Reconstructing Perceived Images From Human Brain Activities With Bayesian Deep Multiview Learning

Changde Du, Changying Du, Lijie Huang, and Huiguang He, *Senior Member, IEEE*

***Abstract*— Neural decoding, which aims to predict external visual stimuli information from evoked brain activities, plays an important role in understanding human visual system. Many existing methods are based on linear models, and most of them only focus on either the brain activity pattern classification or visual stimuli identification. Accurate reconstruction of the perceived images from the measured human brain activities still remains challenging. In this paper, we propose a novel deep generative multiview model for the accurate visual image reconstruction from the human brain activities measured by functional magnetic resonance imaging (fMRI). Specifically, we model the statistical relationships between the two views (i.e., the visual stimuli and the evoked fMRI) by using two view-specific generators with a shared latent space. On the one hand, we adopt a deep neural network architecture for visual image generation, which mimics the stages of human visual processing. On the other hand, we design a sparse Bayesian linear model for fMRI activity generation, which can effectively capture voxel correlations, suppress data noise, and avoid overfitting. Furthermore, we devise an efficient mean-field variational inference method to train the proposed model. The proposed method can accurately reconstruct visual images via Bayesian inference. In particular, we exploit a posterior regularization technique in the Bayesian inference to regularize the model posterior. The quantitative and qualitative evaluations conducted on multiple fMRI data sets demonstrate the proposed method can reconstruct visual images more accurately than the state of the art.**

***Index Terms*— Deep neural network (DNN), image reconstruction, multiview learning, neural decoding, variational Bayesian inference.**

I. Introduction

NEURAL encoding and decoding are two fundamental aspects to understand human visual processing system [1], [2]. Encoding models aim to predict the brain responses according to the presented external stimuli, e.g., visual images. In contrast, decoding models aim to predict the external visual stimuli information by analyzing the evoked brain signals [3]–[6]. Although significant progresses have been made in this field [7]–[10], most of them are based on linear models and only focus on either the brain activity pattern classification [11]–[13] or visual stimuli identification [4], [5], [14]. Accurately reconstructing the perceived images from the human brain activities measured by functional magnetic resonance imaging (fMRI) still remains challenging. The reasons are typically threefold: 1) linear mappings between the visual images and the evoked brain activities have limited representation power; 2) we only have a small number of paired data (stimulus-response); and 3) the brain activities recorded by fMRI are high-dimensional data often degraded by complex noise.

Traditional visual image reconstruction approaches [15]–[17] generally suffer from the above-mentioned issues and hence yield unsatisfactory results. For example, Fujiwara *et al.* [18] developed a Bayesian canonical correlation analysis (BCCA) model for estimating reversible mappings between the visual images and the evoked brain activities. However, its linear architecture greatly limits its ability to learn the hierarchical visual features from images. In addition, its spherical covariance assumption cannot capture the correlations among fMRI voxels, making it susceptible to voxel noise.

In recent years, deep learning methods [19], [20], especially the deep neural networks (DNNs), revolutionized several fields of machine learning, ranging from computer vision [21], [22] to speech recognition [23], [24]. The architectures of

Manuscript received December 11, 2017; revised May 5, 2018, September 17, 2018 and October 17, 2018; accepted November 9, 2018. Date of publication December 12, 2018; date of current version July 17, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 91520202, 61602449, in part by the CAS Scientific Equipment Development Project under Grant YJKYYQ20170050, in part by the Beijing Municipal Science and Technology Commission under Grant Z181100008918010, in part by the Youth Innovation Promotion Association CAS, and in part by the Strategic Priority Research Program of CAS. *(Corresponding author: Huiguang He.)*

C. Du is with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: duchangde@gmail.com).

C. Du is with the Laboratory of Parallel Software and Computational Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China, and also with the 360 Search Lab, Beijing 100015, China (e-mail: changying@iscas.ac.cn).

L. Huang is with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lijie.huang@ia.ac.cn).

H. He is with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: huiguang.he@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2882456

DNNs are loosely inspired by the computational principles of the biological nervous system [25]. For example, the hierarchical layers of DNNs can resemble feedforward visual representations in the human brain visual system [26]. A number of studies [14], [27]–[29] have revealed that the neural activations in human visual cortex show correspondence with the outputs of DNN layers. Therefore, it is reasonable to explore the DNNs' capability in neural encoding and decoding. Furthermore, Higgins *et al.* [30] demonstrated that the unsupervised deep generative models such as variational autoencoders (VAEs) [31], [32] can learn the disentangled image representations that correspond to distinct visual concepts. This is critical to neural decoding, since the visual concepts learned by these unsupervised deep generative models may also be perceived by the human brain.

Motivated by the aforementioned discussions, we propose a deep generative multiview model (DGMM) to reconstruct perceived images from human brain activities. For modeling the statistical relationships between the visual images and the evoked fMRI activity patterns, we assume these two data views can be generated from a shared latent space via two view-specific generative models. Specifically, we apply a deep generative model to visual images, while a sparse linear generative model to fMRI activity patterns. On the one hand, the DNN architecture in the deep generative model can capture the stages of human visual processing [14], [27], [28] and, hence, provides better representations than the linear models. On the other hand, as the brain activities are the high-dimensional fMRI data and the sample size is usually small, using a sparse linear model for brain activity generation can effectively avoid overfitting. Furthermore, to capture the correlations among fMRI voxels, we impose a full-covariance matrix on the distribution of fMRI activity. But this assumption results in severe computational issues. To reduce the computational complexity, we further impose a low-rank assumption on this full-covariance matrix by introducing a set of auxiliary latent variables. Inspired by recent advances in scalable variational methods [31], [32], we train the proposed model by using an efficient mean-field variational inference method. After training, the proposed DGMM method can accurately reconstruct the visual images via Bayesian inference. In particular, we regularize the model posterior via posterior regularization [33], which is a technique for regularizing models by encoding specific prior knowledge into the model posteriors. As a result, the posterior regularization can force the latent representations of the test samples to be close to that of their neighbors from the training set.

Compared with the deterministic deep multiview learning methods [34], [35], the proposed Bayesian framework enjoys the inherent advantage of avoiding overfitting. By harnessing the posterior regularization, DGMM is capable of incorporating specific prior knowledge (e.g., the similarity information between different brain activity patterns) into the Bayesian inference of model posterior. We apply the proposed DGMM method to three public fMRI data sets, including binary contrast patterns [15], handwritten digits [16], and handwritten characters [17]. The quantitative and qualitative evaluations demonstrate that the proposed DGMM can reconstruct visual images more accurately than the state of the art. Our main contributions can be summarized as follows.

1) We describe a new deep generative multiview framework for neural decoding by employing the fusion of probabilistic modeling and DNNs. The generation and inference procedures in the deep generative model naturally support the cognitive phenomena of imagination [36].
2) We impose a full-covariance matrix on the distribution of fMRI activity to capture the correlations among voxels. To reduce the computational complexity, we further impose a low-rank assumption on this full-covariance matrix by introducing a set of auxiliary latent variables.
3) We derive a predictive distribution for perceived images, which takes the uncertainty of data into account. In particular, we show that posterior regularization can be introduced into neural decoding to improve the prediction performance.
4) We devise a mean-field variational inference method to train the proposed model efficiently.
5) The quantitative and qualitative evaluations demonstrate that our approach can reconstruct visual images more accurately than the state of the art.

Our study has strong connections to the research in DNNs and related learning systems. First of all, the proposed deep generative multiview architecture is a novel design of DNNs. Furthermore, perceived image reconstruction from human brain activities using the proposed DGMM is an effective application of DNNs. Finally, we show that DNN models have the potential capability to mimic human visual processing. Researchers can utilize brain-inspired methods to design more efficient neural networks and learning systems.

II. Related Works

A. Neural Decoding

Following the pioneering work in [37], a number of neural decoding approaches [3]–[6], [11]–[14] have been proposed in the past decade. It can be roughly divided into three categories, depending on the decoding type: 1) *brain activity pattern classification* determines which category of stimulus elicits the observed brain signals; 2) *visual stimuli identification* identifies a specific stimulus (from a candidate set) that best explains the observed brain signals; and 3) *visual stimuli reconstruction* reconstructs the corresponding visual stimuli according to the observed brain signals.

Although previous studies have made significant progresses in brain activity pattern classification [11]–[13], [37] and visual stimuli identification [4], [5], [14], the performance of accurate visual image reconstruction [6], [15], [18] still needs to be improved. For example, Miyawaki *et al.* [15] proposed a multiscale image bases method to reconstruct the binary contrast patterns. However, its results are not optimal because the image bases used in this method have the predefined shapes. To overcome this limitation, Fujiwara *et al.* [18] developed the BCCA model to learn the image bases automatically. Nevertheless, BCCA still has two critical drawbacks. First of all, its linear architecture has limited representation power in extracting the hierarchical visual features from images. In addition, its spherical covariance assumption cannot effectively capture the correlations among fMRI voxels. We argue that exploring the correlations among voxels is critical to visual image reconstruction.

Different from previous works, we employ a DNN architecture to extract the hierarchical visual features from images. Furthermore, we design a sparse linear model with a full-covariance matrix assumption to fit the high-dimensional fMRI data, thereby allowing us to capture the correlations among voxels.

B. Deep Generative Models

There has been a surge of research interest in deep generative models in recent years. Two of the most commonly used approaches are VAE [31] and generative adversarial network (GAN) [38]. VAE aims at maximizing the variational lower bound of the data likelihood and GAN aims at achieving an equilibrium between a generator and a discriminator. The most important use of deep generative models is to generate high-quality images [39]–[41].

Recently, techniques for applying VAEs and GANs to neural decoding have emerged [42]–[44]. For example, Du *et al.* [42] proposed a neural decoding model based on the VAE framework. However, the authors only took the visual images into account in the posterior inference phase. Unlike [42], we condition the posterior distribution of the latent variables on both the visual images and the evoked brain activities. Furthermore, Güçlütürk *et al.* [43] combined probabilistic inference with the GAN idea and successfully reconstructed face images from evoked brain activities. Nevertheless, its two-stage design makes it difficult to converge to the global optima.

The proposed DGMM is motivated by the great success of deep generative models in image generation [39]–[41]. To the best of author knowledge, this is the first to introduce the Bayesian deep learning to neural decoding study.

III. Proposed Approach

Suppose that $\mathbf{X} \in \mathbb{R}^{D_x \times N}$ and $\mathbf{Y} \in \mathbb{R}^{D_y \times N}$ denote the visual images and the evoked fMRI activity patterns, respectively. Here, D_x and D_y denote the dimensions of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and N denotes the size of the training set. The training set consists of N paired samples, which can be denoted by $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ and $\mathbf{y}_i \in \mathbb{R}^{D_y}$ for $i = 1, \ldots, N$. For modeling the statistical relationships between the visual images and the evoked fMRI activity patterns, we develop a DGMM, where the two data views are assumed to be generated from a shared latent space via two view-specific generative models. The illustration of DGMM is shown in Fig. 1. Specifically, DGMM consists of a bottom-up inference model and two top-down generative models. In the inference model, a DNN is introduced to infer the shared latent variables $\mathbf{Z} \in \mathbb{R}^{D_z \times N}$ from $\mathbf{X}$ and $\mathbf{Y}$, where D_z is the dimension of $\mathbf{Z}$. Given the shared latent variables $\mathbf{Z}$, another DNN is adopted to generate the visual images, while a sparse linear model is adopted to generate the fMRI activity patterns. Overall, DGMM can be efficiently optimized within the autoencoding variational Bayes framework [31], [32].

