# Variational Autoencoder: An Unsupervised Model for Modeling and Decoding fMRI Activity in Visual Cortex

**Kuan Han**[2,3], **Haiguang Wen**[2,3], **Junxing Shi**[2,3], **Kun-Han Lu**[2,3], **Yizhen Zhang**[2,3], **Zhongming Liu**[*1,2,3]

[1]Weldon School of Biomedical Engineering

[2]School of Electrical and Computer Engineering

[3]Purdue Institute for Integrative Neuroscience

Purdue University, West Lafayette, Indiana, 47906, USA

*Correspondence

Zhongming Liu, PhD

Assistant Professor of Biomedical Engineering

Assistant Professor of Electrical and Computer Engineering

College of Engineering, Purdue University

206 S. Martin Jischke Dr.

West Lafayette, IN 47907, USA

Phone: +1 765 496 1872

Fax: +1 765 496 1459

Email: zmliu@purdue.edu

# Abstract

Goal-driven convolutional neural networks (CNN) have been shown to be able to predict and decode cortical responses to natural images or videos. Here, we explored an alternative deep neural network, variational auto-encoder (VAE), as a computational model of the visual cortex. We trained a VAE with a five-layer encoder and a five-layer decoder to learn visual representations from a diverse set of unlabeled images. Inspired by the "free-energy principle" in neuroscience, we modeled the brain's bottom-up and top-down pathways using the VAE's encoder and decoder, respectively. Following such conceptual relationships, we found that the VAE was able to predict cortical activities observed with functional magnetic resonance imaging (fMRI) from three human subjects watching natural videos. Compared to CNN, VAE resulted in relatively lower prediction accuracies, especially for higher-order ventral visual areas. On the other hand, fMRI responses could be decoded to estimate the VAE's latent variables, which in turn could reconstruct the visual input through the VAE's decoder. This decoding strategy was more advantageous than alternative decoding methods based on partial least square regression. This study supports the notion that the brain, at least in part, bears a generative model of the visual world.

**Keywords**: neural encoding, variational autoencoder, generative model, visual reconstruction

# Introduction

Humans readily make sense of the visual world through complex neuronal circuits. Understanding the human visual system requires not only measurements of brain activity but also computational models with built-in hypotheses about neural computation and learning (Kietzmann, McClure and Kriegeskorte 2017). Models that truly reflect the brain's working in natural vision should be able to explain brain activity given any visual input (encoding), and decode brain activity to infer visual input (decoding) (Naselaris et al. 2011). Therefore, evaluating the model's encoding and decoding performance serves to test and compare hypotheses about how the brain learns and organizes visual representations (Wu, David and Gallant 2006).

In one class of hypotheses, the visual system consists of feature detectors that progressively extract and integrate features for visual recognition. For example, Gabor and wavelet filters allow extraction of such low-level features as edges and motion (Hubel and Wiesel 1962, van Hateren and van der Schaaf 1998), and explain brain responses in early visual areas (Kay et al. 2008, Nishimoto et al. 2011). More recently, convolutional neural networks (CNNs) encode multiple layers of features in a brain-inspired feedforward network (LeCun, Bengio and Hinton 2015), and support human-like performance in image recognition (Simonyan and Zisserman 2014, He et al. 2016, Krizhevsky, Sutskever and Hinton 2012). Such models bear hierarchically organized representations similar as in the brain itself (Khaligh-Razavi and Kriegeskorte 2014, Cichy et al. 2016), and shed light on neural encoding and decoding during natural vision (Yamins et al. 2014, Horikawa and Kamitani 2017, Eickenberg et al. 2017, Guclu and van Gerven 2015, Wen et al. 2017b). For these reasons, goal-driven CNNs are gaining attention as favorable models of the visual system (Kriegeskorte 2015, Yamins and DiCarlo 2016). Nevertheless, biological learning is not entirely goal-driven but often unsupervised (Barlow 1989), and the visual system has not only feedforward (bottom-up) but feedback (top-down) connections (Salin and Bullier 1995, Bastos et al. 2012).

In another class of hypotheses, the visual world is viewed as the outcome of a probabilistic and generative process (Fig. 1A): any visual input results from a generative model that samples the hidden "causes" of the input from their probability distributions (Friston 2010). In light of this view, the brain behaves as an inference machine: recognizing and predicting visual input through "analysis by synthesis" (Yuille and Kersten 2006). The brain's bottom-up process analyzes visual input to infer the "cause" of the input, and its top-down process predicts the input from the brain's internal representations. Both processes are optimized by learning from visual experience in order to avoid the "surprise" or error of prediction (Rao and Ballard 1999, Friston and Kiebel 2009). This hypothesis attempts to account for both feedforward and feedback connections in the brain, align with the humans' ability to construct mental images, and offer a basis for unsupervised learning. Thus, it is compelling for both computational neuroscience (Friston 2010, Rao and Ballard 1999, Bastos et al. 2012, Yuille and Kersten 2006) and artificial intelligence (Hinton et al. 1995, Lotter, Kreiman and Cox 2016, Mirza, Courville and Bengio 2016).

In line with this notion is the so-called variational autoencoder (VAE) (Kingma and Welling 2013). VAE uses independently distributed "latent" random variables to code the causes of the visual world. VAE learns the latent variables from images via an encoder, and samples the latent variables to generate new images via a decoder, where the encoder and decoder are both neural networks that can be trained from a large dataset without supervision (Doersch 2016). Hypothetically, VAE offers a potential model of the brain's visual system, if the brain also captures the causal structure of the visual world. As such, the latent variables in VAE should match (up to linear projection) neural representations given naturalistic visual input; the generative component in VAE should enable an effective and generalizable way to decode brain activity during either visual perception or imagery (Du, Du and He 2017, Güçlütürk et al. 2017, van Gerven, de Lange and Heskes 2010a).

