# 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 and feedforward-only 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 used VAE to predict or decode cortical activity observed with functional magnetic resonance imaging (fMRI) from three human subjects passively watching natural videos. Compared to CNN, VAE resulted in relatively lower accuracies for predicting the fMRI responses to the video stimuli, especially for higher-order ventral visual areas. However, VAE offered a more convenient strategy for decoding the fMRI activity to reconstruct the video input, by first converting the fMRI activity to the VAE's latent variables, and then converting the latent variables to the reconstructed video frames through the VAE's decoder. This strategy was more advantageous than alternative decoding methods, e.g. partial least square regression, by reconstructing both the spatial structure and color of the visual input. Findings from this study support the notion that the brain, at least in part, bears a generative model of the visual world.



**Keywords**: neural coding, variational autoencoder, Bayesian brain, visual reconstruction

# Introduction

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

In one class of hypotheses, the visual system consists of feature detectors that progressively extract and integrate features for pattern recognition. For example, Gabor and wavelet filters model low-level features (Hubel and Wiesel, 1962; van Hateren and van der Schaaf, 1998), and explain responses in early visual areas (Kay et al., 2008; Nishimoto et al., 2011). In contrast, convolutional neural networks (CNNs) encode hierarchical features in a feedforward model (LeCun et al., 2015), and support high performance in image recognition (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Recent studies have shown that CNNs bear similar representations as does the brain (Cichy et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014), and yield high performance in neural encoding and decoding of natural vision (Eickenberg et al., 2017; Guclu and van Gerven, 2015; Horikawa and Kamitani, 2017; Wen et al., 2017b; Yamins et al., 2014). For these reasons, supervised CNN models are gaining attention as favorable models of the visual cortex (Kriegeskorte, 2015; Yamins and DiCarlo, 2016). However, biological learning is not always supervised towards a single goal, but often unsupervised (Barlow, 1989). The visual

cortex has not only feedforward but also feedback pathways (Bastos et al., 2012; Salin and Bullier, 1995). As such, CNNs are not an ideal computational account of human vision.

In another class of hypotheses, the brain builds a causal model of the world, through which it tries to infer what generates the sensory input in order to make proper perceptual and behavioral decisions (Friston, 2010). In this scenario, 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 infers the "cause" of the input, and its top-down process predicts the input (Fig. 1.A). Both processes are optimized by learning from visual experience in order to avoid the "surprise" or error of prediction (Friston and Kiebel, 2009; Rao and Ballard, 1999). This hypothesis takes into account both feedforward and feedback pathways. It aligns with the humans' ability to construct mental images, and offers a basis for unsupervised learning. Thus, it is compelling for both computational neuroscience (Bastos et al., 2012; Friston, 2010; Rao and Ballard, 1999; Yuille and Kersten, 2006) and artificial intelligence (Hinton et al., 1995; Lotter et al., 2016; Mirza et al., 2016).

In line with this notion, variational auto-encoder (VAE) uses independent "latent" variables to code the causes of the visual world (Kingma and Welling, 2013). VAE learns the latent variables from images via an encoder, and samples the latent variables to generate new images via a decoder. Both the encoder and the decoder are neural networks trainable without supervision from unlabeled images (Doersch, 2016). Hypothetically, VAE offers a potential model of the brain's visual system, and may enable an effective way to decode brain activity during either visual perception or imagery (Du et al., 2017; Güçlütürk et al., 2017; Shen et al., 2017; van Gerven et al., 2010a). In this study, we explored the relationships between the VAE's learning objective and the "free-energy" principle in neuroscience (Friston, 2010). To test VAE as a model of the visual cortex, we built

and trained a VAE for unsupervised learning of visual representations, and evaluated the use of VAE for encoding and decoding functional magnetic resonance imaging (fMRI) responses to naturalistic movie stimuli (Fig. 1B).

# Methods and Materials

### *Theory: variational auto-encoder*

In general, 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 neural nets. The encoder learns latent variables from the input, and the decoder generates outputs similar to the input from samples of the latent variables. Given large training datasets, the encoder and the decoder are trained altogether 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 hidden causes or attributes of the images.

Mathematically, let $\boldsymbol{z}$ be the latent variables and $\boldsymbol{x}$ be an image. 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.

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

Since the KL 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 $L(\boldsymbol{\theta}, \boldsymbol{\varphi}; \boldsymbol{x})$ given the training samples of $\boldsymbol{x}$.

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

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 ($\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$) with the re-parameterization trick (Kingma and Welling, 2013). The parameters in VAE could be optimized iteratively using gradient-descent algorithms with the Adam optimizer (Kingma and Ba, 2014).

Similar concepts have been put forth in computational neuroscience theories, for example the free-energy principle (Friston, 2010). In the free-energy principle, the brain's perceptual system includes bottom-up and top-down pathways. Like the encoder in VAE, the bottom-up pathway infers the causes of sensation as probabilistic representations. Like the decoder in VAE, the top-down pathway predicts the sensation from its causes inferred by the brain. Both the bottom-up and top-down pathways are shaped by experiences, 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 the free energy principle are similar, both aiming to minimize the lower bound of the marginal likelihood (or the free energy),

which depends on the difference (or the KL divergence) between the inferred and hidden causes of sensory data while maximizing the likelihood of the sensory data given the inferred causes.

### *Training VAE with diverse natural images*

We designed a VAE with 1,024 latent variables, and the encoder and the decoder were both convolutional neural nets 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 to 128×128×3. To enlarge the amount of training data, the original training images were randomly flipped in the horizontal direction, 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 [0, 1]; the normalized intensity was viewed as the probability of color emission (Gregor et al., 2015). To train the VAE, the Adam optimizer (Kingma and Ba, 2014) was used with a learning rate of 1e-4, as 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 according to the guidelines and regulations in the 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 13.7 hours. The videos were 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 encoding or decoding models. The videos in the training and testing datasets were independent for unbiased model evaluation. Both the training and testing movies were further split into 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 segments of the testing movie were used to test the encoding model. One of the five segments of the testing movie was used to test the decoding models for visual reconstruction, because this segment contained video clips that were continuous over relatively long periods (mean$\pm$std: 13.3$\pm$4.8 s).

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 blood oxygenation level dependent (BOLD) fMRI 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***

After training, VAE extracted the latent representation of any video by a feed-forward pass of every video frame into the encoder, and reconstructed every video frame by a feedback pass of

the latent representation into 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. The voxel-wise fMRI signal was estimated as a linear combination of all unit activities in 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-wise activity given the training movies. Then, the layer-wise activity was concatenated across layers; PCA was applied again to the concatenated activity to keep 99% of the variance of the activity from all layers given the training movies. See details in our earlier paper (Wen et al., 2017a). Following the dimension reduction, the principal components of unit-activity were down-sampled by the sampling rate of fMRI and were used as the regressors to predict the fMRI signal at each voxel through a linear regression model specifically estimated for the voxel.