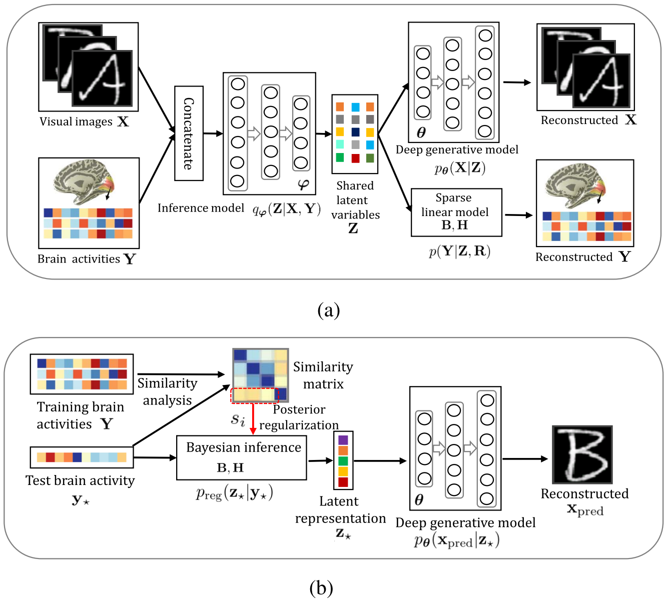

Visual images $\mathbf{X}$
Brain activities $\mathbf{Y}$
Concatenate
Inference model $q_{\varphi}(\mathbf{Z}|\mathbf{X},\mathbf{Y})$
φ
Shared latent variables $\mathbf{Z}$
$\boldsymbol{\theta}$
Deep generative model $p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})$
Sparse linear model $\mathbf{B}, \mathbf{H}$
$p(\mathbf{Y}|\mathbf{Z},\mathbf{R})$
Reconstructed $\mathbf{X}$
Reconstructed $\mathbf{Y}$
(a)
Training brain activities $\mathbf{Y}$
Similarity analysis
Similarity matrix
s_i
Posterior regularization
Test brain activity $\mathbf{y}_\star$
Bayesian inference $\mathbf{B}, \mathbf{H}$
$p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star)$
Latent representation $\mathbf{z}_\star$
$\boldsymbol{\theta}$
Deep generative model $p_{\boldsymbol{\theta}}(\mathbf{x}_{\text{pred}}|\mathbf{z}_\star)$
Reconstructed $\mathbf{x}_{\text{pred}}$
(b)

Fig. 1. Illustration of the proposed DGMM framework. (a) Training. $\mathbf{X}$ and $\mathbf{Y}$ are fed into the inference model to obtain $\mathbf{Z}$, which is used to reconstruct $\mathbf{X}$ and $\mathbf{Y}$ via different generative models. (b) Prediction. By using Bayesian inference, the testing brain activity $\mathbf{y}_\star$ is first decoded to the latent representation $\mathbf{z}_\star$. Given $\mathbf{z}_\star$, we can reconstruct the visual image $\mathbf{x}_{\text{pred}}$ through the pretrained deep generative model. In particular, we regularize the posterior inference of $\mathbf{z}_\star$ by utilizing the similarity information between the test instance $\mathbf{y}_\star$ and the training instances $\mathbf{Y}$.

In the prediction stage, the testing brain activity $\mathbf{y}_\star$ is first decoded to the latent representation $\mathbf{z}_\star$. Given $\mathbf{z}_\star$, we can reconstruct the visual image $\mathbf{x}_{\text{pred}}$ through the pretrained deep generative model. The details are described in the following. For the sake of readability, we list the frequently used symbols and their definitions in Table I.

A. Deep Generative Multiview Model

In deep generative model, we expect that each dimension of the latent variable contains its own semantic information independently, which makes the model more interpretable. Therefore, the prior distribution over the shared latent variables is assumed to be a product of isotropic Gaussian distributions

$$p(\mathbf{Z}) = \prod_{i=1}^{N} \mathcal{N}_{D_z}(\mathbf{z}_i|\mathbf{0}, \mathbf{I}) \tag{1}$$

which is consistent with previous studies in image generation [31], [32], [38].

Because the visual images $\mathbf{X}$ and the evoked fMRI activity patterns $\mathbf{Y}$ are assumed to be generated from the same $\mathbf{Z}$ via two view-specific generative models, we have two likelihood functions. One is for $\mathbf{X}$, and the other is for $\mathbf{Y}$.

1) Deep Generative Model for Perceived Images: We assume the image pixels follow a multivariate Gaussian distribution with zero-mean and diagonal covariance. Therefore, the likelihood function can be written as

$$p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}) = \prod_{i=1}^{N} \mathcal{N}_{D_x}\left(\mathbf{x}_i|\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}_i), \mathrm{diag}\left(\boldsymbol{\sigma}_{\mathbf{x}}^2(\mathbf{z}_i)\right)\right) \tag{2}$$

where $\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}_i)$ and $\boldsymbol{\sigma}_{\mathbf{x}}^2(\mathbf{z}_i)$ denote the mean and covariance, respectively. Note that $\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}_i)$ and $\boldsymbol{\sigma}_{\mathbf{x}}^2(\mathbf{z}_i)$ are obtained by

TABLE I

DEFINITION OF FREQUENTLY USED SYMBOLS

Symbol	Definition
N	Number of training instances
$\mathbf{X}$	Matrix of visual images
$\mathbf{Y}$	Matrix of fMRI activity patterns
$\mathbf{Z}$	Matrix of shared latent variables
$\mathbf{R}$	Matrix of auxiliary latent variables
$\mathbf{x}_i$	i-th visual image
$\mathbf{y}_i$	i-th fMRI activity pattern
$\mathbf{z}_i$	i-th shared latent variable
$\mathbf{r}_i$	i-th auxiliary latent variable
D_x	Dimension of visual image
D_y	Dimension of fMRI activity pattern
D_z	Dimension of shared latent variable
D_r	Dimension of auxiliary latent variable
$\boldsymbol{\theta}$	Parameters of deep *generative network*
$\boldsymbol{\varphi}$	Parameters of deep *inference network*
$\mathcal{N}_D(\cdot)$	D-dimensional Gaussian distribution
$\mathcal{W}(\cdot)$	Wishart distribution
$\mathcal{G}(\cdot\|\alpha,\beta)$	Gamma distribution with parameters α and β
$\mathbf{B}$	Projection matrix $\mathbf{B}\in\mathbb{R}^{D_z\times D_y}$
$\mathbf{H}$	Projection matrix $\mathbf{H}\in\mathbb{R}^{D_r\times D_y}$
$\mathbf{I}$	Identity matrix with corresponding shape
$\boldsymbol{\Psi}$	Full covariance matrix $\boldsymbol{\Psi}\in\mathbb{R}^{D_y\times D_y}$
$\boldsymbol{\tau}$	Precision variables in the prior distribution of $\mathbf{B}$
$\boldsymbol{\eta}$	Precision variables in the prior distribution of $\mathbf{H}$
γ	Precision variable in the likelihood of $\mathbf{Y}$
L	Number of Monte-Carlo sampling
ρ	Posterior regularization parameter
$\mathcal{R}(\cdot)$	Regularization term
Ξ	$\Xi=\{\boldsymbol{\tau},\boldsymbol{\eta},\gamma\}$
Θ	$\Theta=\{\mathbf{B},\mathbf{H},\mathbf{Z},\mathbf{R}\}$
$\mathcal{P}$	Space of probability distributions
$q^*(\cdot)$	Optimal variational distribution
$\\|\cdot\\|$	ℓ_2-norm operator
$\langle\cdot\rangle$	Expectation operator
k	Number of nearest neighbors
t	Free parameter in similarity measure
s_i	Similarity measure between instances $\mathbf{y}_i$ and $\mathbf{y}_\star$

different nonlinear transformations with respect to $\mathbf{z}_i$. In practice, we implement these nonlinear transformations using DNNs, which we refer to as the *generative network*. Parameters in the *generative network* are denoted by $\boldsymbol{\theta}$. Compared with the linear decoding model BCCA [18], our DNN architecture can capture the hierarchical visual features from images, which resembles the stages of human visual processing [14], [27], [28], [45]. Although the diagonal covariance structure does not take into account the relationships between image pixels, previous studies in image generation [31], [46], [47] have shown that it is a simple but effective assumption.[1]

2) Sparse Bayesian Linear Model for Brain Activities: Although nonlinear transformations are more powerful than linear transformations (in terms of the types of features they can learn), they have the risk of overfitting when we only have a small amount of high-dimensional fMRI data. Previous multivoxel pattern analysis studies have shown that linear models perform well in fMRI data analysis [48]. Therefore, we assume the distribution of the observed fMRI voxels has a linear form

$$p(\mathbf{Y}|\mathbf{Z}) = \prod_{i=1}^{N}\mathcal{N}_{D_y}(\mathbf{y}_i|\mathbf{B}^\top\mathbf{z}_i,\boldsymbol{\Psi}) \tag{3}$$

where $\mathbf{B}\in\mathbb{R}^{D_z\times D_y}$ is a projection matrix, and $\boldsymbol{\Psi}\in\mathbb{R}^{D_y\times D_y}$ is a full-covariance matrix. In general, the fMRI voxels are not independent but affecting each other. The correlations among fMRI voxels can naturally reflect the characteristics of corresponding visual stimuli [49]. The full-covariance matrix $\boldsymbol{\Psi}$ in (3) is expected to capture these correlations. Therefore, employing the full covariance will be benefit to suppress voxel noise and improve model performance. Compared with our method, most previous works simply used a spherical [18] or a diagonal [17] covariance, thus ignoring the correlations among fMRI voxels.

To avoid overfitting in the analysis of high-dimensional fMRI data, we impose a sparseness constraint on the projection matrix $\mathbf{B}$. With the sparseness constraint, the model is expected to automatically select a small number of voxels, which are most relevant to the decoding prediction. Naively, one can employ the automatic relevance determination (ARD) prior [50] and Wishart distribution for $\mathbf{B}$ and $\boldsymbol{\Psi}^{-1}$, respectively, that is,

$$\begin{aligned} p(\boldsymbol{\tau}) &= \prod_{j=1}^{D_y}\mathcal{G}(\tau_j|\alpha_\tau,\beta_\tau)\\ p(\mathbf{B}|\boldsymbol{\tau}) &= \prod_{j=1}^{D_y}\mathcal{N}_{D_z}\big(\mathbf{b}_j|\mathbf{0},\tau_j^{-1}\mathbf{I}\big)\\ p(\boldsymbol{\Psi}^{-1}) &= \mathcal{W}(\boldsymbol{\Psi}^{-1}|\mathbf{V},n_0) \end{aligned} \tag{4}$$

where $\mathcal{G}(\cdot|\alpha,\beta)$ denotes the gamma distribution with shape parameter α and rate parameter β. In addition, $\mathbf{V}$ and n_0 are the hyperparameters in Wishart distribution.

