
Article:
Zhang, Y, Bu, T, Zhang, J et al. (2022) Decoding Pixel-Level Image Features from Two-Photon Calcium Signals of Macaque Visual Cortex. Neural Computation, 34 (6). pp. 1369-1397. ISSN: 0899-7667

https://doi.org/10.1162/neco_a_01498

Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex

Yijun Zhang
yijzhang@sjtu.edu.cn
Department of Computer Science and Engineering, Shanghai Jiao Tong University, China; Department of Computer Science and Technology, Peking University, China
Tong Bu
putong30@pku.edu.cn
Department of Computer Science and Technology, Peking University, China
Jiyuan Zhang
jyzhang@stu.pku.edu.cn
Department of Computer Science and Technology, Peking University, China
Shiming Tang
tangshm@pku.edu.cn
School of Life Sciences and Peking-Tsinghua Center for Life Sciences, Peking University, China
Zhaofei Yu
yuzf12@pku.edu.cn
Department of Computer Science and Technology, Peking University, China; Institute for Artificial Intelligence, Peking University, China
Jian K. Liu
j.liu9@leeds.ac.uk
School of Computing, University of Leeds, United Kingdom
Tiejun Huang
tjhuang@pku.edu.cn
Department of Computer Science and Technology, Peking University, China; Institute for Artificial Intelligence, Peking University, China; Beijing Academy of Artificial Intelligence, China

Abstract Images of visual scenes comprise essential features important for visual cognition of the brain. The complexity of visual features lies at different levels, from simple artificial patterns to natural images with different scenes. It has been a focus of using stimulus images to predict neural responses. However, it remains unclear how to extract features from neuronal responses. Here we addressed this question by leveraging two-photon calcium neural data recorded from the visual cortex of awake macaque monkeys. With stimuli including various categories of artificial patterns and diverse scenes of natural images, we employed a deep neural network decoder inspired by image segmentation technique. Consistent with the notation of sparse coding for natural images, a few neurons with stronger responses dominated the decoding performance, whereas decoding of artificial patterns needs a large number of neurons. When decoding natural images using the model pre-trained on artificial patterns, salient features of natural scenes can be extracted, as well as the conventional category information. Altogether, our results give a new perspective on studying neural encoding principles using reverse-engineering decoding strategies.

1 Introduction

In daily life, a substantial body of sensory information is transferred to the brain, which is then processed by neurons to generate a series of coping behaviors. Among these various types of sensory sources, vision is arguably the dominant contributor to the interaction between the external environment and brain. Understanding how the brain detects, interprets, and responses to visual information representing the external environment is a major question in neuroscience (DiCarlo and Cox, 2007; DiCarlo et al., 2012). It is also a critical obstacle to the development of artificial vision used by intelligent machines (Yu et al., 2020). Fundamentally, two types of computational approaches have been proposed. From the encoding perspective, the goal is to predict the neural responses to arbitrary visual stimuli. From the decoding view, instead, the purpose is to restore the original stimuli from neural responses as similar as possible. Most of the researches treat these two aspects separately. However, it is meaningful to leverage a decoding approach to analyse some characteristics of the neural encoding, such as the notation of neural sparse coding in primary visual cortex (V1) (Tang et al., 2018b; Carlson et al., 2011; Vinje and Gallant, 2000), where sparse encoding of neural scenes has been investigated in detail. Yet, it remains unclear what types of visual features can be extracted from neural responses, beyond the simple decoding of visual categories (Yamins and DiCarlo, 2016; Zeiler and Fergus, 2014; DiCarlo et al., 2012).

For the neural decoding problem, there have been two subdivided targeted questions: classification of a finite set of pre-defined stimuli, or reconstruction of pixel-by-pixel images. Notably, the pixel-level reconstruction is more chal-

lenging. Visual reconstruction from neural signals has been studied over many years. The widely used neural signal data is functional magnetic resonance image (fMRI) activity of the visual cortex (Naselaris et al., 2009; Nishimoto et al., 2011; Qiao et al., 2018; Thirion et al., 2006; Wen et al., 2018), while most recently, fine neural signals have been studied, including neural spikes (Botella-Soler et al., 2018; Gollisch and Meister, 2008; Parthasarathy et al., 2017; Marre et al., 2015; Zhang et al., 2020; Brackbill et al., 2020), and calcium imaging signals (Yoshida and Ohki, 2020; Garasto et al., 2019, 2018).

Commonly assumed, the patterns of neural activity in V1 encode local orientation components of high-order patterns (Jones and Palmer, 1987; Livingstone and Hubel, 1984; Movshon et al., 1978; Hubel and Wiesel, 1959). In the standard hierarchical model of visual object recognition, V1 neurons are often thought as simple oriented feature detectors, whose elements are then taken conjunctions of by the subsequent visual areas in the brain. The overall hierarchical architecture theory for object recognition is the inspiration for convolutional neural networks in recent deep learning technologies (LeCun et al., 2015; Riesenhuber and Poggio, 1999). However, many studies argued about the V1 neuronal coding mechanism. It was suggested by some studies that V1 neurons may also encode complex features (Slllito et al., 1995; Hegdé and Van Essen, 2007). The barriers of V1 neuron coding research are partially because of the difficulty of modeling or interpreting their responses under more complicated stimuli, e.g., natural scenes (Victor et al., 2006; Vinje and Gallant, 2000). Moreover, possible biases and a limited number of tested stimuli in neural sampling experiments impede a comprehensive understanding of the V1 neuronal functions (Olshausen and Field, 2005; Carandini et al., 2005). Benefiting from recent developments in larger-scale two-photon calcium imaging techniques (Denk et al., 1990; Li et al., 2017), researchers are capable of characterizing the V1 neuronal coding more thoroughly (Li et al., 2017; Olshausen and Field, 2005). This unique advantage of calcium imaging allows people to investigate the selectivity and specificity of V1 neurons more comprehensively by testing an extensive set of visual shapes and features (Tang et al., 2018a,b).

Decoding certain features from neural responses, such as faces, has been investigated intensively on monkeys (Chang and Tsao, 2017; Chang et al., 2021; Koyano et al., 2021) and humans (Guntupalli and Gobbini, 2017; Wang et al., 2021). However, it is also important to know how the visual neurons can decode more general features of artificial patterns and natural images. In this study, leveraging neural responses of two-photon calcium imaging recorded in V1 from awake macaque monkeys under a rich set of images, including various categories of artificial pattern (AP) and complex natural scene (NS) images (Tang et al., 2018a,b), we investigated neural coding from a perspective of decoding pixel-level visual features from neural responses and compared the different encoding mechanisms of these two stimulus categories respectively. We first proposed an end-to-end model inspired by deep learning neural networks, especially the image segmentation U-net, to restore the origin stimulus images from neural responses. We then tested the performance of our model on decoding of AP and NS. For more comprehensive inspection, we used six different metrics for image qual-

ity estimation and found that our model can achieve state-of-the-art performance, compared to other models. To find out the effect of sparse coding in V1 neurons, we conducted a series of examinations over AP and NS, with different number of neurons chosen by two different ways: randomly choosing and sorted according to neural response. The reconstruction performance is increasing with more cells for AP, but not for NS. Particularly, the decoding ability of NS is saturated with a small set of cells. When separating both AP and reconstructed NS images into different categories according to the feature of simple patterns, we found the decoding performance is in line with the category, such that the category with less complexity of visual features has better reconstruction results for both AP and NS images. Furthermore, with a subset of cells that have responses to both AP and NS images, we intended to discover some connections between AP and NS responses. We cross validated the decoding in three scenarios: testing decoding of all images with models trained on AP, NS, and both. Remarkably, the model pre-trained with AP can capture the salient feature of natural images manifesting the category information. Additionally, we corroborated our results illustrating the low-dimension distribution of both types of images. Taken together, our results provide new insights into decoding pixel-level visual features from neural responses. Besides accurate reconstruction of images with individual pixels, our model grants a new perspective of reverse-engineering approaching to visual coding, and serves as a novel way of quantifying visual accuracy for visual prostheses of brain-machine interfaces and other neuromorphic vision systems.

2 Methods

2.1 Stimulus images

There are two sets of stimulus images used for neural calcium imaging recordings: artificial pattern (AP) and natural scene (NS) images. In experiments, a total of 9500 artificial patterns were presented to awake monkeys. Patterns were generated from 138 basic prototypes with different rotations and locations within receptive fields of neurons.

2.2 Neural response

The calcium imaging data used in this work were collected from layer 2/3 of awake macaque monkeys, with single-cell two-photon fluorescence microscope. Experimental details of data collection can be found in the original studies (Tang et al., 2018a,b). We denote $r^n_{t,i}$ as the neural response of the n-th neuron for t-th stimulus in the i-th trial ($t = 1, ..., 5$), r^n_t to denote the average response over trials, and $\vec{r}^n$to denote all the average neural responses for this neuron as a vector. Specifically, we have n = 1,...,1142, t = 1,...,5000 for AP, and n = 1,...,1225, t = 1,...,2250 for NS. Based on the previous work (Tang et al., 2018a), neurons whose maximum trial-averaged responses were not above 0.5(max $\vec{r}^n \leq 0.5$ were discarded as their responses were too weak and might be unreliable. When sorting

neural responses, we ranked neurons in terms of max $\vec{r}^n$. As is shown in Fig 5A, the distributions of max neural responses under AP and NS stimuli are different and respectively sensitive to AP or NS stimuli.

Visual stimuli were presented to the appropriate retinotopic position in the visual field of the subject monkey. The subject monkey performed a fixation task while stimuli were displayed. Every AP stimulus was repeated five to six times, and NS stimulus was repeated three times. All V1 neurons' responses were collected during one stimulus presentation as follows: a blank screen for one second, and another second for stimuli display. To quantify every activated cell's response, the ratio of fluorescence change ($\Delta F/F0$) was computed through the differential image of two periods, where $\Delta F = F - F0$. *F0* is the neural activity during blank screen, and ΔF is the fluorescence activity during stimulus display in the trial. Finally, 1142 activated cells for AP and 1225 activated cells for NS were identified in the subject monkey. All these cells were used for training and testing models within their stimulus images. Meanwhile, 599 cells were found both activated in AP and NS experiment process, which were used for training and testing models across AP and NS images.