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

$$y_i^{(j)} = \boldsymbol{w}_i^T \mathbf{z}^{(j)} + b_i + \epsilon_i \tag{4}$$

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).

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

where $N$ 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 parameter $\lambda_i$ was optimized, the model was refitted with the entire training dataset and the optimized parameter.

***Evaluation of encoding performance***

After the above model training, we tested the voxel-wise encoding models with the testing movies, which were different from the training movies to ensure unbiased model evaluation. For each voxel, the encoding performance was evaluated 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-wise permutation test with a block size of 24-sec and 100,000 permutations and corrected at false discovery rate (FDR) $q < 0.01$, as described in our earlier papers (Shi et al., 2017; Wen et al., 2017b).

We compared the encoding performance against the so-called "noise ceiling". It indicated the upper-limit of predictability given the presence of "noise" unrelated to the visual stimuli (David and Gallant, 2005; Nili et al., 2014). The noise ceiling was estimated using the method described elsewhere (Kay et al., 2013). Briefly, the noise was assumed to follow a Gaussian distribution with a zero mean and an unknown variance that varied across voxels. The response and the noise were assumed to be independent and additive. The variance of the noise was estimated as the squared standard error of the mean of fMRI signal (averaged across the 10 repeated sessions of each testing movie). The variance of the response was taken as the difference between the variance of the data and the variance of the noise. From the signal and noise distributions, the samples of the response and the noise were drawn by Monte Carlo simulation for 1,000 random trials. For each trial, the

signal was simulated as the sum of the simulated response and noise; its correlation coefficient with the simulated response was calculated. This resulted in an empirical distribution of the correlation coefficient for each voxel. We calculated the mean of the distribution as the noise ceiling $r_{NC}$, and interpreted it as the upper limit of the prediction accuracy for any encoding model.

In terms of the encoding performance, we compared the encoding models based on VAE against those based on CNN, which have been explored in recent studies (Eickenberg et al., 2017; Guclu and van Gerven, 2015; Wen et al., 2017b). For this purpose, we used a 18-layer residual network (ResNet-18) (He et al., 2016). Relative to AlexNet (Krizhevsky et al., 2012) or ResNet-50 (He et al., 2016), ResNet-18 had an intermediate level of architectural complexity in terms of the number of layers and the total number of units. Thus, ResNet-18 was a suitable benchmark for comparison with VAE, which had a comparable level of complexity. Briefly, ResNet-18 consisted of 18 hidden layers organized into 6 blocks. The 1$^{st}$ block was a convolutional layer followed by max-pooling; the 2$^{nd}$ through 5$^{th}$ blocks were residual blocks, each being a stack of convolutional layers with a shortcut connection; the 6$^{th}$ block performed the multinomial logistic regression for classification. Typical to CNNs, ResNet-18 encoded increasingly complex visual features from lower to higher layers.

We similarly built and trained voxel-wise regression models to project the representations in ResNet-18 to voxel responses in the brain, using the same training procedure and data as above for VAE-based encoding models. Then, we compared VAE against CNN (ResNet-18) in terms of the encoding performance evaluated in the level of voxels or regions of interest (ROI). In the voxel level, the encoding accuracy was converted from the correlation coefficient to the z score by the Fisher's z-transform for the VAE or CNN-based encoding models. Their difference in the voxel-wise z score was calculated by subtraction. For the ROI-level comparison, multiple ROIs were

selected from existing cortical parcellation (Glasser et al., 2016), including 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 divided by the noise ceiling, and then averaged over all voxels in the ROI. The average prediction accuracy of each ROI was compared between VAE and ResNet-18.

***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 fMRI response pattern to the latent variables in VAE through a linear regression model, and 2) transforming the latent variables to pixel patterns through the VAE's decoder. Here we used a cortical mask that covered the visual cortex, and used the voxels within the mask as the input to the decoding model as in our previous study (Wen et al., 2017b).

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 means of the latent variables given the visual input $\boldsymbol{x}^{(j)}$. As Eq. (6), a multivariate linear regression model was defined to predict $\boldsymbol{z}^{(j)}$ given $\boldsymbol{y}^{(j)}$.

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

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.

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.

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

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 the mini-batch stochastic gradient-descent algorithm with a batch size of 100 and a learning rate of 1e-7.

It follows that the testing movie was reconstructed frame by frame by passing the estimated latent variables through the decoder in VAE, as expressed by Eq. (8)

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

where $\Theta$ is the nonlinear mapping from latent variables to the visual reconstruction defined by the VAE decoder.

***Evaluation of decoding performance***

To evaluate the decoding performance in visual reconstruction, we calculated the Structural Similarity index (SSIM) (Wang et al., 2004) between every reconstructed video frame and the true video frame, yielding a measure of the similarity in the pattern of pixel intensity. The SSIM was further averaged across all video frames in the testing movie.

In addition, we evaluated how well movie reconstruction preserved the color information in the original movie. 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 represented as a vector and the hue vectors were concatenated sequentially across all video

frames to represent the color information in the entire movie. The circular correlation in the concatenated hue vector between the original and reconstructed frames was calculated for each subject. 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 100,000 times permutation (Adolf et al., 2014).

We also compared the performance of the VAE-based decoding method with a previously published decoding method (Cowen et al., 2014). In 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 the variance in the pixel pattern of the movie. 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 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)). As such, the training procedure was identical for both methods, except that the feature space for decoding was different: the latent variables for VAE, and eigen-images for PLSR.

Moreover, 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, 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 a significant loss in image content, structure and color, albeit blurred details (Fig. 2B). The VAE-generated images showed comparable quality for different types of input images (Fig. 2.B). As such, the latent representations in VAE were generalizable across various types of visual objects, or their combinations.

### *VAE predicted movie-induced cortical responses*

Given natural movies as visual input, we further asked 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 fitting the voxel response to a training movie as a linear combination of the VAE's unit responses to the same movie. Then, the trained voxel-wise encoding model was tested with a new testing movie (not used for training) to evaluate the model's prediction accuracy (i.e. the correlation between the predicted and measured fMRI responses). For a large area in the visual cortex (Fig. 3), the VAE-based encoding models could predict the movie-evoked responses with statistically significant accuracy (FDR $q<0.01$). In particular, early visual areas (V1/V2/V3) showed the highest prediction accuracy, whereas the prediction accuracy was relatively lower for higher visual areas along the ventral or dorsal stream (Fig. 3). The VAE-predictable areas were relatively larger 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.

### *Comparing encoding performance for VAE vs. CNN*

We further compared VAE against CNN, which was found to predict cortical responses to natural picture or video stimuli (Eickenberg et al., 2017; Guclu and van Gerven, 2015; Wen et al., 2017b). Fig. 4 shows the prediction accuracies at different ROIs, given the encoding models based on VAE and ResNet-18 – a particular type of CNN. For VAE, the encoding performance was the highest in V1, and decreased progressively towards higher visual areas. ResNet outperformed VAE in all ROIs, especially for higher visual areas along the ventral stream (e.g. FFA/PPA/TPJ), but marginally for early visual areas. Similar findings were observable in the voxel level. Fig. 5 shows the voxel-wise prediction accuracy for VAE or ResNet. ResNet outperformed VAE for most of the visual cortex. Their difference (by subtraction) was much more notable in the ventral-stream areas than early visual areas or those in the dorsal-stream areas. In sum, VAE was in general less predictive of visual cortical activity than was CNN.

