REPRESENTATION LEARNING OF FMRI DATA USING VARIATIONAL AUTOENCODER

by

Jung-Hoon Kim

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Weldon School of Biomedical Engineering
West Lafayette, Indiana
May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Zhongming Liu, Ph.D., Chair

Weldon School of Biomedical Engineering

Dr. Edward L Bartlett, Ph.D.

Weldon School of Biomedical Engineering

Dr. Alexander Chubykin, Ph.D.

Department of Biological Science

Dr. Yunjie Tong, Ph.D.

Weldon School of Biomedical Engineering

Approved by:

Dr. George R. Wodicka

To my loving family,

ACKNOWLEDGMENTS

I still remember the moment I joined the Laboratory of Integrated Brain Imaging (LIBI) in 2016, now looking in retrospective I see how much many people have helped me and inspired me throughout this process.

First and foremost, I am deeply grateful to my major advisor, Dr. Zhongming Liu, for his professional guidance and support in developing my scientific thinking, communication, and analytical skills that I never imagined I could have reached all by myself. As a strong scientist and engineer, he has shown me what the right attitude and mindset in devoting to science is. His valuable advice, “the simpler the better” will always be my guidance in research. I also want to express my deep gratitude to other members of my thesis committee, Dr. Edward Bartlett, Dr. Alexander Chubykin, and Dr. Yunjie Tong, for their critical comments and feedback at various stages of my PhD research.

Second, I want to thank all my colleagues in the LIBI for their invaluable advice and comments on this dissertation. Everyone I was fortunate enough to work with has been smart, hardworking, and thoughtful. The joyful conversations we had have been the resort during painstaking research. Thank you Ulrich, Yizhen, Jiayue, Kuan, Xiokai, and Minkyu, and former lab members Haiguang, Jun Young, Lauren, Kun-Han, Nishant, Junxing, and Steven. It was my honor to have worked with you. I also would like to express my gratitude to Dr. Nicholas Barbaro and Dr. Robert Worth, for their constructive discussions and support on the intracranial electroencephalogram experiments in Indiana University Hospital. In addition, I would like to thank Dr. David Brang for sharing his neuroscience knowledge and useful advice to my dissertation work.

Finally, I would like to thank my parents, my brother, my nephew, and my niece for their support and love. Without it, I would not have been able to accomplish the PhD course. In addition, I want to show my deep and sincere gratitude to my best friend and life partner María. Your unconditional support and love made me the person who I am today. My PhD journey would not have been possible if not for your support as well as our families’, and I dedicate this milestone to them.

TABLE OF CONTENTS

LIST OF TABLES 8

LIST OF FIGURES 9

LIST OF ABBREVIATIONS 15

ABSTRACT 17

1. INTRODUCTION 18

1.1 Defining the Problem 18

1.2 Source of fMRI Activity 19

1.3 Functional Connectivity of Brain 19

1.4 Brain Decoding and Encoding Using Deep Learning 20

1.5 Brain Encoding Using Unsupervised Deep Learning 22

2. BETA VARIATIONAL AUTOENCODER 25

2.1 Motivation 25

2.2 Input Structure of β-VAE Model 25

2.3 Model Architecture 26

2.4 Data and Training 27

2.5 Model Parameter 27

3. REPRESENTATIONAL LEARNING OF RESTING STATE FMRI WITH VARIATIONAL AUTOENCODER 35

3.1 Introduction 35

3.2 Methods and Materials 37

3.2.1 Testing Data 37

3.2.2 Synthesizing Resting State fMRI Functional Connectivity 37

3.2.3 Defining a Principal Basis Set in the Latent Space 38

3.2.4 Clustering in the Latent Space 39

3.2.5 Individual Variation 40

3.2.6 Subject Identification 40

3.2.7 Comparison with Linear Latent Space 42

3.3 Results 42

3.3.1 VAE Compressed Resting State fMRI Maps 42

3.3.2 VAE Synthesized Correlated fMRI Activity 43
3.3.3 Latent Variables Reflected Network Dynamics 44
3.3.4 Clusters in the Latent Space 45
3.3.5 Individual Variation of Latent Representation 46
3.3.6 Individual Identification 47
3.4 Discussion 49
4. LEARNING TASK-EVOKED REPRESENTATION OF FMRI UNDER NATURALISTIC MOVIE WATCHING 64
4.1 Introduction 64
4.2 Methods and Materials 66
4.2.1 Subjects and Data 66
4.2.2 Movie Stimulus Paradigm 67
4.2.3 Eye-Tracking Data 67
4.2.4 Extracting Video Features from Movie Stimuli 68
4.2.5 Generalizability of Pretrained VAE Model on Movie-Watching fMRI Data 69
4.2.6 Linear Superposition Between Task-Evoked Activity and Spontaneous Activity in the Latent Space 70
4.2.7 Task-Evoked Latent Representation 71
4.2.8 Defining Basis Functions of Task-Evoked Gradient 71
4.2.9 Reproducibility of Task-Evoked Latent Representations 72
4.2.10 Correlating Latent Trajectory with Low-, Middle-, High-Level Aspects of Movie Stimuli 73
4.2.11 Task-Relevant Gaze Position Explains Variance of Latent Gradients 74
4.2.12 Individual-Wise Cortical Activity Depends on the Gaze Location 74
4.3 Results 74
4.3.1 Pretrained VAE Successfully Compressed Unseen Movie-Watching fMRI Data 74
4.3.2 Linear Superposition Between Task-Evoked Activity and Spontaneous Ongoing Activity in the Latent Space 76
4.3.3 Geometry of Task-Evoked Latent Representation 77
4.3.4 Displacement Magnitude of Latent Representation Is Specific to Movie Contents .. 77
4.3.5 Individual Variation of Latent Trajectory 78

4.3.6 Task-Evoked Latent Variable Is Highly Reproducible 79
4.3.7 Cortical Mapping of Task-Evoked Latent Gradient 80
4.3.8 Semantic Meaning of Task-Evoked Latent Gradient 81
4.3.9 Gaze Information Is Related to Task-Evoked Latent Gradient 82
4.4 Discussion 84
5. CONCLUSION 108
5.1 Conclusion 108
5.2 Future Applications 109
REFERENCES 111
PUBLICATIONS 133

LIST OF TABLES

Table 2.1. Algorithm of β-VAE for learning the representation of fMRI data 34

Table 3.1 Subject identification accuracy across different conditions 63

Table 4.1 Description of movie stimuli 106

Table 4.2 Description of semantic labels 107

LIST OF FIGURES

Figure 2.1. Geometric reformatting. The cortical distribution of fMRI activity is converted into a spherical surface and then to an image by evenly resampling the spherical surface with respect to sin(e) and a, where e and a indicate elevation and azimuth, respectively. 28

Figure 2.2. Grayordinates (# = 29,696) in the left hemisphere (left panel) is projected into 2D space (middle panel) based on their azimuth- and elevation levels. Each color stands for different brain atlas based on the literature (Glasser et al., 2016). The imbalance in data density along the varying elevation level (upper in right panel) is alleviated by further applying elevation to Sine function (bottom in right panel). 29

Figure 2.3. Architecture of VAE. Simplified block diagram of VAE model (upper panel in B). An encoder network samples latent variables given an input image under the inference model while a decoder network generates a genuine input image from under the generative model. Details of VAE model (bottom panel in B). Both encoder and decoder network contain 5 convolutional layers. In the encoder network, the size of output image of each layer (from left to right) is 96×96×64 (32 channels per hemisphere), 48×48×128, 24×24×128, 12×12×256, and 6×6x×56; for the decoder network, 6×6×256, 12×12×256, 24×24×128, 48×48×128, and 96×96×64 (32 channels per image), from left to right. The dimension of latent variables is 256. The convolution operations are defined as: 1: convolution (kernel size=8, stride=2, padding=3) with rectified nonlinearity, 2-5: convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 6: fully-connected layer with re-parametrization, 7: fully-connected layer with rectified nonlinearity, 8-11: transposed convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 12: transposed convolution (kernel size=8, stride=2, padding=3). Blue and red boxes stand for the input images from left and right hemispheres, respectively. 30

Figure 2.4. Validation error of VAE at varying beta values. Trade-off between reconstruction loss and Kullback–Leibler divergence is visualized for different beta values. Red color stands for the value we used for analysis. 31

Figure 2.5. Validation curve of VAE at different layers. (a) Layer 6; Both encoder and decoder network contain 2 convolutional layers. In the encoder network, the size of output image of each layer (from left to right) is 48x48x128 (64 channels per hemisphere), and 12x12x256; for the decoder network, 12x12x256, and 48x48x128 (64 channels per image), from left to right. Layer 8; 3 convolutional layers for encoder and decoder networks. The size of output image of each encoder layer is 96x96x64 (32 channels per image), 24x24x128, and 6x6x256; decoder network: 6x6x256, 24x24x128, and 96x96x64 (32 channels per image), from left to right. (b) The validation curve as a function of training epochs.. 32

Figure 2.6. Validation curve of VAE at varying learning rate... 33

Figure 3.1. Image reconstruction using VAE. A series of cortical patterns are reconstructed through the VAE model. Among them, five original cortical patterns (upper panel) and their corresponding reconstruction through VAE (bottom panel) are visualized for comparison. For an example region (green circle), the time series of the original activity (black line) and the reconstructed activity (red line) are plotted for comparison. 53

Figure 3.2. Resting-state fMRI data compression and reconstruction with VAE vs. PCA and ICA. (a) For illustration, three example maps of fMRI activity, before (1st row) and after (2nd row) being smoothed (FWHM=6mm), are shown in comparison with the corresponding maps reconstructed with VAE (3rd row), PCA (4th row), and ICA (5th row) trained to compress and reconstruct the training data from 100 subjects with 256 variables or components. (b) For quantitative comparison, the reconstruction performance, in terms of the percentage of variance in the fMRI images as explained by the model reconstruction, is shown for VAE, PCA, and ICA as a function of FWHM (from 1 to 10 mm) applied to the spatial smoothing of the fMRI images. The error bar stands for the standard error of mean. 54

Figure 3.3. VAE synthesizes correlated fMRI activity. (a) Seed-based correlations of VAE-synthesized fMRI data (top row) vs. experimental fMRI data (bottom row) with the seed location (green circle) at V1 (left), IPS (middle), or PCC (right). (b) Spatial correlations between the seed-based functional connectivity based on VAE-synthesized data and those based on measured fMRI data concatenated across 1, 5, 10, 50, or 100 subjects. The colors indicate different seed locations (V1: black; IPS: red; PCC: blue). Similarly, (c) shows the spatial correlation between the synthesized vs. measured functional connectivity among 360 cortical parcels. The error bar indicates the standard error of the mean averaged across 20 repeated trials. 55

Figure 3.4. Latent variables drive the dynamics of latent representation. The latent variables are defined as the principal directions for the dynamic change of latent representation across adjacent time points. (a) Visualization of the 1st latent variable as a cortical pattern of the signed standard deviation. (b) For each of the four cortical locations, denoted as i through iv and shown as green circles in (a), the activity change is shown as a function of the 1st latent variable, indicating a varying nonlinear relationship. (c) The percentage of the variance that each latent variable explains the first order dynamics of latent representation. The inset shows the percentage of the total variance explained by top 10, 20, 50, or 100 latent variables. (d) The visualization of the 2nd through 10th latent variables as cortical patterns.. 56

Figure 3.5. Clusters of latent representations. (a) The sum of within-cluster distance is shown as a function of the number of clusters (or K). The choice of k=21 is about where an elbow is observable in the plot. (b) Representational dissimilarity matrix of the 1-cosine distance between instantaneous latent representations (from 100 subjects with 1,200 time points per subject) reordered by cluster membership. (c) Hierarchical clustering of 21 cluster centroids. The clusters are grouped into three super-clusters, provisionally labeled as sensorimotor (red), default mode (green), and task positive (blue) networks, based on the cortical visualization of individual clusters in (d). (d) Cortical patterns decoded from every cluster centroid. The number shows the cluster index. The scale of each pattern is normalized by its maximal absolute value. (e) The group-averaged cluster-wise activity, described as the cosine affinity of instantaneous representation to the centroid of each cluster. Each thin line corresponds to one cluster; the color indicating the super-cluster (sensorimotor: red; default mode: green; task positive: blue) that each cluster belongs 57

Figure 3.6. Individual variation of latent representation obtained with VAE vs. PCA. (a-b) Subject-wise latent representations visualized in a 2-D space obtained with t-SNE, when (a) VAE or (b) PCA is used to extract representations of rsfMRI activity from 20 subjects. (c) The Silhouette value shows how similar a representation is similar to each other within the same subject as opposed to between different subjects for VAE (left) or PCA (right). (d) The top-1, 5, and 10 accuracy of using the time-averaged representation as the feature to identify individuals in a large group of (n=500) subjects, for the representations obtained with VAE (black) or PCA (blue). 59

Figure 3.7. Individual identification based on functional connectivity between latent variables or cortical parcels. (a) Density distributions of z-transformed correlations between 60

Figure 3.8. Individual identification with nonlinear vs. linear representations. Each plot shows the histogram of the similarity in the representational geometry between sessions within the same subject (red) vs. across different subjects (black), for representations in the nonlinear latent space obtained by VAE (top) or in the linear latent space obtained by PCA (middle) or ICA (bottom). The similarity reported is based on the inter-session correlation coefficient (or r). The histogram is discretized by bins with a width of 0.02... 62

Figure 4.1. Example of group-level eye gaze heat map. Four 7-seconds movie clips with group-level gaze heatmap were presented. Each row stands for each movie clip. The temporal resolution of the movie scene is 1 second. 88

Figure 4.2. Semantic label of movie stimuli. (a) Appearance frequency of semantic label (bottom) and co-appearance frequency between semantic labels (top). (b) Few scenic examples of semantic labels. 89

Figure 4.3. Definition of representation, gradient, and trajectory in the latent space. (Left) From the original latent space (# of dimension=256), principal subspaces explaining the task-evoked representations are defined using PCA. (Right) Task-evoked latent representation stands for a latent representation of a single fMRI time point averaged across subjects. Task-evoked latent trajectory is a trend how latent representations are traveling over the progression of movie stimulus. Task-evoked latent gradient, i.e., positional displacement of task-evoked representation, stands for the subtraction between temporally adjunct latent representations. 90

Figure 4.4. Illustration of VAE model and task-evoked latent representation. (a) VAE Model architecture. The VAE model consists of two compartments, the encoder (coded as yellow) and the decoder (coded as green). The convolution operations are defined as: 1: convolution (kernel size=8, stride=2, padding=3) with rectified nonlinearity, 2-5: convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 6: fully-connected layer with re-parametrization, 7: fully-connected layer with rectified nonlinearity, 8-11: transposed convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 12: transposed convolution (kernel size=8, stride=2, padding=3). Blue and red boxes stand for the input images from left and right hemispheres, respectively. (b) VAE encoder with learnt parameters ϕ outputs latent representations (z) given individual's fMRI data (x). Latent representations (z) evoked by the movie-watching task are estimated by averaging latent representations across 173 subjects. 91

Figure 4.5. Generalizability of pretrained VAE model. Three different datasets (red, 7T movie-watching fMRI data; blue, 7T resting-state fMRI; green, 3T resting-state fMRI) were reconstructed by VAE and independent component analysis (ICA), as a linear counterpart of VAE model. Reconstruction performances of fMRI cortical patterns are measured by squared correlation coefficient. The null distributions were estimated by estimating the similarity between synthesized cortical patterns and smoothed cortical input patterns. *: Bonferroni-corrected p<0.01, two-sample t-test for testing across different datasets (black lines) and paired t-test for testing within the datasets (colored lines), n=40 subjects. n.s: not significant, Bonferroni-corrected p>0.05. 92

Figure 4.6. Smoothing effect of VAE vs. ICA. The reconstruction performance, in terms of the percentage of variance in the fMRI images as explained by the model reconstruction, is shown for VAE, ICA, and ICA estimated from 7T rsfMRI, as a function of FWHM (from 1 to 10 mm) applied to the spatial smoothing of the fMRI images. The error bar stands for the standard error of mean. 93

Figure 4.7. Linear superposition between task-evoked activity and spontaneous activity in different latent spaces. fMRI data under movie-watching task (1st column) and resting state (2nd column) are compressed into latent spaces defined by VAE (top), or latent-spaces by cortical parcels (middle) or ICA (bottom). For each latent space, covariance matrix between latent variables is measured. Additionally, task-evoked covariance matrix is acquired by estimating correlation between group-level latent variables. Similarity between condition-specific covariance matrix (3rd column, movie-watching vs. rest) and task-evoked covariance matrix is quantified by measuring squared correlation value between off-diagonal elements of two matrices. 94

Figure 4.8. Linear superposition between task-evoked activity and spontaneous activity in the partitioned dataset. By partitioning 173 subjects into ten subsamples (n=17), similarity between condition-specific correlation matrix and task-evoked correlation matrix is quantified by measuring squared correlation value between off-diagonal elements of two matrices. Error bar stands for standard deviation of mean. Explained variance by VAE-derived latent space is significantly higher than parcel- or ICA-based latent space (p<0.01 for parcel; p<0.01 for ICA). There is also significant difference between parcel-based latent space and ICA-based latent space (p<0.01). One-way ANOVA, *F*=863.34, p<0.01, post-hoc paired t-test, Bonferroni correction. 95

Figure 4.9. Geometry of task-evoked latent representation and spontaneous latent representation. (a) Cross correlation matrix between latent variables, under different brain states. (b) Variance explained by principal components. The number of principal components required to explain 90 % of latent variables are counted (dashed line). (c) Variance explained by each principal component. The black arrow stands for the crossover. 96

Figure 4.10. Displacement magnitude of latent representation is specific to movie contents. (a) Representation, gradient, and trajectory defined in the latent space. (b) Magnitude of task-evoked latent gradient as a function of time. Gray box stands for inter-movie resting period (20 seconds). Four recording runs are concatenated. Blue line stands for the boundary between different sessions. The number stands for the numbering of short movie clips (purple: from Vimeo, orange: from Hollywood movie). One movie clip (5, 5’, 5’’, and 5’’’) is repeated over recordings. (c) Average magnitude of task-evoked latent gradient for different movie clips. There is a statistically significant difference in the mean magnitude between video clips (one-way ANOVA, F= 39.96, p<0.001). (d) The dynamics of magnitude over the progression of inter-movie rest periods (n=15), including 5 seconds earlier and 10 seconds later of rest period. Dashed line stands for averaged magnitude during movie stimuli present. *: q<0.05, FDR-corrected. 97

Figure 4.11. Example of scenic changes related to the magnitude of task-evoked latent gradient. From two short video clips (1: Two men; top and 11: Garden; bottom), movie scenes corresponding to lower or higher magnitude of latent gradient are visualized, after considering hemodynamic delay (5 seconds).......... 98

Figure 4.12 The trajectory of latent representations of individuals. t-SNE map of latent representations during watching the movie (Northwest, 104 seconds after the starting of movie), rest, or the transition from rest to movie-watching (bottom). Representations are color-coded by the task conditions (left), and subject indices (right). 99

Figure 4.13. Reproducibility of task-evoked latent gradient. (a) scenes of a short video clip (5; test-retest) are visualized at the pace of 1 seconds. Speaker image stands for the starting of audio. Red box and red number stand for the time points that have big jumps in panel (b). Hemodynamic delay (5 seconds) is considered. (b) t-SNE map of task-evoked representations during watching test-rest video clips (color-coded, 4 sessions). Gray dots stand for task-evoked representations from other movie-watching periods. (c) Dynamics of top-5 basis functions explaining the task-evoked latent gradient. Big jumps found in (b) are marked as dashed lines and red numbers. (d) The inter-session reproducibility of dynamics of basis functions. Gray thin line stands for each pair between sessions (n=6). Black thick line stands for the averaged reproducibility across 6 pairs of sessions. (e) The variance of latent gradient explained by each basis. 100

Figure 4.14. Reproducibility of basis functions defined by repeated datasets. The cumulative variance of latent gradient explained by basis functions. For four sessions, 9 basis functions explained 99% variance (dashed line). (b) Inter-session similarity of basis functions (black upper triangle) and their dynamics (red lower triangle). (c) Similarity between basis functions estimated from the whole data and basis functions defined from repeated dataset. 101

Figure 4.15. Cortical mapping of task-evoked gradient in the latent space and its predictability by video features. (Top) Cortical mapping of 24 basis functions defining the task-evoked latent gradient. (Bottom) Per short video clip, the dynamics of basis functions are correlated with low- (left panel) and middle-level video features (right panel). Error bar stands for the standard error of mean. The number in bottom right stands for basis functions having Fisher's *z*-values different from zero. *: q<0.05, FDR-correction, one-sample t-test. 102

Figure 4.16. Semantic meaning of task-evoked latent gradient. Top-6 basis functions having highest explained variance by regression model consisting of HRF-convolved semantic features (from left top to right bottom). Per basis function, each semantic label is color-coded by its coefficient weight in the fitted ordinary regression model. 103

Figure 4.17. Task-specific gaze position is related to task-evoked latent gradient. (a) and (d) The variance of each basis function explained by regression model only with semantic features (gray), with semantic features and changes in eye-open duration feature (green), or with semantic features and changes in horizontal gazing position feature (dark red). The red star stands for the basis function that shows highest explained variance by gaze-informed feature, as estimated in panel (b). (b) The improvement of explained variance by each semantic feature or eye-open duration feature, in basis function 5. (c) and (e) The cortical mapping of basis functions showing great improvement by gaze-informed features.. 104

Figure 4.18. Individual variability in gaze position explains individual variability in brain response. (a) Exemplified movie scenes having lower (left) or higher (right) inter-subject variability in gaze position. (b) Per fMRI data point, Euclidean distance between individual's gaze position (upper triangle) or between individual's brain response in the latent space (lower triangle). These distance matrices are acquired when the specific scene (right) presents. (c) Per time point of fMRI data, the representational similarity between two distance matrices is measured and plotted as a histogram. Red line stands for the average representational similarity. The center of distribution (red line) is not 0 (p<0.01, one-sample t-test). (d) The representational similarity is grouped based on the presence of face or during inter-session rest period. One-way ANOVA, F= 144.43, p<0.01, post-hoc two-sample t-test, FDR-corrected. 105