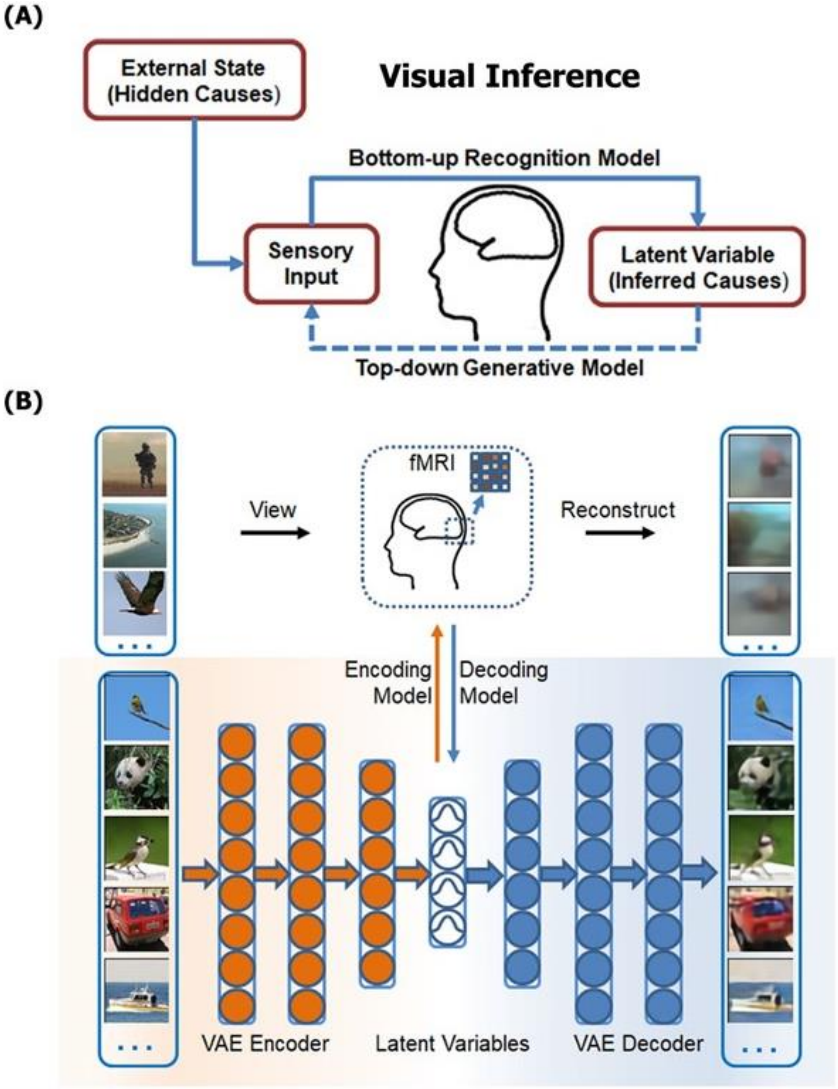

**Figure 1 | The unsupervised inference model of vision. (A) Regarding the brain as an inference machine.** The brain analyzes sensory data by approximating the causes that generate the sensations. Its bottom-up process maps from input to sensory causes and top-down process predicts the input based on inferred causes. **(B) Encoding and decoding cortical activities with variational autoencoder.** For encoding, cortical activities are predicted by using VAE responses given visual stimuli; For decoding, latent variables are predicted from fMRI measurements and mapped to visual reconstruction through VAE decoder.

This led us to investigate VAE as a candidate model of the visual cortex for unsupervised learning of visual representations (Fig. 1B). In this study, we addressed the relationship between

VAE and the brain's "free-energy" principle (Friston 2010), and then tested the degree to which VAE could be used to predict and decode cortical responses observed with functional magnetic resonance imaging (fMRI) during naturalistic movie stimuli.

## Methods and Materials

### *Theory: variational autoencoder*

In general, variational autoencoder (VAE) is a type of deep neural networks that learns representations from complex data without supervision (Kingma and Welling 2013). A VAE includes an encoder and a decoder, both of which are implemented as neural nets. The encoder learns latent variables from the input data, and the decoder learns to generate similar input data from samples of the latent variables. Given large input datasets, the encoder and the decoder are trained together by minimizing the reconstruction loss and the Kullback-Leibler (KL) divergence between the distributions of the latent variables and independent standard normal distributions (Doersch 2016). When the input data are natural images, the latent variables represent the causes of the images. The encoder serves as an inference model that attempts to infer the latent causes of any input image; the decoder serves as a generative model that generates new images by sampling latent variables.

Mathematically, let $\boldsymbol{z}$ be the latent variables and $\boldsymbol{x}$ be images. The encoder parameterized with $\boldsymbol{\varphi}$ infers $\boldsymbol{z}$ from $\boldsymbol{x}$, and the decoder parameterized with $\boldsymbol{\theta}$ generates $\boldsymbol{x}$ from $\boldsymbol{z}$. In VAE, both $\boldsymbol{z}$ and $\boldsymbol{x}$ are random variables. The likelihood of $\boldsymbol{x}$ given $\boldsymbol{z}$ under the generative model $\boldsymbol{\theta}$ is denoted as $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. The probability of $\boldsymbol{z}$ given $\boldsymbol{x}$ under the inference model $\boldsymbol{\varphi}$ is denoted as $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$. The marginal likelihood of data can be written as the following form.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = D_{KL}\left[q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] + L(\boldsymbol{\theta},\boldsymbol{\varphi};\boldsymbol{x}) \tag{1}$$

Since the Kullback-Leibler divergence in Equation (1) is non-negative, $L(\boldsymbol{\theta},\boldsymbol{\varphi};\boldsymbol{x})$ can be regarded as the lower-bound of data likelihood and also be rewritten as Eq. (2). For VAE, the learning rule is to optimize $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ by maximizing the following function given the training samples of $\boldsymbol{x}$.

$$L(\boldsymbol{\theta},\boldsymbol{\varphi};\boldsymbol{x}) = -D_{KL}\left[q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z})\right] + E_{\boldsymbol{z}\sim q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] \tag{2}$$

In this objective function, the first term is the KL divergence between the distribution of $\boldsymbol{z}$ inferred from $\boldsymbol{x}$ and the prior distribution of $\boldsymbol{z}$, both of which are assumed to follow a multivariate normal distribution. The second term is the expectation of the log-likelihood that the input image can be generated by the sampled values of $\boldsymbol{z}$ from the inferred distribution $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$. When $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$ is a multivariate normal distribution with unknown expectations $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$, the objective function is differentiable with respect to ($\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$), which can therefore be optimized iteratively through gradient-descent algorithms (Kingma and Welling 2013).

Similar concepts have been put forth in computational neuroscience theories, for example the free-energy principle (Friston 2010). In free-energy principle, the brain's perceptual system includes bottom-up and top-down pathways. Like the encoder in VAE, the bottom-up pathway maps from sensory data to their inferred causes as probabilistic representations. Like the decoder in VAE, the top-down pathway predicts the sensation from what the brain infers as the causes of sensations. Both the bottom-up and top-down pathways are optimized together, such that the brain infers the causes of the sensory input, and generates the sensory prediction that matches the input with the minimal error or surprise. Mathematically, the learning objectives in both VAE and free energy principle are similar, both aiming to minimize the difference between the inferred and hidden causes of sensory data (measured by Kullback–Leibler divergence) while maximizing the likelihood of the sensory data given the inferred causes.

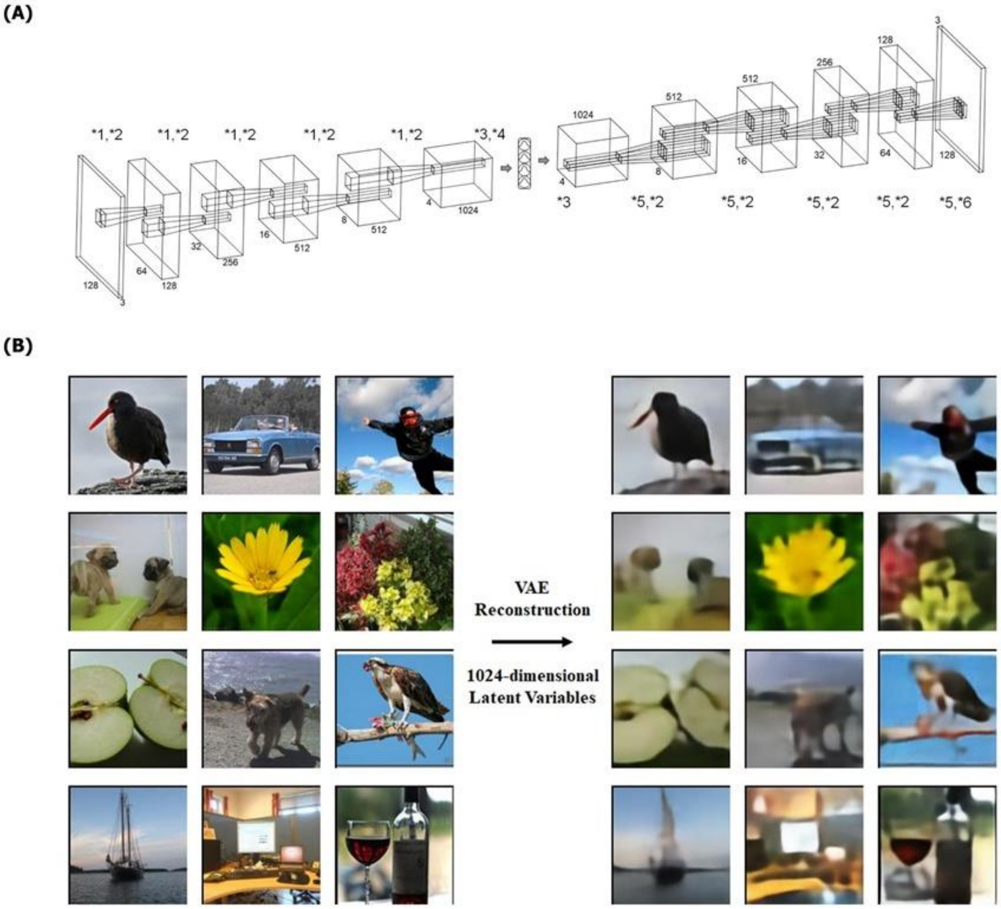

**Figure 2 | VAE architecture and reconstruction examples. (A) VAE architecture.** The encoder and the decoder both contained 5 layers. The dimension of latent variables was 1024. Operations were defined as: *1 convolution (kernel size=4, stride=2, padding=1), *2 rectified nonlinearity, *3 fully connected layer, *4 re-parameterization trick, *5 transposed convolution (kernel size = 4, stride = 2, padding = 1), *6 sigmoid nonlinearity. **(B) Examples of VAE reconstruction.** Each testing image was coded 1024-dimensional latent variables with VAE encoder. The reconstruction from VAE decoder (blurred images on right side) retained basic and condensed information similar to the original input (clear images on left side).