2.3 Decoding models

We developed a deep neural network to decode the V1 neuron responses. To reconstruct AP and NS visual stimuli, the decoding process consists of two stages: the signal converter part sampling neural signals to pixels, and the U-Net part usually called as auto-encoder. The input calcium signal is an array consisting of M vectors, each length is N, where M is the number of stimuli and N is the number of all activated neurons. We first used a multi-layer perceptron to convert the input calcium signal into a vector of the same size as the reconstruction target visual stimuli, so as to map every cell response to every pixel in reconstruction images. In our experiments, the central region of images is taken as reconstruction object, which is ensured to be covered within monkey's receptive fields. Then, in the U-Net part, the converted vector is used as input, and after an auto-encoder similar to U-Net, the reconstructed visual stimulus is generated. Inspired from a typical full convolutional network usually used for image segmentation, in the first half of the autoencoder, the input vector is convolved and down-sampled to fully extract the signal features. In the second half, the network starts upsampling and convolution, and finally recovers a reconstructed image of the same size as the visual stimulus. It should be noted that in the network structure of the autoencoder part, the 'skip connection' structure similar to U-Net is added. Skip connections between network layers of different depths can merge the low-level features (shallow network) and high-level features (deep network) in the process of reconstruction (see Fig 1), which can be both captured in visual stimuli. As a result, more details and accurate positions can be decoded under the condition of small datasets. According to the characteristics of calcium signal and visual stimulus in our data, we used three skip connections between different layers of the network structure. Besides, In order to ensure smooth gradient propagation, we used the batch normalization layer behind activation layer following convolution operation. Spatial dropout

layer was also incorporated in U-net part for preventing overfitting. In order to optimize our network, we used the back-propagation algorithm to perform end-to-end training on the calcium signal responses and corresponding visual stimuli, and the objective function is the mean square error (MSE). Benefiting from the end to end feature, our decoder can decode visual stimuli directly from the neural responses, without intermediate processing.

The presented image sizes in the experiments of AP and NS are different. While AP stimuli are 160 * 160 pixels, NS stimuli are 236 * 236 pixels. To ensure that reconstructed stimuli are covered by the receptive fields of subject monkey's V1 neurons, we cropped the input stimuli into the central region of images, i.e., 40*40 for artificial patterns and 60*60 for natural stimuli. Then, our decoding model structure was modified at layer level according to two datasets. For cross-stimulus decoding, we cropped images into a size of 60 * 60. For the signal converter part, the input shape was set as the neuron response array size in respective dataset. Number of neurons in middle dense layer is 512, followed by corresponding batch normalization, activation, and dropout operations. The output shape of the signal converter was set as the target reconstructed image size, i.e., 1600 for AP and 3600 for NS. For the U-Net autoencoder part, the whole information processing procedure can be split into two stages. In the first stage, we used convolution and down-sampling to process and decrease the size of the input with the target-image size. The kernel sizes of four layers in the first stage are (40,40,64), (20,20,64), (20,20,128), (10,10,256) for AP, (60,60,64), (30,30,64), (30,30,128), (15,15,256) for NS. Another four layers in the second stage coincide with the reversed order of structure in their respective first stage. The function of the second stage is to recover the down-sampled images to the target reconstructed size through up-sampling. From introduced structure change above, we can see the stride operation size in our model is (2,2) for down and up sampling. The down-sampling operation in the first stage is realized by MaxPooling2D function while the up-sampling operation in the second stage is realized by Upsampling2D function in tensorflow. ReLU is chosen as the activation function in our overall model structure. All convolution layers in our model are followed by a batch normalization layer. Besides, in the middle layers (the ones of smallest kernel size), SpatialDropout2D layers were added behind the batch normalization layer so as to ameliorate overfitting.

We trained both AP and NS model with Adam method and batch size of 80 to update network parameters, while the learning rate of AP is 0.003, NS is 0.0001. The training epochs of AP are controlled by early-stop mechanism, i.e., when the validation error (MSE in test set) is not decreasing, the model parameter update will be stopped. While the training epochs of NS is artificially controlled according to the reconstruction result, for the reason that the convergence of NS model is difficult to reach. The model was implemented with Keras deep learning library, Tensorflow as backend, employed on Nvidia v100 super graphics card. Learning rate-customized Adam method was used to train the model (Kingma and Ba, 2017).

We compared our model with several state-of-the-art neural decoding models.

1) Spike-Image Decoder (SID) (Zhang et al., 2020): The SID implemented

end-to-end training with reconstruction constraints. The experimental data used in its original article is spike trains from retinal ganglion cells in the salamander.

2) Deep Generative Multi-view Model (DGMM) (Du et al., 2019): The DGMM used two view-specific generators with a shared latent space. Its original paper used this method to reconstruct visual stimuli from the human brain activities measured by functional magnetic resonance imaging (fMRI).

3) Conditional Generative Adversarial Network (CGAN) (Shen et al., 2019): The CGAN is a variant of generative adversarial networks (GAN) (Dosovitskiy and Brox, 2016). The fMRI data vector was used as the condition to reconstruct the corresponding stimulus image, which would be generated through the generator network in CGAN.

4) Bayesian canonical correlation analysis (BCCA) (Fujiwara et al., 2013): BCCA is a multi-view linear generative model designed for neural encoding and decoding, in which mappings between a set of pixels in a visual image and a set of voxels in an fMRI acitivity pattern are estimated.

2.4 Image reconstruction metrics

In neural stimulus reconstruction, it is not always easy to perceive the differences between reconstructed images and original stimuli. For this reason, we evaluated the quality of reconstruction using different image assessment metrics. Six full-reference metrics were applied to compare reconstructed images with original images. Referring to development of image quality assessment field, this set of metrics would evaluate the similarity from different aspects. Individual characteristics of six metrics are briefly introduced as follows.

1) Mean Square Error (MSE): MSE equals the final expectation of the squared error between desired and original values. Given an original image I and its reconstruction K, MSE is defined as:

$$MSE = \frac{1}{m\,n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \tag{1}$$

2) Peak Signal-to-Noise Ratio (PSNR): The PSNR (unit is decibel) is defined as:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \tag{2}$$

MAX_I is the maximum possible pixel value of the original image, that is 255. Besides, the larger PSNR, the better image quality, and the range of PSNR is not limited.

3) Structural Similarity Index Metric (SSIM) (Zhou Wang et al., 2004): SSIM is designed based on the assumption that human visual processing system is able to extract structural information in scenes highly adaptively. SSIM index is calculated on various windows of an image. Luminance (l), contrast (c), and structure (s) are included when measuring two windows x and y. SSIM value is in the range $[0, 1]$. The larger SSIM value is, the more similar the reconstructed image

is with the original ones.

$$\begin{cases} l(x,y) = \dfrac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x,y) = \dfrac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\ s(x,y) = \dfrac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{cases} \tag{3}$$

μ_x and σ are the mean and the variance for the corresponding window. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are the constant, $c_3 = \frac{c_2}{2}$. L is located in the range of pixel value range, i.e., $[0\,255]$. $k_1 = 0.01$ and $k_2 = 0.03$. We can get SSIM equation as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^\alpha \cdot [c(\mathbf{x},\mathbf{y})]^\beta \cdot [s(\mathbf{x},\mathbf{y})]^\gamma \tag{4}$$

where α, β and γ equal to 1.

4) Most Apparent Distortion (MAD) (Larson and Chandler, 2010): MAD attempts to rate image quality from two strategies: detection-based perceived distortion in high-quality images, appearance-based perceived distortion in low-quality images. A combination of these two measures is thought effective in predicting subjective ratings of image quality. MAD should be a nonnegative value. The larger MAD value is, the worse the image quality is. The equation is given by:

$$\mathrm{MAD} = (d_{\text{detect}})^\alpha \, (d_{\text{appear}})^{1-\alpha} \tag{5}$$

d_{detect} and d_{appear} are measures by specific processes for distortion in high-quality and low-quality image levels, respectively. The weight $\alpha \in [0, 1]$ is chosen according to overall level of distortion.

5) Feature Similarity Index (FSIM) (Zhang et al., 2011): FSIM is inspired from the fact that human visual system (HVS) understands an image according to its low-level features. Specifically, the phase congruency (PC), a dimensionless measure of the significance of a local structure (Morrone et al., 1986), is used as the primary feature in FSIM. Considering the contrast invariance property of PC, the image gradient magnitude (GM) is employed as the secondary feature in FSIM. The range of FSIM value is the same as SSIM. FSIM is defined as:

$$\mathrm{FSIM} = \frac{\sum_{\mathrm{x}\in\Omega} S_L(\mathrm{x}) \cdot PC_m(\mathrm{x})}{\sum_{\mathrm{x}\in\Omega} PC_m(\mathrm{x})} \tag{6}$$

where $S_L(\mathbf{x}) = [S_{PC}(\mathbf{x})]^\alpha \cdot [S_G(\mathbf{x})]^\beta$, usually $S_L(\mathbf{x}) = \mathrm{S}_{PC}(\mathbf{x}) \cdot \mathrm{S}_G(\mathbf{x})$ for simplicity. $S_L(\mathbf{x})$ represents the similarity at each location $\mathbf{x}$ combining PC similarity $S_{PC}(\mathbf{x})$ and GM similarity S_G, whose computation process is omitted here for briefness. $PC_m(\mathbf{x}) = \max(PC_1(\mathbf{x}), PC_2(\mathbf{x}))$ is used as weight for the importance of $S_L(\mathbf{x})$ in overall similarity between two images.

6) Gradient Similarity (GSM) (Liu et al., 2012): Gradients convey important

visual information and are crucial to scene understanding. Structural and contrast changes can be captured through gradients. In addition, luminance changes affect image quality a lot. GSM, whose value is lying in $[0, 1]$, integrates changes in luminance and contrast-structure via an adaptive method to obtain overall image quality score. The larger GSM value is, the better reconstructed image quality is. The proposed gradient similarity is defined as:

$$g(x, y) = \frac{2g_x g_y + C}{g_x^2 + g_y^2 + C} \tag{7}$$

where g_x and g_y are the gradient values for the central pixel of image x and y. C is the small constant to avoid the denominator being zero.