### *Direct visual reconstruction by decoding fMRI activity*

We further explored the use of VAE for decoding the fMRI activity to reconstruct the visual input. For this purpose, a decoding model was trained and used to convert the fMRI activity to the VAE's latent-variable representation, which was in turn converted to a pixel pattern through the VAE's decoder. In comparison with the original videos, Fig. 6 shows the visual input reconstructed from fMRI activity based on VAE and the decoding models, which were trained and tested with data from either the same or different subjects. Although the visual reconstruction was too blurry to fully resolve details or discern visual objects, it captured basic information about the dynamic

visual scenes, including the coarse position and shape of objects, and the rough color and contrast. The quality of visual reconstruction was better when the decoding models were trained and tested for the same subject than for different subjects.

We assessed the quality of visual reconstruction by quantifying the structural similarity (as SSIM) (Wang et al., 2004) between the reconstructed and original movies. The VAE-based decoding method yielded a much higher SSIM (about 0.5) than the eigen-image-based benchmark models with either partial least squares regression (Cowen et al., 2014) or $L_1$-regularized linear regression (Fig. 7A). Moreover, the VAE-based visual reconstruction preserved the color information and its dynamics in the movie, showing statistically significant (permutation test, $p < 0.001$) correlations in color index (Hue-value) around 0.25.

Lastly, we decoded the fMRI activity of Subject 2 or 3 based on the decoding model trained from Subject 1 (Fig. 6). By visual inspection, the quality of reconstruction was lower as the model was trained and tested with different subjects. Nevertheless, the reconstruction was qualitatively similar to what was attained with the decoding models trained/tested with the same subject, while preserving basic patterns in the original video frames (Fig. 6). Therefore, the VAE-based decoding model was transferable across subjects.

# Discussion

We designed and trained a variational autoencoder (VAE) to learn visual representations by reconstructing diverse natural images. Unlike CNN, VAE does not require labeled samples, and can be trained by unsupervised learning based on variational inference (Kingma and Welling, 2013). Non-observable causes of visual input are inferred through joint optimization of bottom-up

and top-down components in VAE. In line with the notion of Bayesian brain (Yuille and Kersten, 2006), the bottom-up process infers the causes of any given visual input, and the top-down process generates a "virtual version" of the visual input based on the inferred cause. As a candidate model of the visual cortex, VAE is able to decode fMRI scans for visual reconstruction, and predict fMRI responses to video stimuli. The decoding strategy is generative in nature, yielding encouraging performance in reconstructing videos, while it is likely applicable to decoding of mental images. The encoding performance of VAE is relatively higher in early visual areas than higher visual areas, but overall not as high as the performance of CNN. In summary, the findings in this paper support the Bayesian brain hypothesis, and highlight a generative strategy for decoding naturalistic and diverse visual experiences.

***Bayesian Brain and Free energy.***

This work is inspired by the Bayesian brain theory (Ma et al., 2006; Yuille and Kersten, 2006). This theory has been used to model the brain in various functional domains, including cognitive development (Perfors et al., 2011; Tenenbaum et al., 2011), perception (Knill and Pouget, 2004; Lee and Mumford, 2003), and action (Friston, 2010). This theory assumes that the brain learns a generative model of the world and uses it to infer the hidden cause of sensation (Friston, 2012). Under this theory, the brain runs computation similar to Bayesian inference (Knill and Richards, 1996); visual perception is viewed as a probabilistic process (Pouget et al., 2013), capable of dealing with noisy, incomplete, or ambiguous visual input (Tenenbaum et al., 2011; Yuille and Kersten, 2006).

VAE rests on a very similar notion. The encoder computes latent variables as Gaussian distributions, from which the samples are drawn by the decoder. The latent variable distribution can be used to generate not only the samples already learned, but also samples that are unknown

to the model. VAE also embodies the inference strategy called "analysis by synthesis" (Yuille and Kersten, 2006). In this regard, the top-down generative component is important for optimizing the bottom-up component, to find what cause the input. Given the sensory input, the bottom-up process makes proposals of possible causes that generate the sensation (Yuille and Kersten, 2006), and then the proposals are updated by the top-down generative model with the direct comparison between sensory input and its generated "virtual version" (Friston and Kiebel, 2009). In VAE, this notion is represented by the learning rule of minimizing the reconstruction error – the difference between the input to the encoder and the output from the decoder. This learning rule allows the (bottom-up) encoder and the (top-down) decoder to be trained all together (Kingma and Welling, 2013).

VAE also contributes to the computational approximation of "free-energy", a neuroscience concept to measure the discrepancy between how the sensation is represented by the model, and the way it actually is (Friston, 2010; Hinton and Zemel, 1994). In Bayesian inference, minimizing this discrepancy (also called "surprise") is important for updating the model but is difficult to calculate. Therefore free-energy is proposed as the upper-bound of the sensory surprise which can be minimized coherently by minimizing free-energy (Friston, 2009). Mathematically, free-energy is expressed as the sensory surprise plus the non-negative KL divergence between the inferred causes given the input, and the true hidden causes. This formulation simplifies the inference process to an easier optimization problem by approximating the posterior distributions of visual causes (Friston, 2010). VAE bears the same idea in an artificial neural network. As shown in Eq. (1), the learning objective $L(\boldsymbol{\theta}, \boldsymbol{\varphi}; \boldsymbol{x})$ can be decomposed into two parts: the marginal likelihood of the data $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, minus the KL divergence between inferred latent variables and its ground truth $D_{KL}\left[q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]$. As the marginal likelihood is the negative format of "surprise"

(Friston et al., 2012), VAE is actually optimizing the probability of generated data through minimizing "free-energy".

***VAE vs. CNN***

Our results suggest that VAE, as an implementation of Bayesian brain, turns out to partially explain and decode brain responses to natural videos. Therefore, it lends support to the Bayesian brain theory. However, its encoding and decoding performance is not perfect, or even worse than CNN, which outperforms VAE in nearly all visual areas. While both trained with the same data with different learning objective, the difference in encoding performance between VAE and CNN was most notable in higher visual areas in the ventral stream, which has been known to play an essential role in object/scene recognition (Ungerleider and Haxby, 1994). CNN is explicitly driven by object categorization, because its training is supervised by categorical labels of images. VAE is trained without using any label, thereby with unsupervised learning.

