Zhao, Y. Wang, A. M. Martinez, A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 3059–3066 (2017).
23. T. Vetter, A. Hurlbert, T. Poggio, View-based models of 3D object recognition: Invariance to imaging transformations. *Cereb. Cortex* **5**, 261–269 (1995).
24. G. E. Hinton, P. Dayan, B. J. Frey, R. M. Neal, The "wake-sleep" algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).
25. A. Stuhlmuller, J. Taylor, N. Goodman, Learning stochastic inverses, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2013), pp. 3048–3056.
26. H. W. Lin, M. Tegmark, D. Rolnick, Why does deep and cheap learning work so well? *J. Stat. Phys.* **168**, 1223–1247 (2017).
27. A. S. Jackson, A. Bulat, V. Argyriou, G. Tzimiropoulos, Large pose 3D face reconstruction from a single image via direct volumetric CNN regression, in *Proceedings of the International Conference on Computer Vision* (IEEE, 2017).
28. W. A. Freiwald, D. Y. Tsao, Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851 (2010).
29. D. A. Leopold, I. V. Bondar, M. A. Giese, Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572–575 (2006).
30. L. Chang, D. Y. Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
31. O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in *Proceedings of the British Machine Vision Conference (BMVC)* (BMVA Press, 2015).
32. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, Cambridge, MA, 1982), vol. 2.
33. P. J. Hancock, V. Bruce, A. M. Burton, Recognition of unfamiliar faces. *Trends Cogn. Sci.* **4**, 330–337 (2000).
34. B. Hartung, P. R. Schrater, H. H. Bülthoff, D. Kersten, V. H. Franz, Is prior knowledge of object geometry used in visually guided reaching? *J. Vis.* **5**, 2 (2005).
35. R. L. Gregory, Knowledge in perception and illusion. *Philos. Trans. R. Soc. Lond. B* **352**, 1121–1127 (1997).
36. P. Grimaldi, K. S. Saleem, D. Tsao, Anatomical connections of the functionally defined "face patches" in the macaque monkey. *Neuron* **90**, 1325–1342 (2016).
37. J. Z. Leibo, Q. Liao, F. Anselmi, W. A. Freiwald, T. Poggio, View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation. *Curr. Biol.* **27**, 62–67 (2017).
38. S. Sabour, N. Frosst, G. E. Hinton, Dynamic routing between capsules, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2017).
39. Y. Ganin, T. Kulkarni, I. Babuschkin, S. Eslami, O. Vinyals, Synthesizing programs for images using reinforced adversarial learning. arXiv:1804.01118 [cs.CV] (3 April 2018).
40. L. T. Germine, B. Duchaine, K. Nakayama, Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition* **118**, 201–210 (2011).
41. C. J. Parde, C. Castillo, M. Q. Hill, Y. I. Colon, S. Sankaranarayanan, J.-C. Chen, A. J. O'Toole, Face and image representation in deep cnn features, in *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (IEEE, 2017), pp. 673–680.
42. J. Zhan, O. G. B. Garrod, N. van Rijsbergen, P. G. Schyns, Modelling face memory reveals task-generalizable representations. *Nat. Hum. Behav.* **3**, 817–826 (2019).
43. D. Y. Tsao, S. Moeller, W. A. Freiwald, Comparing face patch systems in macaques and humans. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 19514–19519 (2008).
44. S. Grossman, G. Gaziv, E. M. Yeagle, M. Harel, P. Mégevand, D. M. Groppe, S. Khuvis, J. L. Herrero, M. Irani, A. D. Mehta, R. Malach, Deep convolutional modeling of human face selective columns reveals their role in pictorial face representation. *bioRxiv* 444323 [**Preprint**]. 19 October 2018. https://doi.org/10.1101/444323.
45. S. M. Landi, W. A. Freiwald, Two areas for familiar face recognition in the primate brain. *Science* **357**, 591–595 (2017).
46. H. Hong, D. L. Yamins, N. J. Majaj, J. J. DiCarlo, Explicit information for categoryorthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
47. X. Zhang, Z. Zhang, C. Zhang, J. B. Tenenbaum,W. T. Freeman, J.Wu, Learning to Reconstruct Shapes from Unseen Classes, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2018).
48. D. L. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
49. I. Yildirim, M. Janner, M. Belledonne, C. Wallraven, W. Freiwald, J. B. Tenenbaum, Causal and compositional generative models in online perception, *Annual Conference of the Cognitive Science Society* (CBMM, 2017).
50. A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, F. Moreno-Noguer, GANimation: Anatomically-aware facial animation from a single image, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 818–833.
51. P. W. Battaglia, J. B. Hamrick, J. B. Tenenbaum, Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18327–18332 (2013).
52. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2014), pp. 2672–2680.
53. P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, *A 3D Face Model for Pose and Illumination Invariant Face Recognition* (IEEE, 2009).
54. I. Murray, R. P. Adams, D. J. C. MacKay, Elliptical slice sampling. arXiv:1001.0175 [stat.CO] (31 December 2009).
55. B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, A. Oliva, Places: An image database for deep scene understanding. arXiv:1610.02055 [cs.CV] (6 October 2016).
56. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. De Vito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, Eds. (Curran Associates, Inc., 2019), pp. 8024–8035.
57. E. M. Meyers, M. Borzello, W. A. Freiwald, D. Tsao, Intelligent information loss: The coding of facial identity, head pose, and non-face information in the macaque face patch system. *J. Neurosci.* **35**, 7069–7081 (2015).
58. E. M. Meyers, The neural decoding toolbox. *Front. Neuroinform.* **7**, 8 (2013).
59. H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, N. Kriegeskorte, A toolbox for representational similarity analysis. *PLOS Comput. Biol.* **10**, e1003553 (2014).
60. J. Diedrichsen, N. Kriegeskorte, Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Comput. Biol.* **13**, e1005508 (2017).
61. C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, Understanding disentangling in β-VAE. arXiv:1804.03599 [stat.ML] (10 April 2018).
62. D. P. Kingma, M. Welling, Auto-encoding variational bayes, in *Proceeding of the Advances in Neural Information Processing Systems* (NIPS, 2015).
63. A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, in *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2014), pp. 512–519.
64. I. Helland, *Partial Least Squares Regression* (John Wiley & Sons Inc., 2006).
65. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgments

Funding: 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.F.); the New York Stem Cell Foundation (to W.F.); ONR MURI N00014-13-1-0333 (to J.T.); a

grant from Toyota Research Institute (to J.T.); and a grant from Mitsubishi MELCO (to J.T.). W.F. is a New York Stem Cell Foundation–Robertson Investigator. A high-performance clustering environment for computations (OpenMind) was provided by the McGovern Institute for Brain Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. **Author contributions:** I.Y., W.F., and J.T. contributed to research goals and aims and to the planning of the project. I.Y. and J.T. designed the models. I.Y. and M.B. implemented the computer programs and performed experiments. W.F. and J.T. mentored the research and provided financial support. I.Y., M.B., W.F., and J.T. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 5 April 2019
Accepted 11 December 2019
Published 4 March 2020
10.1126/sciadv.aax5979

Citation: I. Yildirim, M. Belledonne, W. Freiwald, J. Tenenbaum, Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979 (2020).

ScienceAdvances

Efficient inverse graphics in biological face processing

Ilker Yildirim, Mario Belledonne, Winrich Freiwald and Josh Tenenbaum

Sci Adv **6** (10), eaax5979.
DOI: 10.1126/sciadv.aax5979