Top-down generation of low-precision representations improves the perception and imagination of fine-scale visual information

Zedong Bi

Institute for Future, Qingdao University, Shandong 266071, China

Email: zedong.bi@outlook.com

Abstract

Top-down generation of neural representations is essential for both predictive-coding perception and imagination. In particular, predictive-coding perception or imagination requires top-down generated neural representations in primary visual cortex (V1) to reconstruct or simulate the V1 representations in response to the seen images. However, top-down generated representations are found to have low coding precision in both perception and imagination tasks. Why and how does the brain use top-down generated low-precision representations to reconstruct or simulate bottom-up stimulated high-precision representations? How can information of fine spatial scale be perceived or imagined using the top-down generated low-precision representations? By modeling visual system using variational auto-encoder, we reveal that training to generate low-precision representations facilitates higher-order cortex to form representations smooth to shape morphing of stimuli, thereby improving perceptive accuracy and robustness as well as imaginative creativity. Fine-scale information can be faithfully inferred from low-precision representations if fine-scale information is sparse. Our results provide fresh insights into visual perception and imagination as well as the sparseness of V1 activity, and suggest that generating low-precision representations that contain sparse fine-scale information is a strategy that the brain uses to improve the perception and imagination of fine-scale information.

Introduction

Analysis-by-synthesis is a long-standing perspective of perception, which proposes that instead of passively responding to external stimuli, the brain actively predicts and explains its sensations [1, 2]. Neural network models in this line use a generative network to reconstruct the external stimuli, and use this reconstruction error to guide the updating of the neuronal dynamic states [3] or synaptic weights [4, 5]. This analysis-by-synthesis perspective of perception has received experimental supports in both visual and auditory systems [2, 6]. Additionally, neural representations are generated in a top-down cascade during visual creative imagination [7], and these representations are highly similar to those activated during perception [8, 9]. These results above suggest that the brain performs like a top-down generator in both perception and creativity tasks.

Top-down generation is believed to be an essential basis of consciousness [10, 11]. Feedback transmission needs to be ignited for the awareness of visual stimuli [12, 13, 14]. High-level mental operations such as perceptual integration, working memory and executive control, can happen unconsciously [15, 16, 17], because unawared stimuli may be transmitted feedforwardly without triggering the feedback pathway. Global workspace theory proposes that consciousness happens when a mental model of sensory information is broadcasted to the whole brain by top-down connections, with the broadcasted mental model updated according to the mismatch between model-predicted and observed data [14]. Moreover, imagination, the generation of fictional representation using the internal model of the real world through top-down connections [18], plays an important role in prospective planning and retrospective analysis [10, 19, 20], and is key in model-based reinforcement learning [21].

However, in both perception and imagination, top-down generated representations have lower coding precision than the representations stimulated by seen images. It has been found that V1 neurons have larger receptive fields (RFs) and are tuned to lower spatial frequencies in imagined images than in seen images [22, 23]. If one agrees that imagination requires top-down generated representations to simulate the representations stimulated by seen images [8, 18, 9], one should be curious about why and how the brain uses low-precision generated representations to imagine, especially the visual information with fine spatial scale which is believed to be coded by V1 neurons that have small RFs and prefer to high spatial frequencies. During perception, the responses of V1 neurons to stimuli in classical RFs are modulated by extra-classical RFs that surround classical RFs [24]. This surround modulation, which is mediated by feedback connections [25], is believed to reflect the reconstruction of V1 activity by higher-order cortex in predictive-coding perception [3]. However, extra-classical RFs also have lower precision than classical RFs. Extra-classical RFs have larger size (i.e., lower spatial precision) than classical RFs, which means that a visual stimulus can influence more V1 neurons through feedback connections to V1. The modulation of classical RFs by extra-classical RFs is insensitive to grating phase even for simple cells, and has broader tuning to spatial frequencies than classical RFs [26]. If one agrees the predictive-coding theory of perception [3], one should be curious about why and how the brain uses the generated low-precision representations in extra-classical RFs to reconstruct the stimulated high-precision representations in classical RFs, especially for the visual information with fine spatial scale coded in the small-sized classical RFs of the phase-sensitive simple cells. Overall, the computation underlying the mismatch between the coding precision of the top-down generated and bottom-up stimulated representations remains unclear. Understanding this computation is important due to the fundamental role of perception and imagination in cognition (discussed in the previous paragraph).

In this paper, we address these problems by modeling the visual system as a variational auto-encoder (VAE) [27]. We show that training to generate low-precision representations facilitates higher-order cortex to form representations smooth to shape morphing of stimuli, thereby improving visual perceptive accuracy and robustness as well as imaginative creativity. Fine-scale information can be faithfully inferred from low-precision representations if fine-scale information is sparse. Our results suggest that generating low-precision representations that contain sparse fine-scale information is an important strategy the brain uses to improve perception and imagination of fine-scale information, and reveals the computational advantage of sparse code in V1 from top-down generative perspective.

Results

Low-precision representations can be better simulated or reconstructed through top-down pathway

Model setup

We modeled the visual system as a variational auto-encoder (VAE) [27], whose output is optimized to reconstruct the input (**Fig. 1a**). In response to any input, the bottleneck state is sampled from a Gaussian distribution whose mean and variance are determined by the output of the encoder part (**Fig. 1a**). This Gaussian distribution is encouraged to be close to standard normal distribution during training. VAE mimics the visual system in the following four aspects: (1) the bottleneck layer $\mathcal{B}$, which is low-dimensional, can be regarded as the prefrontal cortex, because experiments suggest that working memory is maintained in a low-dimensional subspace of the state of the prefrontal cortex [28, 29]; (2) the input $\mathcal{I}$ and output $\mathcal{O}$ layers can be regarded as V1, with the $\mathcal{I} \to \mathcal{B}$ and $\mathcal{B} \to \mathcal{O}$ connections (i.e., the encoder and decoder parts) representing the bottom-up and top-down pathways respectively; (3) random sampling is a designed ingredient of VAE, analogous to neuronal stochasticity; and (4) after training, VAE can generate new outputs through the top-down pathway by sampling states of the bottleneck layer from standard normal distribution, mimicking the top-down generation in mental imagination [8, 18], and perception can also be modeled through an error-updating process (see below). This fourth point makes VAE be a generative model of the visual system.

Modeling the visual system using VAE has also been suggested by other authors [10]. The fourth point above suggests that VAE provides a unified model for imagination and perception. The cost function that VAE optimizes is the variational free energy [27]. It has been proposed that minimizing variational free energy is a unified theoretical framework to understand perception and learning [1]. Modeling the visual system using VAE expands the explanation power of this framework to imagination.

We considered three toy models of V1 activity, in each of which V1 neurons are located on a 1-dimensional line. In Model 1, only a single neuron has non-zero activity, while all the other neurons are silent (**Fig. 1b, left**). In Model 2, the profile of neuronal activities is an

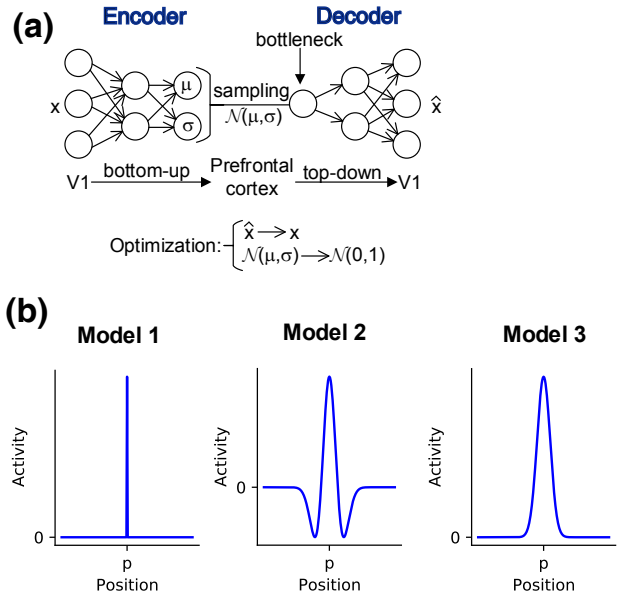

Figure 1: The models to study the generation of representations on a 1-dimensional line. (**a**) Schematic of the variational auto-encoder (VAE). The value z of the bottleneck (which models the prefrontal cortex) is sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, with μ and σ determined by the two output channels of the bottom-up pathway. This network is optimized so that the output of the top-down pathway $\hat{x}$ is close to the input stimulus x, and at the same time $\mathcal{N}(\mu, \sigma)$ is close to $\mathcal{N}(0, 1)$. (**b**) Profiles of the three models of V1 activity. The peak (Model 1), wavelet (Model 2) and bump (Model 3) are parameterized by their positions p on a 1-dimensional line.

oscillating wavelet (**Fig. 1b, middle**). In Model 3, the profile of neuronal activities is a Gaussian bump (**Fig. 1b, right**). The shapes of the three activity profiles are parameterized by their positions on a one-dimensional line. We trained VAE to generate the activity patterns of the three models, and compared the quality of the generated patterns in imagination and perception tasks.

These three models can be biologically interpreted as follows:

(1) The neuronal activities in the three models can be regarded as the responses of three types of V1 simple cells to a dot stimulus on a 1-dimensional line. The sizes of the receptive fields of the simple cells in Model 1 are close to zero, smaller than those in Model 3. By comparing the VAE-generated patterns after training VAEs using the activity patterns in Model 1 and Model 3, we can study the influence of receptive-field size to imagination and perception. Additionally, the receptive fields in Model 2 have the same spatial scale of Gaussian decay with the receptive fields in Model 3, but have higher spatial frequency. So we can study the influence of spatial-frequency preference to imagination and perception by comparing the images generated by VAEs trained by Model 2 and Model 3. Experiments have found that V1 neurons have larger receptive fields and are tuned to lower spatial frequencies in imagined images than in seen images [22, 23]. Comparing Model 3 with Models 1 and 2 brings understanding on this phenomenon.

(2) We can also interpret stimuli in the spatial-frequency domain. Neurons in Model 1 have sharp spatial-frequency tuning, whereas neurons in Model 3 have broad spatial-frequency tuning. Experiments have found that the top-down modulated extra-classical receptive fields have broader spatial-frequency tuning than the bottom-up classical receptive fields [26]. Comparing these two models brings understanding on this phenomenon.

(3) Neurons in Model 3 can be regarded as the complex cells corresponding to the simple cells in Model 1 and Model 2. So by comparing Model 1 or Model 2 with Model 3, we can also study the influence of phase sensitivity to imagination and perception. Experiments have found that extra-classical receptive fields are insensitive to graph phase even for simple cells [26], which can be understood by comparing Model 3 with Models 1 and 2.