3 Results

3.1 Reconstruction stimulus images from two-photon calcium neural signals

We first developed a deep neural network model to reconstruct stimulus images from two-photon calcium imaging neural signals recorded from layer 2/3 V1 neurons in awake macaque monkeys (Tang et al., 2018b). The traditional CNN utilization is mostly inclined to decision-oriented tasks, of which the purpose is the extraction of key knowledge. On the contrary, the reconstruction is a process restoring the original information, where the extracted features need to be supplemented. The U-Net structure has been widely used in medical image segmentation (Ronneberger et al., 2015), providing high-accuracy pixel-level segmentation. Inspired by this, we added the skip connection in our decoding model, allowing the latter layers to have access to the more complete information in the former layers. As illustrated in Fig 1, our model can be mainly regarded as two parts: the signal converter and the U-Net autoencoder. For the signal converter part, the input shape was set as the neuron response array size in the respective dataset. The output shape of the signal converter was set as the target reconstructed image size. For the U-Net autoencoder part, the whole information processing procedure can be split into two stages. In the first stage, we used convolution and down-sampling to process and decrease the size of the input with the target-image size. Another four layers in the second stage coincide with the reversed order of structure in their respective first stage. The function of the second stage is to recover the down-sampled images to the target reconstructed size through up-sampling (see Methods).

The experimental data were collected in a previous study (Tang et al., 2018b), including two sets of visual stimulus images: simple artificial pattern (AP) and complex natural scene (NS) images (see Methods). AP dataset has 5000 images and 1142 activated neurons, while NS dataset has 2250 natural images and 1225 activated neurons. We divided both datasets into 9:1, 90% for training and 10% for test. We compared the capability of reconstruction with our model

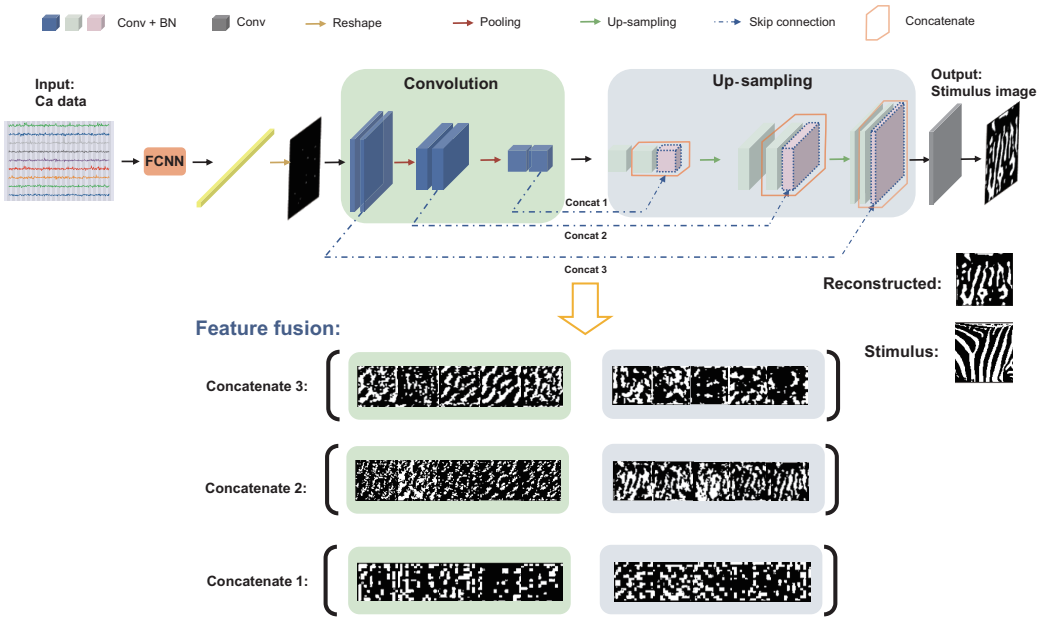

Figure 1: **Illustration of reconstruction network model.** The input calcium response vector is converted to one vector whose length is the product of length and width of stimulus images, through a multi-layer fully connected neural network (FCNN). The output is then transferred to U-Net like autoencoder. Cubes of different colors indicate different feature maps through different operation layers in the network for convolution (Conv), batch normalization (BN), spatial dropout and concatenate layers. The output of autoencoder represents the stimulus image reconstruction. Notably, three skip connections were employed during the autoencoder process. Respective features at symmetrical layers are merged through skip connections (Concatenate 1,2, and 3 indicating three feature mergence). Using a strip image for illustration, the low-level features (shallow network) and high-level features (deep network) during reconstruction complement each other for a better reconstruction effect.

over AP and NS datasets with several state-of-the-art neural decoding models: Spike-Image Decoder (SID) (Zhang et al., 2020); Deep Generative Multi-view Model (DGMM) (Du et al., 2019); Conditional Generative Adversarial Network (CGAN) (Shen et al., 2019); and Bayesian canonical correlation analysis (BCCA) (Fujiwara et al., 2013). Out of which, SID is a recent model developed to reconstruct stimulus images from neural spiking signals, which also works well for fMRI signals (Zhang et al., 2020). Fig 2 shows the outcome decoding results of different models. To quantify the performance of image reconstruction, we used six different metrics (MSE, PSNR, SSIM, GSM, FSIM, and MAD, see Methods) for image quality assessment as a reconstruction index. Besides common measures of MSE, PSNR and SSIM, we selected another three more novel measures (GSM, FSIM, and MAD), since that it is still debated that which metric gives a more reasonable and accurate description of image quality (Dosselmann and Yang, 2011; Pedersen and Hardeberg, 2012; Hore and Ziou, 2010; Wang et al., 2002). With the normalized values of these metrics (Fig 2B, see the raw values in Table S1 and Table S2),

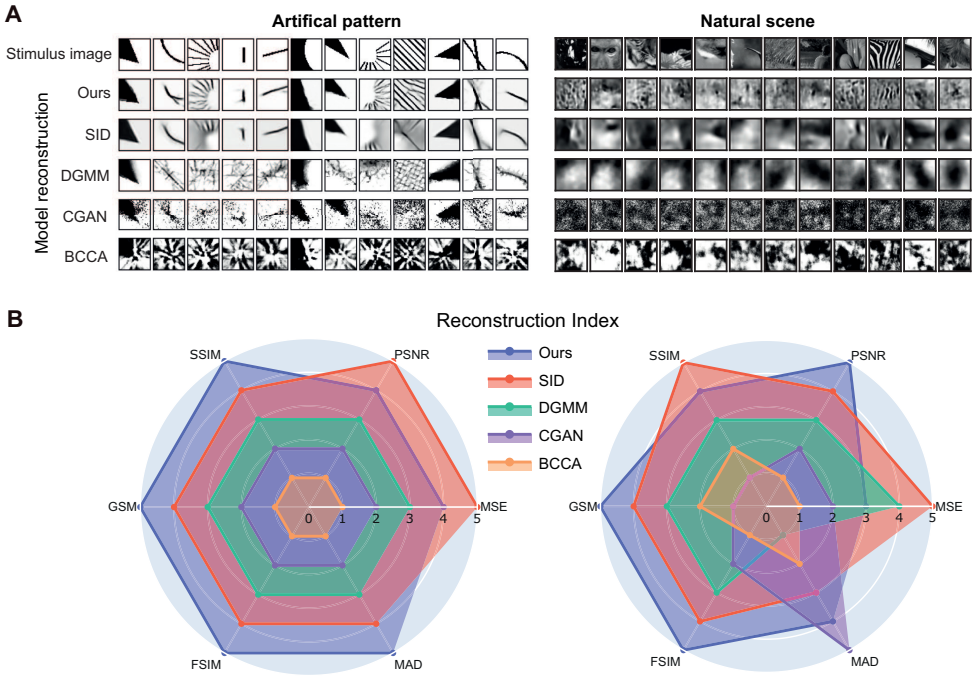

Figure 2: **Reconstructed stimuli and indices of reconstructions from different neural decoders.** (A) Illustration of reconstructed artificial pattern (left) and natural scene(right) images through different neural decoding models. (B) Radar plots of reconstruction indices for artificial pattern (left) and natural scene (right) from different decoders. In each radar plot, each polar axis shows one metric with the outer coordinate indicating better performance.

our current model produces better reconstructions than the compared methods for both AP and NS images, whereas SID is the second best model. In addition, reconstruction in AP test set shows our model is more accurate in artificial shapes. Because of the complexity of natural images and relatively small samples of NS dataset, reconstructions of our model in NS test set are not so clear. Nevertheless, our model still outperforms other methods. Meanwhile, some natural textures, e.g., the zebra stripes shown in Fig 2A, are relatively better reconstructed. These results imply that our current model has a capability for reconstructing pixel-level images using calcium imaging neural signals. Furthermore, we conducted ablation experiments showing that better performance was achieved by skip connections compared to the networks without them (Table S3).

Recent studies suggested low-pass spatial features of stimulus images could be efficiently decoded through a neural decoder of retinal spikes (Kim et al., 2020), where the neural decoder resembled a low-pass image filter. Here we also examined this relationship between our decoded images and low-pass filtered ones. We processed the original natural images in the AP and NS dataset through low-pass filters at different frequencies, in which the low frequency, such as 1 Hz, filter would blur the details of images while keeping the global feature (Fig 3A). Using reconstructed images as the reference images, we quantified the similarity and computed the image quality metrics between them and low-pass filtered images.

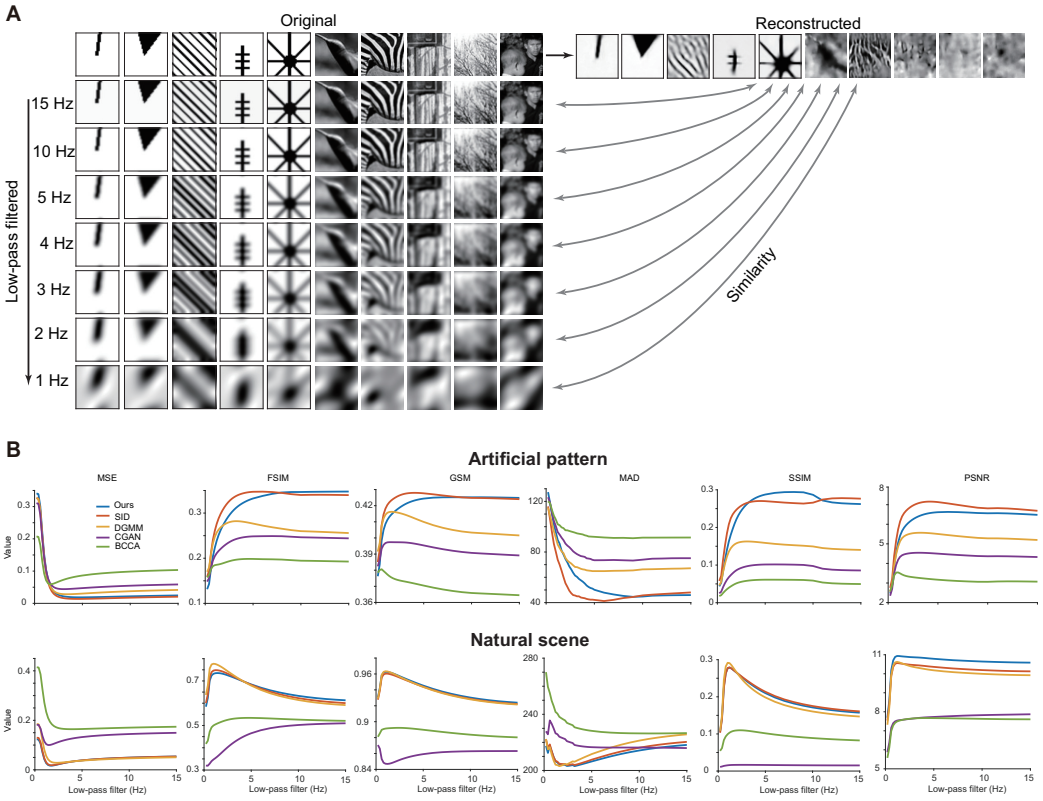

Figure 3: **Frequency low-pass filter and V1 neural decoding.** (A) Illustration of artificial patterns and natural images after frequency low-pass filtering. The corresponding frequency values were labeled on the left. Original images were put in the first row, and reconstructed images through our neural decoder were put on the right side. (B) Image reconstruction metric change through low-pass frequency reference images. Different neural decoders were indicated with different colors. Values on the x-axis are the corresponding low-pass frequencies.

