Top-down perceptual inference shaping the activity of early visual cortex

Ferenc Csikor*, Balázs Meszéna, and Gergő Orbán†

Department of Computational Sciences
HUN-REN Wigner Research Centre for Physics
Budapest, Hungary

Abstract

Deep discriminative models provide remarkable insights into hierarchical processing in the brain by predicting neural activity along the visual pathway. However, these models differ from biological systems in their computational and architectural properties. Unlike biological systems, they require teaching signals for supervised learning. Moreover, they rely on feed-forward processing of stimuli, which contrasts with the extensive top-down connections in the ventral pathway. Here, we address both issues by developing a hierarchical deep generative model and show that it predicts an extensive set of experimental results in the primary and secondary visual cortices (V1 and V2). We show that the widely documented nonlinear sensitivity of V2 neurons to texture statistics is a consequence of learning a hierarchical representation of natural images. Further, we show that top-down influences are inherent to inference in hierarchical generative models, and explain neuronal responses to illusory contours and systematic modulations of noise correlations in V1.

Hierarchical processing of visual information is a fundamental property of the visual cortex. This hierarchical processing machinery has been shaped by multiple adaptive pressures. First, it is the natural image statistics that shapes the sensitivities of individual neurons, i.e., the features these neurons are sensitive to. Second, the training objective determines how these features are combined to achieve the goal of the adaptation, i.e., how the activities of feature-sensitive neurons are contributing to computations. Recently, deep learning models have emerged as key tools to investigate the hierarchical processing of visual information [1]. These goal-oriented image models postulate that the relevant training objective of visual cortical computations is to perform specific tasks, such as classification of inputs into discrete categories, hence referred to as deep discriminative models. These models have the capacity to be trained on natural images, and have been immensely successful in predicting neuronal responses to these natural images, with depth of processing in the model corresponding to the depth of predicted neurons along the cortical hierarchy [2], as well as in designing synthetic images that achieve specific response patterns in populations of visual cortical neurons [3, 4]. While early deep learning approaches remained agnostic about the importance of the task being optimized, more recent studies have started to explore the differences between the predictions of alternative models optimized for different perceptual tasks [5, 6]. These studies highlight that the sensitivities of neurons in intermediate layers of the computational hierarchy is shaped by the task that the network is optimized for. Interestingly, more recently a number of studies started to explore training objectives that are more general than the execution of a specific task [5, 7, 8, 9].

*Corresponding author. E-mail: csikor.ferenc@wigner.hun-ren.hu
†Corresponding author. E-mail: orban.gergo@wigner.hun-ren.hu

Goal-oriented models achieve remarkable computational efficiency by learning invariances over similar images differing in nuisance parameters (e.g., differences in the postures of an animal) but to achieve this, models rely on teaching signals (so-called supervision). In light of these properties of learning in goal-directed models, it is instructive to review the desiderata of biological learning systems. Biological learning imposes two seemingly contradicting requirements on learning in the visual cortex. First, representations need to be learned without the strong requirements of supervision, i.e., training the visual cortex should be performed without a signal what its output should be. Second, in contrast with goal-directed models that are trained to perform inferences about the trained task variables only, e.g., deciding whether seeing a leopard or a tiger in the bushes, the visual system is required to fulfill flexible demands, such as deciding if we see an alert or relaxed animal based on the partial view of its torso through the shrubs. Importantly, learned invariances mean that high-level task variables do not necessarily carry the information required for other tasks and lower-level variables need to be queried for optimal inferences instead. However, optimization for only one set of task variables (distinguishing different species) does not provide guarantees for efficient computations relying on other variables (evaluating the posture). These requirements prompt turning to a goal-general framework for hierarchical processing in the visual cortex. Still, the burning question is how to reconcile the desiderata of richer inference with radically sparser teaching signals?

We propose that the ingredient that provides remedy to the dual challenge of learning hierarchical representations on natural images is provided by top-down interactions between the processing stages, ubiquitous in the visual cortex (Fig. 1) but lacking a well-established role in goal-directed deep learning systems. As we demonstrate in our paper, top-down signals have a central role in a class of goal-general learning systems, which has recently achieved considerable success in machine learning applications, deep generative models. Conceptually, top-down interactions establish expectations to support inference [10]: representing and computing with expectations is a key topic in theories of perception [11, 12, 13] and was shown to have both neuronal and behavioral correlates [10, 14, 15]. Top-down interactions in the hierarchy ensure that instead of a simple feed-forward pass of information to perform inference at the top of the processing hierarchy, inferences in intermediate layers of processing are supported by contextual information fed back from higher levels of the hierarchy. Formally, top-down interactions establish a contextual prior which is used to integrate with the feed-forward information channel to perform hierarchical inference.

In this study we investigate the neural representations of the features learned in the early visual system and the way these features are used to perform hierarchical inference. We seek to study a model of the visual system that is trained on natural images, as this is the key driving force behind the adaptation of the visual cortex. In order to explore the defining characteristics of this model, we study texture representations, which is motivated by three factors. First, textures are highly relevant features of natural images and critically contribute to higher-level visual processes such as segmentation and object recognition. Second, texture representations have been established in mid-level visual cortices, and in particular in V2, across multiple species, ranging from humans [16], through non-human primates [17, 18, 19], to rodents [20], suggesting that textures are not only relevant from a machine learning point of view but from a neuroscience point of view as well. Third, while some features that the V2 is sensitive to can be identified with linear computations, textures can only be learned through non-linear computations, a critical component of the successes of deep learning models.

Based on fundamental principles of hierarchical inference, we extend the feed-forward deep learning framework to perform goal-general hierarchical inference (Fig. 1). The adopted Variational Autoencoder (VAE) framework allows unsupervised learning of a nonlinear model of natural images. A non-hierarchical version of VAEs has earlier been applied in a restricted domain (learning about human faces) to address the activity profiles of neurons in the inferotemporal cortex [21]. We develop the hierarchical TDVAE (Top-Down VAE) model, which we

train end-to-end on natural image patches. First we investigate the properties of the learned representations and compare them to the responses of neurons in the early visual cortex of macaques. Next, we investigate the learned properties of top-down connections through experiments that investigate contextual influences of V2 on V1 responses. Through the development of a Variational Autoencoder model, our work provides a normative interpretation of the extensive top-down network present in the cortical hierarchy and demonstrates signatures of top-down interactions both in the mean and correlations of neural activities.

Results

To investigate the contribution of top-down connections to computations in a normative framework, we introduce a goal-general hierarchical model of natural images. We introduce our goal-general framework by contrasting it with alternative frameworks. The objective of hierarchical computations can be phrased in two different ways. First, the objective can be to learn to perform inference about the presence of high-level visual features. The most common form of this objective is a goal-directed objective, in which high-level features are object categories occurring in natural images (Fig. 1b). Here, $p(\text{object} \,|\, \text{image})$ is calculated such that object category information is provided during the training to learn a flexible and non-linear mapping from images to object categories [1, 2, 22]. A related objective in which high-level features are inferred is a self-supervised paradigm in which the goal is to identify features through learning invariances of individual images [7, 9]. In both of these models, intermediate sets of features are learned, with increasingly more complex features identified in the processing pipeline, which serve the inference of the features at the top level (in the case of goal-driven models, object categories). Here, neurons in the ventral stream of the visual cortex are identified with the different stages of a feed-forward pipeline (Fig. 1c). Second, a more general objective can be established by requiring the learning machinery to infer not only the presence of objects but also the presence of lower-level features constituting objects (Fig. 1d). This is phrased as learning a joint distribution over objects and features of different complexities: $p(\text{features}_\text{L}, \text{features}_\text{M}, \text{features}_\text{H}, \text{object} \,|\, \text{image})$. This is a more fine-grained form of inference that enables performing not only one particular task, such as viewpoint invariant recognition of objects, but also learning to infer the different aspects of different observations of the same object. Given that a goal-general learning is taking place, no category labels are provided at any level of the computational hierarchy. To establish a connection between the model and the activity of cortical neurons, we consider feature activations as the activation of model neurons. We identify the activity of model neurons to that of neurons in the ventral stream of the visual cortex, with features at different levels of the model hierarchy corresponding to neurons in different cortical regions (Fig. 1e).

To gain insight into the computations necessary for goal-general hierarchical inference we focus on inference on two levels of features: $p(\text{features}_\text{L}, \text{features}_\text{M} \,|\, \text{image})$. Assuming that low-level features are represented at V1, while mid-level features are represented at V2, we can obtain predictions for V1 activity by calculating

$$p(\text{V1} \,|\, \text{image}) = \int d\text{V2}\; p(\text{V1}, \text{V2} \,|\, \text{image}). \tag{1}$$

Here, $p(\text{V1} \,|\, \text{image})$ assigns probabilities to different response intensities of the neurons of the V1 population upon observing a stimulus such that high-probability responses predict the recorded activities. Probability theory helps to simplify the calculation of response distributions in individual areas. It indicates two possible partitioning of the joint distribution $p(\text{V1}, \text{V2} \,|\, \text{image})$: one that is solely relying on feed-forward progression of information (see also Methods): $p(\text{V1} \,|\, \text{image}) \cdot p(\text{V2} \,|\, \text{V1})$ or one that relies on top-down interactions between processing stages: $p(\text{V1} \,|\, \text{V2}, \text{image}) \cdot p(\text{V2} \,|\, \text{image})$. Note that in the latter formulation the activity in V1 not only depends on the stimulus, but also on the activity in V2, a signature of top-down

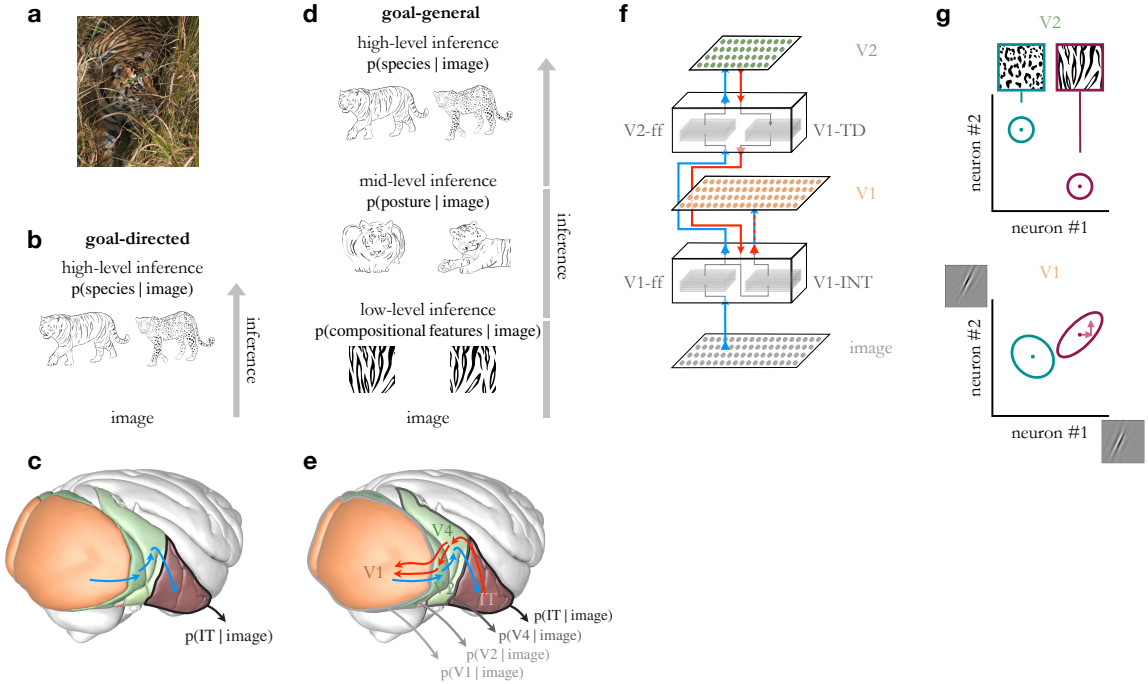


Fig. 1. Hierarchical inference in goal-general models. **a**, Example natural image. © istockphoto/Getty images **b**, Inference in a goal-directed model. The objective of goal-directed models is to identify objects. **c**, In goal-directed models the visual hierarchy serves the inference by progressively transforming the stimuli through predominantly feed-forward computations such that categories can be efficiently read out (i.e., using linear classification). **d**, Inference in a hierarchical goal-general model. The objective of hierarchical goal-general models is to infer a hierarchy of features that can efficiently account for variations observed in stimuli. Features are not directly identified during the training of the model. Inference can be performed on both high-level features such as animal species, and intermediate features, such as the posture of the animal, or the texture of their fur. **e**, In hierarchical goal-general models the visual computations along the hierarchy of the visual cortex correspond to inferences performed for features at different levels of the model hierarchy. Theoretical considerations suggest an interaction between feed-forward (blue arrows) and top-down (red arrows) information flow [13]. **f**, Neuronal circuitry of a Variational Autoencoder for V1 (brown) and V2 (green) of the ventral stream, as required by TDVAE. Features are represented in a layer of neurons (planes with brown and green disks that correspond to individual neurons). Layers of neurons (gray planes) transform feed-forward (V1-ff, V2-ff, blue arrows) and top-down information (V1-TD, red arrows) to ensure precise inference of the activation of individual features. Note the blue/red dashed line, which corresponds to a component for the integration of feed-forward and top-down information (V1-INT). Inference of V1 relies on high-level inferences performed at V2, which can be understood as relying on contextual priors established by V2. Information reaches V2 through layers of neurons potentially residing also in V1 that precede the layer accommodating the features represented in V1. **g**, Feature-selective neurons at higher levels in the cortical hierarchy establish priors for lower-level features, thus constraining their activities and introducing biases in response intensities as well as patterns in response correlations.

influences. These formulations are mathematically equivalent but their efficiency differs once the intractable inference is performed through approximations [23]. This, together with the top-down interactions characteristic of the anatomy of the visual cortex [24] motivates us to focus on hierarchical inference that recruits top-down interactions besides feed-forward processing of stimuli (Fig. 1e).

In order to train a network to perform hierarchical inference, we turn to the Variational Autoencoder formalism [25, 26]. Variational Autoencoders were developed to learn nonlinear goal-general image models and learn a pair of models: the recognition model supports inference, i.e., it describes the probability that a particular feature combination underlies an observed image; and the generative model, which summarizes the 'mechanics' of the environment, i.e., how the combination of features produces observations (see Methods for details). Here we focus on the characteristics of the recognition model: we set out to characterize cortical neural activities through the activities of model neurons in the recognition model. Relying on advances in hierarchical versions of VAEs [27, 28, 29], we developed Top-Down VAE, or TDVAE for short (Fig. 1f). When constructing the TDVAE model, we parameterized a VAE relying on two principles. First, we capitalized on neuroscience intuitions about V1 and V2. Second, the architecture is determined by the principles of probability theory, but the rest of TDVAE was left as general as possible such that its properties were determined solely by natural image statistics. Along this line, in the generative model we assumed an energy-efficient representation of the natural environment [30] that encouraged sparseness in population responses (Methods). Further, we assumed a linear relationship between the features that V1 neurons are sensitive to and the pixels of natural images [30, 31, 32]. Next, we relied on anatomical insights when implementing inference: the top-down formulation $p(\mathrm{V1} \,|\, \mathrm{V2}, \mathrm{image}) \cdot p(\mathrm{V2} \,|\, \mathrm{image})$ distinguishes between two components: $p(\mathrm{V1} \,|\, \mathrm{V2}, \mathrm{image})$, and $p(\mathrm{V2} \,|\, \mathrm{image})$. As these two components are implemented by neural networks, we took into account the architectural constraints of the visual system that information does not bypass V1 before reaching V2. Therefore the two components shared the entry layers of the network initially processing the stimulus information, which then forked into a feed-forward neural network converging onto V2, and a feedback neural network that integrated the information coming from the stimulus with the top-down component (Fig. 1f, see also Methods). At the level of V2 the generative model allowed flexible nonlinear computations such that more complex computations and more complex feature sets can be learned. The circuitry of the recognition model was determined by the principles of probability theory and the goal-general training objective, termed Expectation Lower Bound, ELBO (Fig. 1f, see Methods for details).