### *Training VAE with diverse natural images*

We designed a VAE with 1,024 latent variables while the encoder and the decoder were both convolutional neural networks with five hidden layers (Fig. 2A). Each convolutional layer included nonlinear units with a Rectified Linear Unit (ReLU) function (Nair and Hinton 2010), except the last layer in the decoder where a sigmoid function was used to generate normalized pixel values between 0 and 1. The model was trained on the ImageNet ILSVRC2012 dataset

(Russakovsky et al. 2015). Training images were resized into 128×128×3. To enlarge the amount of training data, the original training images were randomly flipped in the horizontal direction to yield additional training images, resulting in >2 million training samples in total. The training data were divided into mini-batches with a batch size of 200. For each training example, the pixel intensities were normalized to a range from 0 to 1, such that the normalized intensity could be interpreted as the probability of color emission (Gregor et al. 2015). To train the VAE, the Adam optimizer (Kingma and Ba 2014) was used with the learning rate of 1e-4. This VAE was implemented in *PyTorch* (http://pytorch.org/).

### *Experimental data*

Three healthy volunteers (all female, age: 23-26) participated in this study with informed written consent according to a research protocol approved by the Institutional Review Board at Purdue University. All experiments were performed in accordance with relevant guidelines and regulations as described in this approved protocol. As described in detail elsewhere (Wen et al. 2017b), the experimental design and data were briefly summarized as below. Each subject watched a diverse set of natural videos for a total length up to 13.3 hours. The videos were selected and downloaded from Videoblocks and YouTube, and then were separated into two independent sets. One data set was for training the models to predict the fMRI responses based on the input video (i.e. the encoding models) or the models to reconstruct the input video based on the measured fMRI responses (i.e. the decoding models). The other data set was for testing the trained the encoding or decoding models. The videos in the training and testing datasets were entirely distinctive to ensure unbiased model evaluation. Both the training and testing movies were split into different 8-min segments. Each segment was used as visual stimulation ($20.3^{o}\times20.3^{o}$) along with a central fixation cross ($0.8^{o}\times0.8^{o}$) presented via an MRI-compatible binocular goggle during a single fMRI session. The training movie included 98 segments (13.1 hours) for Subject 1, and 18 segments (1.6 hours) for Subject 2 & 3. The testing movie included 5 segments (40 mins in total). Each subject watched the testing movie 10 times. All five testing movies were used to test the encoding model. One of the testing movies contained video clips that were continuous over long periods (mean ± std: 13.3 ± 4.8 s) was used to test the decoding model for visual reconstruction.

MRI/fMRI data were collected from a 3-T MRI system, including anatomical MRI ($T_1$ and $T_2$ weighted) of 1mm isotropic spatial resolution, and BOLD functional MRI with 2-s temporal resolution and 3.5mm isotropic spatial resolution. The fMRI data were registered onto anatomical MRI data, and were further co-registered on a cortical surface template (Glasser et al. 2013). The fMRI data were preprocessed with the minimal preprocessing pipeline released for the human connectome project (Glasser et al. 2013).

### *VAE-based encoding models*

The trained VAE extracted the latent representations of any video by a feed-forward pass of every video frame into the encoder, and generated the reconstruction by a feed-back pass with the decoder. To predict cortical fMRI responses to the video stimuli, an encoding model was defined separately for each voxel as a linear regression model, through which the voxel-wise fMRI signal was estimated as a function of VAE model activity (including both the encoder and the decoder) given the input video. Every unit activity in VAE was convolved with a canonical hemodynamic response function (HRF). For dimension reduction, PCA was applied to the HRF-convolved unit activity for each layer, keeping 99% of the variance of the layer-specific activity given the training movies. Then, the layer-wise activity was concatenated across layers; PCA was applied to the concatenated activity to keep the 99% of the variance of the activity from all layers given the training movies. See more details in our earlier paper (Wen et al. 2017a). Following the dimension reduction, the principal components of unit-activity were down-sampled to match the frequency of fMRI and used as the regressors to predict the fMRI signal at each voxel through the linear regression model specific to the voxel.

The voxel-wise regression model was trained based on the data during the training movie. Mathematically, for any training sample, let $\boldsymbol{x}^{(j)}$ be the visual stimuli at the $j$-th time point, $\boldsymbol{y}_{\boldsymbol{i}}^{(j)}$ be the fMRI response at the $i$-th voxel, $\boldsymbol{z}^{(j)}$ be a vector representing the predictors for the fMRI signal derived from the $\boldsymbol{x}$ through VAE, as described above. The voxel-wise regression model is expressed as Eq. (4).

$$\boldsymbol{y}_{\boldsymbol{i}}^{(j)} = \boldsymbol{w}_i^T \mathbf{z}^{(j)} + \boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{4}$$

where $\boldsymbol{w}_i$ is a column vector representing the regression coefficients, $\boldsymbol{b}_i$ is the bias term, and $\boldsymbol{\epsilon}_i$ is the error unexplained by the model. The linear regression coefficients were estimated using least-squares estimation with $L_2$-norm regularization, or minimizing the loss function as Eq. (5).

$$\langle \widehat{w}_i, \widehat{b}_i \rangle = \underset{w_i, b_i}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \left( \boldsymbol{y}_{\boldsymbol{i}}^{(j)} - \boldsymbol{w}_i^T \boldsymbol{z}^{(j)} - \boldsymbol{b}_i \right)^2 + \lambda_i \|\boldsymbol{w}_i\|_2^2 \tag{5}$$

where $\boldsymbol{N}$ is the number of training samples. The regularization parameter $\lambda_i$ was optimized for each voxel to minimize the loss in three-fold cross-validation. Once the optimal parameter $\lambda_i$ was determined, the model was refitted with the entire training set and the optimal regularization parameter to finalize the model.

After the above model training, we tested the voxel-wise encoding models with the testing movies. The testing movies were independent of the training movies to ensure unbiased model evaluation. The model performance was evaluated separately for each voxel as the correlation between the measured and predicted fMRI responses to the testing movie. The significance of the correlation was assessed by using a block permutation test (Adolf et al. 2014) with a block size of 24-sec and 30,000 permutations and corrected at false discovery rate (FDR) $q < 0.05$.

### *Encoding model comparison with supervised CNN*

We also built up a control model to predict cortical responses with supervised information. A 18-layer pretrained residual network (ResNet-18) was used as a benchmark to compare the VAE encoding performance with the performance of a supervised CNN. The model architecture was elaborated in the original ResNet paper (He et al. 2016). Briefly, ResNet-18 was consisting of 6 layers: the first layer was a convolutional layer followed by max-pooling, yielding location and orientation selective features; The last layer was a logistic regression layer, yielding the classification output; The 4 intermediate layers were consisting of stacked residual blocks, yielding the progressive abstraction from low level features to semantic features. The output units of first five layers in ResNet-18 were extracted as model activities given input video. Then the unit activities were processed in the same way as the preprocessing steps of VAE.