LIST OF ABBREVIATIONS

Abbreviations	Explanation
AD	Alzheimer’s Disease
ADHD	Attention-Deficit/Hyperactivity Disorder
BOLD	Blood Oxygen Level Dependent
β-VAE	Beta-Variational Autoencoder
CNN	Convolutional Neural Network
EEG	Electroencephalography
FC	Functional Connectivity
FDR	False Discovery Rate
FFA	Fusiform Face Area
FWHM	Full Width at Half Maximum
GAN	Generative Adversarial Network
HCP	Human Connectome Project
HRF	Hemodynamic Response Function
ICA	Independent Component Analysis
iEEG	intracranial EEG
IFJ	Inferior Frontal Junction
iPAT	image acceleration factor
IPS	Intraparietal Sulcus
ISFC	Inter-Subject Functional Connectivity
IS-RSA	Inter-Subject Representational Similarity Analysis
LFP	Local Field Potential
MCI	Mild Cognitive Impairment
MEG	Magnetoencephalography
MNI	Montreal Neurological Institute
MRI	Magnetic Resonance Imaging
MST	Medial Superior Temporal
PCA	Principle Component Analysis
PCC	Posterior Cingulate Cortex

RNN	Recurrent Neural Network
ROI	Region of Interest
SLA	Supplementary Language Area
rsfMRI	Resting state functional Magnetic Resonance Imaging
t-SNE	t-Distributed Stochastic Neighbor Embedding
V1	Primary visual area
VAE	Variational Autoencoder

ABSTRACT

Functional imaging data of the brain using Magnetic Resonance Imaging (MRI) – fMRI data exhibits complex but structured patterns. This fMRI data has opened a new venue for understanding the brain system at the whole-brain scale. However, the underlying origins of fMRI data are unclear and entangled. In this dissertation, I establish a variational auto-encoder, a generative model trainable with an unsupervised learning algorithm, to disentangle the unknown sources of fMRI activity. After being trained with large fMRI data in cooperation with a new reformatting strategy of input fMRI data, the model has learned the representations of cortical activity using latent variables. In Chapter 3, I found that the latent representation and its trajectory represented the spatiotemporal characteristics of fMRI activity under resting state. The latent variables reflected the principal gradients of the latent trajectory and drove activity changes in cortical networks. Latent representations were clustered by both individuals and brain states. Representational geometry captured as the covariance between latent variables, rather than cortical connectivity, was used as a more reliable feature to accurately identify subjects from a large group, even if only a short period of data was available per subjects. In Chapter 4, I further applied the VAE model pretrained with fMRI data in the resting state to new fMRI data from subjects watching naturalistic movies. I further validated that my VAE model was highly generalizable to fMRI data under different brain conditions and different scanning parameters. Additionally, I showed the task-evoked brain activity and spontaneous brain activity could be linearly separable in the VAE-derived latent space. Task-evoked latent representations and trajectory were employed to understand the dynamics of brain networks during naturalistic movie stimuli. I found that the principal gradients of the task-evoked latent trajectory were related to many aspects of the movie stimuli: low-, middle-, high-level video features. Cortical mapping of principal gradients showed the interactions between distributed cortical networks spanning from low-level sensory to high-level cognitive. Taken together, the VAE model proposed in this dissertation is a novel and effective tool that can potentially be used for understanding cortical dynamics in different brain conditions and disease conditions.

1. INTRODUCTION

1.1 Defining the Problem

For centuries, the way the human brain receives, processes, and reacts to sensory information from the real world has been of great interest. Thanks to recent advancements in recording and imaging modalities e.g., Electroencephalogram (EEG), Magnetoencephalogram (MEG), intracranial EEG (iEEG), and Magnetic Resonance Imaging (MRI), it now has been revealed that the human brain is a sophisticated network consisting of interareal communications spanning from micro-scale (e.g. neural microcircuit), to meso-scale (e.g. cortical layers), and to macro-scale (e.g. sensory systems) (Gilbert and Li, 2013; Hirabayashi et al., 2013; Rauschecker and Scott, 2009; Van Kerkoerle et al., 2014). Despite the numerous advances that have been made by painstaking efforts from great scientists, there are still many gaps that remain to be filled in understanding the systematic mechanism of the brain.

Of other modalities for brain activity, functional MRI has been favored as a primary imaging tool to investigate brain activity due to its full brain coverage with millimeter-scale (Glover, 2011). A seminal fMRI study done by (Biswal et al., 1995) has shown how brain regions remotely located in the anatomical cortical space exhibited synchronized brain activities – considered as functionally "wired". Followed by, imaging studies have shown canonical wiring patterns – Functional Connectivity (FC), and those are observable under different brain states such as awake (Damoiseaux et al., 2006; Horovitz et al., 2008; Smith et al., 2009; Vincent et al., 2007), sleep (Curtis et al., 2016; Larson-Prior et al., 2009), and even under the anesthetized state (Hutchison et al., 2013b; Kiviniemi et al., 2000; Vincent et al., 2007). Naturally, the "functional connectivity" analysis of fMRI data has been a major analysis technique.

The underlying assumption of functional connectivity is a collinearity of measured signals between remote brain regions by ignoring the non-linearity between cortical sources and observed activity from fMRI modality, and the possible non-linearity of the brain functionality. **There has been increasing evidence that the measured response of fMRI modality can be highly non-linear due to the complex hemodynamic relationship between neural activity and blood-oxygen-level-dependent signal that is used in fMRI modality** (Friston et al., 1998; Liu et al., 2010; Sheth et al., 2004). Therefore, an advanced analytical technique that acknowledges the evident non-linear nature of fMRI activity is demanded to improve our understanding of the brain system.

1.2 Source of fMRI Activity

Now, we know the underlying sources of the BOLD signal in fMRI primarily result from synaptic inputs to neurons (Logothetis, 2002; Logothetis et al., 2001). The synaptic activity also causes local metabolic and hemodynamic changes observable with fMRI sensitized to blood-oxygenation-level-dependent (BOLD) contrast (Buxton, 2009; Metea and Newman, 2006). The neurovascular coupling acts as a temporally low-pass filter, limiting the temporal resolution and specificity of fMRI.

This notion is supported by the evidence obtained with simultaneously recorded neural and fMRI signals (Arthurs and Boniface, 2002; Goense and Logothetis, 2008; Logothetis et al., 2001). The BOLD fMRI signal accompanies, but lags behind, changes in local field potentials (LFP) that reflect synaptic input to neuronal ensembles (Buzsáki et al., 2012). The LFP-fMRI coupling is not confined to a single frequency band, but applies to many, if not all, frequency components of neural activity (Goense and Logothetis, 2008). The fMRI signal is also coupled with electrocorticography (ECoG) and EEG in a similar fashion (Goldman et al., 2002; Goncalves et al., 2006; Mukamel et al., 2005; Niessing et al., 2005; Wan et al., 2006; Yuan et al., 2010). Such a non-linear, complex relationship between neural activity and fMRI signal has been modeled through several studies (Friston et al., 2000; Lindquist et al., 2009; Logothetis et al., 2001; Martin et al., 2006), called as a hemodynamic response function. While hemodynamic response models can be useful for fMRI data under simple experimental designs e.g., a block-design, it does not apply to fMRI data with no or limited behavioral models.

1.3 Functional Connectivity of Brain

Functional Connectivity (FC) captures a linear dependency between different brain regions and/or between networks. As correlation analysis is purely data-driven, model-free, and conceptually simple, it has been one of the most widely used techniques to characterize the functional organization of the brain under task-free conditions e.g., resting-state (van den Heuvel and Hulshoff Pol, 2010). As FC stems from the anatomical “hard” wiring (Honey et al., 2009), FC patterns observable during resting-state are repeatedly observed not only during performing behavioral tasks (Elliott et al., 2019; Shah et al., 2016; Yuan et al., 2015) but also across different states of alertness, sleep states, and even anesthetized states (Curtis et al., 2016; Fukunaga et al., 2006; Horovitz et al., 2009; Horovitz et al., 2008; Hutchison et al., 2013b; Kaufmann et al., 2006; Kiviniemi et al., 2000; Larson-Prior et al., 2009; Martuzzi et al., 2010; Sämann et al., 2011; Vincent et al., 2007; Zhao et al., 2008). Yet, FC can be effectively altered by different brain states or different disease progressions. Given such characteristics, various applications have shown the possibility of FC patterns as an accurate classifier of behavioral states (Gonzalez-Castillo et al., 2015), neural "fingerprinting" (Finn et al., 2015), and predictors of disease progression (Drysdale et al., 2017; Zeng et al., 2012; Zhou et al., 2010).

Several labs including our lab have questioned whether such alternation introduced by the differences in brain conditions will be linearly additive (or deductive) (Bianciardi et al., 2009a; Churchland et al., 2010; He, 2013; Monier et al., 2003; Ponce-Alvarez et al., 2013). Contrary to popular belief, task-evoked brain activity is not independent of spontaneous ongoing brain activity. Rather, engaging to tasks suppresses spontaneous brain activity – named as the negative task-rest interaction. On top of the complex, non-linear relation between observed fMRI signal and neural activity, the existence of complex task-rest interaction suggests there is a significant amount of non-linear information of fMRI data, which has been overlooked by conventional linear analysis. Thus, I design a new unsupervised machine-learning model capable of capturing meaningful representations not only from linear information but also from non-linear information of fMRI data, leading to the systematic understanding of the brain mechanism (**Chapter 2**).

1.4 Brain Decoding and Encoding Using Deep Learning

Since the remarkable success of deep learning in the computer vision field (Krizhevsky et al., 2017), there have been enormous efforts attempted to utilize deep learning models to accelerate our understanding of the brain. The deep learning models mimicking human perception to visual and auditory input have provided deep insights into how the brain receives and processes sensory information, and reacts to the real world – brain encoding (LeCun et al., 2015; Schmidhuber, 2015). To name a few, the Convolutional Neural Network (CNN) model trained to classify static natural images has shown that the human vision is a hierarchical cascade of non-linear yet simple processors (Khosla et al., 2019c; Richards et al., 2019; Yamins and DiCarlo, 2016). The Recurrent

Neural Network (RNN) model that incorporate the temporal information of animated images has further broaden our knowledge that the human vision system stores dynamics of the movie, named as the temporal receptive window, by processing through recurrences between brain hierarchies (Güçlü and van Gerven, 2017; Hardy and Buonomano, 2018; Shi et al., 2018). Recently, deep learning models directly implementing the predictive coding theory further improved the object recognition performance under the same computational resources as the CNN model, suggesting the human vision system is composed of bottom-up feedforward and top-down feedback system (Han et al., 2018; Wen et al., 2018b), backed by electrophysiology studies (Bastos et al., 2015; Bubic et al., 2010; Fries, 2015; Kawato, 1999; Klink et al., 2017; Mejias et al., 2016; Michalareas et al., 2016; Pickering and Clark, 2014; Rauschecker and Scott, 2009; Scheeringa et al., 2016; Van Kerkoerle et al., 2014).

As opposed to brain encoding using deep learning models, there has been another lane of studies directly utilizing deep learning models to analyze the brain signal – brain decoding (Naselaris et al., 2011). Most decoding models employed supervised learning that was trained to perform specific tasks given the input of brain signals. Based on their objectives, previous brain encoding studies can be divided roughly into two categories, characterizing unique traits of healthy individuals, or predicting disease phenotypes. As applications detecting unique traits of individuals, deep learning studies tried to identify individuals among the population (Chen and Hu, 2018; Wang et al., 2019a), and to characterize different brain conditions (Jang et al., 2017; Koppe et al., 2019; Li and Fan, 2018; Oota et al., 2019; Qiao et al., 2019; Vu et al., 2020; Wang et al., 2020), age (Gadgil et al., 2020; Wen et al., 2020; Xia et al., 2019), sex (Fan et al., 2020; Gadgil et al., 2020), cognitive traits (Fan et al., 2020), and emotion (Kim et al., 2019). As clinical applications of brain encoding, significant efforts were made in finding phenotypes of Alzheimer's disease (Ebrahimi-Ghahnavieh et al., 2019; Feng et al., 2020; Goceri, 2019; Kam et al., 2019; Liu et al., 2020a; Meszlényi et al., 2017; Qureshi et al., 2019; Suk et al., 2016; Wang et al., 2019b; Yang et al., 2019) and autism spectrum disorder (Ahmed et al., 2020; Bengs et al., 2020; D'Souza et al., 2019; Dvornek et al., 2018a; El-Gazzar et al., 2019; Guo et al., 2017; Sharif and Khan, 2019), as well as Attention Deficit Hyperactivity Disorder (ADHD) (Riaz et al., 2020), schizophrenia (Chen et al., 2020; Matsubara et al., 2019; Yan et al., 2019), and Parkinson's Disease (Zhang et al., 2018). It is noteworthy that there are ongoing efforts that are trying to model hemodynamic response models (Cui et al., 2019), or to segment brain networks such as the default mode network (Zhao et al.,

2018) or hippocampus (Liu et al., 2020a), in a data-driven fashion. Lastly, instead of fMRI data, there are also deep learning studies using other brain signals such as EEG (Dubreuil-Vall et al., 2020; Gao et al., 2020b; Jang et al., 2018; León et al., 2020; Zeng et al., 2020) or functional near-infrared spectroscopy (Ho et al., 2019; Xu et al., 2020), to predict brain condition, identify individual's traits, or identify disease phenotypes. Among them, the deep learning application using the EEG signal as a seizure predictor is of interest since fMRI activity might fail to capture useful information of seizure prediction due to the poor temporal resolution (Zeng et al., 2020).

However, labels or behavioral information that are central to the supervised learning regime are commonly limited compared to the spatial dimension of input i.e., fMRI data. To alleviate such limitation, most studies compressed the input i.e., fMRI data at the region-of-interest level (Chen and Hu, 2018; Dvornek et al., 2018b; Koppe et al., 2019; Matsubara et al., 2019; Suk et al., 2016; Wang et al., 2019a; Wang et al., 2020), or at the network level (D'Souza et al., 2019; Fan et al., 2020; Kawahara et al., 2017; Kim and Lee, 2016; Riaz et al., 2020; Seo et al., 2019; Venkatesh et al., 2019; Yang et al., 2019; Zhao et al., 2018). Nevertheless, it is uncertain how much information, especially non-linear information of fMRI data, will remain after the linear data dimension reduction. Therefore, it has been debated whether supervised deep neural networks are superior to conventional and simpler machine-learning methods (He et al., 2020).

1.5 Brain Encoding Using Unsupervised Deep Learning

Given the limitation of supervised learning models, an unsupervised learning strategy is a reasonable alternative for fMRI data. There have been many unsupervised deep learning methods and their models were tested in different tasks e.g., a short-time prediction of fMRI data (Brown et al., 2020; Huang et al., 2017; Kashyap and Keilholz, 2020; Khosla et al., 2019a; Ravi et al., 2019), predicting brain conditions (Huang et al., 2017; Oota et al., 2019; Zhao et al., 2019), identifying disease phenotypes (Gao et al., 2020a; Guo et al., 2017; Liu et al., 2020b; Lostar and Rekik, 2020; Matsubara et al., 2019; Oh et al., 2019; Ravi et al., 2019; Saeed et al., 2019; Seo et al., 2019; Suk et al., 2015; Suk et al., 2016; Zhao et al., 2020), or predicting demographic traits (Wen et al., 2020; Xia et al., 2019). Same as supervised deep learning studies, however, most of the abovementioned studies limited their input by manually compressing fMRI data at the region-of-interest level or at the FC level, leading to the same concern as supervised methods. It is noteworthy that there are a couple of unsupervised deep learning studies that used the raw 3D volumetric fMRI data as their input, but those studies had to limit the depth of their deep learning models due to the heavy computational load coming from the input size (Brown et al., 2020; Liu et al., 2020b). Different from natural images, cortical morphology is very convoluted including several layers of gyrus and sulcus, requiring enough depth to address the distinction between the brain anatomy and the brain functionality. Therefore, I sought to design a new input format of fMRI data that can balance between the information of fMRI data for being a useful application and the size of input for being a practical application (**Chapter 2**).

Most of the existing unsupervised deep learning methods are based on the "auto-encoder" strategy consisting of two compartments: 1) an encoder; compressing complex, high-dimensional input to simple, low-dimensional space i.e., latent space, and 2) a decoder; reconstructing input from representations in latent space. On top of the "auto-encoder" strategy, additional constraints can be added, yielding various models with different behaviors. Among them, the Variational AutoEncoder (VAE) model became one of the most widely used deep generative models thanks to its simple yet powerful theoretical concept and the reliable stability against diverse applications. Additionally, the superior interpretability of latent space defined by VAE has been an attractive property of VAE model. Given that, **in Chapter 2**, I designed a VAE model (Higgins et al., 2017; Kingma and Welling, 2013), for the first time, specialized to learn representations of rsfMRI spatial patterns. I also designed a new reformatting strategy of input of VAE – fMRI data, instead of conventional compressing methods. Through exploring the hyperparameters of β-VAE, I determined the model architecture and hyperparameters, and the VAE model was trained and validated using the population-level fMRI data under the resting state.

In Chapter 3, As a starting example of the VAE model for fMRI data, I applied the VAE model to fMRI data under resting state. I characterized the time-evolving trajectory of latent representations and factorized its gradients by principal components. I also visualized the representational gradients, clusters, and geometries within and across individuals, as a way to characterize brain networks and their dynamic interactions. Lastly, I tested the use of this model for characterizing individual variations and identifying individuals from the population.

In Chapter 4, I further tested the generalizability of the VAE model pretrained in Chapter 3 on fMRI under different brain conditions and different recording parameters. I found the VAE model was highly generalizable to fMRI data when subjects were watching naturalistic movies. Additionally, I found generative factors of task-evoked brain activity and generative factors responsible for spontaneous ongoing brain activity were delineated in the latent space. Based on that, I showed that the trajectory of task-evoked brain activity was dependent on scenic changes in movie stimuli. Lastly, I showed that the positive and negative interactions between brain networks were the principal bases forming the reorganization of the brain during watching movies.

2. BETA VARIATIONAL AUTOENCODER

2.1 Motivation

The central concept of the unsupervised generative model can be viewed as “a task reconstructing the input can be used to understand the generative factors underlying the input”. In VAE, implementation of this idea was done by coupling two parametrized models: an encoder (or recognition model) and a decoder (or generative model). These two models reinforce each other; the encoder tries to pass good approximations of input in terms of latent variables to the decoder, whereas the decoder tries to learn meaningful representations of data under given latent variables (Higgins et al., 2017; Kingma and Welling, 2013). As in the original paper (Kingma and Welling, 2013), the objective of VAE can be simplified as minimizing reconstructing error of input under the constraint making latent variables independent to each other. The defining difference of VAE compared to other autoencoder models comes from this constraint, leading to better interpretability of learned latent variables. Similar to VAE, β-VAE also employs the same objective but adds a hyperparameter (called as β originally) emphasizing the importance of better interpretability over the reconstruction performance (Burgess et al., 2018; Higgins et al., 2017). β-VAE has proved its excellent interpretability of latent variables representing underlying generative factors of simple images (Higgins et al., 2017), and complex images (Burgess et al., 2019) while to our best knowledge, there has been no study utilizing β-VAE as a tool for brain encoding. Thus, we chose β-VAE as our base model of unsupervised representational learning of fMRI cortical patterns.

2.2 Input Structure of β-VAE Model

Instead of limiting the input data to activity at the Region of Interest (ROI) level or network level, we converted the rsfMRI data from 3-D cortical surfaces to 2-D grids in order to structure the rsfMRI pattern as an image to ease the application of convolutional neural networks. As illustrated in Figure 2.1, we inflated each hemisphere to a sphere by using FreeSurfer (Fischl, 2012). For each location on the spherical surface, we used cart2sph.m in MATLAB to convert its cartesian coordinates (x, y, z) to spherical coordinates (a, e), which reported the azimuth and elevation angles in a range from -π to π and from -π/2 to π/2, respectively. We defined a 192×192 grid to resample the spherical surface with respect to azimuth and sin(elevation) such that the resampled locations were uniformly distributed at approximation (Figure 2.2). We used the nearest-neighbor interpolation to convert data from the 3-D surface to the 2-D grid, and vice versa.

2.3 Model Architecture

We designed a β-VAE model to learn representations of rsfMRI. This model included an encoder and a decoder (Figure 2.3). The encoder converted an fMRI map to a probabilistic distribution of 256 latent variables. Each latent variable was a Gaussian random variable with a mean and a standard deviation. The decoder sampled the latent distribution to reconstruct the input fMRI map or generate a new map, which appeared similar to what would be observable with fMRI. The encoder stacked five convolutional layers and one fully connected layer. Every convolutional layer applied linear convolution and rectified its output (Nair and Hinton, 2010). The first layer applied 8×8 convolution separately to the input from each hemisphere and concatenated its output. To the feature maps concatenated across both hemispheres, the 2nd through 5th layers applied 4×4 convolution. Since a spherical pattern is circularly continuous with respect to the azimuth, we applied circular padding to the boundaries of the azimuth for the flattened 2-D map but applied zero paddings to the boundaries of elevation. Such padding was intended to avoid artifacts when applying convolution near those boundaries. The fully connected layer applied linear weighting and yielded the mean and standard deviation that described the normal distribution of each latent variable. The decoder used nearly the same architecture as the encoder but it connected the layers in the reverse order for transformation from the latent space back to the input space. Figure 2.3 illustrates the model architecture.

