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

Zedong Bi[1,2,*]

[1]Institute for Future, Qingdao University, Qingdao 266071, China

[2]School of Automation, Shandong Key Laboratory of Industrial Control Technology, Qingdao University, Qingdao 266071, China

[*]Email: zedong.bi@outlook.com

Abstract

Predictive-coding perception or imagination requires top-down generated neural representations in the primary visual cortex (V1) to reconstruct or simulate the V1 representations bottom-up stimulated by the seen images. However, top-down generated representations 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 the visual system using variational auto-encoder, we reveal that training to generate low-precision representations facilitates high-order cortex to form representations smooth to shape morphing of stimuli. Fine-scale visual information, if sparse, can be more accurately and robustly perceived or creatively imagined by inferring from generated representations with low (instead of high) coding precision. Therefore, generating low-precision representations is not a shortcoming, but instead a strategy that the brain uses to improve the perception and imagination of fine-scale information.

1 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 updating 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 imagination 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], if stimuli are transmitted feedforwardly without triggering the feedback pathway. Global workspace theory proposes that consciousness happens when top-down connections broadcast a mental model of sensory information to the whole brain. The broadcasted mental model is then 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 essential role in prospective planning and retrospective analysis [10, 19, 20], and is critical in model-based reinforcement learning [21].

However, in both perception and imagination, top-down generated representations have lower coding precision than the representations bottom-up 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]. Suppose one agrees that imagination requires top-down generated representations to simulate the representations stimulated by seen images [8, 18, 9]. In that case, one should be curious about why and how the brain uses the generated low-precision representations to imagine, especially the visual information with a fine spatial scale that is believed to be coded by V1 neurons with 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, mediated by feedback connections [25], is believed to reflect the reconstruction of V1 activity by the higher-order cortex in predictive-coding perception [3]. However, extra-classical RFs also have lower precision than classical RFs. Extra-classical RFs have larger sizes (i.e., lower spatial precision) than classical RFs. 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]. Suppose one agrees the predictive-coding theory of perception [3]. In that case, 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 of visual information with fine spatial scale coded by phase-sensitive simple cells with small-sized classical RFs. 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).

This paper addresses these problems by modeling the visual system as a variational auto-encoder (VAE) [27]. We show that training to generate low-precision representations facilitates high-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 containing sparse fine-scale information is an important strategy that the brain uses to improve the perception and imagination of fine-scale information, and also reveal the computational advantage of sparse code in V1 from a top-down generative perspective.

Results

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

Model setup

We model the visual system as a variational auto-encoder (VAE) [27], whose output is optimized to reconstruct the input (**Figure 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 (**Figure 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 $\mathcal{I} \rightarrow \mathcal{B}$ and $\mathcal{B} \rightarrow \mathcal{O}$ 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 the standard normal distribution, mimicking the top-down generation in mental imagination [8, 18], and perception can also be modeled through an error-correcting 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 minimizes is the variational free energy [27]. 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 (**Figure 1b, left**). In Model 2, the profile of neuronal activities is an

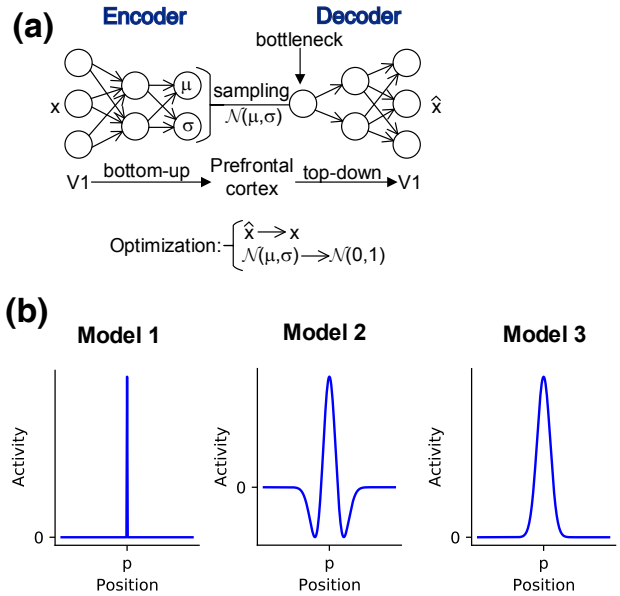

Figure 1: The models to study the generation of representations on a 1-dimensional line. (**a**) Schematic of the variational auto-encoder (VAE). The value z of the bottleneck (which models the prefrontal cortex) is sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, with μ and σ determined 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 (**Figure 1b, middle**). In Model 3, the profile of neuronal activities is a Gaussian bump (**Figure 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.

We can biologically interpret these three models in various ways:

(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 receptive fields of the neurons in Model 1 have smaller sizes than those in Model 3. The receptive fields in Model 2 have the same spatial scale of Gaussian decay as the receptive fields in Model 3, but have a higher spatial frequency.

(2) We can also interpret stimuli in the spatial-frequency domain. Neurons in Model 1 (or Model 3) have sharp (or broad) spatial-frequency tuning.

(3) Neurons in Model 3 can be regarded as the phase-insensitive complex cells corresponding to the phase-sensitive simple cells in Models 1 and 2.

Therefore, by comparing these three models, we can understand the computational advantage of large receptive-field size, low spatial-frequency tuning, and low grating-phase sensitivity in top-down signals in imagination and perception tasks [22, 23, 26].

Visual imagination

Visual imagination is modeled by sampling the bottleneck state z of a trained VAE from the standard normal distribution, so that the decoder part of VAE (**Figure 1a**) outputs various activity patterns. To get a detailed understanding of the relationship between z and the generated pattern, we let z be 1-dimensional, and investigated how the generated pattern changed with z. In Model 1, sharp peaks were never generated by VAE, suggesting the failure of pattern generation (**Figure 2a**). In Model 2, abrupt changes of wavelet positions sometimes happened with the continuous change of z (**Figure 2b**). In Model 3, Gaussian bumps were generated, with the bump position smoothly changing with z (**Figure 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.

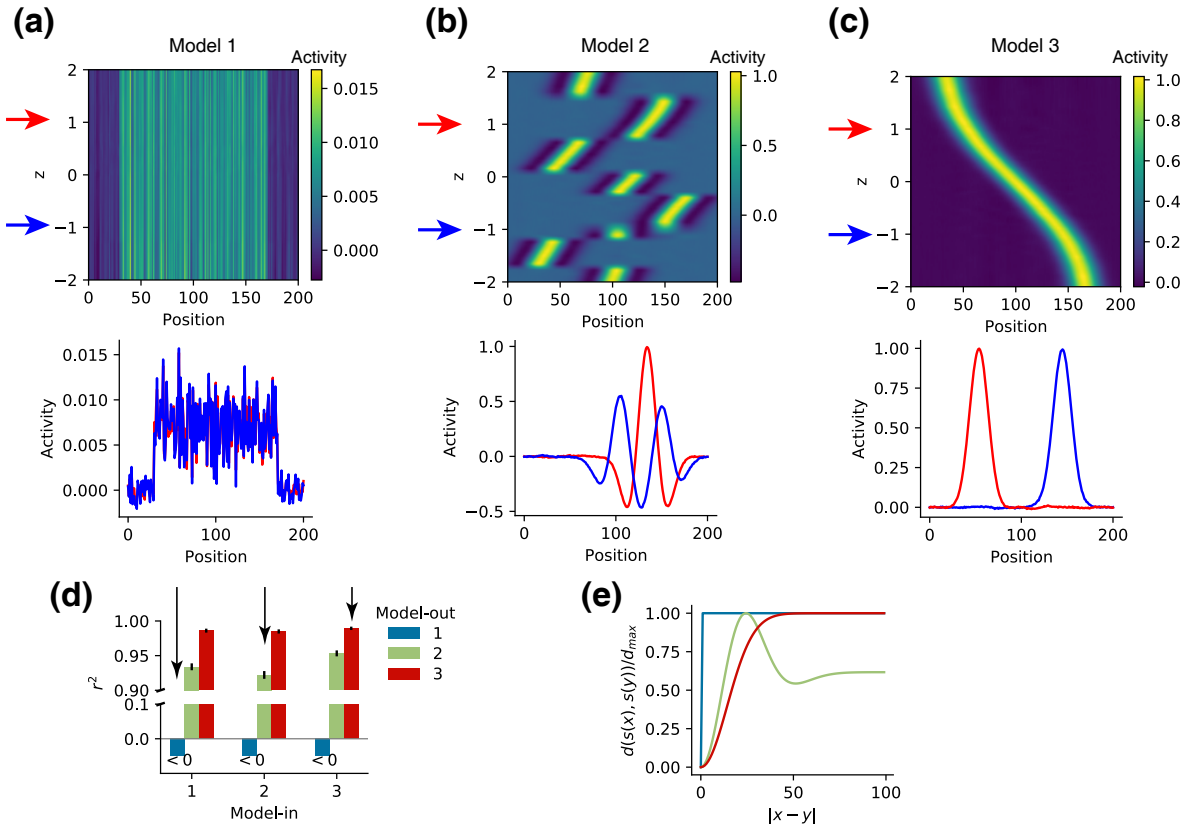

Figure 2: VAE is better at generating low-precision representations. (**a**) Upper panel: generated activity patterns as a function of the state z of the bottleneck variable. Lower panel: blue and red curves respectively represent the two generated patterns when z takes the two values (-1 and 1) indicated 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 **Figure 1a**) be the activity pattern of Model-in, but train the output of VAE (i.e., $\hat{x}$ in **Figure 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 the 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 indicate how well a VAE-generated pattern 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 **Figure 2d**). This result suggests that low-precision representations can be better simulated through the top-down pathway.