It is likely that supervised learning is required for a model to fully explain ventral-stream activity. A previous study has reached a similar conclusion with different models and an analysis method based on representational similarity (Khaligh-Razavi and Kriegeskorte, 2014). However, this conclusion should be taken with caution. There are many potential learning objectives that can be used for unsupervised learning. The argument on unsupervised versus supervised learning still awaits future studies to fully resolve.

***Towards Robust and Generalizable Natural Visual Reconstruction***

Researchers have long been trying to render evoked brain activities into the sensory input, including edge orientation (Kamitani and Tong, 2005), face images (Cowen et al., 2014; Güçlütürk et al., 2017; Nestor et al., 2016), contrast patterns (Miyawaki et al., 2008), digits (Du et al., 2017;

Qiao et al., 2018; van Gerven et al., 2010b) and handwritten characters (Schoenmakers et al., 2013). Other studies have tried to extend the reconstruction from artificial stimuli to diverse natural stimuli. The reconstruction of diverse natural images or movies has been shown by using the Bayesian reconstruction method (Naselaris et al., 2009; Nishimoto et al., 2011), where the reconstruction is matched to the images or movies in the prior dataset. Recently deep generative models have been used for decoding natural vision. A deep convolutional generative adversarial network (DCGAN) (Radford et al., 2015) enables the reconstruction of handwritten characters and natural images (Seeliger et al., 2017). Besides, a deep generator network (DGN) (Nguyen et al., 2016) has been introduced to add naturalistic constraints to image reconstruction based on decoded human fMRI data (Shen et al., 2017).

Our method is different by its training strategy. Unlike the Bayesian reconstruction method, VAE-based decoding model is generalizable without relying on the natural image prior. Besides, both DCGAN and DGN are optimized with adversarial training (Goodfellow et al., 2014). Whereas adversarial training may constrain visual reconstruction to be naturalistic (Shen et al., 2017), the VAE-based decoding model is purely rooted in variational inference. In addition, our goal is not only to optimize neural decoding for visual reconstruction, but also to test the plausibility of using VAE to model the brain. Although VAE and GAN are both emerging techniques for learning deep generative models (Dumoulin et al., 2016), these two types of networks are not mutually exclusive; instead, they might be combined for potentially better and more generalizable visual reconstruction.

### *Limitations and Future Directions*

Although our results support the Bayesian brain theory, some limitations of VAE should be noted. VAE offers a computational account of Bayesian inference and free-energy minimization

in the brain. However, the biological inference might be implemented in a more complex and dynamic process. In this sense, the brain infers hierarchically organized sensory causes possibly with predictive coding (Huang and Rao, 2011; Rao and Ballard, 1999). Higher-level neural systems attempt to predict the inputs to lower-level ones, and prediction errors of the lower-level are propagated to adapt higher-level systems to reduce the prediction discrepancy (Clark, 2013). Therefore, hierarchical message passing with recurrent and feedback connections might be essential for effective implementation of Bayesian inference in the brain (Bastos et al., 2012; Friston, 2010; Friston and Kiebel, 2009). While VAE cannot offer hierarchically organized sensory causes and temporal processing, these properties might be important and should be incorporated into the future development of similar models.

## Acknowledgements

The authors would like to recognize the inputs from Dr. Eugenio Culurciello, Dr. Alfredo Canziani and Abhishek Chaurasia for their constructive discussions on deep neural networks. The research was supported by NIH R01MH104402 and Purdue University.

# Figure Captions

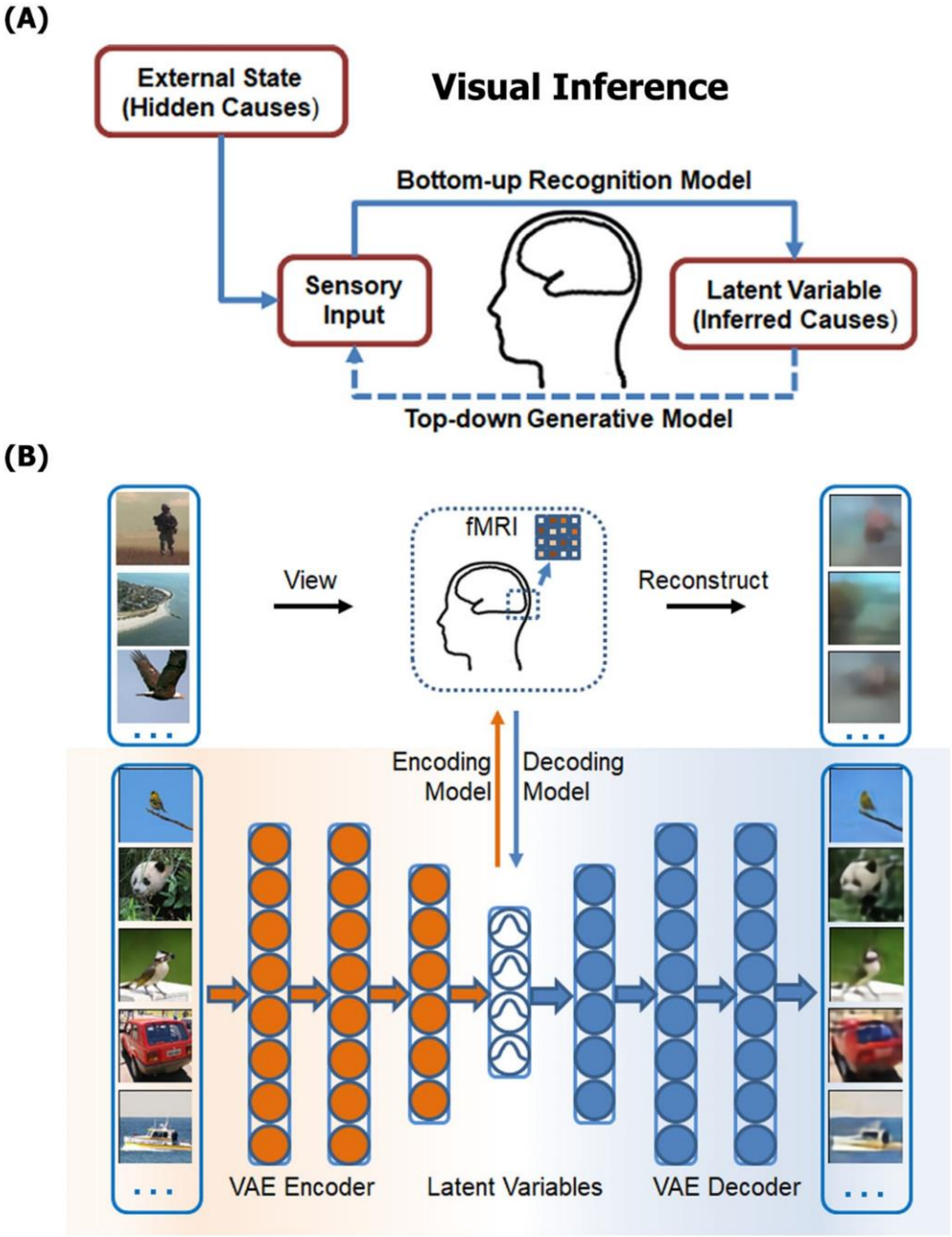


**Figure 1 | A Bayesian model of human vision. (A) The brain models the causal structure of the world for visual inference.** The brain analyzes the sensory input to infer the hidden causes of the input through its bottom-up processes, and predicts the sensory input through its top-down processes. **(B) Encoding and decoding visually-evoked cortical fMRI responses by using VAE as a model of the visual cortex.** For encoding, cortical responses to any visual stimuli are predicted as a linear projection of the VAE responses to the same stimuli. For decoding, visual stimuli were reconstructed by first estimating the VAE's latent variables as a linear function of the fMRI responses, and then generating pixel patterns from the estimated through the VAE's decoder.

