Generative perspective of the primary visual cortex

Zedong Bi

Institute for Future, Qingdao University, Shandong 266071, China

Email: zedong.bi@outlook.com

Abstract

According to analysis-by-synthesis theories of perception, the primary visual cortex (V1) reconstructs visual stimuli through top-down pathway, and higher-order cortex reconstructs V1 activity. Experiments also found that neural representations are generated in a top-down cascade during visual imagination. What code does V1 provide higher-order cortex to reconstruct or simulate to improve perception or imaginative creativity? What unsupervised learning principles shape V1 for reconstructing stimuli so that V1 activity eigenspectrum is power-law with close-to-1 exponent? Using computational models, we reveal that reconstructing the activities of V1 complex cells facilitate higher-order cortex to form representations smooth to shape morphing of stimuli, improving perception and creativity. Power-law eigenspectrum with close-to-1 exponent results from the constraints of sparseness and temporal slowness when V1 is reconstructing stimuli, at a sparseness strength that best whitens V1 code and makes the exponent most insensitive to slowness strength. Our results provide fresh insights into V1 computation.

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 synaptic weights [3, 4] or the neuronal dynamic states [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.

According to the generative perspective above, V1 plays two fundamental cognitive roles: (1) V1 provides code for higher-order cortex to reconstruct for perception or to simulate for creativity, and (2) V1 reconstructs the external stimuli for perception. In this paper, we would like to ask two questions regarding to these two roles: (1) What code does V1 provide higher-order cortex to reconstruct or simulate in order to improve perception or creativity? (2) What unsupervised learning principles shape the activity of V1 when V1 is reconstructing stimuli, so that the statistics of V1 code resembles experimental observations?

Since Hubel and Wiesel proposed their V1 circuit model that complex cells receive from simple cells [10] (see Ref. [11] for experimental validation), most modeling works suppose complex cells to be the output of V1 (e.g., Ref. [12, 13]). Deep convolutional neural network, which consists of alternate stacking of convolution and pooling layers and can be regarded as an engineering realization of the stacking of the Hubel-Wiesel circuits, has seen a huge success in artificial intelligence [14]. It is believed that the computational function of complex

cells is to establish code invariant to local spatial translation, which benefits object recognition performed in down-stream cortices [15, 12]. No studies, as far as we know, however, address the computational function of complex cells from top-down generative perspective.

Understanding the unsupervised learning principles that guide the development of V1 has been a fruitful research direction. It has been shown that sparse coding results in Gabor-like receptive fields of simple cells [4], temporal slowness may guide the formation of complex cells [16], temporal prediction can be used to understand the spatio-temporal receptive fields of simple cells [17]. Recent experiment showed that the variance of the nth principal component of the activity of V1 decays in power law $n^{-\alpha}$ with $\alpha \to 1^+$ [18], which provides a new challenge for computational explanation. This eigenspectrum may result from the compromise between efficient and smooth coding [18], but how this compromise can be realized through biologically plausible unsupervised learning principles in neural networks remains unknown.

In this paper, we addressed the above two problems by training neural network models. By modeling the visual system as a variational auto-encoder [19], we show that reconstructing the activities of complex cells facilitates higher-order cortex to form representations smooth to shape morphing of stimuli, thereby improving visual creativity and perception. Using a parsimonious generative model in which V1 is continuously reconstructing the input temporal stimuli, we show that coding sparseness and temporal slowness together explain the power-law eigenspectrum of V1 activity. The close-to-1 exponent is realized at a sparseness strength that best whitens V1 code and makes the exponent most insensitive to slowness strength. Our results provide fresh insights into V1 computation.

Results

Understanding the top-down function of complex cells: toy models

Model setup

Our V1 circuit model is a two-layer network receiving from stimuli on a one-dimensional line (**Fig. 1b, right panel**). The feedforward connection W_2 (magenta arrow in **Fig. 1b, right panel**) between the simple cell preferring stimulus at position X_{sim} and the complex cell preferring X_{com} Gaussianly decays with $|X_{sim} - X_{com}|$. Experiments showed that lateral inhibition sharpens the tuning of simple cells [20, 21, 22]. We model this tuning sharpening by a winner-take-all (WTA) effect, so that only the simple cell with the highest activity remains active, while all the others are silent.

To study the computational function of complex cells, we asked two questions: (1) what will happen if the complex cells are removed, leaving only simple cells; (2) why V1 is functionally like a two-layer structure (with simple cells and complex cells), instead of a simple linear filter? To answer these questions, we compared three toy models: (Model 1) a single-layer model with lateral inhibition (which results in WTA) to model the simple-cell-only case (**Fig. 1b, left panel**); (Model 2) a single-layer model without lateral inhibition to model the case of a linear filter (**Fig. 1b, middle panel**); and (Model 3) a two-layer model with both simple and complex cells (**Fig. 1b, right panel**), which is the model we introduced above. In all the three models, the stimuli are wavelets parameterized by their positions on a one-dimensional line, and the feedforward connection W_1 as a function of $X_{stim} - X_{sim}$ (where X_{stim} and X_{sim} are respectively the preferred stimulus positions of a stimulus receptor and a simple cell) is also in the shape of the wavelet (**Fig. 1a**). In response to a stimulus, the V1 output is a delta peak in Model 1, an oscillating wavelet in Model 2, and a Gaussian bump in Model 3 (**Fig. 1c**).

We model the visual system as a variational auto-encoder (VAE) [19], whose output is optimized to reconstruct the input, under the regularization that the state of the bottleneck layer is encouraged to distribute close to standard normal distribution in response to any input (**Fig. 1d**). 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 [23, 24]; (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) stochasticity 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 Gaussian distribution, consistent with the experimental finding that V1 participates in mental imagination of new images [9]. This fourth point enables us to study the creativity of the visual system using VAE.

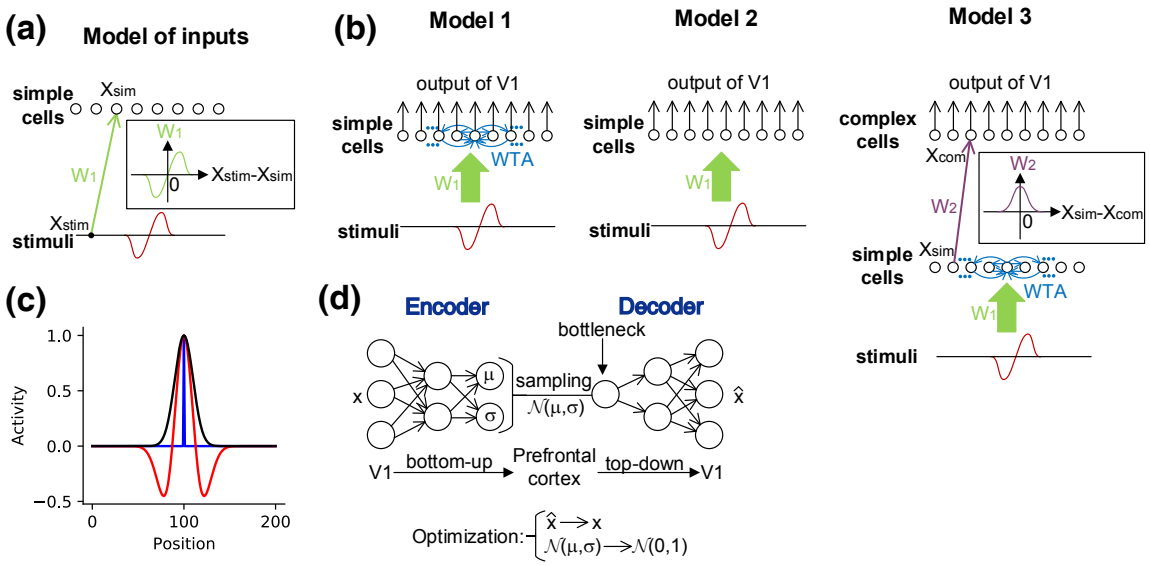

Figure 1: Schematic of the toy models manifesting the top-down function of complex cells. (**a**) Schematic of the input model. Stimulus input (red) is a wavelet along a one-dimensional line. The feedforward connections W_1 (green) as a function of the difference $X_{stim} - X_{sim}$ of the preferences of the stimulus receptors and simple cells share the same wavelet form as the stimuli. (**b**) Schematic of the three V1 models we study. Model 3 has both complex cells and simple cells with lateral inhibition (blue arrows), with the connection W_2 from a simple cell to a complex cell Gaussianly decays with the difference between their stimulus preferences. Model 1 has only simple cells. Model 2 also has only simple cells, but without lateral inhibition. (**c**) Examples of V1 output in Model 1 (blue), Model 2 (red) and Model 3 (black). (**d**) 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 $\mathcal{N}(0, 1)$.