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 (**Figure 2c**), because the generation quality is bad around the abrupt change points in Model 2 (blue curve in **Figure 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 the higher-order cortex receives from V1, or from a 'good' pattern that the 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

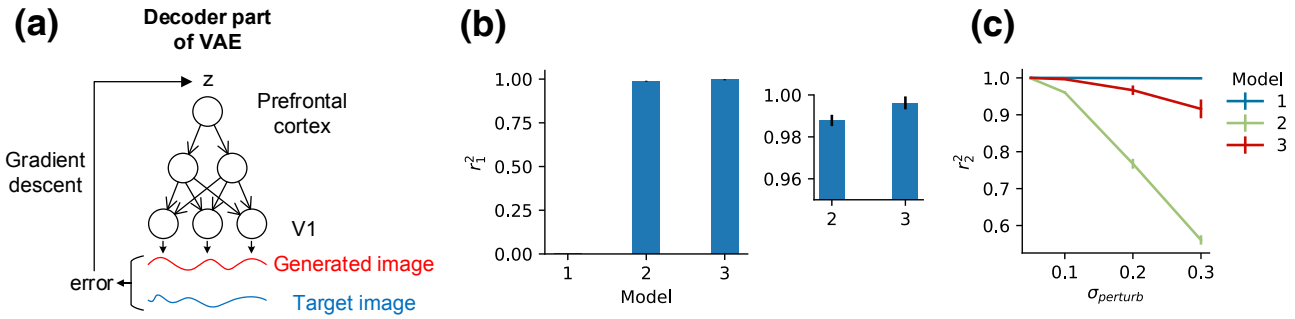

Figure 3: VAE is better at perceiving low-precision representations. (**a**) Schematic of the perception process at the later stage. Only the decoder of a trained VAE is used. The bottleneck state z is updated by gradient descent to minimize the error between the target image (red) to be perceived and the generated image (blue). (**b**) r_1^2 for the VAEs trained by the three models (**Figure 1b**). Inset: results for Models 2 and 3 only. (**c**) r_2^2 as a function of the perturbation strength $\sigma_{perturb}$ for the three models. Error bars represent s.e.m. over 6 VAE configurations.

$\mathcal{P}_a$ (**Figure 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 $s(x)$ and $s(y)$ as a function of $|x-y|$, where $s(x)$ and $s(y)$ are respectively the activity patterns stimulated by two dot stimuli on the 1-dimensional line at positions x and y. 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 $|x-y|$ (**Figure 2e**). This property of Model 3 is important for its advantage in top-down generation. To see this, suppose two nearby bottleneck states z_1 and z_2 ($z_1 \approx z_2$) generate two patterns s_1 and s_2, which correspond 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 $\{s(x)\}_x$. VAE performs stochastic sampling to determine the bottleneck state, so the bottleneck state may be sampled to be z_2 when s_1 is provided as the input to the encoder and the target output of the decoder during training. By doing so, VAE tries to reduce the distance $d(s_1, s_2)$ between the output patterns s_1 and s_2 that correspond to z_1 and z_2 respectively. In Model 3, due to the monotonicity of $d(s_1, s_2)$ with $|x_1 - x_2|$, reducing $d(s_1, s_2)$ reduces $|x_1 - x_2|$. 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 words, the bottleneck state represents the spatial geometry of the stimuli. In Model 2, $d(s_1, s_2)$ is not monotonic with $|x_1 - x_2|$, so reducing $d(s_1, s_2)$ does not always reduce $|x_1 - x_2|$, sometimes instead increases $|x_1 - x_2|$. In Model 1, $d(s_1, s_2)$ remains constant when $s_1 \neq s_2$, so VAE has no gradient clue about how to update s_2 close to s_1.

Visual perception

Perception is a two-stage process: information is transmitted feedforwardly at the early stage and processed recurrently at the later stage, reflecting the shift of computational demand from responding speed to representation precision [7]. We model the early stage by inputting the to-be-perceived image I_{perceive} into the encoder of a VAE, and then setting the bottleneck state of the VAE to be the μ-channel of the encoder. We model the later stage as a predictive-coding process [3], in which the bottleneck state updates its value to minimize the error between I_{perceive} and the image I_{generate} generated by the VAE decoder (**Figure 3a**). We denote the bottleneck state at the end of the updating to be z_{opt} below. Perception accuracy was quantified by r_1^2, which is the ratio of the variance of I_{perceive} explained by I_{generate} after the recurrent updating. We also studied the robustness of perception to the non-optimal estimation of the bottleneck state at the early stage of perception by setting the bottleneck state z to be a perturbed value from z_{opt} and recurrently updating z again. Perception robustness was quantified by r_2^2, which is the the ratio of the variance of I_{perceive} explained by I_{generate} after the second recurrent updating. We found that Model 3 has the highest r_1^2 value (**Figure 3b**), and that Model 3 also has a higher r_2^2 value than that Model 2 (**Figure 3c**), which respectively demonstrate the advantage of low-precision generation to perception accuracy and robustness. See Methods.

Notice that the high r_2^2 value of Model 1 (**Figure 3c**) does not indicate its advantage, because the VAE trained by Model 1 fails to generate

delta peak patterns (**Figure 2a**). The low r_2^2 value of Model 2 is consistent with the representation roughness of the wavelets in the bottleneck state shown in the upper panel of **Figure 2b**: this roughness hinders the bottleneck state from finding the optimal value by gradient descent starting from a state far from z_{opt}.

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

The previous subsection suggests that the 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 inferable 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 (**Figure 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 the low-coding-precision neurons, how to infer the corresponding pattern of the stimulus receptors?

The most straightforward method lets 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 the stimulus receptors to the low-coding-precision neurons. However, this method is unsuccessful (**Figure 4b**): $\boldsymbol{W}_{forward}$ has very small eigenvalues corresponding to the eigenvectors that encode fine spatial details (**Figure 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 **Figure 4b**), impairing the inference.

The inference difficulty results from the fact that a 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 $\Delta\theta_{stim} = |\theta_1 - \theta_2|$ (**Figure 4a**). When $\Delta\theta_{stim}$ is small, these 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 the 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 (**Supplementary Information Section 1** and **Supplementary Figure 1**). 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 $\mathcal{P}_{stim}$ (**Figure 4d**). In both algorithms, all the three indexes indicated that the inference error was small when $\Delta\theta_{stim}$ was large (**Figure 4e,f**, **Supplementary Figure 1**). 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 the imagination and perception of skeleton MNIST digits using low-precision generation

The previous subsections show that the brain is good at generating low-precision representations through the top-down pathway, and that fine-scale information can be inferred from low-precision representations if fine-scale information is sparse. In this subsection, we will show that generating low-precision (instead of high-precision) representations that contain sparse fine-scale information improves imaginative creativity as well as perceptive accuracy and robustness of fine-scale information.