The encoding accuracy of VAE was compared with supervised CNN in two ways. First, regions of interest (ROI) were selected from various levels of visual hierarchy: V1, V2, V3, V4, lateral occipital (LO), middle temporal (MT), fusiform face area (FFA), para-hippocampal place area (PPA) and temporo-parietal junction (TPJ). In each ROI, the correlation coefficient of each voxel was corrected by noise-ceiling, and then averaged over all voxels and all subjects. The averaged prediction accuracies of each ROI were compared between VAE and ResNet-18. Second, for each voxel, the voxel-wise correlation coefficient was transformed to z score through Fisher transformation for both VAE and ResNet-18. Then the difference between two models were calculated as subtracting the z score of VAE from z score of ResNet-18, yielding the voxel-wise difference in prediction accuracy.

The process of calculating noise-ceiling was elaborated in a previous study (Kay et al. 2013). Briefly, for each subject, the noise was assumed to be zero-mean, and the variance of the noise was estimated as the mean square of standard errors in the testing data across 10 repetitions. The variance of the response was taken as the difference between the variance of the data and the variance of the noise. From the estimated signal and noise distributions, the sample of response and the sample of noise was drawn by Monte Carlo simulation. The simulated $R^2$ value between simulated response and noisy data was calculated over 1000 repetitions, yielding the distribution of noise-ceiling of each voxel.

### *VAE-based decoding of fMRI for visual reconstruction*

We trained and tested the decoding model for reconstructing visual input from distributed fMRI responses. The model contained two steps: 1) transforming the spatial pattern of fMRI response to latent variables through a linear regression model, 2) transforming latent variables to pixel patterns through the VAE's decoder.

Mathematically, let $\boldsymbol{Y}^{(j)}$ be a column vector representing the measured fMRI map at the $j$-th time point, and $\boldsymbol{z}^{(j)}$ be a column vector representing the latent variables given the unknown visual input $\boldsymbol{x}^{(j)}$. As Eq. (6), a multivariate linear regression model was defined to predict $\boldsymbol{z}^{(j)}$ given $\boldsymbol{Y}^{(j)}$.

$$\boldsymbol{z}^{(j)} = \mathbf{U}\boldsymbol{Y}^{(j)} + \boldsymbol{c} + \boldsymbol{\varepsilon} \qquad (6)$$

where $\mathbf{U}$ is a weighting matrix representing the regression coefficients to transform the fMRI map to the latent variables, $\boldsymbol{c}$ is the bias term, and $\boldsymbol{\varepsilon}$ is the error term unexplained by the model. This model was estimated based on the data during the training movie, by using $L_1$-norm regularization least-squares estimation, or minimizing the loss function defined as below.

To estimate parameters of the decoding model, we minimized the objective function as Eq. (7) with $L_1$-regularized least-squares estimation to prevent over-fitting.

$$\langle\, \widehat{\mathbf{U}}, \hat{\boldsymbol{c}} \rangle = \arg\min_{\mathbf{U},\boldsymbol{c}} \frac{1}{N} \sum_{j=1}^{N} \left(\boldsymbol{z}^{(j)} - \mathbf{U}\boldsymbol{Y}^{(j)} - \boldsymbol{c}\right)^2 + \lambda \|\mathbf{U}\|_1^1 \qquad (7)$$

where N is the number of data samples used for training the model. The regularization parameter, $\lambda$, was optimized to minimize the loss in three-fold cross-validation. To solve Eq. (7), we used stochastic gradient-descent algorithm with a batch size of 100 and a learning rate of 1e-7.

After the estimation of latent variables, the visual input in the testing movie was reconstructed by passing the estimated latent variables through the decoder in VAE, as expressed by Eq. (8)

$$\hat{\boldsymbol{x}}^{(j)} = \Theta\left(\hat{\boldsymbol{z}}^{(j)}\right) = \Theta\left(\widehat{\mathbf{U}}\boldsymbol{Y}^{(j)} + \hat{\boldsymbol{c}}\right) \tag{8}$$

To evaluate the decoding performance in visual reconstruction, we calculated the Structural Similarity index (SSIM) (Wang et al. 2004) between every video frame of the testing movie and the corresponding frame reconstructed on the basis of the decoded fMRI signals. We averaged the SSIM over all frames in the testing movie (resampled by TR), as a measure of the spatial similarity between the reconstructed and actual movie stimuli. In addition, we further evaluated the degree to which color information was preserved in the movie reconstructed from fMRI data. For this purpose, the color information at each pixel was converted from the RGB values to a single hue value. The hue maps of the reconstructed movie frames were compared with those of the original movie frames. Their similarity was evaluated in terms of the circular correlation (Berens 2009, Jammalamadaka and Sengupta 2001). Hue values in the same frame were flattened into a vector and the hue vectors of different frames were concatenated sequentially as a monolithic vector including all testing frames. Then the circular correlation of monolithic hue vectors between original and reconstructed frames for each subject was calculated. The statistical significance of the circular correlation between the reconstructed and original color was tested by using the block-permutation test with 24-sec block size and 30,000 times permutation (Adolf et al. 2014).

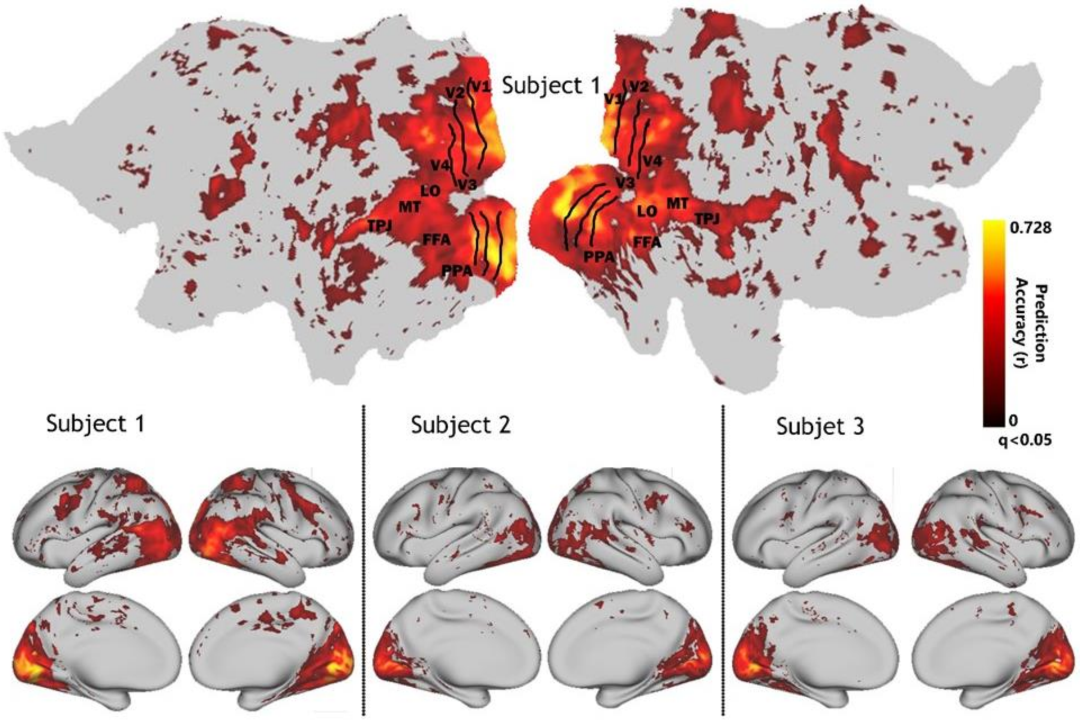

**Figure 3 | Prediction accuracy with VAE-based encoding model.** The accuracy was measured by the average Pearson's correlation coefficient (r) between the predicted and the observed fMRI responses over five testing movies (q<0.05, FDR correction). The result of subject 1 was shown with flat and stereoscopic views (top and bottom-left). Results of other subjects were shown with stereoscopic views.