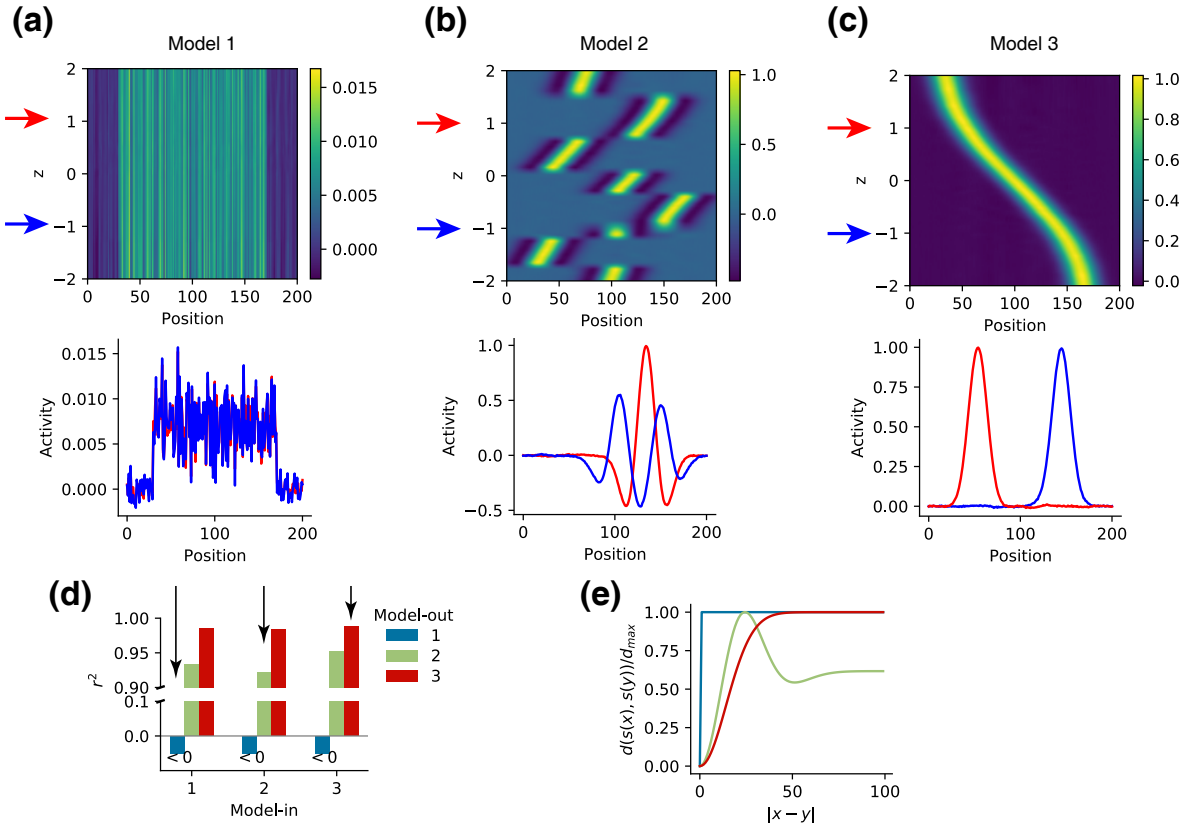

Figure 2: Complex cells facilitate creativity, toy models. (**a**) Upper panel: generated activity patterns as a function of the state z of the bottleneck variable. Lower panel: blue and red curves respectively represent the two generated patterns when z takes the two values (-1 and 1) indicated by the red and blue arrows in the upper panel. VAE is trained using the output of Model 1. (**b, c**) Similar to panel **a**, except that VAE is trained using the output of Model 2 (panel **b**) or 3 (panel **c**). (**d**) We let the input of VAE (i.e., x in **Fig. 1d**) be the output of Model-in, but train the output of VAE (i.e., $\hat{x}$ in **Fig. 1d**) to approach the output of Model-out. Both Model-in and Model-out can be Model 1, 2 or 3. r^2 quantifies how well a generated pattern by VAE looks like an output pattern of Model-out. 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** and **e**. (**e**) Euclidean distance $d(s(x), s(y))$ between two output patterns s as a function of the distance $|x - y|$ between the positions x and y of two stimuli, normalized by the maximal value of d, for Models 1 (blue), 2 (green) and 3 (red). The bottleneck state of VAE has only 1 dimension.

Visual creativity

To study the function of complex cells in visual creativity, we trained VAEs to generate the output patterns of V1 in the three models above (**Fig. 1b, c**), and compared the quality of the generated patterns in the three models. After training, we investigated how the generated pattern changed with the bottleneck state z, which is one-dimensional in our model. 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 the peaking positions sometimes happen with the continuous change of z (**Fig. 2b**). In Model 3, Gaussian peaks are generated, with the peaking position smoothly changing with z (**Fig. 2c**).

To quantify the quality of the generated patterns, we defined an index r^2 to quantify how well a generated pattern by VAE looks like an output 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**). Therefore, VAE trained by Model 3 learns to generate patterns looking like the output of Model 3. This advantage of Model 3 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 (blue curve in **Fig. 2b, lower panel**).

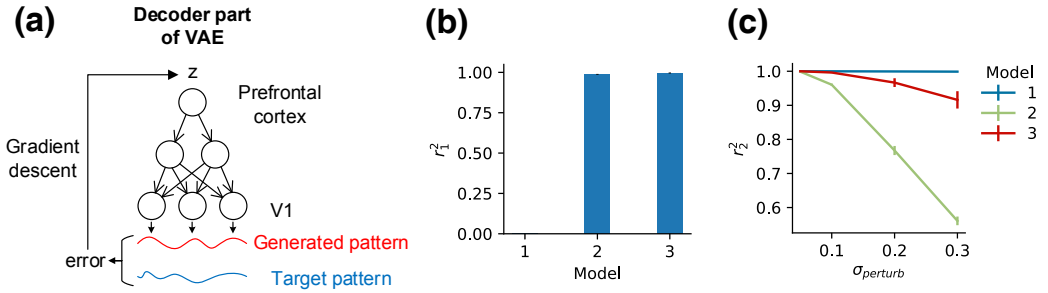

Figure 3: Complex cells facilitate perception, toy models. (**a**) Schematic of the perception model. Only the decoder of VAE after training is used. The bottleneck state z is updated by gradient descent to minimize the error between the generated pattern (red) and the target pattern (blue). (**b**) 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 represent s.e.m..