Visual imagination

Visual imagination can be modeled by sampling the bottleneck state z of a trained VAE from standard normal distribution, so that the decoder part of VAE (**Fig. 1a**) outputs various activity patterns. To get a detailed understanding on the relationship between z and the generated pattern, we let z to be 1-dimensional, and investigated how the generated pattern changed with z. In Model 1, sharp peaks are never generated by the VAE, suggesting the failure of pattern generation (**Fig. 2a**). In Model 2, abrupt changes of wavelet positions sometimes happen with the continuous change of z (**Fig. 2b**). In Model 3, Gaussian bumps are generated, with the bump position smoothly changing with z (**Fig. 2c**). These results suggest that generating low-precision representations in V1 (i.e., the representations resulting from large receptive fields, low spatial-frequency preference and phase invariance) facilitates higher-order cortex to form representations smooth to shape morphing (which is here the translational movement of the activity profiles along the 1-dimensional line) of stimuli.

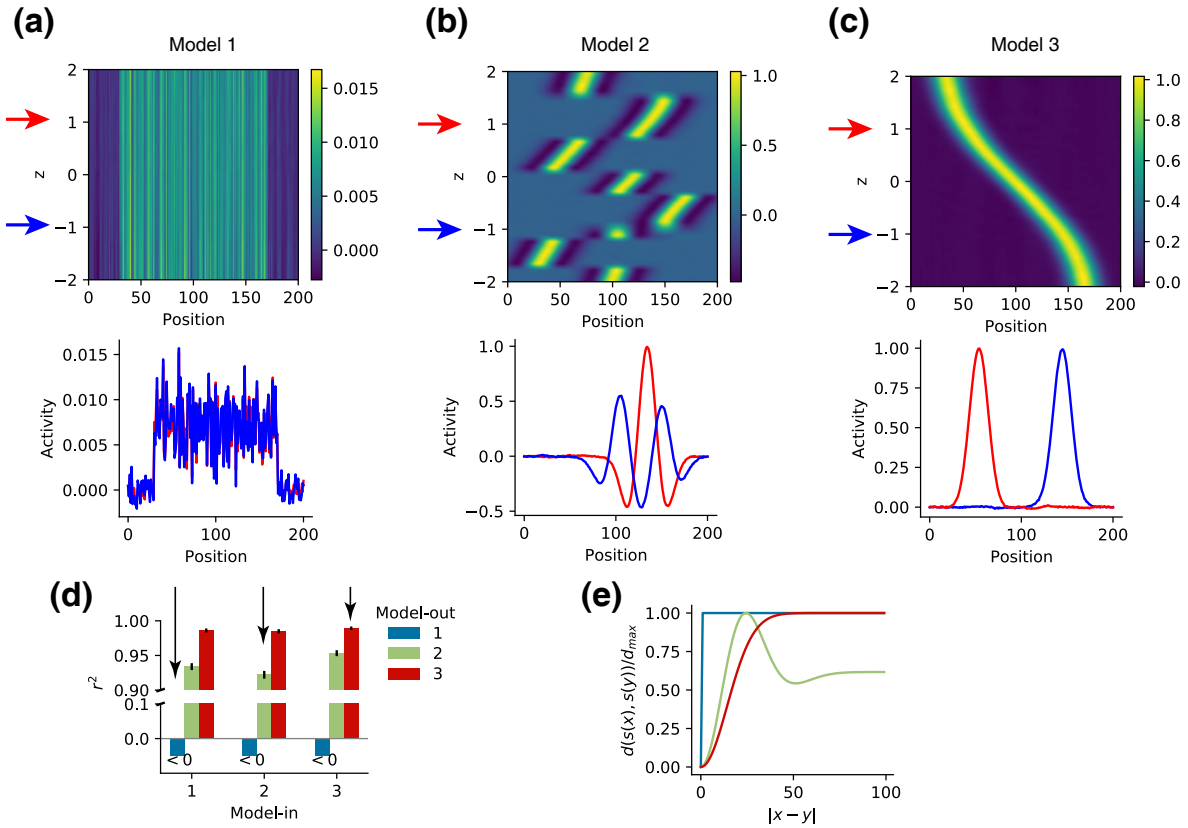

Figure 2: VAE is better at generating low-precision representations. (**a**) Upper panel: generated activity patterns as a function of the state z of the bottleneck variable. Lower panel: blue and red curves respectively represent the two generated patterns when z takes the two values (-1 and 1) indicated by the red and blue arrows in the upper panel. VAE is trained using the activity pattern of Model 1. (**b, c**) Similar to panel **a**, except that VAE is trained using the activity pattern of Model 2 (panel **b**) or 3 (panel **c**). (**d**) We let the input of VAE (i.e., x in **Fig. 1a**) be the activity pattern of Model-in, but train the output of VAE (i.e., $\hat{x}$ in **Fig. 1a**) to approach the activity pattern of Model-out. Both Model-in and Model-out can be Model 1, 2 or 3. We do not accurately show the value of r^2 when $r^2 < 0$. Arrows indicate the cases when Model-in equals Model-out, which are the cases in panels **a-c**. Error bars represent standard error of mean (s.e.m.) over 1200 samples. (**e**) Euclidean distance $d(s(x), s(y))$ between two activity patterns s as a function of the distance $|x - y|$ between the positions x and y of two dot stimuli on the 1-dimensional line, normalized by the maximal value of d, for Models 1 (blue), 2 (green) and 3 (red).

To quantify the quality of the generated patterns, we defined an index r^2 to quantify how well a pattern generated by VAE looks like an activity pattern of the model used to train the VAE (see Methods). We found r^2 is maximal for Model 3, intermediate for Model 2, but is small (even negative) for Model 1 (see the bars indicated by arrows in **Fig. 2d**). This result suggests that low-precision representations can be better simulated through top-down pathway.

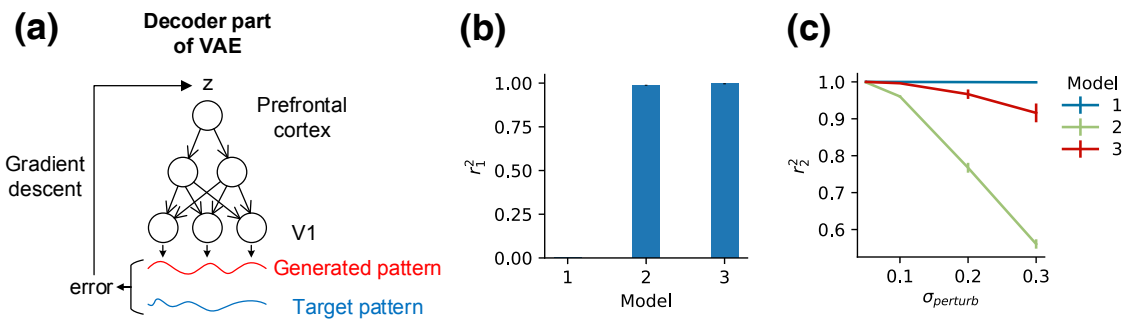

Figure 3: VAE is better at perceiving low-precision representations. (**a**) Schematic of the perception model. After training, only the decoder of VAE is used. The bottleneck state z is updated by gradient descent to minimize the error between the generated pattern (red) and the target pattern (blue). (**b**) r_1^2 for the VAEs trained by the three models (**Fig. 1b**). (**c**) r_2^2 as a function of $\sigma_{perturb}$ for the three models. Error bars (which are too small to recognize in panel **b**) represent s.e.m. over 6 VAE configurations.

The advantage of Model 3 over Model 2 is closely related to the smooth transition of z to the shape morphing (i.e., the translational movement) of the generated bumps (**Fig. 2c**), because the generation quality is bad around the abrupt change points in Model 2 (blue curve in **Fig. 2b, lower panel**).

In the study above, we input $\mathcal{P}_a$ to VAE and trained VAE to generate $\mathcal{P}_a$ in the output, with $\mathcal{P}_a$ being the activity pattern of Model a ($a = 1, 2, 3$). Now we ask whether the advantage of Model 3 results from a 'good' pattern that higher-order cortex receives from V1, or from a 'good' pattern that higher-order cortex is to generate through the top-down pathway. To answer this question, we input $\mathcal{P}_a$ to VAE but trained VAE to generate $\mathcal{P}_b$ ($b \neq a$). We found that the quality r^2 of the generated images strongly depended on $\mathcal{P}_b$, but hardly on $\mathcal{P}_a$ (**Fig. 2d**). Therefore, the advantage of low-precision representations is a top-down effect, and cannot be understood from a bottom-up perspective.

To understand the reason for the advantage of Model 3, we studied the Euclidean distance $d(s(x), s(y))$ between the activity patterns s of these three models as a function of $|x - y|$, where x and y are the positions of two dot stimuli on the 1-dimensional line respectively. In Model 1, $d(s(x), s(y))$ sharply jumps from 0 to a constant value at $|x - y| = 0$; in Model 2, $d(s(x), s(y))$ is not monotonic; and in Model 3, $d(s(x), s(y))$ monotonically and gradually increases with $|x - y|$ (**Fig. 2e**). This property of Model 3 may be important for its advantage in top-down generation. To see this, suppose two spatially nearby bottleneck states z_1 and z_2 ($z_1 \approx z_2$) generate two patterns s_1 and s_2, which corresponds to two dot stimuli at positions x_1 and x_2 respectively. For illustration, we constrain that s_1 is fixed during training, and s_2 changes within the manifold $\{s(x)\}_x$. By introducing stochastic sampling during training, VAE tries to minimize the distance between s_1 and s_2. In Model 3, the updating of s_2 can be guided by the gradient $\frac{\partial d(s_1, s_2)}{\partial s_2}$, with the monotonicity of $d(s_1, s_2)$ leading s_2 closer to s_1, so that x_2 gets closer to x_1. In this way, two nearby bottleneck states z_1 and z_2 correspond to two nearby spatial positions x_1 and x_2 respectively; in other worlds, the bottleneck state represents the spatial topology of the stimuli. In Model 2, $d(s_1, s_2)$ is not monotonic with s_2, so minimizing $d(s_1, s_2)$ does not always lead s_2 close to s_1, sometimes instead far away from s_1. In Model 1, $\frac{\partial d(s_1, s_2)}{\partial s_2}$ remains zero when $s_1 \neq s_2$, so it cannot provide clues to the updating of s_2.