The complexity of inferring high-dimensional covariance matrix $\boldsymbol{\Psi}$ is $\mathcal{O}(D_y^3)$, thus resulting in severe computational issues in practice. To address the computational issues, we propose to introduce the following auxiliary latent variables $\mathbf{R}\in\mathbb{R}^{D_r\times N}$:

$$p(\mathbf{R}) = \prod_{i=1}^{N}\mathcal{N}_{D_r}(\mathbf{r}_i|\mathbf{0I}) \tag{5}$$

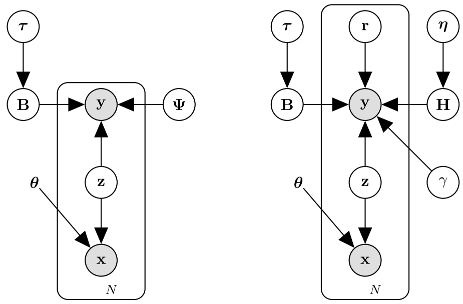
and rewrite the likelihood function in (3) as

$$p(\mathbf{Y}|\mathbf{Z},\mathbf{R}) = \prod_{i=1}^{N}\mathcal{N}_{D_y}(\mathbf{y}_i|\mathbf{B}^\top\mathbf{z}_i+\mathbf{H}^\top\mathbf{r}_i,\gamma^{-1}\mathbf{I}). \tag{6}$$

Similarly, we impose the ARD prior on the extra projection matrix $\mathbf{H}\in\mathbb{R}^{D_r\times D_y}$, and the gamma prior on the variance parameter γ, that is,

$$\begin{aligned} p(\boldsymbol{\eta}) &= \prod_{j=1}^{D_y}\mathcal{G}(\eta_j|\alpha_\eta,\beta_\eta)\\ p(\mathbf{H}|\boldsymbol{\eta}) &= \prod_{j=1}^{D_y}\mathcal{N}_{D_r}(\mathbf{h}_j|\mathbf{0},\eta_j^{-1}\mathbf{I})\\ p(\gamma) &= \mathcal{G}(\gamma|\alpha_\gamma,\beta_\gamma). \end{aligned} \tag{7}$$

[1] Actually, we can set a full-covariance structure for $p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})$, but this will increase the amount of computation dramatically. Therefore, the independence assumption for image pixels is just a tradeoff between precision and efficiency and not a limitation of our method.



(a) (b)

Fig. 2. Probabilistic graphical models of the proposed method. The gray nodes **x** and **y** denote observable variables. All other nodes are unobservable variables. (a) DGMM with the full-covariance matrix $\boldsymbol{\Psi}$. (b) DGMM with the low-rank assumption on $\boldsymbol{\Psi}$.

where D_r is the dimension of the auxiliary latent variables. The probabilistic graphical models of the proposed method are shown in Fig. 2. It can be shown that (6) is equivalent to imposing a low-rank assumption on the covariance matrix $\boldsymbol{\Psi}$ (i.e., $\boldsymbol{\Psi} = \mathbf{H}^\top\mathbf{H} + \gamma^{-1}\mathbf{I}$, see Appendix A for details). On the one hand, this low-rank assumption effectively reduces the computational complexity. On the other hand, the variations in the fMRI activity patterns are factorized into two parts. One part is the shared variables $\mathbf{Z}$, and the other part is the private variables $\mathbf{R}$ (e.g., the noisy component of brain activities). The ability to learn these two different types of latent variables makes our model more flexible than the existing methods. Note that we impose sparsity-inducing priors on $\mathbf{B}$ and $\mathbf{H}$ under the assumption that only a small number of fMRI voxels are relevant to the decoding task. As a result, we can effectively prune away the irrelevant projections in $\mathbf{B}$ and $\mathbf{H}$ by assigning suitable values to the hyperparameters $(\alpha_\tau, \beta_\tau)$ and $(\alpha_\eta, \beta_\eta)$.

In the following, $\boldsymbol{\Omega} = \{\alpha_\tau, \beta_\tau, \alpha_\eta, \beta_\eta, \alpha_\gamma, \beta_\gamma\}$ denotes the hyperparameters, $\Xi = \{\boldsymbol{\tau}, \boldsymbol{\eta}, \gamma\}$ denotes the priors, and $\Theta = \{\mathbf{B}, \mathbf{H}, \mathbf{Z}, \mathbf{R}\}$ denotes the remaining variables. For clarity, we have omitted the dependence on $\boldsymbol{\Omega}$ in the following equations. Then, the joint posterior distribution can be obtained by using Bayes' rule

$$p_{\boldsymbol{\theta}}(\Theta, \Xi|\mathbf{X}, \mathbf{Y}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})p(\mathbf{Y}|\mathbf{Z}, \mathbf{R})p(\Theta|\Xi)p(\Xi)}{p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})} \tag{8}$$

where $p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ is a normalization constant.

B. Optimization

Given the above-mentioned multiview generative model, the exact posterior inference is intractable. Alternatively, we first introduce an explicit inference model for the latent variables $\mathbf{Z}$, and then we devise an efficient mean-field variational inference method to optimize the proposed DGMM algorithm.

1) Explicit Inference Model for Latent Variables $\mathbf{Z}$*:* Due to the nonlinear architecture in the deep generative model of visual images, we cannot estimate the generative parameter $\boldsymbol{\theta}$ directly. Inspired by the autoencoding variational Bayes framework [31], [32], we define a fixed-form inference model $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$ for the latent variables $\mathbf{Z}$, and optimize the inference parameter $\boldsymbol{\varphi}$ as well as the generative parameter $\boldsymbol{\theta}$ jointly by maximizing the variational lower bound on the marginal likelihood. Specifically, we define $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$ as

$$q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{N} \mathcal{N}_{D_z}\big(\mathbf{z}_i|\boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}_i, \mathbf{y}_i),\ \mathrm{diag}\big(\boldsymbol{\sigma}^2_{\mathbf{z}}(\mathbf{x}_i, \mathbf{y}_i)\big)\big) \tag{9}$$

where the mean $\boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}_i, \mathbf{y}_i) = [\mu_{\mathbf{z}i1}, \ldots, \mu_{\mathbf{z}iD_z}]^\top$ and the covariance $\mathrm{diag}(\boldsymbol{\sigma}^2_{\mathbf{z}}(\mathbf{x}_i, \mathbf{y}_i)) = \mathrm{diag}(\sigma^2_{\mathbf{z}\,i1}, \ldots, \sigma^2_{\mathbf{z}\,iD_z})$ are the outputs of DNN with parameters $\boldsymbol{\varphi}$. We call this network the *inference network*. In practice, we regard the concatenation of $\mathbf{x}_i$ and $\mathbf{y}_i$ as the input of the *inference network*.

2) Objective Function: The mean-field variational inference method is based on two basic assumptions: 1) the joint variational distribution $q(\Theta, \Xi)$ is fully factorable, that is,

$$q(\Theta, \Xi) = q(\mathbf{B})q(\mathbf{H})q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})q(\mathbf{R})q(\boldsymbol{\tau})q(\boldsymbol{\eta})q(\gamma) \tag{10}$$

and 2) all factor distributions are free-form except for $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$. The objective function can obtained by minimizing the Kullback–Leibler (KL) divergence between the approximated posterior $q(\Theta, \Xi)$ and the target posterior $p_{\boldsymbol{\theta}}(\Theta, \Xi|\mathbf{X}, \mathbf{Y})$, that is,

$$\min_{\boldsymbol{\varphi},\ \boldsymbol{\theta},\ q(\Theta,\Xi)\in\mathcal{P}} \mathrm{KL}(q(\Theta, \Xi)\,\|\,p_{\boldsymbol{\theta}}(\Theta, \Xi|\mathbf{X}, \mathbf{Y})) \tag{11}$$

where $\mathcal{P}$ denotes the space of probability distributions. The above-mentioned objective function can be optimized via an iterative strategy. Specifically, we first appropriately initialize the moments of all factor distributions in $q(\Theta, \Xi)$. Then, we update each factor distribution, in turn, using the latest estimates of other factor distributions. Because the KL divergence is convex with respect to each factor distribution, the convergence can be expected after enough iterations.

3) Learning $\boldsymbol{\theta}$*,* $\boldsymbol{\varphi}$*, and the Optimal Distribution* $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$*:* We first fix the moments of $q(\mathbf{B}), q(\mathbf{H}), q(\mathbf{R}), q(\boldsymbol{\tau}), q(\boldsymbol{\eta})$, and $q(\gamma)$, and attempt to optimize $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$. It can be implemented by maximizing the variational lower bound on the marginal likelihood

$$\begin{aligned}
&\log p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})\\
&\geq \mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\left[\log\frac{p(\mathbf{Z})}{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})} + \log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}) + \log p(\mathbf{Y}|\mathbf{Z})\right]\\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})]}_{\text{likelihood term for visual images}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}[\log p(\mathbf{Y}|\mathbf{Z})]}_{\text{likelihood term for fMRI patterns}}\\
&\quad - \underbrace{\mathrm{KL}(q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})\,\|\,p(\mathbf{Z}))}_{\text{KL divergence term}}\\
&\equiv \mathcal{L}(\mathbf{X}, \mathbf{Y}).
\end{aligned} \tag{12}$$

Intuitively, the first likelihood term reflects the reconstruction error of the visual images and the second one reflects the reconstruction error of the fMRI activity patterns. Recent advances in variational training procedures such as the stochastic backpropagation [32] and the reparametrization trick [31] have made the optimization of DNN parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ feasible and efficient (see Appendix B for more details about the optimization of $\mathcal{L}(\mathbf{X}, \mathbf{Y})$).

4) Learning the Optimal Distributions of $\mathbf{B}, \mathbf{H}, \mathbf{R}$, *and* Ξ*:* In this section, we fix the moments of $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$, and optimize $q(\mathbf{B}), q(\mathbf{H}), q(\mathbf{R}), q(\boldsymbol{\tau}), q(\boldsymbol{\eta})$, and $q(\gamma)$ in turn. For any factor π (e.g., $\mathbf{B}$), it can be shown that the optimal variational distribution $q^*(\pi)$ satisfies

$$q^*(\pi) \propto \exp\{\mathbb{E}_{q(\{\Theta,\Xi\}\setminus\pi)}[\log p(\mathbf{Y}, \Theta, \Xi)]\}. \tag{13}$$

Using the conjugacy between the likelihood and the prior, we can iteratively perform fast Bayesian updating for each factor. The updating rules are given in the following.

Update $q(\mathbf{B})$ *and* $q(\mathbf{H})$*:* The updating rules for the projection parameters $\mathbf{B}$ and $\mathbf{H}$ can be found as

$$q^*(\mathbf{B}) = \prod_{j=1}^{D_y} \mathcal{N}_{D_z}(\mathbf{b}_j|\boldsymbol{\mu}_{\mathbf{b}_j}, [\langle\tau_j\rangle\mathbf{I} + \langle\gamma\rangle\langle\mathbf{Z}\mathbf{Z}^\top\rangle]^{-1}) \tag{14}$$

$$q^*(\mathbf{H}) = \prod_{j=1}^{D_y} \mathcal{N}_{D_r}(\mathbf{h}_j|\boldsymbol{\mu}_{\mathbf{h}_j}, [\langle\eta_j\rangle\mathbf{I} + \langle\gamma\rangle\langle\mathbf{R}\mathbf{R}^\top\rangle]^{-1}) \tag{15}$$

where $\langle\cdot\rangle$ denotes the expectation over its current optimal distribution, and

$$\boldsymbol{\mu}_{\mathbf{b}_j} = \boldsymbol{\Sigma}_{\mathbf{b}_j} \sum_{i=1}^{N} \langle\gamma\rangle\big(y_{ij} - \big\langle\mathbf{h}_j^\top\big\rangle\langle\mathbf{r}_i\rangle\big)\langle\mathbf{z}_i\rangle \tag{16}$$

$$\boldsymbol{\mu}_{\mathbf{h}_j} = \boldsymbol{\Sigma}_{\mathbf{h}_j} \sum_{i=1}^{N} \langle\gamma\rangle\big(y_{ij} - \big\langle\mathbf{b}_j^\top\big\rangle\langle\mathbf{z}_i\rangle\big)\langle\mathbf{r}_i\rangle \tag{17}$$

where $\boldsymbol{\Sigma}_{\mathbf{b}_j}$ and $\boldsymbol{\Sigma}_{\mathbf{h}_j}$ are the corresponding covariance in (14) and (15), respectively.