Since the visual cortex was adapted to the natural environment [33, 34, 35, 36], the TDVAE model was trained end-to-end on natural image patches (Methods), such that the learned representations (i.e., the represented sets of features) are determined by the natural image statistics. The primary determinants of the learned model are the properties of the generative model while the recognition model was set to be sufficiently complex to accommodate the computations required for inference. To demonstrate the invariance of our results on the specific choice of the generative and recognition models, we trained an array of models with different parametrizations (see Methods). The goodness of these alternative models was established through calculating the evidence lower bound (ELBO; Fig. 2a, Extended Data Table 3, Methods) and the one having the highest ELBO was used in the main text of the study, while alternatives are presented as Extended Data. Apart from these architectural choices, TDVAE has no free parameters.

Activity of variables at the level of V1 and V2 in TDVAE are assumed to correspond to activations of V1 and V2 neurons up to a linear combination of the variables, standard in earlier approaches [2]. In particular, the mean of $p(\mathrm{V1} \,|\, \mathrm{image})$ and $p(\mathrm{V2} \,|\, \mathrm{image})$, referred to as posteriors, corresponds to the mean response of the V1 and V2 neurons. Width of the posteriors can be identified with response variance, while correlations in the posterior can be identified with noise correlations. In this paper response variances are not investigated in detail but we refer the reader to earlier work [37, 38], and later sections discuss response correlations. Activity of model neurons are not constrained to be positive and therefore are identified with membrane potentials, similar to earlier approaches [37]. Without loss of generality, the presented results can be translated to firing rates by transforming membrane potentials with a threshold linear transformation [37, 39].

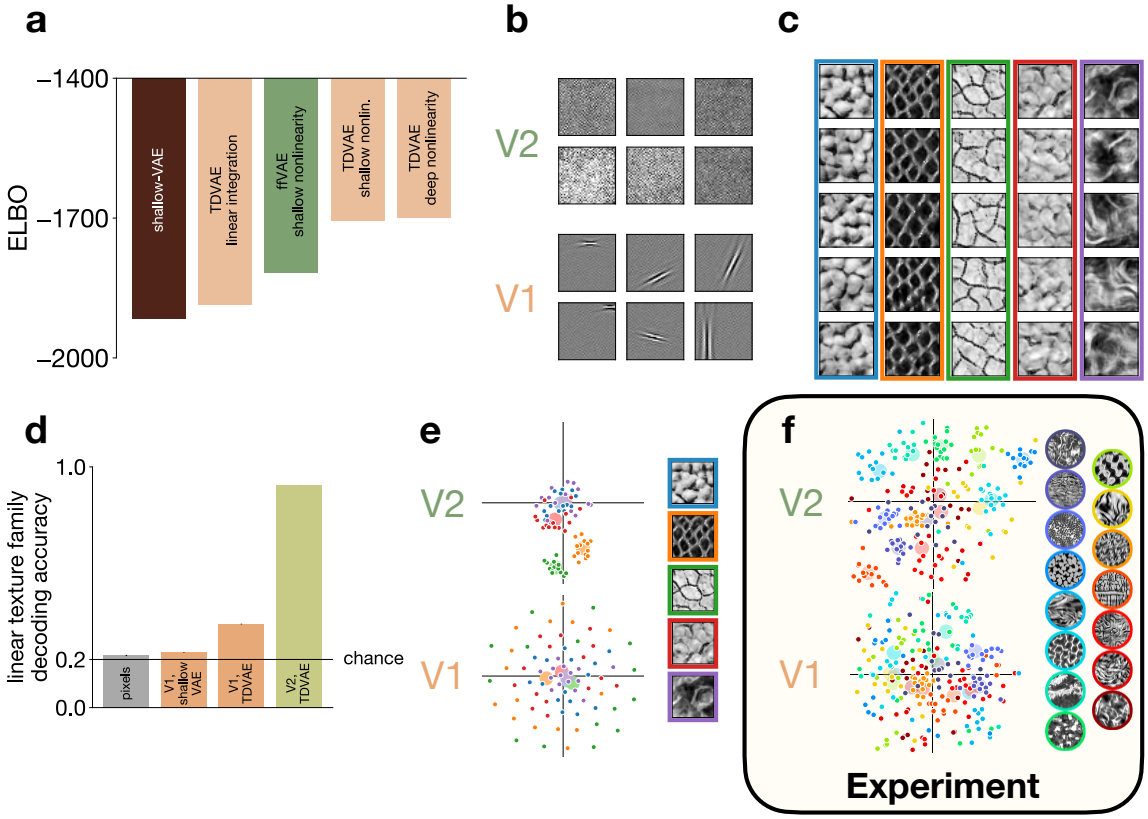


Fig. 2. Hierarchical representation in V1 and V2 learned by TDVAE. a, Value of the training objectives of alternative models. *Shallow-VAE*: non-hierarchical VAE constrained to V1; *ffVAE*: hierarchical VAE relying on feed-forward processing; *TDVAE*: three variants of hierarchical VAEs that feature top-down computational architecture. TDVAEs differ in the details of their computational architecture. *Linear integration*: feed-forward and top-down pathways are linearly combined to reach V1; *shallow nonlinearity*: the nonlinearity assumed in the generative model between V1 and V2 is implemented with a single processing layer; *deep nonlinearity*: a double processing layer in the generative model between V1 and V2, permitting the learning of more complex nonlinearities. **b**, Example receptive fields of V1 and V2 neurons, as characterized by white noise analysis. White noise analysis reveals linear response properties. As V2 features nonlinear computing capabilities, no linear characteristics are revealed but a wider coverage of the visual field is evident. **c**, Example texture patches from the five texture families (colors) used for probing the representation learned by model neurons. **d**, Decoding accuracies of texture families from different layers of the processing hierarchy. Bar heights: means for $n = 5$ random seeds; error bars: s.d. Solid black line indicates chance. *pixels*: decoder is directly applied to the pixels of texture patches; *V1, shallow-VAE*: decoding from V1 neurons of a non-hierarchical VAE; *V1, TDVAE* and *V2, TDVAE*: V1 and V2 neurons of the TDVAE model, respectively. **e**, Low dimensional visualization (t-SNE) of the mean responses of model V1 and V2 neurons to a number of randomly sampled example images (dots) from the five texture families (colors). Disks: means of the two-dimensional transformed mean responses per texture family. **f**, Same as **e** in V1 and V2 recordings from macaques for 15 texture families. Reproduced with permission from [18](PNAS).

Hierarchical representation of natural images. First, we investigate the representations that emerge in TDVAE. At the level of V1, we found that TDVAE robustly learns a complete dictionary of Gabor-like filters (Fig. 2b, Extended Data Fig. 5, Extended Data Table 3). The receptive fields of individual model neurons are localized, oriented, and bandpass (Extended Data

Fig. 6), qualitatively matching those obtained by training a VAE lacking a V2 layer (Extended Data Fig. 7). This result confirms earlier studies using single layer linear generative models featuring sparsity constraint on neural activations [30, 31, 32]. Thus, learning a hierarchical representation complete with a V2 layer on top of V1 left the qualitative features of the V1 representation intact in the TDVAE model.

The representation at the V2 layer is characterized by larger receptive fields than the compact ones found at the V1 layer (Fig. 2b). This is in qualitative agreement with electrophysiological data [16]. The learned representation is nonlinear therefore standard techniques suitable for establishing linear response properties [40] are not suitable to characterize the sensitivities of model V2 neurons. Inspired by extensive evidence supporting that V2 neurons are sensitive to texture-like structures [16, 20, 41], we used synthetic texture patches to explore the properties of the V2 representation that TDVAE learned on natural images [42].

We selected five natural textures characterized by dominant wavelengths compatible with the image patch size used in the study (Fig. 2c). We matched low-level texture family statistics (Methods), thus making texture families linearly indistinguishable in the pixel space (Fig. 2d). Since model V1 neurons learn a linear mapping from the neural space to the pixel space, we expect that texture information cannot be read out with a linear decoder from V1 response intensities either. Decoders that queried the texture family that a particular image was sampled from but ignored the identity of the image showed near chance performance on model V1 neurons (Fig. 2d), similar to a model that completely lacked V2 (Fig. 2d).

Nonlinear components in the generative model endow the V2 to build a representation that is capable of distinguishing texture families such that texture information is linearly decodable from the responses. Texture was reliably decodable from the mean responses of V2 neurons in TDVAE (Fig. 2d). This texture representation was very robust as it was qualitatively similar for all investigated hierarchical architectures (Extended Data Table 4). Responses of the V2 population displayed clustering by the texture family (Fig. 2e), similar to the representations identified in V2 neuron populations in macaque (Fig. 2f). In contrast, no such clustering could be identified in V1 (Fig. 2e), reminiscent of V1 neuron populations of the same macaque recordings (Fig. 2f). The texture representation discovered by TDVAE was low dimensional, irrespective of the architectural details of the model (Extended Data Table 4). This finding is surprising but recent experiments have confirmed that even though textures are represented by large neuron populations in the V2 of macaques, the effective dimensionality of this representation is very limited [20]. Intriguingly, the TDVAE architecture best fitting natural images only featured texture-selective dimensions, yielding a very compact representation exclusively focussing on textures (Extended Data Tables 3, 4).

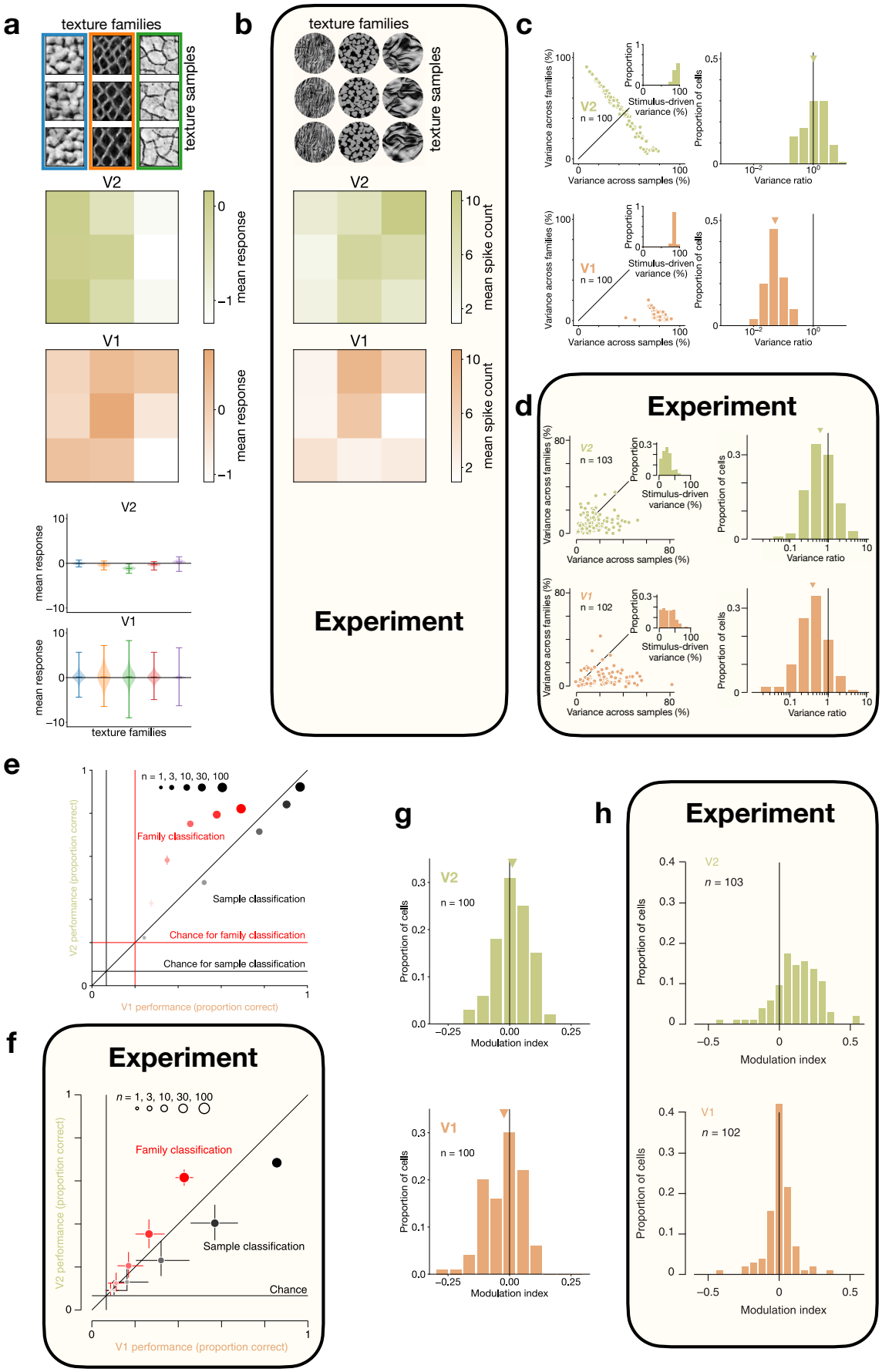
Further details of the representation of V1 and V2 can be revealed by investigating the learned invariances. Mean responses of individual V1 neurons displayed a high level of variability both across instances of texture images belonging to the same texture family and those belonging to different texture families (Fig. 3a). This was similar to the sensitivities displayed by V1 neurons in the macaque visual cortex (Fig. 3b). In contrast, model V2 neurons displayed a higher level of invariance across texture images belonging to the same texture family (Fig. 3a), again reflecting the properties of macaque V2 neurons (Fig. 3b). To quantify this observation, across-population statistics was calculated by dissecting across-sample and across-family variances in mean responses both in V1 and V2 neurons of TDVAE using a nested ANOVA method [18]. Note, that in order to make our analysis more general by permitting an arbitrary linear mapping between model V2 neurons and biological neurons, we used a random rotation of the V2 neurons to perform the analysis (Methods). The relative magnitude of response variance across families versus across samples was significantly higher in V2 than in V1 (Fig. 3c), corroborating earlier experimental findings (Fig. 3d). Higher invariance of responses to different stimuli belonging to the same texture family in V2 than in V1 can also be captured by a linear decoding analysis (Fig. 3e). Efficiency of discrimination between images in the same texture family is weaker

in V2 than that of V1 (Fig. 3e). Such increasing invariance to lower-level features along the visual hierarchy is a hallmark of gradual compression and can be identified at the population level in the primate visual cortex as well (Fig. 3f). In contrast with within-family stimulus identity decoding, decoding of texture family is more efficient from V2 than from V1 in TDVAE (Fig. 3e), a finding confirmed by macaque recordings (Fig. 3f).

Texture families can be defined through a set of pairwise statistics over the co-activations of linear filters represented in V1 [42]. To demonstrate that the learned V2 representation reflects these quintessential features of a proper texture representation, the TDVAE model can be tested with stimuli in which these high-level statistics are selectively manipulated while keeping the low-level statistics intact. To achieve this, we used phase scrambling of images [17]. As texture sensitive neurons are assumed to be sensitive to high-level statistics, their removal is expected to reduce model V2 neuron activations. This specific reduction is observed in texture-sensitive units in V2 (Fig. 3g), in line with findings in macaque recordings (Fig. 3h).