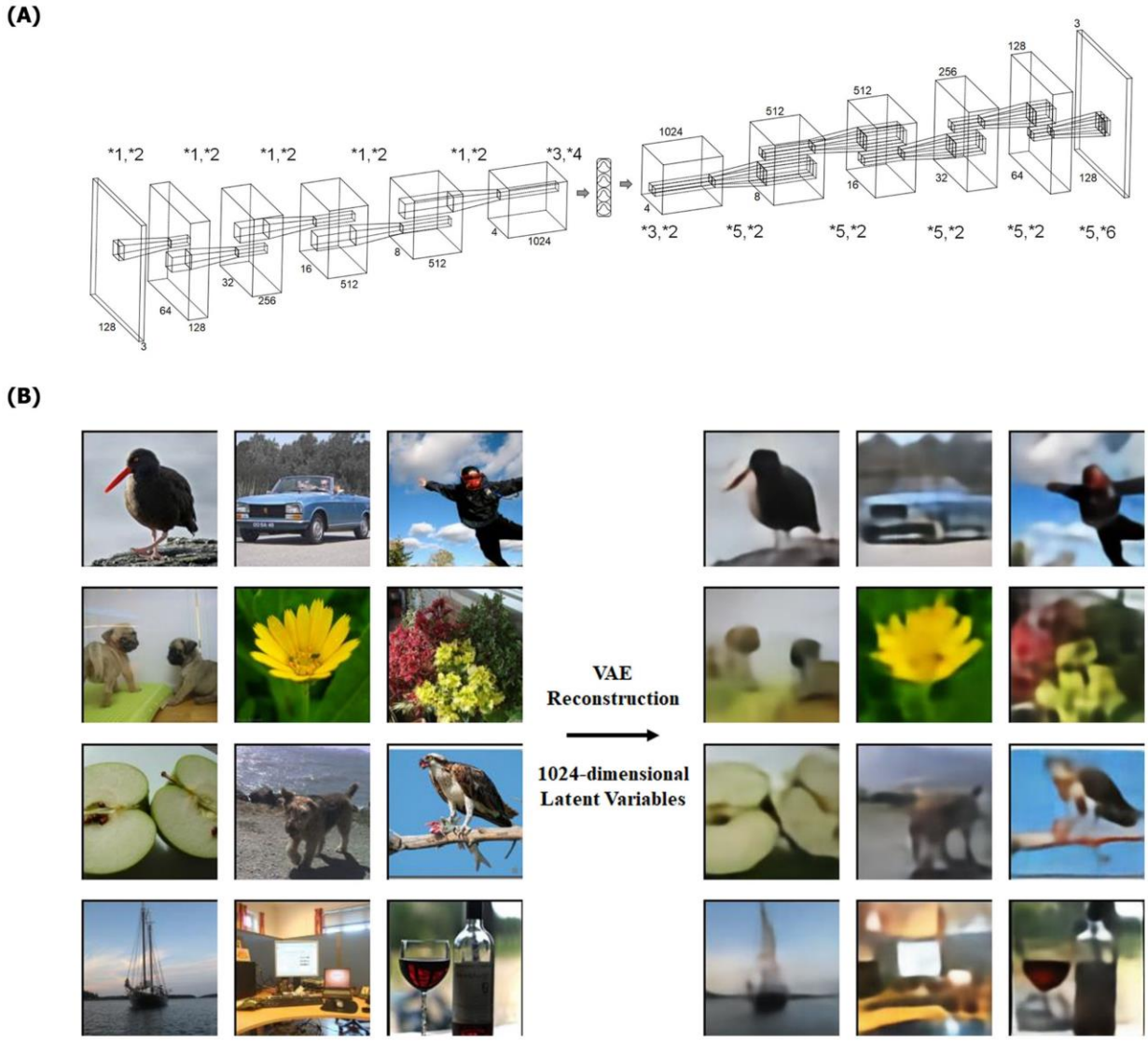


**Figure 2 | Variational auto-encoder. (A) The architecture of VAE for natural images.** The encoder and the decoder both contained 5 layers. The dimension of latent variables was 1024. Operations were defined as: *1 convolution (kernel size=4, stride=2, padding=1), *2 rectified nonlinearity, *3 fully connected layer, *4 re-parameterization, *5 transposed convolution (kernel size = 4, stride = 2, padding = 1), *6 sigmoid nonlinearity. **(B) Reconstruction of natural images by VAE.** For any image (left), its information was encoded by 1024 latent variables by passing it through the VAE encoder. From the latent variables, the VAE decoder generated an image (right) as the reconstruction of the input image, despite blurred details.