Update $q(\mathbf{R})$*:* Similarly, the updating rule for the auxiliary latent variables $\mathbf{R}$ can be found as

$$q^*(\mathbf{R}) = \prod_{i=1}^{N} \mathcal{N}_{D_r}(\mathbf{r}_i|\boldsymbol{\mu}_{\mathbf{r}_i}, [\mathbf{I} + \langle\gamma\rangle\langle\mathbf{H}\mathbf{H}^\top\rangle]^{-1}) \tag{18}$$

where $\boldsymbol{\mu}_{\mathbf{r}_i} = \boldsymbol{\Sigma}_{\mathbf{r}_i}\sum_{j=1}^{D_y}\langle\gamma\rangle(y_{ij} - \langle\mathbf{b}_j^\top\rangle\langle\mathbf{z}_i\rangle)\langle\mathbf{h}_j\rangle$, and $\boldsymbol{\Sigma}_{\mathbf{r}_i}$ is the corresponding covariance in (18).

Update $q(\boldsymbol{\tau})$*,* $q(\boldsymbol{\eta})$*, and* $q(\gamma)$*:* Finally, the updating rules for the precision parameters can be found as

$$q^*(\boldsymbol{\tau}) = \prod_{j=1}^{D_y} \mathcal{G}\left(\tau_j|\alpha_\tau + \frac{D_z}{2}, \beta_\tau + \frac{1}{2}\big\langle\mathbf{b}_j^\top\mathbf{b}_j\big\rangle\right)$$

$$q^*(\boldsymbol{\eta}) = \prod_{j=1}^{D_y} \mathcal{G}\left(\eta_j|\alpha_\eta + \frac{D_r}{2}, \beta_\eta + \frac{1}{2}\big\langle\mathbf{h}_j^\top\mathbf{h}_j\big\rangle\right)$$

$$q^*(\gamma) = \mathcal{G}\left(\gamma|\alpha_\gamma + \frac{ND_y}{2}, \beta_\gamma + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D_y}\delta_{ij}^2\right) \tag{19}$$

where $\delta_{ij} = y_{ij} - \langle\mathbf{b}_j^\top\rangle\langle\mathbf{z}_i\rangle - \langle\mathbf{h}_j^\top\rangle\langle\mathbf{r}_i\rangle$.

We sequentially update $\boldsymbol{\varphi}$, $\boldsymbol{\theta}$, $q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$, $q^*(\mathbf{B})$, $q^*(\mathbf{H})$, $q^*(\mathbf{R})$, $q^*(\boldsymbol{\tau})$, $q^*(\boldsymbol{\eta})$, and $q^*(\gamma)$ until convergence. The detailed training procedures are summarized in Algorithm 1.

Algorithm 1 DGMM

```
▶ Training
Input:
    • Visual images $\mathbf{X} \in \mathbb{R}^{D_x \times N}$
    • fMRI activity patterns $\mathbf{Y} \in \mathbb{R}^{D_y \times N}$
Initialize $\boldsymbol{\varphi}$, $\boldsymbol{\theta}$ and the moments of all random variables $\mathbf{Z}, \mathbf{R}, \mathbf{B}, \mathbf{H}$ and $\Xi$
for number of training iterations do
    Update $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$ and the moments of $\mathbf{Z}$ by maximizing the variational lower bound $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ (Eq. (12)).
        $\max_{\boldsymbol{\theta},\boldsymbol{\varphi}} \mathcal{L}(\mathbf{X}, \mathbf{Y})$
    Update the moments of $\mathbf{B}$ by using Eq. (14)
    Update the moments of $\mathbf{H}$ by using Eq. (15)
    Update the moments of $\mathbf{R}$ by using Eq. (18)
    Update the moments of $\Xi$: by using Eq. (19)
end for
Output:
    • Inference parameters: $\boldsymbol{\varphi}$
    • Generative parameters for visual image: $\boldsymbol{\theta}$
    • Generative parameters for fMRI activity pattern: $\mathbf{B}, \mathbf{H}$ and $\Xi$
▶ Prediction
Input: fMRI activity pattern $\mathbf{y}_\star$ (not available in training)
Draw L samples $\{\mathbf{z}_\star^{(l)}\}_{l=1}^{L}$ from $p_{\mathrm{reg}}(\mathbf{z}_\star|\mathbf{y}_\star)$ according to Eq. (28).
Reconstruct the visual image by $\mathbf{x}_{\mathrm{pred}} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}_\star^{(l)})$
Output: The reconstructed image $\mathbf{x}_{\mathrm{pred}}$
```

C. Prediction

Given a new fMRI pattern $\mathbf{y}_\star$ (not available in training), we can derive a predictive distribution $p(\mathbf{x}_{\mathrm{pred}}|\mathbf{y}_\star)$ for the perceived image

$$p(\mathbf{x}_{\mathrm{pred}}|\mathbf{y}_\star) = \int p_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{pred}}|\mathbf{z}_\star)p(\mathbf{z}_\star|\mathbf{y}_\star)d\mathbf{z}_\star \tag{20}$$

where $p(\mathbf{z}_\star|\mathbf{y}_\star)$ is the unnormalized posterior distribution, which can be derived by

$$p(\mathbf{z}_\star|\mathbf{y}_\star) = \int p(\mathbf{y}_\star|\mathbf{z}_\star, \mathbf{r}_\star, \mathbf{B}, \mathbf{H}, \gamma)p(\mathbf{z}_\star)p(\mathbf{r}_\star) \times q^*(\mathbf{B})q^*(\mathbf{H})q^*(\gamma)d\mathbf{r}_\star d\mathbf{B}d\mathbf{H}d\gamma. \tag{21}$$

In (21), we simply infer the latent representation $\mathbf{z}_\star$ from the given fMRI pattern $\mathbf{y}_\star$. Actually, we can further explore the similarity information between $\mathbf{y}_\star$ and the training data $\mathbf{Y}$ to obtain more expressive $\mathbf{z}_\star$. In the following, we implement this by using a posterior regularization technique [33].

1) Posterior Regularization: Posterior regularization [33] is a technique for regularizing models by encoding specific prior knowledge into model posteriors. Compared with the specially designed priors, the posterior regularization technique can more naturally integrate the prior knowledge into the Bayesian model. The purpose of using the posterior regularization is to restrict the space of $p(\mathbf{z}_\star|\mathbf{y}_\star)$ so that the image reconstruction results are more accurate.

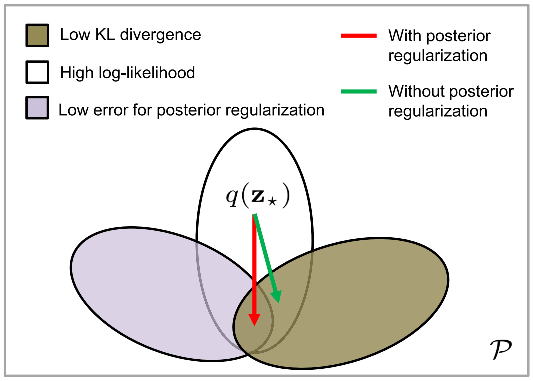

▨ Low KL divergence
□ High log-likelihood
▨ Low error for posterior regularization
With posterior regularization
Without posterior regularization
$q(\mathbf{z}_\star)$
$\mathcal{P}$

Fig. 3. Trajectories of posterior $q(\mathbf{z}_\star)$ are illustrated in a schematic probability space $\mathcal{P}$. Green line: trajectory of standard Bayesian inference without the posterior regularization. Red line: trajectory of Bayesian inference with the posterior regularization. By incorporating data-dependent constraints into the Bayesian inference, our posterior regularization strategy guides the model posterior toward the desired probability space.

Given (21), we can equivalently obtain $p(\mathbf{z}_\star|\mathbf{y}_\star)$ by minimizing the following KL divergence:

$$\min_{q(\mathbf{z}_\star)\in\mathcal{P}} \mathrm{KL}(q(\mathbf{z}_\star)\|p(\mathbf{z}_\star|\mathbf{y}_\star)). \tag{22}$$

Expanding the KL divergence, (22) can be rewritten as

$$\min_{q(\mathbf{z}_\star)\in\mathcal{P}} \mathrm{KL}\left(q(\mathbf{z}_\star)\|p(\mathbf{z}_\star)\right) - \mathbb{E}_{q(\mathbf{z}_\star)}[\log p(\mathbf{y}_\star|\mathbf{z}_\star)]. \tag{23}$$

Then, we add a regularization term to (23)

$$\min_{q(\mathbf{z}_\star)\in\mathcal{P}} \underbrace{\mathrm{KL}\left(q(\mathbf{z}_\star)\|p(\mathbf{z}_\star)\right)}_{\text{KL divergence term}} - \underbrace{\mathbb{E}_{q(\mathbf{z}_\star)}[\log p(\mathbf{y}_\star|\mathbf{z}_\star)]}_{\text{likelihood term}} + \rho \underbrace{\mathcal{R}(q(\mathbf{z}_\star))}_{\text{regularization term}} \tag{24}$$

where $\mathcal{R}(q(\mathbf{z}_\star))$ is the introduced regularization term, and the parameter $\rho > 0$ controls the expected scale. As illustrated in Fig. 3, the regularization term can guide the model posterior toward the desired probability space. Specifically, we define

$$\mathcal{R}(q(\mathbf{z}_\star)) = \mathbb{E}_{q(\mathbf{z}_\star)}\left[\sum_{i=1}^{N} s_i\|\mathbf{z}_\star - \mathbf{z}_i\|^2\right] \tag{25}$$

where s_i is a similarity measure between $\mathbf{y}_i$ and $\mathbf{y}_\star$, and $\|\cdot\|$ denotes the ℓ_2-norm. Here, we further define

$$s_i = \begin{cases} \exp\left(-\dfrac{\|\mathbf{y}_\star - \mathbf{y}_i\|^2}{2t^2}\right), & \mathbf{y}_i \in \mathcal{K}(\mathbf{y}_\star) \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{K}(\mathbf{y}_\star)$ denotes the k-nearest neighbors of $\mathbf{y}_\star$ and t is a free parameter. In other word, s_i is calculated by a radial basis function kernel function when $\mathbf{y}_i$ is in the k-nearest neighbor space of $\mathbf{y}_\star$. Intuitively, s_i can capture the local geometry structure of the fMRI pattern space. As a result, the regularization term will force $\mathbf{z}_\star$ to be close to their nearest neighbors from the training set.

Let $h(\mathbf{z}_\star|\rho,\mathbf{s}) = \exp\{-\rho\sum_{i=1}^{N} s_i\|\mathbf{z}_\star - \mathbf{z}_i\|^2\}$, then (24) can be rewritten as

$$\min_{q(\mathbf{z}_\star)\in\mathcal{P}} \mathrm{KL}\left(q(\mathbf{z}_\star)\|p(\mathbf{z}_\star)\right) - \mathbb{E}_{q(\mathbf{z}_\star)}[\log p(\mathbf{y}_\star|\mathbf{z}_\star)] - \mathbb{E}_{q(\mathbf{z}_\star)}[\log h(\mathbf{z}_\star|\rho,\mathbf{s})] \tag{26}$$

and hence the regularized posterior distribution satisfies

$$p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star) = \int p(\mathbf{y}_\star|\mathbf{z}_\star,\mathbf{r}_\star,\mathbf{B},\mathbf{H},\gamma)p(\mathbf{z}_\star)h(\mathbf{z}_\star|\rho,\mathbf{s})p(\mathbf{r}_\star) \times q^*(\mathbf{B})q^*(\mathbf{H})q^*(\gamma)d\mathbf{r}_\star d\mathbf{B}d\mathbf{H}d\gamma. \tag{27}$$

Equation (27) is intractable due to the multiple integral with respect to $\mathbf{r}_\star$, $\mathbf{B}$, $\mathbf{H}$, and γ. To address this problem, we replace the random variables $\mathbf{B}$, $\mathbf{H}$ and γ with the mean of $q^*(\mathbf{B})$, $q^*(\mathbf{H})$, and $q^*(\gamma)$, respectively. Now, we have

$$p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star) = \int p(\mathbf{y}_\star|\mathbf{z}_\star,\mathbf{r}_\star)p(\mathbf{z}_\star)h(\mathbf{z}_\star|\rho,\mathbf{s})p(\mathbf{r}_\star)d\mathbf{r}_\star \sim \mathcal{N}_{D_z}\left(\mathbf{z}_\star|\boldsymbol{\mu}_{\mathbf{z}_\star}\boldsymbol{\Sigma}_{\mathbf{z}_\star}\right) \tag{28}$$

where $\boldsymbol{\mu}_{\mathbf{z}_\star}$ and $\boldsymbol{\Sigma}_{\mathbf{z}_\star}$ can be find in Appendix C.