In summary, V1 and V2 representations of TDVAE display key features of V1 and V2

Fig. 3 *(following page)*. Signatures of learning the invariances of natural textures in V1 and V2. **a**, Mean responses of a selected model neuron from V1 and V2 to three sample images from three texture families. V2 responses vary less within than across families, in contrast to V1 responses. Below: violin plots of mean responses of the same V1 and V2 model neurons to a large number of texture samples (12,800 samples per family; center lines: means, whiskers: extrema). Variations of per-family means are comparable to per-family variances in V2 unlike in V1 where they are much smaller. **b**, Same as **a** in V1 and V2 recordings from macaque monkeys. Reproduced with permission from [18](PNAS). **c**, Partitioning single unit response variances into a component across families, a component across samples (within families) and a residual component (across stimulus repetitions) with nested ANOVA (top: $n = 100$ model neurons from V2, bottom: $n = 100$ model neurons from V1). Left: the first two components of variance for each neuron (dots). Insets: distribution of the sum of these two components. Right: distribution of the ratio of the same two components. Geometric means of variance ratios are 0.053 for V1 and 1.022 for V2 (triangles). The difference is significant (one-sided t-test in the log domain, $P = 1.2 \times 10^{-64}$, $t = -25.4$, $df = 198$). **d**, Same as **c** in 102 V1 and 103 V2 units from 13 macaque monkeys. Variance ratio geometric means are 0.4 in V1 and 0.63 in V2. Their difference is significant (t-test in the log domain, $P < 0.001$). Reproduced with permission from [18](PNAS). **e**, Decoding texture family (red) and within-family stimulus identity (black) from the same 100 V1 units (horizontal axis) and 100 V2 units (vertical axis) as in **c**. Dot sizes: number of randomly selected model units used for the decoding analysis ($n = 1, 3, 10, 30, 100$). Chance performance for the family classification task was 1/5 and for the sample classification task 1/15. Dot centers represent means and error bars represent 95% confidence intervals of the bootstrapped distribution over included neurons and cross-validation partitioning. **f**, Same as **e** from the single unit V1 and V2 macaque recordings in **d**. Here, chance performance was 1/15 for both the family and sample classification tasks. Reproduced with permission from [18](PNAS). **g**, Phase scrambling induced modulation of the magnitude of mean responses to texture images in the same 100 V1 units (bottom) and 100 V2 units (top) as in **c**. Per-neuron per-family modulation index was defined as the difference between sample- and presentation-averaged absolute responses to undisturbed texture images and their phase scrambled counterparts divided by their sum. Histograms show per-neuron modulation indices averaged over all texture families. Mean modulation indices are -0.0114 in V1 and 0.0046 in V2 (triangles). Their difference is significant (one-sided t-test, $P = 0.0050$, $t = -2.60$, $df = 198$). **h**, Same as **g** from the above V1 and V2 recordings in macaques. The mean modulation indices differed significantly (t-test on signed modulations, $P < 0.0001$). Reproduced with permission from [17] (Nature Neuroscience, reproduced with permission from SNCSC)


a
texture families
texture samples
V2
V1
mean response
b
texture families
texture samples
mean spike count
Experiment
c
Variance across families (%)
Variance across samples (%)
Stimulus-driven variance (%)
Proportion
Proportion of cells
Variance ratio
V2
n = 100
V1
n = 100
d
Experiment
V2
n = 103
V1
n = 102
e
n = 1, 3, 10, 30, 100
Family classification
Sample classification
Chance for family classification
Chance for sample classification
V2 performance (proportion correct)
V1 performance (proportion correct)
f
Experiment
n = 1, 3, 10, 30, 100
Family classification
Sample classification
Chance
g
V2
n = 100
V1
n = 100
Modulation index
h
Experiment
V2
n = 103
V1
n = 102

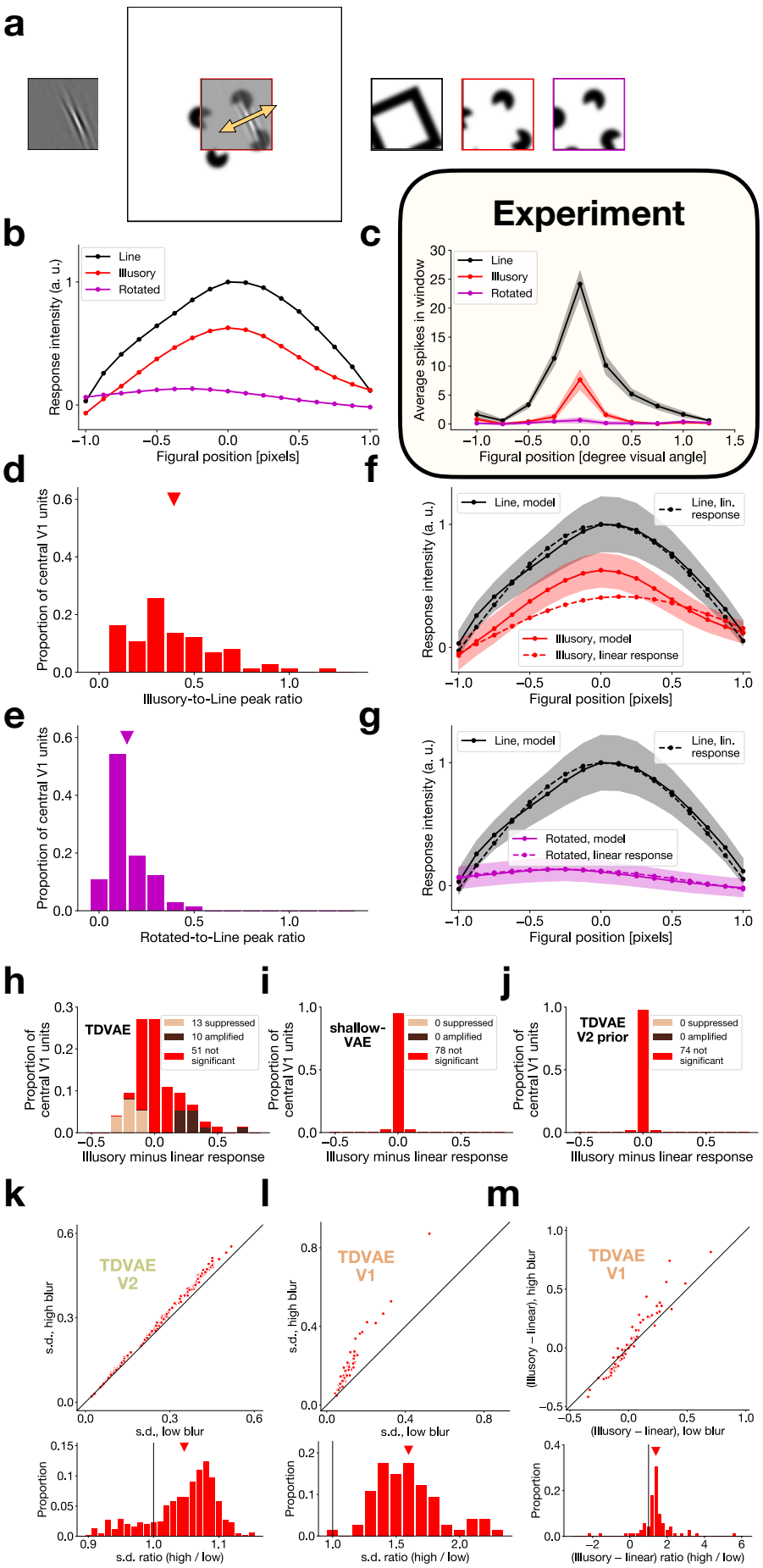
representations of the macaque visual cortex. In particular, an elaborate texture representation is a salient feature of hierarchical representations of natural images.

Top-down contributions to hierarchical inference. In the goal-general model, the representation serves to infer the features underlying the observations and, importantly, in a hierarchical model, inference can be performed on all levels of computing (Fig. 1d). According to equation (1), top-down influences affect the inference of V1 activities, which can be interpreted as establishing a contextual prior for V1 (Fig. 1g). Importantly, this prior is not an invariant component of computations as expected in simple Bayesian computations, instead the prior depends on the context, that is on the higher-level interpretation of the scene. In the following sections we seek to identify signatures of this contextual prior in V1 response statistics: first on mean responses, second on response correlations.

Top-down influences can be expected when local interpretation of images depends on the

---

Fig. 4 *(following page)*. Top-down influence induced contextual priors result in illusory contour responses. **a**, Illustration of the illusory contour experiment. Left: receptive field of an example model V1 neuron. Middle: Illustration of the Kanizsa square stimulus aligned to the filter. Red border indicates the actual boundary of the stimulus. Arrow indicates the direction of shifts of the stimulus relative to the neural receptive field. Right: real square ('Line'), Kanizsa illusory contour ('Illusory') and incongruent ('Rotated') stimuli for gauging model neuron responses. **b**, Mean responses of the model V1 neuron in **a** to 10,000 presentations of the three stimuli in **a** as a function of stimulus displacement. Shaded regions (invisibly small): s.e.m. **c**, Same as **b** for a selected unit in the V1 of a macaque monkey. Number of trials is unknown. Reproduced from [43] with permission (PNAS, Copyright (2001) National Academy of Sciences, U.S.A.). **d**, Ratios of mean V1 responses to 'Illusory' and 'Line' stimuli at the 'Line' response peak for central, localized, medium wavelength V1 filters fitting into the image patch ($n = 74$). Mean peak ratio: 0.39 (triangle). **e**, Same as **d** for 'Rotated' vs. 'Line' stimuli. Mean peak ratio: 0.15 (triangle) which is significantly less than the mean peak ratio in **d** (one-sided t-test, $P = 3.1 \times 10^{-14}$, $t = 8.3$, $df = 146$). **f**, Same as **b** for the 'Line' and 'Illusory' stimuli, together with linear responses of the V1 filter to the same stimuli (dashed lines). Magnitudes of the mean and linear response sets were scaled together at the 'Line' mean response peak. Shaded regions: s.d. (as opposed to s.e.m. in **b**). **g**, Same as **f** for the 'Line' and 'Rotated' stimuli. **h**, Differences between the mean and linear responses to the 'Illusory' stimulus for the same V1 filters as in **d**. Out of the $n = 74$ filters studied, 13 showed significant suppression and 10 showed significant amplification compared to the linear response (colors). Significance was defined as the mean response being lower (suppression) or higher (amplification) than the linear response in at least 95% of $n = 10,000$ trials. **i**, Same as **h** in the non-hierarchical model. None of the $n = 78$ central, localized, medium wavelength V1 filters fitting into the image patch displays a significant illusory effect. **j**, Same as **h** with the stimulus-dependent V2 contextual prior replaced with the stimulus-independent V2 prior. The illusory effect disappears. **k**, Top: V2 response variances to the $n = 74$ aligned 'Illusory' stimuli in **d** filtered with high vs. low Gaussian blur along each active V2 dimension in the TDVAE model (dots). Bottom: distribution of the ratios of the high and low blur V2 response variances. The mean ratio 1.047 is significantly larger than 1 (one-sided t-test, $P = 3.1 \times 10^{-61}$, $t = 19.4$, $df = 443$). **l**, Top: increased Gaussian blur (as in **k**) also affects the response variances of the $n = 74$ V1 filters in **d** to the 'Illusory' stimuli. Bottom: the mean V1 response variance ratio 1.60 is significantly larger than 1 (one-sided t-test, $P = 1.2 \times 10^{-30}$, $t = 19.3$, $df = 73$). **m**, Top: same as **l** for the illusory effect size, i.e., the difference between the mean and linear V1 responses. Bottom: the mean illusory effect strength ratio 1.39 is significantly larger than 1 (one-sided t-test, $P = 4.7 \times 10^{-4}$, $t = 3.45$, $df = 71$).


a
b
Line
Illusory
Rotated
Response intensity (a. u.)
Figural position [pixels]
c
Experiment
Average spikes in window
Figural position [degree visual angle]
d
Proportion of central V1 units
Illusory-to-Line peak ratio
e
Rotated-to-Line peak ratio
f
Line, model
Line, lin. response
Illusory, model
Illusory, linear response
g
Rotated, model
Rotated, linear response
h
TDVAE
13 suppressed
10 amplified
51 not significant
Illusory minus linear response
i
shallow-VAE
0 suppressed
0 amplified
78 not significant
j
TDVAE V2 prior
0 suppressed
0 amplified
74 not significant
k
TDVAE V2
s.d., high blur
s.d., low blur
Proportion
s.d. ratio (high / low)
l
TDVAE V1
m
TDVAE V1
(Illusory – linear), high blur
(Illusory – linear), low blur
(Illusory – linear) ratio (high / low)

surroundings. Establishing a wider context is directly supported by the growing receptive field sizes in higher hierarchical layers of TDVAE. A particularly strong form of contextual effect can be established by attempting to selectively block direct stimulus effects from designated V1 neurons and therefore top-down effects can be studied in isolation by investigating the emerging V1 activation. This 'illusory' activity in V1 neurons resonates with the interpretation of illusions as the contribution of priors to perception under uncertainty [44, 45]. In practice, using stimuli that evoke V2 activations by relying on characteristic features of textures such as co-alignment of distant line segments but are blank within the receptive field of a V1 neuron can be particularly strong in our learned hierarchical representation. Illusory contours arising in Kanizsa-illusions are good candidates for such stimuli. We constructed Kanizsa square stimuli (Fig. 4a) aligned to individual V1 filters and measured mean responses (Fig. 4b). We compared Kanizsa stimuli-elicited responses to both unobstructed images of squares and to control stimuli that were built from identical elements to the Kanizsa stimuli but in configurations not congruent with the percept of a square. Responses of individual neurons to an illusory edge was similar to the response to an unobstructed square stimulus, albeit the gain of the response was lower for the illusory edge (Fig. 4b), similar to responses of neurons in V1 of macaques [43] (Fig. 4c). Confirming expectations, responses to Kanizsa-incongruent stimulus elements were muted (Fig. 4b), again reproducing experimental findings (Fig. 4c). This tendency was consistent across the model neurons investigated using stimuli that were tailored to their response properties (see Methods, Fig. 4d,e).

Using the TDVAE model, we designed a stricter control. Responses of model neurons to the Kanizsa stimuli could be boosted by any overlap between the receptive field of the neuron and the flanking Kanizsa elements. To isolate the effect of the top-down computations, we calculated linear responses of the neurons to the Kanizsa stimuli (Fig. 4f). We found significant boosting of the linear response in TDVAE for the example neuron. In the whole population of analyzed neurons ($n = 74$) not only boosting ($n = 10$) but also suppression ($n = 13$) and modulation-free examples ($n = 51$) could be identified (Fig. 4h). Thus, besides contextual boosting of neural responses, contextual suppression is also present in the population. Incongruent Kanizsa stimuli produced limited modulation in model V1 neurons (significant boosting: $n = 1$; significant suppression: $n = 6$) (Extended Data Fig. 8). Confirming the contributions of top-down connections to illusory activity, modulation of V1 responses relative to the linear responses was absent in a model that lacked V2 (Fig. 4i).