In our case, the objective of the VAE model aimed to maximize the marginal log-likelihood of the observed fMRI data combined across the left and right hemispheres x over the ground-truth generative factors of z:

$$\max_{\phi,\theta}\ \mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\theta}\,(\boldsymbol{x}|\boldsymbol{z})], \qquad (1)$$

where, ϕ and θ stand for the learnable parameters of the encoder and the decoder, respectively. As desire that inferred latent factors $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ are informative (or disentangled), additional constraint with adjustable hyperparameter β is introduced to force $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ to match to be isotropic unit Gaussian $N(0, I)$. In short, the objective can be re-written as:

$$L(x) = \|x - x'\|_2^2 + \beta \cdot D_{KL}[N(\mu_z, \sigma_z) \parallel N(0, I)], \qquad (2)$$

where x is the, x' is the reconstructed input, $N(\mu_z, \sigma_z)$ is the posterior normal distribution of the latent variables, z, with their mean and standard deviation denoted as μ_z and σ_z, D_{KL} measures the Kullback-Leibler (K-L) divergence between the posterior and prior distributions.

Part of the medial cortical surface that corresponds to the corpus callosum (i.e. white matter) was excluded from training such that the learned model was intended to merely represent the activity of cortical gray matter.

2.4 Data and Training

Here, we used rsfMRI data consisting of 150 healthy subjects randomly chosen from the Q2 release by HCP (Van Essen et al., 2013). For each subject, we used two sessions of rsfMRI data acquired from different days with either the right-to-left or left-to-right phase encoding. Each session included 1,200 time points separated by 0.72s. The imaging protocol of rsfMRI data were followed: gradient-echo echo-planar imaging (EPI) sequence with the following parameters: repetition time (TR) = 720 ms, echo time (TE) = 33.1 ms, flip angle = 52 deg, field of view (FOV) = 208 x 180 mm, matrix = 104 x 90, spatial resolution = 2.0mm^3, number of slices = 72, multiband factor = 8, echo spacing = 0.58 ms, bandwidth = 2290 Hz/Px. Following minimal preprocessing (Glasser et al., 2013) and automatic denoising with ICA (or the ICA-FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014), we applied voxel-wise detrending (regressing out a 3rd-order polynomial function), bandpass filtering (from 0.01 to 0.1 Hz), and normalization (to zero mean and unitary variance). We further separated the data into two sets, including 100 or 50 subjects for training and validating the VAE model, respectively. The validation dataset was used to determine the hyperparameters used in the VAE model. To train the model, we used stochastic gradient descent (batch size=128, initial learning rate=10^{-4}, and 100 epochs) and Adam optimizer (Kingma and Ba, 2014) implemented in PyTorch (v1.2.0). The learning rate was decayed by a factor of 10 every 20 epochs. See Table. 2.1 for the training algorithm.

2.5 Model Parameter

We determined the hyperparameters by exploring and testing different parameter settings with the validation dataset. Specifically, we explored four values (1, 5, 10, 15) for β and chose β=10 to balance the reconstruction performance vs. the disentanglement (or independence) of latent variables (Figure 2.4) – the two terms in the loss function shown in Eq. (2). We also explored several options for the number of layers (# of layers = 6 and 8; full model: 12). Since shallower models were not able to reduce reconstruction loss when β=10, we set β=1 and compared the validation loss as a function of epochs only for the comparison purpose. As expected, shallower models had poorer reconstruction performance compared to one obtained using the full model (Figure 2.5). Moreover, only VAE model with 12 layers was able to reduce both reconstruction loss and D_{KL} when β=10. Collectively, this result suggested the current model architecture having 12 layers is a reasonable choice. Lastly, we explored three values of learning rate (10^{-3}, 10^{-4}, and 10^{-5}) as a trend of validation loss over the progression of training epochs. Given the current setting, we found when the learning rate is too high (=10^{-3}), the model was not able to be converged whereas too low learning rate (=10^{-5}) was stuck in local minima. Therefore, we chose the learning rate as 10^{-4} in the subsequent analysis.

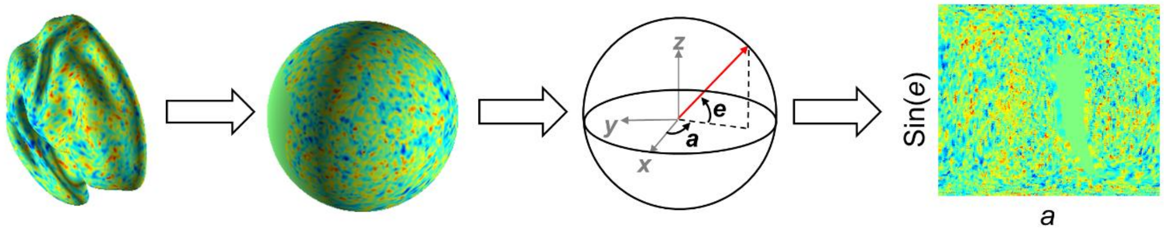


Figure 2.1. Geometric reformatting. The cortical distribution of fMRI activity is converted into a spherical surface and then to an image by evenly resampling the spherical surface with respect to sin(e) and a, where e and a indicate elevation and azimuth, respectively.

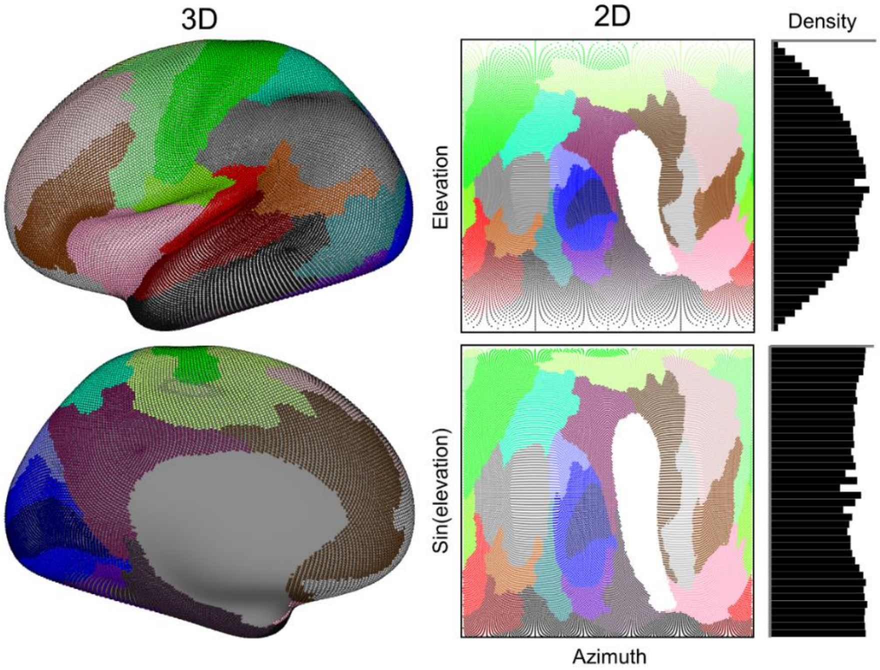


Figure 2.2. Grayordinates (# = 29,696) in the left hemisphere (left panel) is projected into 2D space (middle panel) based on their azimuth- and elevation levels. Each color stands for different brain atlas based on the literature (Glasser et al., 2016). The imbalance in data density along the varying elevation level (upper in right panel) is alleviated by further applying elevation to Sine function (bottom in right panel).

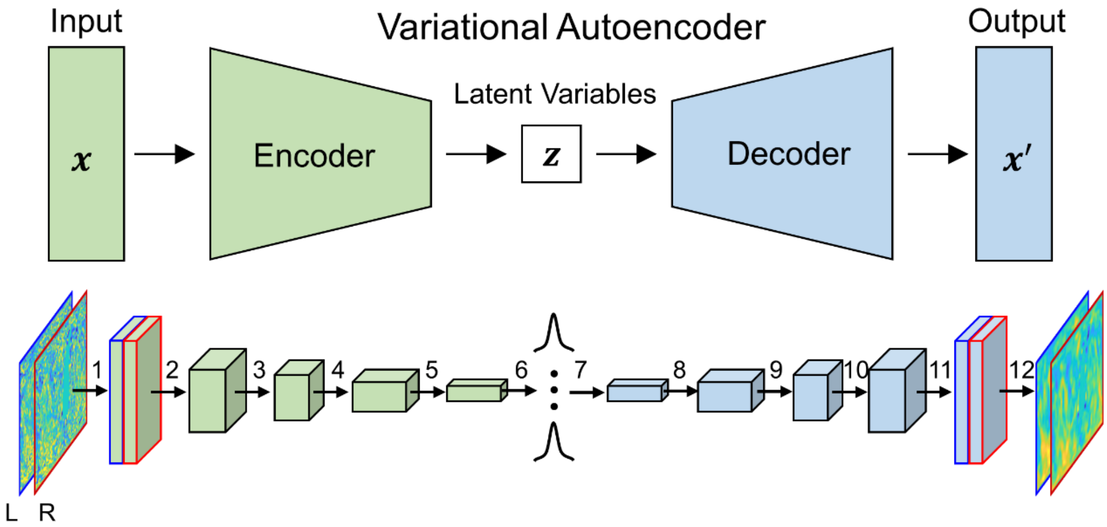


Figure 2.3. Architecture of VAE. Simplified block diagram of VAE model (upper panel in B). An encoder network samples latent variables given an input image under the inference model while a decoder network generates a genuine input image from under the generative model. Details of VAE model (bottom panel in B). Both encoder and decoder network contain 5 convolutional layers. In the encoder network, the size of output image of each layer (from left to right) is 96×96×64 (32 channels per hemisphere), 48×48×128, 24×24×128, 12×12×256, and 6×6x×56; for the decoder network, 6×6×256, 12×12×256, 24×24×128, 48×48×128, and 96×96×64 (32 channels per image), from left to right. The dimension of latent variables is 256. The convolution operations are defined as: 1: convolution (kernel size=8, stride=2, padding=3) with rectified nonlinearity, 2-5: convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 6: fully-connected layer with re-parametrization, 7: fully-connected layer with rectified nonlinearity, 8-11: transposed convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 12: transposed convolution (kernel size=8, stride=2, padding=3). Blue and red boxes stand for the input images from left and right hemispheres, respectively.

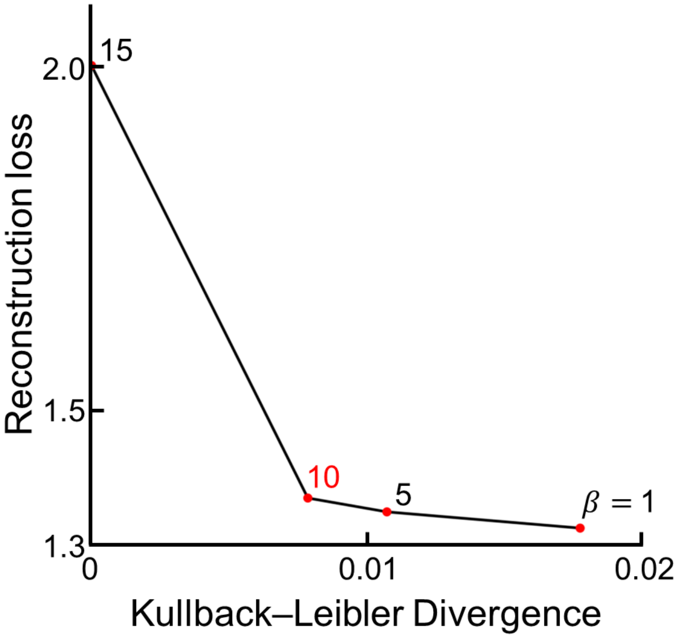


Figure 2.4. Validation error of VAE at varying beta values. Trade-off between reconstruction loss and Kullback–Leibler divergence is visualized for different beta values. Red color stands for the value we used for analysis.

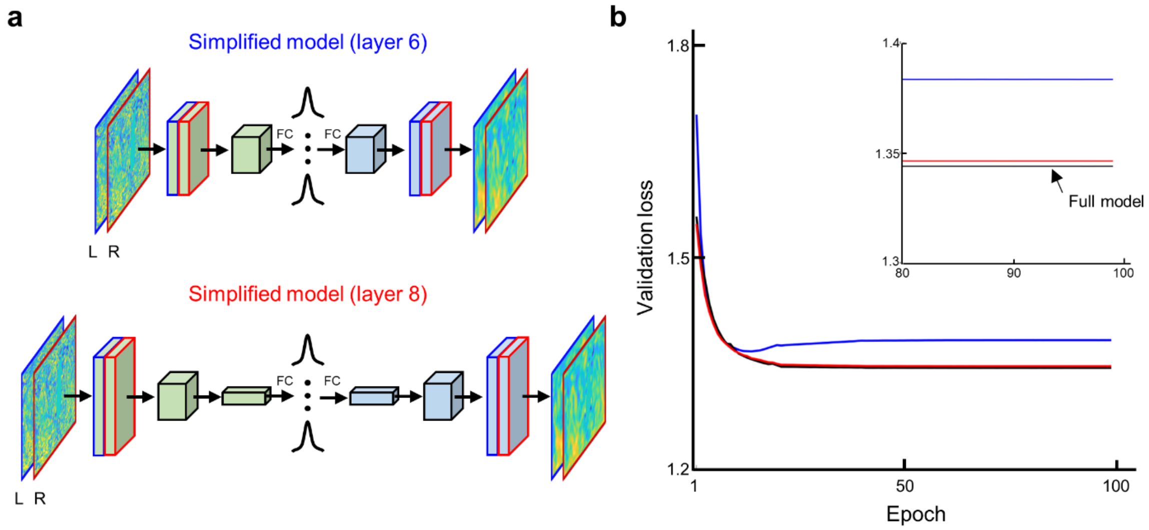


Figure 2.5. Validation curve of VAE at different layers. (a) Layer 6; Both encoder and decoder network contain 2 convolutional layers. In the encoder network, the size of output image of each layer (from left to right) is 48x48x128 (64 channels per hemisphere), and 12x12x256; for the decoder network, 12x12x256, and 48x48x128 (64 channels per image), from left to right. Layer 8; 3 convolutional layers for encoder and decoder networks. The size of output image of each encoder layer is 96x96x64 (32 channels per image), 24x24x128, and 6x6x256; decoder network: 6x6x256, 24x24x128, and 96x96x64 (32 channels per image), from left to right. (b) The validation curve as a function of training epochs.

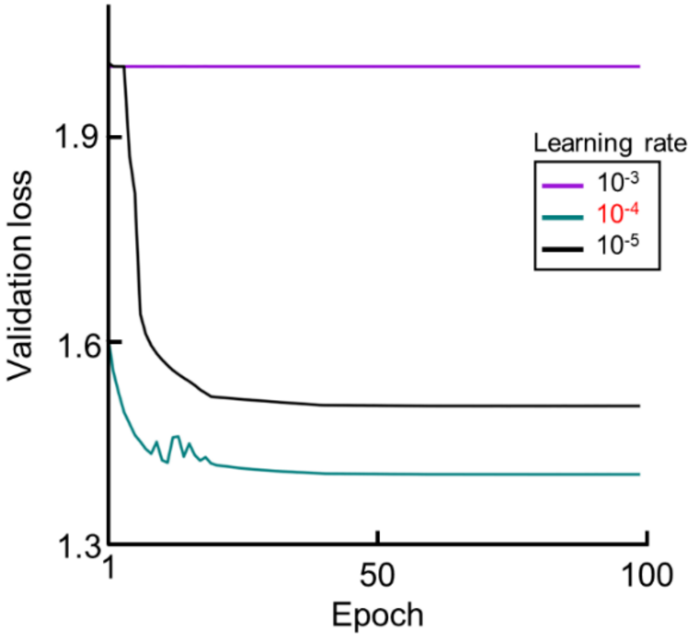


Figure 2.6. Validation curve of VAE at varying learning rate

Table 2.1. Algorithm of β-VAE for learning the representation of fMRI data

Algorithm Beta-VAE for learning representations of fMRI data

Input fMRI data combined across the left and right hemisphere $\mathbf{X}$

```
Randomly Initialize $\theta$, $\phi$
For k = 1 to K epochs do
    Sample a batch $\boldsymbol{B}$ from $\mathbf{X}$
    For $b \in \boldsymbol{B}$ do
        Compute $z_b$ given $b$
        Compute gradient $\nabla_\theta L$ and $\nabla_\phi L$ given $z_b$
    end
    Average gradients from a batch
    Update $\theta$ and $\phi$ using stochastic gradient descent
end
return $\theta$, $\phi$
```

3. REPRESENTATIONAL LEARNING OF RESTING STATE FMRI WITH VARIATIONAL AUTOENCODER

* Modified and formatted for dissertation from the article that have been submitted for review with NeuroImage.

3.1 Introduction

The brain is active even at rest, showing complex activity patterns measurable with resting state fMRI (rsfMRI) (Fox and Raichle, 2007). It is widely recognized that rsfMRI activity is shaped by how the brain is wired, or the brain connectome (Sporns et al., 2005). Inter-regional correlations of rsfMRI activity are often used to report functional connectivity (Biswal et al., 1995) and map brain networks for individuals (Finn et al., 2015) or populations in various behavioral (Smith et al., 2009) or disease states (Fox et al., 2014). However, it remains largely unclear where rsfMRI activity comes from (Leopold and Maier, 2012; Lu et al., 2019), whereas understanding its origins is critical to interpretation of any rsfMRI pattern or dynamics (Winder et al., 2017).

Prior findings suggest a multitude of sources (or causes) for rsfMRI activity (Bianciardi et al., 2009b), including but not limited to fluctuations in neurophysiology (Mantini et al., 2007), arousal (Chang et al., 2016), unconstrained cognition (Chou et al., 2017), non-neuronal physiology (Birn et al., 2008), head motion (Power et al., 2014) etc. These sources only partially account for rsfMRI activity and may be entangled not only among themselves but also with other sources that are left out simply because they are hard to specify or probe in a task-free state (Leopold and Maier, 2012). An inclusive study would benefit from using a data-driven approach to uncover and disentangle all plausible but hidden sources from rsfMRI data itself, without having to presume the sources to whatever are experimentally observable. To be effective, such an approach should be able to infer sources from rsfMRI data and generate new rsfMRI data from sources, while being able to account for complex and nonlinear relationships between the sources and the data.

These requirements lead us to deep learning, or representation learning with deep neural networks (LeCun et al., 2015), as a nonlinear method for blind source separation, in contrast to its linear counterparts, e.g. independent component analysis (Beckmann and Smith, 2004; Calhoun et al., 2001; Smith et al., 2012). For brain research, deep learning models has provided testable models of the brain in terms of neural computation for sensory and language processing (Han et al., 2019; Kell et al., 2018; Khaligh-Razavi and Kriegeskorte, 2014; Richards et al., 2019; Wen et al., 2018a; Yamins and DiCarlo, 2016; Zhang et al., 2020). Deep learning has also been increasingly used as a generic family of machine learning tools to learn features from fMRI data. See (Khosla et al., 2019c) for a review. Most applications are in the regime of supervised learning. Typically, a neural network takes an fMRI-based input data and is trained to generate an output that optimally matches the ground truth for a task, such as individual identification (Chen and Hu, 2018; Wang et al., 2019a), prediction of gender, age, or intelligence (Fan et al., 2020; Gadgil et al., 2020; Plis et al., 2014), disease classification (Seo et al., 2019; Suk et al., 2016; Wang et al., 2020; Yang et al., 2019; Zou et al., 2017). The labels required for supervised learning are often orders of magnitude smaller in size than the fMRI data itself, which has a high dimension in both space and time. As a result, the prior studies often limit the model capacity by using a shallow network and/or limit the input data to activity at the region of interest (ROI) level (Chen and Hu, 2018; Dvornek et al., 2018b; Koppe et al., 2019; Matsubara et al., 2019; Suk et al., 2016; Wang et al., 2019a; Wang et al., 2020) or reduce it to functional connectivity (D'Souza et al., 2019; Fan et al., 2020; Kawahara et al., 2017; Kim and Lee, 2016; Riaz et al., 2020; Seo et al., 2019; Venkatesh et al., 2019; Yang et al., 2019; Zhao et al., 2018). It is also uncertain to what extent representations learned for a specific task would be generalizable to other tasks. It is further debatable whether deep neural networks with supervised learning are currently superior to more conventional and simpler methods (He et al., 2020)

For these considerations, unsupervised learning is more preferable for uncovering the underlying causes that drive intrinsic brain activity regardless of any task or disease. We choose to use the Variational Auto-Encoder (VAE) (Higgins et al., 2017; Kingma and Welling, 2013), for unsupervised learning of the increasing "big data" in rsfMRI without requiring any label or narrowly focusing on any downstream task. Unlike auto-encoder, VAE is a generative model capable of synthesizing new data similar to the training data, and it regularizes the latent space with a priori spherical Gaussian distributions. These properties allow the representation learned to be expressed in terms of latent variables that encode the disentangled causes of the data. Our emphasis on disentangling latent representations sets this work apart from several prior work based on the auto-encoder implemented in various forms of deep neural networks (Cui et al., 2019; Huang et al., 2017; Liu et al., 2020a; Makkie et al., 2019; Suk et al., 2016; Zhao et al., 2018). Briefly in this study, we designed and trained a VAE model to represent rsfMRI data in terms of its latent sources and tested its ability to explain and generate rsfMRI data. We characterized the time evolving trajectory of latent representation and factorized its gradients by principal components. We also visualized the representational gradients, clusters, and geometries within and across individuals, as a way to characterize brain networks and their dynamic interactions. Lastly, we tested the use of this model for characterizing individual variations and identifying individuals from their rsfMRI data (Finn et al., 2015)as a starting example of its applications.