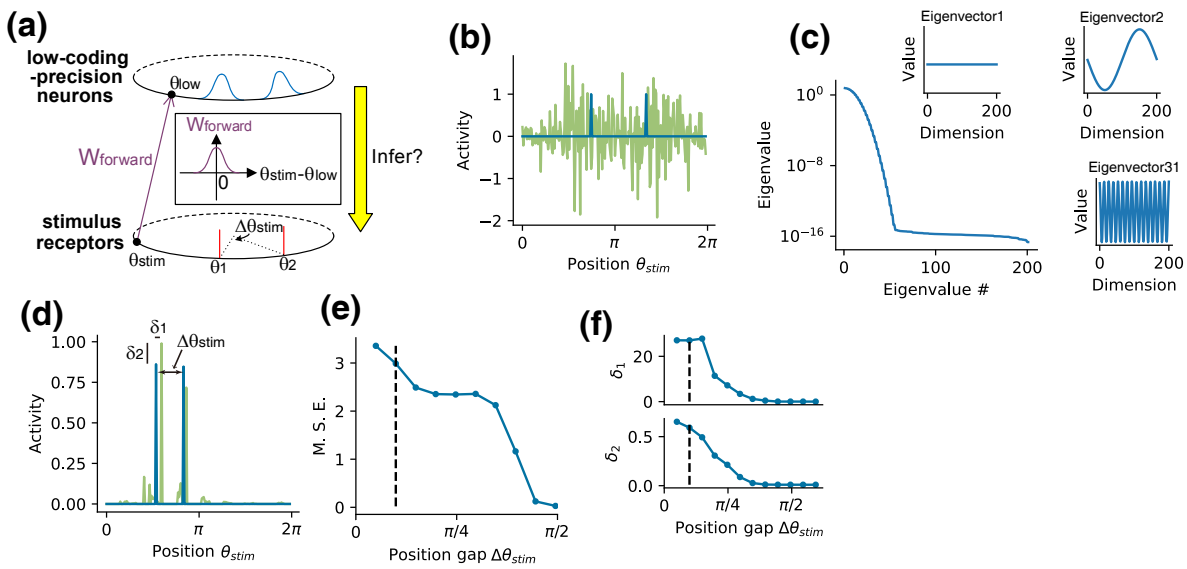

Figure 4: Sparse fine-scale information can be inferred from the activities of low-coding-precision neurons. (**a**) Schematic of the model. Stimulus receptors and low-coding-precision neurons are located on a circle. The feedforward connections $\boldsymbol{W}_{forward}$ (magenta) from stimulus receptors 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 the stimulus receptors from the activities of the low-coding-precision neurons (yellow arrow). (**b**) Blue: the activity of stimulus receptor as a function of θ_{stim}. Green: inferred activity using $\boldsymbol{W}_{forward}^{-1}$. (**c**) Eigenspectrum of $\boldsymbol{W}_{forward}$. Insets are examples of eigenvectors. (**d**) The original activities (blue) of the stimulus receptors and the inferred activities (green) using a matching-pursuit algorithm. δ_1 and δ_2 respectively represent the differences in the positions and heights of the original and inferred peaks. (**e**) Mean-square error between the original and inferred activities 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 functions of $\Delta\theta_{stim}$. In panels **e, f**, error bars representing s.e.m. over 8192 samples are smaller than the plot markers.

Model setup

We conducted our study using the skeleton MNIST dataset [32]. Images in this dataset (denoted as dataset $\mathcal{DS}_1$ below) represent digits using lines of 1-pixel width (**Figure 5a, second column**), which are the fine-scale information we would like to study. Besides, images in $\mathcal{DS}_1$ are also analogous to the activity patterns of V1 simple cells: 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 $\mathcal{DS}_3$, with $\mathcal{DS}_2$ and $\mathcal{DS}_3$ processed from $\mathcal{DS}_1$. In $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$, the pixel intensities delta peak, oscillating decay or Gaussianly decay along the direction perpendicular to a line respectively (**Figure 5a**), so that $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$ are regarded as the displacement of Model 1, 2 or 3 respectively for the 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 if the intensity of a pixel is larger (or smaller) than θ_{thres}, then the intensity is set to be 1 (or 0), then we skeletonized the images (see Methods), resulting in binary images with lines of 1-pixel width (**Figure 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.

As we mentioned in the last subsection (**Figure 4**), sparseness is required to infer fine-scale information from low-precision representations. Similarly, 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 usual. However, despite its thickness, only a single 1-pixel-width line (instead of two parallel 1-pixel-width lines) results from the skeletonization process illustrated in **Figure 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 a good generation of digital images of 1-pixel-width lines satisfies three conditions. (1) Realisticity (quantified by $\mathcal{H}_{real}$): a generated image should look like a digit in dataset $\mathcal{DS}_1$. (2) Cross-category variety (quantified by $\mathcal{H}_{xcat}$): 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 (quantified by D_{incat}): the shapes of the images of the same digit should be various. Realistic and variable generation of digital images are manifested by small $\mathcal{H}_{real}$, large $\mathcal{H}_{xcat}$ and large D_{incat} values. See Methods.

We investigated $\mathcal{H}_{real}$, $\mathcal{H}_{xcat}$ and D_{incat} with the change of the binarization threshold θ_{thres} (**Figure 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 $\mathcal{DS}_2$ (**Figure 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 a large range of parameters θ_{thres} and λ_{KL} (**Figure 5d-f**, **Supplementary Figure 3**). These results suggest that generating low-precision representations facilitates the visual system to imagine fine-scale information.

Similar to the 1-dimensional-image case (**Figure 2a-c**), we also investigated the series of the generated images when the bottleneck state z was continuously changing (**Figure 5g**). The images generated by the VAE trained by $\mathcal{DS}_3$ have two advantages compared 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 **Supplementary Figure 2** for more examples). This investigation gains insight into the facilitation of imaginative creativity by low-precision generation.