2) Predictive Distribution: Substituting (28) into (20) yields an expression for the predictive distribution

$$p(\mathbf{x}_{\text{pred}}|\mathbf{y}_\star) = \int p_{\boldsymbol{\theta}}(\mathbf{x}_{\text{pred}}|\mathbf{z}_\star)p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star)d\mathbf{z}_\star. \tag{29}$$

Because the likelihood function of visual image $p_{\boldsymbol{\theta}}(\mathbf{x}_{\text{pred}}|\mathbf{z}_\star)$ is characterized by the complex DNN, the integral with respect to $\mathbf{z}_\star$ cannot be solved analytically. Alternatively, we use Monte Carlo sampling method to approximately solve this integral. Specifically, we assume $\mathbf{z}_\star^{(l)}$ is the lth sample randomly sampled from $p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star)$, and $\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}_\star^{(l)})$ is the output of the pretrained *generative network* characterized by $p_{\boldsymbol{\theta}}(\mathbf{x}_{\text{pred}}|\mathbf{z}_\star)$, where $l \in \{1, 2, \ldots, L\}$. Then, the reconstructed visual image can be obtained by taking the average of all L predictions, that is,

$$\mathbf{x}_{\text{pred}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{x}_{\text{pred}}^{(l)} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}_\star^{(l)}). \tag{30}$$

The prediction procedures are summarized in Algorithm 1.

IV. Experiments

A. Experimental Setup

1) Data Sets: We conduct the experiments on three publicly available fMRI data sets, and we briefly describe them in the following.

1) *Binary Contrast Patterns*[2] *[15]:* The visual images in this data set can be divided into two types. One type is random images, which are used in the training phase. The other type is the figure images including geometric shapes and alphabetical letters, which are used for testing (see Fig. 4). The resolution of the visual images is 10 × 10. The corresponding fMRI data are given for each visual image. In the experiments, we use voxels from the V1 area of subject 1 (S1).

[2] Data are available at http://brainliner.jp/data/brainliner

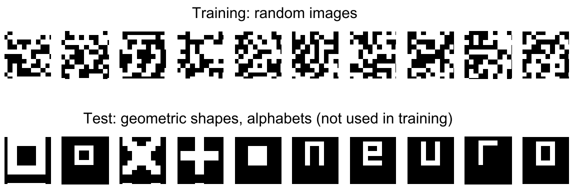


Fig. 4. Examples of the binary contrast patterns used in training and test.

TABLE II

PROPERTIES OF THE DATA SETS USED IN THE EXPERIMENTS

Datasets	Instances	Pixels	Voxels	ROIs	Training
Binary contrast patterns	1400	100	797	V1	1320
Handwritten digits	100	784	3092	V1, V2, V3	90
Handwritten characters	360	784	2420	V1, V2	330

2) *Handwritten Digits*[3] *[16]:* This data set contains 100 gray-scale handwritten digit images (equal number of 6's and 9's). The image resolution is 28 × 28. The corresponding fMRI data contain voxels from the V1, V2, and V3 areas.

3) *Handwritten Characters*[4] *[17]:* This data set contains 360 gray-scale handwritten character images (equal number of B's, R's, A's, I's, N's, and S's) taken from [51]. The image resolution is 56 × 56. In the experiments, the visual images are downsampled to 28 × 28. The corresponding fMRI data contain voxels from the V1 and V2 areas of all three subjects.

More descriptions about these three fMRI data sets can be found in the original publications [15]–[17]. We briefly summarize their main properties in Table II.

2) Compared Methods: The following methods are used as baselines.

1) *Miyawaki*[5]*:* A specially designed multiscale image reconstruction method proposed by Miyawaki *et al.* [15]. A fatal limitation of this method is that the shapes of image bases are fixed, thereby losing the flexibility.

2) *BCCA*[6]*:* A multiview linear generative model designed for neural encoding and decoding [18]. However, its linear architecture and spherical covariance assumption may greatly limit its performance.

3) *Deep canonically correlated autoencoder (DCCAE)*[7]*:* DCCAE was originally proposed to learn the deep representations from multiview data [34]. It consists of two basic autoencoders. The objective function of DCCAE is a combination of two components. The first component is the canonical correlation between the two learned bottleneck representations and the second one is the reconstruction errors of both autoencoders. DCCAE can be applied to cross-view reconstructions as well as neural encoding and decoding. However, DCCAE only considers the intraview reconstruction errors and the correlation between the bottleneck representations while it ignores the interview reconstruction errors.

4) *Deconvolutional Neural Network:* It is a two-stage cascade neural decoding method [29]. It first decodes the fMRI activity pattern to the high-level feature maps using the multivariate linear regression [52]. Then, the predicted high-level feature maps are fed into the pretrained deconvolution neural network, whose outputs are the expected reconstructions. Such an approach is limited in that it requires two stages and each stage is optimized separately, making the whole decoding pipeline often suboptimal.

3) Parameter Setting: To obtain sparse projection matrices $\mathbf{B}$ and $\mathbf{H}$, we set the hyperparameters $(\alpha_\tau, \beta_\tau) = (\alpha_\eta, \beta_\eta) = (10^{-10}, 10^{-10})$ and $(\alpha_\gamma, \beta_\gamma) = (1, 1)$, which are consistent with the previous sparse Bayesian method [53]. Once the hyperparameters are fixed, the model parameters can be learned automatically on the specific data set. Furthermore, the regularization parameter ρ should be tuned separately on different data sets. Specifically, we conducted fivefold cross-validation on the training sets to choose the best ρ from $[0.05, 0.1, 0.5, 1, 5]$, where DGMM[8] achieves the experimental results. We empirically set the free parameter $t = 10$ and the nearest neighbor parameter $k = 10$ in constructing the nearest neighbor graphs. This setting can ensure the values of s_i are evenly distributed between 0 and 1, which is useful for selecting the top-k nearest neighbors. Finally, we empirically find that setting $L = 100$ is enough to get clear reconstructions. In addition, we consider multiple layer perceptrons (MLPs) as the type of the *generative network* and *inference network*. Specifically, the architecture of the *inference network* was set to "100-200," "784-256-128-10," and "784-256-128-6" on the three data sets, respectively. The *generative network* has a symmetrical architecture with the *inference network*. In particular, we consider two different cases for DCCAE: 1) DCCAE-A has an asymmetric architecture (DNN for visual images, while the single-layer neural network for fMRI) and 2) DCCAE-S has a symmetric architecture (i.e., DNNs for both data views), which can explore the deep transformations of fMRI activity patterns. The Adam optimizer [54] with learning rate 0.0003 is utilized for the training of all DNN-based models.

4) fMRI Voxel Selection: The fMRI activity patterns are high-dimensional data, in which a lot of voxels may not respond to the visual stimuli. Therefore, it is necessary to remove the unrelated voxels before the decoding experiments. First, we randomly divide the training set into 10-folds, where 9-folds are used to train the proposed DGMM, and the rest 1-fold is used for evaluation. Then, we use the coefficient of determination (R^2) as a metric to evaluate the goodness of fit between measured activations and model predictions for each voxel. The final result of R^2 is an average of 10 runs with different data splits. We select the voxels with positive R^2 from all voxels for downstream decoding study.

[3] Data are available at http://www.ccnlab.net/data/

[4] Data are available at http://sciencesanne.com/research/

[5] Code are available at http://brainliner.jp/data/brainliner

[6] Code are available at https://github.com/KamitaniLab/VBCCA

[7] Code are available at http://ttic.uchicago.edu/%7Ewwang5/dccae.html

[8] Code are available at https://github.com/ChangdeDu/DGMM

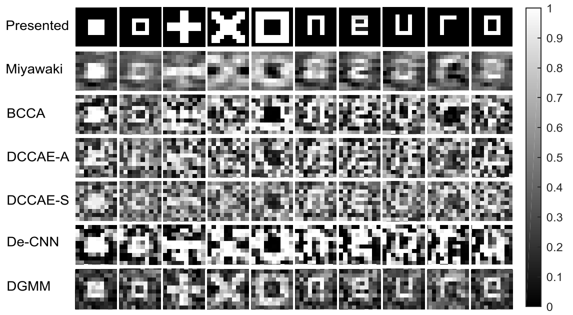


Fig. 5. Image reconstructions of distinct binary contrast patterns.

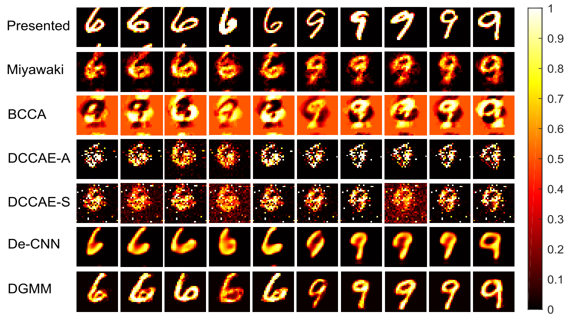


Fig. 6. Image reconstructions of distinct handwritten digits.

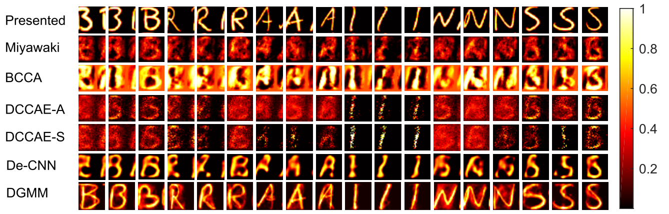


Fig. 7. Image reconstructions of distinct handwritten characters taken from subject 3.

B. Experimental Results

1) Qualitative Analysis: The reconstructed results on three different data sets are shown in Figs. 5–7, respectively. In each figure, the top row shows the presented visual images, while the following rows show the reconstructed results of all compared methods.

As can be seen from Figs. 5–7, DGMM produces better reconstructions than the compared methods, especially on the handwritten digits and characters data sets. In Fig. 5, the reconstructed images can roughly capture the presented shapes. In Figs. 6 and 7, the reconstructed handwritten digits and characters are very similar to the original images. The subtle differences between the presented images and the reconstructed ones may be caused by the posterior regularization. Compared with our DGMM, the performances of Miyawaki and BCCA are coarse on all three data sets. Their reconstructions are often polluted by noises. Furthermore, the results of DCCAE-A and DCCAE-S are disappointing too. Their reconstructions not only have a lot of noises but also lack the basic features of the original images. This may be caused by the fact that the nonlinear transformations of fMRI data are prone to overfitting. Although deconvolutional neural network obtains the comparable results, there are slight ambiguity and distortion in its reconstructions. Because the reconstructions produced by our DGMM are based on Monte Carlo sampling [see (30)], the noises in the results could be reduced through the averaging operation.