Visual perception

According to predictive coding theory of perception [3], higher-order cortex adjusts its state using the error of its recontruction of the activity of lower-order cortex. In our model, this is performed by adjusting the bottleneck state z to minimize the error between the pattern generated by the decoder of VAE after training and the target pattern (**Fig. 3a**).

Two conditions are required for good perception performance: (1) there exists a state z_0 at which the generated pattern well resembles the target pattern; and (2) the representations of the target patterns in the bottleneck state should be 'smooth' so that starting from a state z_1 ($z_1 \neq z_0$), the optimal state z_0 can be reached by error updating using, in our model, gradient descent algorithm (**Fig. 3a**).

Guided by these two understandings, we studied the perception performance of a VAE trained by Model a ($a = 1, 2, 3$) in two steps:

(1) We set the target pattern $\mathcal{T}$ (i.e., blue curve in **Fig. 3a**) to be an activity pattern of Model a, and set the initial bottleneck state to be the value of the μ-channel (see **Fig. 1d**) of the VAE encoder, which is the optimal bottleneck state estimated by the encoder; then we updated the bottleneck state to minimize the error between $\mathcal{T}$ and the decoder output. We denote the bottleneck state at the end of the updating as $\hat{z}_0$.

(2) We set the target pattern $\mathcal{T}$ to be the decoder output when the bottleneck state took $\hat{z}_0$, and set the initial bottleneck state to be $\hat{z}_0 + \epsilon$, with ϵ being a Gaussian perturbation with standard deviation $\sigma_{perturb}$, and updated the bottleneck state again.

We quantified the reconstruction quality in the above two steps using r_1^2 and r_2^2, respectively representing the ratio of variance of the target pattern explained by the decoder output at the end of each step. r_1^2 quantifies how optimal the bottleneck state can be to reconstruct an activity pattern of Model a ($a = 1, 2, 3$). r_2^2 quantifies how smooth the representations of the output patterns in the bottleneck state can be so that the optimal bottleneck state can be easily found by error updating. We found that Model 2 and Model 3 correspond to high r_1^2 values (**Fig. 3b**), and r_2^2 for Model 3 is higher than that for Model 2 (**Fig. 3c**).

The high r_2^2 value for Model 1 (**Fig. 3c**) may be due to the failure of pattern generation (**Fig. 2a**), so that the target pattern $\mathcal{T}$ in the second step is not a delta peak as the activity pattern of Model 1. The low r_2^2 value for Model 2 implies the ruggedness of the representations of the wavelets in the bottleneck state, consistent with the fragmentary representations shown in the upper panel of **Fig. 2b**.

The first step above models a typical two-stage process of perception: information is transmitted feedforwardly at the early stage of perception, but processed recurrently at the later stage [7], which reflects the shift of computational demand from responding speed to representation precision. The value of the μ-channel is the fast-estimation of the bottleneck state by the feedforward pathway before refinement. The second step above studies the case when the feedforward fast-estimation is inaccurate, which may be caused by a fast unexpected change of the stimulus. Our results suggest that the perception of low-precision representations (i.e., the activity pattern of Model 3) is more accurate, and can be more quickly adapted to unexpected situations.

Inferring fine-scale information from the activities of low-coding-precision neurons

The previous subsection suggests that higher-order cortex is better at simulating or reconstructing low-precision representations for imagination or perception. For fine-scale information in images to be imagined or perceived by the brain, fine-scale information must be able to be inferred from the top-down generated low-precision representations. Here we study this coarse-to-fine inference problem.

We consider a population of low-coding-precision neurons (**Fig. 4a**) receiving stimuli from a one-dimensional circle. The feedforward connection from the stimulus receptor at position θ_{stim} to the low-coding-precision neuron preferring θ_{low} Gaussianly decays with $|\theta_{stim} - \theta_{low}|$. We study the following problem: given an activity pattern $\mathcal{P}_{low}$ of low-coding-precision neurons, how to infer the corresponding pattern of stimulus receptors?

The stimulus receptors in this model have higher coding precision that the neurons in the second layer. By studying the inference of the activities of stimulus receptors from the activities of the neurons in the second layer, we can gain insight into the inference of fine-scale visual information from low-precision representations.

The most straightforward method is to let the inferred stimulus be $\hat{\mathcal{P}}_{stim} = \boldsymbol{W}_{forward}^{-1}\mathcal{P}_{low}$, where $\boldsymbol{W}_{forward}$ is the feedforward connection matrix from stimulus receptors to low-coding-precision neurons. However, this method is not successful (**Fig. 4b**): $\boldsymbol{W}_{forward}$ has very small eigenvalues corresponding to the eigenvectors that encode fine spatial details (**Fig. 4c**), so $\boldsymbol{W}_{forward}^{-1}$ has very large eigenvalues. These large eigenvalues make $\boldsymbol{W}_{forward}^{-1}$ strongly amplify random noises (which were induced by limited numeric precision in **Fig. 4b**), impairing the inference.

The inference difficulty results from the fact that the group $\mathcal{G}$ of stimulus receptors with similar position preferences give output to overlapping low-coding-precision neurons, so that it is hard to tell which stimulus receptor causes an activity pattern of the low-coding-precision neurons. If the stimulus receptors in $\mathcal{G}$ have sparse activities, so that in the ideal case only a single receptor in $\mathcal{G}$ is active, all the others being silent, then the inference can be achieved by exploiting the sparseness of $\mathcal{G}$, similar to the compressed sensing scenario [30].

To manifest the advantage of the sparseness of $\mathcal{G}$ for inference, we supposed that all stimulus receptors except those preferring two positions θ_1 and θ_2 were silent, and studied how the inference quality depended on $\Delta\theta_{stim} = |\theta_1 - \theta_2|$ (**Fig. 4a**). When $\Delta\theta_{stim}$ is small, these

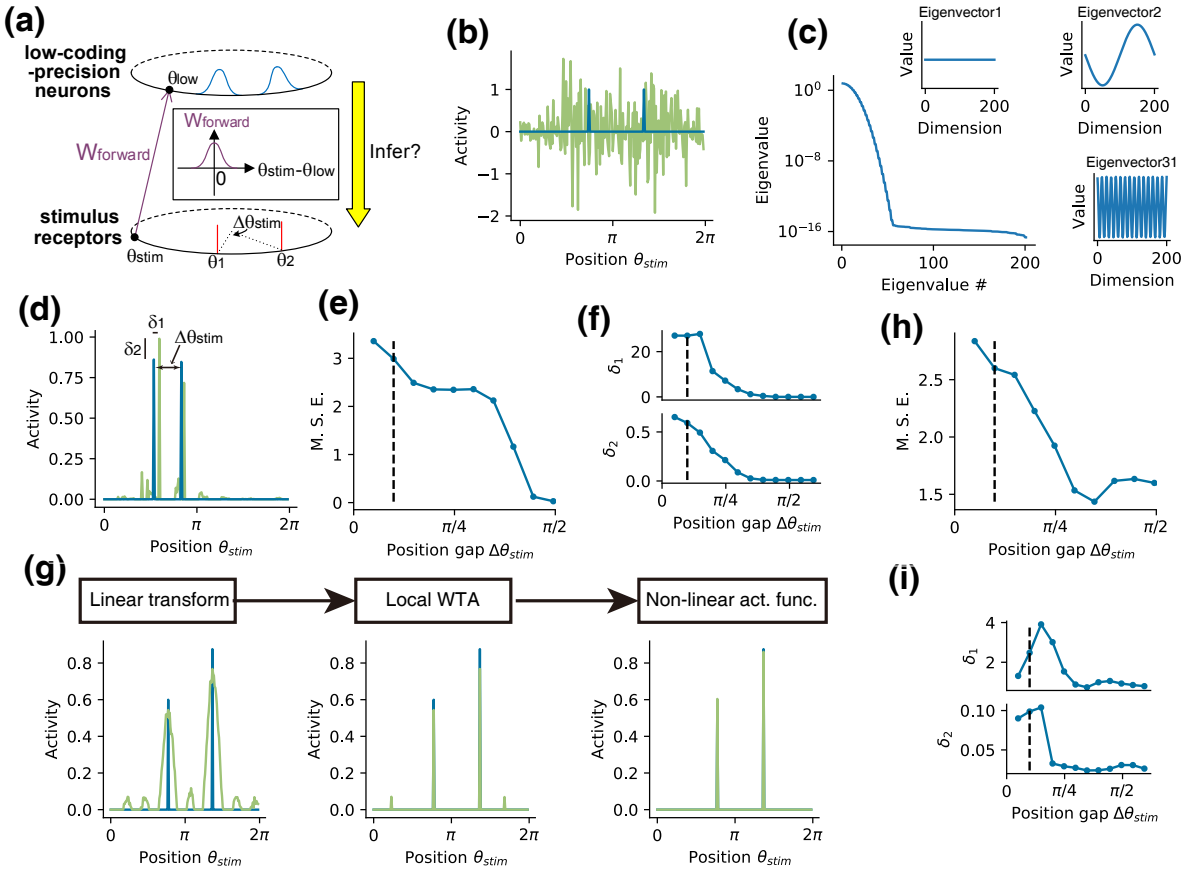

Figure 4: Sparse fine-scale information can be inferred from the activities of low-coding-precision neurons. (**a**) Schematic of the model. Stimulus receptors and low-coding-precision neurons are located on a circle. The feedforward connections $\boldsymbol{W}_{forward}$ (magenta) from stimulus receptors to low-coding-precision neurons Gaussianly decays with the difference $|\theta_{stim} - \theta_{low}|$ between their position preferences. The activities of stimulus receptors are delta-peaked (red) at θ_1 and θ_2 (with $\Delta\theta_{stim} = |\theta_1 - \theta_2|$). The activities of low-coding-precision neurons are Gaussianly bumped (blue). We study the inference of the activities of stimulus receptors from the activities of low-coding-precision neurons (yellow arrow). (**b**) Blue: the activity of stimulus receptors as a function of θ_{stim}. Green: inferred activity using $\boldsymbol{W}^{-1}_{forward}$. (**c**) Eigenspectrum of $\boldsymbol{W}_{forward}$. Insets are examples of eigenvectors. (**d**) The original activity (blue) and the inferred activities (green) of stimulus receptors using matching-pursuit algorithm. δ_1 and δ_2 respectively represent the differences of the positions and heights of the original and inferred peaks. (**e**) The mean-square error of the inferred activity as a function of the gap $\Delta\theta_{stim}$ between the positions of the two peaks. Vertical dashed line is the standard deviation of the Gaussian decay in $\boldsymbol{W}_{forward}$. (**f**) δ_1 and δ_2 as a function of $\Delta\theta_{stim}$. (**g**) Flow chart of the set of operations that can be fulfilled using a network with feedforward and recurrent connections. Three panels below respectively illustrate the real (blue) and inferred (green) activity pattern of stimulus receptors after each of the three steps. Note that local winner-take-all (WTA) sharpens the peaks, nonlinear activation function suppresses small peaks and also makes the heights the inferred peaks closer to the original ones. (**h, i**) Similar to panels **e** and **f**, but for the set of operations manifested in panel **g**. In panels **e, f, h, i**, error bars representing s.e.m. over 8192 samples are smaller than the plot markers.

two stimulus receptors have similar position preferences, so they can be considered to belong to the same group $\mathcal{G}$; when $\Delta\theta_{stim}$ gets larger, these two stimulus receptors have farther different preferences, so a group $\mathcal{G}$ which contain stimulus receptors with similar position preferences have sparser activity.