The change course of six metrics over a range of filter frequencies revealed that the reconstructed images indeed match to filtered images with a low frequency of 1-3 Hzs, depending on the decoding models. Notably, these tendencies also separate different decoding models: our current model and SID model have the highest performance and behave closely in both AP and NS images. DGMM is close to ours and SID in NS images only, while CGAN and BCCA are significantly different. These analyses suggest that different decoding models may capture different sets of features, similar to the way different details of image textures or contexts are captured by different image assessment metrics. We will leave out other models, and focus on the behaviours of our current model in detail below.

3.2 The effect of sparse encoding on reconstruction

It is generally believed that the number of input samples influences decoding models. Particularly for the model based on deep learning is highly demanding on input data. For decoding of stimulus in neuroscience, the number of cells has also

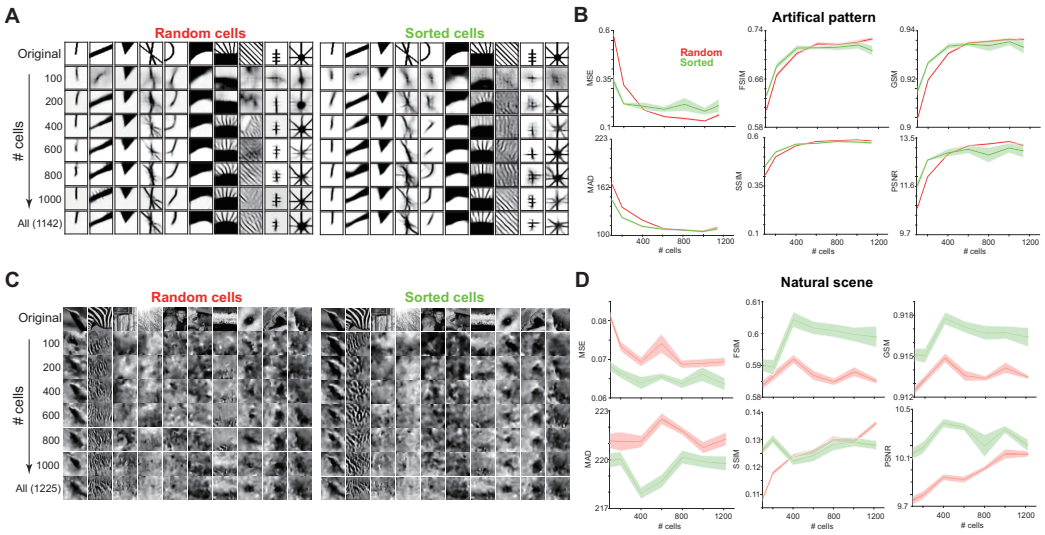

Figure 4: **Image reconstruction using different numbers of cells.** (A) Reconstructed artificial patterns with different numbers of cells selected randomly (left) and sorted according to their maximum response to all stimuli (right). (B) Change of the reconstruction metrics over different cell numbers used for decoding. The solid lines are the average metric values over all 10 runs, and the spread areas are the standard error. (C) (D) Similar to (A) (B) but for natural scene images. For comparison, the same scale was used for visualizing metric values of both types of stimulus images.

been playing a role in the decoding performance (Zhang et al., 2020; Yoshida and Ohki, 2020) and more cells are better for improving the decoding outcome. Here, we tested the reconstructions using different number of cells. Our model used the input calcium signal as an array comprising M vectors whose length is N, where M is the number of stimuli and N is the number of all activated neurons, where N is 1142 in AP and 1225 in NS. In the all-cell decoding described above, N remained unchanged. We now changed N on different scenarios and examine the effect of decoding. The simplest scenario is to use a different number of cells chose randomly from all cells. Another scenario is that we first sorted all activated cells according to their responses over all stimulus images from high to low. The resulting reconstructions and metrics of AP and NS images in both scenarios are shown in Fig 4. For AP images, the reconstructed image quality grew better as the number of cells increased, over both random and sorted cell scenarios. The change course of all six metrics in Fig 4B shows that metrics are increasing as the cell number increased and saturate when there are around 500 cells used. In contrast, for NS images, the performance metrics are not changing significantly over a large range of cell numbers considering the vertical scale in Fig 4. Specifically, metrics of the overall random case were not good as the sorted case. But in SSIM/PSNR, the reconstructed image quality grew better as the number of cells increased while fluctuated for the sorted case.

Compared to the random case, metrics have shown that the sorted cell case attained a better reconstruction than the random ones. To further examine this in

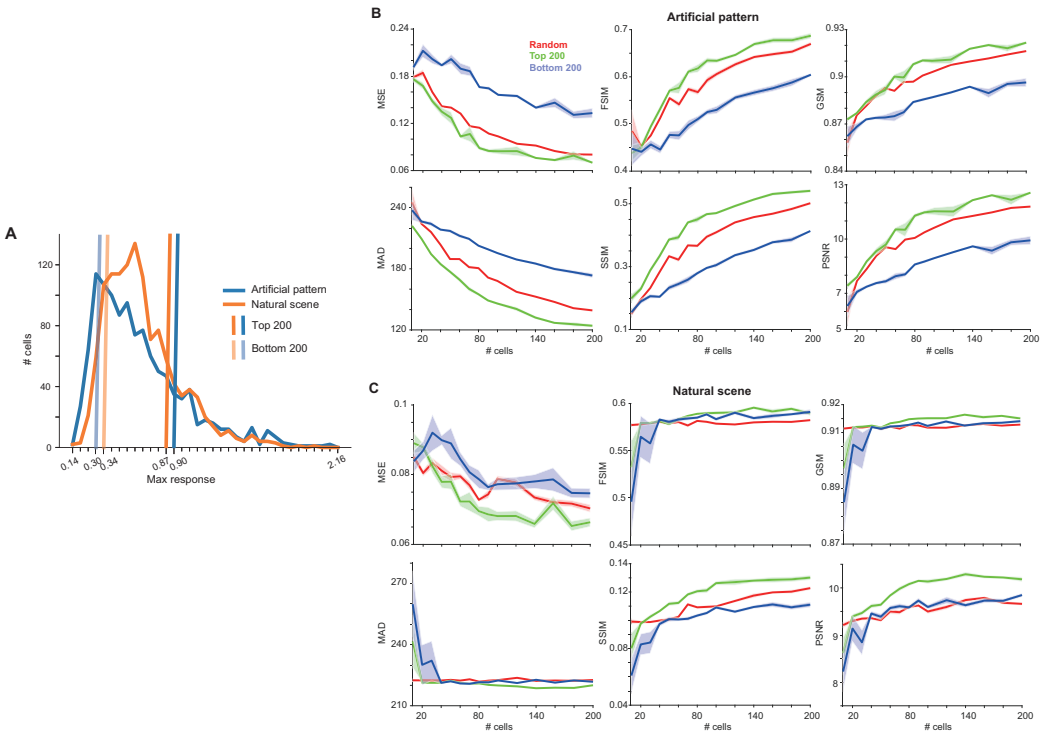

Figure 5: **Reconstruction metrics decoded by 200 cells with different settings.** (A) The distribution of maximum responses over all stimulus images: artificial pattern (orange) and natural scene (blue). Bottom 200 cells are indicated by light colored vertical lines; Top 200 cells by dark. (B) Reconstruction metrics for artificial pattern images over 200 cells in three conditions: randomly selected, top 200, and bottom 200. The solid lines show the average metric values over all 10 runs, and the shadow area are the standard error. (C) Similar to (B) but for natural scene images.

detail, we refine three more scenarios in a more detailed input range: random 200 cells, sorted all cells to utilize top 200 cells with stronger responses and bottom 200 cells with weaker responses. The distribution of neural maximal responses to any of AP and NS stimulus images shows that AP generates weaker responses while NS triggers stronger responses, shown in Fig 5A, consisting with the view that V1 neurons are more selective to respond to natural scenes (Yoshida and Ohki, 2020; Tang et al., 2018b; Yoshida and Ohki, 2018). The change course of all six metrics in this refined scenarios revealed a slightly different between neural encoding of AP and NS. For AP images, metrics are rather interpolated between three cases: Top 200 is the best and Bottom 200 is the worst, while Random 200 is ranked between these two. This shows that V1 neural responses to AP are distributed over all cells, e.g., a population code. In contrast, for NS images, Top 200 is the best, while the other two cases are intervened. Thus, cells with stronger responses contain more information for decoding. This is a sign that V1 neurons are more selective to decoding natural scenes, consisting with the notation that V1 neurons are sparse coding for natural images (Tang et al., 2018b). These results suggest that our decoding model can capture the characteristic of neural sparse