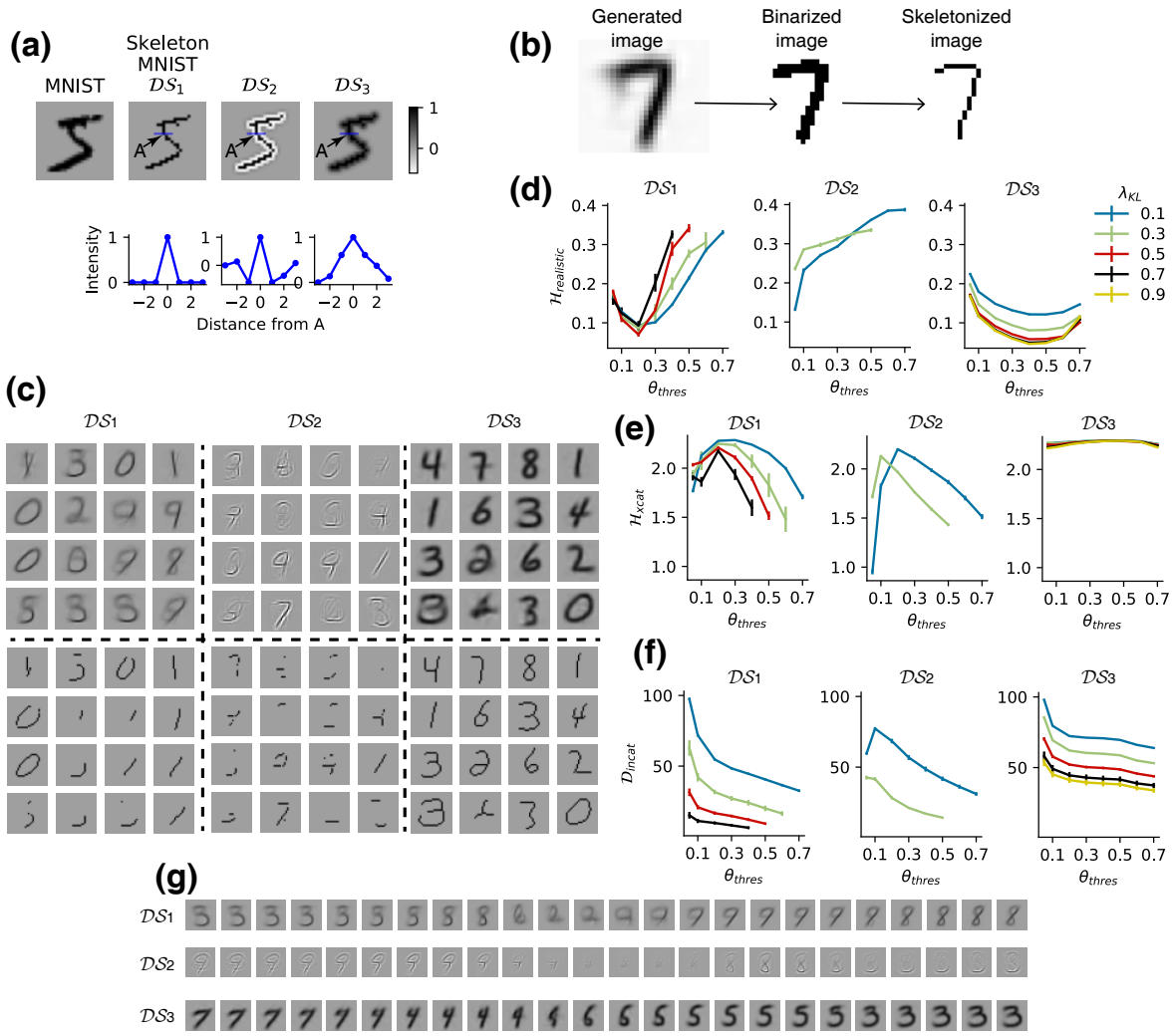

Figure 5: Low-precision generation facilitates the imaginative creativity of skeleton MNIST digits. (**a**) Upper panels: example images in the MNIST, skeleton MNIST ($\mathcal{DS}_1$), $\mathcal{DS}_2$, and $\mathcal{DS}_3$ datasets. Lower panels: pixel intensity along the blue bar in the upper panel 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 datasets $\mathcal{DS}_1$, $\mathcal{DS}_2$ 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 **Supplementary Figure 2**. In panels **d, e, f**, error bars representing s.e.m. over 8 VAE configurations. In panel **c**, $\theta_{thres} = 0.4$. In panels **c** and **g**, $\lambda_{KL} = 0.7$ for $\mathcal{DS}_1$ and $\mathcal{DS}_3$, $\lambda_{KL} = 0.3$ for $\mathcal{DS}_2$.

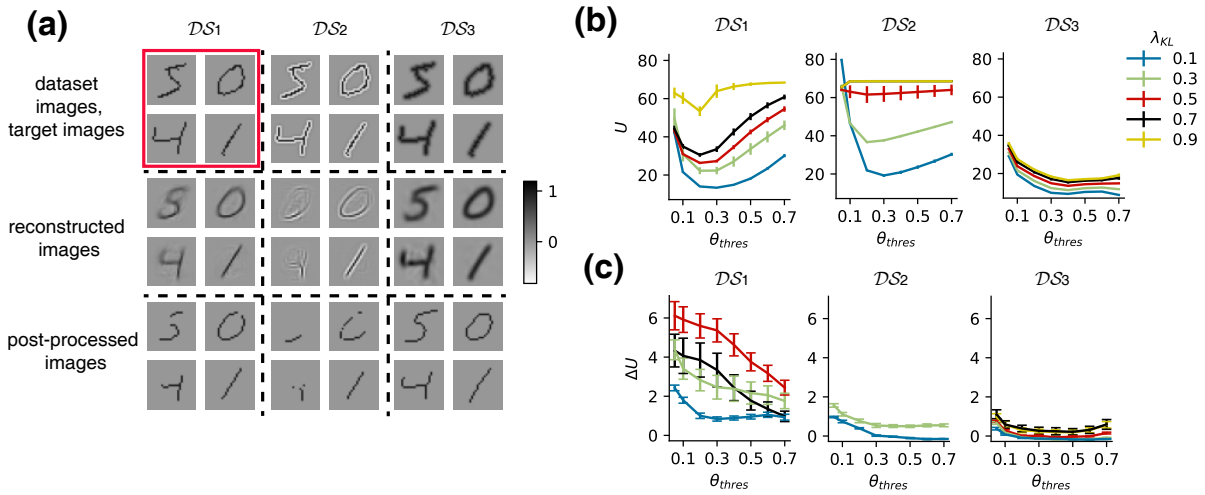

Figure 6: Low-precision generation facilitates the perception of skeleton MNIST digits. (**a**) Examples of the dataset images (first row) which are also the target images I_t for perception, the reconstructed images I_g by VAE (second row), and the images post-processed from the reconstructed images by binarizing and skeletonizing (third row). Note that the post-processed images for $\mathcal{DS}_3$ have the highest similarity to the original images (those in the red box). (**b**) U as a function of θ_{thres} for the three datasets when λ_{KL} takes different values. (**c**) ΔU is the increase of U value caused by the perturbation of the bottleneck state. Some data points for $\mathcal{DS}_1$ and $\mathcal{DS}_2$ are missing, see Methods for explanation. Error bars represent s.e.m. over 8 VAE configurations. In panel **a**, $\lambda_{KL} = 0.3$, $\theta_{thres} = 0.4$.

Visual perception

To model the perception of an image I_1 in $\mathcal{DS}_1$, we updated the bottleneck state of a VAE to minimize the difference between the generated image I_g (**Figure 6a, second row**) and the target image I_t (**Figure 6a, first row**), with I_t being I_1 itself, or the image in $\mathcal{DS}_2$ or $\mathcal{DS}_3$ that corresponds to I_1, depending on whether the VAE was trained by $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$. To evaluate the performance of the VAE to perceive I_1, we post-processed the VAE-generated images by binarizing and skeletonizing to infer the 1-pixel-width lines (**Figure 5b**), and then compared the post-processed images (**Figure 6a, third row**) with the original image I_1 from $\mathcal{DS}_1$ (**Figure 6a, red box**).

The evaluation of perception was performed in two steps. In the first step, we initialized the bottleneck state of the VAE to be the μ-channel of the encoder output in response to I_t, and then optimized the bottleneck state to minimize the mean square difference between I_t and I_g. We found that U, the difference between the post-processed image and I_1 (see Methods), was lower for VAE trained by $\mathcal{DS}_3$ than for VAE trained by $\mathcal{DS}_1$ or $\mathcal{DS}_2$ (**Figure 6b**, **Supplementary Figure 4a**). In the second step, we perturbed the bottleneck state from the optimal state found in the previous step by a Gaussian noise, and then optimized the bottleneck state again (see Methods). We found that the increase of U caused by this perturbation was smaller for $\mathcal{DS}_3$ than for $\mathcal{DS}_1$ or $\mathcal{DS}_2$ in a large parameter range (**Figure 6c**, **Supplementary Figure 4b**). Therefore, more accurate 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).

The imagination and perception of the fine-scale information in textures

The digits shown in skeleton MNIST images have a variety of shapes, but lack internal structures. To examine whether the advantage of low-precision generation persists for the imagination and perception of internal structures of images, we made a database of textures of 1-pixel-width lines synthesized from OpenSimplex2 noise (**Figure 7a**, $\mathcal{DS}_1$) [38], see Methods. Image patches were extracted from this database, processed using similar methods to those shown in **Figure 5a**, resulting in $\mathcal{DS}_1$, $\mathcal{DS}_2$ and $\mathcal{DS}_3$ (**Figure 7a**), which were then used to train VAE.

Imagination was quantified by sliced Wasserstein distance (SWD) [39, 40] between the post-processed VAE-generated images (**Figure 7b, lower row**) and the original images (i.e., $\mathcal{DS}_1$). Perception accuracy was quantified by the difference U between the post-processed