In the study above, we input $\mathcal{P}_a$ to VAE and trained VAE to construct $\mathcal{P}_a$ in the output, with $\mathcal{P}_a$ being the output pattern of Model a ($a = 1, 2, 3$). Now we ask whether the advantage of Model 1 results from a 'good' code that higher-order cortex receives from V1, or from a 'good' code that higher-order cortex is to reconstruct through the top-down pathway. To answer this question, we input $\mathcal{P}_a$ to VAE but trained VAE to construct $\mathcal{P}_b$ ($b \neq a$). We found that quality of the generated images strongly depended on $\mathcal{P}_b$, but hardly on $\mathcal{P}_a$ (**Fig. 2d**). Therefore, the advantage of complex cells 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 then studied the Euclidean distance $d(s(x), s(y))$ between the output patterns s as a function of $|x - y|$, where x and y are the positions of two stimuli 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 output patterns s_1 and s_2, which corresponds to two stimuli at positions x_1 and x_2 respectively. For simplicity, 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 encourages s_2 to be closer to s_1, so that x_2 gets closer to x_1, which means that the bottleneck state represents the spatial topology (i.e., the translational movement) of the stimuli. In Model 3, this can be realized using the gradient $\frac{\partial d(s_1,s_2)}{\partial s_2}$. In Model 2, $d(s_1, s_2)$ is not monotonic with s_2, so $\frac{\partial d(s_1,s_2)}{\partial s_2}$ 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 [5], higher-order cortex adjusts its state using the error of its recontruction of the activity of lower-order cortex. In our model, perception is performed by adjusting the bottleneck state z to minimize the error between the generated pattern 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 complex-cell pattern; and (2) the representation of the complex-cell 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. First, we set the target pattern $\mathcal{T}$ (i.e., blue curve in **Fig. 3a**) to be a complex-cell 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 $\hat{z}_0$. Second, 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 noise 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 a bottleneck state can be to reconstruct a complex-

cell pattern. r_2^2 quantifies how smooth the representation of the output in the bottleneck state can be so that an 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**). These results suggest the advantage of complex cells for visual perception. The low r_2^2 value for Model 2 may be closely related to the fragmentary representation in the bottleneck state (**Fig. 2b, upper panel**).

Understanding the top-down function of complex cells: skeleton MNIST dataset

Model setup

We then studied the advantage of complex cells for top-down image generation using skeleton MNIST dataset [25]. Images in this dataset represent digits using lines of 1 pixel width (**Fig. 4a, second column**). We used this dataset (denoted as dataset $\mathcal{DS}_1$ below) to displace the output of Model 1 (**Fig. 1b, left**) for 2-dimensional-image case, because a 1-pixel-width line is analogous to the activities of simple cells under local winner-take-all inhibition along the direction perpendicular to the line (comparing the second column of **Fig. 4a** with the blue peak in **Fig. 1c**). Biologically, simple cells are selective to the orientation of bars. Our model does not has the ingredient of orientation selectivity, but uses images of thin lines, mimicking the enhanced representation of contours in V1 due to collinear facilitation [26]. Our model also shares a property with the simple cells with sharpened orientation tuning due to lateral inhibition [20]: the representations of two parallel bars hardly overlap with each other. To displace Model 2 (**Fig. 1b, middle**) for 2-dimensional case, we set the pixel intensities oscillatorily decaying along the direction perpendicular to a line in the skeleton MNIST images (dataset $\mathcal{DS}_2$, **Fig. 4a, third column**). To displace Model 3 (**Fig. 1b, right**), we blurred the skeleton images so that pixel intensities Gaussianly decayed along the direction perpendicular to a line (dataset $\mathcal{DS}_3$, **Fig. 4a, fourth column**). We trained VAE using the three datasets, and compared the quality of the generated images (see Methods).

Visual creativity

The images generated by VAE trained by different datasets have different styles (**Fig. 4c, upper panels**). To fairly compare the quality of the generated images by the three datasets, we post-processed the generated images by first binarizing 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 skeletonizing the images (see Methods), resulting in binary images with lines of 1-pixel width (**Fig. 4b and lower panels of c**), similar to the images in the skeleton MNIST dataset.

People have proposed that creativity is a mental process of generating worthwhile and novel ideas [27]. Inspired by this definition of creativity, we propose that good generation of skeleton-MNIST images should satisfy three conditions. (1) Realisticity: a generated image should look like a digit in the skeleton MNIST dataset. (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 skeleton MNIST dataset, 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 [28]. 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 [28]. To quantify within-category variety, we calculated the intrinsic dimensionality $D_{incat} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}$ [29], 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. 4b**) and a parameter λ_{KL} in VAE which controlled the regularization strength onto the distribution of the bottleneck state variable (see Methods). We found that VAEs trained by dataset $\mathcal{DS}_3$ generated images with better eye-looking quality than VAEs trained by dataset $\mathcal{DS}_1$ and $\mathcal{DS}_2$ (**Fig. 4c**), consistent with the quantitive indication of smaller $\mathcal{H}_{real}$, larger $\mathcal{H}_{xcat}$ and larger D_{incat} in large range of parameters θ_{thres} and λ_{KL} (**Fig. 4d-f**). These results suggest that complex cells facilitate the visual system to generate diverse realistic-looking images, supporting the functional role of complex cells in visual creativity.

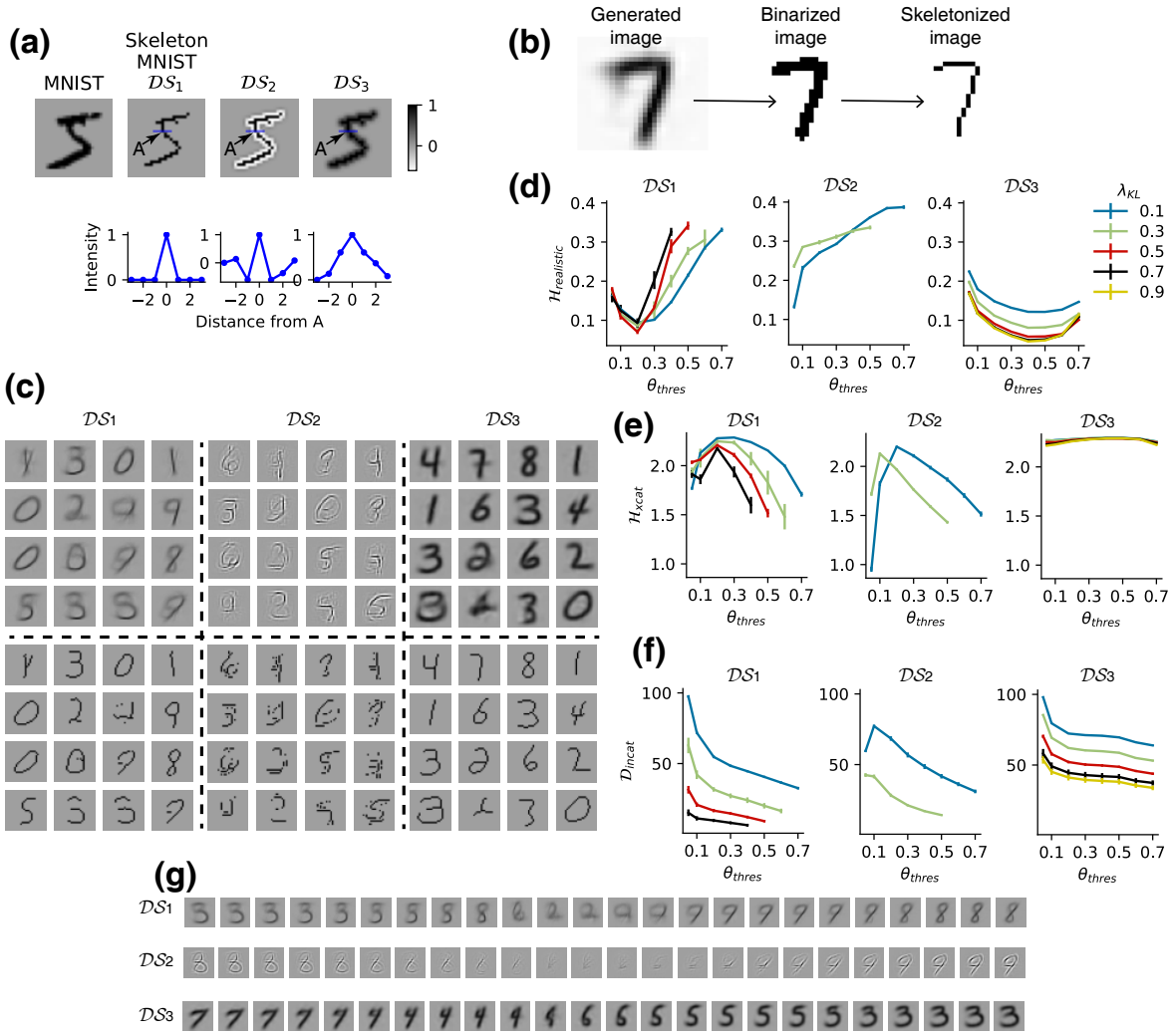

Figure 4: Complex cells facilitate creativity, skeleton MNIST dataset. (**a**) Upper panels: example images in the MNIST, skeleton MNIST ($\mathcal{DS}_1$), $\mathcal{DS}_2$ and $\mathcal{DS}_3$ datasets. Lower panels: pixel intensity along the blue bar in the upper panel as a function of the distance from point A (i.e., the middle point of the bar). (**b**) An example manifesting the post-processing of the generated image 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. (**e, f**) Similar to panel **d**, but for $\mathcal{H}_{xcat}$ and D_{incat}. (**g**) Examples of the generated images when the bottleneck state continuously changes its value, see more examples in **Supplementary Fig. 1**. In panels **d, e, f**, error bars representing s.e.m.. The bottleneck state of VAE has 20 dimensions.