3.2 Methods and Materials

3.2.1 Testing Data

We used rsfMRI data from 500 healthy subjects randomly chosen from the Q2 release by HCP (Van Essen et al., 2013). For each subject, we used two sessions of rsfMRI data acquired from different days with either the right-to-left or left-to-right phase encoding. Each session included 1,200 time points separated by 0.72s. The imaging protocol of rsfMRI data were followed: gradient-echo echo-planar imaging (EPI) sequence with the following parameters: repetition time (TR) = 720 ms, echo time (TE) = 33.1 ms, flip angle = 52 deg, field of view (FOV) = 208 x 180 mm, matrix = 104 x 90, spatial resolution = 2.0mm^3, number of slices = 72, multiband factor = 8, echo spacing = 0.58 ms, bandwidth = 2290 Hz/Px. Following minimal preprocessing (Glasser et al., 2013) and automatic denoising with ICA (or the ICA-FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014), we applied voxel-wise detrending (regressing out a 3rd-order polynomial function), bandpass filtering (from 0.01 to 0.1 Hz), and normalization (to zero mean and unitary variance). The testing data were neither seen nor used by the model during training or validation. This held-out data was used to test the generalizability of the model across different datasets. For an exploratory analysis, we additionally tested the model with rsfMRI data that did not go through denoising with ICA-FIX to evaluate the model performance against presumably noisier rsfMRI data.

3.2.2 Synthesizing Resting State fMRI Functional Connectivity

We used the trained VAE to synthesize rsfMRI data from random samples of latent variables. To synthesize a vector in the latent space, we drew a random sample of every latent variable independently from a standard normal distribution. The synthesized vector passed through the decoder in VAE, generating a cortical pattern. Repeating this process, we synthesized 12,000 cortical patterns as data used for seed-based correlation analysis. As examples, we explored three seed locations within primary visual cortex (V1), intraparietal sulcus (IPS), and posterior cingulate cortex (PCC) and calculated the functional connectivity to each seed based on the Pearson correlation coefficient. The MNI coordinates of the seed in V1, IPS, and PCC were (7, -83, 2), (26, -66, 48), and (0, 57, 27), respectively (Jarrett, 2009). In addition, we performed a similar analysis without limiting to the seed locations. Instead, we calculated the functional connectivity between each pair of parcels as defined in a 360-parcel atlas of the whole cortex (Glasser et al., 2016).

For comparison, we similarly calculated seed-based or parcel-to-parcel functional connectivity with experimental rsfMRI data concatenated across a varying number (1, 5, 10, 50, and 100) of subjects in HCP. We compared the functional connectivity pattern observed with synthesized and experimental data and repeated the comparison 20 times. At each time, we generated a different set of synthesized data while using experimental data from a different subset of subjects. The comparison was thus randomly repeated.

3.2.3 Defining a Principal Basis Set in the Latent Space

By our design, the VAE model encodes the spatial pattern of fMRI activity and does not represent the temporal dynamics explicitly. The distribution of every latent variable is constrained to be close to a standard normal distribution independent of one another, for the K-L divergence term in the loss function in Eq. (2). This implies that the latent variables in the VAE model are not unique. An arbitrary rotation of a tentative set of latent variables would arrive at a new set of latent variables that span the same latent space and satisfy the same learning objective.

To identify a unique set of latent variables, we exploited the trajectory of the latent representation. Specifically, for the fMRI data in the testing set (concatenated across 500 subjects), we encoded the fMRI pattern observed at every time into a point (or vector) embedded in the latent space. As time progressed, this point moved in the latent space along a trajectory that represented the temporal dynamics of fMRI activity.

In a first-order differential analysis, we evaluated the displacement (or difference) of the latent representation from every time point to its next. To this time-difference vector (or the latent gradient), we further applied singular vector decomposition and used the singular vectors to define a unique basis set of the latent space. As such, each singular vector was a re-defined latent variable, while the corresponding singular value indicated its importance in explaining the latent gradient of cortical activity. In other words, the trajectory was more likely to move along the direction represented by a singular vector with a larger singular value than that with a smaller singular value.

We further interpreted and visualized the top-10 latent variables defined as the singular vectors with the largest 10 singular values. For this purpose, we decoded each of these latent variables onto the cortical surface by using the decoder in the VAE model. Note that the latent variables were related to cortical patterns through nonlinear functions. We evenly sampled each latent variable of interest from -5 to 5, while keeping other latent variables to zero. We mapped the decoded cortical pattern and characterized its variation due to the variation of a single latent variable. We quantified the variation separately for each cortical location in terms of the standard deviation multiplied by a sign. The sign of standard deviation map was determined by measuring the sign of Pearson correlation coefficient between the decoded values of each cortical location and the samples of the given latent variable.

3.2.4 Clustering in the Latent Space

We encoded the rsfMRI spatial pattern at every time point for 500 testing subjects, yielding 600,000 vectors in the latent space. We used k-means clustering with 1-cosine distance (based on "kmeans" in Matlab) to group those vectors to 21 clusters. The choice of k=21 was empirical but made intentionally to be consistent to a prior study with a similar motivation (Smith et al., 2012). This choice was within a reasonable range for the number of resting state networks as reported in literature (Smith et al., 2009; Yeo et al., 2011). Beyond this single choice, we explored other numbers of clusters to ensure that k=21 was a reasonable choice for the distribution of latent representation. Specifically, we varied k from 1 to 100. Given each choice, we ran the k-means clustering for the testing data (n=500 subjects), identified the clusters, calculated the centroid of each cluster, and summed the distance to the centroid within every cluster. We further plotted the sum of distance as a function of k and ensured that k=21 was around the "elbow" of the plot, as a useful, but not strict, rule of thumb.

Given k-means clustering with k=21, we re-ordered individual time points by their cluster membership and compared the distance between time points within and between clusters. We also evaluated the relationships among different clusters by calculating the distance between the centroids of individual clusters and we further grouped clusters into super-clusters organized in a multi-level hierarchy visualized as a dendrogram (by using “linkage” in Matlab).

To visualize and interpret each cluster, we further converted the cluster centroid to a corresponding cortical pattern by using the VAE’s decoder. The resulting cortical pattern was scaled such that its maximal absolute value equaled 1. This pattern was considered as a functional cortical network. To further evaluate how each cortical network changed its activity in time, we defined and evaluated the cluster-wise activity as the cosine affinity between the centroid of each cluster and the latent representation of fMRI activity at every time. As such, a cluster increased its activity when the latent representation moved toward the centroid of that cluster or decreased its activity when the representation moved away from that centroid but towards the centroid of another cluster. After this analysis was done separately for each session and subject, we averaged the cluster-wise activity across subjects. Then we compared the group-level activity between sessions and across clusters, and tested the statistical significance with a non-parametric permutation test (false discovery rate $q<0.01$), for which the time points were randomly shuffled for 10,000 trials to yield a null distribution.

3.2.5 Individual Variation

To evaluate the individual variation, we compared the latent representations of the fMRI data from different individuals. In an exploratory analysis, we randomly selected a small ($n=20$) subset of subjects. We chose 20 subjects to ease visualization and intuitive demonstration, before scaling up the analysis to 500 subjects. For each of the 20 subjects, we converted the fMRI activities, instance by instance, to the representations in the latent space. To visualize and compare subject-wise representations, we used the t-distributed Stochastic Neighbor Embedding (t-SNE) method to visualize the 256-dimensional latent representations (color-coded by subjects) in a two-dimensional space. We calculated the Silhouette index to measure how similar the latent representation was within the same subject vs. between different subjects.

3.2.6 Subject Identification

After the exploratory analysis above, we evaluated the individual variation across $n=500$ subjects. For the distribution of subject-wise latent representation, the first moment was the mean and the second moment was the covariance, which indicated the location and geometry of the subject-wise latent representation, respectively. We tested the use of the first moment (mean) or the second moment (covariance) as the subject-identifying feature.

In the testing data set, every individual had rsfMRI data acquired for two separate sessions. From the first session, we extracted the feature from every subject and stored it as the subject-identifying key in a database that included a population of 500 subjects. Given this database, we tested the accuracy of retrieving any subject's identity by using the feature extracted from the second session as a query to match against all keys in the database. The goodness of match was evaluated as the cosine similarity or the Pearson correlation coefficient when the query and the key were based on the first moment (mean) or the second moment (covariance) of the subject-wise representation, respectively. The accuracy of individual identification was evaluated as the percentage by which the correct identity was retrieved as one of the best 1, 5, or 10 matches, yielding the namely top-1, 5, or 10 accuracy.

For comparison, we compared the performance of individual identification based on the above latent-space feature vs. the similar feature evaluated in the cortical space. The cortical-space features extracted with a similar method as previously reported in (Finn et al., 2015). Specifically, the FC between brain regions (or connectome) was calculated as features for individual identification. It is worth noting that the cortical connectome and covariance of latent representation, although they are nominally different terms, can both be viewed as the representational geometry of brain activity in the cortical space (for the connectome) or the latent space (for the covariance of latent representation). In addition, we may also cast both notions as the functional connectivity profile in the cortical space or the latent space. Given such conceptual connections, we evaluated the FC between every pair of 360 cortical parcels defined in an established atlas (Glasser et al., 2016) and used the FC-based connectome as the feature for individual identification (Finn et al., 2015). We compared the connectome-based identification accuracy with that based on the FC profile (or representational geometry) in the latent space for a varying population size (from n=5 to 500 subjects) or a varying length of data per subject (from 9 to 180 s). We repeated the above analysis 100 times, each time with a different subset of the testing data and averaged the identification accuracy across the repeated tests.

3.2.7 Comparison with Linear Latent Space

The VAE model described herein provided nonlinear mapping from the cortical space to the latent space (through the encoder) and in reverse (through the decoder). Such reversible mapping could be conventionally done through linear matrix operations, such as the principal component analysis (PCA) and independent component analysis (ICA). Hence, we compared the distribution and geometry of the rsfMRI representation in the nonlinear latent space obtained with VAE vs. the linear latent space obtained with PCA or ICA. For such comparison, we used PCA or ICA trained with the training data to represent rsfMRI data in the testing dataset, while keeping the linear latent space of the same (256) dimension as its nonlinear counterpart. We compared the performance of reconstructing fMRI patterns from their latent representations (see results in Figure 3.2). In addition, we also compared PCA or ICA vs. VAE for characterizing individual variation or performing individual identification by using the representation in the PCA or ICA-derived linear latent space for the same analyses as used for the representation in the VAE-based nonlinear latent space (see results in Figs. 3.6 and 3.8).

3.3 Results

3.3.1 VAE Compressed Resting State fMRI Maps

Inspired by its success in artificial intelligence (Higgins et al., 2017; Kingma and Welling, 2013), we designed a VAE model in order to disentangle the generative factors underlying rsfMRI activity. The model was trained to represent and reconstruct rsfMRI data with a set of latent variables that were constrained to be as independent as possible. The hyper-parameter, β, which expressed the weighting of independence among latent variables in the overall learning objective, was initially explored for different values (1, 5, 10, 15) before being finalized to β=10 – a setting that led to a reasonable trade-off of the model performance vs. constraint as demonstrated with the validation dataset (Figure 2.4).

The model used a pair of convolutional and deconvolutional neural networks in an encoder-decoder architecture (Figure 2.3). The encoder transformed any rsfMRI pattern, formatted as an image on a regular 2D grid (Figure 2.1), to the probability distributions of 256 latent variables. The decoder used samples of the latent variables to reconstruct or generate an fMRI map. Using data from HCP (WU-Minn HCP Quarter 2) (Van Essen et al., 2013), we first trained the model with rsfMRI maps from 100 subjects and then tested it with rsfMRI data from 500 other subjects.

After being trained, the model could compress any fMRI map to a low-dimensional latent space and restore the map from the latent representation separately for every time point (Figure 3.1). The compression resulted in spatial blurring comparable to the effect of spatial smoothing with 4-6 mm full width at half maximum (FWHM) (Figure 3.2). Given fMRI data spatially smoothed to a varying extent (FWHM from 1 to 10 mm), VAE showed either comparable or better performance of reconstruction than its linear counterparts (PCA and ICA), when VAE, PCA, and ICA all used the same dimension (256) for their latent spaces (Figure 3.2.a). The difference in reconstruction performance between VAE and PCA or ICA was marginal but statistically significant (repeated measures ANOVA followed by post-hoc paired t-test, false discovery rate $q<0.05$), for all smoothing levels except FWHM=1 mm (Figure 3.2.b). These results suggest that the latent representation obtained with VAE preserved the spatial and temporal characteristics of rsfMRI, despite a modest but acceptable loss in spatial resolution and specificity.

3.3.2 VAE Synthesized Correlated fMRI Activity

We asked whether the decoder in the VAE, as a generative model of fMRI activity, had learned the putative mechanisms by which rsfMRI activity patterns arise from brain networks. To address this question, we randomly sampled every latent variable from a standard normal distribution and used the decoder to synthesize 12,000 rsfMRI maps (equivalent to 10 subjects at 1,200 time points per subject).

We calculated the seed-based correlations by using the VAE-synthesized data and compared the resulting maps of correlations with those obtained with rsfMRI data concatenated across a different number of subjects. Figure 3.3.a shows three examples with the seed region in the primary visual cortex (V1), intraparietal sulcus (IPS), or posterior cingulate cortex (PCC). For each of the three seed locations, the synthesized fMRI data showed a similar correlational map as that based on length-matched rsfMRI data obtained from 10 subjects (Figure 3.3.a), and the correlational map was consistent with the literature (Yeo et al., 2011). The measured FC patterns were more similar to the synthesized FC patterns, when the measured FC was based on data from increasingly more subjects, regardless of whether the FC was evaluated and compared with respect to a specific seed location (Figure 3.3.b) or across all cortical parcels (Figure 3.3.c). These results suggest that the VAE provided a computational account for the generative process of resting state activity and could synthesize realistic rsfMRI activity patterns and preserve inter-regional correlations as are experimentally observable at a group or population level. However, it is worth mentioning that the temporal ordering of the synthesized data is not meaningful, since the VAE model does not explicitly model the temporal dynamics.

3.3.3 Latent Variables Reflected Network Dynamics

We also examined the time-evolving trajectory of the latent representation and re-defined the latent variables such that they reflected the dynamic changes of fMRI activity. We first evaluated the displacement of the latent representation from every time point to its next. Then we applied singular value decomposition and used the resulting singular vectors to redefine the latent variables as a new basis set that spanned the latent space. These redefined latent variables, ranked in a descending order by their singular values, represented the principal directions in which the instantaneous latent representation tended to move along its time-evolving trajectory.

We chose the top-10 latent variables for further visualization and interpretation. For each latent variable, we uniformly sampled its value in a range from -5 to 5 and visualized each sample by decoding it to a cortical pattern. We found that as the latent variable increased its value linearly, the decoded cortical pattern changed in a non-linear way that differed across cortical locations (see illustrative examples in Figure 3.4.b). To visualize how each latent variable controlled the activity at each cortical location, we calculated the standard deviation of the voxel-wise activity change given an increasing value for the given latent variable and multiplied by a sign (+1 or -1) depending on whether the activity tended to increase or decrease as the latent variable increased. For example, the 1st latent variable was visualized as a cortical pattern that resembled the default mode network (Figure 3.4.a). Using the same visualization method, we found that the 2nd through 10th latent variables all corresponded to distinct but partially overlapping cortical patterns (Figure 3.4.d). However, the top-10 latent variables were found to be inadequate to explain the dynamics of latent representations. The percentage of the variance explained by each latent variable was around 1% or less, and the total variance collectively explained by top 10, 20, 50, and 100 latent variables were 9.6, 17.5, 37.6, and 62.3% (Figure 3.4.c). These results suggest that the dynamics of rsfMRI is complex and high-dimensional in nature. Nevertheless, the latent variables derived from the above analysis represent distinctive factors that drive the dynamic change in resting state activity.

3.3.4 Clusters in the Latent Space

We further characterized the distribution of latent representation and attempted to identify clusters in the latent space. We used the VAE to encode the rsfMRI pattern observed at every time point from 500 subjects, clustered the time points by applying k-means clustering (k=21) to the latent representations, and decoded the cluster centroids to corresponding cortical maps. The number of clusters (k=21) was close to the "elbow" indicative of a reasonable balance between reducing variation within clusters and avoiding too many clusters (Figure 3.5.a). The 1-cosine dissimilarity between latent representations reordered by their cluster membership shows not only close affinity within every cluster but also a varying level of affinity between different clusters (Figure 3.5.b). This motivated us to hierarchically merge clusters into super-clusters (or "clusters of clusters") based on the cosine affinity between cluster centroids (Figure 3.5.c).

For each of the 21 clusters, we decoded and visualized the cluster centroid as a cortical pattern as shown in Figure 3.5.d. Among the 21 clusters, 5 clusters (Cluster 4, 6, 11, 12, 18) showed activity increase (positive) at one or multiple regions in the default mode network (Buckner et al., 2008; Greicius et al., 2003; Raichle et al., 2001), alongside activity decrease (negative) at other regions. Similarly, we found 5 clusters with activity increase in the so-called frontoparietal control network (Cluster 8) (Dixon et al., 2018), cingulo-opercular network (Cluster 7 and 9) (Dosenbach et al., 2007), cognitive control network (Cluster 1) (Cole and Schneider, 2007), and dorsal attention networks (Cluster 10) (Fox et al., 2006) – collectively referred to as "the task positive network" (Fox et al., 2005). In addition, cluster 13 and 16 showed activity decrease in the whole brain, thereby a signature of global signal fluctuation (Murphy et al., 2009; Schölvinck et al., 2010; Wen and Liu, 2016). Cluster 5 and 17 showed widespread synchrony across sensory systems. Cluster 2 and 21 showed the networks for sensorimotor control of the limbs and of the mouth, pharynx, and visceral organs, respectively. Whereas most clusters were bilaterally symmetric, Cluster 15 and 3 were unilateral to the right and left prefrontal cortex, respectively. A common observation for many clusters was that a cluster could highlight the positive interactions among a set of well-defined cortical regions alongside their negative interactions with a different set of regions. Given the above interpretation of individual clusters, we further interpreted the three super-clusters as sensorimotor, default mode, and task positive networks (Figure 3.5.c).

In addition, we evaluated the temporal dynamics of latent representation in terms of the dynamics of individual clusters or their corresponding cortical networks. Intuitively, we considered the time-evolving trajectory of the latent representation as the movement towards or away from each cluster. In this regard, we expressed the cluster-wise activity as the time series of cosine affinity between the instantaneous latent representation and the cluster centroid. During a rsfMRI session, different clusters expressed similar activity levels (Figure 3.5.e), except in the initial period of the session. In that period of 20 seconds, clusters presumably related to task positive networks showed a transition from a high activity level to a lower steady state; the clusters related to sensorimotor networks showed a transition from a low activity level to a higher steady state; in contrary, clusters related to the default mode network remained roughly unchanged. These (somewhat incidental) observations were consistent and reproducible across individuals and sessions. On one hand, this result suggests that the first 20 seconds in a rsfMRI session are not necessarily the steady state under a resting condition. On the other hand, this exploratory analysis shows the feasibility of using the VAE-extracted latent representations to identify brain networks and reveal their individual dynamics.

3.3.5 Individual Variation of Latent Representation

Whereas the aforementioned analyses focused on the group-level characteristics of the latent representations, we further asked how the distribution and geometry of latent representation varied across individuals. Only for the sake of demonstration, we randomly selected 20 subjects in the testing dataset and visualized their individual representations in the latent space after reducing its dimension from 256 to 2 by using t-SNE (Figure 3.6.a). Strikingly, the latent representations were grouped by and separable across individuals. The clustering by individuals was noticeable in the nonlinear latent space obtained with VAE (Figure 3.6.a), but not in the linear latent space obtained with PCA (Figure 3.6.b). Such distinctions were quantitatively confirmed (Figure 3.6.c) by using the Silhouette value to measure the degree of clustering by individuals. The Silhouette value for VAE (mean ± std: s = 0.044±0.002, 50 bootstrapping trials) was significantly higher ($p<0.001$, two-sample t-test) than that for PCA (s = -0.020±0.015). Using the center of latent representation as the subject-identifying feature, we found that subject identity could be retrieved with a reasonably high accuracy when the latent representation was extracted by VAE, whereas the linear representation by PCA failed the same task nearly entirely (Figure 3.6.d). These results suggest the feasibility of using VAE to characterize and reveal individual variations of resting state activity in the non-linear latent space.

3.3.6 Individual Identification

From the t-SNE based visualization (Figure 3.6.a), it was noticeable that subject-wise representations exhibited different geometries. Some were more elongated or scattered than others. This observation motivated us to ask whether the representational geometry (Kriegeskorte and Kievit, 2013) could be an individual-specific feature (or "fingerprint") to allow for more accurate individual identification. Specifically, we calculated the covariance between every pair of latent variables and assembled the pair-wise covariance into a vector as the feature of the representational geometry and evaluated the similarity in this feature between two sessions within or between subjects. The representational geometry evaluated in this way could be interpreted as the functional connectivity (FC) between latent variables. This interpretation related this approach to a conceptually similar approach: the "connectome-based fingerprinting" (Finn et al., 2015; Venkatesh et al., 2020), in which the functional connectivity was evaluated between cortical parcels. So, we evaluated the use of either the latent-space or cortical-space FC for individual identification in comparison.

As shown in Figure 3.7.a, FC between any pair of cortical areas was mostly positive (mean ± std of z-transformed correlation: z=0.26±0.3) and highly reproducible not only within the same subject (r=0.66) but also between different subjects (r=0.45). On the other hand, FC between latent variables had both positive and negative values (mean ± std of covariance: σ^2=0.00±0.13) and its reproducibility was high only within the same subject (r=0.33) but not between different subjects (r=0.07). The FC profile was more distinctive across subjects when it was evaluated between latent variables rather than cortical areas (Figure 3.7.b). In the latent space, the FC profile was significantly more consistent within a subject than between subjects (two-sample t-test, t(249,998)=254.05, two-sided p<0.001). The distribution of within-subject correlations was in nearly complete separation from that of between-subject correlations (Figure 3.7.b, bottom).