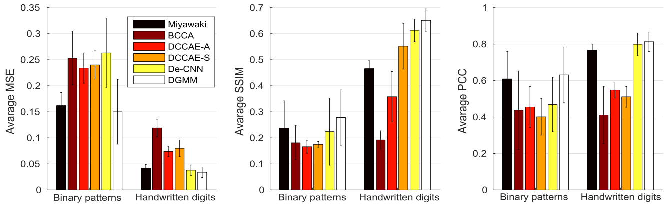


Fig. 8. Quantitative comparisons between DGMM and the baselines on the binary contrast patterns and handwritten digit data sets. Results are averaged over 20 random seeds. Error bars represent standard error.

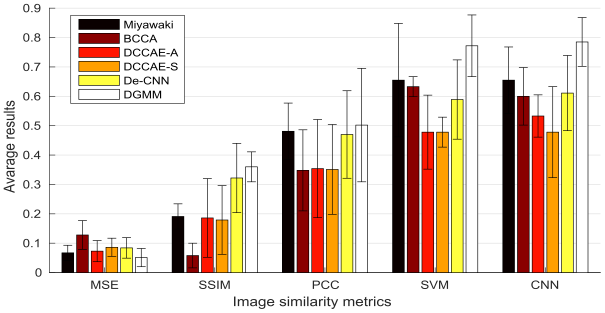


Fig. 9. Quantitative comparisons between DGMM and the baselines on the handwritten characters data set. Results are averaged across three subjects. Error bars represent standard error.

2) Quantitative Analysis: For supporting quantitative evaluation, we use the following metrics to evaluate the similarity between the presented images and the reconstructed ones.

1) Pearson's correlation coefficient.
2) Mean squared error.
3) *Structural Similarity Index (SSIM):* The SSIM [55] correlates well with human visual perception due to taking image texture into account. Its values are between [0, 1], higher is better.
4) *Accuracy (ACC)–Support Vector Machine (SVM)/ Convolutional Neural Network (CNN):* We train the linear SVM and CNN on the presented visual images. The pretrained models are regarded as gold-standard classifiers to classify the reconstructed images. The labels of test data are used as ground truth to calculate the classification ACC.

The quantitative comparisons between DGMM and other methods are shown in Figs. 8 and 9. From them, we can find that DGMM consistently outperforms the baselines. For example, our SSIM remarkably surpass the baselines on all three data sets. Furthermore, the performance of BCCA is greatly limited by its linear architecture and spherical covariance assumption. The low performance of DCCAE-A and DCCAE-S may be partly due to the fact that DCCAE method only considers the intraview reconstruction errors and the correlation between bottleneck representations, thus ignoring the interview reconstruction errors. In addition, the deep

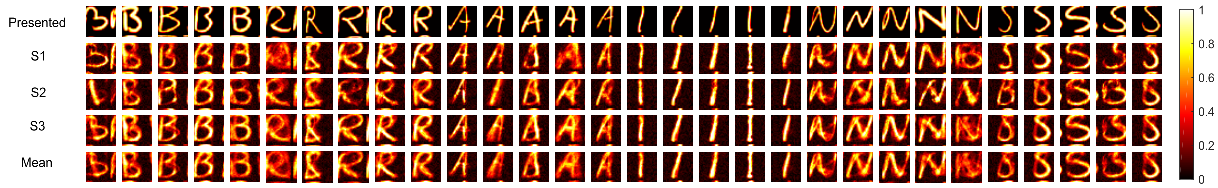
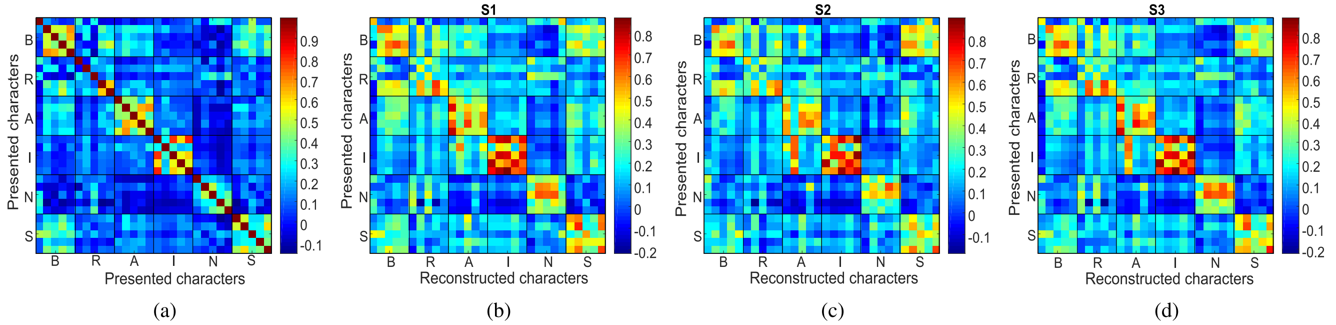


Fig. 10. Reconstructions produced by DGMM for all three subjects. Reconstruction procedures are independently conducted for S1, S2, and S3. The mean of three subjects' reconstructions about the same character is shown in the bottom row.



(a)

(b)

(c)

(d)

Fig. 11. Correlation matrices for all three subjects (S1, S2, and S3) computed on the test set containing 30 characters (6 letter classes, 5 samples in each class). (a) Elements indicate the self-correlations between the presented visual images. (b)–(d) Elements indicate the correlations between the presented images and the reconstructed ones for S1, S2, and S3, respectively. Dark lines are used to separate different letter classes.

transformations of fMRI data make DCCAE-S easy to overfitting. Compared with DGMM, De-CNN cannot be trained in an end-to-end manner. Therefore, its performance is moderate. On the other hand, we see that DGMM achieves remarkably higher classification ACC on the handwritten characters data set. This evaluation metric demonstrates the superiority of the proposed method again.

3) Individual Differences Between Subjects: By using DGMM, the reconstructed handwritten characters for all three subjects are shown in Fig. 10, where the top row is the presented visual stimuli, and the following rows are the reconstructed images obtained from the individual subject and the mean results averaged across all three subjects.

In most cases, the obtained reconstructions are unique. Also, the reconstructions obtained from different subjects show subtle differences. The reason is that different people have different brain responses, even for an identical stimulus. Occasionally, some of the reconstructions are not accurate. We attribute this to the fact that the synchronization between presented stimuli and observed brain signals is not so good due to data processing problems. Overall, the results of DGMM are of high quality.

The correlation between the presented image and the reconstructed one can provide a quantitative measure of the quality of individual reconstructions. Here, we use the difference between three correlation matrices as a metric to evaluate the individual differences between the three subjects. For each subject, we calculate the correlation matrix between the presented images and the reconstructed ones on the test set containing 30 characters (6 letter classes, 5 samples in each class), and illustrated in Fig. 11. Intuitively, elements in Fig. 11(a) indicate the self-correlations between the presented visual images, while elements in Fig. 11(b)–(d) indicate the correlations between the presented image and reconstructed one for S1, S2, and S3, respectively. Fig. 11(a) can be regarded as a ceiling of the reconstruction performance of S1, S2, and S3. The obvious difference between Fig. 11(b)–(d) indicates the individual difference between the three subjects. For example, S1 and S3 are better at reconstructing "N" than S2. The block diagonal structures of Fig. 11(b)–(d) indicate that our algorithm has successfully decoded the category information of brain signals. Note that the letter classes "B" and "S" are easily confused, because their appearance difference is relatively small.

C. Interpretation of Latent Representations

Automatically learn the basic visual concepts is important to visual image reconstruction. Without disentangled and symbolic visual concepts, it is difficult to interpret or reuse representations across tasks as no single component of the representation vector has a semantic meaning by itself. In order to find disentangled factors in the latent space corresponding to specific semantic concepts or visual features in the image space [30], [56], we inspect the latent space of a trained DGMM model. The idea is to reconstruct the visual images using each dimension of the latent representation. Fig. 12 depicts 18 distinct reconstructions taken from the testing set of S3. Though not perfect, we clearly see that each dimension of the latent representation captures some semantic concepts in the image space such as "R," "N," and so on. We also see that each dimension roughly captures two kinds of visual features, which are the basic components of the original visual stimuli.

D. Estimated fMRI Voxel Weights

Fig. 13 illustrates the distribution of fMRI voxels selected from S3. Obviously, not all voxels are needed in image

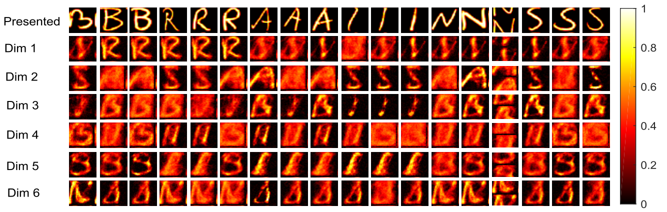


Fig. 12. Reconstructions (taken from S3) by using only one dimension of the learned latent representation. The top row is the original visual stimuli, and the following rows are the reconstructed results by each dimension of the latent representation inferred from the testing brain response.

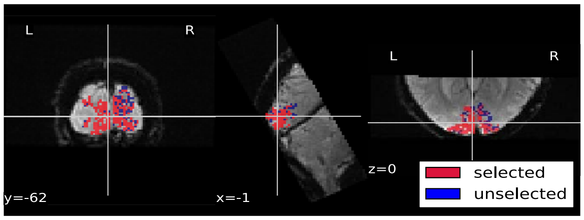


Fig. 13. Distribution of fMRI voxels selected from S3. Only the selected voxels were involved in reconstruction.

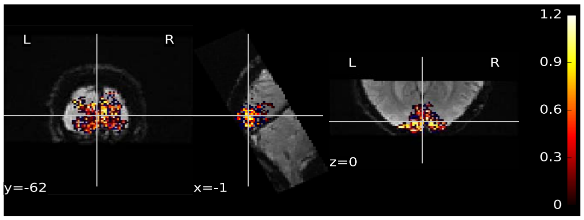


Fig. 14. Estimated fMRI voxel weights for S3. Higher voxel weight means higher contribution to image reconstruction.

reconstruction. To further investigate the importance of each selected voxel in image reconstruction, the estimated voxel weights were mapped to the cortical surface and shown in Fig. 14. Specifically, the absolute values of the estimated voxel weights were projected on the visual areas V1 and V2. It seems that the most important voxels come from area V1 rather than V2, and the voxels with high weights tend to be spatially localized in V1 and V2. These results suggest that the sparsity-inducing priors imposed on the projection matrices play an important role in estimating physiologically meaningful fMRI voxels.

E. Full Versus Diagonal Covariance

Comparing the proposed DGMM to the previous approaches [17], [18], one core difference is the full-covariance matrix, which is introduced to capture the correlations among voxels. Here, we investigate the effects of utilizing spatial voxel covariance in the perceived image reconstruction. To implement our DGMM with a diagonal covariance, we ignored the terms with respect to the projection matrix $\mathbf{H}$, the precision variable $\boldsymbol{\eta}$, and the auxiliary latent variable $\mathbf{r}$ in the generative and inference procedures. Handwritten digits and characters reconstructed by DGMM with a full/diagonal covariance were shown in Fig. 15.

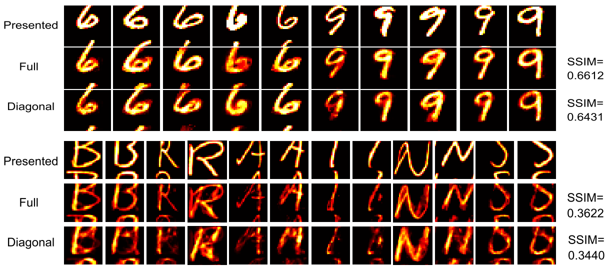


Fig. 15. Reconstructions produced by our DGMM with a full or diagonal covariance. The handwritten characters were taken from S3.