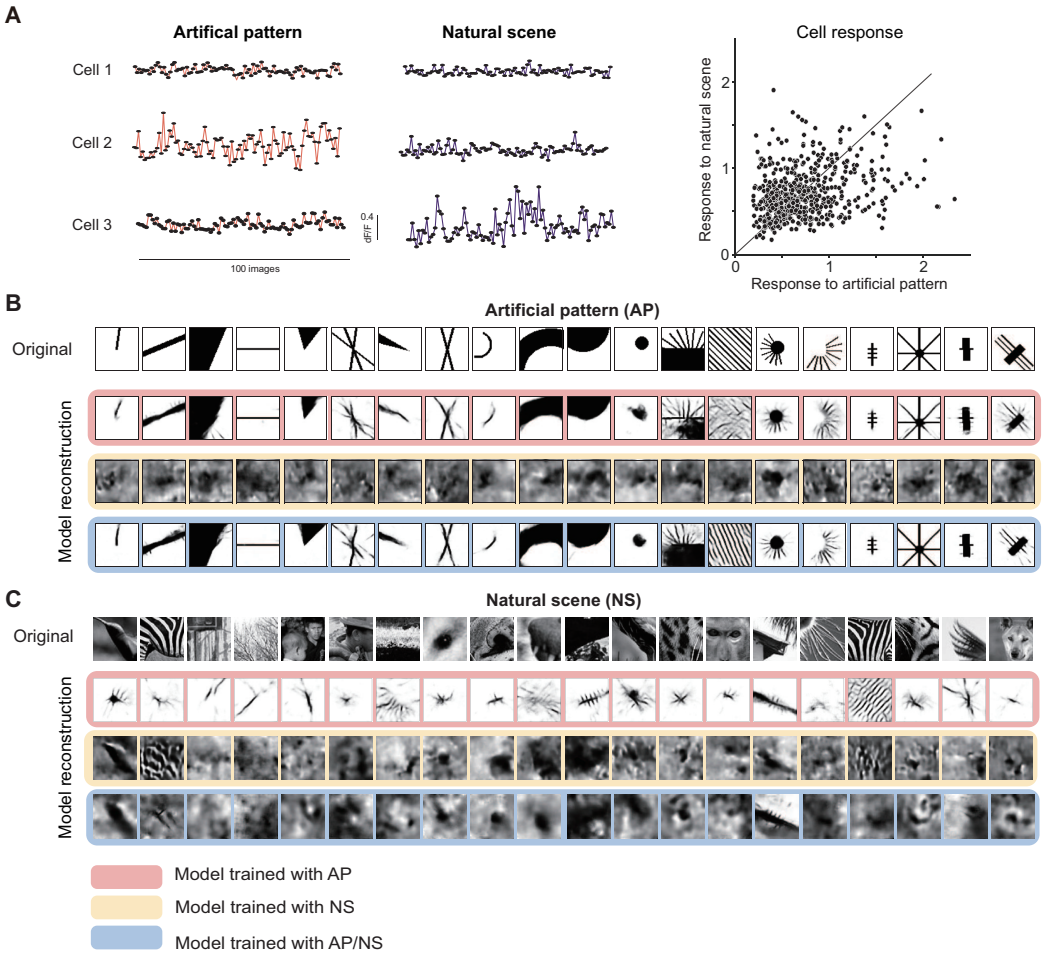

Figure 6: **Image reconstruction with cells responded to both artificial pattern and natural scene.** (A) Three example cells show diverse responses to stimuli of 100 AP and 100 NS images. (Right) The distribution of 599 cell responses with AP and NS image stimuli. (B) Reconstructed AP images using the models trained by AP images (top), NS images (middle), and mixed images. (C) Similar to (B) but for NS images.

encoding and provide a way to differentiate population coding from sparse coding with a given set of stimulus images.

3.3 Salient features of natural scenes revealed by AP-trained models

With the notation of distinguished profiles of neural encoding of AP and NS, we further investigated the detailed visual features encoded by neurons using our decoding model. The benefit of our dataset is that there are a subset of neurons recorded under both AP and NS images. It enables us to have access to the responses of the same V1 cells over different types of stimuli. We identified 599 cells both activated to AP and NS images and exploited how these cells encode different visual features combining with different decoding scenarios. The num-

ber of stimuli in the AP and NS dataset used here is the same as above, while we only focus on these selected cells for decoding. Fig 6A shows the overview of neural responses to AP and NS images with three typical cells showing no preference between AP and NS (Cell 1) and more activated for AP (Cell 2) or NS (Cell 3). The population plot with mean responses to either AP or NS shows that cells have slightly more response to AP images.

We then considered three cases of either training the decoding model using AP or NS, or both mixed, while tests were done by using all the mixed images. Not surprisingly, the model trained by mixed AP and NS generates generated both AP and NS images similar to those trained by each type of images (Fig S1). Interestingly, reconstructions of NS with the model trained by AP reveals a set of features resembling to salient features of that particular natural scene (shown in Fig 6B and C). For instance, the natural image whose content is zebra stripes are in line with similar artificial patterns. However, reconstructions of AP with the NS-trained model yield over-representation of image textures and far from the targeted APs.

To further characterize these effects, we categorized all AP images into six classes: composition (CO), cross (CX), orientation (OT), curvature (CV), and corner (CN), according to their specific textures. Then using the salient features of NS images obtained by the decoding model trained by AP, we then classified all reconstructed NS images into the same five categories through dimension-reduction t-SNE (t-distributed stochastic neighbor embedding) technique (Van der Maaten and Hinton, 2008) (Fig 7), which has been utilized in neuroscience studies (Wang et al., 2021; Xu et al., 2021). In the end, both types of images were separated into these five categories. In this way, the natural image whose content contains, for example, a sharp shape has generated a CN (corner) type artificial pattern. This is in line with the encoding of local low- and high-order features by V1 neurons (Tang et al., 2018a). When measuring the reconstructed AP images using different metrics, different AP categories have different levels of metrics (Fig 7B, see more metrics in Fig S2). Particularly, as the level of scene complexity decreases, from the highest CO, to lower ones: CX, OT, CV, and CN, reconstruction metrics show the same trends as scene complexity, e.g., CO has the worse image quality than OT. Similarly, different categories of NS images also have distinguished reconstructed metrics (Fig 7C and Fig S2). Altogether, our model reveals the properties of neural encoding from the perspective of neural decoding. These results show that our decoding model provides a way to characterize visual features embedded in images pixels through reading out the code using neural responses.

3.4 Low-dimensional image distribution

Due to the difficulty of direct distinction of reconstructed NS images, we compared the low-dimensional embedding representations of NS and reconstructed NS images using the t-SNE technique. For the illustration clarity, 225 images (10 percent of all images) were randomly chosen from the NS dataset. Then these 225 NS images and the corresponding reconstructions were pooled and visualized by t-SNE. We found that their low-dimensional embedding distributions were close

to each other (Fig 8A), and the main visual patterns of some reconstructed NS images are similar to their respective original NS ones, in different regions of both distributions. This observation corroborates the effectiveness of our NS reconstruction.

We then studied the embedding result when AP and NS images are all pooled into the t-SNE dimension reduction process. Specifically, to ensure the NS images can be sufficiently surrounded by the AP images as many as possible, 225 randomly chosen NS images and all 5000 AP images were pooled. As seen in Fig 8B. The large number of AP and NS images are respectively clustered with their own categories. Although the AP images are very different with the NS ones over the image complexity, there are still two close clusters formed with AP and NS images shown by some examples of the AP and NS images from these two clusters (Fig 8B). Among the images in each cluster, the AP and NS images stay close, whose salient features are similar, similar to the illustration in Fig 7A. Taken together, the t-SNE dimension reduction corroborated our analysis of decoding AP and NS connections.

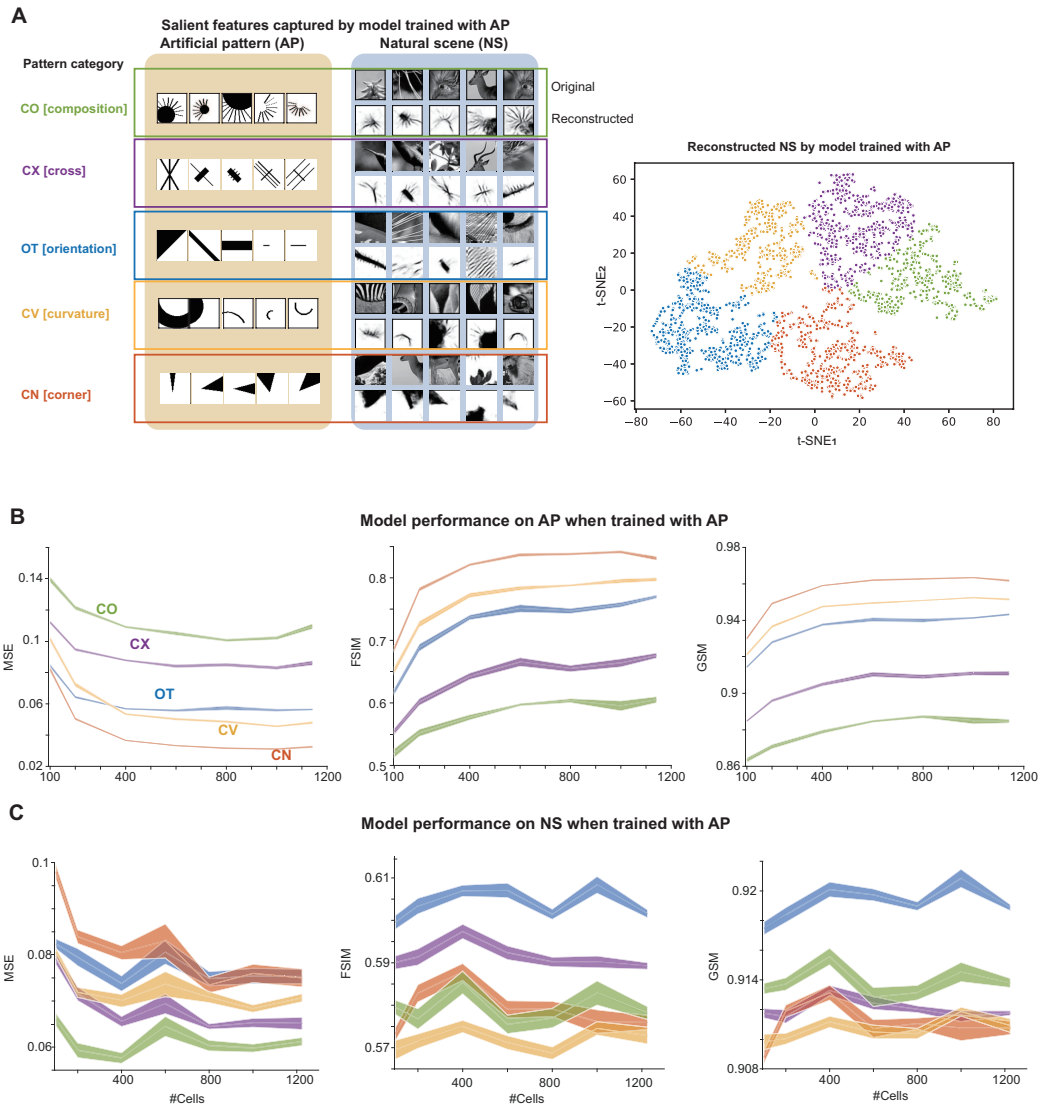

Figure 7: **Salient features of natural scenes revealed by AP-trained models.** (A) (Left) Examples of reconstruction of NS images with a model trained with AP images for five categories (CO, CX, OT, CV, and CN). (Right) Reconstructed NS images are clustered into the same categories visualized by t-SNE. (B) Change of metrics of reconstructed AP images on five categories over cells. The solid lines are the average metric values over all 10 runs, and the shadow areas are the standard errors. (C) Similar to (B), but for reconstructed NS images.