We studied two inference algorithms: (1) a classical matching-pursuit algorithm [31], which aims to explain a given pattern of low-coding-precision neurons using as few active stimulus receptors as possible; (2) a set of operations that can be fulfilled using neural networks with only feedforward and recurrent connections, consisting of a linear transformation, a local winner-take-all operation, and a nonlinear activation function (see Methods). We used three indexes to quantify the inference quality: (1) mean-square error of the inferred stimulus-receptor activity pattern $\hat{\mathcal{P}}_{stim}$ from the original pattern $\mathcal{P}_{stim}$, (2) the difference δ_1 between the positions of the peaks in $\hat{\mathcal{P}}_{stim}$ and $\mathcal{P}_{stim}$, and (3) the difference δ_2 between the heights of the peaks in $\hat{\mathcal{P}}_{stim}$ and $\mathcal{P}_{stim}$ (**Fig. 4d**). In both algorithms, all the three indexes indicated that the inference error was small when $\Delta\theta_{stim}$ was large (**Fig. 4e,f,h,i**). These results suggest that if fine-scale information is sparse, then fine-scale information can be inferred from low-coding-precision neurons. In particular, those fine-scale features that cause similar activity patterns of low-coding-precision neurons should have sparse strength.

Improving imagination and perception of fine-scale information using low-precision generation

The previous subsections show that the brain is good at generating low-precision representations through top-down pathway, and that fine-scale information can be inferred from low-precision representations if fine-scale information is sparse. These results imply that fine-scale visual information can be generated in two steps: (1) generating the activities of low-precision neurons; and (2) inferring fine-scale information from low-precision neurons. In this subsection, we will show that comparing with generating high-precision representations, generating low-precision representations improves imaginative creativity and perceptive accuracy and robustness of fine-scale information.

Model setup

We conducted our study using skeleton MNIST dataset [32]. Images in this dataset (denoted as dataset $\mathcal{DS}_1$ below) represent digits using lines of 1 pixel width (**Fig. 5a, second column**). There are two reasons why we study $\mathcal{DS}_1$. First, images in $\mathcal{DS}_1$ are made up of 1-pixel-width lines, which are the fine-scale information we would like to study. Second, images in $\mathcal{DS}_1$ are analogous to the activity patterns of V1 simple cells, which are to be generated by higher-order cortex through top-down pathway: the thin lines in $\mathcal{DS}_1$ mimic the enhanced representation of contours in V1 due to collinear facilitation [33], whose saliency attracts attention; the thin lines in $\mathcal{DS}_1$ also mimic the sharpened tuning of V1 simple cells under attention [34]. To study the best way to imagine or perceive these 1-pixel-width lines, we trained VAEs using three datasets $\mathcal{DS}_1$, $\mathcal{DS}_2$ and $\mathcal{DS}_3$, explained below:

(1) $\mathcal{DS}_1$ can be regarded as the displacement of Model 1 (**Fig. 1b, left**) for 2-dimensional-image case, because a 1-pixel-width line is analogous to the delta-peak pattern in Model 1 along the direction perpendicular to the line (comparing the second column of **Fig. 5a** with **Fig. 1b, left**).

(2) $\mathcal{DS}_2$ is obtained by processing $\mathcal{DS}_1$, so that the pixel intensities oscillatorily decay along the direction perpendicular to a line in the skeleton MNIST images (**Fig. 5a, third column**). $\mathcal{DS}_2$ can be regarded as the displacement of Model 2 (**Fig. 1b, middle**) for 2-dimensional-image case.

(3) $\mathcal{DS}_3$ is obtained by processing $\mathcal{DS}_1$, so that the 1-pixel-width lines are blurred, with the pixel intensities Gaussianly decaying along the direction perpendicular to a line (**Fig. 5a, fourth column**). $\mathcal{DS}_3$ can be regarded as the displacement of Model 3 (**Fig. 1b, right**) for 2-dimensional-image case.

To infer the 1-pixel-width lines from the VAE-generated images, we post-processed the VAE-generated images in the following way: we first binarized the images using a threshold θ_{thres}, such that pixels with intensities larger (or smaller) than θ_{thres} were set to 1 (or 0), then skeletonized the images (see Methods), resulting in binary images with lines of 1-pixel width (**Fig. 5b and lower panels of c**). By comparing the quality of the post-processed images resulting from the three datasets ($\mathcal{DS}_1$, $\mathcal{DS}_2$ and $\mathcal{DS}_3$), we can know the best way to imagine or perceive the fine-scale information of 1-pixel-width lines.

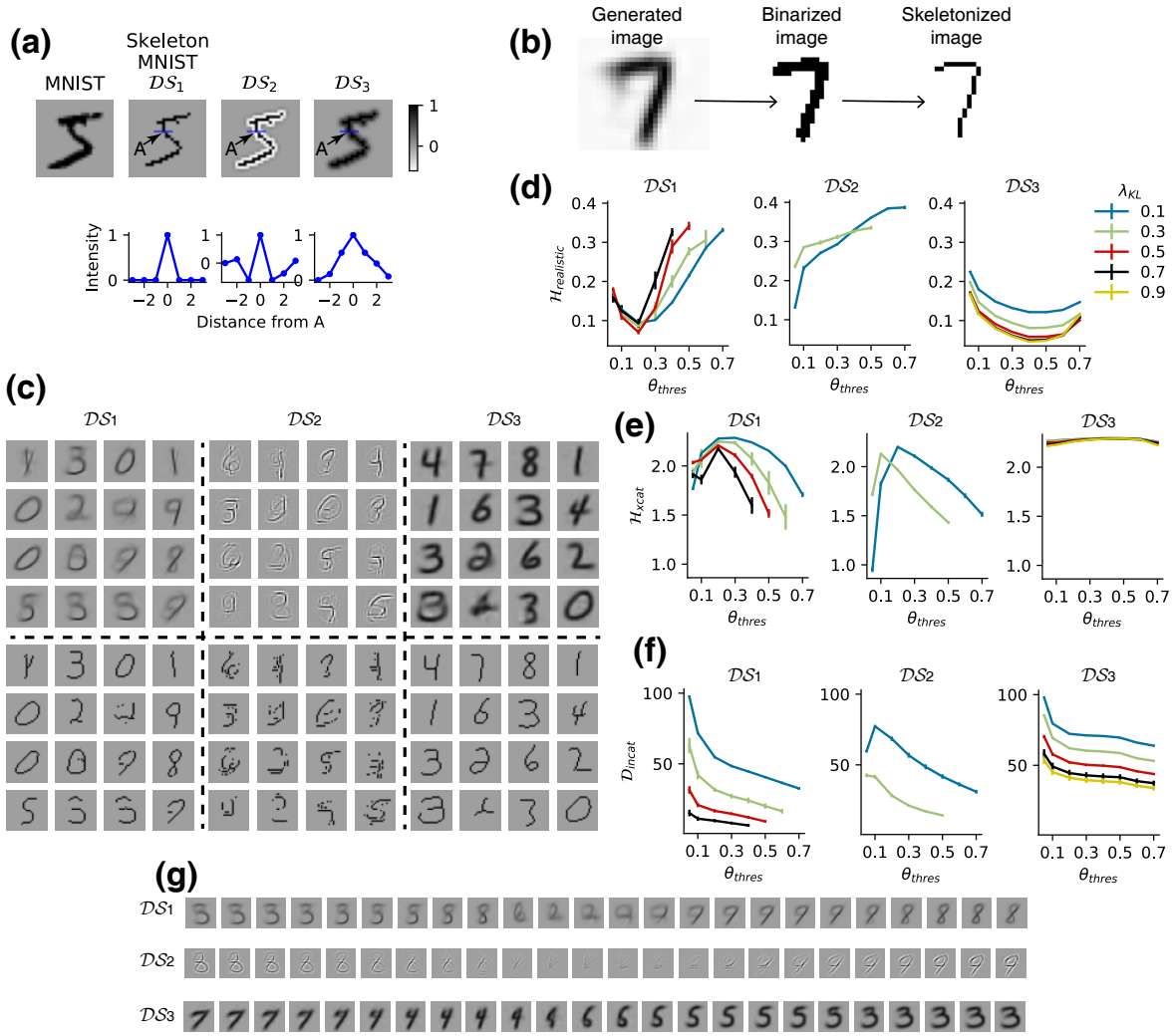

Figure 5: Low-precision generation facilitates imaginative creativity of fine-scale information. (**a**) Upper panels: example images in the MNIST, skeleton MNIST ($\mathcal{DS}_1$), $\mathcal{DS}_2$ and $\mathcal{DS}_3$ datasets. Lower panels: pixel intensity along the blue bar in the upper panel as a function of the distance from point A (i.e., the middle point of the bar). (**b**) An example manifesting the post-processing of the images generated by VAE. (**c**) Upper panels: examples of the generated images when using dataset $\mathcal{DS}_1$, $\mathcal{DS}_2$ and $\mathcal{DS}_3$ to train VAE. Lower panels: the generated images after post-processing. Note that $\mathcal{DS}_3$ results in better image quality than $\mathcal{DS}_1$ and $\mathcal{DS}_2$. (**d**) $\mathcal{H}_{realistic}$ as a function of the binarization threshold θ_{thres} when the parameter λ_{KL} in VAE (see Methods) takes different values, when VAE is trained using $\mathcal{DS}_1$, $\mathcal{DS}_2$ and $\mathcal{DS}_3$ respectively. Some data points for $\mathcal{DS}_1$ and $\mathcal{DS}_2$ are missing, see Methods for explanation. (**e, f**) Similar to panel **d**, but for $\mathcal{H}_{xcat}$ and D_{incat}. (**g**) Examples of the generated images when the bottleneck state slowly and linearly changes its value, see more examples in **Fig. S1**. In panels **d, e, f**, error bars representing s.e.m. over 8 VAE configurations.