To demonstrate that the illusory contour responses fulfill some fundamental requirements of probabilistic inference, we performed additional experiments. First, since the contextual prior results from the stimulus-specific activation of V2 neurons, we expect that when V2 activity is not constrained by the stimulus, i.e., when V2 is only reflecting the prior, then the activity of V1 neurons return to the linear response. This is exactly what we observe in the TDVAE model (Fig. 4j). Next, as probability theory suggests that inference relies on the prior more when the evidence about observations is degraded, we investigated if degradation of stimulus quality results in predictable changes in illusory contour responses. We introduced Gaussian blur to the Kanizsa stimuli with increasing variance. V2 model responses displayed increased noise with higher levels of blur (Fig. 4k). Our prediction is that the top-down effect will be stronger with increasing blur, since the network relies increasingly on the contextual prior. Indeed, the top-down effect became stronger with increased blur, as measured by the deviation of V1 model neuron responses from the linear response (Fig. 4m), while becoming more noisy (Fig. 4l).

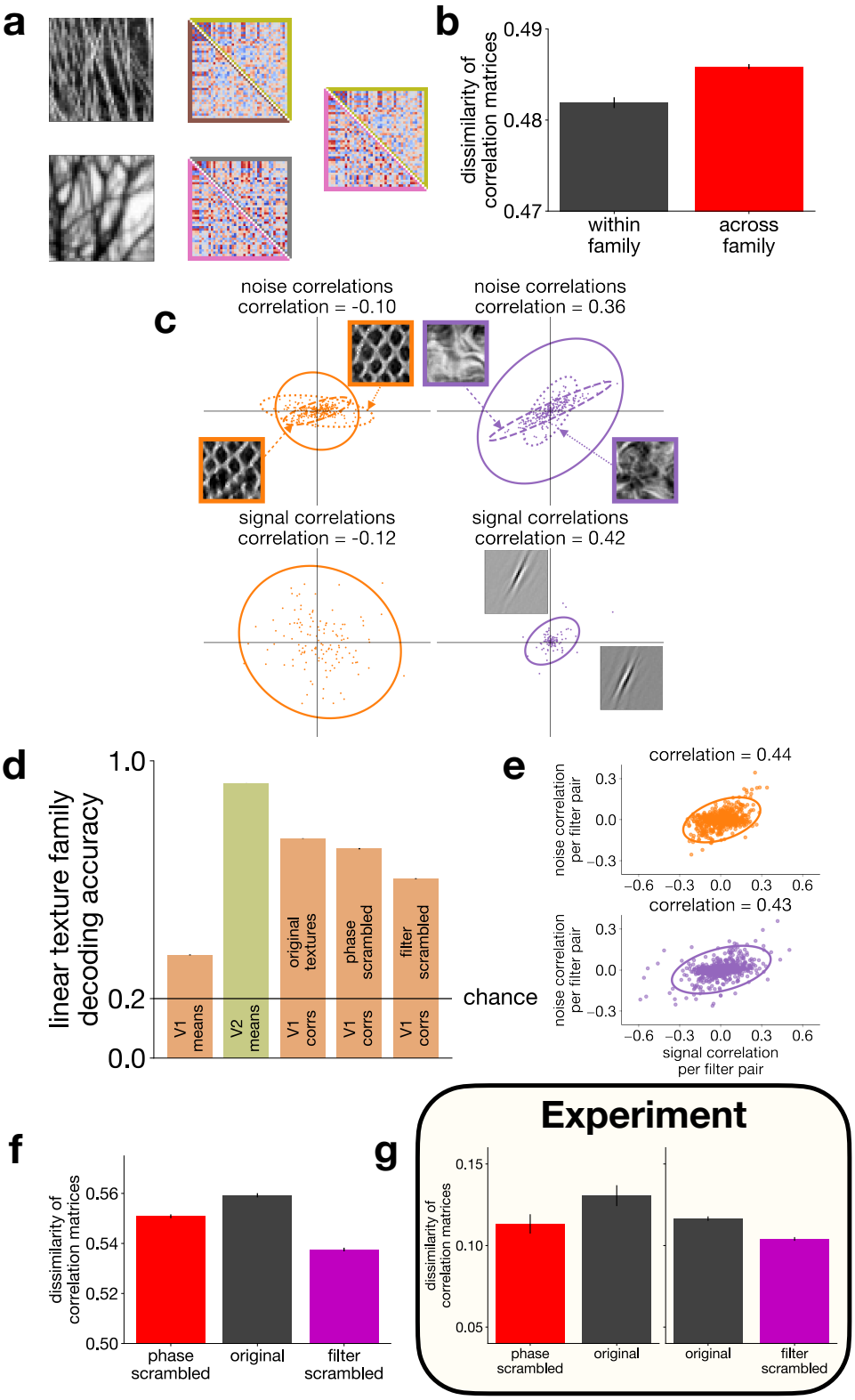
Response correlations emerging through contextual priors. Illusory contours demonstrate contextual effects in which low-level features represented in V1 neurons are affected by more complex features that are represented higher in the visual hierarchy, V2 in our specific case. Such contextual modulations constitute a specific form of a contextual prior: contextual priors carry information about the regularities in the surroundings of a neuron's receptive field,

such as the tendency of another neuron to be jointly active in the local environment. These regularities in the local environment of the receptive field can change from image to image and thus the changing contextual prior dictates different co-activation patterns between neurons (Fig. 5a, Extended Data Figs. 9a,b). Such stimulus-specific co-activation patterns are captured in noise correlations between neuronal responses and thus noise correlations can inform us about contextual priors. Importantly, if regularities in a set of images are similar, the contextual prior and thus the noise correlations are not changing. In particular, texture statistics are defined through a range of across-space, across-spatial scale correlations between the localized filters represented in V1 [42] therefore texture family-specific activity in V2 is expected to induce texture family-specific noise correlations in V1 (Fig. 5b,c).

We measured the texture-family specificity of noise correlations in the TDVAE model by constructing linear decoders that we applied to noise correlation matrices of a small model V1 neuron population (Fig. 5d). For this, we calculated noise correlations between pairs of

---

Fig. 5 *(following page)*. Contribution of top-down influences to response correlations in V1. **a**, V1 noise correlations for $n = 40$ central, localized, medium wavelength V1 filters (rows and columns) in the TDVAE model for two example natural images (left). Center: lower and upper triangles correspond to correlations calculated from two independent sets of trials ($n = 64$ each) for the image in the same row. Differences are due to sampling error. Right: across-image noise correlation differences. Data taken from the central column (border colors). **b**, Across-image variability of noise correlation matrices (each calculated from $n = 64$ trials each for the same 40 V1 neurons as in **a**) within and across texture families ($n = 32$ sample images per family). Mean within-family correlation dissimilarity 0.482 is significantly smaller than mean across-family correlation dissimilarity 0.486 (one-sided t-test, $P = 8.6 \times 10^{-11}$, $t = -6.4$, df = 12,718). **c**, Top: mean responses (dots) of two model V1 neurons (see to the bottom right) to two images (insets) from two texture families (colors; $n = 128$ trials for each image). Dashed and dotted ellipses: single-image covariance ellipses. Solid ellipses: average covariance ellipses for $n = 128$ images per texture family. Text labels: per-family mean noise correlation coefficients. Bottom: trial-averaged mean responses of the same V1 units ($n = 128$ trials per image) to $n = 128$ images (dots) per texture family (colors). Covariance ellipses, text labels: signal correlations calculated from these data. Note that respective noise (top) and signal (bottom) correlation coefficients are similar. **d**, Linear decodability of texture family from V1 and V2 mean responses (like in Fig. 2d) and V1 noise correlations. Decoders were constructed for true texture images (left three bars), as well as phase and filter scrambled versions of texture images (right two bars). Bar heights: means for $n = 5$ random seeds; error bars: s.d. Solid black line indicates chance. **e**, Relationship between the noise correlations and signal correlations for a population of model neuron pairs (dots) in two texture families. While both noise and signal correlations change across families, there is a strong dependence between signal and noise correlations that is characteristic across the population. Colors and data as on panel **c**. **f**, The effect of phase and filter scrambling on across-stimuli noise correlation variations (40 natural images, four noise correlation matrices per image, each calculated from 32 trials). Mean correlation dissimilarities (original: 0.559, phase scrambled: 0.551, filter scrambled: 0.538) are significantly different (original vs. phase scrambled: one-sided t-test, $P = 3.8 \times 10^{-14}$, $t = 7.48$, df = 24,958; original vs. filter scrambled: one-sided t-test, $P = 1.7 \times 10^{-86}$, $t = 19.75$, df = 24,958). Error bars: s.e.m. **g**, Same as **f** in V1 multiunit recordings from macaque monkeys. Mean correlation dissimilarities (phase scrambled: 0.113, original to the left: 0.131; original to the right: 0.116, filter scrambled: 0.104) are significantly different pairwise (original to the left vs. phase scrambled: t-test, $P = 2.12 \times 10^{-2}$, $t = 2.31$, df = 644; original to the right vs. filter scrambled: t-test, $P = 4.17 \times 10^{-10}$, $t = 6.26$, df = 7468). Reproduced from Figs. S9 and 5c in [46] with permission (PNAS).


a
b
dissimilarity of correlation matrices
0.49
0.48
0.47
within family
across family
c
noise correlations
correlation = -0.10
noise correlations
correlation = 0.36
signal correlations
correlation = -0.12
signal correlations
correlation = 0.42
d
linear texture family decoding accuracy
1.0
0.2
0.0
chance
V1 means
V2 means
V1 corrs
original textures
V1 corrs
phase scrambled
V1 corrs
filter scrambled
e
correlation = 0.44
correlation = 0.43
noise correlation per filter pair
signal correlation per filter pair
0.3
0.0
−0.3
−0.6
0.6
f
dissimilarity of correlation matrices
0.56
0.54
0.52
0.50
phase scrambled
original
filter scrambled
g
Experiment
0.15
0.10
0.05
phase scrambled
original
original
filter scrambled

V1 neurons through repeated presentation of the same texture image and we repeated this calculation for a large number of texture images from five texture families. Critically, using the correlation decoders, we sought to capture the texture-related information in V1 response correlations, as we also did by training decoders on the mean responses of V2 neurons. Indeed, we found reliable decoding of texture family in both V1 correlations and V2 means (Fig. 5d). In contrast to V1 response correlations, V1 response means are worse suited to linear decoding of texture family information (Fig. 5d).

The contextual prior reflects the regularities of the environment that are specific to the particular context. For texture families, the tendency of a pair of neurons to be jointly active for different texture images, referred to as signal correlation, is expected to be captured by the contextual prior of a well trained model. Indeed, Karklin and Lewicki argued that the correlations between mean filter responses is characteristic to texture families [47], and we also found that such texture family specific signal correlations were present between an example pair of neurons (Fig. 5c). Thus, the contextual prior is related to two different forms of joint response statistics: the signal correlations, measuring correlations between mean responses across images, and the noise correlations, measuring correlations between neurons following multiple presentations of the same image. Therefore, we measured for every texture family the relationship between these two quantities for a large number of neuron pairs of model V1 neurons. We found strong dependence of noise correlations on signal correlations in the responses of pairs of V1 neurons (Fig. 5e, Extended Data Fig. 9c).

We test the consequences of the proposal that noise correlations are shaped by contextual priors in a modified setting. We argued that stimulus-specific noise correlations arise due to top-down feedback that carries information about high-level inferences. Consequently, removal of high-level structure is expected to result in reduced stimulus-specificity of noise correlations in V1. To test this, we used the same stimulus manipulations that we used for exploring the properties of V1 and V2 representations: phase and filter scrambling of natural stimuli (Methods). We calculated the dissimilarity of noise correlation matrices for natural images and for phase- and filter-scrambled versions of the same natural images (Methods). We found reduced dissimilarity for phase and filter scrambled versions of natural images (Fig. 5f). Similar reduction in noise correlations for phase- and filter-scrambled versions of natural images was found in the V1 of macaques [46] (Fig. 5g). Such patterns in noise correlations provide support that contextual priors emerge as a consequence of hierarchical inference.

Discussion

In this paper we introduced TDVAE, a deep learning model of natural images, which permits the joint study of the representations and computations in the early visual areas of the ventral stream. TDVAE is a departure from the goal-directed deep learning framework, which has proven highly successful in predicting neural activity in the ventral stream, inspired by the computational requirements of the biological brain: enabling richer inferences while relying on considerably sparser teaching signals. A key characteristic of the resulting deep generative modeling framework is the recruitment of top-down interactions to perform the computations, a defining feature of the cortical organization. Crucially, the proposed TDVAE model is end-to-end trained on natural images, leaving no free parameter for tuning. This parameter-free approach permitted to account for subtle details of the texture representation identified in macaques and could also help identify signatures of top-down computations in the means and the correlations of neuronal responses. Beyond reproducing a wide range of experimental findings, we demonstrate that the framework opens a new window on exploring the representation, computations, and the circuitry of the early visual system.

Perception has long been argued to be supported by generative models of the environment. Generative models provide an essential framework for learning in an unsupervised manner.

Conceptually, generative models summarize our knowledge about the world and describe the way observations are produced as a result of the interactions between (a hierarchical set of) features. Perception is assumed to invert the generative process by inferring the features that underlie observations [11, 35]. However, tools to investigate perceptual processes related to interpreting natural scenes through generative models have been scarce. Considerable successes were achieved with non-hierarchical and essentially linear models in understanding the computations in the primary visual cortex (V1) [30, 37, 38, 47, 48, 49], as well as with manually designed hierarchical extensions [49, 50]. However, training models that are both hierarchical and nonlinear using natural images has remained an unresolved problem. We proposed to overcome this obstacle by recruiting recently developed tools in machine learning, Variational Autoencoders (VAEs) [25, 26] and addressed learning and computations in the first two stages of the visual cortical processing hierarchy, V1 and V2. We developed the TDVAE architecture such that it overcomes a major limiting factor of earlier generative models used for biological vision, the limit to constrain generative models to linear computations. VAEs have previously been applied to account for highly nonlinear computations in the visual cortex, face perception [21]. A fundamental difference between that account and our treatment is that the TDVAE model is capable of performing hierarchical inference, a feature central to the investigation of top-down influences. A further difference concerns the way the model was trained. The model in [21] investigated a constrained setting where the learned face representation was investigated by training the model on face images, while in our case the training was on natural images, thus the learned representations reflect the general properties of the natural environment. The proposed TDVAE architecture was developed such that it relies on a limited set of assumptions while remaining as faithful as possible to the principles of probabilistic inference. VAEs rely on an approximate form of inference (amortized variational inference). In order to stick with the guarantees of VAEs, we did not rely on extensions of the framework that encourage disentanglement [21, 51], regularization methods that serve more stable learning but might distort the learned representations (see Methods), or methods that help to limit the number of learned parameters/amount of data used for training the model (convolution). Our two-layer generative model can be extended to higher levels of the hierarchy but this requires both practical considerations (e.g., limiting the inflation of the parameter space by relying on convolutional architectures), and exploring further inductive biases that can disentangle the computational role of further nonlinear computations.