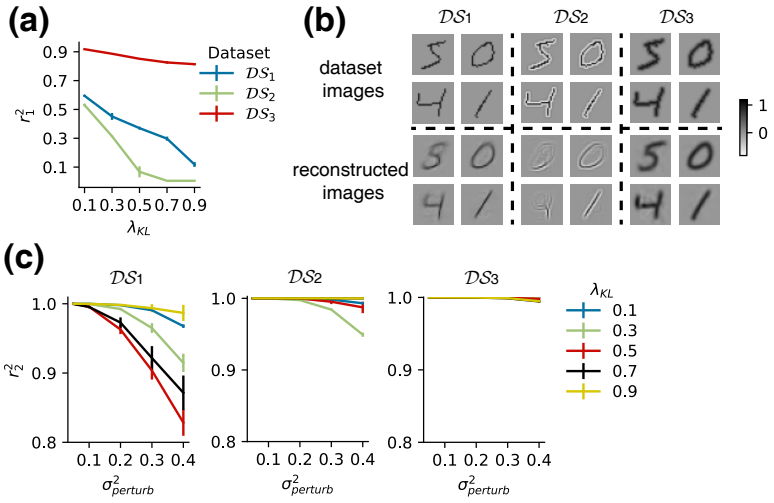

Figure 5: Complex cell facilitates perception, skeleton MNIST dataset. (**a**) r_1^2 for the VAEs trained by the three datasets. (**b**) Examples of the dataset images and the reconstructed images by VAE, note that $\mathcal{DS}_3$ result in best reconstruction quality. (**c**) r_2^2 as a function of $\sigma_{perturb}$ for the three datasets when λ_{KL} takes different values. Error bars represent s.e.m..

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. 4g**). 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 **Supplementary Fig. 1** for more examples). This investigation gains insights into the facilitation of creativity by complex cells.

Model 1 in **Fig. 1** can hardly generate realistic-looking samples (**Fig. 1e**); the generated digits trained by $\mathcal{DS}_1$, although are worse than those trained by $\mathcal{DS}_3$ in quality, are sometimes recognizable (**Fig. 4c**). This is possibly because in 1-dimensional case, there is no delta peak p_3 that can interpolate two delta peaks p_1 and p_2, so that $d(p_1, p_3) < d(p_1, p_2)$ and $d(p_2, p_3) < d(p_1, p_2)$, where $d(\cdot, \cdot)$ is the Euclidean distance between the representations of two peaks; in 2-dimensional case, however, we can draw many 1-pixel-width thin lines that interpolate two parallel thin lines. In both the 1- and 2-dimensional cases, the models corresponding to the complex cells perform best.

Visual perception

We also studied the perception performance of VAEs trained by the three datasets using a similar scheme to that for 1-dimensional case (**Fig. 3a**). We found that both r_1^2 and r_2^2 have highest values for $\mathcal{DS}_3$ (**Fig. 5**), supporting that complex cells facilitate perception.

Sparseness and slowness together explain the power-law eigenspectrum of V1

Experiments found power-law eigenspectrum of V1 activity with exponent close to 1 [18]. Here we would like to explain this phenomenon using a parsimonious generative model. Our working hypothesis is that V1 is trying to reconstruct the input temporal stimuli through a top-down pathway, keeping the activity of V1 as sparse and slow-changing as possible. Specifically, we minimized the following cost function:

$$E(\{\boldsymbol{w}_i\}_i, \{x_{i,t}\}_{i,t}) = \frac{1}{2}\sum_t (\boldsymbol{I}_t - \sum_i \boldsymbol{w}_i x_{i,t})^2 + \lambda_{sparse}\sum_{t,i}|x_{i,t}| + \lambda_{slow}\sum_{t,i}(x_{i,t} - x_{i,t-1})^2, \quad (1)$$

where $\boldsymbol{I}_t$ is the input stimulus at time t, $x_{i,t}$ is the activity of the ith neuron in V1 at time t, $\boldsymbol{w}_i$ is the top-down reconstruction weight from neuron i, and λ_{sparse} and λ_{slow} respectively control the strengths of sparseness and temporal slowness. We generated temporal stimuli by

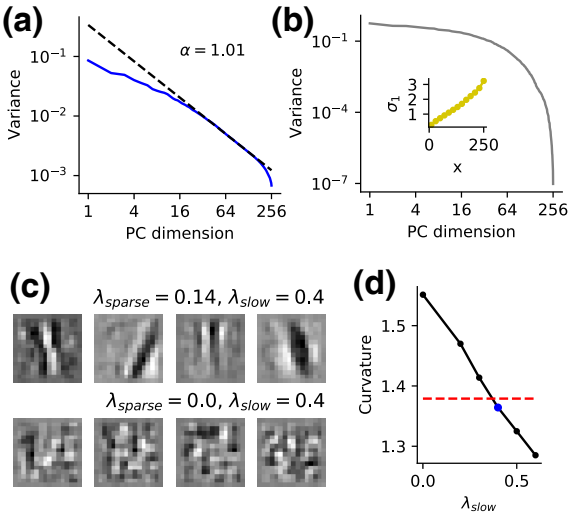

Figure 6: Sparseness and temporal slowness explain the power-law eigenspectrum of V1 in a generative model. (**a**) Eigenspectrum when $\lambda_{sparse} = 0.14$, $\lambda_{slow} = 0.4$. The dashed black line denotes the linear fit of $n^{-\alpha}$. (**b**) Eigenspectrum of the partially whitened image used to train the model. Inset: $\sigma_1(x)$ quantifies the whiteness (i.e., equal variance along all PCs) in the subspace spanned by the first x PCs. Smaller σ_1 indicates more equal variances. (**c**) Examples of generation weights ($\boldsymbol{w}_i$ in **eq. 1**) after training, when $\lambda_{sparse} = 0.14$, $\lambda_{slow} = 0.4$ (upper panels) and $\lambda_{sparse} = 0.0$, $\lambda_{slow} = 0.4$ (lower panels). (**d**) Curvature of temporal trajectory of V1 state as a function of λ_{slow} when $\lambda_{sparse} = 0.14$. Dashed red horizontal line indicates the curvature of the images used to train the model. The blue dot represents the $\lambda_{slow} = 0.4$ value that panel **a** takes.

sliding a small spatial window on static natural images, imitating the experimental protocol of Ref. [18], where animals were seeing static natural images, and the temporal change of stimuli resulted from head or eye movement of the animals.

At suitable values of λ_{sparse} and λ_{slow}, the variance of the nth principal component (PC) of samples of $\{x_{i,t}\}_i$ decay as a power law $n^{-\alpha}$, with exponent $\alpha \approx 1$ (**Fig. 6a**), with successive PC dimensions encoding finer spatial features (**Supplementary Fig. 4a**). Similar to the result in Ref. [18], this power-law scaling is not inherited from the eigenspectrum of input stimuli (**Fig. 6b**), because we partially whitened the natural images in the preprocessing step, modeling the function of the retina and lateral geniculate nucleus [30]. After training, $\boldsymbol{w}_i$ exhibited Gabor-like shape under the sparseness constraint (**Fig. 6c**), which implies Gabor-like receptive fields of the neurons [4]. When the slowness constraint is strong, the curvature of the temporal trajectory of the neuronal population activity $\{\boldsymbol{x}_t\}_t$ (we denote $\boldsymbol{x}_t = \{x_{i,t}\}_i$) can be smaller than the curvature of the temporal trajectory of the stimuli $\{\boldsymbol{I}_t\}_t$ (**Fig. 6d**), consistent with experimental observations [13].