As we mentioned in the last subsection, sparseness is required for the inference of fine-scale information from low-precision representations. The inference of 1-pixel-width lines above is also based on the sparseness of 1-pixel-width lines. To understand this point, suppose in a VAE-generated image, there is a very thick line, whose thickness is twice as much as normal. Despite its thickness, however, only a single 1-pixel-width line (instead of two parallel 1-pixel-width lines) results from the skeletonization process illustrated in **Fig. 5b**. Therefore, the inference of 1-pixel-width line by skeletonization is based on the sparseness assumption, and is inaccurate if 1-pixel-width lines are densely aggregated together.

Imaginative creativity

People have proposed that creativity is a mental process of generating worthwhile and novel ideas [35]. Inspired by this definition of creativity, we propose that good generation of digital images of 1-pixel-width lines should satisfy three conditions. (1) Realisticity: a generated image should look like a digit in dataset $\mathcal{DS}_1$. (2) Cross-category variety: the numbers of the generated images looking like different digits should be almost the same. In other words, it is not good if the VAE can generate images looking like the same digit. (3) Within-category variety: the shapes of the images of the same digit should be various. To quantify the image-generation quality, we trained a neural network to classify the images in $\mathcal{DS}_1$, so that the network output a label distribution $p(x|I)$ ($x = 0, 1, \cdots, 9$) after receiving a post-processed generated image I (see Methods). Realisticity requires that $p(x|I)$ has low entropy $\mathcal{H}_{realistic}$ for each I [36]. Cross-category variety requies that the marginal $\int_{I\in A} p(x|I)\mathrm{d}I$ has high entropy $\mathcal{H}_{xcat}$, with A being the set of all the post-processed generated images [36]. To quantify within-category variety, we calculated the intrinsic dimensionality $D_{incat} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}$ [37], where λ_i is the eigenspectrum of the post-processed generated images A_0 belonging to the same category. D_{incat} has maximal value if all principal components (PCs) of the set A_0 have equal variance, and has small value if a single PC dominates.

We investigated $\mathcal{H}_{real}$, $\mathcal{H}_{xcat}$ and D_{incat} with the change of the binarization threshold θ_{thres} (**Fig. 5b**) and a parameter λ_{KL} in VAE which controls the regularization strength onto the distribution of the bottleneck state variable (see Methods). We found that VAEs trained by $\mathcal{DS}_3$ generated images with better eye-looking quality than VAEs trained by $\mathcal{DS}_1$ or $\mathcal{DS}_2$ (**Fig. 5c**), consistent with the quantitive indications of smaller $\mathcal{H}_{real}$, larger $\mathcal{H}_{xcat}$ and larger D_{incat} of the post-processed images in large range of parameters θ_{thres} and λ_{KL} (**Fig. 5d-f**). These results suggest that generating low-precision representations facilitates the visual system to imagine fine-scale information.

Similar to the 1-dimensional-image case (**Fig. 2a-c**), we also investigated the series of the generated images when the bottleneck state z was continuously changing (**Fig. 5g**). The images generated by the VAE trained by $\mathcal{DS}_3$ have two advantages comparing with those resulting from $\mathcal{DS}_1$ and $\mathcal{DS}_2$: (1) when the change of z is not large so that the generated images look like the same digit, the images generated by $\mathcal{DS}_3$ undergo more smooth and flexible shape morphing, whereas the images generated by $\mathcal{DS}_1$ and $\mathcal{DS}_2$ are more rigid; (2) when the change of z is large so that the generated images experience digit transition, the images generated by $\mathcal{DS}_3$ look more realistic during the transition period, whereas the images generated by $\mathcal{DS}_1$ and $\mathcal{DS}_2$ are more unrecognizable during the transition period (see **Fig. S1** for more examples). This investigation gains insights into the facilitation of imaginative creativity by low-precision generation.

Visual perception

To model the perception of an image $\mathcal{I}_1$ in $\mathcal{DS}_1$, we updated the bottleneck state of a VAE to minimize the difference between the generated image and the target image (**Fig. 3a**), with this target image being $\mathcal{I}_1$ itself, or the image in $\mathcal{DS}_2$ or $\mathcal{DS}_3$ that corresponds to $\mathcal{I}_1$, depending on whether the VAE was trained by $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$. The perception performance was quantified from two aspects:

(1) The similarity between the generated images (**Fig. 6a, second row**) and the target images (**Fig. 6a, first row**).

(2) We post-processed the VAE-generated images by binarizing and skeletonizing to infer the 1-pixel-width lines (**Fig. 5b**), then we compared the post-processed images (**Fig. 6a, third row**) with the original image $\mathcal{I}_1$ in $\mathcal{DS}_1$ (**Fig. 6a, red box**).

The first aspect was quantified in a similar way to that used for 1-dimensional images (**Fig. 3b, c**). We found that $\mathcal{DS}_3$ resulted in the highest values of r_1^2 and r_2^2 (**Fig. 6b, c**), supporting that the images with blurred lines can be reconstructed more faithfully, and that this reconstruction is more robust to perturbations to the bottleneck state of VAE.

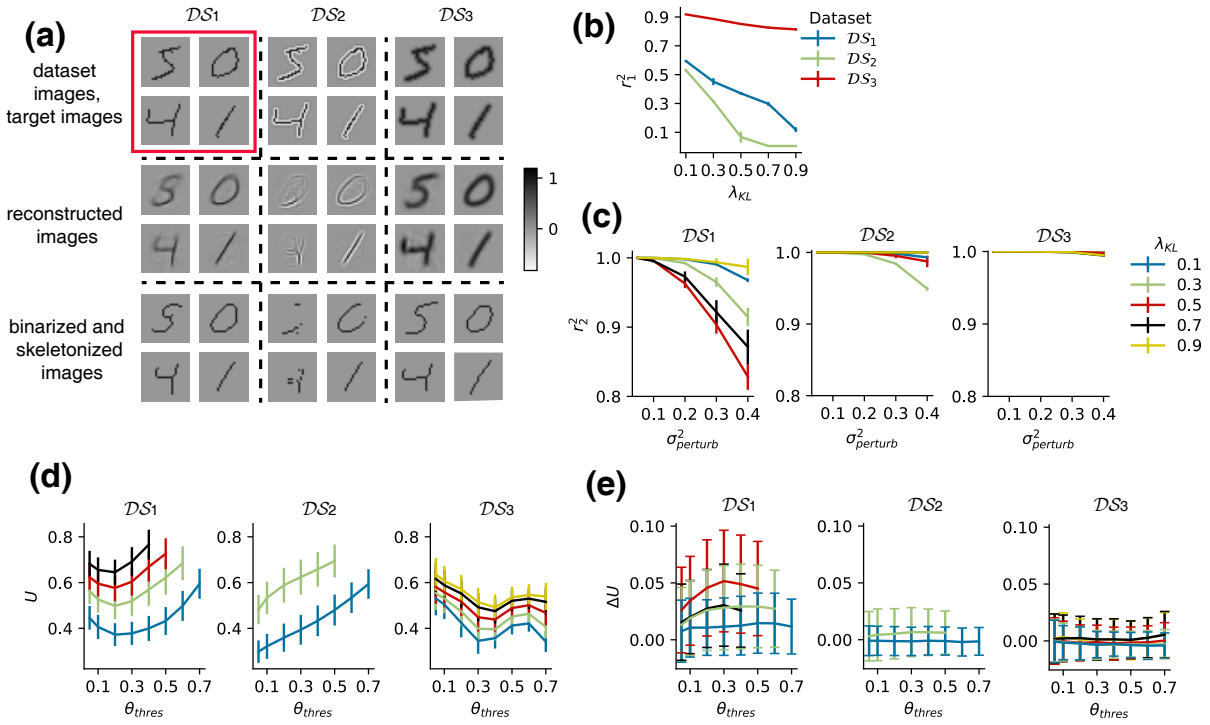

Figure 6: Low-precision generation facilitates perception of fine-scale information. (**a**) Examples of the dataset images (first row) which are also the target images for perception, the reconstructed images by VAE (second row), and the images post-processed from the reconstructed images (third row). Note that $\mathcal{DS}_3$ result in the best quality of reconstructed images, and the post-processed images for $\mathcal{DS}_3$ also have the highest similarity to the orginal images (those in the red box). (**b**) r_1^2 as a function of λ_{KL} for the three datasets. (**c**) r_2^2 as a function of $\sigma_{perturb}$ for the three datasets when λ_{KL} takes different values. (**d**) U as a function of θ_{thres} for the three datasets when λ_{KL} takes different values. (**e**) ΔU is the increase of U value caused by the perturbation of bottleneck state. Error bars represent s.e.m. over 8 VAE configurations.

To quantify the second aspect, we calculated the portion U of the pixels in 1-pixel-width lines in the original images in $\mathcal{DS}_1$ that could not be reconstructed by the post-processed images. We first set the initial bottleneck state to be the μ-channel of the encoder output, optimized the bottleneck state, and found that U for VAEs trained by $\mathcal{DS}_3$ had relatively lower values in large parameter range than that for $\mathcal{DS}_1$ or $\mathcal{DS}_2$ (**Fig. 6d**). We then studied the perception robustness by perturbing the bottleneck state of VAE from the optimal state found in the previous step before again optimizing the bottleneck state (see Methods), and found that the increase of U caused by this perturbation for $\mathcal{DS}_3$ was smaller than that for $\mathcal{DS}_1$ and was not larger than that for $\mathcal{DS}_2$ (**Fig. 6e**). Therefore, better and more robust perception of fine-scale information (here, 1-pixel-width lines in $\mathcal{DS}_1$ images) can be achieved by inferring fine-scale information from the reconstruction of low-precision representations (here, $\mathcal{DS}_3$ images).

Discussion

In this paper, we show that the perceptive accuracy and robustness as well as imaginative creativity of visual information with fine spatial scale can be improved by top-down generation of low-precision representations. This improvement results from two reasons: (1) low-precision representations of fine-scale visual information can be better reconstructed or simulated through top-down pathway than high-precision representations, and (2) fine-scale information can be inferred from low-precision representations if the fine-scale information is sparse. These results suggest that the best way in perceiving or imagining sparse fine-scale information is to generate low-precision representations that contain fine-scale information, instead of generating high-precision representations.