We also compared the performance of the VAE-based decoding method with a previously published decoding method (Cowen, Chun and Kuhl 2014). For this alternative method (Cowen et al. 2014), we applied PCA to the training movie and obtained its principal components (or eigen-images), which explained 99% of its variance. The partial least square regression (PLSR) (Tenenhaus et al. 2005) was used to estimate the linear transformation from fMRI maps to eigen-images given the training movie. Using the estimated PLSR model, the fMRI data during the testing movie was first converted to representations in eigen-images, which in turn were recombined to reconstruct the visual stimuli (Cowen et al. 2014). As a variation of this PLSR-based model, we also explored the use of $L_1$-norm regularized optimization to estimate the linear transform from fMRI maps to eigen-images, in a similar way as used for the training of our decoding model (see Eq. (7)).

We also explored whether the VAE-based decoding models could be generalized across subjects. For this purpose, we used the decoding model trained from one subject to decode the fMRI data observed from the other subjects while watching the testing movie.

# Results

## *VAE provided vector representations of natural images*

By design, the VAE aimed to form a compressed and generalized vector representation of any natural image. In VAE, the encoder converted any natural image into 1,024 independent latent variables; the decoder reconstructed the image from the latent variables (Fig. 2.A). After training it with >2 million natural images in a wide range of categories, the VAE could regenerate natural images without significant loss in image content, structure and color, albeit blurred details (Fig. 2B). Note that the VAE-generated images showed comparable quality for different types of input images (Fig. 2.B). In this sense, the latent representations in the VAE were generalizable across various types of visual objects, or their combinations.

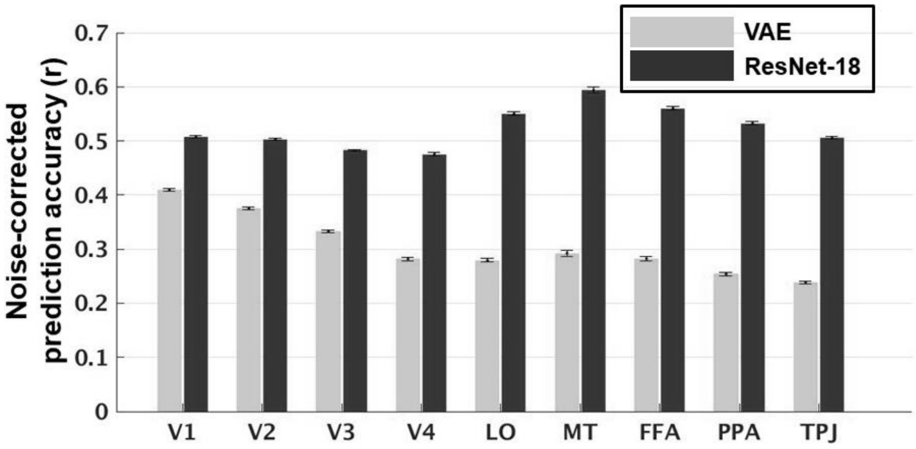

**Figure 4 | Model comparison with group prediction accuracies averaged within different ROIs.** Each bar represented the mean ± SE of the prediction accuracy (noise-corrected Pearson' s correlation coefficient averaged over 5 testing movies) across voxels within a ROI and across subjects. Results from VAE were shown with light color, with gradually decreasing accuracies from low level to high level. Results from ResNet-18 were shown with dark color, with accuracies in high level ROIs generally higher than those of low level ROIs.

### *VAE predicted movie-induced cortical responses of visual cortex*

Given natural movies as visual input, we further asked whether and to what extent the model dynamics in VAE could be used to model and predict the movie-induced cortical responses. Specifically, a linear regression model was trained separately for each voxel by optimally fitting the voxel response time series to a training movie as a linear combination of VAE's unit responses to the movie. Then, the trained voxel-wise encoding model was tested with a new testing movie (not used for training) to evaluate its prediction accuracy (i.e. the correlation between the predicted and measured fMRI responses). We found significantly reliable predictions (q<0.05, FDR correction) based on VAE-encoding model in a reasonable fraction of cortical areas (Fig. 3). The primary visual area (V1) showed highest prediction accuracy and intermediate visual areas (V2/V3/V4) were also predictable but the prediction accuracy was decreasing gradually, either along ventral or dorsal stream (Fig. 3). The VAE-predictable areas extended to a relatively larger scope when more data (~12-hour movie) were used for training the encoding models in Subject 1 than Subject 2 & 3 for whom less training data (2.5-hour movie) were available.

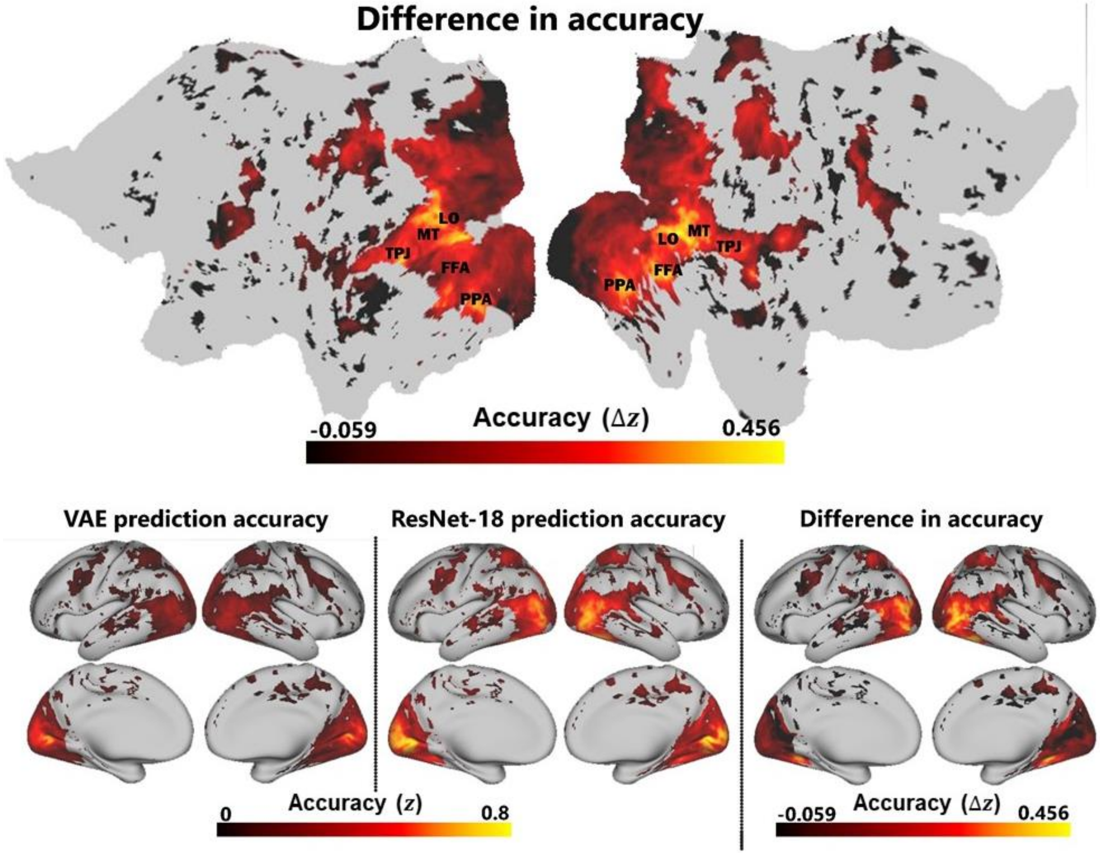

**Figure 5 | Comparing encoding prediction accuracy of VAE and ResNet-18.** The voxel-wise correlation coefficient was transformed to z-score through Fisher transformation. The z-scores of VAE based encoding model ($z_{VAE}$), ResNet-18 based encoding model ($z_{Res}$) and their difference ($\Delta z = z_{Res} - z_{VAE}$) were shown.