Then we compared the performance of individual identification on the basis of the FC profile in the latent vs. cortical space. To identify 1 out of 500 subjects, we compared a target subject's FC profile in the 1st session (as a query) against every subject's FC profile in the 2nd session (as a key) and chose the best match between the query and the key in terms of the Pearson correlation coefficient. As such, the choice was correct if the correlation with the target subject was higher than the largest correlation with any non-target subject. We found that the FC profile in the cortical space could support 69.3% top-1 accuracy while identification was often made with marginal confidence relative to the decision boundary (Figure 3.7.c). Using the FC in the latent space allowed us to reach 97.8% top-1 accuracy. The evidence for correct identification was apparent with a large margin from the decision boundary (Figure 3.7.d). The use of FC in the latent space supported reliable and robust performance in top-1 identification given an increasingly larger population (Figure 3.7.e) or when the data were limited to a short duration (Figure 3.7.f), being notably superior to the use of FC in the cortical space.

We further tested to what extent the performance of individual identification relied on the use of ICA-FIX to preprocess and denoise the rsfMRI data. For this purpose, we applied ICA-FIX to one or both of the two sessions in every subject and then tested the individual identification with n=500 subjects. As shown in Table 3.1, when the FC profile in the latent space was derived from the (ICA-FIX denoised) clean data for both the keys and queries, the identification has the highest accuracy (97.5%). When the key and the query were both based on noisy data (without denoising), the accuracy dropped to 91.3%. When the key and the query were unpaired as denoising applied to one but not the other, the accuracy further dropped to about 88%. Nevertheless, this performance obtained with the latent-space FC was still notably higher than the performance based on the cortical-space FC. For the latter, the use of unpaired preprocessing for the query and the key significantly dropped the identification performance from 69.3% to 47.5%. Counter-intuitively, when the denoising was applied to neither the query nor the key, the identification accuracy with the cortical-space FC increased to 76.9%, but still significantly lower than the accuracy of 91.3% obtained with the latent-space FC.

Lastly, we explored whether the representational geometry (based on the profile of the covariance between latent variables) would yield a similar level of distinction across individuals for linear latent spaces obtained with PCA or ICA. As shown in Figure. 3.8, PCA or ICA was not as effective as VAE. The top-1 accuracy of individual identification was 61.1% for PCA, 63.6% for ICA, in contrast to 97.8% for VAE. The within-subject vs. between-subject similarity in the geometry of linear representation obtained with PCA or ICA exhibited largely overlapping distributions, whereas the corresponding distributions were separated nearly completely for the nonlinear representations obtained with VAE.

3.4 Discussion

Here, we present a method for unsupervised representation learning of cortical rsfMRI activity. Our results suggest that this method is able to disentangle generative factors underlying spontaneous brain activity, discover overlapping brain networks, capture individual characteristics or variation, and support accurate individual identification. We expect this method to be a valuable addition to the existing tools for investigating the origins of resting state activity, mapping functional brain networks, and potentially supporting individualized prediction of disease phenotypes and progression. Next, we discuss our findings from the joint perspective of methodology, neuroscience, and applications.

VAE is trainable with unsupervised learning (without any label) (Higgins et al., 2017; Kingma and Welling, 2013), which is appealing for learning representations of rsfMRI data. Since rsfMRI measures spontaneous brain activity unconstrained by any task, labels as required for supervised learning are either unavailable or far fewer than the data itself. Unsupervised learning with VAE can leverage the ever-increasing amount of rsfMRI data (Van Essen et al., 2013). The latent representations extracted from VAE can serve as the input to other algorithms to further support more specific goals such as classification of brain disorders and prediction of their phenotypes (Garrity et al., 2007; Moradi et al., 2015; Shen et al., 2010; Zhang et al., 2011).

The method herein can be extended in multiple ways. Although it is trained with rsfMRI data, we hypothesize that the VAE model can encode and decode both rsfMRI and task-fMRI data but with different latent distributions. If this is true, one may use this model to classify different perceptual, behavioral, or cognitive states and to reveal the distinctive network interactions underlying various states (Gonzalez-Castillo et al., 2015). The fact that the VAE can synthesize new data (Figure 3.3) is also appealing. It can be used as a post-processing strategy for data augmentation and interpolation, when data is short or corrupted, of interest for evaluation of dynamic functional connectivity (Allen et al., 2014; Chang and Glover, 2010) and correction for head motion (Power et al., 2014). It also supports the notion that the learned latent space captures the origins of rsfMRI and the VAE decoder captures the computational account for how rsfMRI arises from its origins.

It is worth mentioning two limitations of the VAE model in its current form. First, the model focuses on cortical patterns but excludes sub-cortical and white-matter voxels. This design is not only for the ease of model implementation but also for the predominant role of the neocortex in brain functions (Rakic, 2009). However, this precludes the model from accounting for subcortical networks or their interactions with the cortex. Addressing this limitation awaits future studies to redesign the model as a 3-D neural network that takes volumetric fMRI data as the input. Second, the VAE model only represents spatial patterns but ignores temporal dynamics inherent to rsfMRI data. Modeling the temporal dynamics is desirable but non-trivial, since it is highly irregular, complex and variable. To fill this gap, we direct future studies to designing a recurrent neural network (Chen and Hu, 2018; Cui et al., 2019; Shi et al., 2018; Sutskever et al., 2014; Zhao et al., 2019), as an add-on to VAE, to further learn sequence representation, for example, with a self-supervised predictive learning strategy (Kashyap and Keilholz, 2020; Khosla et al., 2019b).

Although VAE does not explicitly model the temporal dynamics, the representation obtained with VAE preserves the temporal dynamics (Figure 3.1). The trajectory of the latent representation describes the temporal behavior of brain networks, as opposed to voxels or regions. This trajectory is amenable to the use of many methods previously described for voxel-wise or region-wise analysis. To note a few examples explored in this study, the first-order temporal difference in the latent representation captures the gradient of latent trajectory that drives the brain to change its activity pattern from one time point to the next. As the latent gradient is also represented as a vector in the latent space, the length of this vector measures the displacement in the latent space and presumably the magnitude of network activity, and the direction of this vector encodes a pattern of network interaction that drives the instantaneous change of brain activity. The principal components of the displacement in representation uncover the important hidden factors that drive the temporal dynamics of brain networks (Figure 3.4). Similar analysis or notion has also been explored in two independent studies discussed in two very recent papers published or in preprint during the peer review of our paper (Brown et al., 2020; Liu et al., 2020b). These initial analyses are expected to merit and direct future studies upon predictive modeling of the trajectory of the VAE-derived latent representation, for example, by using Multivariate Auto-Regressive models (Liégeois et al., 2019; Rogers et al., 2010), Hidden Markov Models (Eavani et al., 2013; Suk et al., 2016).

VAE provides a new tool for mapping overlapping functional networks in the brain. A brain region may be involved in multiple networks each supporting a distinctive function (Liu and Duyn, 2013; Smith et al., 2012). However, existing network analyses still tend to group brain regions into non-overlapping networks (Yeo et al., 2011). VAE allows us to discover overlapping networks as clusters in the latent space spanned by independent latent variables. As such, VAE is conceptually similar to temporal ICA (Smith et al., 2012) but allows for nonlinear relationships between latent variables and the input data they represent (Khemakhem et al., 2019). Arguably, finding clusters in the low-dimensional latent space is more desirable than doing so in the higher-dimensional voxel space (Liu et al., 2013). Not only is it more computationally efficient, but representations are also more disentangled in the latent space than in the voxel space to readily reveal the underlying organization. However, it is not readily straightforward to attribute a cluster in the latent space to a distinct brain state (Hutchison et al., 2013a) or an individual (Xie et al., 2018). Both are plausible. Our results show that individual variation manifests itself as the latent representation is in part clustered by subject (Figure 3.6), suggesting individual variation is a contributing factor to the clustering of latent representation. Our results also suggest that cluster-wise activity shows a consistent pattern across all subjects, in particular for the first 20 seconds of each session (Figure 3.5). Moreover, the clusters seem to group themselves hierarchically into presumably functional domains: sensorimotor, default-mode and task-positive networks (Figure 3.5). Together These results lead us to speculate that variation in brain states and individuals both contribute to the clustering of brain activity in the latent space. It is challenging to fully separate them and awaits future studies.

Central to this study is the efficacy of using VAE to disentangle what causes resting state activity. In the VAE model, the sources are the latent variables; the decoder describes how the sources generate the observed activity; the encoder models the inverse inference of the sources from the activity. Since the latent variables are data-driven, it is currently unclear how to interpret them as specific physiological processes, many of which are not observable. Nevertheless, we expect the latent variables extracted by VAE to provide the computational basis for further understanding the origins of resting state activity. We hypothesize that the truly disentangled physiological origins, whether observable or not, are individually describable as the latent variables up to linear and sparse projection. This hypothesis awaits confirmation by future studies.

In the latent space, functional connectivity between latent variables describes the geometry of the latent representation of rsfMRI activity. This is a new perspective different from the functional connectivity among observable voxels, regions or networks (Biswal et al., 1995; Yeo et al., 2011). If the VAE model has fully disentangled the sources in a population level, functional connectivity should be near zero between different latent variables and thus reflect a spherical geometry. In other words, the model sets a nearly null population-level baseline, against which individual variation stands out. The latent-space functional connectivity given data from a single subject becomes a unique feature of that subject. Supporting this notion, the use of functional connectivity in the latent space allows for a significantly improved accuracy, robustness, and efficiency in individual identification, compared to the use of functional connectivity among cortical parcels (Amico and Goñi, 2018; Byrge and Kennedy, 2019; Finn et al., 2015; Mejia et al., 2018; Venkatesh et al., 2020).

Note that our main purpose is not to push for a higher identification accuracy but to understand the distribution and geometry of data representations in the feature space. Therefore, we opt for minimal preprocessing and the simplest strategy for individual identification. There is room for methodological development to further improve the identification accuracy or to extend it for many other tasks, including classification of the gender or disease states, prediction of behavioral and cognitive performances, to name a few examples. We expect that such applications would be fruitful and potentially impactful to cognitive sciences and clinical applications.

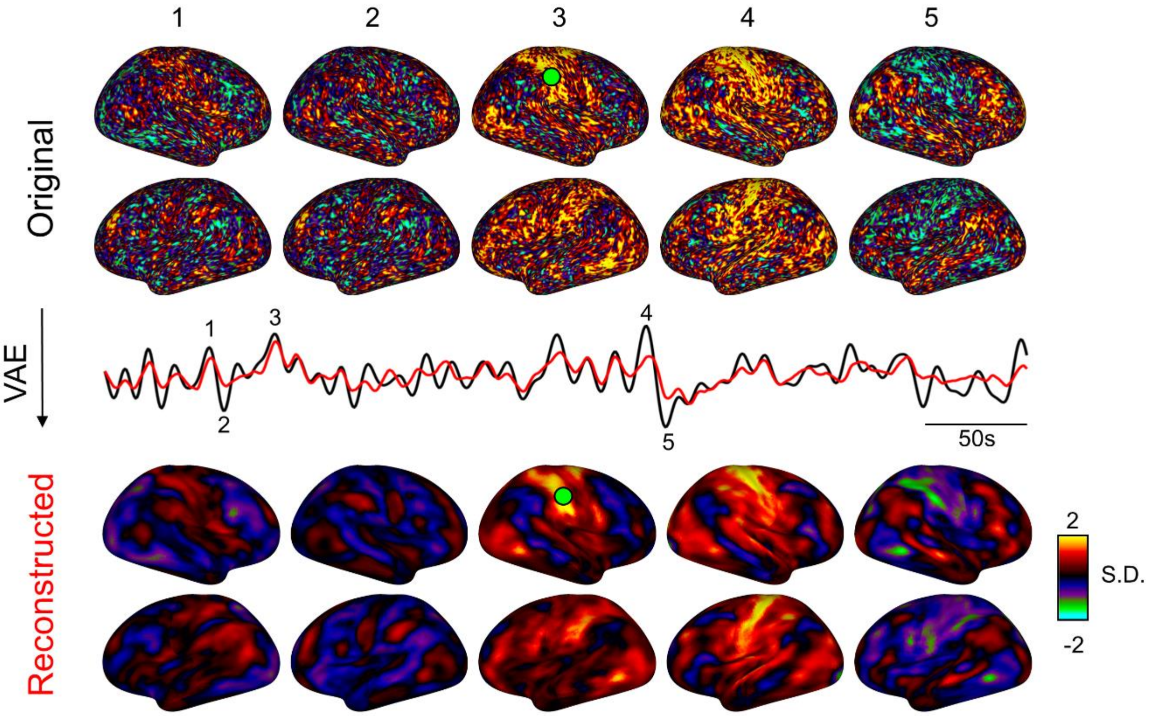


Figure 3.1. Image reconstruction using VAE. A series of cortical patterns are reconstructed through the VAE model. Among them, five original cortical patterns (upper panel) and their corresponding reconstruction through VAE (bottom panel) are visualized for comparison. For an example region (green circle), the time series of the original activity (black line) and the reconstructed activity (red line) are plotted for comparison.

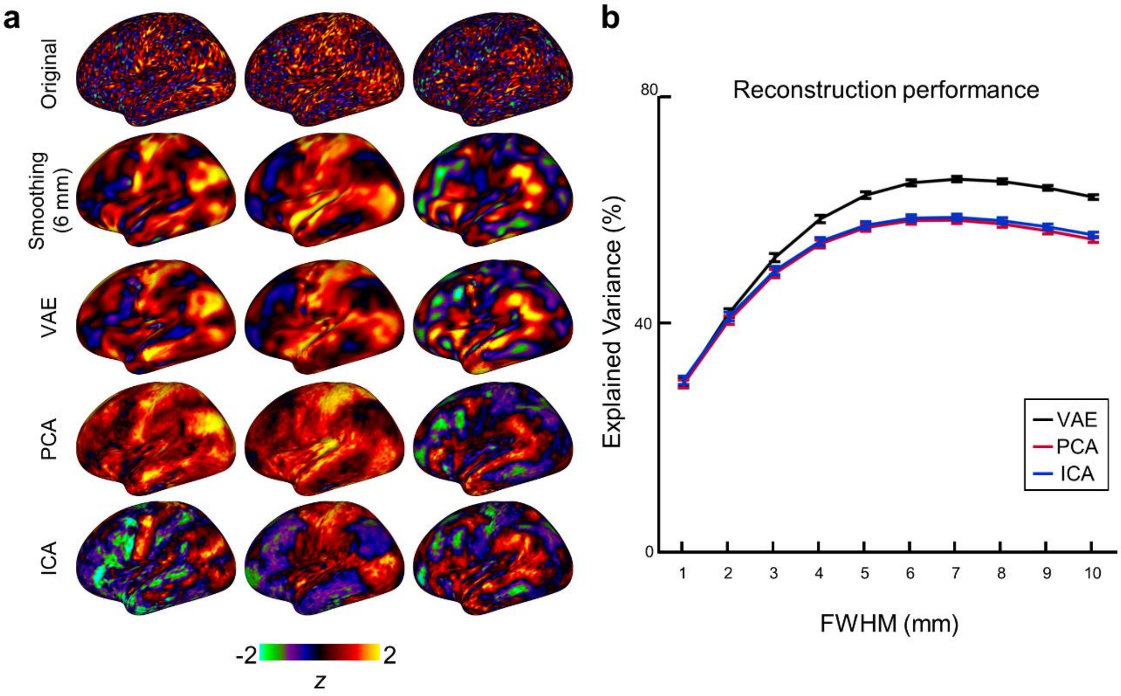


Figure 3.2. Resting-state fMRI data compression and reconstruction with VAE vs. PCA and ICA. (a) For illustration, three example maps of fMRI activity, before (1st row) and after (2nd row) being smoothed (FWHM=6mm), are shown in comparison with the corresponding maps reconstructed with VAE (3rd row), PCA (4th row), and ICA (5th row) trained to compress and reconstruct the training data from 100 subjects with 256 variables or components. (b) For quantitative comparison, the reconstruction performance, in terms of the percentage of variance in the fMRI images as explained by the model reconstruction, is shown for VAE, PCA, and ICA as a function of FWHM (from 1 to 10 mm) applied to the spatial smoothing of the fMRI images. The error bar stands for the standard error of mean.

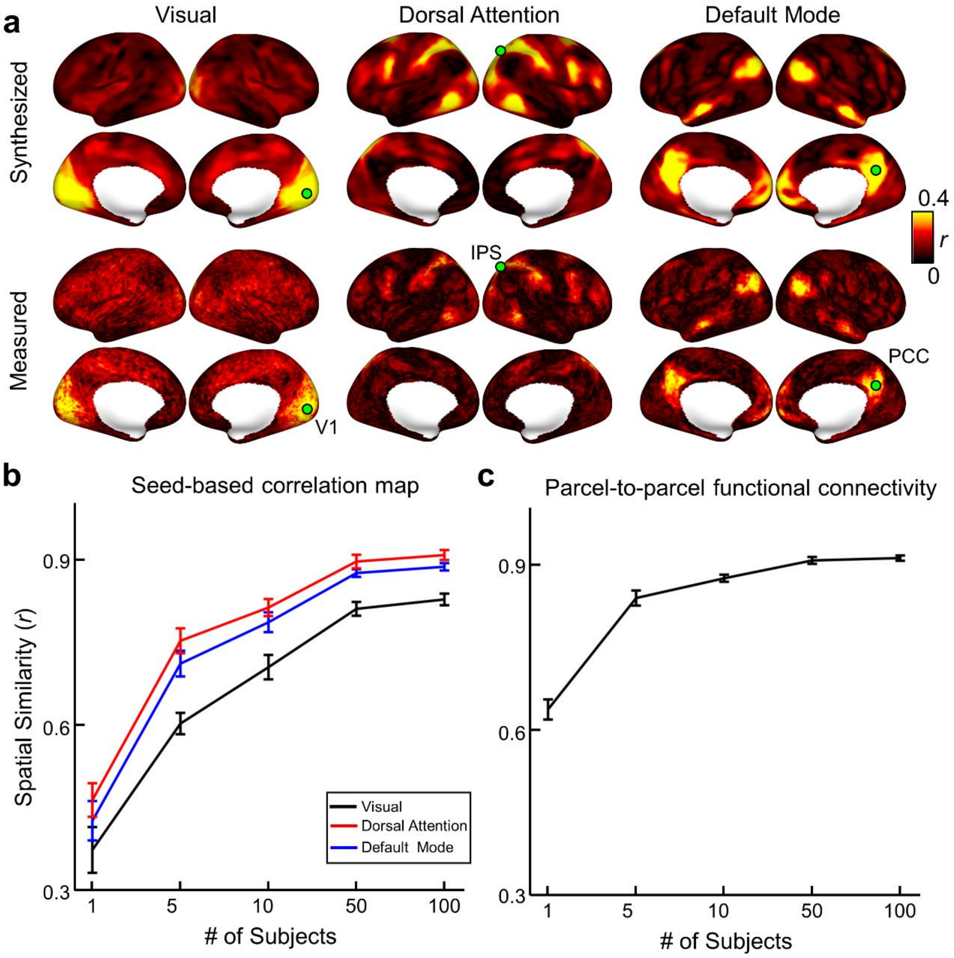


Figure 3.3. VAE synthesizes correlated fMRI activity. (a) Seed-based correlations of VAE-synthesized fMRI data (top row) vs. experimental fMRI data (bottom row) with the seed location (green circle) at V1 (left), IPS (middle), or PCC (right). (b) Spatial correlations between the seed-based functional connectivity based on VAE-synthesized data and those based on measured fMRI data concatenated across 1, 5, 10, 50, or 100 subjects. The colors indicate different seed locations (V1: black; IPS: red; PCC: blue). Similarly, (c) shows the spatial correlation between the synthesized vs. measured functional connectivity among 360 cortical parcels. The error bar indicates the standard error of the mean averaged across 20 repeated trials.