Low-coding precision has been found in extra-classical receptive fields of V1 neurons in perception [26] and the internally generated

activities in imagination [22, 23]. Considering that sparseness is a basic computational principle of V1 activity [5], and that fine-scale information can be further sparsified by attention-based selection [38, 39], our results suggest that top-down generating low-precision representations in V1 is not a shortcoming, but instead a strategy that the brain uses to improve the perception and imagination of fine-scale information.

Representations smooth to shape morphing

Our study suggests that the advantage to generate low-precision neural representations is closely related to the formation of neural representations in high-order cortex that are smooth to shape morphing of fine-scale information (**Figs. 2c, 5g**). Smoothness implies that the semantic meaning of the state of the high-order cortex can be interpolated. States of the high-order cortex that fail in this meaningful interpolation may generate unmeaningful representations during imagination, and cannot guide the updating of the state of the high-order cortex during perception (see the fragmentary represention in **Fig. 2b** and the corresponding inferior performance in imagination and perception). Representations that are smooth to shape morphing have been widely observed in neural network models that generate realistic-looking images [27, 40]. Such representations can be used to do algebraic operations on the semantic meanings of the generated images [41], and can also be exploited to design semi-supervised learning algorithms that train discriminators using only a few labeled data [42, 43].

It is well known that VAE, which tries to construct representations smooth to shape morphing of stimuli in the bottleneck state, tends to generate blurred images; vanilla auto-encoder, which does not constrain the structure of representations in the bottleneck state, does not have the tendency to output blurred images [44]. Therefore, the low precision of the top-down generated representations may be a price that the brain has to pay for structured representations of the world in high-order cortex. Such structured representations are important for cognition [45, 46].

Engineering neural networks to generate images with high-resolution details is a non-trivial task. VAE tends to generate blurred images [44]. Another well-known generative model, generative adversarial network (GAN), also struggles to generate high-resolution details [47] due to a gradient problem [48]. Our results suggest that blurring these high-resolution details may result in better creativity and perception performance, as long as these high-resolution details can be later inferred from the generated low-resolution representations. The implementation of this idea in engineering VAE or GAN for image generation or deep predictive coding network for image recognition is an interesting research direction.

In [22], it is found that the receptive fields of V1 neurons in imagined images have larger sizes and lower spatial-frequency preferences than in seen images. A direct prediction of our study is that V1 neurons also have less sensitivity to grating phases in imagined images, which may be tested using the experimental protocol of [22].

Inference of sparse fine-scale information

The inference of fine-scale information from low-precision representations requires the sparseness of fine-scale information, especially for those fine-scale features which stimulate highly overlapping low-precision patterns (**Fig. 4**). Sparseness has long been proposed as a computational principle of V1 activity [5]. Besides, this sparse requirement can also be realized by attention. On the one hand, attention selects salient or task-relevant sensory inputs [38, 39], sparsifying the fine-scale information for perception. On the other hand, attention also selects features to imagine: anticipation or intention reflects a simulation of the real world before things happen or actions are executed [49, 50], and attention selects features under anticipation or intention before actually sensing the stimuli [51, 52]. For example, such attention guides us to search for a featured object in a cluttered scene [52].

Although we have illustrated the inference of fine-scale information from low-precision representations in **Figs. 4 and 5b**, one should note that our purpose here is to show that low-precision representations have the ability to code fine-scale information. The transformation of top-down generated low-precision representations to high-precision representations may not physically happen in the brain: we have shown that better perception results from updating the state of the high-order cortex according to the error calculated from low- (instead of high-) precision representations (**Figs. 3, 6**). Therefore, instead of transforming top-down generated low-precision representations to high precision, it is better to transform bottom-up stimulated high-precision representations to low precision before calculating the bottom-up and

top-down mismatch for predictive-coding perception. Such high-to-low precision transformation can easily be done through feedforward projection [30]. It has been found that top-down prediction error is feedforwardly transmitted in predominantly high (gamma) frequency range [53, 54], so the coding precision of prediction error may be understood by investigating the receptive fields of V1 neuron responses in gamma frequency range.

The power-law eigenspectrum of V1 activity

It has been found that the variance of the nth principal component of V1 activity decays in power law $n^{-\alpha}$ with $\alpha \to 1^+$ [55]. The value of this exponent α ensures the differentiability of neural code, such that V1 responses smoothly change with the positions of stimuli [55]. Our results support the importance of code differentiability for top-down generation. In the 1-dimensional-image case (**Figs. 1, 2**), the activity pattern of Model 1 is non-differentiable to the position p of the dot stimulus (**Figs. 1b, blue line of 2e**); consistently, Model 1 results in the worst VAE-generation performance (**Figs. 2a-c**). In the 2-dimensional-image case (**Fig. 5**), $\mathcal{DS}_2$ has the worst differentiability among the three datasets, because the intensity of a pixel immediately adjacent to a 1-pixel-width line in $\mathcal{DS}_2$ has the largest difference from the intensity of a pixel on the 1-pixel-width line (see the second row of **Fig. 5a**); consistently, $\mathcal{DS}_2$ results in the worst VAE-generation performance (**Fig. 5d-f**). Therefore, the power-law spectrum of V1 activity [55] may be closely related to the top-down generative function of the visual system. A possible way to understand the importance of code differentiability to top-down generation is as follows. Suppose that higher-order cortex represents a stimulus at position x using firing-rate code $z(x)$, and the V1 firing-rate code is $v_1(x)$. Predictive-coding perception and realistic imagination require that $v_1(x) \approx f(z(x))$, where $f(\cdot)$ represents the top-down network, which is differentiable in the biologically plausible context. Therefore, if $z(x)$ is to be differentiable to x (which means that higher-order cortex forms representations differentiable to shape morphing of stimuli, see **Fig. 2c**), $v_1(x)$ must also be differentiable to x to better fulfill the equation $v_1(x) \approx f(z(x))$.

Method

The imagination and perception of 1-dimensional images

VAE structure

A variational auto-encoder (VAE) [27] contains two parts: an encoder and a decoder (**Fig. 1a**). The encoder is a feedforward network that receives input $\boldsymbol{x}$ and gives two channels of output $\boldsymbol{\mu}$ and $\log(\boldsymbol{\sigma}^2)$. Then a random number $\boldsymbol{z}$ is generated according to the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, and is input to the decoder which outputs $\hat{\boldsymbol{x}}$. VAE is trained to minimize the following cost function:

$$E_{VAE} = \sum_{i=1}^{D_x} (x_i - \hat{x}_i)^2 - \frac{\lambda_{KL}}{2} \sum_{j=1}^{D_z} (1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2), \tag{1}$$

where D_x is the dimension of the input and the output, D_z is the dimension of the random variable $\boldsymbol{z}$. Minimizing the first term of this equation reduces the reconstruction error, minimizing the second term (which is the KL-divergence of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ from the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$) makes $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ close to $\mathcal{N}(\mathbf{0}, \mathbf{1})$. λ_{KL} is a parameter controlling the relative strengths of these two terms.

In the VAE used in **Figs. 2 and 3**, the encoder was a multilayer perceptron (MLP) that had three hidden layers with sizes 100, 50 and 20 respectively. The input layer was of size 201, and the output layer had two channels each with size 1. The decoder was another MLP that had three hidden layers with sizes 20, 50 and 100 respectively. Adjacent layers were all-to-all connected. We used leaky relu as the activation functions of the hidden layers. VAE was trained using Adam optimizer [56] using Pytorch.

The three models of 1-dimensional images

In the three models in **Fig. 1b**, 201 neurons are positioned on a one-dimensional line. In Model 1, the activity pattern is a delta peak $f_1(x; a) = \delta_{x,a}$ in which only a single neuron at position a has activity 1, whereas all the other neurons have zero activity. In Model 2,

this pattern is a Gabor wavelet $f_2(x;a) = C\exp(-\frac{(x-a)^2}{2\sigma^2})\sin(\frac{2\pi}{T}(x-a))$, where $\sigma = 10$, $T = 80$, C is a normalization factor such that $\max_x f_2(x;a) = 1$. In Model 3, this pattern is a Gaussian bump $f_3(x;a) = C\exp(-\frac{(x-a)^2}{2\sigma^2})$, where $\sigma = 10$, C is a normalization factor such as $\max_x f_3(x;a) = 1$. In $f_1(x;a)$, $f_2(x;a)$ and $f_3(x;a)$, a is a random integer from 31 to 171.

The quality of imagination

To quantify the quality of the generated patterns (**Fig. 2d**), for any generated pattern p, we defined $r^2(p) = \max_x(r_0^2(p, s(x)))$, where $s(x)$ is the activity pattern of Model-out in response to the dot stimulus at position x on the 1-dimensional line, and $r_0^2(p, s(x))$ is the ratio of the variance of p that can be explained by $s(x)$ (i.e., coefficient of determination). Specifically, $r_0^2(p, s)$ is defined by

$$r_0^2(p, s) = \frac{\sum_i (p_i - s_i)^2}{\sum_i (p_i - \langle p \rangle)^2} \tag{2}$$

where p_i and s_i are respectively the activities of the ith neuron in the patterns p and s, $\langle p \rangle$ is the mean activity of pattern p.

The quality of perception

In **Fig. 3**, the bottleneck state was optimized to minimize the error between the target pattern and the generated pattern using Adam optimizer [56]. Both r_1^2 and r_2^2 are the coefficient of determination to quantify how much variance of the target pattern can be explained by the generated pattern.

The imagination and perception of 2-dimensional images

VAE structure

The VAE used in **Fig. 5** had a similar structure with that used in **Fig. 1**, except that the size of the input and output layers was $28 \times 28 = 784$, and the sizes of the three hidden layers of the encoder (or decoder) were 512, 256 and 128 (or 128, 256 and 512) respectively. The size of each of the two output channels of the encoder was 20.

The three datasets of 2-dimensional images

The dataset $\mathcal{DS}_1$ in **Fig. 5a** is the skeleton MNIST dataset [32]. The intensity of a pixel in an image in $\mathcal{DS}_1$ is binary (1 or 0) depending on whether this pixel belongs to a line of 1-pixel width.

An image I_2 in $\mathcal{DS}_2$ was obtained using an image I_1 in $\mathcal{DS}_1$ in the following way. To determine the intensity $T_2(x_2, y_2)$ of a pixel $I_2(x_2, y_2)$ at the position (x_2, y_2) in I_2, we defined a box $\mathcal{B}(x_2, y_2) = \{I_1(\eta, \xi) : |\eta - x_2| \leq 2, |\xi - y_2| \leq 2\}$ in I_1. We looked for a pixel $I_1(x_1, y_1) \in \mathcal{B}(x_2, y_2)$ such that its intensity $T_1(x_1, y_1) = 1$ and the distance $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ was minimized. Then we set $T_2(x_2, y_2) = a\exp(-d^2/2)$, where $a = -1$ if $\max(|x_1 - x_2|, |y_1 - y_2|) = 1$, and $a = 1$ otherwise. If all pixels in $\mathcal{B}(x_2, y_2)$ had intensity 0, then $T_2(x_2, y_2) = 0$.