### *Comparing prediction accuracies between VAE and ResNet-18 based encoding models*

VAE based encoding model explained visual area activities with significant accuracies. Furthermore, an investigation was given to compare VAE and ResNet-18, addressing their relative prediction performance. Fig. 4 showed prediction accuracies of different ROIs based on two models. VAE showed the highest prediction accuracy for V1, and other accuracies gradually decreased from intermediate visual levels (V2/V3) to higher levels. Differently, ResNet-18 showed higher accuracies for high level ROIs (LO/MT/FFA/PPA/TPJ) than early visual areas. As for the accuracy difference, VAE provided similar performance to ResNet-18 in V1, but was gradually outperformed by ResNet-18 to a significant extent in ROIs of high levels.

We also tested the voxel-wise accuracy difference between two models. In this sense, the correlation coefficient of each voxel was transformed to z score. Then the differences between VAE based z scores and ResNet-18 based z scores were calculated. Generally, ResNet-18 gave better predictions than VAE for most of the voxels. Fig. 5 indicated a gradient of the z score subtraction: for early visual areas, prediction performance of VAE was similar to or slightly lower than ResNet-18; for high-level visual areas, stronger performance gap was found, suggesting that features from an unsupervised model might not be enough to explain voxel dynamics especially as the cortical level goes higher.

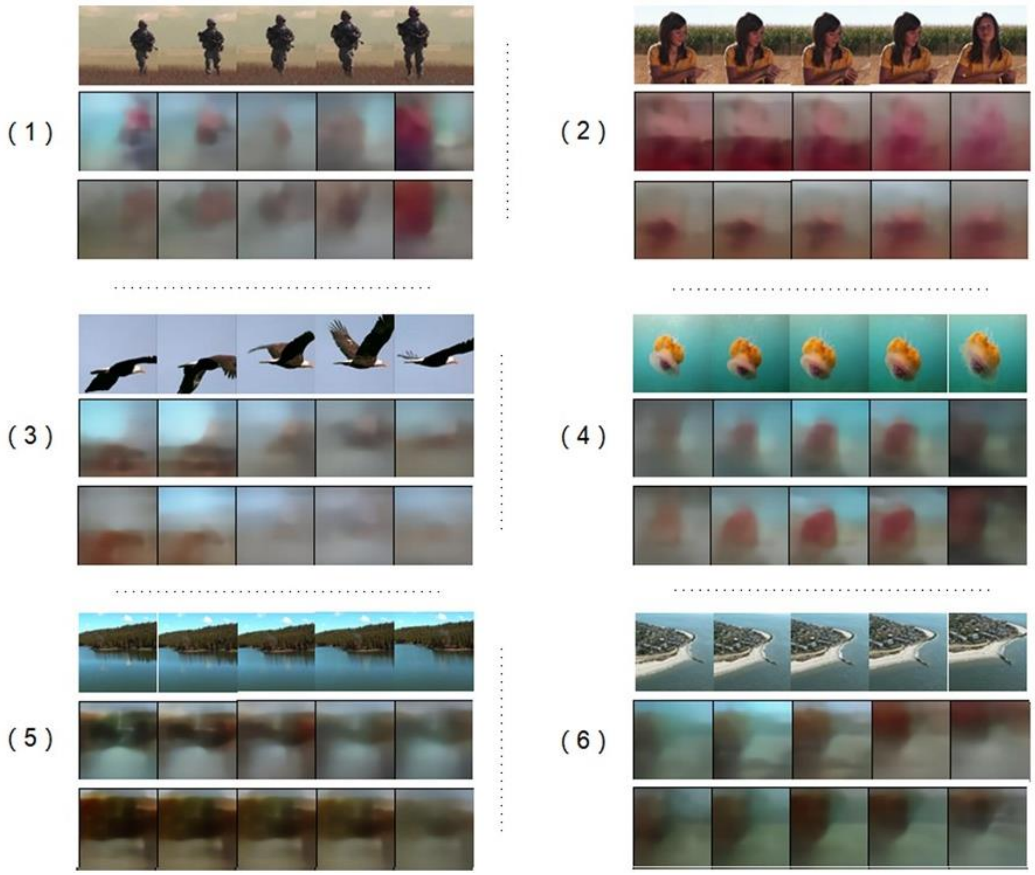

**Figure 6 | Reconstruction result and evaluation.** Reconstruction results of 6 different movie clips were shown. In each group, the first line showed original movie frames and the second line showed within-subject reconstruction. For cross-subject reconstruction, the decoding model was trained from subject 1 and results of subject 2 (clips 1, 2 & 3) and 3(clips 4, 5 & 6) were shown on the third line.

### *Direct visual reconstruction through decoding latent representation from the brain*

To see how well could the VAE model match to the brain, another way was to evaluate how well could the latent variables predicted from fMRI signals fit the generative model of VAE and reconstruct the visual input. We estimated the latent variables on visual stimuli from fMRI signals and fed the latent representation to VAE decoder. The reconstructed results generally reflected structure information of original input, including relative position and shape information of objects in the scene (Fig. 6). During the transition of consecutive movie clips, the reconstructed frames seemed to be unclear and affected by the content of both nearby clips. Inside the same continuous clip the result looks more stable and better. For the entire reconstructed movie frames of all 3 subjects, please see the enclosed video.

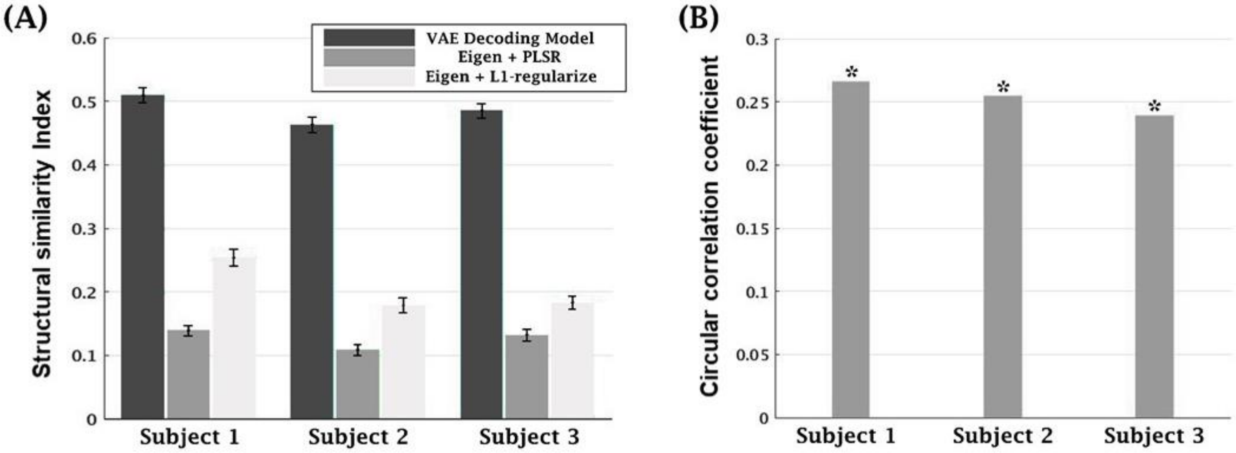

**Figure 7 | Quantitative evaluation of reconstruction. (A) Comparison of structural similarity index (SSIM).** SSIM scores of VAE-based decoding model and control models were compared for all 3 subjects. Each bar represented mean ± SE SSIM score over all frames in the testing movie. **(B) Color component correlation**. The circular correlation coefficient between original and reconstructed hue components was calculated and the p-value was given through permutation test (*, p<0.001).