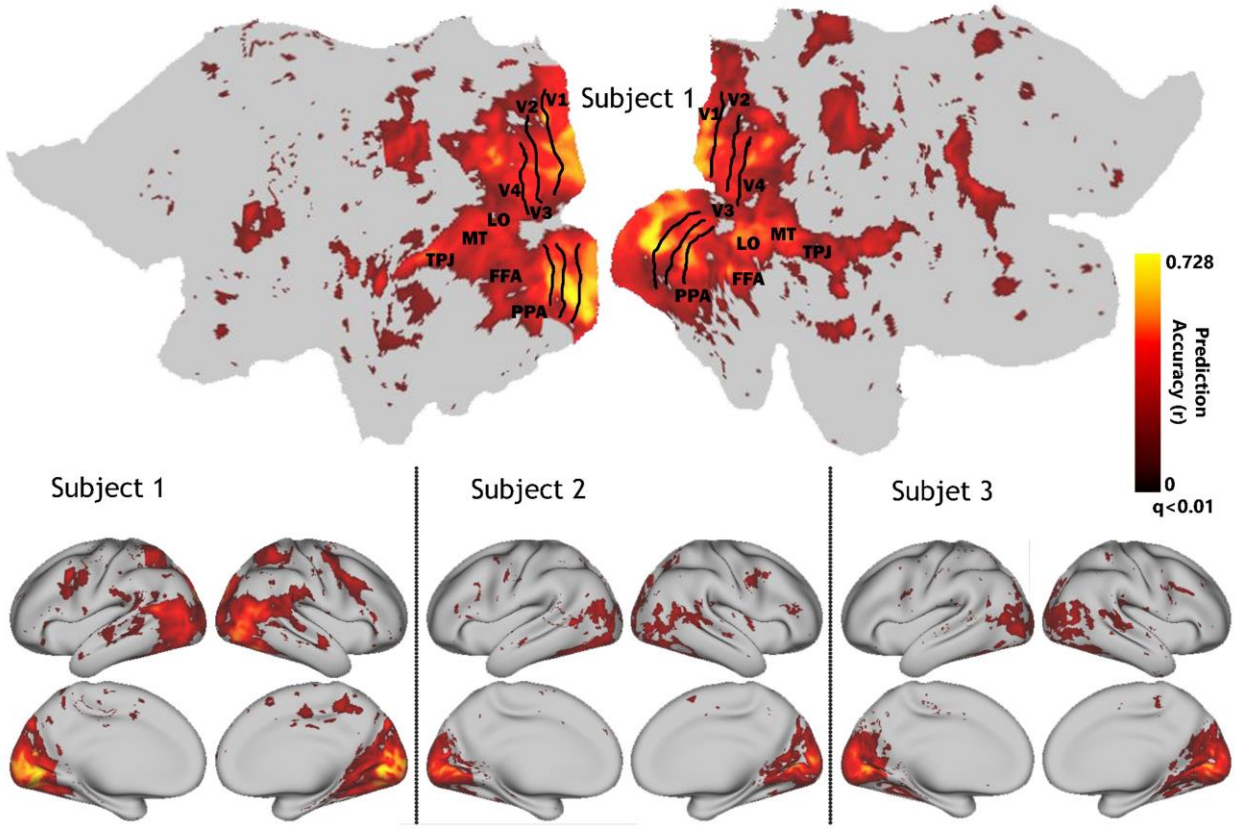


**Figure 3 | Prediction accuracy with VAE-based encoding models.** The accuracy was measured by the Pearson's correlation coefficient (r) between the model-predicted response and the actual fMRI response. The map shows the r value averaged across the five testing movies. The map was thresholded by statistical significance (FDR q<0.01). The results are shown on the flattened (only for Subject 1) and inflated cortical surfaces for every subject.

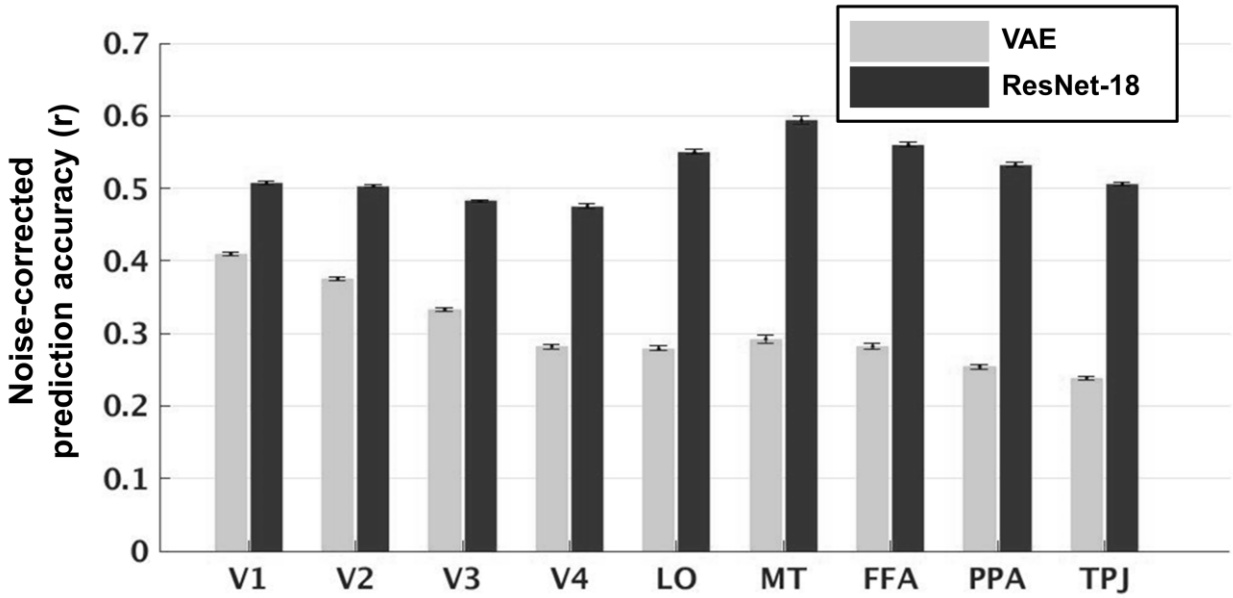


**Figure 4 | The ROI-level encoding performance of VAE vs. ResNet-18.** For either VAE (light array) or ResNet-18 (dark gray), the encoding model's accuracy in predicting the fMRI response to the testing movie was normalized by the noise ceiling and summarized for each of the nine pre-defined ROIs. Arranged from the left to the right, individual ROIs are located in increasingly higher levels of the visual hierarchy. The bar chart was based on the mean±SEM (standard error of the mean) of the voxel-wise prediction accuracy (divided by the noise ceiling) averaged across all the voxels in each ROI, and across different testing movies and subjects.