$\mathcal{DS}_3$ was generated using $\mathcal{DS}_1$ in a similar way to above, except that $a = 1$ all the time.

The quality of imaginative creativity

The images generated by VAE were post-processed in two steps (**Fig. 5b**). First, images were binarized such that pixels with intensities larger (or smaller) than a threshold θ_{thres} were set to 1 (or 0). Second, the images were skeletonized using the 'skeletonize' routine of the skimage python package.

To quantify the quality of the post-processed images, we trained a MLP to classify the skeleton MNIST dataset (i.e., $\mathcal{DS}_1$). This MLP contained a hidden layer of size 1000 with leaky-relu activation function. After receiving a post-processed VAE-generated image I, this

MLP output a label distribution $p(x|I)$ ($x = 0, 1, \cdots, 9$). In **Fig. 5d**, $\mathcal{H}_{realistic} = E_I[-\sum_x p(x|I) \ln p(x|I)]$, where $E_I[\cdot]$ means average over all post-processed VAE-generated images [36]; in **Fig. 5e**, $\mathcal{H}_{xcat} = -\sum_x E_I[p(x|I)] \ln E_I[p(x|I)]$ [36]. To plot **Fig. 5f**, we first chose the post-processed VAE-generated images with high realisticity (i.e., $\max_x p(x|I) > 0.9$), then for all the images belonging to a category x, we calculated the variance $\lambda_i(x)$ of the ith principal component (PC), D_{incat} was defined as $D_{incat} = E_x[\frac{(\sum_i \lambda_i(x))^2}{\sum_i \lambda_i^2(x)}]$ [37]. **Fig. 5d-f** show how $\mathcal{H}_{realistic}$, $\mathcal{H}_{xcat}$ and D_{incat} change with the binarization threshold θ_{thres} and the parameter λ_{KL} in **eq. 1**. Note that if θ_{thres} is high, the image after post-processing may be very sparse (i.e., only a few pixels are nonzero), especially when λ_{KL} also takes a large value. In this case, the MLP network has an artifact that $p(x|I)$ strongly peaks at $x = 1$, and $p(x \neq 1|I)$ has very small value. Because of this artifact, in **Fig. 5d-f**, we excluded the data points at which the percentage of nonzero pixels in the post-processed images was smaller than 1%. Some data points when $\lambda_{KL} = 0.9$ for $\mathcal{DS}_1$ and when $\lambda_{KL} = 0.5, 0.7, 0.9$ for $\mathcal{DS}_2$ resulted in images with sparsity a little larger than 1%, but we also excluded these data points, because the quality of the generated images was really bad. These artifacts are weak for $\mathcal{DS}_3$ in our range of parameters, so we plotted the whole parameter range for $\mathcal{DS}_3$.

Fig. 5g was plotted by gradually and linearly changing the bottleneck states of VAEs from $\boldsymbol{z} = [1.5, 1.5, \cdots, 1.5]$ to $[-1.5, -1.5, \cdots, -1.5]$.

The quality of perception

As we mentioned in the main text, the perception of 2-dimensional images was quantified from two aspects.

The first aspect quantified how well and robust the VAE-generated images can approximate the images in $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$. This aspect was quantified by r_1^2 and r_2^2 in a similar way to that in the 1-dimensional-image case. Specifically, to calculate r_1^2, we initialized the bottleneck state to be the μ-channel of the VAE encoder after inputting an image I in $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$ to the encoder, depending on whether the VAE was trained by $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$. Then we optimized the bottleneck state by gradient descent to minimize the difference between I and the image generated by the VAE decoder. r_1^2 is the variance of I that can be explained by the generated image. We denote the image generated by the VAE decoder at the end of this step as I_g. To calculate r_2^2, we perturbed the optimized bottleneck state in the previous step by a Gaussian noise with standard deviation $\sigma_{perturb}$, and updated the bottleneck state again to minimize the difference between the generated image and I_g. r_2^2 is the variance of I_g that can be explained by the generated image.

The second aspect quantified how well and robust the VAE-generated image post-processed by binarizing and skeletonizing (**Fig. 5b**) can approximate a given image I_1 in $\mathcal{DS}_1$. We quantified the second aspect in the following two steps:

(1) Given an image I_1 in $\mathcal{DS}_1$, we set the initial bottleneck state to be the μ-channel of the VAE encoder in response to I_1 or the corresponding image in $\mathcal{DS}_2$ or $\mathcal{DS}_3$, depending on whether the VAE was trained by $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$. We then set the target image to be I_1 or the corresponding image in $\mathcal{DS}_2$ or $\mathcal{DS}_3$, depending on whether the VAE was trained by $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$, and optimized the bottleneck state of the VAE to minimize the difference between the image generated by the decoder and the target image. Then we post-processed the generated image by binarizing and skeletonizing, and calculated the portion U of the pixels in the 1-pixel-width lines of I_1 that could not be reconstructed by the post-processed image. Specifically, U is defined by:

$$U = \langle \frac{\sum_i I_{1,i}(I_{1,i} - I_{postprocess,i})^2}{\sum_i I_{1,i}} \rangle_{I_1}, \tag{3}$$

where $I_{1,i}$ and $I_{postprocess,i}$ are respectively the ith pixel of I_1 and the post-processed generated image, and $\langle \cdot \rangle_{I_1}$ means to take average over I_1 (i.e., the images in $\mathcal{DS}_1$). Both $I_{1,i}$ and $I_{postprocess,i}$ take binary values (0 or 1).

(2) We then perturbed the optimized bottleneck state in the above step by a Gaussian noise with standard deviation $\sigma_{perturb} = 0.4$, optimized the bottleneck state again to minimize the difference between the image generated by the VAE decoder and the target image, and calculated U using eq. 3 again. **Fig. 6e** plots the increase of U value in this second step from that in the first step.

Inferring fine-scale information from the activities of low-coding-precision neurons

In our simulation for **Fig. 4a**, both the stimulus receptor layer and the low-coding-precision neuron layer contained $N = 201$ neurons lying on a one-dimensional circle. Only two stimulus receptors were active, separated by

$$\Delta N = \Delta\theta_{stim} N/2\pi \tag{4}$$

neurons, with $\Delta N = 5, 10, 15, \cdots, 60$. The activities of the two active stimulus receptors were set to two random values in the range $[0.5, 1]$, and the activities of all the other stimulus receptors were zero.

The feedforward connection from a stimulus receptor preferring $\theta_{stim} \in [0, 2\pi)$ to a low-coding-precision neuron preferring $\theta_{low} \in [0, 2\pi)$ was

$$W_{forward}(\theta_{stim}, \theta_{low}) = C \exp(-\frac{(\min(2\pi - |\theta_{stim} - \theta_{low}|, |\theta_{stim} - \theta_{low}|))^2}{2\sigma^2}),$$

where $\sigma = 2\pi \times 10/N$, C is a normalization factor so that $\sqrt{\sum_{n=1}^{201} W_{forward}(\theta_{stim}, 2\pi n/N)^2} = 1$.

Given the activities $\{x_1, x_2, \cdots, x_N\}$ of stimulus receptors, we calculated the activities $\{y_1, y_2, \cdots, y_N\}$ of low-coding-precision neurons using the feedforward matrix $\boldsymbol{W}_{forward}$, and then we studied how to infer the activities $\{\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_N\}$ of stimulus receptors using $\{y_1, y_2, \cdots, y_N\}$.

We first tried matching-pursuit algorithm [31] (**Fig. 4d-f**). Specifically, we first defined the remnant $r_i = y_i$, and set $\hat{x}_i = 0$ ($i = 1, 2, \cdots, N$), then iterated the following steps for 1000 times:

(i) Calculate $c_i = \sum_{j=1}^{N} r_j (W_{forward}^T)_{ji}$, then set $c_k = 0$ for $k \neq \arg\max_i |c_i|$.

(ii) Let $\hat{x}_i \leftarrow \hat{x}_i + c_i$, $r_i \leftarrow r_i - \sum_{j=1}^{N} c_j (W_{forward})_{ji}$.

To manifest the biological possibility of the inference, we then performed this inference using a set of operations that can be fulfilled using neural networks that contain only feedforward and recurrent connections, consisting of a linear transformation, a local winner-take-all (WTA) operation, and a nonlinear activation function (**Fig. 4g-i**). The linear transformation was defined by $g_{L,i} = \sum_j w_{ij} y_j$, where y_j is the activity of the jth low-precision neuron, $g_{L,i}$ is the ith output of the linear transformation, w_{ij} is the transformation matrix. We determined w_{ij} by training. Specifically, we minimized the following cost function:

$$E_L = \sum_i (x_i + \epsilon)(g_{L,i} - x_i)^2, \tag{5}$$

where $\epsilon = 0.01$. This cost function means that we would like $g_{L,i}$ to be close to x_i, emphasizing this closeness for large x_i values. The output $g_{W,i}$ of the local WTA operation was calculated by initializing $g_{W,i} = g_{L,i}$ and the set $\mathcal{P} = \emptyset$ and then iterating the following steps: selecting the $i \notin \mathcal{P}$ with maximal $g_{W,i}$, let $g_{W,j} = 0$ for $0 < |j - i| \leq \sigma_{WTA}$ (where σ_{WTA} is the spatial scale of the local WTA), and let $\mathcal{P} = \mathcal{P} \cup \{i\}$. In our simulation, we set $\sigma_{WTA} = \text{floor}((\Delta N + 1)/2)$, where floor$(x)$ is the maximal integer that is no larger than x, and ΔN is the distance between the two peaks in the activity pattern of the stimulus-receptor layer, defined in **eq. 4**. The output $g_{N,i}$ of the nonlinear activation function was calculated as

$$g_{N,i} = \begin{cases} 0, & g_{W,i} < \theta_{th} \\ a g_{W,i} + b, & g_{W,i} \geq \theta_{th} \end{cases}. \tag{6}$$

We set $\theta_{th} = 0.2$, which is well below the smallest possible height of a peak in the stimulus-receptor layer (note that in the first paragraph of this subsection of Methods, we mentioned that the height of a peak in an activity pattern of the stimulus receptors was a random number from the range $[0.5, 1]$). We found that the output $g_{W,i}$ of local WTA contained some small peaks that did not correspond to the two peaks in the activity pattern of stimulus receptors (**Fig. 4g, middle panel**). By setting $g_{N,i} = 0$ when $g_{W,i} < \theta_{th}$, we removed these small peaks (**Fig. 4g, right panel**).