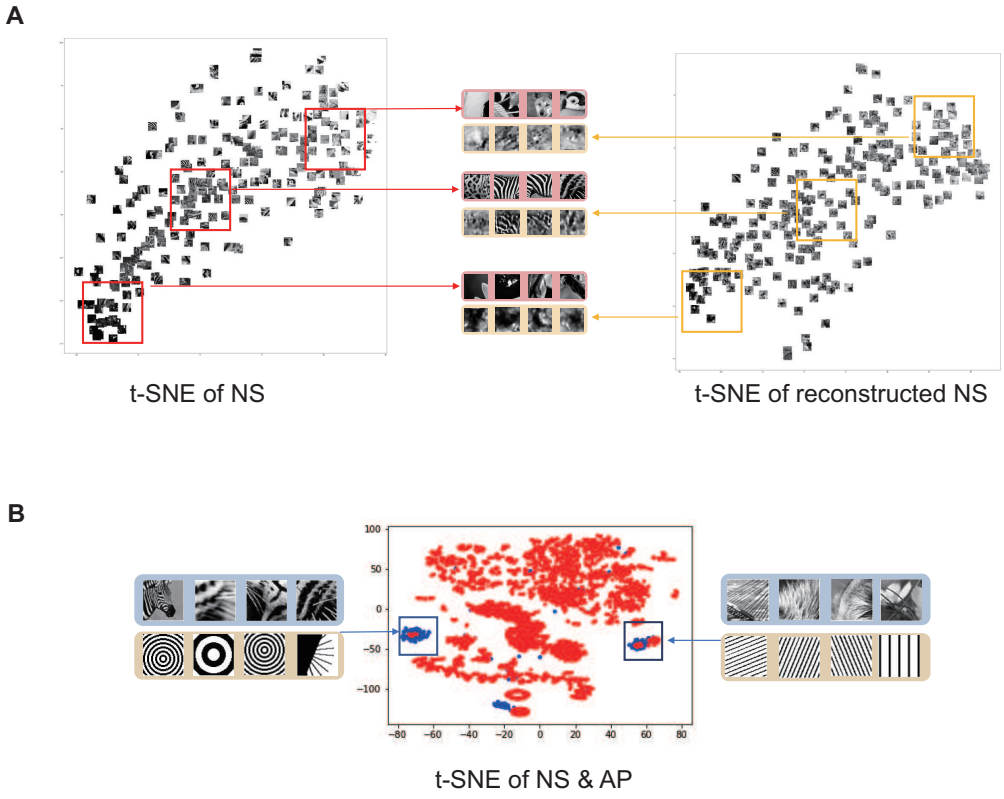

Figure 8: **Distribution of image structures using t-SNE.** (A) The persevered low-dimensional distribution of NS images and reconstructed NS images. (Left) The NS images visualized by t-SNE technique. (Middle) Example images of both types. (Right) The reconstructed NS images visualized by t-SNE. (B) The persevered low-dimensional distribution of the original NS and AP images shown by t-SNE (red for AP, blue for NS). Example images on left and right chosen from the corresponding clusters formed by both AP and NS images.

4 Discussion

In this work, leveraging a decoding model with a set of image reconstruction metrics, we analysed neural coding of visual features using a large scale two-photon calcium signal data collected in awake macaque monkeys. We found that a sparse set of V1 neurons with high responses contributed more valuable information than randomly selected neurons. We also decoded a latent relationship between artificial patterns and natural scenes that is not easily perceptible. The responses of natural images in V1 neurons contain salient features of complex scenes, which can be revealed by the reconstructed NS images through a AP-trained decoding model. These results suggest a valuable approach for studying of neural encoding from a perspective of neural decoding.

4.1 Decoding models

In the last decades, varied methods of neural decoding of stimulus images have been developed. Traditionally, a neural decoder can be optimized with linear and nonlinear statistical methods (Stanley et al., 1999; Marre et al., 2015; Yoshida and Ohki, 2020; Botella-Soler et al., 2018). Due to the limited representation power of these structures, the decoding tasks were usually simple or performances were not so satisfied to reveal pixel-level information from fMRI neural signals (Naselaris et al., 2009; Fujiwara et al., 2013; Nishimoto et al., 2011).

In recent years, deep learning techniques, especially deep neural networks (DNNs), are implemented for neural decoding. It was suggested that the DNN computing has correspondence with biological visual systems (Horikawa and Kamitani, 2017; Güçlü and van Gerven, 2015; Wen et al., 2018; Cichy et al., 2016; Du et al., 2019; Baek et al., 2019,?; Yan et al., 2020; Wang et al., 2021; Zheng et al., 2021). For neural signals with single-cell resolution, DNNs were used to reconstruct natural images directly (Parthasarathy et al., 2017; Zhang et al., 2020; Brackbill et al., 2020). The concept of simple and complex cell in V1 has given an indication to computation model development from a micro view (Fukushima and Miyake, 1982; Riesenhuber and Poggio, 2000), e.g., the widely used CNN sub-module in deep learning. Our results here could bring some inspiration for neural decoding, brain-computer interface and neural prosthesis from a different perspective. Some recent studies use a combined approach of synchronized encoding and decoding process through dual deep generative model optimizing both encoding and decoding performance (Zhou et al., 2020), which could be an interesting topic for future work. Given the recent advances of neural recording techniques for brain-machine interface, it is highly demanding to develop decoding methods that are able to reconstruct pixel-level images with high precision, particularly for visual neuroprothesis devices (Shah and Chichilnisky, 2020; Yu et al., 2020).

4.2 Visual decoding from calcium imaging data

Visual decoding from calcium imaging data has a relatively sparse history than other neural signals. Recently the calcium activity collected from small numbers of high-responding neurons has been used to decode movie scenes with nearest mean classification (Kampa et al., 2011). At the population level, with the calcium imaging data from a larger (˜500) population of V1 neurons under natural and phase-scrambled movie stimuli, simple linear classification has been implemented for a classification decoding task to reveal those holistic activities of primary visual cortical neurons in anesthetic and awake mouse (Froudarakis et al., 2014). When exploiting different machine learning architectures on calcium data under 118 unique natural scenes, one can improve the classification accuracy (Ellis and Michaelides, 2018).

Besides classification, the full-stimulus reconstruction has also been explored recently. Calcium imaging of 103 neurons in mouse primary visual cortex were used to reconstruct the complex natural stimuli with an optimal linear estimator and revealed that V1 neurons display linear readout properties with low information in the joint distribution of neural activity (Garasto et al., 2018). Further, with the help of receptive field data and simulations from a linear-nonlinear Poisson model, reconstruction performance could be ameliorated if the receptive field are more uniformly sampled in the subject's visual field (Garasto et al., 2019). Besides, with single neurons' activity recorded from anesthetized mouse V1, a linear regression method was used to extract visual feature values contributed by each cell's calcium imaging responses and reconstruct original natural stimuli (Yoshida and Ohki, 2020). In line with our results, it was demonstrated that natural images can be reliably represented by a few highly responsive neurons (Yoshida and Ohki, 2020; Tang et al., 2018b).

4.3 Salient features of natural scenes related to artificial patterns

With both artificial and natural images responded by the same neurons in the macaque visual cortex, these responses must contain a certain quantity of information that could be helpful for primates to distinguish from different visual feature patterns. The results in our experiment opportunely explored these points from the perspective of neural decoding. Despite those relatively clear reconstructions with NS responses on model trained with AP dataset, reconstructions with AP responses on model trained with NS dataset is difficult to be identified (Fig 6). Presumably the poor capability of reconstructing complex natural images is resulted from the limited biological dataset contributed to an unsatisfied result in the decoding model. Ideally, if the biological dataset is large enough which results in a dramatically increasing reconstruction ability, the distinguishable artificial patterns could be generated with corresponding AP responses, benefiting from the reconstruction ability strong to generate stimuli of natural image complexity. This can be seen that reconstructions with AP or NS response with a model trained on both were almost same as those trained separately. Thus, the

mixture training enriched the reconstruction feature space.

When a natural image whose main context is some artificial pattern, then the V1 neural responses would strive to embody the sensitivity corresponding to this artificial pattern. The reconstruction whose information source is these responses would generate an image that reflects the perceptual pattern to the greatest extent. Thus, with a model trained by rich sets of APs, it enables us to extract salient features of natural scenes per se. A more refined picture comes from the different categories of AP images are in line with those salient features of NS images.

A related line of research is focusing on face recognition (Baek et al., 2019; Wang et al., 2021) establishing the relationship between artificial DNN face recognition system and real human face recognition. The DNN with face-selective neurons can perform well when the original celebrity faces are transformed to various cartoon styles and the response profile was similar to that with the original stimuli (Wang et al., 2021). It was found that face-selective neurons can be emerged in networks trained for non-facial natural images (Baek et al., 2019) so that in face-selective neurons, face-like shapes including components such as eyes, nose, and mouth were observed in preferred feature images. Our results along with these findings have revealed that, there is a strong correlation between complex natural image perception and artificial pattern recognition in primates visual system.