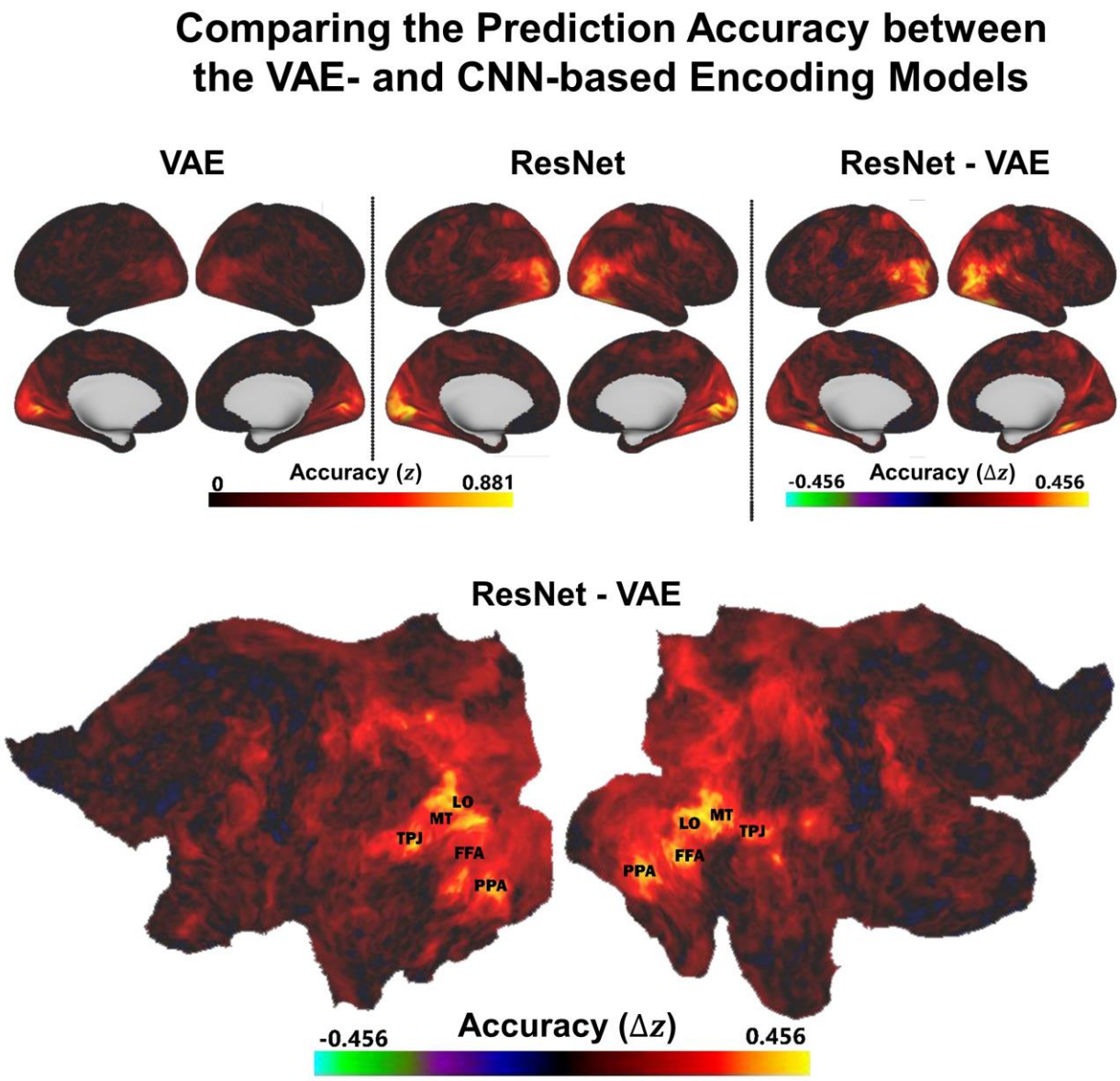


**Figure 5 | Encoding performance with VAE vs. ResNet-18.** The prediction accuracy (the z-transformed correlation between the predicted and measured fMRI responses) is displayed on inflated cortical surfaces for the encoding models based on VAE (top-left) and ResNet-18 (top-middle). Their difference (ResNet – VAE) in the prediction accuracy is displayed on both inflated (top-right) and flattened (bottom) cortex.

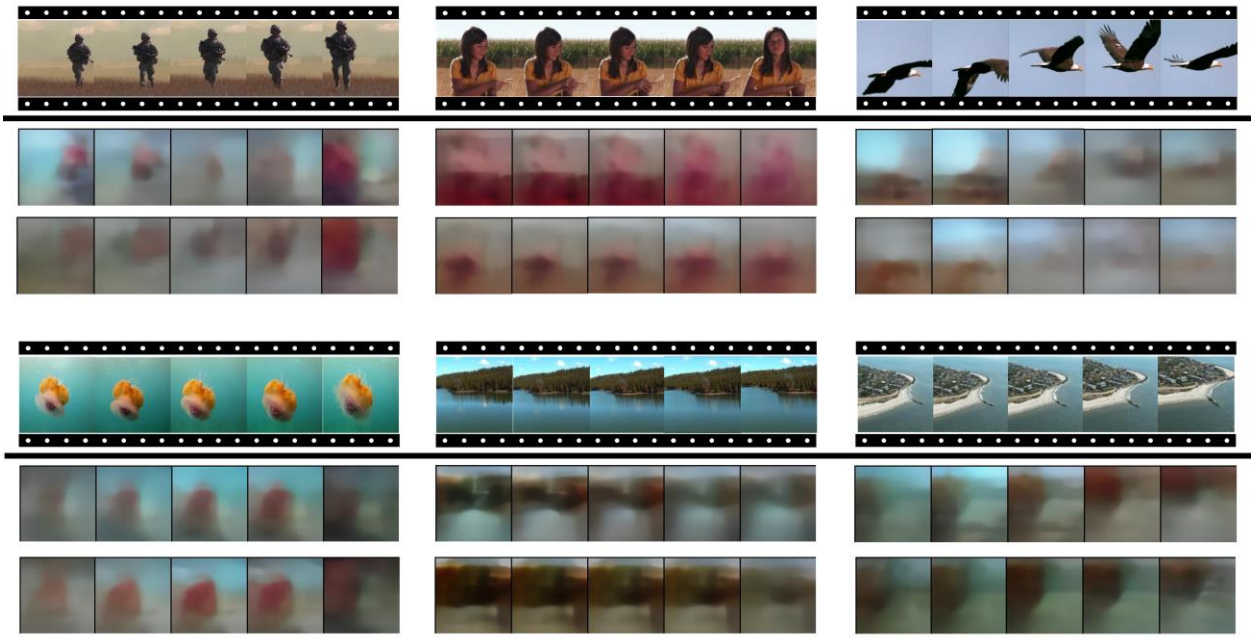

**Figure 6 | Visual reconstruction based on VAE and fMRI.** For each of the 6 example video clips in the testing movie, the top row shows the original video frames, the middle or bottom rows show the frames reconstructed from the measured fMRI responses, based on VAE and the decoding model trained and tested within the same subject (Subject 1), or across different subjects, respectively. The cross-subject decoding models were trained with data from Subject 1, but tested on data from Subject 2 (the top 3 clips) or Subject 3 (the bottom 3 clips).