We then searched the exponent α in a range of λ_{sparse} and λ_{slow}, and found that α is most close to 1 at an optimal sparseness strength λ_{sparse}^{optim} at which α is most insensitive to λ_{slow} when $\lambda_{slow} \neq 0$ (**Fig. 7a, b**).

How to understand the good property of λ_{sparse}^{optim}? Ref. [18] proprosed that the power-law scaling of PC variances is a compromise between the whitening and differentiability of neural code. The slowness constraint improves the differentiability, manifested by the straightening of temporal trajectories (**Fig. 6d**), so as long as λ_{slow} is large enough, optimal code should be obtained when the sparseness constraint most whitens the code. Closer investigation of the eigenspectrum unveiled that at the λ_{sparse} and λ_{slow} values resulting in $\alpha > 1$, the eigenspectrum decays in a power-law manner toward the end (blue arrow in **Fig. 7c**). In this case, the exponent $\alpha > 1$ manifests the undersize of the variances of the last several PCs. When $\alpha < 1$, however, there is a sharp drop toward the end of the eigenspectrum (red arrow in **Fig. 7d**), also manifesting the undersize of the variances of the last several PCs. Therefore, both $\alpha > 1$ and $\alpha < 1$ result from the non-whiteness of the code. To test this hypothesis, we let $\lambda_{slow} = 0$ and quantified the sparseness-induced whiteness through two indexes σ_1 and σ_2 (see Methods). Consistent with our presumption, both indexes get their smallest values around λ_{sparse}^{optim} (**Fig. 7f**), indicating the optimal whiteness.

To get some understanding on the optimal sparseness-induced whitening at $\lambda_{sparse} \approx \lambda_{sparse}^{optim}$, we studied the whiteness of the reconstructed stimuli (i.e., $\sum_i \boldsymbol{w}_i x_{i,t}$ in **eq. 1**) when $\lambda_{slow} = 0$. The link between the whiteness of the reconstructed stimuli and that of the V1

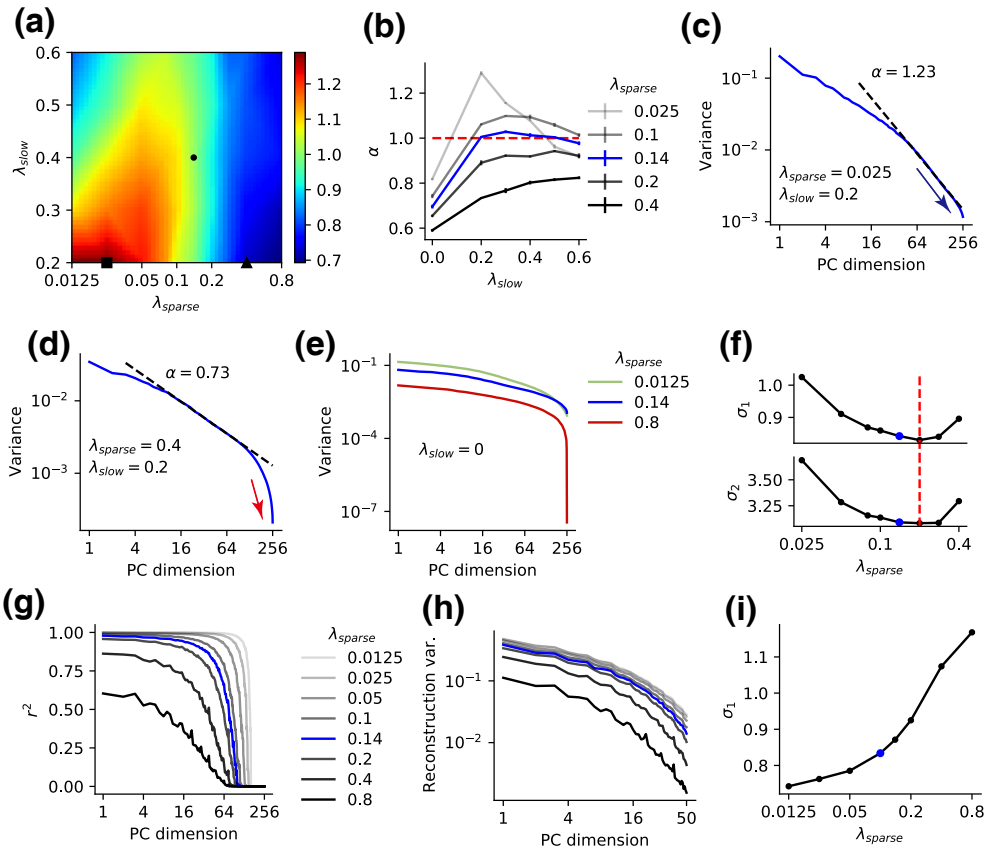

Figure 7: The experimental exponent $\alpha \approx 1$ is realized around a sparseness strength that best whitens the V1 code. (**a**) α as a function of λ_{sparse} and λ_{slow}. The black dot represents the $\lambda_{sparse} = 0.14$ and $\lambda_{slow} = 0.4$ value pair that panel **a** of **Fig. 6** takes. The black square and triangle respectively represent the $(\lambda_{sparse}, \lambda_{slow})$ value pairs used in panels **c** and **d** of this figure. (**b**) α as a function of λ_{slow} at different λ_{sparse} values. The blue curve represents the cases when $\lambda_{sparse} = \lambda_{sparse}^{optim} = 0.14$, which is also the λ_{sparse} value indicated by the black dot in panel **a**. The dashed red line represents $\alpha = 1$. (**c, d**) Eigenspectra when λ_{sparse} and λ_{slow} takes the indicated values. The blue and red arrows are explained in the text. (**e**) Eigenspectra when λ_{sparse} takes different values with $\lambda_{slow} = 0$. (**f**) σ_1 and σ_2 as functions of λ_{sparse}. Smaller σ_1 and σ_2 values indicate more equal variance along different PCs, which means more whitened V1 activity. The blue dot represents $\lambda_{sparse} = \lambda_{sparse}^{optim} = 0.14$. The dashed red line indicates the λ_{sparse} value that minimizes σ_1 and σ_2. (**g**) r^2 quantifies how well the reconstruction can explain the variance of the stimuli along a specific PC dimension of the stimuli. The blue line represents $\lambda_{sparse} = \lambda_{sparse}^{optim} = 0.14$. (**h**) The variance of the reconstruction along PC dimensions of the stimuli. (**i**) σ_1 of the reconstructed images as a function of λ_{sparse}. Smaller σ_1 indicates that the reconstructed images are more whitened. In panels **e-i**, $\lambda_{slow} = 0$.

activity $\{x_{i,t}\}_i$ is strictly valid when the reconstruction weights $\{\boldsymbol{w}_i\}_i$ are orthonormal. In our simulation, we constrained $\|\boldsymbol{w}_i\|^2 = 1$, the approximate orthogonality between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ $(i \neq j)$ is numerically manifested in **Supplementary Information Section 4** and **Supplementary Fig. 4b**.