4.4 Neural responses to artificial patterns and natural scenes

In the research on biologic visual mechanism, it is common to see that natural scenes or artificial stimuli have been used as the visual input (Liu et al., 2017; Liu and Gollisch, 2021). However, in most of these works, researchers always investigated visual mechanism through these two types of stimuli separately. As a result, the observation of biological neural response may be not comprehensive enough due to the limitation in the capability of neural representation for the visual inputs. Researchers have tried to build the relations between natural images and simple stimuli in primary visual cortex through experiments. It was observed that in awake mice, the early visual experience under different exposure conditions of natural scene or grating stimulus would induce behaviorally divergent discriminability in favor of natural scene or grating stimulus (Kowalewski et al., 2021). However, population responses to gratings and natural scenes were suggested to be similar (Miller et al., 2014).

More fine relations between artificial and natural stimuli have been revealed through modeling neural encoding process. It is known that the deep learning technique is inspired by the hierarchical visual feature extraction from brain (DiCarlo et al., 2012). When using neural signal data collected from subjects, i.e., human or animals, more distinct and convincing clues for hierarchical visual feature mechanism have been shown, from low-level feature (simple stimulus) to high-level feature (natural image) in various stages of visual systems in primates (Cadena et al., 2019; Zhang et al., 2019; Ikezoe et al., 2018). In line with these studies, our results revealed different coding strategies of artificial and natural images embedded in V1 neurons, particularly sparse coding, while using a decoding approach.

The classification of AP and NS image representations in brain has a long tradition of studying with neurons at different layers with visual cortex. Some studies suggested that AP and NS images would cause different responses and have mutual effect with each other using different experimental techniques (Jessen et al., 2019; Habib et al., 2015; Rosset et al., 2010; Takacs and Bus, 2016). In our work, the main focus was pixel-level reconstruction, and the classification decoding was not involved. Also, it was found that neurons have distinguished dynamics from different layers of visual cortex (Wang et al., 2020). Future work is need to investigate neural coding of AP and NS images through a series of experiments and analysis in detail.

Acknowledgments

We would like to thank Fang Liu and Shenghui Zhang for helpful discussions of experimental data. This work was supported by the National Natural Science Foundation of China (62176003, 62088102, 61961130392) and Royal Society Newton Advanced Fellowship of UK (NAF-R1-191082).

References

Baek, S., Song, M., Jang, J., Kim, G., and Paik, S.-B. (2019). Spontaneous generation of face recognition in untrained deep neural networks. *Biorxiv*, page 857466.

Botella-Soler, V., Deny, S., Martius, G., Marre, O., and Tkačik, G. (2018). Nonlinear decoding of a complex movie from the mammalian retina. *PLOS Computational Biology*, 14(5):e1006057.

Brackbill, N., Rhoades, C., Kling, A., Shah, N. P., Sher, A., Litke, A. M., and Chichilnisky, E. (2020). Reconstruction of natural images from responses of primate retinal ganglion cells. *eLife*, 9.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597.

Carlson, E. T., Rasquinha, R. J., Zhang, K., and Connor, C. E. (2011). A sparse object coding scheme in area V4. *Current Biology*, 21(4):288–293.

Chang, L., Egger, B., Vetter, T., and Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, 31(13):2785–2795.e4.

Chang, L. and Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, 169(6):1013–1028.e14.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755.

Denk, W., Strickler, J. H., and Webb, W. W. (1990). Two-photon laser scanning fluorescence microscopy. *Science*, 248(4951):73–76.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Dosovitskiy, A. and Brox, T. (2016). Generating Images with Perceptual Similarity Metrics based on Deep Networks. *arXiv:1602.02644 [cs]*.

Dosselmann, R. and Yang, X. D. (2011). A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5(1):81–91.

Du, C., Du, C., Huang, L., and He, H. (2019). Reconstructing Perceived Images From Human Brain Activities With Bayesian Deep Multiview Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8):2310–2323.

Ellis, R. J. and Michaelides, M. (2018). High-accuracy Decoding of Complex Visual Scenes from Neuronal Calcium Responses. Preprint, Neuroscience.

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D., Saggau, P., Bethge, M., and Tolias, A. S. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature Neuroscience*, 17(6):851–857.

Fujiwara, Y., Miyawaki, Y., and Kamitani, Y. (2013). Modular encoding and decoding models derived from Bayesian canonical correlation analysis. *Neural Computation*, 25(4):979–1005.

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

Garasto, S., Bharath, A. A., and Schultz, S. R. (2018). Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex. *bioRxiv*, page 300392.

Garasto, S., Nicola, W., Bharath, A. A., and Schultz, S. R. (2019). Neural Sampling Strategies for Visual Stimulus Reconstruction from Two-photon Imaging of Mouse Primary Visual Cortex. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 566–570.

Gollisch, T. and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, 319(5866):1108–1111.

Güçlü, U. and van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014.

Guntupalli, J. S. and Gobbini, M. I. (2017). Reading Faces: From Features to Recognition. *Trends in Cognitive Sciences*, 21(12):915–916.

Habib, K., Soliman, T., et al. (2015). Cartoons' effect in changing children mental response and behavior. *Open Journal of Social Sciences*, 3(09):248.

Hegdé, J. and Van Essen, D. C. (2007). A comparative study of shape representation in macaque visual areas V2 and V4. *Cerebral Cortex*, 17(5):1100–1116.

Hore, A. and Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE.

Horikawa, T. and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591.

Ikezoe, K., Amano, M., Nishimoto, S., and Fujita, I. (2018). Mapping stimulus feature selectivity in macaque V1 by two-photon Ca2+ imaging: Encoding-model analysis of fluorescence responses to natural movies. *NeuroImage*, 180:312–323.

Jessen, S., Fiedler, L., Münte, T. F., and Obleser, J. (2019). Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie. *NeuroImage*, 202:116060.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.

Kampa, B. M., Roth, M. M., Göbel, W., and Helmchen, F. (2011). Representation of visual scenes by local neuronal populations in layer 2/3 of mouse visual cortex. *Frontiers in Neural Circuits*, 5:18.

Kim, Y. J., Brackbill, N., Batty, E., Lee, J., Mitelut, C., Tong, W., Chichilnisky, E., and Paninski, L. (2020). Nonlinear decoding of natural images from large-scale primate retinal ganglion recordings. *Neural Computation*, 33(7):1719–1750.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kowalewski, N., Kauttonen, J., Stan, P., Jeon, B., Fuchs, T., Chase, S., Lee, T., and Kuhlman, S. (2021). Development of Natural Scene Representation in Primary Visual Cortex Requires Early Postnatal Experience. *Current Biology*, 31(2):369–380.e5.

Koyano, K. W., Jones, A. P., McMahon, D. B. T., Waidmann, E. N., Russ, B. E., and Leopold, D. A. (2021). Dynamic Suppression of Average Facial Structure Shapes Neural Tuning in Three Macaque Face Patches. *Current Biology*, 31(1):1–12.e5.

Larson, E. C. and Chandler, D. M. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, M., Liu, F., Jiang, H., Lee, T. S., and Tang, S. (2017). Long-term two-photon imaging in awake macaque monkey. *Neuron*, 93(5):1049–1057.

Liu, A., Lin, W., and Narwaria, M. (2012). Image Quality Assessment Based on Gradient Similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512.

Liu, J. K. and Gollisch, T. (2021). Simple model for encoding natural images by retinal ganglion cells with nonlinear spatial integration. *bioRxiv*.

Liu, J. K., Schreyer, H. M., Onken, A., Rozenblit, F., Khani, M. H., Krishnamoorthy, V., Panzeri, S., and Gollisch, T. (2017). Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature Communications*, 8(1):149.

Livingstone, M. S. and Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, 4(1):309–356.

Marre, O., Botella-Soler, V., Simmons, K. D., Mora, T., Tkačik, G., and Berry II, M. J. (2015). High accuracy decoding of dynamical motion from a large retinal population. *PLoS Comput Biol*, 11(7):e1004304.

Miller, J.-e. K., Ayzenshtat, I., Carrillo-Reid, L., and Yuste, R. (2014). Visual stimuli recruit intrinsically generated cortical ensembles. *Proceedings of the National Academy of Sciences of the United States of America*, 111(38):E4053–E4061.

Morrone, M. C., Ross, J., Burr, D. C., and Owens, R. (1986). Mach bands are phase dependent. *Nature*, 324(6094):250–253.

Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978). Receptive field organization of complex cells in the cat's striate cortex. *The Journal of physiology*, 283(1):79–99.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6):902–915.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19):1641–1646.

Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, 17(8):1665–1699.

Parthasarathy, N., Batty, E., Falcon, W., Rutten, T., Rajpal, M., Chichilnisky, E. J., and Paninski, L. (2017). Neural networks for efficient bayesian decoding of natural images from retinal neurons. In *Advances in Neural Information Processing Systems*, pages 6434–6445.

Pedersen, M. and Hardeberg, J. Y. (2012). Full-reference image quality metrics: Classification and evaluation. *Foundations and Trends in Computer Graphics and Vision*, 7(1):1–80.

Qiao, K., Zhang, C., Wang, L., Chen, J., Zeng, L., Tong, L., and Yan, B. (2018). Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture. *Frontiers in Neuroinformatics*, 12:62.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Riesenhuber, M. and Poggio, T. (2000). Computational models of object recognition in cortex: A review.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rosset, D. B., Santos, A., Da Fonseca, D., Poinso, F., O'Connor, K., and Deruelle, C. (2010). Do children perceive features of real and cartoon faces in the same way? evidence from typical development and autism. *Journal of clinical and experimental neuropsychology*, 32(2):212–218.

Shah, N. P. and Chichilnisky, E. J. (2020). Computational challenges and opportunities for a bi-directional artificial retina. *Journal of Neural Engineering*, 17(5):055002.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. (2019). End-to-End Deep Image Reconstruction From Human Brain Activity. *Frontiers in Computational Neuroscience*, 13:21.

Slllito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., and Davls, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378(6556):492–496.

Stanley, G. B., Li, F. F., and Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *Journal of Neuroscience*, 19(18):8036–8042.

Takacs, Z. K. and Bus, A. G. (2016). Benefits of motion in animated storybooks for children's visual attention and story comprehension. an eye-tracking study. *Frontiers in Psychology*, 7:1591.

Tang, S., Lee, T. S., Li, M., Zhang, Y., Xu, Y., Liu, F., Teo, B., and Jiang, H. (2018a). Complex pattern selectivity in macaque primary visual cortex revealed by large-scale two-photon imaging. *Current Biology*, 28(1):38–48.

Tang, S., Zhang, Y., Li, Z., Li, M., Liu, F., Jiang, H., and Lee, T. S. (2018b). Large-scale two-photon imaging revealed super-sparse population codes in the V1 superficial layer of awake monkeys. *eLife*, 7:e33370.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Victor, J. D., Mechler, F., Repucci, M. A., Purpura, K. P., and Sharpee, T. (2006). Responses of V1 neurons to two-dimensional hermite functions. *Journal of Neurophysiology*, 95(1):379–400.