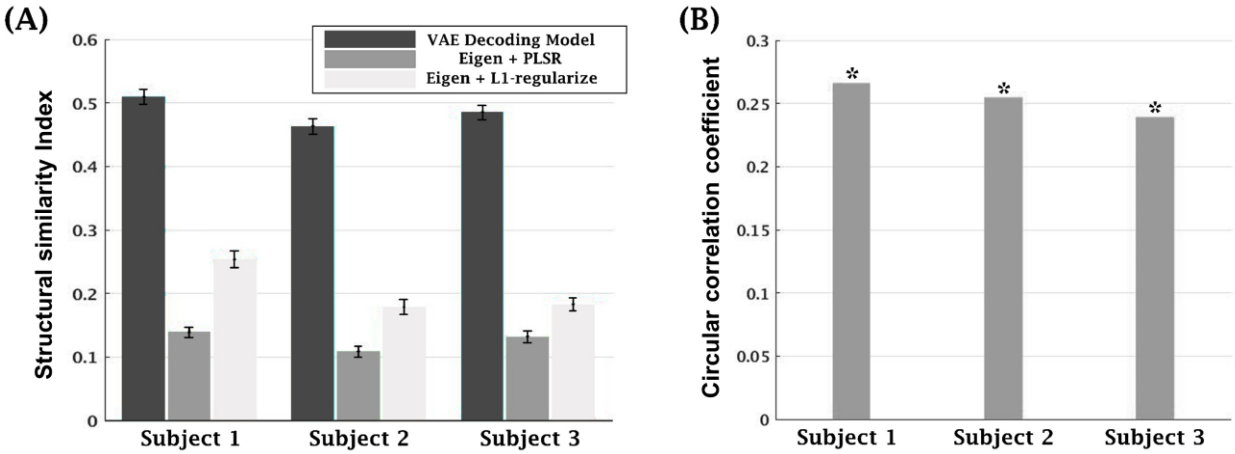


**Figure 7 | Quantitative evaluation of visual reconstruction. (A) Comparison of structural similarity index (SSIM).** SSIM scores of VAE-based decoding model and eigen-image-based models were compared for all 3 subjects. Each bar shows the mean±SE SSIM score over all frames in the testing movie. **(B) Correlation in color (hue-value)**. The (circular) correlation between the original and reconstructed hue components was calculated and evaluated for statistical significance with permutation test (*, $p<0.001$).

# Reference


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

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

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 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., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6, 27755.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36, 181-204.

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

David, S.V., Gallant, J.L., 2005. Predicting neuronal responses during natural vision. Network: Computation in Neural Systems 16, 239-260.

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

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

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., Courville, A., 2016. Adversarially learned inference. arXiv preprint arXiv:1606.00704.

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

Friston, K., 2009. The free-energy principle: a rough guide to the brain? Trends in cognitive sciences 13, 293-301.

Friston, K., 2010. The free-energy principle: a unified brain theory? Nat Rev Neurosci 11, 127-138.

Friston, K., 2012. The history of the future of the Bayesian brain. NeuroImage 62, 1230-1233.

Friston, K., Adams, R., Montague, R., 2012. What is value—accumulated reward or evidence? Frontiers in neurorobotics 6.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364, 1211-1221.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171-178.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, pp. 2672-2680.

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

Guclu, U., van Gerven, M.A., 2015. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J Neurosci 35, 10005-10014.

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

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.
Hinton, G.E., Dayan, P., Frey, B.J., Neal, R.M., 1995. The Wake-Sleep Algorithm for Unsupervised Neural Networks. Science 268, 1158-1161.
Hinton, G.E., Zemel, R.S., 1994. Autoencoders, minimum description length and Helmholtz free energy. Advances in neural information processing systems, pp. 3-10.
Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun 8, 15037.
Huang, Y., Rao, R.P.N., 2011. Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science 2, 580-593.
Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160, 106-154.
Jammalamadaka, S.R., Sengupta, A., 2001. Topics in circular statistics. World Scientific.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nature neuroscience 8, 679-685.
Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352-355.
Kay, K.N., Winawer, J., Mezer, A., Wandell, B.A., 2013. Compressive spatial summation in human visual cortex. Journal of neurophysiology 110, 481-494.
Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10, e1003915.
Kietzmann, T.C., McClure, P., Kriegeskorte, N., 2017. Deep Neural Networks In Computational Neuroscience. bioRxiv, 133504.
Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. TRENDS in Neurosciences 27, 712-719.
Knill, D.C., Richards, W., 1996. Perception as Bayesian inference. Cambridge University Press.
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., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, Chicago, pp. 1097-1105.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436-444.
Lee, T.S., Mumford, D., 2003. Hierarchical Bayesian inference in the visual cortex. JOSA A 20, 1434-1448.
Lotter, W., Kreiman, G., Cox, D., 2016. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104.
Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A., 2006. Bayesian inference with probabilistic population codes. Nature neuroscience 9, 1432-1438.
Mirza, M., Courville, A., Bengio, Y., 2016. Generalizable Features From Unsupervised Learning. arXiv preprint arXiv:1612.03809.
Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y., 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60, 915-929.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400-410.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902-915.
Nestor, A., Plaut, D.C., Behrmann, M., 2016. Feature-based face representations and image reconstruction from behavioral and neural data. Proceedings of the National Academy of Sciences 113, 416-421.
Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J., 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in Neural Information Processing Systems, pp. 3387-3395.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS computational biology 10, e1003553.
Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21, 1641-1646.
Perfors, A., Tenenbaum, J.B., Griffiths, T.L., Xu, F., 2011. A tutorial introduction to Bayesian models of cognitive development. Cognition 120, 302-321.
Pouget, A., Beck, J.M., Ma, W.J., Latham, P.E., 2013. Probabilistic brains: knowns and unknowns. Nature neuroscience 16, 1170-1178.
Qiao, K., Zhang, C., Wang, L., Yan, B., Chen, J., Zeng, L., Tong, L., 2018. Accurate reconstruction of image stimuli from human fMRI based on the decoding model with capsule network architecture. arXiv preprint arXiv:1801.00602.
Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2, 79-87.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211-252.
Salin, P.A., Bullier, J., 1995. Corticocortical connections in the visual system: structure and function. Physiol Rev 75, 107-154.
Schoenmakers, S., Barth, M., Heskes, T., van Gerven, M., 2013. Linear reconstruction of perceived images from human brain activity. NeuroImage 83, 951-961.
Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., van Gerven, M., 2017. Generative adversarial networks for reconstructing natural images from brain activity. bioRxiv, 226688.
Shen, G., Horikawa, T., Majima, K., Kamitani, Y., 2017. Deep image reconstruction from human brain activity. bioRxiv, 240317.
Shi, J., Wen, H., Zhang, Y., Han, K., Liu, Z., 2017. Deep Recurrent Neural Network Reveals a Hierarchy of Process Memory during Dynamic Natural Vision. bioRxiv, 177196.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: Statistics, structure, and abstraction. science 331, 1279-1285.
Tenenhaus, M., Vinzi, V.E., Chatelin, Y.-M., Lauro, C., 2005. PLS path modeling. Computational statistics & data analysis 48, 159-205.
Ungerleider, L.G., Haxby, J.V., 1994. 'What'and 'where'in the human brain. Current opinion in neurobiology 4, 157-165.
van Gerven, M.A., de Lange, F.P., Heskes, T., 2010a. Neural decoding with hierarchical generative models. Neural Comput 22, 3127-3142.
van Gerven, M.A.J., de Lange, F.P., Heskes, T., 2010b. Neural decoding with hierarchical generative models. Neural Computation 22, 3127-3142.

van Hateren, J.H., van der Schaaf, A., 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. Proc Biol Sci 265, 359-366.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600-612.
Wen, H., Shi, J., Chen, W., Liu, Z., 2017a. Deep Residual Network Reveals a Nested Hierarchy of Distributed Cortical Representation for Visual Categorization. bioRxiv, 151142.
Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., Liu, Z., 2017b. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. Cerebral Cortex.
Wu, M.C., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29, 477-505.
Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111, 8619-8624.
Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience 19, 356-365.
Yuille, A., Kersten, D., 2006. Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10, 301-308.