Top-down interactions in TDVAE were the result of hierarchical inference. Top-down interactions have also been identified in an alternative computational framework, predictive coding [52, 53]. The two frameworks have strong parallels: both are unsupervised, rely on a generative model of the inputs, and are hierarchical. The two frameworks can be distinguished based on the goal the top-down connections serve. In our account, top-down interactions emerge as a consequence of fundamental principles of probability theory [13], and they contribute to rich inferences that can serve flexible task execution [23]. In the predictive coding framework top-down connections serve an efficient representation of the stimuli, which can ensure that neural expenditures are minimized when the changes in the environment are limited [54]. A critical difference between the two frameworks is how inference is performed. An important computational property of the hierarchical inference framework proposed here is that inferences made about the environmental features are probabilistic. Probabilistic inference augments classical inference frameworks by representing the uncertainty associated with the interpretation of the incoming stimuli [15, 55]. Building a probabilistic representation is an efficient way to deal with ambiguous observations, limited data, and machine learning arguments highlight the necessity of such computations when dealing with the natural environment [56]. The ability to perform probabilistic computations has gained strong support from behavioral studies [14, 57, 58]. More importantly, the representation of uncertainty has been identified in neuronal responses in V1 [55] and such probabilistic computations have explained a wide range of phenomena, including

response variability, noise correlations, oscillatory activity, onset transients [37, 38, 59, 60, 61]. Importantly, representation of uncertainty is not part of the existing formulations of predictive coding models (inference is performed at the maximum a posteriori level [62]), albeit proposals have been made to extend the original framework of predictive coding in this direction [63]. In practice, our results on stimulus-dependent correlations are tightly linked to a probabilistic representation and are therefore beyond the scope of predictive coding accounts. A second motif that distinguishes predictive coding models from the hierarchical inference model presented here is the role of error coding units. Predictive coding postulates the existence of two distinct classes of neurons [64]. Neurons in the deeper 5/6 layers of the cortex are responsible for internal state representation while 2/3 layer neurons are error detecting neurons representing the difference of sensory input and top-down prediction. In the presented VAE framework the computational role of the latter type of neurons was not discussed.

The results presented here provide further evidence that response statistics beyond the mean carries meaningful information about the computations taking place in the visual cortex [37, 46, 65, 66, 67]. How neural circuits implement these computations is a critical question. A class of recurrent neural networks, stabilized supralinear networks were shown to produce response statistics with well controlled correlation and variance patterns [68], which framework has also been shown to be particularly apt to perform probabilistic computations in V1 [60].

Our argument about the role of top-down interactions emphasized their role in inferring perceptual variables: the top-down inputs helped establish a contextual prior for lower-level features through a higher-level, more global interpretation of the stimulus. Nonetheless, a wide array of studies consider attentional modulation as an important factor contributing to top-down interactions [69]. Reconciling these alternative interpretations is a critical challenge. We argue that considering task-relevant variables besides perceptual variables can be interpreted in the same inference framework [70]. Indeed, top-down delivered information in the early visual cortex was shown to be specific to the task being performed in humans [71]. Also, it has been argued that attentional effects can rely on a similar computational scheme, in particular hierarchical inference, to that proposed here for perceptual inference [66, 72]. Interestingly, the top-down modulation of noise correlation patterns discussed in our paper has also been demonstrated in the context of changing the task being performed [67]. This establishes strong parallels of the influences of task variables [69] and task-general higher-level perceptual variables on computations at the level of V1. The question how perceptual and task-related variables are integrated in the circuitry that establishes top-down interactions remains a largely open question but recent studies indicate that they may share the same communication pathways [73].

We argued that the emergence of illusions can distinguish goal-driven models with essentially feed-forward architectures from goal-general models that perform probabilistic inference through establishing contextual priors. Indeed, emergence of other illusions have been reported to be at odds with goal-directed models [74]. Beyond illusions, another form of challenge to goal-directed models are adversarial images [75]. Excitingly, similar adversarial images were found to match those stimuli that displayed signatures of the recruitment of computations different from feed-forward computations [76]. It was proposed that recurrent connections that supplement the feed-forward architecture can resolve the issue of adversarial images and could provide a motivation for the architectures observed in biological systems. While there is indication that at the level of inferotemporal cortex top-down input from prefrontal cortex contributes to visual computations [76], it remains an open question if recurrent neural network models that are trained in a data driven manner implement similar computations that are implied by the principles of hierarchical probabilistic inference. While illusions are key signatures of the contextual priors that emerge through top-down interactions, these interactions have wider ranging effects. Recently, modulation of V1 neurons by the visual context of their receptive field has received support from a hierarchical inference account [50]. Our approach to use end-to-end

natural image training provides a general framework to investigate such effects, as also revealed by the rich modulation patterns we identified for Kanizsa stimuli (Fig. 4h). These illusions are very specific examples of top-down modulations but top-down effects are widespread in perceptual processes. These effects are harder to disentangle from feed-forward contributions but novel tools promise to break this impasse by providing a quantitative testbed for theories [77].

In our study of hierarchical computations, we found a robust texture representation in a layer homologous to V2 of the cortical hierarchy. In our approach we did not encourage learning texture-like representations, since we trained the model on natural images. The motivation to explore texture-related responses came from neuroscience. Still the question remains: is there something specific about textures that yields the emergence of our texture representation? We believe that there are multiple contributors for the specific representation that was learned in TDVAE. The statistics relevant to textures have a characteristic scale. Intuitively, more complex features might have statistical signatures spanning even larger spatial scales. As a consequence, it can be conjectured that the limited patch size actually affected the features TDVAE developed sensitivity to. Another factor that can affect the learned representation is the computational complexity of the generative model. As we have shown, the inductive bias implemented by the linearity of $p(\text{image} \,|\, \text{V1})$ (along with the shape of the prior) critically contributed to Gabor-like features at the level of V1 in the model. Certainly, such inductive biases are present in the ventral stream, limiting the complexity of computations performed in V2. Further, dynamics of features could also shape the representations at different levels of the hierarchy [78]. Taken together, our results demonstrate a very robust texture representation that emerges for natural images in a nonlinear generative model. The constraints implicit in our implementation of TDVAE contributed to this robustness, and provide the necessary insights to build generative models with deeper hierarchies that capture more complex features and higher-level processing in the visual cortex.

References

[1] Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nature Neuroscience* 22, 1761–1770 (2019).

[2] Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience* 19, 356–365 (2016).

[3] Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* 364, eaav9436 (2019).

[4] Walker, E. Y. *et al.* Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience* 22, 2060–2065 (2019).

[5] Conwell, C. *et al.* Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex. *Advances in Neural Information Processing Systems* 34, 5590–5607 (2021).

[6] Nayebi, A. *et al.* Goal-driven recurrent neural network models of the ventral visual stream. *bioRxiv* (2021). URL `https://doi.org/10.1101/2021.02.17.431717`.

[7] Nayebi, A. *et al.* Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology* 19, e1011506 (2023).

[8] Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications* 13, 491 (2022).

[9] Zhuang, C. *et al.* Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences* 118, e2014196118 (2021).

[10] De Lange, F. P., Heilbron, M. & Kok, P. How do expectations shape perception? *Trends in Cognitive Sciences* 22, 764–779 (2018).

[11] Von Helmholtz, H. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, vol. 9 (Voss, 1867).

[12] Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271–304 (2004).

[13] Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *JOSA A* 20, 1434–1448 (2003).

[14] Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* 427, 244–247 (2004).

[15] Berkes, P., Orbán, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* 331, 83–87 (2011).

[16] Freeman, J. & Simoncelli, E. P. Metamers of the ventral stream. *Nature Neuroscience* 14, 1195–1201 (2011).

[17] Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience* 16, 974–981 (2013).

[18] Ziemba, C. M., Freeman, J., Movshon, J. A. & Simoncelli, E. P. Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences* 113, E3140–E3149 (2016).

[19] Nurminen, L., Merlin, S., Bijanzadeh, M., Federer, F. & Angelucci, A. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nature Communications* 9, 1–13 (2018).

[20] Bolaños, F. *et al.* Efficient coding of natural images in the mouse visual cortex. *bioRxiv* (2022). URL `https://www.biorxiv.org/content/early/2022/09/17/2022.09.14.507893`.

[21] Higgins, I. *et al.* Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications* 12, 1–14 (2021).

[22] Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology* 10, e1003915 (2014).

[23] Csikor, F., Meszéna, B., Szabó, B. & Orban, G. Top-down effects in an early visual cortex inspired hierarchical Variational Autoencoder. In *SVRHM 2022 Workshop @ NeurIPS* (2022). URL `https://openreview.net/forum?id=8dfboOQfYt3`.

[24] Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nature Reviews Neuroscience* 14, 350–363 (2013).

[25] Kingma, D. P. & Welling, M. Auto-encoding Variational Bayes (2013). URL `https://doi.org/10.48550/arXiv.1312.6114`.

[26] Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. & Jebara, T. (eds.) *Proceedings of the 31st International Conference on Machine Learning*, vol. 32 of *Proceedings of Machine Learning Research*, 1278–1286 (PMLR, Bejing, China, 2014). URL `https://proceedings.mlr.press/v32/rezende14.html`.

[27] Hazami, L., Mama, R. & Thurairatnam, R. Efficient-VDVAE: Less is more (2022). URL `https://doi.org/10.48550/arXiv.2203.13751`.

[28] Child, R. Very deep VAEs generalize autoregressive models and can outperform them on images (2021). URL `https://doi.org/10.48550/arXiv.2011.10650`.

[29] Vahdat, A. & Kautz, J. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems* 33, 19667–19679 (2020).

[30] Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609 (1996).

[31] Bell, A. J. & Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Research* 37, 3327–3338 (1997).

[32] Barello, G., Charles, A. S. & Pillow, J. W. Sparse-coding variational auto-encoders. *bioRxiv* (2018). URL `https://www.biorxiv.org/content/early/2018/08/29/399246`.

[33] Knill, D. C. & Richards, W. *Perception as Bayesian inference* (Cambridge University Press, 1996).

[34] Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience* 14, 926–932 (2011).

[35] Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences* 10, 301–308 (2006).

[36] Geisler, W. S. Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192 (2008).

[37] Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* 92, 530–543 (2016).

[38] Festa, D., Aschner, A., Davila, A., Kohn, A. & Coen-Cagli, R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nature Communications* 12, 1–11 (2021).

[39] Carandini, M. Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS Biology* 2, e264 (2004).

[40] Abbott, L. F. & Dayan, P. The effect of correlated variability on the accuracy of a population code. *Neural Computation* 11, 91–101 (1999).

[41] Yu, Y., Stirman, J. N., Dorsett, C. R. & Smith, S. L. Selective representations of texture and motion in mouse higher visual areas. *Current Biology* 32, 2810–2820 (2022).

[42] Portilla, J. & Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision* 40, 49–70 (2000).

[43] Lee, T. S. & Nguyen, M. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences* 98, 1907–1911 (2001).

[44] Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nature Neuroscience* 5, 598–604 (2002).

[45] Notredame, C.-E., Pins, D., Deneve, S. & Jardri, R. What visual illusions teach us about schizophrenia. *Frontiers in Integrative Neuroscience* 8, 63 (2014).

[46] Bányai, M. *et al.* Stimulus complexity shapes response correlations in primary visual cortex. *Proceedings of the National Academy of Sciences* 116, 2723–2732 (2019).

[47] Karklin, Y. & Lewicki, M. S. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* 457, 83–86 (2009).

[48] Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature Neuroscience* 4, 819–825 (2001).

[49] Coen-Cagli, R., Kohn, A. & Schwartz, O. Flexible gating of contextual influences in natural vision. *Nature Neuroscience* 18, 1648–1655 (2015).

[50] Fu, J. *et al.* Pattern completion and disruption characterize contextual modulation in mouse visual cortex. *bioRxiv* (2023). URL `https://www.biorxiv.org/content/early/2023/03/14/2023.03.13.532473`.

[51] Burgess, C. P. *et al.* Understanding disentangling in β-VAE (2018). URL `https://doi.org/10.48550/arXiv.1804.03599`.

[52] Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2, 79–87 (1999).

[53] Lotter, W., Kreiman, G. & Cox, D. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence* 2, 210–219 (2020).

[54] Ali, A., Ahmad, N., de Groot, E., van Gerven, M. A. J. & Kietzmann, T. C. Predictive coding is a consequence of energy efficiency in recurrent neural networks. *Patterns* 3 (2022).

[55] Walker, E. Y., Cotton, R. J., Ma, W. J. & Tolias, A. S. A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience* 23, 122–129 (2020).

[56] Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* 521, 452–459 (2015).

[57] Kording, K. Decision theory: What "should" the nervous system do? *Science* 318, 606–610 (2007).

[58] Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433 (2002).

[59] Hoyer, P. & Hyvärinen, A. Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in Neural Information Processing Systems* 15 (2002).

[60] Echeveste, R., Aitchison, L., Hennequin, G. & Lengyel, M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience* 23, 1138–1149 (2020).

[61] Aitchison, L. & Lengyel, M. The Hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Computational Biology* 12, e1005186 (2016).

[62] Boutin, V., Zerroug, A., Jung, M. & Serre, T. Iterative VAE as a predictive brain model for out-of-distribution generalization. In *NeurIPS 2020 Workshop SVRHM* (2020). URL `https://openreview.net/forum?id=jE6SlVTOFPV`.

[63] Aitchison, L. & Lengyel, M. With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology* 46, 219–227 (2017).

[64] Rao, R. P. & Jiang, L. P. Predictive coding theories of cortical function. In *Oxford Research Encyclopedia of Neuroscience* (2022).

[65] Goris, R. L., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. *Nature Neuroscience* 17, 858–865 (2014).

[66] Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* 90, 649–660 (2016).

[67] Bondy, A. G., Haefner, R. M. & Cumming, B. G. Feedback determines the structure of correlated variability in primary visual cortex. *Nature Neuroscience* 21, 598–606 (2018).

[68] Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M. & Miller, K. D. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron* 98, 846–860 (2018).

[69] Lange, R. D. & Haefner, R. M. Task-induced neural covariability as a signature of approximate Bayesian learning and inference. *PLoS Computational Biology* 18, e1009557 (2022).

[70] Bányai, M. & Orbán, G. Noise correlations and perceptual inference. *Current Opinion in Neurobiology* 58, 209–217 (2019).

[71] Harel, A., Kravitz, D. J. & Baker, C. I. Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences* 111, E962–E971 (2014).

[72] Lange, R. D. & Haefner, R. M. Characterizing and interpreting the influence of internal variables on sensory activity. *Current Opinion in Neurobiology* 46, 84–89 (2017).

[73] Lazar, A. *et al.* Paying attention to natural scenes in area V1. *bioRxiv* (2023). URL `https://www.biorxiv.org/content/early/2023/08/26/2023.03.21.533636`.

[74] Paulun, V. C., Zheng, K. & Kar, K. Distributed population activity in the macaque inferior temporal cortex but not current deep neural networks predict the ponzo illusion. *Journal of Vision* 22, 3354–3354 (2022).

[75] Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples (2015). URL `https://doi.org/10.48550/arXiv.1412.6572`.

[76] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience* 22, 974–983 (2019).

[77] Młynarski, W., Hledík, M., Sokolowski, T. R. & Tkačik, G. Statistical analysis and optimality of neural systems. *Neuron* 109, 1227–1241 (2021).

[78] Hénaff, O. J., Goris, R. L. & Simoncelli, E. P. Perceptual straightening of natural videos. *Nature Neuroscience* 22, 984–991 (2019).

Methods

Generative models. We study Markovian two-layer hierarchical latent variable generative models that learn the joint distribution of observations and latent variables in the form

$$p_\theta(\text{image}, \text{V1}, \text{V2}) = p_\theta(\text{image} \,|\, \text{V1}) \cdot p_\theta(\text{V1} \,|\, \text{V2}) \cdot p_\theta(\text{V2}) \tag{2}$$

where image denotes the pixel intensities of the modelled images, V1 and V2 denote stochastic model layers modeling biological neurons in visual cortical areas V1 and V2, respectively, and θ denotes the parameters of the generative model. To make inference tractable, we train Variational Autoencoders (VAEs) [25, 79] which supplement the above generative model with a separate recognition model $q_\Phi(\text{V1}, \text{V2} \,|\, \text{image})$ with parameters Φ that establishes a variational approximation of the true posterior distribution $p_\theta(\text{V1}, \text{V2} \,|\, \text{image})$.