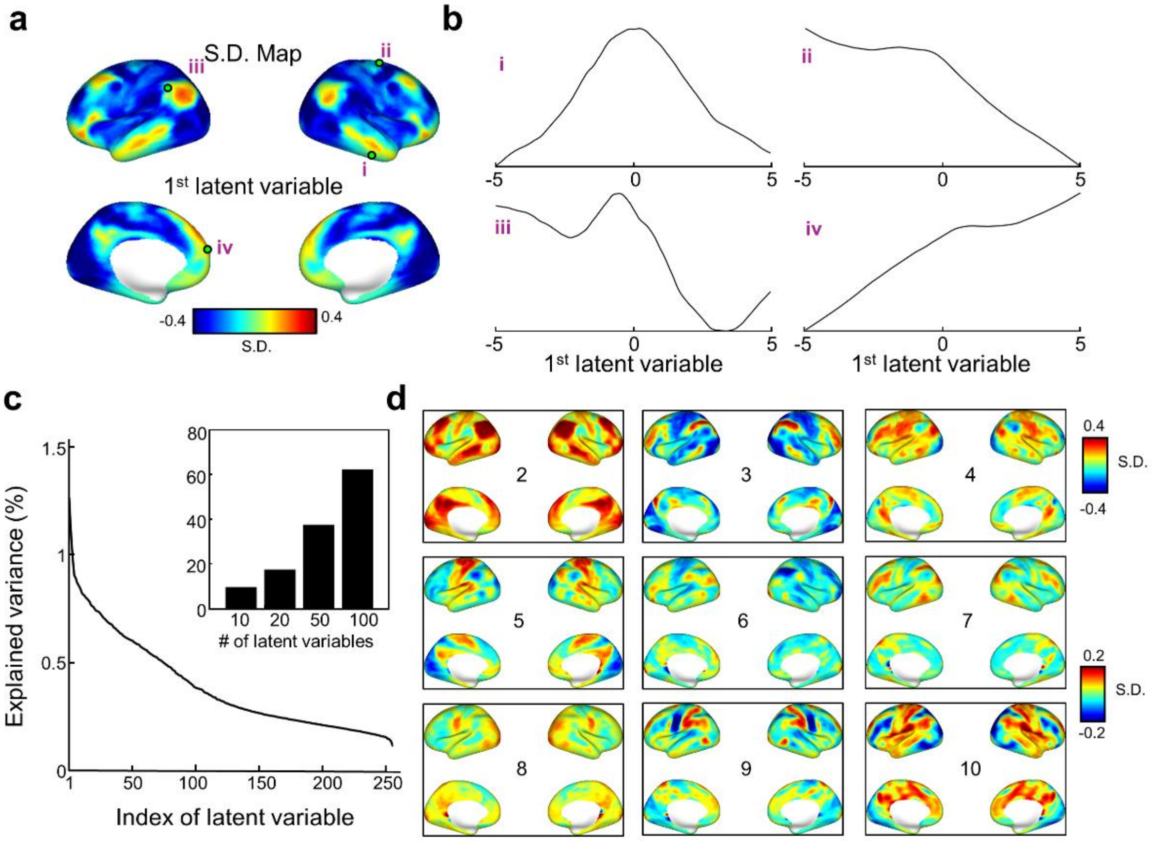
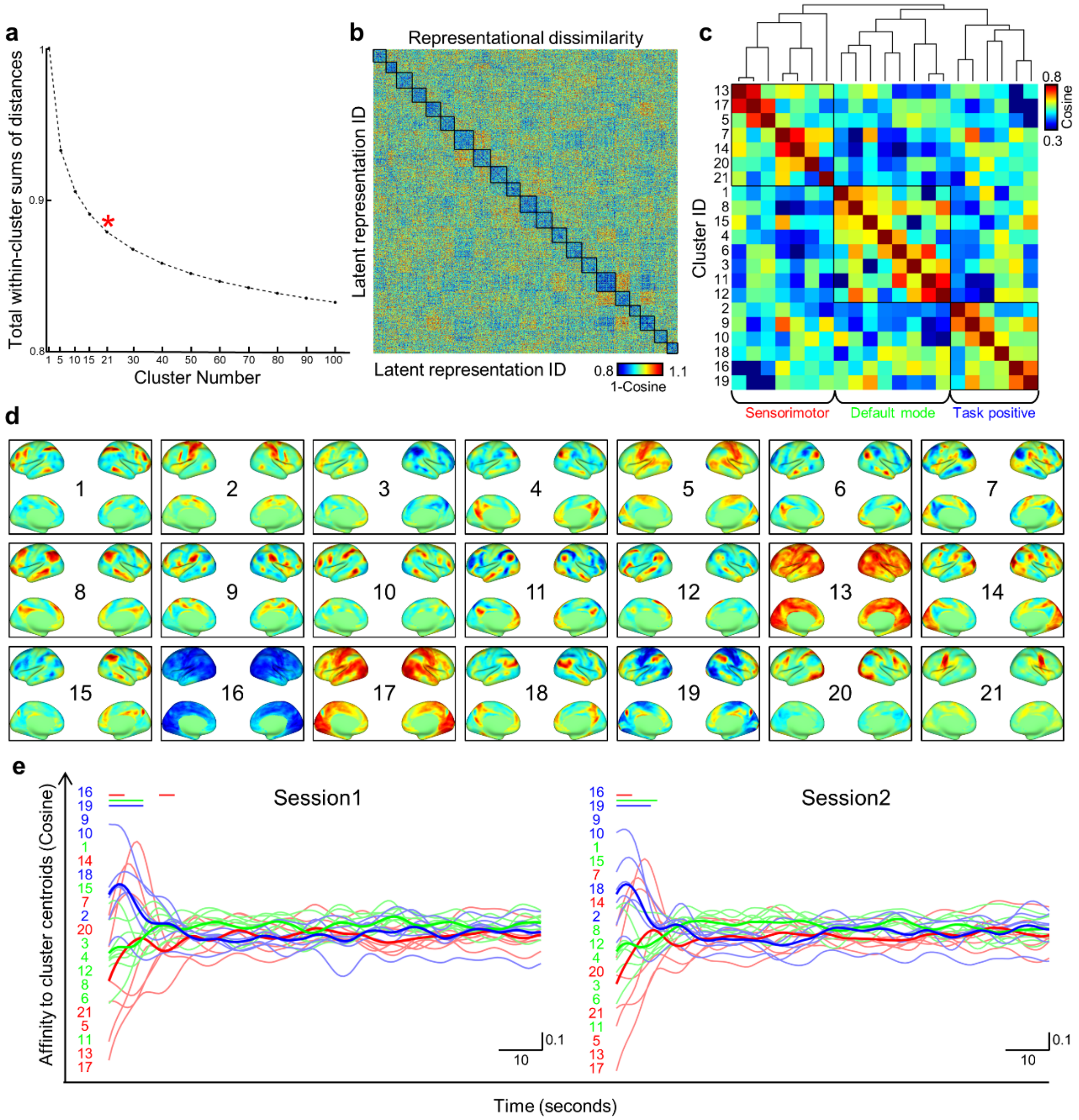


Figure 3.4. Latent variables drive the dynamics of latent representation. The latent variables are defined as the principal directions for the dynamic change of latent representation across adjacent time points. (a) Visualization of the 1st latent variable as a cortical pattern of the signed standard deviation. (b) For each of the four cortical locations, denoted as i through iv and shown as green circles in (a), the activity change is shown as a function of the 1st latent variable, indicating a varying nonlinear relationship. (c) The percentage of the variance that each latent variable explains the first order dynamics of latent representation. The inset shows the percentage of the total variance explained by top 10, 20, 50, or 100 latent variables. (d) The visualization of the 2nd through 10th latent variables as cortical patterns.

Figure 3.5. Clusters of latent representations. (a) The sum of within-cluster distance is shown as a function of the number of clusters (or K). The choice of k=21 is about where an elbow is observable in the plot. (b) Representational dissimilarity matrix of the 1-cosine distance between instantaneous latent representations (from 100 subjects with 1,200 time points per subject) reordered by cluster membership. (c) Hierarchical clustering of 21 cluster centroids. The clusters are grouped into three super-clusters, provisionally labeled as sensorimotor (red), default mode (green), and task positive (blue) networks, based on the cortical visualization of individual clusters in (d). (d) Cortical patterns decoded from every cluster centroid. The number shows the cluster index. The scale of each pattern is normalized by its maximal absolute value. (e) The group-averaged cluster-wise activity, described as the cosine affinity of instantaneous representation to the centroid of each cluster. Each thin line corresponds to one cluster; the color indicating the super-cluster (sensorimotor: red; default mode: green; task positive: blue) that each cluster belongs to. The thick lines show the average within super-clusters. The left and right panels show the activity patterns averaged across all subjects for session 1 and session 2, respectively. Colored lines on the top of panel (e) highlight the periods in which the sensorimotor (red), default-mode (green) or task positive (blue) super-cluster was statistically significant in the group level (permutation test, false discovery rate q<0.01).

Figure 3.5 continued



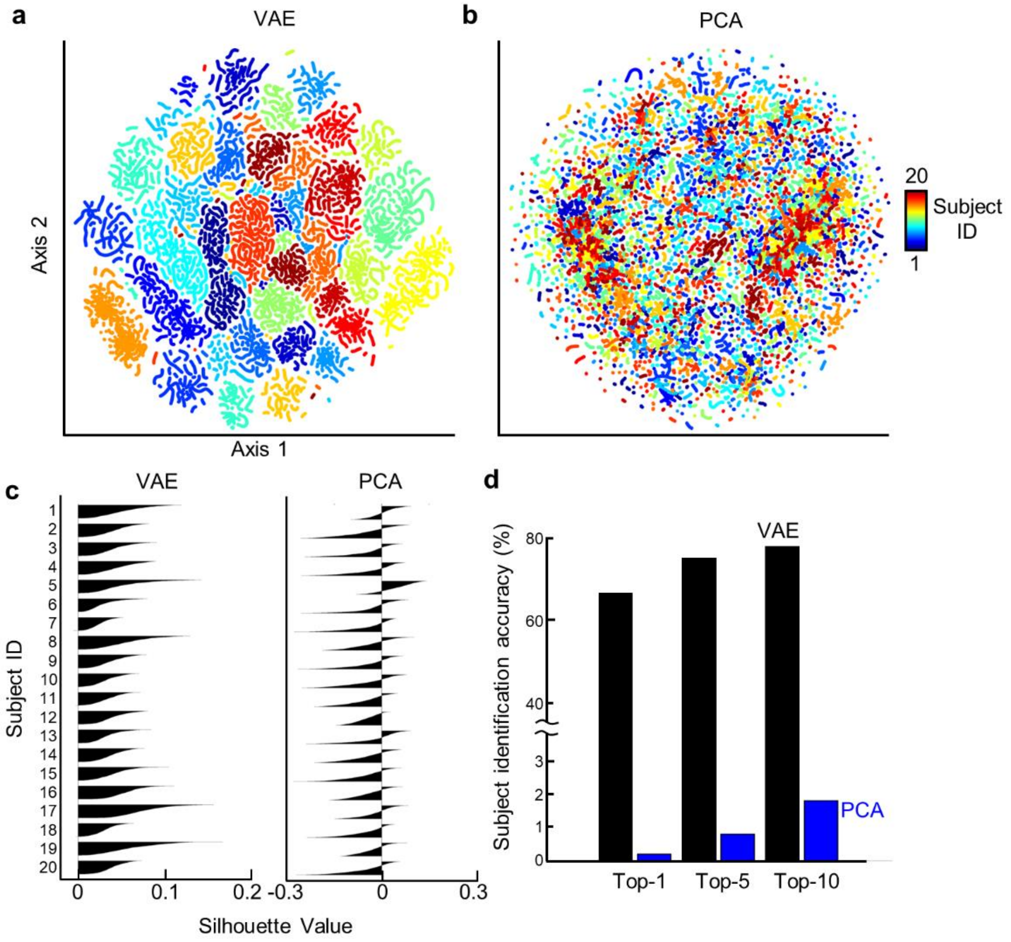
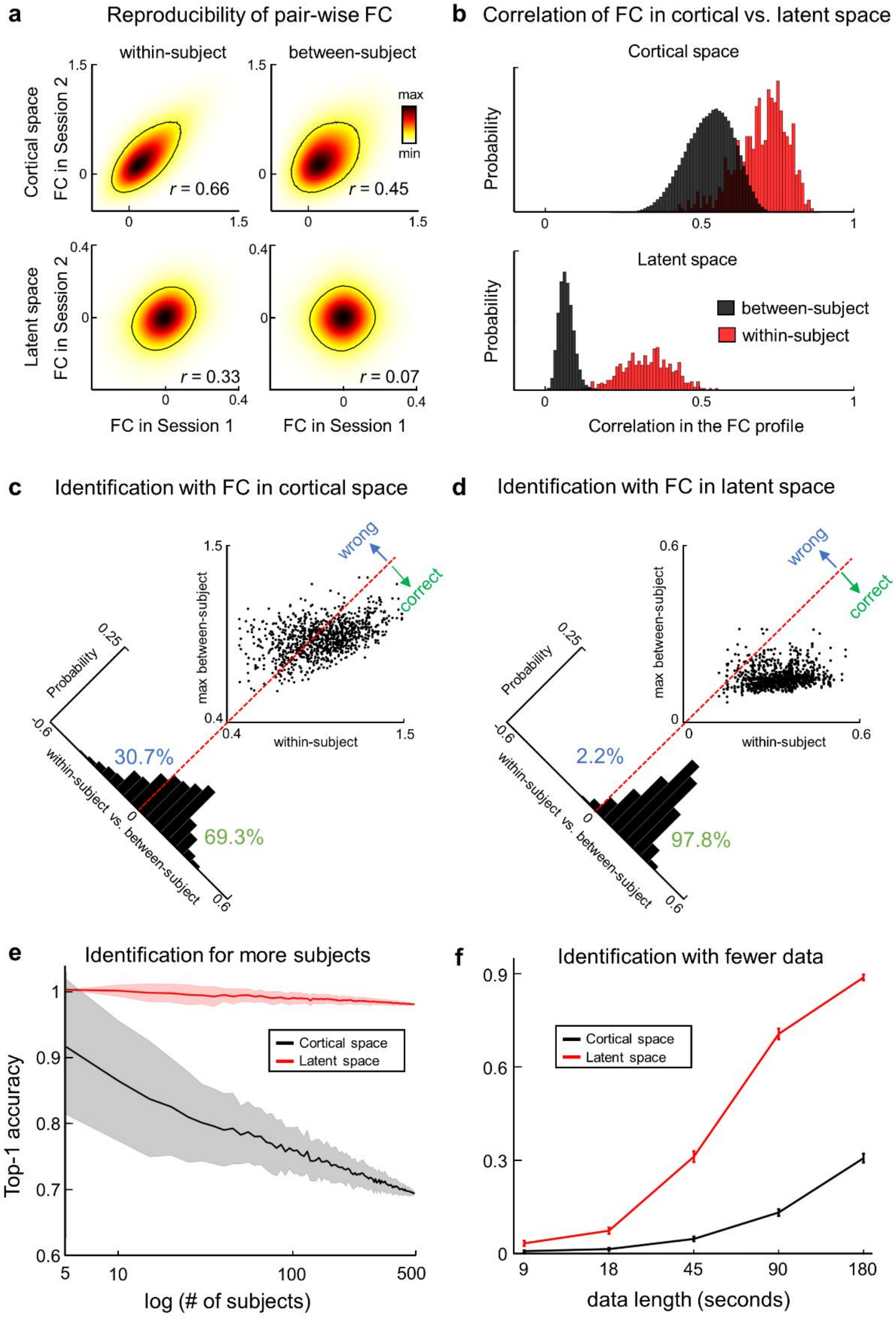


Figure 3.6. Individual variation of latent representation obtained with VAE vs. PCA. (a-b) Subject-wise latent representations visualized in a 2-D space obtained with t-SNE, when (a) VAE or (b) PCA is used to extract representations of rsfMRI activity from 20 subjects. (c) The Silhouette value shows how similar a representation is similar to each other within the same subject as opposed to between different subjects for VAE (left) or PCA (right). (d) The top-1, 5, and 10 accuracy of using the time-averaged representation as the feature to identify individuals in a large group of (n=500) subjects, for the representations obtained with VAE (black) or PCA (blue).

Figure 3.7. Individual identification based on functional connectivity between latent variables or cortical parcels. (a) Density distributions of z-transformed correlations between every pair of cortical parcels (top) or covariance between every pair of latent variables (bottom). For each pair, the correlation and covariance in one session is plotted against the corresponding correlation in the other session for the same subject (within-subject, left) or different subjects (between-subject, right) given the testing dataset with n=500 subjects. Contour line stands for 20% of the maximal density. (b) Within-subject (red) and between-subject (black) correlations in the FC among cortical parcels (top) or latent variables (bottom) are shown as histograms with the width of each bin at 0.01. (c) In the scatter plot, each dot indicates one subject, plotting the maximal correlation in the cortical FC profile between that subject and a different subject against the corresponding correlation within that subject. The red-dashed line indicates y=x, serving as a decision boundary, across which identification is correct (x>y) or wrong (y>x). The histogram shows the distribution of y-x (0.05 bin width) with the decision boundary corresponding to 0. Similarly, (d) presents the results obtained with latent-space FC in the same format as (c). (e) Top-1 identification accuracy evaluated with an increasing number of subjects (n=5 to 500) given the latent-space (red) or cortical-space (black) FC profile. The solid line and the shade indicate the mean and the standard deviation of the results with different testing data. (f) Top-1 identification accuracy given rsfMRI data of different lengths (from 9s to 180s). The line and the error bar indicate the mean and the standard deviation with different testing data.

Figure 3.7 continued



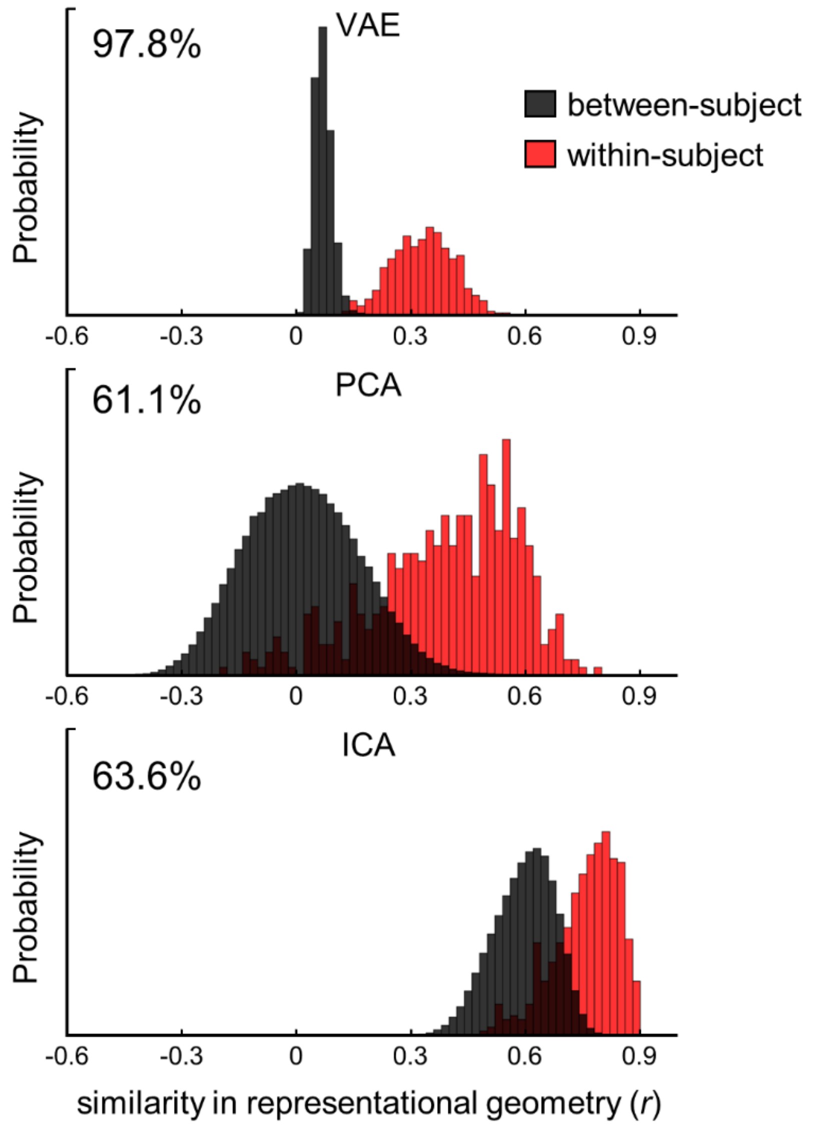
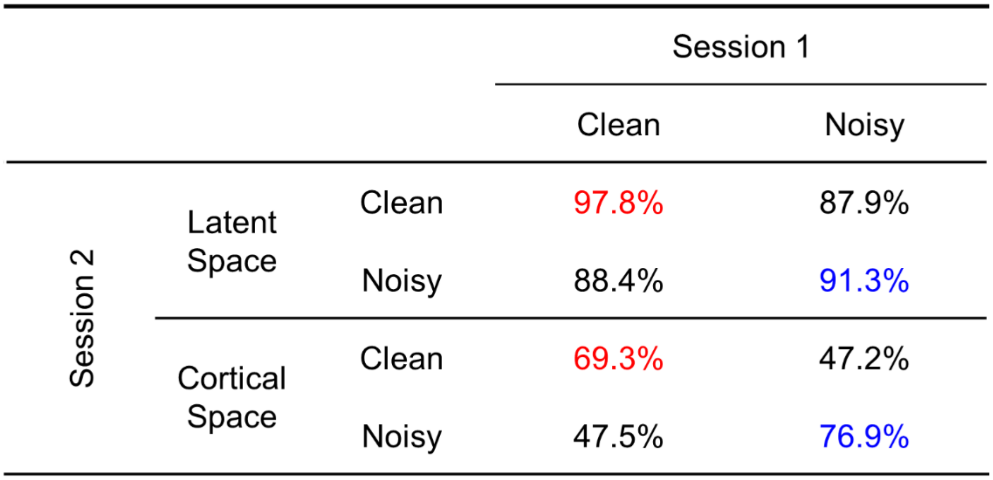


Figure 3.8. Individual identification with nonlinear vs. linear representations. Each plot shows the histogram of the similarity in the representational geometry between sessions within the same subject (red) vs. across different subjects (black), for representations in the nonlinear latent space obtained by VAE (top) or in the linear latent space obtained by PCA (middle) or ICA (bottom). The similarity reported is based on the inter-session correlation coefficient (or r). The histogram is discretized by bins with a width of 0.02.

Table 3.1 Subject identification accuracy across different conditions



			Session 1	
			Clean	Noisy
Session 2	Latent Space	Clean	97.8%	87.9%
		Noisy	88.4%	91.3%
	Cortical Space	Clean	69.3%	47.2%
		Noisy	47.5%	76.9%

4. LEARNING TASK-EVOKED REPRESENTATION OF FMRI UNDER NATURALISTIC MOVIE WATCHING

4.1 Introduction

Watching naturalistic audiovisual movies is the paradigm aiding to evoke naturalistic neural response patterns occurring at our daily life. Animal electrophysiology studies have shown more reliable and reproducible neural activity under naturalistic paradigms than under the laboratory-designed artificial stimuli (Belitski et al., 2008; Mechler et al., 1998; Yao et al., 2007). Similarly, human fMRI study during movie-watching tasks have shown reliable and reproducible brain patterns with a nearly full brain coverage (Hasson et al., 2010; Hasson et al., 2004). This defining feature of the naturalistic movie-watching paradigm, therefore, has led many neuroscientists to new understandings of the brain system engaged in the naturalistic task (Betzel et al., 2020; Bolton et al., 2020; Kauppi et al., 2010; Mandelkow et al., 2016; Vanderwal et al., 2017). Unlike simple and artificial stimuli, highly complex and continuous naturalistic paradigms is hard to be modeled, making it more difficult to map the brain networks evoked by the stimuli. Hence, in most studies, fMRI data under the naturalistic movie-watching task has been analyzed through the model-free non-parametric methods, for example, inter-subject Functional Connectivity (FC) analysis (Betzel et al., 2020; Hasson et al., 2004) or condition-specific FC analysis (Demirtaş et al., 2019; Vanderwal et al., 2017).