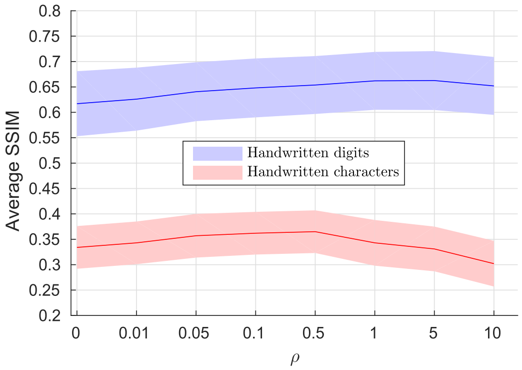


Fig. 16. Average SSIM of reconstructed handwritten digits (blue) and characters (red) with different regularization parameters ρ.

The average SSIM values computed under the full/diagonal covariance cases were shown on the right side of the figure. In most cases, we can observe that DGMM with a full covariance produced better results, which indicate that our DGMM approach has benefitted from taking into account the correlations among voxels.

F. Effects of Posterior Regularization Strategy

Another important feature of the proposed DGMM approach is its ability to exploit the similarity information between different brain activities via the posterior regularization strategy. If two brain activities are similar, then the posterior regularization will make their reconstructions similar too. We conducted image reconstruction experiments with different regularization parameters ρ, and displayed the results in Fig. 16. From the figure, we can observe that the best regularization parameter ρ can be chosen from [0.05, 0.1, 0.5, 1, 5], where DGMM achieves good results.

V. Conclusion

We presented a DGMM for perceived image reconstruction from human brain activities. DGMM models the statistical relationships between the visual image and the evoked fMRI by using two view-specific generators with a shared latent space. On the one hand, we adopted a DNN architecture for visual image generation. On the other hand, we designed a sparse Bayesian linear model for brain activity generation. Our model disentangles the factors of variation in the latent space, corresponding to specific semantic concepts or visual features in the image space. Furthermore, the sparsity-inducing priors imposed on the projection matrices facilitated the selection of the meaningful voxels, and the full-covariance matrix we adopted is benefit to capture the correlations among fMRI voxels. Finally, our posterior regularization strategy incorporated the similarity between different brain activities into the Bayesian inference of latent variables, which also improved the reconstruction ACC.

Although the power of our framework has been verified in this paper, there are some promising future directions.

1) Given the image categories (or attributes), each stimulus-response pair has a label indicator vector. This vector captures the high-level semantic information about the visual image. Therefore, creating an auxiliary generation pathway for the semantic view allows us to decode the brain activity patterns into a low-level pixel space and a high-level semantic space simultaneously [57]–[59].
2) Human's visual experiences are often dynamic, how to reconstruct the dynamic visual experiences from the measured human brain activities is a more challenging task [6], [60]. Under the proposed framework, replacing the MLPs by the recurrent neural networks [61] allows us to explore the reconstruction of the dynamic visual scenes.

Appendix A
Low-Rank Assumption on $\boldsymbol{\Psi}$

According to the definitions in (5) and (6), we have $p(\mathbf{r}) = \mathcal{N}_{D_r}(\mathbf{r}|\mathbf{0},\mathbf{I}) \propto \exp[-(1/2)\mathbf{r}^\top\mathbf{r}]$ and

$$\begin{aligned}p(\mathbf{y}|\mathbf{z},\mathbf{r}) &= \mathcal{N}_{D_y}(\mathbf{y}|\mathbf{B}^\top\mathbf{z}+\mathbf{H}^\top\mathbf{r},\gamma^{-1}\mathbf{I})\\ &\propto \exp\bigg\{-\frac{\gamma}{2}[\mathbf{r}^\top\mathbf{H}\mathbf{H}^\top\mathbf{r}-2\mathbf{r}^\top\mathbf{H}(\mathbf{y}-\mathbf{B}^\top\mathbf{z})\\ &\qquad +(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}.\end{aligned}$$

Then, $\int p(\mathbf{y}|\mathbf{z},\mathbf{r})p(\mathbf{r})\,d\mathbf{r}$ can be computed by

$$\begin{aligned}&\int p(\mathbf{y}|\mathbf{z},\mathbf{r})p(\mathbf{r})\,d\mathbf{r}\\ &\propto \int \exp\bigg\{-\frac{\gamma}{2}[\mathbf{r}^\top\mathbf{H}\mathbf{H}^\top\mathbf{r}-2\mathbf{r}^\top\mathbf{H}(\mathbf{y}-\mathbf{B}^\top\mathbf{z})\\ &\qquad +(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}\exp\bigg[-\frac{1}{2}\mathbf{r}^\top\mathbf{r}\bigg]d\mathbf{r}\\ &\propto \int \exp\bigg\{-\frac{1}{2}\bigg[\mathbf{r}^\top\underbrace{(\gamma\mathbf{H}\mathbf{H}^\top+\mathbf{I})}_{U}\mathbf{r}-2\mathbf{r}^\top\underbrace{\gamma\mathbf{H}(\mathbf{y}-\mathbf{B}^\top\mathbf{z})}_{V}\\ &\qquad +\gamma(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})\bigg]\bigg\}d\mathbf{r}\\ &\propto \int \exp\bigg\{-\frac{1}{2}[\mathbf{r}^\top U\mathbf{r}-2\mathbf{r}^\top V+V^\top U^{-1}V-V^\top U^{-1}V\\ &\qquad +\gamma(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}d\mathbf{r}\\ &\propto \underbrace{\int \exp\bigg\{-\frac{1}{2}[(\mathbf{r}-U^{-1}V)^\top U(\mathbf{r}-U^{-1}V)]\bigg\}d\mathbf{r}}_{1}\end{aligned}$$

$$\begin{aligned}&\cdot\exp\bigg\{-\frac{1}{2}[-V^\top U^{-1}V+\gamma(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}\\ &\propto \exp\bigg\{-\frac{1}{2}[-V^\top U^{-1}V+\gamma(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}\\ &\propto \exp\bigg\{-\frac{1}{2}[-[\gamma\mathbf{H}(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]^\top U^{-1}[\gamma\mathbf{H}(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\\ &\qquad +\gamma(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}\\ &\propto \exp\bigg\{-\frac{1}{2}[(\mathbf{y}-\mathbf{B}^\top\mathbf{z})^\top\underbrace{(-\gamma^2\mathbf{H}^\top U^{-1}\mathbf{H}+\gamma\mathbf{I})}_{\boldsymbol{\Psi}^{-1}}(\mathbf{y}-\mathbf{B}^\top\mathbf{z})]\bigg\}\\ &\sim \mathcal{N}_{D_y}(\mathbf{y}|\mathbf{B}^\top\mathbf{z},\boldsymbol{\Psi})\\ &= p(\mathbf{y}|\mathbf{z}).\end{aligned}$$

We have proved the model in (3) is equivalent to the model in (5) and (6). In the following, we prove that $\boldsymbol{\Psi} = \mathbf{H}^\top\mathbf{H}+\gamma^{-1}\mathbf{I}$.

We first approximate U^{-1} by using Taylor expansion

$$\begin{aligned}U^{-1} &= (\gamma\mathbf{H}\mathbf{H}^\top+\mathbf{I})^{-1}\\ &= \mathbf{I}-\gamma\mathbf{H}\mathbf{H}^\top+(\gamma\mathbf{H}\mathbf{H}^\top)^2-(\gamma\mathbf{H}\mathbf{H}^\top)^3+\cdots\\ &\approx \mathbf{I}-\gamma\mathbf{H}\mathbf{H}^\top\end{aligned}$$

where we used the Taylor expansion for the inversion of the sum of two matrices. Then, we can obtain

$$\begin{aligned}\boldsymbol{\Psi} &= (-\gamma^2\mathbf{H}^\top U^{-1}\mathbf{H}+\gamma\mathbf{I})^{-1}\\ &= [-\gamma^2\mathbf{H}^\top(\mathbf{I}-\gamma\mathbf{H}\mathbf{H}^\top)\mathbf{H}+\gamma\mathbf{I}]^{-1}\\ &= [\gamma\mathbf{I}-\gamma^2\mathbf{H}^\top\mathbf{H}+\gamma^3(\mathbf{H}^\top\mathbf{H})^2]^{-1}\\ &\approx [\gamma\mathbf{I}-\gamma^2\mathbf{H}^\top\mathbf{H}]^{-1}\\ &= \gamma^{-1}(\mathbf{I}+\gamma\mathbf{H}^\top\mathbf{H}+(\gamma\mathbf{H}\mathbf{H}^\top)^2+\cdots)\\ &\approx \mathbf{H}^\top\mathbf{H}+\gamma^{-1}\mathbf{I}.\end{aligned}$$

Appendix B
Optimizing the Variational Lower Bound $\mathcal{L}(\mathbf{X},\mathbf{Y})$

The variational lower bound on the marginal likelihood is

$$\begin{aligned}\mathcal{L}(\mathbf{X},\mathbf{Y}) &= \mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})+\log p(\mathbf{Y}|\mathbf{Z})]\\ &\qquad -\mathrm{KL}(q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})\parallel p(\mathbf{Z}))\end{aligned}$$

where the KL divergence term can be computed exactly as

$$\begin{aligned}&-\mathrm{KL}(q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})\parallel p(\mathbf{Z}))\\ &\qquad = \frac{1}{2}\sum_{i=1}^{N}\sum_{d=1}^{D_z}\left(1+\log\left(\sigma_{\mathbf{z}\,id}^2\right)-\mu_{\mathbf{z}id}^2-\sigma_{\mathbf{z}\,id}^2\right).\end{aligned}$$

On the other hand, the likelihood terms $\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\big]$ and $\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\big[\log p(\mathbf{Y}|\mathbf{Z})\big]$ can be approximated by Monte Carlo approximation

$$\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})] = \frac{1}{L}\sum_{i=1}^{N}\sum_{l=1}^{L}\log p_{\boldsymbol{\theta}}\big(\mathbf{x}_i\big|\mathbf{z}_i^{(l)}\big)$$

$$\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})}[\log p(\mathbf{Y}|\mathbf{Z})] = \frac{1}{L}\sum_{i=1}^{N}\sum_{l=1}^{L}\log p\big(\mathbf{y}_i\big|\mathbf{z}_i^{(l)},\mathbf{r}_i\big)$$

with $\mathbf{z}_i^{(l)} = \boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}_i,\mathbf{y}_i)+\boldsymbol{\sigma}_{\mathbf{z}}(\mathbf{x}_i,\mathbf{y}_i)\odot\boldsymbol{\epsilon}^{(l)}$, where $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}_{D_z}(\mathbf{0},\mathbf{I})$ and $\odot$ denotes elementwise multiplication.

Finally, $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ can be optimized by maximizing $\mathcal{L}(\mathbf{X},\mathbf{Y})$, using stochastic gradient descent method.

APPENDIX C
REGULARIZED POSTERIOR DISTRIBUTION $p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star)$

The regularized posterior distribution $p_{\text{reg}}(\mathbf{z}_\star|\mathbf{y}_\star) = \mathcal{N}_{D_z}(\mathbf{z}_\star|\boldsymbol{\mu}_{\mathbf{z}_\star}, \boldsymbol{\Sigma}_{\mathbf{z}_\star})$, where

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{z}_\star} &= \left[\langle \mathbf{B}\mathbf{T}\mathbf{B}^\top \rangle + \left(1 + \rho \sum_{i=1}^{N} s_i\right)\mathbf{I}\right]^{-1}\\ \boldsymbol{\mu}_{\mathbf{z}_\star} &= \boldsymbol{\Sigma}_{\mathbf{z}_\star}\left[\langle \mathbf{B}\rangle \mathbf{T}\mathbf{y}_\star + \rho \sum_{i=1}^{N} s_i \langle \mathbf{z}_i\rangle\right]\\ \mathbf{T} &= \gamma\,\mathbf{I} - \gamma^2 \langle \mathbf{H}^\top(\mathbf{I} + \gamma\,\langle \mathbf{H}\mathbf{H}^\top\rangle)^{-1}\mathbf{H}\rangle.\end{aligned}$$

ACKNOWLEDGMENT

The authors would like to thank Z.-Y. Liu for his valuable discussions.