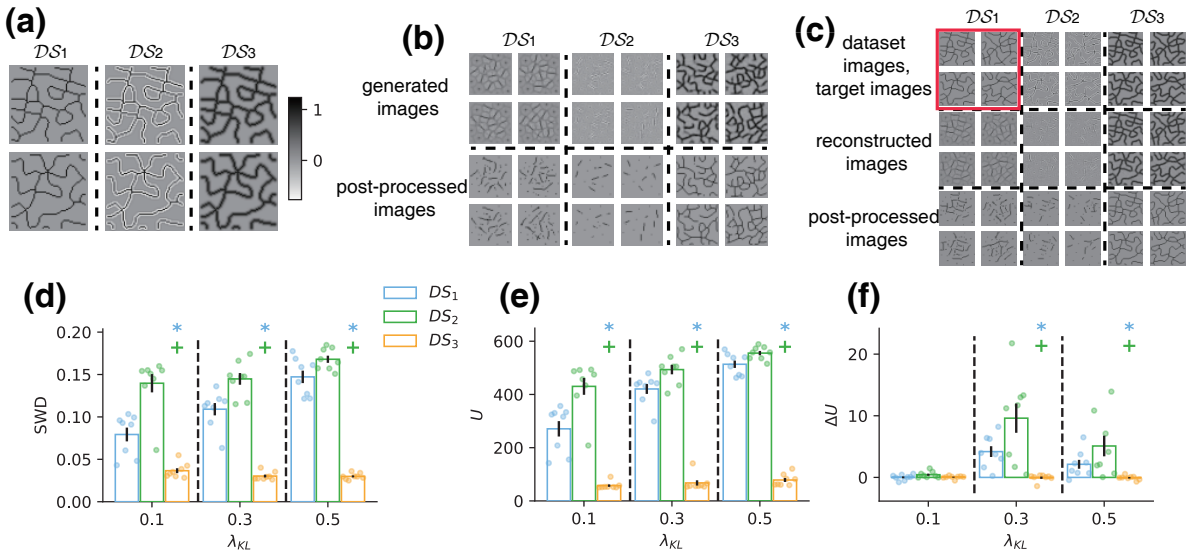

Figure 7: Low-precision generation facilitates the imagination and perception of fine-scale information in textures. (**a**) Examples of image patches synthesized from OpenSimplex2 noise. Image size 64×64. (**b**) Examples of VAE-generated images (upper row) and post-processed images (lower row). Note that $\mathcal{DS}_3$ results in the best imagination quality. (**c**) Examples of targeted images for perception (first row), reconstructed images (second row) and post-processed reconstructed images (third row). Note that $\mathcal{DS}_3$ results in the best perception quality. (**d**) SWD as a function of λ_{KL}, when VAE is trained by $\mathcal{DS}_1$, $\mathcal{DS}_2$ or $\mathcal{DS}_3$. Blue asterisk (or green plus sign) means that SWD for $\mathcal{DS}_3$ is significantly ($p < 0.05$) smaller than that for $\mathcal{DS}_1$ (or $\mathcal{DS}_2$) at the same λ_{KL}. (**e, f**) Similar to panel **d**, but for the perception error U (panel **e**) and the increase ΔU of U caused by perturbation of the bottleneck state (panel **f**). Error bars in panels **d-f** represent s.e.m. over 8 VAE configurations. All panels are results on OpenSimplex2 textures. See **Supplementary Information Section 3** and **Supplementary Figures 5, 6** for more results.

reconstructed images (**Figure 7c, 3rd row**) and the original images (i.e., $\mathcal{DS}_1$, **Figure 7c, red box**), with the bottleneck state of VAE before the iterating reconstruction process (**Figure 3a**) initialized to be the μ-channel of the VAE encoder. Perception robustness was quantified by the increase ΔU of U when the bottleneck state before a second iteration reconstruction was initialized at a randomly perturbed position from the optimal bottleneck state found during the calculation of U. We found that either SWD or U and ΔU for $\mathcal{DS}_3$ are lower than those for $\mathcal{DS}_1$ and $\mathcal{DS}_2$ (**Figure 7d-f**), which means that $\mathcal{DS}_3$-trained VAEs can generate images looking better like the original images (**Figure 7b**), and can perceive fine-scale information with better accuracy and robustness (**Figure 7c**).

We have examined SWD, U and ΔU for a broad range of parameters, including different sizes of images, different post-processing (**Figure 5b**) and VAE parameters, we have also examined another texture database treated from the Brodatz database [41]. We found that SWD, U and ΔU for $\mathcal{DS}_3$ tend to be lower than those for $\mathcal{DS}_1$ and $\mathcal{DS}_2$ in a broad range of parameters, see **Supplementary Figure 6**. These results further support that low-precision top-down generation facilitates the imagination and perception of fine-scale visual information.

The imagination and perception of the contours of CelebA faces

Contours are thin lines that depict the boundaries of objects in images, and are among the most important features associated with images. To further examine the advantage of low-precision top-down generation, we extract contours from the human faces in CelebA dataset [42] (**Figure 8a**, see Methods), and study the imagination and perception of these contours using VAE. Unlike the fact that all pixels on the skeleton lines in **Figures 5 and 7** take the same value, pixels on CelebA contours take variable values.

To infer the original contours ($\mathcal{DS}_1$, **Figure 8a**) from the blurred VAE-generated images (**Figure 8d, left column**), we passed the VAE-generated images through a feedforward deep neural network called post-processor (**Figure 8b, yellow box**); so that the imagination and perception performance of VAE can be quantified by comparing the original contours ($\mathcal{DS}_1$, **Figure 8a**) with the post-processed VAE-generated images (**Figure 8d, middle and right columns**). To train the post-processor, we inputted CelebA contours x_{input} into the VAE encoder, and let the post-processed VAE-output images $x_{\text{post-proc}}$ approach x_{input} by minimizing a loss function (**Figure 8c**). We examined two loss functions, a mean square error (MSE) loss and an MSE loss under adversarial regularization (see Methods). We found that with either loss function, VAE trained by blurred contours ($\mathcal{DS}_3$, **Figure 8a**) results in better imagination and perception than VAE trained by thin contours ($\mathcal{DS}_1$), see **Figure 8d-f**. Therefore, the advantage of low-precision top-down generation persists for CelebA contours, which are more naturalistic-looking fine-scale structures than those in **Figures 5 and 7**.

Discussion

In this paper, we show that top-down generation of low-precision representations improves the perceptive accuracy and robustness as well as imaginative creativity of visual information with fine spatial scale. This improvement results from two reasons: (1) low-precision representations of fine-scale visual information can be better reconstructed or simulated through the 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 to perceive or imagine 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 sparsified by an attention-based selection [43, 44], 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.

Implications of low-precision generation to brain cognition and machine learning

We have shown that generating low-precision neural representations facilitates the formation of neural representations in the high-order cortex that are smooth to shape morphing of fine-scale information (**Figures 2c, 5g**), which has implications for both brain cognition and machine learning.