Specifically, we investigated $r^2(m)$, which is the ratio of the variance of the mth PC of the stimuli $\boldsymbol{I}_t$ explained by the top-down reconstruction $\sum_i \boldsymbol{w}_i x_{i,t}$. We found that the reshaping of the function $r^2(m)$ with the increase of λ_{sparse} experiences two stages, separated largely at $\lambda_{sparse} \approx \lambda_{sparse}^{optim}$ (see the blue curve in **Fig. 7g**). At stage 1 ($\lambda_{sparse} < \lambda_{sparse}^{optim}$), $r^2(m)$ remains close to 1 for small m values, but drops to zero for large m values. This means that at this stage, with the increase of λ_{sparse}, the model gradually abandons the reconstruction of the PCs with small variances, whereas the reconstruction in the subspace $\mathcal{S}$ of the dominating PCs remains largely intact. The stimuli in the subspace $\mathcal{S}$ are well-whitened (**Fig. 6b, inset**). At stage 2 ($\lambda_{sparse} > \lambda_{sparse}^{optim}$), $r^2(m)$ lowers down for small m values, suggesting the deterioration of the reconstruction of $\mathcal{S}$. This deterioration is not uniform along all PCs in $\mathcal{S}$, with the PC with smaller variance deteriorated worse (**Fig. 7h**). The reconstructed stimuli in the subspace $\mathcal{S}$ becomes non-whitened with the increase of λ_{sparse} (**Fig. 7i**). These results imply a scenario that the improvement of the whiteness of V1 code with λ_{sparse} when $\lambda_{sparse} < \lambda_{sparse}^{optim}$ is because V1 gradually focuses to reconstruct a well-whitened subspace $\mathcal{S}$ of stimuli spanned by the dominating PCs, and the deterioration of whiteness with λ_{sparse} when $\lambda_{sparse} > \lambda_{sparse}^{optim}$ is because V1 deteriorates the reconstruction of the well-whitened subspace $\mathcal{S}$ non-uniformly along different PC directions. See **Supplementary Information Section 4** and **Supplementary Fig. 4c** for more supports on the notion that the reconstruction of PC with smaller variance is more impaired with the increase of λ_{sparse}.

Discussion

In this paper, we found that complex cells facilitate visual creativity and perception, and showed that the close-to-1 exponent of the power-law eigenspectrum of V1 is realized at a sparseness strength that best whitens V1 code and makes the exponent most insensitive to slowness strength. Our work provides fresh insights into the cognitive roles of complex cells from top-down generative perspective, establishes a link between the V1 eigenspectrum and the principles of sparseness and slowness, and suggests that there is an optimal sparseness strength that V1 is working at.

Brain as an organ of active explanation

Predictive coding is the dominating theory in the generative theories of perception, which proposes that the brain iteratively updates its explanation for the world using the error of the current explanation transmitted through the feedforward pathway [5] (**Fig. 3a**). However, one should note that error updating is not the only approach to construct this explanation. For example, the encoder of VAE uses a deep feedforward network to construct this explanation in the bottleneck state (**Fig. 1d**). Similar technique has been used to solve lasso regression [31], where a deep neural network is used to approximate an implicit function defined by iterative error updating that maps stimuli to hidden states. The advantage of this deep neural network approximation is computational speed. It has been found that information is transmitted feedforwardly at the early stage of perception, but processed recurrently at the later stage [7, 32, 33], which reflects the shift of computational demand from responding speed to representation precision during the process of perception. Our perception model (**Fig. 3a**) also captures these two stages, where the μ-channel of the feedforward encoder output of VAE is regarded as the initial bottleneck state before recurrent updating (see the text explanation related to **Fig. 3b**).

An important but seldomly discussed question is why the brain is evolved to use the generative explanation scheme instead of the passive sensation scheme. One possible reason is that the brain has to continuously adapt itself to kaleidoscopic task demands. For example, it is believed that the sharpened feature tuning of simple cells by lateral inhibition improves discrimination [20, 22] and the position tolerance of complex cells improves classification [15, 12]. These understandings preset the task demands of discrimination and classification for simple and complex cells respectively. The problem is that a feedforward network optimized for one task may perform badly for another: high discrimination may imply strong amplification of intra-class noises, blurring the clustered structure of the inputs [34], impairing classification task; a good classifier may map all elements in the same class onto a single output, impairing discrimination task. The generative explanation scheme, however, requires the activity of the high-level layer reconstructs that of the low-level layer, which ensures that most information

in the low-level layer gets represented in higher levels. The high-level layer can have representations optimized for different tasks after imposed different priors, but remains sensitive to the low-level information un-used in the current task, ready to switch tasks according to environmental changes.

Complex cells, creativity, and perception

In this paper, we show that temporal slowness is necessary for the power-law eigenspectrum of V1 with exponent $\alpha \to 1^+$ (**Fig. 7b**), and that complex cells facilitate higher-order cortex to form representations smooth to shape morphing of stimuli, improving creativity and perception (**Figs. 1-5**). Previous studies showed that temporal slowness is necessary for the developmental formation of complex cells [35, 36], and that the $\alpha \to 1^+$ power-law exponent ensures the differentiability of neural code [18]. These results together suggest the cognitive role of the close-to-1^+ exponent of the eigenspectrum of V1: facilitating higher-order cortex to form representations differentiable to shape morphing of stimuli thereby improving creativity and perception. To better understand this point, suppose that higher-order cortex represents a stimulus at position x using code $z(x)$, and the V1 code is $v_1(x)$. The generative theory of perception requires that $v_1(x) \approx f(z(x))$, where $f(\cdot)$ represents the deep neural network along the top-down pathway, which is differentiable in the biologically plausible context. Therefore, if $z(x)$ is to be differentiable to x, $v_1(x)$ must also be differentiable to x to better fulfill the equation $v_1(x) \approx f(z(x))$. Consistently, in **Supplementary Information Section 5** and **Supplementary Fig. 5**, we trained VAE to generate V1 activity (i.e., $\{x_{i,t}\}_i$ in **eq. 1**), and found that VAEs trained by V1 activity at $\lambda_{slow} \neq 0$ (we kept $\lambda_{sparse} = 0.14$ so that $\alpha \approx 1$ when $\lambda_{slow} \neq 0$, see **Fig. 7b**) performs better on creativity and perception than those trained by V1 activity at $\lambda_{slow} = 0$.

Engineering neural networks to generating images with high-resolution details is difficult. VAE tends to generate blurred images [37]. Another well-known generative model, generative adversarial network (GAN), also struggles when generating high-resolution details [38] due to a gradient problem [39]. Our results suggest that blurring these high-resolution details may result in better creativity and perception performance ($\mathcal{DS}_3$ in **Figs. 4a** and **5c**). 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.

It is believed that mental creativity mostly involves default mode network and executive control network [27, 40], which include association cortical areas in the frontal, parietal and cingulate cortices. Our results suggest that low-order sensory cortices such as V1 also plays an important role in, or is even designed for, the high-order cognitive task of creativity. But this may not be so surprising: the cost function that VAE optimizes is the variational free energy [19], similar to that during perception and learning [1], two cognitive processes that V1 also participates in. Therefore, creativity, perception and learning are different aspects of the same nature: free-energy minimization.

The neural mechanism of V1 adaptation

In Ref. [18], animals were presented stimuli ensembles with different statistics. Similarly, we also studied our model using stimuli ensembles of a low dimensionality $d = 4$. Consistent with our result with high-dimensional stimuli, λ_{sparse}^{optim}, at which the exponent α is most close to the experimental value 1.62 and insensitive to the change of λ_{slow}, is obtained near a λ_{sparse} value that best whitens the V1 code at $\lambda_{slow} = 0$ (**Supplementary Information Section 2** and **Supplementary Fig. 2**). Therefore, V1 may be able to adapt itself to the statistics of the presented images by adjusting λ_{sparse}, in the time scale of the experiment [18] (i.e., minutes).

In our model, this adaptation to stimuli statistics is realized by adjusting λ_{sparse} to λ_{sparse}^{optim} before learning the generation weight $\boldsymbol{w}_i$ (**eq.1**). λ_{sparse} has the physiological meaning of neuronal firing threshold, and $\boldsymbol{w}_i$ is related to the feedforward and recurrent weights to and within V1 [41, 42]. So it is likely that V1 performs this adaptation by first adjusting global inhibition, and then adjusting feedforward and recurrent weights accordingly. This global inhibition may be adjusted by thalamocortical connections or neuromodulators: it has been found that inactivation (or excitation) of the pulvinar neurons suppresses (or increases) the responses of superficial V1 neurons to visual input [43]; and cholinergic axons from the basal forebrain depolarize cortical interneurons [44]. We found that λ_{sparse}^{optim} for low-dimensional stimuli is smaller than that for high-dimensional stimuli (**Fig. 7f**, **Supplementary Fig. 2g**), which means that this global inhibition is weaker during the presentation of low-dimensional stimuli than during high-dimensional stimuli.