In general, the tractable Expectation Lower Bound (ELBO) for a two-layer VAE can be written as

$$\begin{aligned} \mathcal{F}(\text{image}, \theta, \Phi) &= \mathbb{E}_{q_\Phi(\text{V1},\text{V2} \,|\, \text{image})} \left[\log p_\theta(\text{image} \,|\, \text{V1})\right] \\ &- \text{KL}\left[q_\Phi(\text{V1}, \text{V2} \,|\, \text{image}) \,||\, p_\theta(\text{V1}, \text{V2})\right] \end{aligned} \tag{3}$$

where $\text{KL}\left[\cdot \,||\, \cdot\right]$ denotes the Kullback–Leibler divergence of two probability distributions.

In our main model, TDVAE, we factorize the variational posterior in the following way:

$$q_\Phi(\text{V1}, \text{V2} \,|\, \text{image})^{\text{TDVAE}} = q_\Phi(\text{V1} \,|\, \text{image}, \text{V2}) \cdot q_\Phi(\text{V2} \,|\, \text{image}). \tag{4}$$

This leads to the recognition model architecture depicted in Fig. 1f and the following objective function (ELBO):

$$\begin{aligned} \mathcal{F}_{\text{TDVAE}}(\text{image}, \theta, \Phi) &= \mathbb{E}_{q_\Phi(\text{V1} \,|\, \text{image},\text{V2}) \cdot q_\Phi(\text{V2} \,|\, \text{image})} \left[\log p_\theta(\text{image} \,|\, \text{V1})\right] \\ &- \mathbb{E}_{q_\Phi(\text{V2} \,|\, \text{image})} \left[\text{KL}[q_\Phi(\text{V1} \,|\, \text{image}, \text{V2}) \,||\, p_\theta(\text{V1} \,|\, \text{V2})]\right] \\ &- \text{KL}\left[q_\Phi(\text{V2} \,|\, \text{image}) \,||\, p_\theta(\text{V2})\right]. \end{aligned} \tag{5}$$

In our alternative model ffVAE the top-down interaction is not manifested in the factorization of the variational posterior:

$$q_\Phi(\text{V1}, \text{V2} \,|\, \text{image})^{\text{ffVAE}} = q_\Phi(\text{V2} \,|\, \text{V1}) \cdot q_\Phi(\text{V1} \,|\, \text{image}). \tag{6}$$

This results in the following objective function:

$$\begin{aligned} \mathcal{F}_{\text{ffVAE}}(\text{image}, \theta, \Phi) &= \mathbb{E}_{q_\Phi(\text{V1} \,|\, \text{image})} \left[\log p_\theta(\text{image} \,|\, \text{V1}) + \log p_\theta(\text{V1} \,|\, \text{V2})\right] \\ &+ \text{H}\left[q_\Phi(\text{V1} \,|\, \text{image})\right] \\ &- \mathbb{E}_{q_\Phi(\text{V1} \,|\, \text{image})} \left[\text{KL}[q_\Phi(\text{V2} \,|\, \text{V1}) \,||\, p_\theta(\text{V2})]\right] \end{aligned} \tag{7}$$

where $\text{H}\left[\cdot\right]$ denotes the entropy of a probability distribution.

For baseline we also trained a non-hierarchical VAE which we call shallow-VAE and has the objective function

$$\begin{aligned} \mathcal{F}_{\text{shallow-VAE}}(\text{image}, \theta, \Phi) &= \mathbb{E}_{q_\Phi(\text{V1} \,|\, \text{image})} \left[\log p_\theta(\text{image} \,|\, \text{V1})\right] \\ &- \text{KL}\left[q_\Phi(\text{V1} \,|\, \text{image}) \,||\, p_\theta(\text{V1})\right]. \end{aligned} \tag{8}$$

Architectural details. Inspired by models of V1 activity [30], we chose to implement the V1 layer of latent variables as an overcomplete layer ($\dim(\text{V1}) = 1800 > \dim(\text{image}) = 40^2 = 1600$) with a sparse (Laplace distribution) prior that is in a linear generative relationship to observations: $p_\theta(\text{image} \,|\, \text{V1}) = \mathcal{N}(\text{image}; \mathbf{A} \cdot \text{V1} + \mathbf{b}, \sigma^2_{\text{obs}} \cdot \mathbf{I})$ where σ_{obs} denotes the standard

deviation of the independent Gaussian observation noise for which we used $\sigma_{\text{obs}} = 0.4$ in each presented model. In addition, we chose $\dim(\text{V2}) = 250$ for the hierarchical models TDVAE and ffVAE. While this may seem like an artificially induced compression constraint compared to the 1800 units in V1, you will see in Extended Data Table 3 that 250 dimensions were enough to accommodate the entirety of the learned representations in V2.

The components of the generative and recognition models were parameterized through neural networks with the softplus activation function. We used fully connected networks (multilayer perceptrons, MLPs) instead of convolutional networks (CNNs) to avoid the indirect effects of CNNs on emerging representations. Extended Data Table 1 highlights the number of hidden units for each part of the computational graph in the shallow-VAE and the ffVAE models.

The computational graph of the recognition model for the TDVAE model is nontrivial and contains further inductive biases. There are four MLPs defined for the recognition model as depicted in Fig. 1f. The first neural network V1-ff maps the pixel space to a layer L_x that is shared between the computations of $q_\Phi(\text{V2}\,|\,\text{image})$ and $q_\Phi(\text{V1}\,|\,\text{image}, \text{V2})$. From L_x, a second MLP (V2-ff) computes the means and variances of the $q_\Phi(\text{V2}\,|\,\text{image})$ distribution. The third MLP, V1-TD transforms V2 into a layer L_z. We fuse the information from image and V2 by concatenating L_x and L_z and applying an MLP V1-INT to the concatenated layer to obtain the means and variances of $q_\Phi(\text{V1}\,|\,\text{image}, \text{V2})$. The number of hidden layers and hidden units used in each MLP to calculate the means and standard deviations of the conditional generative and variational posterior distributions is shown in Extended Data Table 2 for the different TDVAE models referenced in the paper.

Model training details. We trained all our models with the Adam optimizer on natural image patches with a minibatch size of 128. We found that while the learned V1 representation was robust against the tested regularization techniques (weight decay, gradient clipping [80], gradient skipping [28]), the learned V2 representation was sensitive to them. To eliminate such regularization artifacts, we turned off weight decay and increased gradient clipping and skipping thresholds to have an activation frequency below 10^{-6}.

For training our hierarchical VAE models TDVAE and ffVAE we applied β annealing, which is a standard technique for training VAEs. This means that in their ELBO losses in equations (5) and (7) we introduced a scaling factor β in front of the last term. We started training with $\beta = 0.1$ for 1000 epochs and then linearly increased β to 1 over the course of 1800 epochs and then continued training with the original ELBO loss. No β annealing was used for the shallow-VAE model.

We found that our two-layer models TDVAE and ffVAE achieved significantly better ELBO values if we changed the learning rate parameter α whenever the validation loss reached a minimum. The following learning rate schedules were found to lead to the highest ELBO values (these are the models reported throughout this paper):

shallow-VAE $\alpha = 3 \cdot 10^{-5}$ for 12,000 epochs

TDVAE linear integration $\alpha = 10^{-5}$ for 40,500 epochs, then $2 \cdot 10^{-5}$ for 2100 epochs, and $2.85 \cdot 10^{-5}$ for 630 epochs

ffVAE shallow nonlinearity $\alpha = 10^{-5}$ for 7400 epochs, then $2 \cdot 10^{-5}$ for 600 epochs, $3 \cdot 10^{-5}$ for 320 epochs, and 10^{-6} for 10 epochs

TDVAE shallow nonlinearity $\alpha = 5 \cdot 10^{-5}$ for 4500 epochs, then $2.5 \cdot 10^{-5}$ for 9000 epochs

TDVAE deep nonlinearity $\alpha = 10^{-5}$ for 205,700 epochs, then $2 \cdot 10^{-5}$ for 2100 epochs, $3 \cdot 10^{-5}$ for 5700 epochs, and $5 \cdot 10^{-5}$ for 900 epochs

The models were trained on an Nvidia GeForce RTX 3080 Ti GPU.

Validation ELBOs at the end of training and the number of active units in the learned representations in V1 and V2 are presented in Extended Data Table 3. Of all presented models, TDVAE with a deep nonlinearity in the higher level of its generative model $p_\theta(\mathrm{V1} \,|\, \mathrm{V2})$ reached the highest ELBO value, therefore it is the primary interest of this paper. For all models, V1 learned a complete linear basis of the training data (see below on the effective dimensionality of the natural training data) while the nonlinear representations learned by V2 had different dimensionalities for different models.

Feed-forward discriminative model. We also trained a feed-forward discriminative model on texture images. The task of the network was to classify texture images into families. We wished to mimic the architecture of the generative counterpart of this model, ffVAE. Therefore, the first two layers contained 2000 hidden units, and the third layer (which is analogous to the V1 latent in ffVAE) contained 1800 hidden units. We then had three more layers with 1000, 500 and 250 units, respectively. On the top of the last layer (which is analogous to the V2 layer in ffVAE) we put a softmax layer with 5 outputs (corresponding to the texture families used throughout this paper). After training the model we found that it classifies texture images very well, but we could not identify localized filters in the third MLP layer as in the generative models. (The receptive fields for this model were determined by the same reverse correlation method as for our generative models.) Therefore, this model is not a suitable model for biological vision.

In a second step, we trained the model using pre-trained weights from the shallow-VAE recognition model for the first three layers. Therefore, this model had localized, oriented filters by construction. We found that training the remaining layers was sufficient to obtain an efficient classifier with 99% classification accuracy.

Natural image datasets. All presented models were trained on a 'training' natural image dataset compiled by the authors in a way similar to [32]. We took 320,000 random 40×40-pixel crops from the first 1000 images of the van Hateren natural image dataset [81], matched their grand total pixel intensity histogram to the standard normal distribution, and applied ZCA whitening while keeping only $25\pi\%$ of the largest-eigenvalue PCA components. This made the effective dataset dimension 1256 instead of the number of pixels in an image patch, 1600. This whitening procedure is motivated by its resemblance to the spatial-frequency response characteristic of retinal ganglion cells [32, 82], and also helps gradient based learning by uniforming variances along the main axes of variation. A statistically independent 'validation' set of 64,000 whitened natural images was created in the same way and used to monitor and report validation losses during model training (Fig. 2a). Another statistically independent 'test' set of 64,000 whitened natural images was also created in an identical way and was used throughout the paper for model evaluation.

Texture image datasets. Our texture image datasets, created solely for model evaluation, were synthesized from five texture family seed images with the algorithm in [42]. But what kinds of starting images could be useful for probing TDVAE? Computational constraints limited the training image patch sizes to 40×40 pixels; therefore, we selected only starting images for which the dominant wavelength of synthesized textures fits into a 40×40 patch. We also wanted to ensure that the selected starting images are well described by localized, oriented linear filters; therefore, we set the additional criterion on starting images that a sparse one-layer VAE with a linear generative model learns localized, oriented linear filters from 40×40 patches randomly sampled from textures synthesized from it. We then randomly cropped a balanced set of 40×40-pixel texture patches from the synthesized texture images, matched the grand total pixel histogram to the standard normal distribution for each texture family separately, and applied the whitening procedure used for natural images to the whole texture

dataset. Fig. 2c shows sample patches from five texture families fulfilling these criteria. Using this procedure, we created a balanced 'train' texture dataset of 320,000 patches in total and a statistically independent balanced 'test' texture dataset of 64,000 patches in total and used these for the texture family decoding experiments in Figs. 2d and 5d. In the rest of the paper, only the latter, 'test' texture dataset was used for model evaluation. (For more details, see [23].)

Phase scrambling. Phase scrambling (Figs. 3g and 5f) was performed on whitened images in the way described in [17]: we took the Fourier transform of the image, randomized the phase of each Fourier component, and took the real part of the inverse Fourier transform of the result. This manipulation did not change the spectral properties of the input images.

Filter scrambling. Filter scrambling (Fig. 5f) was also performed on whitened images, in a way that is equivalent to low-level-synthetic (LL-synthetic) image generation in [46], to which it is compared in Fig. 5g. First, we transformed the image to the complete basis of normalized V1 receptive fields of the shallow-VAE model, randomly permuted filter activations, and transformed the result back with the normalized V1 projective fields of the shallow-VAE model. Finally, we rescaled all filter scrambled images with a common factor which restored the mean image variance of the starting images. This manipulation did not preserve the spectral properties of the input images.

Comparison texture image dataset. Figs. 2e, 3c, 3e, and 3g compare TDVAE model predictions to experiments in [17] and [18]. These two papers are built on the same electrophysiology data set, that we tried to mimic as described here. The papers present spike counts in a certain time window for 102 individual V1 and 103 individual V2 neurons found in 13 macaque monkeys as a response to the same texture stimulus sequence with 15 texture families, 15 samples plus 15 phase scrambled samples from each texture family, and 20 presentations per image. In our comparison dataset, we randomly sampled 100 active V1 dimensions as a model of the 102 V1 neurons in the experiment. The TDVAE model only has 6 active dimensions in V2 which is less than the 103 experimental V2 neurons. As a substitute, we generated 100 random rotations of the 6D space and took the first components of the rotated 6D active V2 space as a model of 100 V2 neurons. (Note that this procedure is consistent with the rotation-invariant unit normal prior in the V2 layer of the TDVAE model.) To avoid correlated responses which are clearly not present in the experimental dataset, we sampled the stimulus-conditional model posteriors separately for all 100 V1 and 100 V2 model neurons, respectively. For stimuli, we used test texture images from the 5 texture families used throughout this paper (this is three times less than the 15 texture families in the experiments), 15 samples plus 15 phase scrambled samples from each texture family, and 20 presentations per image. We named this dataset the 'comparison' dataset.

Illusory stimuli. Stimuli used for illusory contour experiments (Fig. 4) were based on stimuli used in the experimental paper [43] (see Fig. 4c). To compensate for the similarity of patch edge length to V1 receptive field sizes in the TDVAE model (Fig. 4a), we used only a subset of model V1 units the receptive fields of which fulfilled each one of four criteria: 1) centrality: center of mass of the pointwise square of the receptive field falls in the central $(40\%)^2$ of the image patch, 2) locality: size σ of the filter, calculated by fitting a cylindrical Gaussian profile with scale σ to the pointwise absolute value of the V1 receptive field by requiring the two to have the same moment of inertia, falls between 18% and 33% of the patch edge length, 3) middle-wavelength: the dominant wavelength, calculated from the maximum position of the squared magnitude of the 2D Fourier transform of the receptive field, falls between 17% and 28% of the patch edge length, and 4) fitting into patch: the line segment connecting the two pac-man centers in the 'Illusory' stimulus aligned to the receptive field (based on the above definitions of

center of mass and dominant wave vector), extended by 11.1% in both directions, fits into the patch. 37 filters fulfilled these criteria in the TDVAE model and 39 in the shallow-VAE model. Upright and upside-down stimuli were both used, hence the 74 and 78 V1 units for TDVAE and shallow-VAE in Fig. 4, respectively. Receptive-field-fitted 'Line', 'Illusory', and 'Rotated' stimulus sets were first generated in black (-1) and white (1) with 16×16 grid supersampling, then downsampled in a bilinear way, filtered with Gaussian blur, Z scored, and whitened with the PCA basis obtained from natural image whitening above. Whitening with the PCA basis of natural images instead of the illusory contour PCA basis was chosen because the effective dimensionality of the illusory contour stimulus set is much lower than the 1256 dimensions required by the whitening procedure used throughout this paper. Stimulus parameters used in Fig. 4: pac-man radius: 6 pixels; Gaussian blur standard deviations in panels k, l, and m: 0.9 and 1 pixels; Gaussian blur standard deviations in all other panels: 0.9 pixels; step size of stimulus displacement: 0.12 pixels.