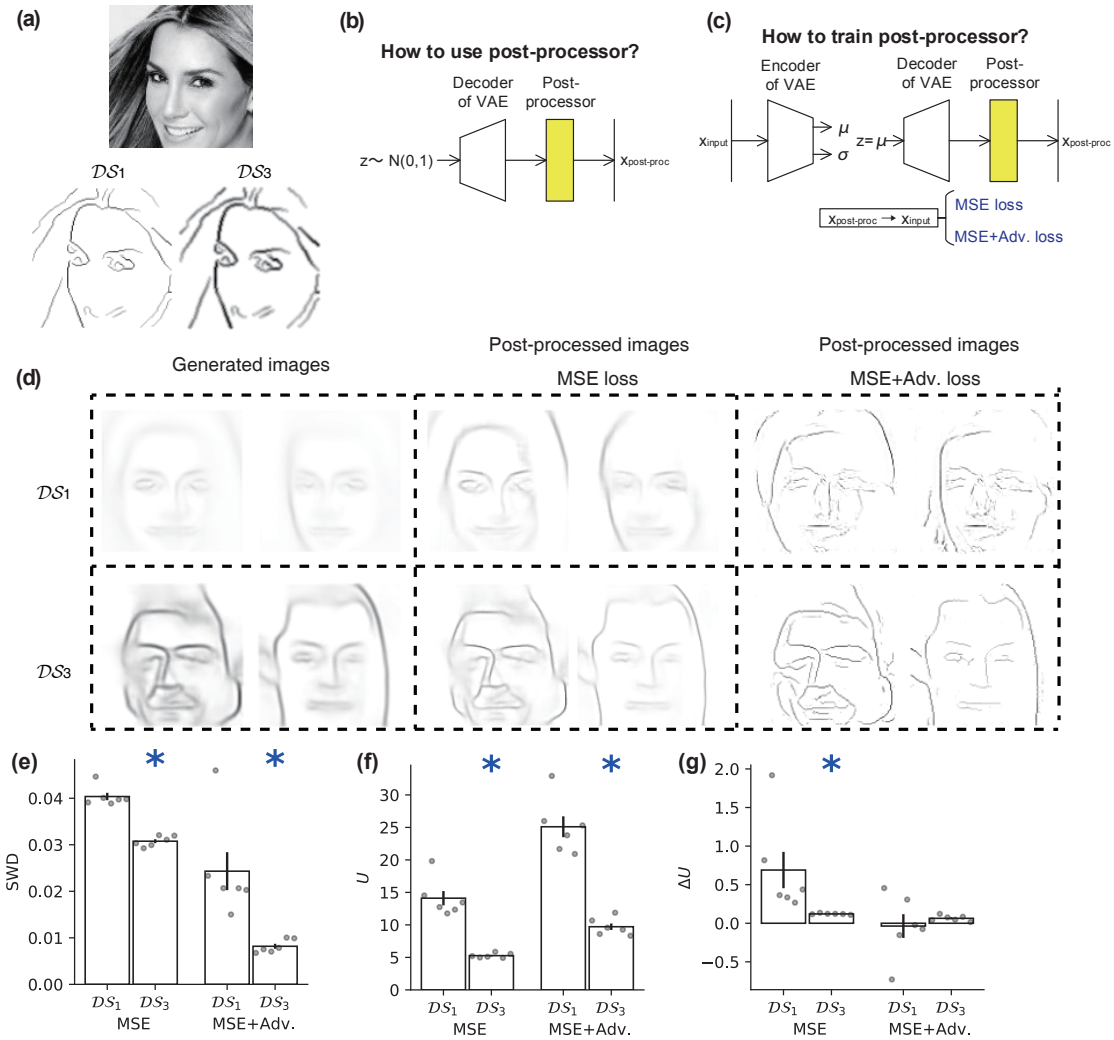

Figure 8: Low-precision generation facilitates the imagination and perception of CelebA contours. (**a**) A gray-scale CelebA image (top), its thin contours ($\mathcal{DS}_1$, bottom left) and its blurred contours ($\mathcal{DS}_3$, bottom right). (**b**) A feedforward network (yellow) is used to post-process the images generated by the decoder of VAE. (**c**) To train the post-processor (yellow), we let the bottleneck value of the VAE be the μ-channel of the encoder, and let the post-processed image $x_{\text{post-proc}}$ approach the input image x_{input} by minimizing MSE loss only or MSE loss under adversarial regularization (blue text). VAE is trained before the post-processor, and is kept unchanged during the training of the post-processor. (**d**) Examples of the VAE-generated images (left column), generated images post-processed by post-processors trained with MSE loss (middle column) and MSE loss under adversarial regularization (right column), when the VAE is trained using thin contours ($\mathcal{DS}_1$, upper row) or blurred contours ($\mathcal{DS}_3$, lower row). Notice that $\mathcal{DS}_3$ results in better post-processed images than $\mathcal{DS}_1$. (**e**) SWD between the post-processed images ($x_{\text{post-proc}}$ in panel **c**) and the thin contours ($\mathcal{DS}_1$) when the VAE is trained by $\mathcal{DS}_1$ or $\mathcal{DS}_3$ and the post-processor is trained by MSE loss only or MSE loss under adversarial regularization. Blue asterisk means that SWD for $\mathcal{DS}_3$ is significantly ($p < 0.05$) smaller than that for $\mathcal{DS}_1$ under the same loss function. (**f, g**) Similar to panel **e**, except for the perception error U (panel **f**) and perturbation-induced increase ΔU of perception error (panel **g**). In panels **f-g**, error bars represent s.e.m. over 6 VAE configurations, $\lambda_{KL} = 0.25$. See **Supplementary Figure 7** for the cases when $\lambda_{KL} = 0.5$ or 0.75.

Results of both brain and machine learning studies support the importance of such smooth shape-morphing representations for cognition. 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 representation in **Figure 2b** and the corresponding inferior performance in imagination and perception). Consistently, it has been found that human brain represents similar items using similar codes [45], and hippocampus activity also represents the geometry of both spatial and abstract knowledge [46]. In machine learning studies, representations smooth to shape morphing have been widely observed in neural network models that generate realistic-looking images [27, 47]. Such representations can be used to do algebraic operations on the semantic meanings of the generated images [48], and can also be exploited to design semi-supervised learning algorithms that train discriminators using only a few labeled data [49, 50].

It is well known that VAE, which tries to construct representations respecting the geometric structure of stimuli in the bottleneck state [51], tends to generate blurred images; vanilla auto-encoder, which does not constrain the structure of representations in the bottleneck state, does not tend to output blurred images [52]. Therefore, the low precision of the top-down generated representations may be a price the brain has to pay for the structured representations of the world in the high-order cortex, which are very important for cognition as discussed in the paragraph above.

Engineering neural networks to generate images with high-resolution details is a non-trivial task. VAE tends to generate blurred images [52]. Another well-known generative model, generative adversarial network (GAN), also struggles to generate high-resolution details [53] due to a gradient problem [54]. Our results suggest that blurring these high-resolution details may result in better imagination performance, as long as these high-resolution details can be later inferred from the generated low-resolution representations. As an example of this idea, the generated CelebA contours (**Figure 8**) can be later used to generate realistic-looking images by depicting colors or gray scales according to the contours [55]. Further implementation of this idea in engineering VAE or GAN is an exciting 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].

Implications on V1 computation

Sparse coding has long been believed as a computational principle of V1 [5]. Here we show that sparseness is required for fine-scale information to be inferred from top-down generated low-precision representations. In other words, for fine-scale information to be represented by the high-order cortex, it must be sparse. Besides the sparseness of V1 activity [5], sparseness can also be realized by attention. Attention not only selects salient or task-relevant sensory inputs [43, 44], sparsifying the fine-scale information for perception, but also selects features to imagine, sparsifying the features under anticipation or intention before actually sensing the stimuli [56, 57].

Our results also gain an understanding of V1 computation during predictive-coding perception. 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 (**Figures 3, 6-8**). Therefore, it is better to first transform the bottom-up stimulated high-precision representations to low precision before calculating the bottom-up and top-down mismatch and transmitting this mismatch feedforwardly for predictive-coding perception. Such high-to-low precision transformation can easily be done through feedforward projections [30]. It has been found that top-down prediction error is feedforwardly transmitted in a predominantly high (gamma) frequency range [58, 59], so the coding precision of prediction error may be understood by investigating the activities of V1 neurons 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 $\alpha \to 1^+$ [60]. The value of this exponent α ensures the differentiability of neural code, such that the V1 response smoothly changes with the position of a stimulus, and the distance between the V1 responses to two stimuli increases with the distance between the two stimuli [60]. Our results support the importance of code differentiability for imagination and perception. In Model 3, which results in the best imagination and

perception performance in the 1-dimensional-image case (**Figures 2, 3**), the Euclidean distance between two activity patterns gradually and monotonically increases with the distance between the two dot stimuli corresponding to the two patterns (**Figure 2e**). Similarly, in $\mathcal{DS}_3$, which results in the best performance in the 2-dimensional-image cases (**Figures 5-8**), the intensity of a pixel gradually and monotonically decays with the distance of this pixel from the center of a blurred line, such that the Euclidean distance between two textures $\mathcal{T}_1$ and $\mathcal{T}_2$ gradually and monotonically increases when $\mathcal{T}_2$ experiences gradual local shape morphing starting from $\mathcal{T}_1$. Therefore, the differentiability of V1 activity [60] may be closely related to the top-down generative function of the visual system, which may be understood as follows. Suppose that the high-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 **Figure 2c**), $v_1(x)$ must also be differentiable to x to better fulfill the equation $v_1(x) \approx f(z(x))$.

2 Methods and Materials

The imagination and perception of 1-dimensional images

VAE structure

A variational auto-encoder (VAE) [27] contains an encoder and a decoder (**Figure 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 **Figures 2 and 3**, the encoder is a multilayer perceptron (MLP) with three hidden layers with sizes 100, 50 and 20 respectively. The input layer is of size 201, and the output layer has two channels each with size 1. The decoder is another MLP that has three hidden layers with sizes 20, 50 and 100 respectively. Adjacent layers are all-to-all connected. We use leaky relu as the activation functions of the hidden layers. VAE was trained using Adam optimizer [61] using Pytorch.

The three models of 1-dimensional images

In the three models in **Figure 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 (**Figure 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

We examined the quality of perception using two indexes r_1^2 and r_2^2 (**Figure 3**). This examination was guided by the intuition that two conditions are required for good perception performance: (1) there exists a state z_0 at which the generated image well resembles the target image; and (2) the representations of the target images 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 (**Figure 3a**).

Guided by these 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 image $\mathcal{T}$ (i.e., blue curve in **Figure 3a**) to be an activity pattern of Model a, and initialized the bottleneck state to be the value of the μ-channel (see **Figure 1a**) of the VAE encoder when inputting $\mathcal{T}$ into the encoder. The value of the μ-channel is the best estimation of the optimal bottleneck state 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 image $\mathcal{T}$ 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 optimized the bottleneck state again.

We defined r_1^2 and r_2^2 as the coefficient of determination to quantify how mu ch variance of the target image can be explained by the generated image. In the above two steps, the bottleneck state was optimized to minimize the error between the target and generated patterns using Adam optimizer [61].

The imagination and perception of skeleton MNIST images

VAE structure

The VAE used in **Figures 5 and 6** has a similar structure to that used in **Figures 1-3**, except that the size of the input and output layers is $28 \times 28 = 784$, and the sizes of the three hidden layers of the encoder (or decoder) are 512, 256 and 128 (or 128, 256 and 512) respectively. The size of each of the two output channels of the encoder is 20.