The values of a and b in **eq. 6** were the same for all is, and were set so that the heights of the highest two peaks in $\{g_{N,i}\}_i$ were close to the heights of the two peaks in $\{x_i\}_i$ (**Fig. 4g, right panel**). To better explain how we fulfill this point, we denote the positions (i.e., indexes) of the highest two peaks in $\{g_{W,i}\}_i$ as i_{gW1} and i_{gW2}, and the positions of the two peaks in $\{x_i\}_i$ as i_{x1} and i_{x2}. We

denote the position of the peak in $\{x_i\}_i$ that is closer to i_{gW1} (or i_{gW2}) to be j_{x1} (or j_{x2}), i.e., $j_{x1} = \arg\min_{i\in\{i_{x1},i_{x2}\}}(|i - i_{gW1}|)$ (or $j_{x2} = \arg\min_{i\in\{i_{x1},i_{x2}\}}(|i - i_{gW2}|)$). We plotted pairs of values $(g_{W,i_{gW1}}, x_{j_{x1}})$ and $(g_{W,i_{gW2}}, x_{j_{x2}})$ for all simulation trials in the same figure, and fit these dots using a linear function. The slop and intercept of this linear function were respectively the parameters a and b in eq. 6.

Acknowledgements

Z.B. thanks Prof. Changsong Zhou and Prof. Shuzhi Sam Ge for comments on the manuscript and helpful discussions. This work was supported by the National Natural Science Foundation of China (32000694 to Z.B.) and the start-up fund of the Institute for Future of Qingdao University.

Declaration of interests

The author declares no competing interests.

References

[1] K. Friston, "The free-energy principle: a unified brain theory?," *Nat. Rev. Neurosci.*, vol. 11, pp. 127–138, 2010.

[2] K. Friston, "Does predictive coding have a future?," *Nat. Neurosci.*, vol. 21, pp. 1019–1021, 2018.

[3] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nat. Neurosci.*, vol. 2, pp. 79–87, 1999.

[4] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The helmholtz machine," *Neural Comput.*, vol. 7, pp. 889–904, 1995.

[5] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[6] M. Heilbron and M. Chait, "Great expectations: Is there evidence for predictive coding in auditory cortex?," *Neuroscience*, vol. 389, pp. 54–73, 2018.

[7] N. Dijkstra, L. Ambrogioni, D. Vidaurre, and M. van Gerven, "Neural dynamics of perceptual inference and its reversal during imagery," *eLife*, vol. 9, p. e53588, 2020.

[8] N. Dijkstra, S. E. Bosch, and M. A. J. van Gerven, "Shared neural mechanisms of visual perception and imagery," *Trends Cogn. Sci.*, vol. 23, pp. 423–434, 2019.

[9] A. M. Albers, P. Kok, I. Toni, H. C. Dijkerman, and F. P. de Lange, "Shared representations for working memory and mental imagery in early visual cortex," *Curr. Biol.*, vol. 23, pp. 1427–1431, 2013.

[10] R. Kanai, A. Chang, Y. Yu, I. M. de Abril, M. Biehl, and N. Guttenberg, "Information generation as a functional basis of consciousness," *Neurosci. Conscious.*, vol. 5, p. niz016, 2019.

[11] W. Wiese, "The science of consciousness does not need another theory, it needs a minimal unifying model," *Neurosci. Conscious.*, vol. 6, p. niaa013, 2020.

[12] H. Supèr, H. Spekreijse, and V. A. Lamme, "Two distinct modes of sensory processing observed in monkey primary visual cortex (V1)," *Nat. Neurosci.*, vol. 4, pp. 304–310, 2001.

[13] B. van Vugt, B. Dagnino, D. Vartak, H. Safaai, S. Panzeri, S. Dehaene, and P. R. Roelfsema, "The threshold for conscious report: Signal loss and response bias in visual and frontal cortex," *Science*, vol. 360, pp. 537–542, 2018.

[14] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene, "Conscious processing and the global neuronal workspace hypothesis," *Neuron*, vol. 105, pp. 776–798, 2020.

[15] L. Mudrik, N. Faivre, and C. Koch, "Information integration without awareness," *Trends Cogn. Sci.*, vol. 18, pp. 488–496, 2014.

[16] D. Soto and J. Silvanto, "Reappraising the relationship between working memory and conscious awareness," *Trends Cogn. Sci.*, vol. 18, pp. 520–525, 2014.

[17] S. V. Gaal and V. A. F. Lamme, "Unconscious high-level information processing: implication for neurobiological theories of consciousness," *Neuroscientist*, vol. 18, pp. 287–301, 2012.

[18] J. Pearson, "The human imagination: the cognitive neuroscience of visual mental imagery," *Nat. Rev. Neurosci.*, vol. 20, pp. 624–634, 2019.

[19] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav. Brain Sci.*, vol. 40, p. e253, 2017.

[20] M. G. Mattar and N. D. Daw, "Prioritized memory access explains planning and hippocampal replay," *Nat. Neurosci.*, vol. 21, pp. 1609–1617, 2018.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge: The MIT Press, 2018.

[22] J. L. Breedlove, G. St-Yves, C. A. Olman, and T. Naselaris, "Generative feedback explains distinct brain activity codes for seen and mental images," *Curr. Biol.*, vol. 30, pp. 1–14, 2020.

[23] S. E. Favila, B. A. Kuhl, and J. Winawer, "Perception and memory have distinct spatial tuning properties in human visual cortex," *bioRxiv*, 2020.

[24] A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, and P. C. Bressloff, "Circuits and mechanisms for surround modulation in visual cortex," *Annu. Rev. Neurosci.*, vol. 40, pp. 425–451, 2017.

[25] A. J. Keller, M. M. Roth, and M. Scanziani, "Feedback generates a second receptive field in neurons of the visual cortex," *Nature*, vol. 582, pp. 545–549, 2020.

[26] G. C. DeAngelis, R. D. Freeman, and I. Ohzawa, "Length and width tuning of neurons in the cat's primary visual cortex," *J. Neurophysiol.*, vol. 71, pp. 347–374, 1994.

[27] D. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[28] J. D. Murray, A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo, and X.-J. Wang, "Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex," *Proc. Natl Acad. Sci. USA*, vol. 114, pp. 394–399, 2017.

[29] A. Parthasarathy, C. Tang, R. Herikstad, L. F. Cheong, S.-C. Yen, and C. Libedinsky, "Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex," *Nat. Commun.*, vol. 10, p. 4995, 2019.

[30] S. Ganguli and H. Sompolinsky, "Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis," *Annu. Rev. Neurosci.*, vol. 14, pp. 485–508, 2012.

[31] M. Rani, S. B. Dhok, and R. B. Deshmukh, "A systematic review of compressive sensing: Concepts, implementations and applications," *IEEE Access*, vol. 6, pp. 4875–4894, 2018.

[32] E. D. de Jong, "Incremental sequence learning," *arXiv:1611.03068*, 2016.

[33] W. Li, V. Piëch, and C. D. Gilbert, "Contour saliency in primary visual cortex," *Neuron*, vol. 50, pp. 951–962, 2006.

[34] K. Anton-Erxleben and M. Carrasco, "Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence," *Nat. Rev. Neurosci.*, vol. 14, pp. 188–200, 2013.

[35] N. D. Pisapia, F. Bacci, D. Parrott, and D. Melcher, "Brain networks for visual creativity: a functional connectivity study of planning a visual artwork," *Sci. Rep.*, vol. 6, p. 39185, 2016.

[36] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2234–2242, 2016.

[37] K. Rajan, L. Abbott, and H. Sompolinsky, "Inferring stimulus selectivity from the spatial structure of neural network dynamics," in *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[38] L. Itti and C. Koch, "Computational modelling of visual attention," *Nat. Rev. Neurosci.*, vol. 2, pp. 194–203, 2001.

[39] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nat. Rev. Neurosci.*, vol. 14, pp. 350–363, 2013.

[40] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *The International Conference on Learning Representations*, 2016.

[41] A. B. L. Larsen, S. K. Sϕnderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of Machine Learning Research*, 2016.

[42] D. P. Kingma, D. J. Rezendey, S. Mohamedy, and M. Welling, "Semi-supervised learning with deep generative models," in *Conference on Neural Information Processing Systems*, 2014.

[43] M. E. Abbasnejad, A. Dick, and A. van den Hengel, "Infinite variational autoencoder for semi-supervised learning," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge: The MIT Press, 2016.

[45] Y. Fang, X. Wang, S. Zhong, L. Song, Z. Han, G. Gong, and Y. Bi, "Semantic representation in the white matter pathway," *PLoS Biol.*, vol. 16, p. e2003993, 2018.

[46] A. O. Constantinescu, J. X. O'Reilly, and T. E. J. Behrens, "Organizing conceptual knowledge in humans with a gridlike code," *Science*, vol. 352, pp. 1464–1468, 2016.

[47] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *International Conference on Machine Learning*, 2017.

[48] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *The International Conference on Learning Representations*, 2017.

[49] G. Pezzulo and C. Castelfranchi, "Intentional action: from anticipation to goal-directed behavior," *Psychol. Res.*, vol. 73, pp. 437–440, 2009.

[50] M. Jeannerod, "The representing brain: neural correlates of motor intention and imagery," *Behav. Brain Sci.*, vol. 17, pp. 187–202, 1994.

[51] A. C. Nobre and F. van Ede, "Anticipated moments: temporal structure in attention," *Nat. Rev. Neurosci.*, vol. 19, pp. 34–48, 2018.

[52] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," *Nat. Hum. Behav.*, vol. 1, p. 0058, 2017.

[53] L. H. Arnal, V. Wyart, and A.-L. Giraud, "Transitions in neural oscillations reflect prediction errors generated in audiovisual speech," *Nat. Neurosci.*, vol. 14, pp. 797–801, 2011.

[54] A. M. Bastos, J. Vezoli, C. A. Bosman, J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P. D. Weerd, H. Kennedy, and P. Fries, "Visual areas exert feedforward and feedback influences through distinct frequency channels," *Neuron*, vol. 85, pp. 390–401, 2015.

[55] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, "High-dimensional geometry of population responses in visual cortex," *Nature*, vol. 571, pp. 361–365, 2019.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *The International Conference on Learning Representations*, 2015.