Linear characterization of model units. Receptive fields of model units (Figs. 2b, 4a, 5c, Extended Data Figs. 5, 6, 7) were calculated with a method customary in visual neuroscience, reverse correlation [40], using 128,000 white noise images where each individual pixel had an independent Gaussian distribution with position and scale matching the grand total mean and standard deviation of the pixels in the 'test' natural image dataset. Projective fields (Extended Data Figs. 5, 7) were calculated with latent traversal along each model unit in V1 and V2 by taking two points symmetrically to 0 that differed only along the given unit dimension, generating a large number (V1: 128, V2: 2560) of mean images from these two points and taking the difference between the mean images for the two points.

Identifying active model units. In a given model, we classified V1 and V2 model units into active and inactive categories based on four criteria (Extended Data Fig. 5). (1) Variance of the mean response to the images in the 'test' natural image dataset was found to be orders of magnitude larger for active units than for inactive units. (2) Mean of the response variance (uncertainty) over the same image dataset also creates two clean clusters for active and inactive units. (3) Mean squared pixel intensities of receptive fields are clearly larger for active units than for inactive units. (4) Mean squared pixel intensities of projective fields are also larger for active units than for inactive units. We found that these four criteria consistently cluster units into the same active and inactive sets (for TDVAE, see Extended Data Fig. 5). The number of active V1 and V2 units is denoted in Extended Data Table 3 for all discussed models. In the paper we only consider active model units.

Equivalence of V1 receptive fields in single-layer and hierarchical models. Learned V1 receptive fields are localized, oriented and bandpass in both the hierarchical TDVAE model and the single-layer shallow-VAE model (Extended Data Figs. 6, 7). We calculated two quantities for each active V1 unit for a more detailed comparison: filter center (estimated as the center of mass of the pointwise squared receptive field) and dominant wave vector (calculated from the maximum position of the squared magnitude of the Fourier transform of the receptive field). To increase the precision of these estimates, we used projective fields instead of receptive fields as they are very similar but less noisy (Extended Data Fig. 7). We found that the empirical distribution of filter centers and dominant wave vectors are very similar in the TDVAE and shallow-VAE models, making them equivalent complete bases in the space of natural training images, at least in terms of first-order probability densities.

Multiunit texture family decoding. Linear decoding accuracies of texture family in Figs. 2d, 5d were calculated with multinomial logistic regression. Each bar reports the mean and standard deviation of test classification accuracies calculated from $n = 5$ fitting sessions started

from different random seeds. In each case, the dependent variable was texture family and independent variables were derived from the 'train' and 'test' texture datasets as follows. *Pixels*: independent variables were image pixel intensities. *V1, shallow-VAE*; *V1, TDVAE* (also: *V1 means*) and *V2, TDVAE* (also: *V2 means*): independent variables were mean responses of all active units in the denoted layer of the denoted model to 'train' and 'test' texture images. *V1 corrs* (*original textures*, *phase scrambled*, *filter scrambled*): independent variables were noise correlation coefficients between the same 100 randomly selected active V1 units, calculated from 16384 presentations of every 10th image in the original, phase scrambled and filter scrambled texture image datasets, respectively.

Robust texture representation in the V2 model layer. The multiunit texture family decoding test was repeated for mean V1 and V2 responses in all models presented in the paper. As it is shown in Extended Data Table 4, decoding accuracies from V1 were around chance level 0.2 for all models (we have five texture families) and decoding accuracies from V2 were close to perfect performance 1 for all hierarchical models, implying no texture representation in V1 and a robust texture representation in V2 for all studied models. Repeating the same analysis for subsets of active V2 units showed that texture family is only decodable from some of the V2 dimensions and not from others in TDVAE models (last two columns of Extended Data Table 4), revealing a low-dimensional (4–6 dimensions) learned texture representation in TDVAE models. In contrast, we found no axis-alignment in the V2 texture representation in the ffVAE model as texture family could be more or less decoded from all active V2 units (Extended Data Table 4). Since our primary model, TDVAE with deep nonlinearity, contains no non-texture-family-coding V2 units, we do not analyze such units in the present paper.

t-SNE visualization. In Fig. 2e we wanted to compare the low dimensional visualization of model responses in the TDVAE model with the low dimensional visualization of neural responses in [18] (Fig. 2f). To this end, we used the 'comparison' dataset defined above and the t-SNE visualization method utilized in [18]. t-distributed stochastic neighbor embedding (t-SNE) [83] attempts to minimize the divergence between the distributions of neighbor probability in the original high dimensional data space and the visualized low dimensional space. Similar to [18], we input a set of 75 data vectors (5 texture families, 15 images per family) to the t-SNE algorithm, each of which collected the mean responses of 100 neurons in a model layer to a stimulus (note, however, that experimental data have 15 texture families instead of 5). Like [18], we normalized the data so that, for each model unit, responses to the 75 images had a mean of 0 and standard deviation of 1, and executed the t-SNE analysis with a perplexity value of 30.

ANOVA. We also repeated the ANOVA study in [18] in the context of the TDVAE model (Figs. 3c and 3d). The utilized nested ANOVA method partitioned the total variance of V1 and V2 model unit responses to the 'comparison' dataset into portions arising across families, across samples within a family, and across repetitions of the same stimulus. The ANOVA generates an F-statistic that captures the ratio of variances between each hierarchical level. For all model neurons we found the F-statistic to be significant for ratios of variance across repetitions and across samples as well as for ratios of variance across samples and across families. Following [18] we performed further analysis using the ratio between partitioned variance components. To obtain the variance ratio, we divided the percent variance across families by the percent variance across samples. To avoid outlying values when either variance was very low, we stabilized the ratio by adding 0.6% variance to both the numerator and denominator for V1 and 6% for V2.

Classification decoding. We also reproduced the decoding study in [18] with the TDVAE model (Figs. 3e and 3f; differences from [18] in parentheses). We used Gaussian decoders

(experiment: Poisson decoders) to classify model responses to the 'comparison' dataset into 15 sample categories per family and, independently, into five texture family categories. On each iteration, we randomly selected a number of units from our model neurons in the V1 or V2 model layer. To compute performance in the sample classification task, we estimated the mean and variance of the response of each selected neuron for each of the 15 samples within each family by computing the sample mean and sample variance over four randomly selected repetitions out of the 20 repetitions in the dataset (experiment used 10 repetitions but that made the task trivial for the model). For the held-out 16 repetitions of each sample, we computed which of the 15 samples was most likely to have produced the population response, assuming independent normal distributions for the model units under the estimated mean and variance. We computed the average performance (% correct) over all samples and families, and repeated this process 100 times to get a performance for each population size (experiment used 10,000 here but estimated errors for the model were already small at 100 repetitions). To compute performance in the family classification task, we estimated the mean and variance of the response for each family over 8 randomly selected samples from the 15 different samples and for all repetitions. For each of the repetitions of the held-out seven samples, we computed which of the five families was most likely to have produced the population response. We computed the average performance over all repetitions and repeated this process 100 times to get a performance for each population size (experiment used 10,000 here but estimated errors for the model were already small at 100 repetitions). We computed performance measures for both tasks using population sizes of 1, 3, 10, 30, and 100 neurons.

Phase scrambling induced modulation of response magnitudes in V1 and V2. We studied how reducing high-level stimulus structure affects model responses by reproducing the electrophysiology experiment in [17] (see Figs. 3g and 3h). First, as the experiment queries changes in response magnitudes, we took the absolute value of all model responses to undisturbed and phase scrambled texture images in the 'comparison' dataset. Following [17], we then calculated per-neuron per-family mean response magnitudes by averaging these absolute responses over per-family samples and stimulus presentations. Next we calculated modulation indices for each 'comparison' V1 and V2 unit and texture family by dividing the difference between the above mean response magnitudes to undisturbed and phase scrambled textures with their sum. As the final step, we averaged these per-neuron modulation indices over the five texture families and plotted separate histograms for V1 and V2 units (Fig. 3g).

Model responses to illusory stimuli. One important goal of modeling illusory contour responses in Fig. 4 was to faithfully compare them with Fig. 3a in the experimental paper [43] (reproduced as Fig. 4c). This Figure in [43] presents mean spike counts in a 50 ms time window, a response averaging procedure we modeled in the following way. Experimentally, autocorrelations of membrane potentials for static stimuli typically decay on a ~20 ms timescale in the V1 cortical area [84]. Therefore, in the context of Fig. 4, we modeled 50 ms response averaging in [43] by averaging three independent samples from V1 model response distributions. Autocorrelation times in higher cortical areas are longer [85, 86, 87, 88], which we modeled by taking the three V1 samples using the very same sample from V2 model response distributions. For consistency, all model responses in Fig. 4 were calculated in this way.

Noise and signal correlations. We calculated noise correlation matrices by recording the mean responses of a number of V1 units to the repeated presentation of the same stimulus and then calculating the Pearson product-moment correlation coefficients between each distinct pair of V1 units over stimulus presentations. We calculated the signal correlation matrix of a texture family by selecting a number of samples from the family, recording the means of the mean responses of a number of V1 units to the repeated presentation of each sample, and

calculating the Pearson product-moment correlation coefficients between each distinct pair of V1 units over the selected samples. Since small patch sizes may have an adverse effect on quantities calculated from two distinct V1 units, we only used those V1 units for calculating correlations that had receptive fields fulfilling the following three criteria: 1) centrality: center of mass of the pointwise square of the receptive field falls in the central $(40\%)^2$ of the image patch, 2) locality: size σ of the filter, calculated by fitting a cylindrical Gaussian profile with scale σ to the pointwise absolute value of the V1 receptive field by requiring the two to have the same moment of inertia, falls between 18% and 33% of the patch edge length, 3) middle-wavelength: the dominant wavelength, calculated from the maximum position of the squared magnitude of the 2D Fourier transform of the receptive field, falls between 17% and 28% of the patch edge length. This filtering procedure selected 40 central, localized, medium wavelength V1 filters in the TDVAE model. On a final note, we defined the dissimilarity of two correlation matrices as the mean absolute difference between their elements which is the definition also used in [46].

References

[79] Kingma, D. P. & Welling, M. An introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning* 12, 307–392 (2019). URL `http://dx.doi.org/10.1561/2200000056`.

[80] Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. In Dasgupta, S. & McAllester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, 1310–1318 (PMLR, Atlanta, Georgia, USA, 2013). URL `https://proceedings.mlr.press/v28/pascanu13.html`.

[81] Van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265, 359–366 (1998).

[82] Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37, 3311–3325 (1997). URL `https://www.sciencedirect.com/science/article/pii/S0042698997001697`.

[83] van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008).

[84] Azouz, R. & Gray, C. M. Cellular mechanisms contributing to response variability of cortical neurons in vivo. *Journal of Neuroscience* 19, 2209–2223 (1999).

[85] Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience* 17, 1661–1663 (2014).

[86] Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* 548, 92–96 (2017).

[87] Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* 592, 86–92 (2021).

[88] Brunec, I. K. & Momennejad, I. Predictive representations in hippocampal and prefrontal hierarchies. *Journal of Neuroscience* 42, 299–312 (2022).

Acknowledgements

The authors thank Mihály Bányai and Dávid G. Nagy for early discussions about the model. G.O. was supported by a grant from the Human Frontiers Science Program, and by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

Author contributions

F.C., B.M, and G.O. conceptualized the study, F.C. performed data curation, F.C. and B.M. wrote simulation software, F.C. performed the simulations and performed the analyses, F.C., and G.O. were designing the visualization, C.F., G.O., and B.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Extended Data Table 1. Architecture of models with bottom-up recognition. Number of hidden units in each MLP layer computing the mean and the standard deviation of each conditional distribution in the bottom-up recognition models.

Model name	$p_\theta(\text{V1} \mid \text{V2})$	$q_\Phi(\text{V1} \mid \text{image})$	$q_\Phi(\text{V2} \mid \text{V1})$
shallow-VAE	N/A	(2000, 2000)	N/A
ffVAE	(2000)	(2000, 2000)	(1000, 500, 250)

Extended Data Table 2. Architecture of models with top-down recognition. Number of hidden units in each MLP layer computing the mean and the standard deviation of each conditional distribution in TDVAE models.

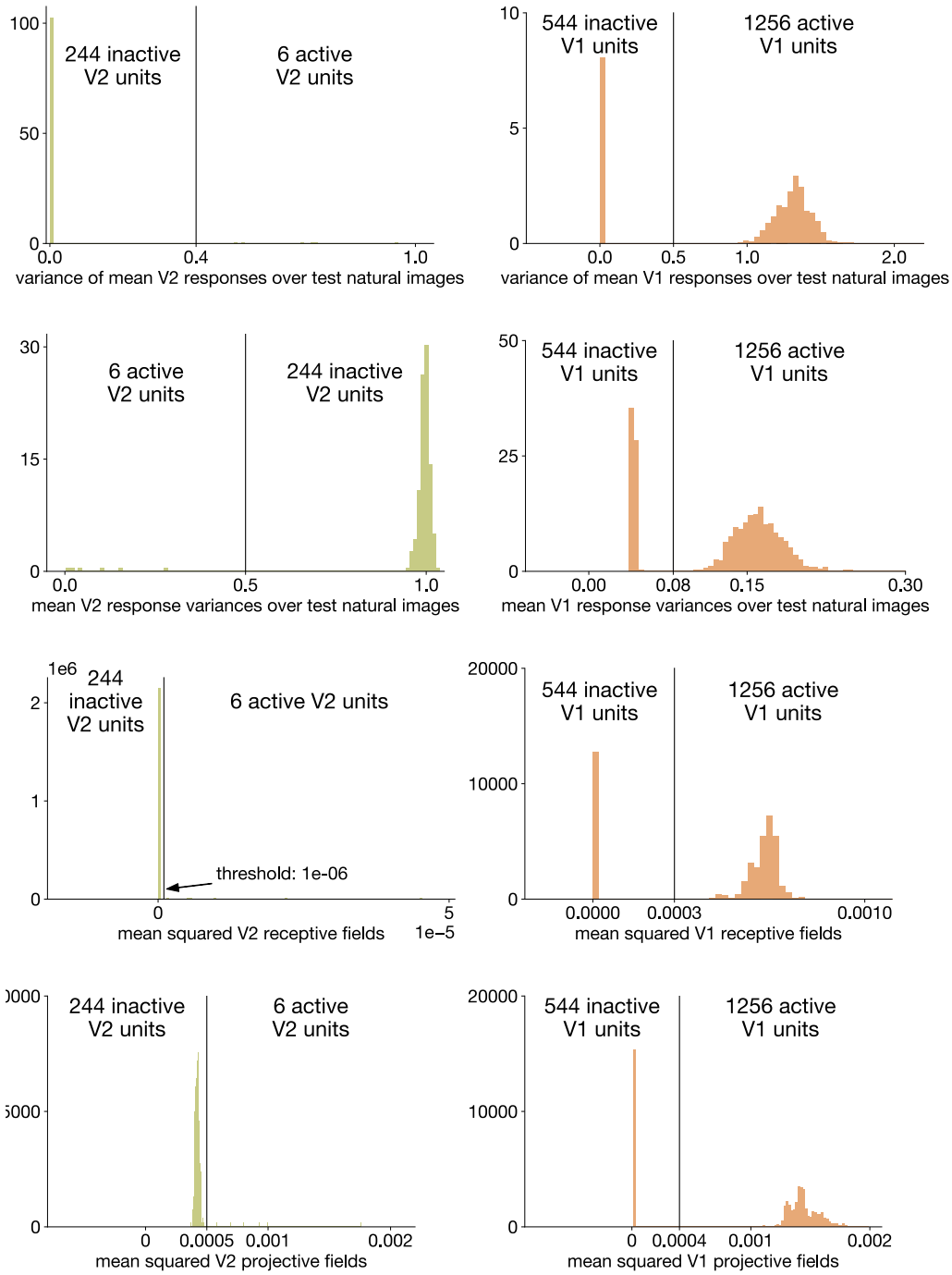
Model name	$p_\theta(\text{V1} \mid \text{V2})$	V1-ff	V2-ff	V1-TD	V1-INT
TDVAE linear integration	(1000, 2000)	(2000)	(1000, 500, 250)	(250, 500, 1000, 2000)	()
TDVAE shallow nonlinearity	(2000)	(2000)	(1000, 500, 250)	(250, 500, 1000, 2000)	(2000)
TDVAE deep nonlinearity	(1000, 2000)	(2000)	(1000, 500, 250)	(250, 500, 1000, 2000)	(2000)

Extended Data Table 3. Basic features of the trained models. Validation ELBO at the end of training and the number of active stochastic units in the trained V1 and V2 model layers for each presented model.