We also gave a quantitative measurement and comparison of the model's performance on structural similarity. Given the decoding model and two control models, the structural similarity between reconstructed frames and visual inputs was evaluated using Structure Similarity Index (SSIM) (Wang et al. 2004). Fig. 7A indicated that the mean SSIM score of the decoding model was higher than the control models (pair sample t-test, $p < 0.001$) for each subject, suggesting the decoding model could get higher group-averaged SSIM score and reconstruct with better structural similarity. The visualization of reconstruction result intuitively showed the retained color information of original frames. For example, the reconstructed frame of the jellyfish (clips indexed 4 in Fig. 6) showed the color of orange, blue and cyan corresponding to true frame. To give a quantitative measurement, we evaluated the circular correlation of hue components with permutation test (Fig. 7B). For Subject 1, 2 & 3, the correlation coefficient values for all frames were 0.2657, 0.2544 & 0.2392 respectively (permutation test, $p < 0.001$ for all three subjects, no correction), suggesting that color variations within or across reconstructed frames reflected the color fluctuations shown by the visual stimuli.

The decoding model was transferable among subjects. We decoded fMRI signals of Subject 2&3 based on the decoding model trained from Subject 1 (Fig 6). The reconstruction result still retained information about structural, shape & color variation of objects and scenes in original frames.

## Discussion

### *Extending learning rules for modeling visual cortex*

Recently a growing body of literatures have used supervised deep-learning approaches to model and understand cortical representations during natural vision, yielding state-of-art performance for neural encoding and representational modeling (Yamins and DiCarlo 2016, Guclu and van Gerven 2015, Eickenberg et al. 2017, Cichy et al. 2016, Wen et al. 2017b, Shi et al. 2017), experimental paradigm simulation (Wen et al. 2017a, Eickenberg et al. 2017), visualizing single-voxel representation (Wen et al. 2017b) and neural decoding (Horikawa and Kamitani 2017, Wen et al. 2017b). This study extends previous findings by using an unsupervised inference model to explain cortical activities, focusing on modeling the unsupervised inference attribute of human brain.

In this study, the VAE-predictable voxels cover a large portion of the visual cortex (Fig. 3), suggesting that unsupervised inference is a compelling learning principle to drive brain-explanatory computational models. Beyond that, one interesting finding of this study is the encoding accuracy contrast between VAE and supervised CNN. While supervised CNN explains sufficient variance over nearly the entire visual cortex, VAE shows comparable encoding performance only in primary visual area. This is consistent with a previous finding which indicates that explaining inferior temporal (IT) cortex requires computational features trained through supervised learning (Khaligh-Razavi and Kriegeskorte 2014). Our result further suggests that the unsupervised learning better matches low level visual areas rather than high level visual areas, while CNN with supervised training matches both well and is more exclusively important for explaining high level visual areas.

### *Accumulating evidence for analysis-by-synthesis hypothesis with natural stimuli*

In the hypothesized “analysis by synthesis” theory of visual processing, the top-down generative component supports the bottom-up processing for inferring the causes of visual input. (Yuille and Kersten 2006, Friston 2010). Previous experimental studies have been accumulating evidences for “analysis by synthesis” hypothesis, including single-cell-like receptive fields (Rao and Ballard 1999), neural suppression due to stimulus predictability (Summerfield et al. 2008, Alink et al. 2010) and visual mismatch negativity (Wacongne, Changeux and Dehaene 2012, Garrido et al. 2009). These are all based on artificial visual stimuli instead of natural visual stimuli. However, it has been suggested that one important aspect of real inference systems is to deal with the complexities and ambiguities of visual information in natural world (Yuille and Kersten 2006). Therefore, this study attempts to evaluate the VAE based encoding and decoding models especially with natural visual stimuli.

Based on VAE computational model, we evaluate the inference property of natural vision through two phases: 1) While the model is driven by unsupervised inference, the unit activities of the whole model can jointly model and predict cortical dynamics, yielding similar or related learning rules in the significantly predicted areas; 2) The inference result of the computational model,  latent variables which represent inferred causes, can be directly extracted from cortical activities and tested for visual reconstruction. The regression techniques for both 1) and 2) are chosen to be linear because we want to guarantee that the encoding and decoding performances are contributed by the similar representations of the model to the brain, instead of the added regression. The encoding results match well to early visual areas (Fig. 5), suggesting the similar visual representations shared by the model dynamics and cortical responses. The decoding results show that the extracted latent representations from brain can support visual reconstruction (Fig. 6), with the performance better than linear reconstruction methods (Fig. 7A), and can likely retain color information of the original visual input (Fig. 7B). One surprising finding is that the decoding model keeps comparable performance after transferred across subjects, suggesting that different individuals organize cortical representations of visual cause information in a similar way. This similarity is likely to result from unsupervised inference representations because different individuals have different supervised learning experience and supervision knowledge, but the factors generating the visual world might be similar over diverse circumstances.

### *Towards generalizable visual stimuli reconstruction*

For decoding cortical activities, the built-in decoder of VAE enables a clear, independent and generalizable way to reconstruct visual input. Previous brain decoding studies mostly generate reconstruction in certain domain rather than diverse natural input, including edge orientation (Kamitani and Tong 2005), faces (Cowen et al. 2014, Nestor, Plaut and Behrmann 2016, Güçlütürk et al. 2017), contrast patterns (Miyawaki et al. 2008), digits (van Gerven, de Lange and Heskes 2010b, Du et al. 2017) and words (Schoenmakers et al. 2013). As for natural visual reconstruction, Bayesian method has been used to combine the structure encoding model, the semantic encoding model and image prior information together to have accurate reconstruction result on natural images (Naselaris et al. 2009). The Bayesian model is further extended with motion-energy filter to reconstruct natural movies (Nishimoto et al. 2011). However, using image prior information requires sampling from a large dataset of natural images and the reconstruction is likely to correspond to an image that is already in the dataset. Therefore, the method potential is limited if the purview of the content of the reconstruction bank is unknown. VAE-based decoding method do not rely on prior samples because the nonlinear mapping from latent variables to the input is provided by the VAE decoder generalized from millions of training samples, without requiring best-matching image or other kinds of prior information.

Some recent studies have proposed generative methods for visual reconstruction. Deep belief network has been used to model hierarchical binary latent causes (van Gerven et al. 2010a). Then the decoding process is consisting of a conditional sampling step to estimate the top-level associative states from cortical activities, and an unconditional sampling step to propagate expectations back to input layer. One study uses deep generative multi-view model to analyze brain response and visual stimulus together. In this regard, visual reconstruction becomes Bayesian inference of missing view in a multi-view latent variable model (Du et al. 2017). Besides, another study introduces perceptual similarity metrics to face reconstruction (Güçlütürk et al. 2017), proposing adversarial training for neural decoding. Our study is complementary to previous generative reconstruction methods for both the application scenario and the model principle. From the application phase, we showed the feasibility of using generative model to reconstruct dynamic natural movies, which is also promising for natural image reconstruction; From the theoretical modeling phase, VAE framework can extend to a wide range of model architectures and is easier to maintain tractability, compared with deep belief networks (Goodfellow, Bengio and Courville 2016). Generally, VAE shows a light and generalizable framework for reconstructing diverse natural movies, and could potentially be extended with other deep learning frameworks. We hypothesize that the integration of multiple generative techniques (especially the techniques mentioned above including adversarial training and extended Bayesian inference methods) would integrate the merits of different models and open a new avenue towards generalizable reconstruction on diverse natural sensations.

### *Future work*

This work extends computational modeling for movie-induced cortical responses from supervised CNNs to an unsupervised inference model, and reconstructs dynamic natural vision with a generalizable framework. As for model architecture, biological inference systems are likely to have hierarchically organized sensory causes (Friston 2010) and updating schemes with temporal information (Friston and Kiebel 2009, Huang and Rao 2011). VAE processes static images and can only serve as a spatial-temporally flattened inference system with end-to-end model architecture (Kingma and Welling 2013). Therefore, an improved inference model with spatial and temporal hierarchy would be needed towards more inference-theory-plausible neural coding. As for visual reconstruction, previous studies have suggested that feature maps of supervised CNN can be predicted from brain activities (Horikawa and Kamitani 2017, Wen et al. 2017b). Since latent variables and feature maps both retain information of visual input, it would be interesting to integrate feature estimation methods with generative models to augment visual reconstruction, and even enable the model to decode mental states, imagery or dreams. Besides, another property of variational autoencoder is feature disentangling (Kingma and Welling 2013, Higgins et al. 2016). However, to our knowledge, in current AI literatures there is no report on disentangling the factors of natural images by using VAE. A technical improvement on this would provide potential avenues to analyze the feature disentangling property inside human brain (DiCarlo, Zoccolan and Rust 2012).