REFERENCES

[1] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, “Encoding and decoding in fMRI,” *NeuroImage*, vol. 56, no. 2, pp. 400–410, May 2011.

[2] M. Chen, J. Han, X. Hu, X. Jiang, L. Guo, and T. Liu, “Survey of encoding and decoding of visual stimulus via FMRI: An image analysis perspective,” *Brain Imag. Behavior*, vol. 8, no. 1, pp. 7–23, Mar. 2014.

[3] J.-D. Haynes and G. Rees, “Decoding mental states from brain activity in humans,” *Nature Rev. Neurosci.*, vol. 7, no. 7, pp. 523–534, Jul. 2006.

[4] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, “Identifying natural images from human brain activity,” *Nature*, vol. 452, no. 7185, pp. 352–355, Mar. 2008.

[5] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, “Bayesian reconstruction of natural images from human brain activity,” *Neuron*, vol. 63, no. 6, pp. 902–915, Sep. 2009.

[6] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, “Reconstructing visual experiences from brain activity evoked by natural movies,” *Current Biol.*, vol. 21, no. 19, pp. 1641–1646, Sep. 2011.

[7] T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani, “Neural decoding of visual imagery during sleep,” *Science*, vol. 340, no. 6132, pp. 639–642, May 2013.

[8] A. S. Cowen, M. M. Chun, and B. A. Kuhl, “Neural portraits of perception: Reconstructing face images from evoked brain activity,” *NeuroImage*, vol. 94, pp. 12–22, Jul. 2014.

[9] S. Schoenmakers, U. Güçlü, M. A. Van Gerven, and T. Heskes, “Gaussian mixture models and semantic gating improve reconstructions from human brain activity,” *Frontiers Comput. Neurosci.*, vol. 8, pp. 173–182, Jan. 2015.

[10] H. Lee and B. A. Kuhl, “Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex,” *J. Neurosci.*, vol. 36, no. 22, pp. 6069–6082, Jun. 2016.

[11] M. A. J. Van Gerven, B. Cseke, F. P. De Lange, and T. Heskes, “Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior,” *NeuroImage*, vol. 50, no. 1, pp. 150–161, Mar. 2010.

[12] S. R. Damarla and M. A. Just, “Decoding the representation of numerical values from brain activation patterns,” *Hum. Brain Mapping*, vol. 34, no. 10, pp. 2624–2634, Oct. 2013.

[13] E. Yargholi and G.-A. Hossein-Zadeh, “Brain decoding-classification of hand written digits from fMRI data employing Bayesian networks,” *Frontiers Hum. Neurosci.*, vol. 10, no. 13, pp. 351–364, Jul. 2016.

[14] T. Horikawa and Y. Kamitani, “Generic decoding of seen and imagined objects using hierarchical visual features,” *Nature Commun.*, vol. 8, p. 15037–15052, May 2017.

[15] Y. Miyawaki *et al.*, “Visual image reconstruction from human brain activity using a combination of multiscale local image decoders,” *Neuron*, vol. 60, no. 5, pp. 915–929, Dec. 2008.

[16] M. A. Van Gerven, F. P. De Lange, and T. Heskes, “Neural decoding with hierarchical generative models,” *Neural Comput.*, vol. 22, no. 12, pp. 3127–3142, Dec. 2010.

[17] S. Schoenmakers, M. Barth, T. Heskes, and M. van Gerven, “Linear reconstruction of perceived images from human brain activity,” *NeuroImage*, vol. 83, no. 1, pp. 951–961, Dec. 2013.

[18] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, “Modular encoding and decoding models derived from Bayesian canonical correlation analysis,” *Neural Comput.*, vol. 25, no. 4, pp. 979–1005, Mar. 2013.

[19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[21] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, “From deterministic to generative: Multimodal stochastic RNNs for video captioning,” *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2018.2851077.

[22] J. Zheng, X. Cao, B. Zhang, X. Zhen, and X. Su, “Deep ensemble machine for video classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2018.2844464.

[23] G. Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[24] A. L. Maas *et al.*, “Building DNN acoustic models for large vocabulary speech recognition,” *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017.

[25] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[26] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, “How does the brain solve visual object recognition?” *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012.

[27] U. Güçlü and M. A. van Gerven, “Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream,” *J. Neurosci.*, vol. 35, no. 27, pp. 10005–10014, Jul. 2015.

[28] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, “Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence,” *Sci. Rep.*, vol. 6, pp. 27755–27768, Jun. 2016.

[29] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, “Neural encoding and decoding with deep learning for dynamic natural vision,” *Cerebral Cortex*, vol. 28, no. 12, pp. 4136–4160, Dec. 2017.

[30] I. Higgins *et al.* (Jun. 2016). “Early visual concept learning with unsupervised deep learning.” [Online]. Available: https://arxiv.org/abs/1606.05579

[31] D. P. Kingma and M. Welling. (Dec. 2013).“Auto-encoding variational Bayes.” [Online]. Available: https://arxiv.org/abs/1312.6114

[32] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *Proc. ICML*, May 2014, pp. 1278–1286.

[33] J. Zhu, N. Chen, and E. P. Xing, “Bayesian inference with posterior regularization and applications to infinite latent SVMs,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1799–1847, 2014.

[34] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in *Proc. ICML*, Jan. 2015, pp. 1083–1092.

[35] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, “Correlational neural networks,” *Neural Comput.*, vol. 28, no. 2, pp. 257–285, Feb. 2016.

[36] D. L. Yamins and J. J. Dicarlo, “Eight open questions in the computational modeling of higher sensory cortex,” *Current Opinion Neurobiology*, vol. 37, pp. 114–120, Apr. 2016.

[37] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, “Distributed and overlapping representations of faces and objects in ventral temporal cortex,” *Science*, vol. 293, no. 5539, pp. 2425–2430, Sep. 2001.

[38] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. NIPS*, 2014, pp. 2672–2680.

[39] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Proc. NIPS*, 2017, pp. 700–708.

[40] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets,” in *Proc. NIPS*, 2016, pp. 2172–2180.

[41] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proc. ICML*, 2017, pp. 214–223.

[42] C. Du, C. Du, and H. He, “Sharing deep generative representation for perceived image reconstruction from human brain activity,” in *Proc. IJCNN*, May 2017, pp. 1049–1056.

[43] Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A. van Gerven, "Reconstructing perceived faces from brain activations with deep adversarial neural decoding," in *Proc. NIPS*, 2017, pp. 4246–4257.

[44] K. Han, H. Wen, J. Shi, K.-H. Lu, Y. Zhang, and Z. Liu, "Variational autoencoder: An unsupervised model for modeling and decoding fMRI activity in visual cortex," *bioRxiv*, 2017.

[45] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, "Seeing it all: Convolutional network layers map the function of the human visual system," *NeuroImage*, vol. 152, pp. 184–194, May 2017.

[46] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. NIPS*, 2014, pp. 3581–3589.

[47] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. NIPS*, 2016, pp. 3738–3746.

[48] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data," *Trends Cognit. Sci.*, vol. 10, no. 9, pp. 424–430, 2006.

[49] O. Yamashita, M.-A. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *NeuroImage*, vol. 42, no. 4, pp. 1414–1429, Oct. 2008.

[50] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. 3, pp. 211–244, Jun2001.

[51] L. Van der Maaten, "A new benchmark Dataset for handwritten character recognition," *Tilburg Univ.*, Apr. 2009, pp. 2–5.

[52] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. ICCV*, Nov. 2011, pp. 2018–2025.

[53] M. Gonen, "Bayesian efficient multiple kernel learning," in *Proc. ICML*, Jun. 2012, pp. 1–8.

[54] D. P. Kingma and J. Ba. (Dec. 2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[56] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

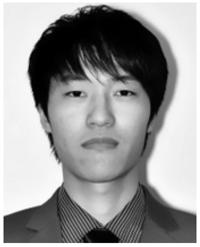
[57] D. B. Walther, E. Caddigan, L. Fei-Fei, and D. M. Beck, "Natural scene categories revealed in distributed patterns of activity in the human brain," *J. Neurosci.*, vol. 29, no. 34, pp. 10573–10581, Aug. 2009.

[58] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain," *Neuron*, vol. 76, no. 6, pp. 1210–1224, Dec. 2012.

[59] A. G. Huth, T. Lee, S. Nishimoto, N. Y. Bilenko, A. T. Vu, and J. L. Gallant, "Decoding the semantic content of natural movies from human brain activity," *Frontiers Syst. Neurosci.*, vol. 10, pp. 81–97, Oct. 2016.

[60] E. Chong, A. M. Familiar, and W. M. Shim, "Reconstructing representations of dynamic visual objects in early visual cortex," *Proc. Nat. Acad. Sci.*, vol. 113, no. 5, pp. 1453–1458, 2016.

[61] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Proc. NIPS*, 2015, pp. 2980–2988.



Changde Du received the B.E. degree in automatic control from the Beijing Information Science and Technology University, Beijing, China, in 2013, and the M.S. degree in data mining from the University of Chinese Academy of Sciences, Beijing, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences, Beijing.

His current research interests include machine learning, Bayesian statistics, brain-inspired intelligence, computational neuroscience, and applications in computer vision.

Mr. Du was a recipient of the AMD Science and Technology Award in 2016, and the Third Prize of Final Contest of National Collegiate Contest and International Invitational Tournament for Brain-Inspired Computing and Application in 2017.

Changying Du received the B.Sc. degree from Central South University, Changsha, China, in 2008, and the Ph.D. degree in machine learning from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2015.

From 2013 to 2014, he was a Visiting Scholar with Purdue University, West Lafayette, IN, USA. He is currently an Associate Professor with the Institute of Software, CAS. He has authored or co-authored over 20 peer-reviewed research papers in prestigous conferences and journals. His current research interests include machine learning, data mining, and Bayesian statistics.

Dr. Du serves as a PCC Member/reviewer for IJCAI 2018, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

Lijie Huang received the B.Sc. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the Ph.D. degree in cognitive neuroscience from the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, in 2016.

He is currently an Assistant Professor with the Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing. He is involved in science and technology challenges at the intersection of neuroscience, machine learning, and large-scale data analysis, e.g., how to read our mind using state-of-the-art machine learning techniques.

Huiguang He (M'04–SM'10) received the B.S. and M.S. degrees from Dalian Maritime University (DMU), Dalian, China, in 1994 and 1997, respectively, and the Ph.D. degree (Hons.) in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China.

From 1997 to 1999, he was an Associate Lecturer with DMU. From 2003 to 2004, he was a Post-Doctoral Researcher with the University of Rochester, Rochester, NY, USA. From 2014 to 2015, he was a Visiting Professor with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He is currently a Full Professor with CASIA. His research has been supported by several research grants from the National Science Foundation of China. He has authored or co-authored more than 100 peer-reviewed papers. His current research interests include pattern recognition, medical image processing, and brain computer interface (BCI).

Dr. He is an Excellent Member of Youth Innovation Promotion Association, CAS, in 2016. He was a recipient of the Excellent Ph.D. dissertation of CAS in 2004, the National Science and Technology Award in 2003 and 2004, the Beijing Science and Technology Award in 2002 and 2003, the K.C. Wong Education Prizes in 2007 and 2009, and the Jia-Xi Lu Young Talent Prize in 2009.