The three datasets of 2-dimensional images

The dataset $\mathcal{DS}_1$ in **Figure 5a** is the skeleton MNIST dataset [32]. The intensity of a pixel in an image in $\mathcal{DS}_1$ is 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 (**Figure 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 an MLP to classify the skeleton MNIST dataset (i.e., $\mathcal{DS}_1$). This MLP contained a hidden layer of size 1000 with a 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 **Figure 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 **Figure 5e**, $\mathcal{H}_{xcat} = -\sum_x E_I[p(x|I)] \ln E_I[p(x|I)]$ [36]. To plot **Figure 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 $\lambda_i(x)$ along 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]. **Figure 5d-f** shows 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 a very small value. Because of this artifact, in **Figure 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. This artifact was weak for $\mathcal{DS}_3$ in our range of parameters, so we plotted the whole parameter range for $\mathcal{DS}_3$.

Figure 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

The performance of a VAE to perceive a skeleton MNIST image was quantified in the following two steps:

(1) Given an image I_{origin} in $\mathcal{DS}_1$, we let the target image I_t be I_{origin} itself 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 set the initial bottleneck state to be the μ-channel of the VAE encoder in response to I_t, and optimized the bottleneck state to minimize the mean square error between I_t and the VAE-generated image I_g. I_g was then post-processed by binarizing and skeletonizing, resulting in an image $I_{postprocess}$. The perception quality was then quantified by (**Figure 6b**)

$$U = \langle \sum_i (\text{Blur}(I_{origin})_i - \text{Blur}(I_{postprocess})_i)^2 \rangle_{I_{origin}}, \tag{3}$$

where both $I_{postprocess}$ and I_{origin} are made up of 1-pixel-width lines, and $\text{Blur}(\cdot)_i$ means the ith pixel of a blurred image. This blurring operation was performed using the same method as we produced $\mathcal{DS}_3$ from $\mathcal{DS}_1$. To understand the reason for this blurring operation, suppose two 1-pixel-width lines are taking the same direction in I_{origin} and $I_{postprocess}$ respectively. In this case, $\sum_i (I_{origin,i} - I_{postprocess,i})^2$ is zero if these two parallel lines overlap, but takes a positive value if these two parallel lines do not overlap. This positive value remains unchanged with the distance between these two parallel lines. By blurring I_{origin} and $I_{postprocess}$, U gradually increases with the distance between these two parallel lines, which is consistent with the intuition that the perception error increases with the distance between these two lines.

(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 I_g and I_t, and then calculated U using eq. 3 again. **Figure 6c** plots the increase of U value in this second step from that in the first step. Some data points for $\mathcal{DS}_1$ (when $\lambda_{KL} = 0.9$) and $\mathcal{DS}_2$ (when $\lambda_{KL} = 0.5, 0.7, 0.9$) are missing in **Figure 6c** and **Supplementary Figure 4b**. The reason is that for $\mathcal{DS}_1$ and $\mathcal{DS}_2$, the reconstruction quality is so bad when λ_{KL} takes large values, such that the perception error U already takes large values (**Figure 6b**). In this case, noise perturbation can hardly increase U any more, which makes ΔU close to zero. However, this close-to-zero ΔU does not imply perception robustness, but is instead an artifact in the case of very bad perception. To avoid possible confusions of the readers and let ΔU faithfully represent perception robustness, we deleted these data points in **Figure 6c** and **Supplementary Figure 4b**.

The imagination and perception of texture images

Database preparation

In **Figure 7** and **Supplementary Figures 5, 6**, we studied the generation of textures using two databases: one synthesized from OpenSimplex2 noise [38], and the other from the Brodatz database [41].

The OpenSimplex2 textures were generated using the FastNoiseLite package with the following parameters. Noise type: OpenSimplex2; frequency: 0.05; fractal type: ridged; fractal octaves: 1; fractal lacunarity: -0.29; fractal gain: 0.05; fractal weighted strength: 1.3; domain warp type: OpenSimplex2; domain warp amplitude: 5. The generated images were then binaried through threshold 0.5, and skeletonized by the 'skeletonize' routine of skimage python package, resulting in 60000 images of size 64×64.

The Brodatz textures were generated by treating normalized Brodatz's textures [41] using the command '1-edge(imgaussfilt(data, 2), 'Canny')' of Matlab, binarizing through threshold 0.5, and then skeletonizing, resulting in 112 images of size 640×640.

Image patch preparation

We extracted image patches of different sizes (8×8, 16×16, 32×32 and 64×64) at random positions from the two databases respectively. For each image patch, we set the intensities of the pixels whose distances with boundaries were 1 or 2 to be zero (see below for the reason for this operation), resulting in dataset $\mathcal{DS}_1$. Datasets $\mathcal{DS}_2$ and $\mathcal{DS}_3$ were derived from $\mathcal{DS}_1$ similar to those for skeleton MNIST images.

The quality of imagination

We post-processed the VAE-generated images by binarizing and skeletonizing, and then compared the obtained images of 1-pixel-width lines with the original images in $\mathcal{DS}_1$ to evaluate the imagination quality of VAE. The skeletonization operation (i.e., the 'skeletonize' routine of skimage python package) has an artifact: if both a pixel at the boundary of an image and a nearby pixel not at the boundary take 1 value after binarization, then the boundary pixel is more likely to belong to a 1-pixel-width line after skeletonization than the non-boundary pixel. To avoid this boundary artifact, we need to reduce the intensities of the pixels in the VAE-generated images at the boundary, so that they are unlikely to take 1 value after binarization. To this end, as we mentioned before, we set the intensities of the pixels whose distances with boundaries were 1 or 2 to be zero in dataset $\mathcal{DS}_1$, before $\mathcal{DS}_2$ and $\mathcal{DS}_3$ were derived from $\mathcal{DS}_1$ and used to train VAE.

The imagination quality was quantified using sliced Wasserstein distance (SWD) [39, 40] between the post-processed VAE-generated images and $\mathcal{DS}_1$. SWD is defined by

$$\text{SWD} = \langle \sum_j |\text{sort}_j(\{\boldsymbol{x} \cdot \boldsymbol{n}\}_{\text{batch } \boldsymbol{x}}) - \text{sort}_j(\{\tilde{\boldsymbol{x}} \cdot \boldsymbol{n}\}_{\text{batch } \tilde{\boldsymbol{x}}})| \rangle_{\boldsymbol{n}}, \tag{4}$$

where $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ are respectively the post-processed VAE-generated images and $\mathcal{DS}_1$, $\boldsymbol{n}$ is a randomly selected direction vector with $\|\boldsymbol{n}\|^2 = 1$, $\text{sort}_j(\{\boldsymbol{x} \cdot \boldsymbol{n}\}_{\text{batch } \boldsymbol{x}})$ means the jth element after sorting the value $\boldsymbol{x} \cdot \boldsymbol{n}$ in a batch of $\boldsymbol{x}$, $\langle \cdot \rangle_{\boldsymbol{n}}$ means averaging over $\boldsymbol{n}$. Numerically, the batch sizes of both $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ were 1024, and we averaged over 512 directions of $\boldsymbol{n}$ when calculating SWD.

The quality of perception

The accuracy and robustness of perception were respectively quantified by U defined in eq. 3 and the increase ΔU of U after optimizing the perturbed bottleneck state, similarly to that for skeleton MNIST images.

The network structure of VAE

The VAEs to generate texture images (**Figure 7**) were deep convolutional neural networks with batch normalization and leaky relu activation function following each convolutional layer. In the encoder part of VAE, convolutional kernels had spatial size 4 and stride 2, such that the spatial size of the signal was reduced by a half after each convolutional layer. In the decoder part, transposed convolutional kernals also had

spatial size 4 and stride 2, such that the spatial size of the signal was increased by a time after each transposed convolutional layer. For VAEs used to generate images of size 8×8, 16×16, 32×32 and 64×64, there were 2, 3, 4 and 5 convolutional (or transposed convolutional) layers in the encoder (or decoder) respectively. In the 1st, 2nd, 3rd, 4th and 5th convolutional layers in the encoder (or the last 1st, 2nd, 3rd, 4th and 5th transposed convolutional layers in the decoder), there were 64, 128, 256, 512 and 512 channels respectively (note that some high-order layers did not exist in the VAE if the image size was smaller than 64×64). The bottleneck state has dimensions 8, 16, 32 and 64 respectively.