## Reference

Adolf, D., S. Weston, S. Baecke, M. Luchtmann, J. Bernarding & S. Kropf (2014) Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in neuroinformatics,* 8**,** 72.

Alink, A., C. M. Schwiedrzik, A. Kohler, W. Singer & L. Muckli (2010) Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience,* 30**,** 2960-2966.

Barlow, H. B. (1989) Unsupervised learning. *Neural computation,* 1**,** 295-311.

Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries & K. J. Friston (2012) Canonical microcircuits for predictive coding. *Neuron,* 76**,** 695-711.

Berens, P. (2009) CircStat: a MATLAB toolbox for circular statistics. *J Stat Softw,* 31**,** 1-21.

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

Cowen, A. S., M. M. Chun & B. A. Kuhl (2014) Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage,* 94**,** 12-22.

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

Doersch, C. (2016) Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Du, C., C. Du & H. He (2017) Sharing deep generative representation for perceived image reconstruction from human brain activity. *arXiv preprint arXiv:1704.07575*.

Eickenberg, M., A. Gramfort, G. Varoquaux & B. Thirion (2017) Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage,* 152**,** 184-194.

Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci,* 11**,** 127-38.

Friston, K. & S. Kiebel (2009) Predictive coding under the free-energy principle. *Philos Trans R Soc Lond B Biol Sci,* 364**,** 1211-21.

Garrido, M. I., J. M. Kilner, K. E. Stephan & K. J. Friston (2009) The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology,* 120**,** 453-463.

Glasser, M. F., S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster & J. R. Polimeni (2013) The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage,* 80**,** 105-124.

Goodfellow, I., Y. Bengio & A. Courville. 2016. *Deep learning*. MIT press.

Gregor, K., I. Danihelka, A. Graves, D. J. Rezende & D. Wierstra (2015) DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

Guclu, U. & M. A. van Gerven (2015) Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J Neurosci,* 35**,** 10005-14.

Güçlütürk, Y., U. Güçlü, K. Seeliger, S. Bosch, R. van Lier & M. van Gerven (2017) Deep adversarial neural decoding. *arXiv preprint arXiv:1705.07109*.

He, K., X. Zhang, S. Ren & J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

Higgins, I., L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed & A. Lerchner (2016) Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*.

Hinton, G. E., P. Dayan, B. J. Frey & R. M. Neal (1995) The Wake-Sleep Algorithm for Unsupervised Neural Networks. *Science,* 268**,** 1158-1161.

Horikawa, T. & Y. Kamitani (2017) Generic decoding of seen and imagined objects using hierarchical visual features. *Nat Commun,* 8**,** 15037.

Huang, Y. & R. P. N. Rao (2011) Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science,* 2**,** 580-593.

Hubel, D. H. & T. N. Wiesel (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol,* 160**,** 106-54.

Jammalamadaka, S. R. & A. Sengupta. 2001. *Topics in circular statistics*. World Scientific.

Kamitani, Y. & F. Tong (2005) Decoding the visual and subjective contents of the human brain. *Nature neuroscience,* 8**,** 679-685.

Kay, K. N., T. Naselaris, R. J. Prenger & J. L. Gallant (2008) Identifying natural images from human brain activity. *Nature,* 452**,** 352-5.

Kay, K. N., J. Winawer, A. Mezer & B. A. Wandell (2013) Compressive spatial summation in human visual cortex. *Journal of neurophysiology,* 110**,** 481-494.

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

Kietzmann, T. C., P. McClure & N. Kriegeskorte (2017) Deep Neural Networks In Computational Neuroscience. *bioRxiv***,** 133504.

Kingma, D. & J. Ba (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. & M. Welling (2013) Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kriegeskorte, N. (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu Rev Vis Sci,* 1**,** 417-446.

Krizhevsky, A., I. Sutskever & G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1097-1105. Chicago.

LeCun, Y., Y. Bengio & G. Hinton (2015) Deep learning. *Nature,* 521**,** 436-44.

Lotter, W., G. Kreiman & D. Cox (2016) Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Mirza, M., A. Courville & Y. Bengio (2016) Generalizable Features From Unsupervised Learning. *arXiv preprint arXiv:1612.03809*.

Miyawaki, Y., H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato & Y. Kamitani (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron,* 60**,** 915-929.

Nair, V. & G. E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807-814.

Naselaris, T., K. N. Kay, S. Nishimoto & J. L. Gallant (2011) Encoding and decoding in fMRI. *Neuroimage,* 56**,** 400-10.

Naselaris, T., R. J. Prenger, K. N. Kay, M. Oliver & J. L. Gallant (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron,* 63**,** 902-915.

Nestor, A., D. C. Plaut & M. Behrmann (2016) Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences,* 113**,** 416-421.

Nishimoto, S., A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu & J. L. Gallant (2011) Reconstructing visual experiences from brain activity evoked by natural movies. *Curr Biol,* 21**,** 1641-6.

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

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla & M. Bernstein (2015) Imagenet large scale visual recognition challenge. *International Journal of Computer Vision,* 115**,** 211-252.

Salin, P. A. & J. Bullier (1995) Corticocortical connections in the visual system: structure and function. *Physiol Rev,* 75**,** 107-54.

Schoenmakers, S., M. Barth, T. Heskes & M. van Gerven (2013) Linear reconstruction of perceived images from human brain activity. *NeuroImage,* 83**,** 951-961.

Shi, J., H. Wen, Y. Zhang, K. Han & Z. Liu (2017) Deep Recurrent Neural Network Reveals a Hierarchy of Process Memory during Dynamic Natural Vision. *bioRxiv***,** 177196.

Simonyan, K. & A. Zisserman (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Summerfield, C., E. H. Trittschuh, J. M. Monti, M.-M. Mesulam & T. Egner (2008) Neural repetition suppression reflects fulfilled perceptual expectations. *Nature neuroscience,* 11**,** 1004-1006.

Tenenhaus, M., V. E. Vinzi, Y.-M. Chatelin & C. Lauro (2005) PLS path modeling. *Computational statistics & data analysis,* 48**,** 159-205.

van Gerven, M. A., F. P. de Lange & T. Heskes (2010a) Neural decoding with hierarchical generative models. *Neural Comput,* 22**,** 3127-42.

van Gerven, M. A. J., F. P. de Lange & T. Heskes (2010b) Neural decoding with hierarchical generative models. *Neural Computation,* 22**,** 3127-3142.

van Hateren, J. H. & A. van der Schaaf (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc Biol Sci,* 265**,** 359-66.

Wacongne, C., J.-P. Changeux & S. Dehaene (2012) A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience,* 32**,** 3665-3678.

Wang, Z., A. C. Bovik, H. R. Sheikh & E. P. Simoncelli (2004) Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing,* 13**,** 600-612.

Wen, H., J. Shi, W. Chen & Z. Liu (2017a) Deep Residual Network Reveals a Nested Hierarchy of Distributed Cortical Representation for Visual Categorization. *bioRxiv*, 151142.

Wen, H., J. Shi, Y. Zhang, K.-H. Lu, J. Cao & Z. Liu (2017b) Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*.

Wu, M. C., S. V. David & J. L. Gallant (2006) Complete functional characterization of sensory neurons by system identification. *Annu Rev Neurosci,* 29**,** 477-505.

Yamins, D. L., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert & J. J. DiCarlo (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A,* 111**,** 8619-24.

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

Yuille, A. & D. Kersten (2006) Vision as Bayesian inference: analysis by synthesis? *Trends Cogn Sci,* 10**,** 301-8.