Paradoxically, those two methods reported diverging findings. For example, the study from (Demirtaş et al., 2019) found that most of the significant FCs were confined to the visual regions, or between visual and auditory regions, using the inter-subject FC analysis. On the contrary, the study comparing FC maps under the movie-watching condition and resting-state condition revealed FCs spanning broader brain regions even including cognitive brain networks, and the majority of FCs were suppressed by the movie-watching condition (Lynch et al., 2018). Such contradictory results were discussed by other studies (Bianciardi et al., 2009a; He, 2013; Monier et al., 2003; Ponce-Alvarez et al., 2013). Among other possibilities, the negative task-rest interaction – engagement in tasks suppresses the ongoing brain activity – has been considered as a major cause for this contradiction (Churchland et al., 2010; He, 2013). Therefore, the decreased synchrony between brain regions, i.e., negative FCs, under the naturalistic movie-watching paradigm, have been overlooked and considered as the false-negative results introduced by the negative task-rest interaction. While much has been learned regarding the neural origins of negative FCs observed under the resting state (Chen et al., 2011; Gopinath et al., 2015; Liang et al., 2012), relatively little is known regarding the basis of negative FCs during the movie-watching task. A network analysis technique that can effectively address the non-linear task-rest interaction is required to reveal the functional brain organizations modulated by naturalistic tasks.

Recently, we have proposed an unsupervised and non-linear Variational AutoEncoder (VAE) model to learn deeply embedded representations of resting-state fMRI data (rsfMRI) (Kim et al., 2020). Given the model design, we were able to delineate non-linear generative factors of rsfMRI and presented the complex cortical patterns of rsfMRI in the low-dimensional and linear latent space. In addition, we found that the temporal dependencies between generative factors of rsfMRI (presented as latent variables) were effectively removed, yielding a spherical gaussian null distribution for representations of fMRI data collected from a large number of subjects. Based on this observation, we hypothesized that, in the latent space non-linearly defined by the VAE, the representations of task-evoked activity would be separable from that of spontaneous activity. Furthermore, we hypothesized our VAE model, trained on rsfMRI, would be generalizable to other fMRI data under different brain conditions and different recording parameters. Collectively, here we tested two aspects of VAE using movie-watching fMRI data: 1) generalizability of our pretrained VAE model to fMRI dataset during watching naturalistic movies, and 2) independency between task-evoked brain activity and spontaneous brain activity in the latent space defined by the VAE. Testing was done by employing large fMRI datasets from subjects watching naturalistic movies, including several short video clips interleaved the resting state, which were publicly provided by HCP (Van Essen et al., 2013). Given the observed linear superposition between task-evoked brain activity and ongoing spontaneous brain activity in the latent space, we successfully estimated task-evoked latent variables by simply averaging latent variables of individuals. By defining the principal bases explaining the trajectory of task-evoked latent variables, we showcased new findings: different principal bases of task-evoked latent representations reflected different aspects of video contents and each of principal bases exhibited the unique interaction/anti-interaction between brain networks spanning from low-level sensory networks to high-level cognitive networks.

4.2 Methods and Materials

4.2.1 Subjects and Data

Here, we used task-fMRI and rsfMRI data from 192 healthy subjects released by HCP (Van Essen et al., 2013). Among them, 19 subjects were excluded because their recordings were missing or shortened, resulting in a total of 173 subjects. For each subject, there were four sessions of recordings under resting state or movie watching paradigm. Those runs were recorded with different phase encodings: anterior-posterior or posterior-anterior from different days. We included all four sessions in the analysis. For the resting state, each session included 900 time points and the resolution was 1s. Similarly, each session of movie-watching fMRI data was about 15-mins although the exact length of each session was different due to the different movie lengths. Four sessions (namely, session 1, 2, 3, and 4) had 921, 918, 915, and 901 time points, respectively. Both 7T rsfMRI and movie-watching fMRI data were collected using the same gradient-echo Echo-Planar Imaging (EPI) sequence with the following parameters: repetition time (TR) = 1000 ms, echo time (TE) = 22.2 ms, flip angle = 45 deg, field of view (FOV) = 208 x 208 mm^2, matrix = 130 x 130, spatial resolution = 1.6mm^3, number of slices = 85, multiband factor = 5, image acceleration factor (iPAT) = 2, partial Fourier sampling = 7/8, echo spacing = 0.64 ms, bandwidth = 1924 Hz/Px.

After downloading the data preprocessed with the minimal preprocessing pipeline (Glasser et al., 2013), we applied voxel-wise detrending (regressing out a 3rd-order polynomial function), bandpass filtering (from 0.01 to 0.1 Hz), and normalization (to zero mean and unitary variance). It is noteworthy that 7T data was also pruned through the automatic denoising with ICA (or the ICA-FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014), which was used in 3T fMRI data. Since the pretrained VAE model was utilized, all data was used as a testing dataset of the VAE model. For the purpose of comparison, we also used 3T rsfMRI data. The details of preprocessing steps in 3T rsfMRI can be found elsewhere in (Kim et al., 2020). After the preprocessing, four sessions of movie-watching fMRI and rsfMRI data were concatenated, resulting in a total 3,655 and 3,600 time points, respectively. To prevent the possible confusion between datasets, hereafter we named 3T resting-state and 7T resting-state fMRI as 3T rsfMRI and 7T rsfMRI, respectively.

4.2.2 Movie Stimulus Paradigm

All subjects watched the same movie clips while each session used a different set of movie clips. For example, movie stimuli of session 1 and session 3 included short video clips made freely available under Creative 5 Commons license on Vimeo while movie stimuli of session 2 and session 4 were truncated movie scenes from Hollywood films such as Inception, Home Alone, and Star Wars. The video clips had different length. The length of the shortest video was 63 seconds and the longest data was 4 minutes and 15 seconds (Star Wars). The details of the short clips can be found in Table 4.1. At the end of each session, a short video clip (83 seconds) was repeated for the test-retest purpose. At the beginning and end of each short clip, there were 20 seconds of resting periods presented as a white "REST" text on a black background. To make it consistent with the temporal resolution of fMRI data, we downsampled the video clips to 1 Hz from 24 Hz.

4.2.3 Eye-Tracking Data

Eye-tracking data was acquired during movie-watching tasks using an EyeLink S1000 system 12 (SR Research). Here, we downloaded eye-tracking data from HCP files (for example, 100610_7T_MOV1_eyetrack.asc) and the synchronization information between fMRI data and eye-tracking data was extracted from the summary file (e.g., 100610_7T_MOV1 _eyetrack_summary.csv). All data analyzed in the study is freely downloadable from the HCP website.

Briefly, eye-tracking data provided three types of information, horizontal and vertical positions of the pupil, and the pupil diameter. The sampling rate was either 1000 or 500 Hz. Among 172 subjects, 145 subjects had full availability in eye-tracking data across four sessions of movie-watching fMRI data. Among 145 subjects, 47 subjects having eye-closing (or lost in the pupil trace) periods more than 20% of the total recording period were further excluded from the analysis. Finally, eye-tracking data from 92 subjects was used to extract the gazing information over movie stimuli.

We estimated the gaze heatmap from the gaze information as follows. First, eye-tracking data was divided into 1-sec segments without overlapping. Then, we averaged the gaze location for each segment without considering the eye-closed period. Next, we applied a 2D Gaussian spatial filter to the gaze point. The standard deviation of the filter was determined as a 1-degree radius of vision. Given the information from (Benson et al., 2018), we approximated that the 1-degree radius was similar to 48 pixels in the 1024 by 720 screen resolution. If subjects had no eye-open duration during specific segments, those subjects were excluded in estimating the group-level heatmap. Finally, we acquired movie-related gaze heatmaps by averaging heatmaps of individuals. Figure 4.1. illustrates several examples of estimated movie-related gaze heatmap. The duration of eye-closed period was estimated by counting the periods when measured pupil dilation was 0, per segment. Lastly, we estimated the timeseries of horizontal gaze position by choosing the points having the highest density, per group-level heatmap. Here, the eye-open duration feature was considered as a proxy measurement of the vigilance level, as suggested by (McIntire et al., 2014; Wang et al., 2016).

4.2.4 Extracting Video Features from Movie Stimuli

We extracted diverse visual and audio features spanning from low-level features to high-level features, to utilize the rich audiovisual content of movie stimuli. For low-level features, the image luminosity and audio intensity; for middle-level features, the presence of face, the presence of text, and the presence of speech, were used in the analysis. All video features were extracted from the movie scenes using our in-house code and the MATLAB toolboxes. Image luminosity of each scene was extracted as the following: 1) we converted the image to grayscale using MATLAB function *rgb2gray.m*, and 2) we estimated a mean-square-root of grayscale image. The audio intensity was acquired using MATLAB toolbox *MIRtoolbox*, which is freely available at https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox (Lartillot et al., 2008). As a fair starting point of detecting faces, we employed MATLAB built-in machine-learning algorithm (*vision.CascadeObjectDetector*) with the Viola-Jones algorithm (Viola et al., 2001) and applied the algorithm to downsampled movie scenes. We further visually inspected the quality of the detected features and corrected mislabeled features. Detecting the text from natural images was done using MATLAB built-in function *ocr.m*. After initial searching using the algorithm, we carefully checked the quality of results and edited the mislabeled ones. Lastly, identifying the presence of speech was done manually from the original movie stimuli, and the time-stamps of the labels were matched to the fMRI data.

The analysis of high-level aspects of movie contents was done by utilizing the semantic labels that were made available by the HCP. In the movie stimuli, there were 853 semantic labels presented at least once during the movie stimuli. For example, the “eggplant.n.01” feature appeared only once for the entire video stimuli since the meaning of this feature is very specific and narrow. On the other hand, the “entity.n.01” appeared in nearly every movie scene since it covers very wide semantic meaning. To control such inconsistency in appearing frequencies and specificity of semantic concepts, we chose 23 semantic labels (12 nouns and 11 verbs) that can represent 853 semantic labels. The whole list of categorical labels can be found in Figure 4.2 and look for Table 4.2 for the exact definitions of semantic labels in Wordnet. Since the temporal resolution of semantic features was already matched to the fMRI data, we did not further preprocess the feature sets. Lastly, we convolved all levels of features with hemodynamic response function (HRF), to consider the hemodynamic delay between movie stimulus and brain response. The hemodynamics response function used here was a conventional function consisting of two-gamma functions, provided by the Statistical Parametric Mapping toolbox (Frackowiak, 2004). The hyperparameters of HRF function were defined as: delay of response = 6sec, delay of undershoot = 16sec, dispersion of response=1, dispersion of undershoot=1, ratio of response to undershoot = 6, onset delay = 0sec, and length of function = 32sec. Each HRF-convolved feature was rescaled to have 0 mean and 1 as standard deviation.

4.2.5 Generalizability of Pretrained VAE Model on Movie-Watching fMRI Data

One of our goals in designing the VAE model was to make the VAE model generalizable to various fMRI data under different neural states and/or recorded under different recording parameters. To examine the generalizability of the VAE model, we imported the VAE model pretrained from 3T rsfMRI without applying any further fine-tuning steps. The generalizability of pretrained VAE model was examined by estimating the reconstruction performance on unseen movie-watching fMRI data or unseen 7T rsfMRI under different recording parameters. To specify, we manually smoothed 7T MOVIE and REST fMRI data with varying smoothing effects (FWHM=1, 2, …, and 10 mm). The reconstruction performance was quantified by measuring Pearson correlation between the reconstructed cortical pattern and manually smoothed cortical pattern. Compressing and reconstructing original fMRI data was done through the encoder and the decoder of pretrained VAE model, respectively. The compressing performance of the VAE model was further compared by Independent Component Analysis (ICA), as the linear counterpart of the VAE. We additionally calculated a new ICA basis from 7T REST fMRI data as a fine-tuned version of the linear compressor. To set the baseline of reconstruction performance, we synthesized the cortical patterns from random latent variables (0 mean and unitary variance) and examined the reconstruction performance between synthesized cortical patterns and manually smoothed cortical patterns.

4.2.6 Linear Superposition Between Task-Evoked Activity and Spontaneous Activity in the Latent Space

As fMRI data obtained while subjects were watching the movie was driven by both task-evoked activity and spontaneous ongoing activity, many studies estimated the task-evoked activity (or task-evoked FCs) by averaging fMRI data across subjects (or correlating the task-evoked activity). This averaging strategy was done based on the assumption that the spontaneous cortical activity and the task-evoked cortical activity are independent to each other. Mathematically, this assumption can be expressed by the law of variance sum as:

$$\sigma_{\boldsymbol{X+Y}}^2 = \sigma_{\boldsymbol{X}}^2 + \sigma_{\boldsymbol{Y}}^2 + 2cov_{\boldsymbol{X,Y}}, \qquad (1)$$

where **X** and **Y** stand for spontaneous cortical activity and task-evoked cortical activity, $\sigma_{\boldsymbol{X}}^2$ and $\sigma_{\boldsymbol{Y}}^2$ are their variances, and $cov_{\boldsymbol{X,Y}}$ is the covariance between **X** and **Y**. If the linear superposition between activities holds true, Eq. (1) can be reformulated as:

$$\sigma_{\boldsymbol{X+Y}}^2 = \sigma_{\boldsymbol{X}}^2 + \sigma_{\boldsymbol{Y}}^2. \qquad (2)$$

As, we measured fMRI under movie-watching task, which is **X** + **Y**, and rsfMRI, which is **X**, we directly tested whether Eq. (3) is true or not:

$$\sigma_{\boldsymbol{X+Y}}^2 - \sigma_{\boldsymbol{X}}^2 = \sigma_{\boldsymbol{Y}}^2. \qquad (3)$$

Specifically, we encoded rsfMRI data of 173 subjects (data length of latent variables=3,600 per subject). Given the latent variables, we estimated the covariance matrix between timeseries of 256 latent variables per subjects, and averaged the covariance matrix across subjects, yielding $\sigma_{\boldsymbol{X}}^2$. The identical analysis was applied to fMRI data during watching movies (data length=3,105; excluded inter-movie resting period), yielding another covariance matrix $\sigma_{\boldsymbol{X+Y}}^2$. For the $\sigma_{\boldsymbol{Y}}^2$, we first averaged the latent variables across subjects and estimated the covariance matrix. We measured the similarity between $\sigma_{\boldsymbol{Y}}^2$ and $\sigma_{\boldsymbol{X+Y}}^2 - \sigma_{\boldsymbol{X}}^2$, by estimating squared Pearson correlation between vectorized upper triangular parts of two matrices. The same analysis was repeated using the same fMRI data but in the different latent spaces, 360 cortical parcels defined by (Glasser et al., 2016), or 256 IC maps defined from 3T rsfMRI data.

4.2.7 Task-Evoked Latent Representation

Here, we encoded the individual’s movie-watching fMRI (n=173) using the VAE encoder which was pretrained from 3T rsfMRI data. Then, we averaged latent representations across individuals to estimate the task-evoked latent representations (Figure 4.4). The rationale underlying this procedure was that the representations of spontaneous activity and the representations of task-evoked cortical activity were independent in the latent space non-linearly defined by the VAE. Thus, we assumed the averaging procedure would cancel out latent representations of spontaneous activity while keeping the latent representations evoked by the movie stimuli.

4.2.8 Defining Basis Functions of Task-Evoked Gradient

As our goal was to investigate how the latent representations were traveling throughout the movie stimuli i.e., task-evoked latent trajectory, we asked how much the latent space would be required to explain the task-evoked latent trajectory (Figure 4.3. left). As an initial step, we estimated the representational geometries of latent trajectory for the rsfMRI and the task-evoked fMRI. For spontaneous 7T rsfMRI, we concatenated latent representations of individuals and estimated the cross-correlation matrix, whereas the representational geometry of task-evoked representations was estimated by correlating between latent variables of the task-evoked representations. Quantification of the subspace dimensions was done by applying the PCA to resting-state and task-evoked latent representations (Figure 4.3 right). After projecting the task-evoked latent representations to the re-defined latent subspace, we estimated the latent gradient that can approximate the latent trajectory, by subtracting adjunct latent representations (Figure 4.3. right).

We further examined which aspects of movie stimuli evoked the changes in the task-evoked latent gradient. We calculated the magnitude of the task-evoked latent gradient by estimating the root-mean-square of the latent gradient per timepoint. The mean gradient magnitude was estimated by averaging the magnitude of latent gradient per short clip. Additionally, we segmented the trend of magnitude during inter-session resting periods (20 seconds, n=15), also including 5s before and 10s after the resting periods. The significance level of gradient magnitude over the progression of the resting period was tested using two-sample t-test between samples of each time point (n=15) and samples from the movie-watching task (n=3,101). The multiple comparison correction was done by correcting the false discovery rate at *q*=0.05.

The principal basis functions defining dynamics of task-evoked latent gradients were estimated by applying PCA to the task-evoked latent gradient. Same as the dimension of subspace (=24), 24 principal basis functions were defined. Cortical mapping of each principal basis was done by the following procedure: 1) we multiplied random scaling factors (n=1,000) to the basis function, 2) we reconstructed the cortical patterns using the VAE decoder, and 3) we calculated covariance between random scaling factors and reconstructed fMRI activity, per cortical location. The estimated covariance value of each cortical location was considered as the activation/deactivation level of that location. Since the VAE decoder is highly non-linear, the visualization method proposed here was not guaranteed to reflect the true cortical meaning of principal bases in the latent space, but the results in our study empirically supported that our proposed method was a good approximation strategy mapping the latent representations into the cortical space. Lastly, the cortical map of each basis function was thresholded at the 30% of the maximal absolute value of each map, for the better interpretation of results.

4.2.9 Reproducibility of Task-Evoked Latent Representations

We further asked whether the trajectory of latent representations was reproducible under the same movie stimulus. Only for the visualization purpose, we projected the task-evoked latent trajectory into 2D-space using the t-Distributed Stochastic Neighbor Embedding (t-SNE) method. The distance between latent representations was defined as Euclidean distance and the perplexity, which is a hyperparameter determining the cluster size in t-SNE method, was set to 60.45 (=square root of 3,655). To acquire reliable t-SNE mapping results, the same t-SNE method was repeated 10 times with random initializations, and the trial having the lowest sum of distance was chosen and visualized. Additionally, we segmented the latent trajectory to 4 short trajectories under the repeated movie stimuli (“test-retest”). Among the total length of a short video clip (83 seconds), the last 6 time points were excluded from the analysis to prevent possible contamination from the resting-state period. We further quantified the inter-session reproducibility of basis functions by calculating the Pearson correlation between dynamics of basis functions across different sessions. The significance level of reproducibility was tested using a one-sample t-test after transforming *r* values into Fisher’s z-score. The multiple comparison correction was done by controlling the FDR level at *q*=0.05.

To further validate the reproducibility of our analysis, for each latent gradient of “test-retest” run, we defined new principal basis functions. Since the top-9 subspaces were able to explain 99% of the variance for all four test-retest sessions, we only kept 9 subspaces and evaluated the inter-session reproducibility of principal map and dynamics across different sessions. Finally, we estimated the similarity of the principal maps between the subspaces from the whole data and one from the “test-retest” runs.

4.2.10 Correlating Latent Trajectory with Low-, Middle-, High-Level Aspects of Movie Stimuli

Here we tried to understand the driving force of task-evoked latent gradients using low-, middle-, and high-level audiovisual features of movie stimuli. Since the contents of video clips were varying by each short clip, some middle-level features were absent or rarely appearing for some video clips. To control such bias, for each middle-level feature, we counted the appearing frequency of middle-level feature per short video clip, and applied the correlation analysis only to video clips containing appearances of middle-level features three times or more. This analysis was intended to balance between the statistical power and the reliability of estimated correlation values. Here, the correlation coefficient was estimated using non-parametric Spearman correlation analysis. Correlation values were averaged over different movie clips after transforming *r*-value into Fisher’s z-score, and re-transformed averaged z-score to *r*-value. The statistical significance was tested using a one-sample t-test. Multiple comparison errors were corrected by controlling the FDR at *q*=0.05.

For the high-level semantic features, we employed the ordinary linear regression taking time-series of semantic labels as input to predict the dynamics of each latent basis function. The regression analysis was done after concatenating short movie clips. The coefficient weights of regressors (i.e., semantic features) were used as indicating the importance of semantic meaning of each basis function. This analysis was repeated for each basis function.

4.2.11 Task-Relevant Gaze Position Explains Variance of Latent Gradients

To evaluate the effects of the gaze-informed features on the dynamics of latent gradients, we built another ordinary regression model by combining semantic labels and gaze-informed features, EO-duration or horizontal gaze position features – named as the full model. Then, the improved predictability by each feature was quantified by employing the Jackknife resampling scheme. Specifically, we subtracted the explained variance by the leave-one-feature-out regression model from the explained variance by the full model. These improvements from semantic labels were used as a null distribution of improvement and compared to the improvement by gaze-informed features.

4.2.12 Individual-Wise Cortical Activity Depends on the Gaze Location

We investigated whether inter-subject variability in cortical patterns during watching naturalistic movies can be explained by the variability in their gaze positions. Among 96 subjects who had reliable eye-tracking information, we estimated the individual-wise gaze position after applying the HRF function to the time-series of gaze position. Per time point, we measured the inter-subject Euclidean distance of gaze positions, as the time-resolved inter-subject variability in gaze position. Similarly, per time point, we also calculated the Euclidean distance of latent representations between subjects. Finally, we asked whether the inter-subject variability in latent representations was predicted by the inter-subject variability in gaze positions, by applying the inter-subject representational similarity analysis (IS-RSA). For each timepoint, subjects who had eyes-closed period or gaze position out of display size were excluded from the IS-RSA analysis.