The imagination and perception of CelebA contours

Contour detection

We converted each CelebA image into gray-scale, and cropped the resulted gray-scale image around the center, resulting in a 128×128 image patch. Then, we normalized the mean of each patch to be the mean of all the obtained patches. To detect contours from a normalized gray-scale patch, we followed the standard technique (see https://learnopencv.com/edge-detection-using-opencv/): we first blurred the patch using Gaussian filter with a standard deviation of 2 pixels (we denote the blurred patch as $\mathcal{I}_{\text{blur}}$ below), then used the 'Canny' routine of OpenCV (with parameters threshold1=50, threshold2=100, L2gradient=True) to obtain binary contours corresponding with $\mathcal{I}_{\text{blur}}$. We then skeletonized the binary contours using the 'skeletonize' routine of the skimage package, resulting in binary contours of 1-pixel width, denoted as $\mathcal{I}_{\text{binary}}$ below. An image $\mathcal{I}_{DS1}$ in $\mathcal{DS}_1$ dataset (**Figure 8a, right panel**) was obtained so that if a pixel in $\mathcal{I}_{\text{binary}}$ had zero value, then the corresponding pixel in $\mathcal{I}_{DS1}$ also had zero value; if a pixel in $\mathcal{I}_{\text{binary}}$ had non-zero value, then the value of the corresponding pixel in $\mathcal{I}_{DS1}$ was the magnitude of the intensity gradient at the corresponding pixel in $\mathcal{I}_{\text{blur}}$, calculated using the 'Sobel' routine of OpenCV. Images in $\mathcal{DS}_3$ dataset were obtained by blurring images in $\mathcal{DS}_1$, similar to that for processing skeleton MNIST images introduced above.

Training the post-processor

To train the post-processor, we inputted images x_{input} in $\mathcal{DS}_1$ or $\mathcal{DS}_3$ into the VAE encoder, let the bottleneck state of the VAE be the μ-channel of the encoder, then passed the output images of VAE through the post-processor, getting $x_{\text{post-proc}}$. We adjust the synaptic weights in the post-processor by minimizing the mean-square error (MSE) loss function

$$l_{\text{mse}} = \sum_i (x_{\text{input},i} - x_{\text{post-proc},i})^2 \tag{5}$$

with $x_{\text{input},i}$ being the ith pixel in x_{input}. We then used further trained the post-processor by minimizing an MSE loss function regularized by an adversarial loss

$$l_{\text{mse+adv}} = \sum_i (x_{\text{input},i} - x_{\text{post-proc},i})^2 + \lambda_{\text{adv}} D(x_{\text{post-proc}}, \text{label}_{\text{real}}), \tag{6}$$

where $\lambda_{\text{adv}} = 0.01$, $D(x_{\text{post-proc}}, \text{label}_{\text{real}})$ is the binary cross entropy between the output of a discriminator network and the label indicating that $x_{\text{post-proc}}$ comes from $\mathcal{DS}_1$. The discriminator network was trained to better discriminate $x_{\text{post-proc}}$ and the images in $\mathcal{DS}_1$. This discriminator network is a 6-layer convolutional network with leaky relu activation function [47], with the spatial size of the signal was reduced by half after each convolutional layer.

The network structure of VAE

The network structure for the VAE for generating CelebA contours was similar to that for generating textures. The spatial size of the signal was reduced by half after each convolutional layer in the encoder, and increased by a time after each transposed convolutional layer in the decoder. The bottleneck state has dimension 128.

The network structure of post-processor

The post-processor (yellow boxes in **Figure 8b, c**) had a U-Net structure [62]. The spatial size of the signal was reduced by half after each convolutional layer in the encoder, and increased by a time after each transposed convolutional layer in the decoder. There was a direct link between the ith (counting from the input size) convolutional layer of the encoder and the ith (counting from the output size) transposed convolutional layer of the decoder. The bottleneck state has dimension 128.

The quality of imagination and perception

The quality of imagination was quantified by SWD defined in eq. 4, in a similar way to that for texture images.

The accuracy and robustness of perception were respectively quantified by U defined in eq. 3 and the increase of U after optimizing the perturbed bottleneck state, in a similar way to that for texture images.

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

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

$$\Delta N = \Delta\theta_{stim} N/2\pi \tag{7}$$

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

$$W_{forward}(\theta_{stim}, \theta_{low}) = C \exp(-\frac{(\min(2\pi - |\theta_{stim} - \theta_{low}|, |\theta_{stim} - \theta_{low}|))^2}{2\sigma^2}), \tag{8}$$

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 tried a matching-pursuit algorithm [31] in **Figure 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 $\hat{x}_i \leftarrow \hat{x}_i + c_i$, $r_i \leftarrow r_i - \sum_{j=1}^{N} c_j (W_{forward})_{ji}$.

3 Acknowledgments

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

References

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

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

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

[4] Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The helmholtz machine. *Neural Comput.* **7**, 889–904 (1995).

[5] 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).

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

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

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

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

[10] Kanai, R. *et al.* Information generation as a functional basis of consciousness. *Neurosci. Conscious.* **5**, niz016 (2019).

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

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

[13] van Vugt, B. *et al.* The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360**, 537–542 (2018).

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

[15] Mudrik, L., Faivre, N. & Koch, C. Information integration without awareness. *Trends Cogn. Sci.* **18**, 488–496 (2014).

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

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

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

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

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

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

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

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

[24] Angelucci, A. *et al.* Circuits and mechanisms for surround modulation in visual cortex. *Annu. Rev. Neurosci.* **40**, 425–451 (2017).

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

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

[27] Kingma, D. & Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations* (2014).

[28] Murray, J. D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl Acad. Sci. USA* **114**, 394–399 (2017).

[29] Parthasarathy, A. *et al.* Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex. *Nat. Commun.* **10**, 4995 (2019).

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

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

[32] de Jong, E. D. Incremental sequence learning. *arXiv:1611.03068* (2016). URL `https://github.com/edwin-de-jong/mnist-digits-stroke-sequence-data/wiki/MNIST-digits-stroke-sequence-data`.

[33] Li, W., Piëch, V. & Gilbert, C. D. Contour saliency in primary visual cortex. *Neuron* **50**, 951–962 (2006).

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

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

[36] Salimans, T. *et al.* Improved techniques for training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2234–2242 (2016).

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

[38] Perlin, K. Noise hardware. In Olano, M. (ed.) *Real-Time Shading SIGGRAPH Course Notes* (2001).

[39] Rabin, J., Peyr, G., Delon, J. & Bernot, M. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision (SSVM)*, 435–446 (2011).

[40] Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *The International Conference on Learning Representations* (2018).

[41] Safia, A. & He, D.-C. New brodatz-based image databases for grayscale color and multiband texture analysis. *ISRN Machine Vision* **21**, 876386 (2013).

[42] Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)* (2015).

[43] Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).

[44] Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).

[45] Fang, Y. *et al.* Semantic representation in the white matter pathway. *PLoS Biol.* **16**, e2003993 (2018).

[46] Nieh, E. H. *et al.* Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).

[47] Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *The International Conference on Learning Representations* (2016).

[48] Larsen, A. B. L., Sϕnderby, S. K., Larochelle, H. & Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of Machine Learning Research* (2016).

[49] Kingma, D. P., Rezendey, D. J., Mohamedy, S. & Welling, M. Semi-supervised learning with deep generative models. In *Conference on Neural Information Processing Systems* (2014).

[50] Abbasnejad, M. E., Dick, A. & van den Hengel, A. Infinite variational autoencoder for semi-supervised learning. In *Conference on Computer Vision and Pattern Recognition* (2017).

[51] Higgins, I. *et al.* beta-VAE: Learning basic visual concepts with a constrained variational framework. In *The International Conference on Learning Representations* (2017).

[52] Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (The MIT Press, Cambridge, 2016).

[53] Odena, A., Olah, C. & Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning* (2017).

[54] Arjovsky, M. & Bottou, L. Towards principled methods for training generative adversarial networks. In *The International Conference on Learning Representations* (2017).

[55] Dekel, T., Gan, C., Krishnan, D., Liu, C. & Freeman, W. T. Sparse, smart contours to represent and edit images. *arXiv:1712.08232* (2018).

[56] Nobre, A. C. & van Ede, F. Anticipated moments: temporal structure in attention. *Nat. Rev. Neurosci.* **19**, 34–48 (2018).

[57] Wolfe, J. M. & Horowitz, T. S. Five factors that guide attention in visual search. *Nat. Hum. Behav.* **1**, 0058 (2017).

[58] Arnal, L. H., Wyart, V. & Giraud, A.-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. Neurosci.* **14**, 797–801 (2011).

[59] Bastos, A. M. *et al.* Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* **85**, 390–401 (2015).

[60] Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).

[61] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations* (2015).

[62] Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597* (2015).