What is the neural mechanism the brain uses to guide λ_{sparse} to λ_{sparse}^{optim}? **Fig. 7g** shows that the reconstruction of PCs with small (or large) variance is impaired with the increase of λ_{sparse} when $\lambda_{sparse} < \lambda_{sparse}^{optim}$ (or $\lambda_{sparse} > \lambda_{sparse}^{optim}$). Consistently, the reconstruction error ϵ

increases slowly with λ_{sparse} when $\lambda_{sparse} < \lambda_{sparse}^{optim}$, but starts to quickly soar up after $\lambda_{sparse} > \lambda_{sparse}^{optim}$ (**Supplementary Information Section 3** and **Supplementary Fig. 3**). Therefore, it is possible that the brain monitors ϵ when adjusting λ_{sparse}, and stops the adjustment just at the point where ϵ starts to soar up with λ_{sparse}. Predictive coding theory suggests that ϵ is encoded by the pyramidal neurons in superficial layers [1], experiments found that ϵ is closely related to the gamma-band frequencies [45, 46]. More detailed mechanical insights require experimental studies.

Methods

Manifesting the advantage of complex cells using toy models

A variational auto-encoder (VAE) [19] contains two parts: an encoder and a decoder (**Fig. 1d**). 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:

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

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. 1, 2**, 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 [47].

In **Figs. 1, 2**, the inputs received by VAE (i.e., the outputs of the three models) are positioned on a one-dimensional line of 201 neurons. In Model 1, this input 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 input is a Gabor function $f_2(x; a) = C\exp(-\frac{(x-a)^2}{2\sigma^2})\sin(\frac{2\pi}{T}(x-a))$, where $\sigma = 10$, $T = 80$, C is a normalization factor such that $\max_x f_2(x; a) = 1$. In Model 3, this input is a Gaussian function $f_3(x; a) = C\exp(-\frac{(x-a)^2}{2\sigma^2})$, 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 in the range $[31, 171]$.

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 output pattern of Model-out in **Fig. 2d** in response to the wavelet stimulus at position x, 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).

In **Fig. 3**, the bottleneck state was optimized to minimize the error between the target pattern and the generated pattern using Adam optimizer [47].

Manifesting the advantage of complex cells using skeleton MNIST dataset

The dataset $\mathcal{DS}_1$ in **Fig. 4a** is the skeleton MNIST dataset [25]. 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 generated 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 VAE used in **Fig. 4** 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 images generated by VAE were post-processed in two steps. 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. This MLP contained a hidden layer of size 1000 with leaky-relu activation function. After receiving a post-processed image I generated by VAE, this MLP output a label distribution $p(x|I)$ ($x = 0, 1, \cdots, 9$). In **Fig. 4d**, $\mathcal{H}_{realistic} = E_I[-\sum_x p(x|I) \ln p(x|I)]$, where $E_I[\cdot]$ means average over all the generated images [28]; in **Fig. 4e**, $\mathcal{H}_{xcat} = -\sum_x E_I[p(x|I)] \ln E_I[p(x|I)]$ [28]. To plot **Fig. 4f**, we first chose the generated post-processed 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 $D_{incat} = E_x[\frac{(\sum_i \lambda_i(x))^2}{\sum_i \lambda_i^2(x)}]$ [29]. **Fig. 4d-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. 2**. 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. 4d-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. 4g were plotted by gradually changing the bottleneck state of VAE from $\boldsymbol{z} = [1.5, 1.5, \cdots, 1.5]$ to $[-1.5, -1.5, \cdots, -1.5]$.

The generative model to explain the eigenspectrum of V1

To explain the eigenspectrum of V1, our working hypothesis is that V1 is trying to reconstruct a sequence of input images through a top-down pathway, keeping the activity of V1 as sparse and slow-changing as possible. Specifically, we minimized the following cost function:

$$E(\{\boldsymbol{w}_i\}_i, \{x_{i,t}\}_{i,t}) = \frac{1}{2}\sum_t (\boldsymbol{I}_t - \sum_i \boldsymbol{w}_i x_{i,t})^2 + \lambda_{sparse}\sum_{t,i}|x_{i,t}| + \lambda_{slow}\sum_{t,i}(x_{i,t} - x_{i,t-1})^2, \tag{3}$$

where $\boldsymbol{I}_t$ is the input image at time t, $x_{i,t}$ is the activity of the ith neuron in V1 at time t, $\boldsymbol{w}_i$ is the top-down reconstruction weight from neuron i, and λ_{sparse} and λ_{slow} respectively control the strengths of sparseness and temporal slowness. However, minimizing **eq. 3** for a long sequence $\{\boldsymbol{I}_t\}_t$ of images is computationally costly, so approximation has to be used. Specifically, we minimized the following cost fuction for short sequences $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3\}$ of only three images using an EM algorithm:

$$E_{\text{short\_seq}}(\{\boldsymbol{w}_i\}_i, \{x_{i,t}\}_{i,t=1,2,3}) = \frac{1}{3}\frac{1}{2}\sum_{t=1,2,3}(\boldsymbol{I}_t - \sum_i \boldsymbol{w}_i x_{i,t})^2 + \frac{1}{3}\lambda_{sparse}\sum_{i,t=1,2,3}|x_{i,t}| + \frac{1}{2}\lambda_{slow}\sum_{i,t=2,3}(x_{i,t} - x_{i,t-1})^2. \tag{4}$$

In the E-step, $E_{\text{short\_seq}}$ was minimized respective to $\{x_{i,t}\}_{i,t=1,2,3}$ using a fast iterative shrinkage-thresholding algorithm (FISTA) [48]; in the M-step, $E_{\text{short\_seq}}$ was minimized respective to $\{\boldsymbol{w}_i\}_i$ using Adam optimizer [47]. After training, we fixed $\{\boldsymbol{w}_i\}_i$ and inferred $\{x_{i,t}\}_{i,t}$ from a given image sequence $\{\boldsymbol{I}_t\}_t$ in a Markovian manner: we inferred $\{x_{i,t}\}_{i,t}$ temporally sequentially (i.e., starting from $\{x_{i,t}\}_{i,t=1}$ to $\{x_{i,t}\}_{i,t=2}$ then to $\{x_{i,t}\}_{i,t=3}$, ...); when inferring $\{x_{i,t}\}_{i,t=T}$, we fixed the values of $\{x_{i,t}\}_{i,t<T}$. The off-line training and on-line inferring algorithms model the replay-driven plasticity [49] and the perception of V1 respectively. To calculate the eigenspectra (**Figs. 6a**, **7a-e**), we prepared a number of triplet sequences of three images $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3\}$ (see below), inferred the states in the Markovian manner above, and collected the state $\{x_{i,t}\}_{i,t=3}$ that corresponding to $\boldsymbol{I}_3$. To calculate the curvature of temporal trajectory of states (**Fig. 6d**), we prepared

sequences of four images $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3, \boldsymbol{I}_4\}$, and calculated the curvature c of $\{x_{i,t}\}_{i,t=\{2,3,4\}}$ by [13]

$$c = \arccos\left(\frac{\boldsymbol{x}_3 - \boldsymbol{x}_2}{\|\boldsymbol{x}_3 - \boldsymbol{x}_2\|} \cdot \frac{\boldsymbol{x}_4 - \boldsymbol{x}_3}{\|\boldsymbol{x}_4 - \boldsymbol{x}_3\|}\right), \tag{5}$$

where we have denoted $\boldsymbol{x}_t = \{x_{i,t}\}_i$. In our simulation, the size of image $\boldsymbol{I}_t$ was $16 \times 16 = 256$, and the number of hidden units x_i was 257.