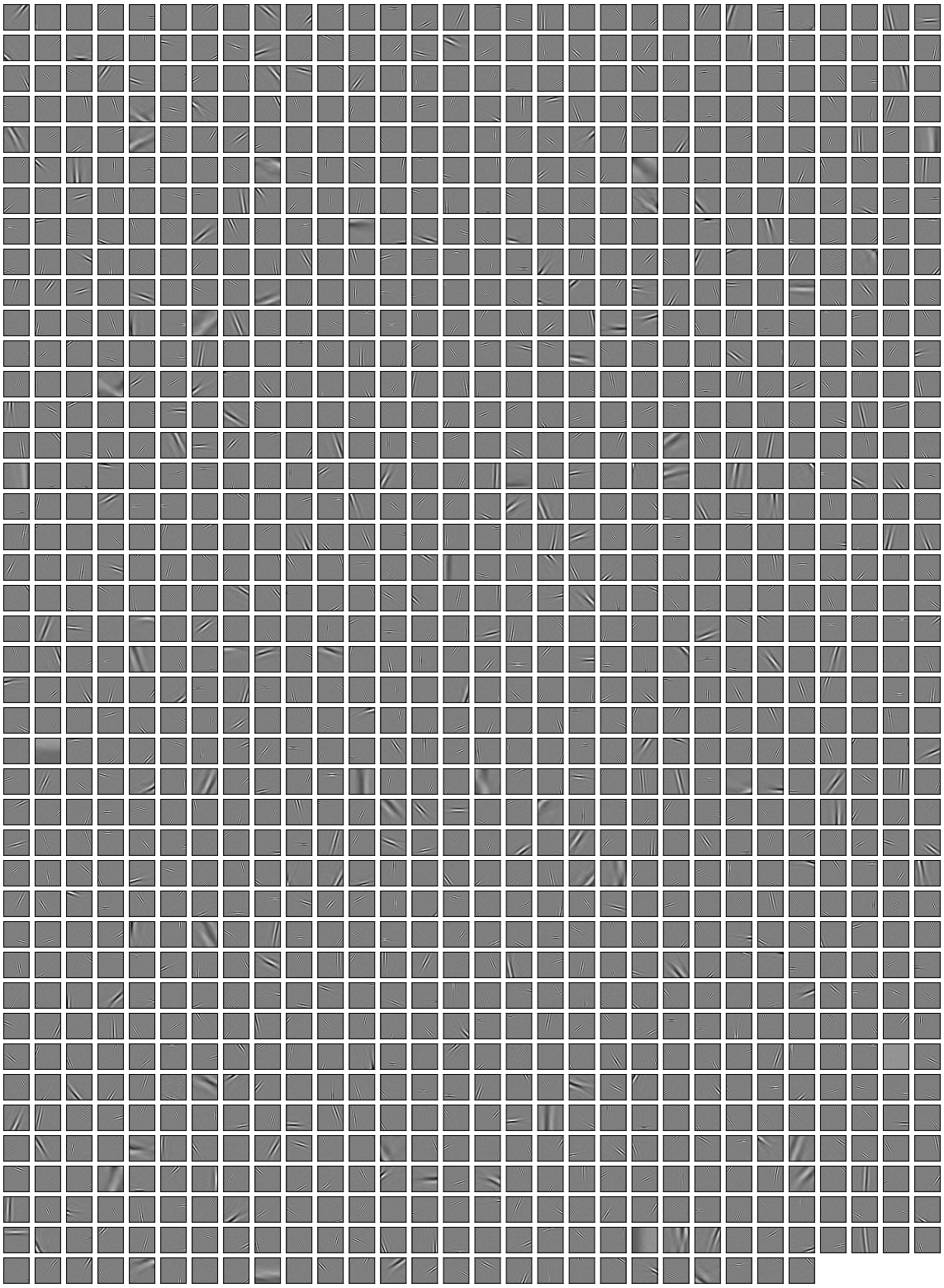
Model name	ELBO	active V1 units	active V2 units
shallow-VAE	−1916	1256	N/A
TDVAE linear integration	−1886	1256	27
ffVAE shallow nonlinearity	−1818	1256	25
TDVAE shallow nonlinearity	−1705	1256	199
TDVAE deep nonlinearity	−1699	1256	6

Extended Data Table 4. Texture representation in the trained models. Texture family decoding accuracies from mean responses of active V1 and active V2 units and the number of texture-family-coding (TF-enc) and non-texture-family-coding (non-TF-enc) active V2 units for all studied models. No clear distinction was found between texture-family-coding and non-texture-family-coding V2 units in the ffVAE model.

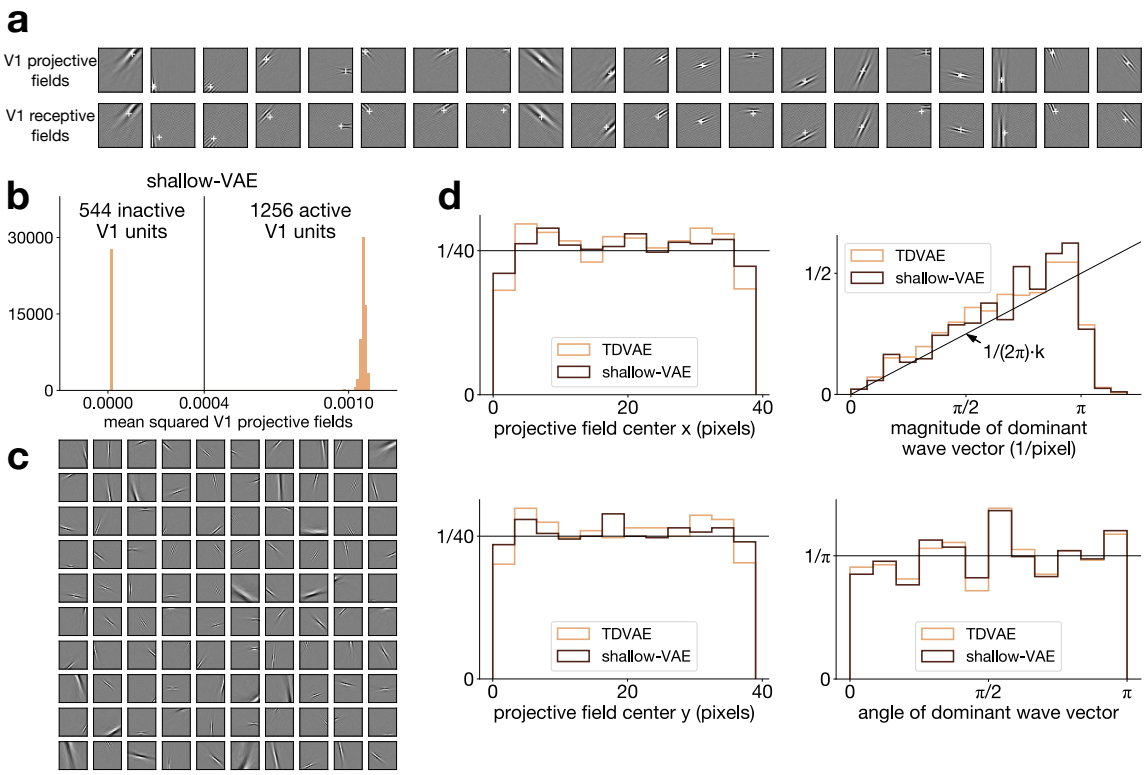
Model name	V1	V2	TF-enc	non-TF-enc
shallow-VAE	0.230	N/A	N/A	N/A
TDVAE linear integration	0.209	0.884	4	23
ffVAE shallow nonlinearity	0.220	0.958	25	
TDVAE shallow nonlinearity	0.290	0.908	5	194
TDVAE deep nonlinearity	0.347	0.925	6	0



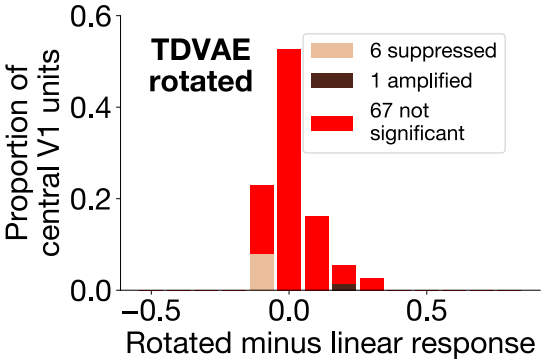
Extended Data Fig. 5. Activity of model units in the trained TDVAE model. We used four criteria for selecting active model units in the V1 and V2 model layers (Methods): (from top to bottom) stimulus dependent responses (variance of the mean response over the 'test' natural image dataset); response confidence (mean of the response variance over the 'test' natural image dataset); response to white noise (mean of the squared receptive field); influence on generated images (mean of the squared projective field). We found that all criteria consistently selected the same six active units in V2 (left) and that, in a similar way, all criteria consistently selected the same 1256 active units in V1 (right). Note that the 'training' natural image dataset has 1256 effective dimensions (Methods), meaning that the V1 layer learned a complete linear basis from and over the training data.



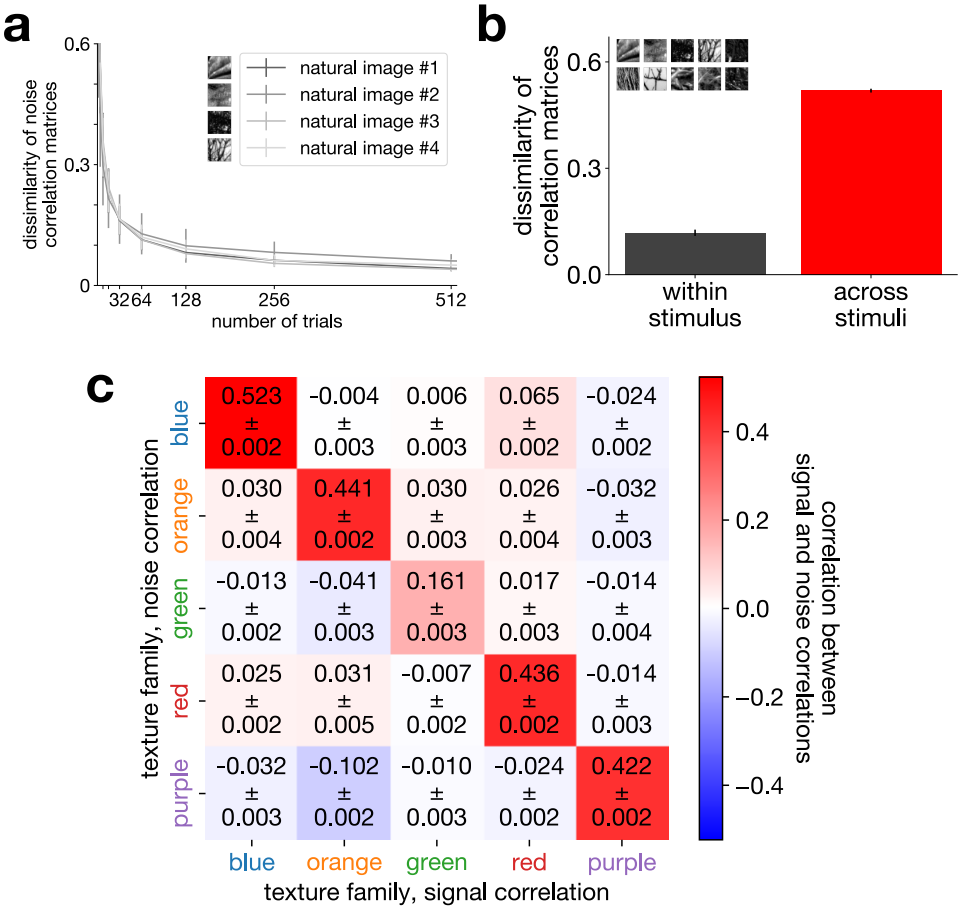
Extended Data Fig. 6. Receptive fields of all active V1 units in the TDVAE model. All V1 receptive fields are localized, oriented, and bandpass, reminiscent of V1 simple cells in primates.



Extended Data Fig. 7. Equivalence of V1 receptive fields in the TDVAE and the shallow-VAE models. a, V1 projective and receptive field centers calculated in the TDVAE model with the same procedure (Methods). Projective fields are very similar to receptive fields but also smoother, making them more reliable for center of mass estimation than receptive fields. In general, we use projective fields for estimating V1 filter parameters throughout the paper. **b**, The shallow-VAE model has 1256 active V1 units, like the TDVAE model. **c**, Example projective fields from the shallow-VAE model. These are localized, oriented, and bandpass, like those in the TDVAE model. **d**, V1 projective field centers and dominant wave vectors (Methods) homogeneously fill the patch area in the TDVAE and the shallow-VAE models, making them equivalent complete bases in the space of natural training images, at least in terms of first-order probability densities.



Extended Data Fig. 8. Differences between the mean and linear responses to the 'Rotated' stimulus. Out of the $n = 74$ V1 filters studied, 6 showed significant suppression and 1 showed significant amplification compared to the linear response (colors). See Fig. 4h for the set of V1 filters and the definition of significance.



Extended Data Fig. 9. Noise and signal correlations in the TDVAE model. a, Estimation of noise correlations improves with the number of trials they are calculated from. Solid lines: mean within-stimulus noise correlation matrix dissimilarities for four natural image examples. Error bars: s.d. In general, this paper presents noise correlation matrices calculated from at least 32 trials. **b**, Noise correlations depend on the stimulus. Two noise correlation matrices ($n = 64$ trials each) were calculated for each of 10 natural images (inset). Mean dissimilarity between noise correlation matrices on the same image (0.118, within stimulus) is significantly smaller than on different images (0.519, across stimuli; one-sided t-test: $P = 4.4 \times 10^{-41}$, t = -17.31, df = 188). Error bars: s.e.m. **c**, Noise and signal correlations are connected through the contextual prior. Pearson correlations between noise and signal correlation coefficients over distinct pairs of $n = 40$ V1 units. Correlation matrices were calculated on 5 texture families (axes), 128 samples per family and 128 trials per sample. The whole process was repeated 5 times. Colors: means. Errors: s.e.m.