Vinje, W. E. and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276.

Wang, J., Cao, R., Brandmeir, N. J., Li, X., and Wang, S. (2021). Face identity selectivity in the deep neural network and human brain. *bioRxiv*.

Wang, T., Li, Y., Yang, G., Dai, W., Yang, Y., Han, C., Wang, X., Zhang, Y., and Xing, D. (2020). Laminar subnetworks of response suppression in macaque primary visual cortex. *The Journal of Neuroscience*, 40(39):7436–7450.

Wang, Z., Bovik, A. C., and Lu, L. (2002). Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160.

Xu, Q., Shen, J., Ran, X., Tang, H., Pan, G., and Liu, J. K. (2021). Robust transcoding sensory information with neural spikes. *IEEE Transactions on Neural Networks and Learning Systems*.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yan, Q., Zheng, Y., Jia, S., Zhang, Y., Yu, Z., Chen, F., Tian, Y., Huang, T., and Liu, J. K. (2020). Revealing fine structures of the retinal receptive field by deep-learning networks. *IEEE Transactions on Cybernetics*, pages 1–12.

Yoshida, T. and Ohki, K. (2018). Robust representation of natural images by sparse and variable population of active neurons in visual cortex. *bioRxiv*, page 300863.

Yoshida, T. and Ohki, K. (2020). Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, 11(1):872.

Yu, Z., Liu, J. K., Jia, S., Zhang, Y., Zheng, Y., Tian, Y., and Huang, T. (2020). Toward the next generation of retinal neuroprosthesis: Visual computation with spikes. *Engineering*, 6(4):449–461.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386.

Zhang, Y., Jia, S., Zheng, Y., Yu, Z., Tian, Y., Ma, S., Huang, T., and Liu, J. K. (2020). Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, 125:19–30.

Zhang, Y., Lee, T. S., Li, M., Liu, F., and Tang, S. (2019). Convolutional neural network models of V1 responses to complex patterns. *Journal of Computational Neuroscience*, 46(1):33–54.

Zheng, Y., Jia, S., Yu, Z., Liu, J. K., and Huang, T. (2021). Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks. *Patterns*, 2(10):100350.

Zhou, Q., Du, C., Li, D., Wang, H., Liu, J. K., and He, H. (2020). Simultaneous neural spike encoding and decoding based on cross-modal dual deep generative model. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhou Wang, Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Supplemental Materials: Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex

Table S1: Reconstruction image index for artificial patterns. Raw values of image quality assessment metrics of all models measured over artificial patter images, corresponding to the radar plot in Fig 2C.

	MSE	PSNR	SSIM	GSM	FSIM	MAD
Ours	0.068	13.158	0.625	0.937	0.764	95.699
SID	0.059	13.167	0.519	0.931	0.718	107.972
DGMM	0.108	10.569	0.306	0.882	0.538	155.127
CGAN	0.147	8.936	0.213	0.855	0.527	161.857
BCCA	0.247	6.337	0.120	0.801	0.411	204.523

Table S2: Reconstruction image index for natural images. Raw values of image quality assessment metrics of all models measured over natural scene images, corresponding to the radar plot in Fig 2C.

	MSE	PSNR	SSIM	GSM	FSIM	MAD
Ours	0.063	10.487	0.137	0.916	0.592	221.628
SID	0.063	10.006	0.141	0.915	0.578	224.067
DGMM	0.058	9.812	0.128	0.914	0.569	229.328
CGAN	0.158	7.931	0.014	0.863	0.516	212.718
BCCA	0.181	7.552	0.072	0.876	0.514	227.103

Table S3: Reconstruction image index of models of three cases: our model, our model without skip connections, SID.

	MSE	PSNR	SSIM	GSM	FSIM	MAD
AP						
Our model	0.0663	12.9775	0.5832	0.9307	0.7364	113.8812
w/o skip connections	0.0612	13.1267	0.5675	0.9279	0.7162	113.7167
SID	0.0578	13.2294	0.5081	0.9272	0.6997	114.2638
NS						
Our model	0.0694	10.1314	0.1361	0.9134	0.5850	221.2789
w/o skip connections	0.0716	9.8401	0.1191	0.9154	0.5974	218.8611
SID	0.1353	9.6314	0.1450	0.9110	0.5567	226.2519

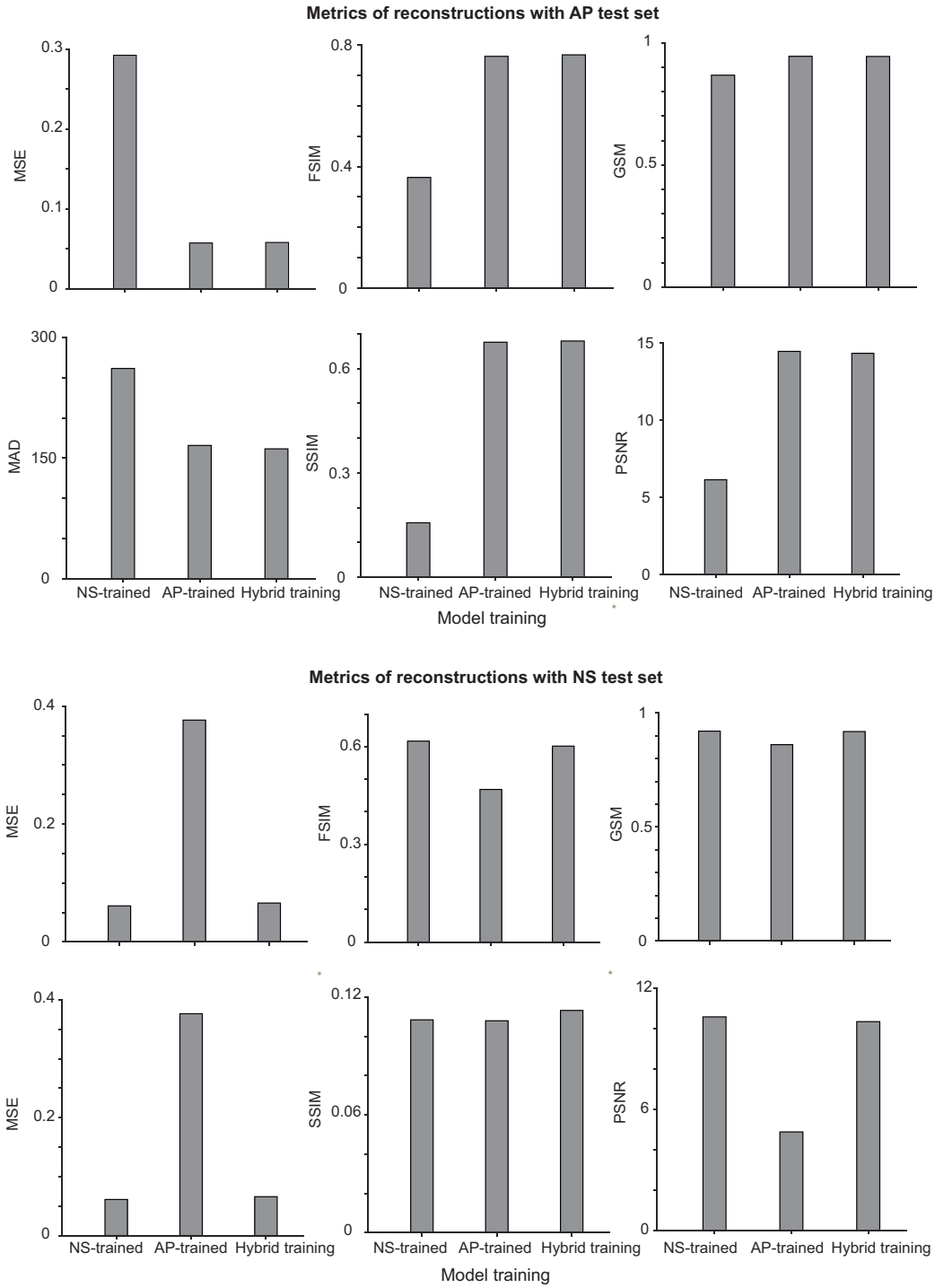

Figure S1: **Metrics of different models with across training.**

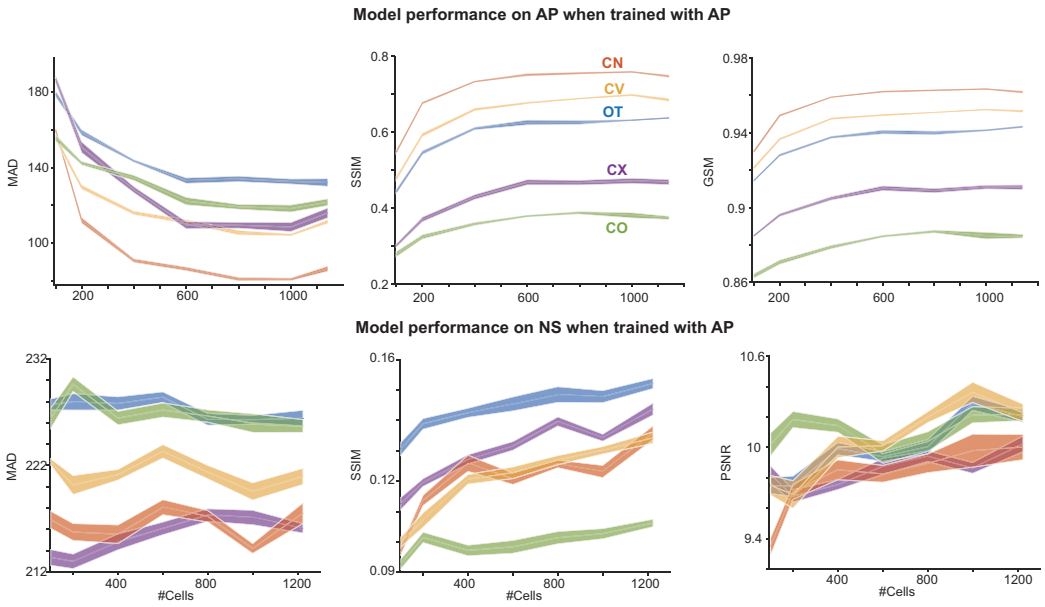

Figure S2: **Additional metrics of the analysis on models trained with APs.** The performance of three additional metrics related to Fig 7B and C.