Image sequence preparation

The image sequences $\{\boldsymbol{I}_t\}_t$ used to train the model (**eq. 4**) were prepared in the following way. We picked 100 images from van Hateren's natural image dataset [50], avoiding images that contained large areas of the sky. We first took logarithm of the intensities of the image pixels, following the suggestion of Ref. [50], and then partially whitened the images using the method in Ref. [4], modeling the image whitening by retina or lateral geniculate nucleus in the upstream of V1 [30]. To get a short sequence $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3\}$ in **eq. 4**, we picked a 16×16 patch from the images preprocessed above, and sliding the position of the patch window by the same vector $\Delta\boldsymbol{P} = (\Delta X, \Delta Y)$ for two successive steps, where ΔX and ΔY were randomly -1, 0 or 1. A caveat here is a boundary effect. To see this, suppose all pixels in $\boldsymbol{I}_1$ have zero intensity, but after the patch window moves by $\Delta\boldsymbol{P}$, the pixels in a boundary of $\boldsymbol{I}_2$ have strong intensities. In this case, $x_{i,t=1} = 0$ for all is, but $|x_{i,t=2}|$ may be large, enlarging the cost for temporal slowness (i.e., the third term at the right-hand side of **eq. 4**). This large cost term does not represent the fast change of the stimulus itself, but is due to the sudden entrance of high-intensity pixels into the small patch window. To alleviate this boundary effect, we multiplied element-wise each image patch $\boldsymbol{I}_t$ by a 16×16 filter $\boldsymbol{F}$. A pixel of $\boldsymbol{F}$ took value 0.05, 0.24, 0.43, 0.62, 0.81 or 1, depending on whether its distance with the boundary was 1, 2, 3, 4, 5 or larger than 5 pixels. This filtering also improves the biological plausibility of the model, because it means that the response of a neuron to a stimulus gradually decays (instead of sharply reducing to zero) when the stimulus is moving away from the center of the receptive field.

Power-law exponent estimation

The power-law exponents of the eigenspectra of V1 states (**Figs. 6** and **7**) were estimated in the following way. We noted that the change of eigenvalue λ_i with the order i of principal component largely has three stages (**Figs. 6a** and **7c, d**): when i takes small values, the decay of λ_i with i is relatively slow and may exhibit zigzag fluctuations during decaying; when i takes intermediate values, the decay of λ_i with i can be best approximated by power law; when i takes large values, λ_i quickly decays with i. These three stages also exhibited in experimental results [18]. The power-law exponent of an eigenspectrum was obtained by linearly fitting the intermediate stage in log-log scale. The intermediate stage was determined by eye-looking, which slightly varied when λ_{sparse} in **eq. 4** took different values, and largely remained the same when λ_{slow} changed in the range of our study. When $\lambda_{sparse} = 0.14$ (which is the value in **Fig. 6a**) , we let $i \in [29, 109]$ be the intermediate range.

Quantifying the whiteness of eigenspectrum

Suppose the variance of the nth principal component (PC) is v_i, to quantify the whiteness of the eigenspectrum, we used two indexes: (1) $\sigma_1 = \text{std}_n(\log(v_n))$, which is the standard deviation over logarithm of the variances; (2) $\sigma_2 = \sum_{n=\{\text{first 20}\}} \log(v_n) - \sum_{n=\{\text{last 20}\}} \log(v_n)$ which is the difference between the summation of $\log(v_n)$ over the first 20 PCs and that over the last 20 PCs. In **Fig. 7f**, we used both σ_1 and σ_2; in the inset of **Fig. 6b**, we used σ_1.

Acknowledgements

Z.B. thanks Prof. Changsong Zhou and Prof. Shuzhi Sam Ge for comments on the manuscript and helpful discussions. Z.B. is supported by the NSF of China (Grant No. 32000694) and the start-up fund of the Institute for Future, Qingdao University.

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

[4] 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.

[5] 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.

[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] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.

[11] J. M. Alonso and L. M. Martinez, "Functional connectivity between simple cells and complex cells in cat striate cortex," *Nat. Neurosci.*, vol. 1, pp. 395–403, 1998.

[12] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, pp. 415–434, 2012.

[13] O. J. Hénaff, R. L. T. Goris, and E. P. Simoncelli, "Perceptual straightening of natural videos," *Nat. Neurosci.*, vol. 22, pp. 984–991, 2019.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[15] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, pp. 1019–1025, 1999.

[16] P. Berkes and L. Wiskott, "Slow feature analysis yields a rich repertoire of complex cell properties," *J. Vis.*, vol. 5, pp. 579–602, 2005.

[17] Y. Singer, Y. Teramoto, B. D. B. Willmore, J. W. H. Schnupp, A. J. King, and N. S. Harper, "Sensory cortex is optimized for prediction of future input," *eLife*, vol. 7, p. e31557, 2018.

[18] 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.

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

[20] S.-H. Lee, A. C. Kwan, S. Zhang, V. Phoumthipphavong, J. G. Flannery, S. C. Masmanidis, H. Taniguchi, Z. J. Huang, F. Zhang, E. S. Boyden, K. Deisseroth, and Y. Dan, "Activation of specific interneurons improves v1 feature selectivity and visual perception," *Nature*, vol. 488, pp. 379–383, 2012.

[21] J. M. Crook, Z. F. Kisvárday, and U. T. Eysel, "Evidence for a contribution of lateral inhibition to orientation tuning and direction selectivity in cat visual cortex: reversible inactivation of functionally characterized sites combined with neuroanatomical tracing techniques," *Eur. J. Neurosci.*, vol. 10, pp. 2056–2075, 1998.

[22] J. S. Isaacson and M. Scanziani, "How inhibition shapes cortical activity," *Neuron*, vol. 72, pp. 231–243, 2011.

[23] 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.

[24] 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.

[25] E. D. de Jong, "https://github.com/edwin-de-jong/mnist-digits-stroke-sequence-data/wiki/mnist-digits-stroke-sequence-data,"

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

[27] 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.

[28] 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.

[29] 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.

[30] L. Zhaoping, *Understanding Vision: Theory, Models, and Data*. Oxford: Oxford University Press, 2014.

[31] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 399–406, 2010.

[32] K. Seeliger, M. Fritsche, U. Güçlü, S. Schoenmakers, J.-M. Schoffelen, S. E. Bosch, and M. A. J. van Gerven, "Convolutional neural network-based encoding and decoding of visual object recognition in space and time," *NeuroImage*, vol. 180, pp. 253–266, 2018.

[33] M. Cauchoix, G. Barragan-Jason, T. Serre, and E. J. Barbeau, "The neural dynamics of face detection in the wild revealed by MVPA," *J. Neurosci.*, vol. 34, pp. 846–854, 2014.

[34] B. Babadi and H. Sompolinsky, "Sparseness and expansion in sensory representations," *Neuron*, vol. 83, pp. 1213–1226, 2014.

[35] G. Matteucci and D. Zoccolan, "Unsupervised experience with temporal continuity of the visual environment is causally involved in the development of V1 complex cells," *Sci. Adv.*, vol. 6, p. eaba3742, 2020.

[36] P. Földiák, "Learning invariance from transformation sequences," *Neural Comput.*, vol. 3, pp. 194–200, 1991.

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

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

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

[40] R. E. Beaty, M. Benedek, P. J. Silvia, and D. L. Schacter, "Creative cognition and brain network dynamics," *Trends Cogn. Sci.*, vol. 20, pp. 87–95, 2016.

[41] C. S. N. Brito and W. Gerstner, "Nonlinear hebbian learningas a unifying principle in receptive field formation," *PLoS Comput. Biol.*, vol. 12, p. e1005070, 2016.

[42] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Res.*, vol. 37, pp. 3311–3325, 1997.

[43] G. Purushothaman, R. Marion, K. Li, and V. A. Casagrande, "Gating and control of primary visual cortex by pulvinar," *Nat. Neurosci.*, vol. 15, pp. 905–912, 2012.

[44] S. Arroyo, C. Bennett, D. Aziz, S. P. Brown, and S. Hestrin, "Prolonged disynaptic inhibition in the cortex mediated by slow, non-α7 nicotinic excitation of a specific subset of cortical interneurons," *J. Neurosci.*, vol. 32, pp. 3859–3864, 2012.

[45] 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.

[46] 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.

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

[48] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIIMS*, vol. 2, pp. 183–202, 2009.

[49] H. Miyamoto and T. K. Hensch, "Bidirectional interaction of sleep and synaptic plasticity: A view from visual cortex," *Sleep Biol. Rhythms*, vol. 4, pp. 35–43, 2006.

[50] J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc. R. Soc. Lond. B*, vol. 265, pp. 359–366, 1998.