4.3 Results

4.3.1 Pretrained VAE Successfully Compressed Unseen Movie-Watching fMRI Data

Inspired by the finding that the VAE model was able to extract meaningful latent features from noisier rsfMRI data even though the model was trained from the clean data (Table 3.1), we asked whether our VAE model can be generalizable enough to extract useful representations from unseen fMRI data under the different neural states and/or under different recording settings. To address the question, we utilized the VAE model pretrained from 3T rsfMRI without further fine-tuning procedures. By compressing and reconstructing 7T movie-watching fMRI data, which was never introduced during training the VAE model, we found that the reconstruction performance was comparable to one from 7T resting-state fMRI (FWHM=6mm, for movie-watching, r^2 = 0.71±0.02; for resting-state, r^2 = 0.71±0.02, mean±s.d.; p=0.20, two-sample t-test), as shown left panel in Figure 4.5. This result suggested that the VAE model remains robust regardless of brain conditions. In addition, VAE showed better reconstruction performance than the ICA method, in any neural states (Figure 4.5). We further asked whether different recording parameters would change the reconstruction performance of VAE. Interestingly, unseen 7T rsfMRI data showed higher reconstruction performance than 3T rsfMRI (Figure. 4.5 right; FWHM=6mm; for 7T, r^2 = 0.71±0.02; for 3T, r^2 = 0.65±0.03, mean±s.d.; two-sample t-test, p<0.01). We speculated this somewhat unexpected result might be partially originated from two aspects: 1) better SNR level due to the high field strength (7T vs. 3T), 2) better spatial SNR by scarifying temporal resolution (TR=1 sec for 7T vs. TR=0.72 sec for 3T; and bandwidth: 1924 Hz/pixel for 7T vs. 2290 Hz/pixel), or both. To set the baseline of comparison, we synthesized the cortical patterns from random latent variables, and estimated the reconstruction performance using the synthesized cortical patterns, as the null distribution of reconstruction performance. For all fMRI data, the null distributions had significantly lower performance than the actual reconstruction.

We further found that the VAE outperformed the linear compressing method, for all different smoothing levels except when FWHM=1mm (Figure 4.6, n= 40 subjects, paired two-sample t-test, FDR-corrected). The reconstructed images via the VAE model were most similar to the fMRI patterns smoothed at the level of FWHM=6mm, confirming the VAE model reconstructs input with a certain level of smoothing effects, regardless of data characteristics. To further test the generalizability of VAE model, we accessed the ICA basis from the 7T resting-state fMRI data – named as the fine-tuned ICA basis and accessed the reconstruction performance using it. Indeed, we found that the reconstruction performance of VAE was significantly worse than one from the fine-tuned ICA basis when FWHM = 1 and 2 mm. However, when the smoothing levels were higher FWHM>4mm, the VAE showed superior reconstruction performance than the fine-tuned ICA (n=40 subjects, paired two-sample t-test, FDR-corrected q=0.05). Collectively, our results presented here strongly supported that our VAE model is highly generalizable to various fMRI data under different neural states and different recording settings.

4.3.2 Linear Superposition Between Task-Evoked Activity and Spontaneous Ongoing Activity in the Latent Space

According to the theoretical concept of the VAE, the VAE is a non-linear data compressor. Therefore, the non-linear and complex relationship between generative factors and observations can be properly deciphered by the VAE model, which is a defining difference from other linear source blind methods e.g., PCA and ICA. Consequently, we investigated whether non-linear, complex interaction between task and rest, which were observed at movie-watching fMRI study (Lynch et al., 2018), would be effectively diminished in the latent space. To test it, we evaluated whether the covariance between task-evoked latent variables (estimated by averaging latent variables across individuals) would be able to explain the difference between group-level covariance matrix of movie watching-fMRI and one from rsfMRI (Figure 4.7).

The rationale behind this analysis was that the second statistics (e.g., a covariance between latent variables) of task-evoked or spontaneous cortical activities are linearly additive/deductive if two signals are independent to each other (see method 4.2.6 for detail). For comparison purposes, we further applied the same analysis on the latent spaces linearly defined at the region-of-interest (ROI) level (cortical parcel) or at the network level (ICA). As shown in Figure 4.7 middle row, we observed that not much difference between covariance matrices under different neural states was explained at the region-of-interest level (i.e., cortical parcel, $r^2 = 0.08$), as in line with findings from (Lynch et al., 2018). Similarly, the latent space linearly defined by ICA also failed to disentangle task-rest interaction (Figure 4.7 bottom row, $r^2 = 0.15$). On the contrary, in the VAE-derived latent space, much higher similarity was observed (Figure 4.7 top row, $r^2 = 0.59$). Such observation was statistically tested by partitioning the data into 10 folds ($r^2 = 0.473 \pm 0.007$ for VAE; $r^2 = 0.078 \pm 0.009$ for parcel; $r^2 = 0.142 \pm 0.005$ for ICA, F=848.16, $p<0.01$, one-way ANOVA), as shown in Figure 4.8. By following the post-hoc paired t-test, we found VAE had better similarity than ICA ($p<0.001$) or parcel ($p<0.001$). There was also a significant difference between ICA and parcel ($p<0.01$). This result clearly suggested that VAE can separate the representations of the spontaneous cortical activity and the one of the task-evoked activity in the latent space, whereas the linear compressing method ICA failed to delineate the spontaneous activity and task-evoked activity. This result formed the solid basis for us to analyze the dynamics of task-evoked brain activity in the latent space, rather than in the cortical space.

4.3.3 Geometry of Task-Evoked Latent Representation

Next, we asked what the geometry of latent representations driven by the movie-watching task was, and how the representational geometry was different from one of spontaneous brain activity.

To address this, we first estimated the latent representations by feeding movie-watching fMRI data of individuals to the VAE encoder. Given the observed independence between task-evoked brain activity and spontaneous brain activity in the latent space (Figure. 4.7 and 4.8), we averaged the latent representations across the whole population (n=173), yielding the task-evoked latent representations.

Then, we approximated the geometry of task-evoked latent representations by estimating the correlation between latent variables (Figure 4.9.a). Interestingly, the representational geometry of spontaneous activity was shaped like a multi-dimensional Gaussian shape having a higher density at the center (mean and standard deviation of $r = 0.00 \pm 0.05$) while the task-evoked latent representations were elongated (mean and standard deviation of $r = -0.00 \pm 0.25$). By applying the PCA analysis to the latent representations, we further confirmed that the task-evoked latent representations occupied roughly 12.7% (24/189) of spaces that required to explain the spontaneous brain activity (n=24 for task-evoked, and n=189 for spontaneous), as shown in Figure 4.9.b. Interestingly, the crossover of the explained variance between task-evoked activity and spontaneous activity was observed at N=11, and the variance explained by principal components was 1.3% (Figure 4.9.c). After projecting the latent representations to the re-defined subspaces, we estimated the latent gradients by subtracting the adjunct latent representations.

4.3.4 Displacement Magnitude of Latent Representation Is Specific to Movie Contents

Given that the geometry of task-evoked latent representation was elongated, we evaluated the trajectory of brain response by utilizing positional displacement of latent representation – latent gradient, by subtracting temporally adjunct latent representations (Figure 4.10.a). To intuitively understand whether the latent gradients were dynamical over the progression of movie stimuli, we estimated the magnitude of latent gradient. The gradient magnitude has fluctuated dynamically during watching movies (Figure 4.10.b), and we found that the mean gradient magnitudes were significantly different across different video clips involving different contents (one-way ANOVA, *F*=39.96, *p*<0.001, Figure 4.10.c). Among them, the video clip (named as “Northwest”) that showed the smallest gradient magnitude was the only video that did not contain any person-related object. We also observed the adaptation effect that the largest gradient magnitude presented when the subjects saw the video for the first time, while the gradient magnitude decreased as they were watching the same movie stimulus repeatedly (5 > 5’ > 5’’ > 5’’’). To further extend our understanding of what caused the changes in the magnitude of latent gradients, we segmented the gradient magnitudes during inter-session resting periods (n=15). Interestingly, the gradient magnitude increased at the beginning resting period with hemodynamic response delay (t=5, 6, 7), and decreased over the progression of resting period (t=11, 12, …, 22), as shown in Figure 4.10.d. To facilitate our observation intuitively, we zoomed-in the dynamics of gradient magnitude during two short video clips (1: two men, and 11: the garden), as shown in Figure 4.11. Generally, gradient magnitudes tended to increase when scenic changes presented. For example, in the “two men” video clip, the first evident increment in gradient magnitude was related to the change in the objects (from human-crafted objects to the natural scene), followed by other scenic changes (e.g., from body part to human face; from human to another person). In “the garden” video clip, we were able to observe more scenic changes including human vs. natural objects, natural objects vs. human-crafted objects, and natural objects vs. human.

Collectively, these results suggested the latent gradient is dynamically varying over the progression of the movie, and its variance is dependent on movie contents or the absence of movie content.

4.3.5 Individual Variation of Latent Trajectory

Here, we further examined whether the geometry of latent presentation varied across individuals and/or across different neural states, watching “Northwest” video or under the rest. We segmented 15-seconds of latent representations during watching the “Northwest” movie clip (104 seconds after the start of movie), or during the resting period before the start of “Northwest”. In addition, we additionally segmented 15-seconds of latent representations after the start of “Northwest” clip to trace the transition between neural states. We visualized their individual representations in the latent space by reducing its dimensionality from 256 to 2 by using t-SNE method (Figure 4.12). Interestingly, the latent representations were grouped not only by neural states (movie-watching vs. rest) but also by individuals (top panel in Figure 4.12). Interestingly, when the behavioral states transit from the resting state to the movie-watching state, the trajectory of latent representations was continuous for a few seconds after the start of the movie, but the latent representations of individuals were grouped into another cluster eventually (bottom panel in Figure 4.12). More interestingly, the latent representations during the transition period were also separable across different subjects (bottom right panel in Figure 4.12).

In line with findings using rs-fMRI data (Figure 3.6), these results suggest the feasibility of using VAE to characterize and reveal individual variations and variations of neural states simultaneously in the latent space defined by the VAE.

4.3.6 Task-Evoked Latent Variable Is Highly Reproducible

In the previous section, we found that the magnitude of the task-evoked latent gradient was sensitive to the contents of movie clips and the existence of stimulus, both at levels of group and individual. However, it remained unclear what aspects of movie clips were encoded by the gradient direction, reproducibility of task-evoked latent gradients under the same movie stimuli. To answer it, we utilized the "test-retest" movie stimulus in this analysis.

First, we visualized movie scenes consisting of test-retest movie stimulus at the interval of 1 second (Figure 4.13.a). To intuitively understand the trajectory of latent gradient, we visualized the trajectory of the task-evoked latent representations by projecting into the 2D space using the t-SNE mapping method (Figure 4.13.b). We found that the trajectory was relatively rippling for varying periods (1~27 seconds) whereas it suddenly jumped (TR: 4, 21, 48, 50, 61, 66, and 67). We found that most of such jumps coincided with scenic changes (21, 48, 61, 66, and 67), the transition (50 and 67), or the moment when audio started (4, marked as a speaker icon), and the trajectories were highly reproducible for all repeated movie stimuli.

While the current results seemed promising, we concerned that the reproduced patterns were simply driven by the fact that too much information of latent representations was lost by projecting high-dimensional latent representations to the 2-D space. To address this concern, we visualized the trajectory of each principal basis function that defined the latent gradient without information loss, under the repeated movie contents (Figure 4.13.c). We were able to successfully observe the same findings as found in the t-SNE method. Strikingly, the jumps observed in the t-SNE map coincided with peaks or pits of various basis functions. Quantitatively, we found that the dynamics of principal bases were reproducible across different sessions, for all principal basis functions (n=6 per principal basis, one-sample t-test, FDR corrected). We also found that basis functions having less importance tended to have weaker inter-session reproducibility (*r* = -0.75, p<0.01, Pearson correlation). Up to the top-8 basis functions, each basis function explained >1% of the total variance of the latent gradient (Figure 4.13.e).

The inter-session reproducibility of latent gradients was further tested by defining new sets of basis functions using latent gradients of repeated video clips. By applying the same PCA analysis to the latent gradient per session, we found that only 9 basis functions were enough to explain the 99% variance of the latent gradient. Such observation was successfully reproduced across different sessions (Figure 4.14.a). Additionally, we found each set of principal bases were highly reproduced as well as their trajectories (Figure 4.14.b). Interestingly, basis functions estimated from just about 1 minute of data were also similar to the top-9 basis functions obtained from the whole data (>1hour), as shown in Figure 4.14.c.

Collectively, the results support that the dynamics of latent gradients defined by principal basis functions were highly reproducible and specific to movie contents.

4.3.7 Cortical Mapping of Task-Evoked Latent Gradient

Finally, we visualized 24 principal basis functions of the latent gradient in the cortical space (Figure 4.14). Unlike clusters observed in rsfMRI data (Figure 3.5), we found most principal bases of latent gradients were related to one of the sensory networks, somatosensory-, auditory-, or visual networks. For example, cluster 1 precisely formed the language network consisting of well-defined language-related brain regions e.g., associated auditory cortex, Broca’s area, Wernicke's area, supplementary language area (SLA), 40a, etc. Similarly, cluster 7 also formed the language network but followed by the separation of the primary auditory area and the inclusion of the primary visual area. Interestingly, the language network was also observed on the basis 11, while the basis 11 showed the right lateralization as opposed to the left lateralization observed in bases 1 and 7. Besides, we found two bases showed activations in the frontoparietal control network (basis 15) (Dixon et al., 2018) and in the default mode network (basis 17) (Buckner et al., 2008; Greicius et al., 2003; Raichle et al., 2001). Similar to the observation from rsfMRI, most clusters were bilaterally symmetric while we observed the lateralization especially in bases associated with language networks (basis 1, 7, and 11) and also with the default mode network (basis 18). Another interesting cortical network was the basis 5. Basis 5 showed the clear separation between foveal

V1 and peripheral V1, as observable through the retinotopic mapping (Benson et al., 2018). Collectively, we found that each principal basis covered not only well-defined sensory networks e.g., visual, auditory, and somatosensory, but also spanned to cognitive networks e.g., the default mode network and the attention network.

Given that each basis showed the unique interaction between brain networks that differed from each other functionally, we further examined whether different principal bases were responding to different aspects of video stimuli (figure 4.14 bottom). To do so, we extracted various features from the movie stimuli. For low-level features, the image luminosity and audio intensity were extracted, and for middle-level features, presences of face, speech, or text were identified throughout the whole video stimuli. We found that none of principal bases were significantly explained by changes of image luminosity, whereas there were significant positive correlations for the changes in audio intensity (r=0.34±0.07, r=0.33±0.05; mean± standard error of mean, basis 1 and 2). Interestingly, we found that all middle-level features predicted the dynamics of specific principal bases significantly. For the presence of speech feature, we found the basis1 solely showed the strong and significant correlation (r=0.59±0.06; mean± standard error of mean). For the presence of face feature, two principal basis functions (1 and 3) were correlated to that feature while their signs were opposite (for basis 1, r=0.23±0.05; for basis 4, r=-0.41±0.07; mean± standard error of mean). Different from other middle-level visual features, we found four basis functions (for basis 1, r=0.34±0.04; for basis 4, r=0.38±0.06; for basis 4, r=-0.20±0.04 for basis 4, r=-0.36±0.08; mean± standard error of mean) were significantly correlated with the presence of text.

Collectively, our results suggested different basis functions were forming the interactions between cortical networks, and dynamics of some basis functions could be predicted by different aspects of the video.

4.3.8 Semantic Meaning of Task-Evoked Latent Gradient

While some principal bases (7 among 24 bases) were predicted by the low- or middle-level aspects of video stimuli, many principal bases of latent gradients (especially later ones) remained unexplained. We hypothesized dynamics of those principal bases could be explained by semantic aspects of movies. We used the semantic labels provided by the HCP, reflecting the high-level semantic meaning of movie stimuli. Briefly, 23 semantic features among 839 semantic features were chosen based on their definitions. The full list of 23 semantic labels can be found in Figure 4.2 and Table 4.2. Given that, we built an ordinary linear regression model that predicted the dynamics of basis function using HRF-convolved time-series of semantic labels as the predictor. In this analysis, all resting periods were excluded from the regression analysis. As a visual confirmation, we exemplified six basis functions that had the best explained variances, and color-coded semantic features with their positive/negative coefficient weights (Figure 4.16). Interestingly, we found the typical biological vs. non-biological relation in the basis 3; biological features, e.g., "person" and "animal", showed negative and large coefficients while positive and high coefficients were related to non-biological objects e.g., vehicle and artifact. Besides, we found the basis 3 formed the suppressed activation in fusiform face area (FFA) and medial superior temporal (MST) area along with increased activation in the general vision-related brain area. Basis 2 also showed the interesting semantic meaning. The dynamics of basis 2 were predicted by "body part" and "covering" along with "look" features, and the cortical pattern covered the multisensory networks including somatosensory, audio, and visual networks, along with the temporoparietal junction known to be related multisensory convergence (Matsuhashi et al., 2004) and social interaction (Decety and Lamm, 2007). In line with findings from middle-level video features, we also found basis 1 and basis 7 were related to "talk" and "written communication" semantic features, which were related to human language, and basis 4 was related to "person" semantic feature. Fascinatingly, we found the basis 6 had a strong negative weight to "travel" and "travel rapidly" semantic label, and its cortical mapping showed the activation in the inferior frontal junction (IFJ), which is regarded to serve a crucial role in top-down modulation of visual features (Brass et al., 2005; Zanto et al., 2010), and both dorsal and ventral streams that are crucial in deciding "what" and "where" to look in human vision system (Fang and He, 2005; Hebart and Hesselmann, 2012; Milner and Goodale, 1995). Collectively, these results clearly suggested that the well-defined cortical networks were interacted dynamically to process the dynamically varying audiovisual information of movie stimuli.

4.3.9 Gaze Information Is Related to Task-Evoked Latent Gradient

By relating semantic features to the dynamics of the latent gradient, we were able to predict the dynamics of principal basis functions. Still, we observed that dynamics of some basis functions were poorly predicted by any of video features (Figure 4.15 and Figure 4.16). Therefore, we assumed that some portions of latent gradients would be related to endogenous aspects of human perception. To test our hypothesis, we asked whether the inclusion of gaze-informed features, eyes-open duration and horizontal gazing position, in the analysis would be able to further improve the predictability of dynamics of the task-evoked latent gradient. Here, the eyes-open duration feature was used as a surrogate measurement of human visual vigilance level, as reported in (Dinges et al., 1998; Johns et al., 2007; Ong et al., 2013). Interestingly, we found there were improvements on explained variances for some basis functions (especially, basis 2, 5, 6, 7, and 21) while most of them showed nearly no improvement (Figure 4.17.a). By utilizing the Jackknife resampling approach, we found the dynamics of basis 5 were explained best by the eye-open duration feature (Figure 4.17.b), and the basis 5 showed the functional separation between foveal V1 and peripheral V1, along with activations in somatosensory regions (Figure 4.17.c). Our result supported the notion that human visual attention is, at least partially, modulated by the dynamic interaction between foveal V1 and peripheral V1 (Ludwig et al., 2014). With the horizontal gazing position feature, we further found that two basis functions (basis 13 and 17) showed great improvements in their explained variances (Figure 4.17.d). Interestingly, the cortical patterns of two basis functions were somewhat opposite to each other, including the default mode network and the precuneus (Figure 4.17.e). This result supported that the lateralization of brain networks is related to where the human gaze towards (Pelphrey et al., 2003).

Under the same situation, individuals behave and react differently. Indeed, many studies have shown the inter-subject variability in their brain responses across different task paradigms (Lund et al., 2005; Stevens et al., 2012; Xiong et al., 2000). However, whether the origin of inter-subject variation is neuronal or artifactual, it remains debatable (Gaxiola-Valdez and Goodyear, 2012; Lund et al., 2005). Therefore, we asked whether the inter-subject variation on brain responses would be explained by the difference in their gaze positions. First, we visually examined whether individuals gaze at different objects under the same movie stimulus (Figure 4.18.a). Interestingly, we found subjects tended to share similar gaze positions when there was a sole object such as a person, a pen, and a moving bridge. On the other hand, gaze positions of subjects were variant when multiple objects that were similarly important in scenes, e.g., a conversation between two people, or a human and visually appealing subtitles. We applied IS-RSA to evaluate whether inter-subject distance map of gaze positions predicted the inter-subject variability in latent representations (Figure 4.18.b). For example, when there were Han Solo and the commander together in the scene, subjects decided to gaze one of them. The results clearly showed that subjects who were sharing their gaze locations tended to have similar latent representations (Pearson correlational analysis, r=0.47, p<0.01). By repeating the analysis for each scene, we found there was a significant mean representational similarity (Figure 4.18.c, n=3,105, average r = 0.12, one-sample t-test). We further found that the strength of predictability was variant by scenic contents (Figure 4.18.d; with person feature; n=2,122, average r=0.13±0.002; without person feature, n=983, average r=0.08±0.002; no feature, n=546, average r=0.06±0.005, one-way ANOVA, F=144.43, two-sample t-test, p<0.01). In sum, our results suggested that the inter-subject variability in the brain response is, at least partially, neural, which can be partially explained by the inter-subject variability in the gaze location.

4.4 Discussion

Here, we present that the VAE model pretrained to learn the representation of cortical rsfMRI activity is generalizable to learn latent representation of cortical activity under the movie-watching task without any further finetuning procedure. Our results suggest that our VAE model can delineate the task-evoked cortical patterns from the spontaneous cortical patterns by effectively suppressing the non-linear task-rest interaction and discover the principal basis functions tracing the dynamics of task-evoked cortical patterns. We further observe that each principal basis function consists of overlapping brain networks spanning from multisensory networks to the cognitive networks, the attention network and the default mode network. We reveal dynamics of the latent gradient can be explained by the appearance/disappearance of the low- and middle-level video-related features e.g., audio intensity, presence of face, presence of speech, and presence of word, while extended predictions are achieved by the semantic meaning of video. Interestingly, we find that the unexplained dynamics of the task-evoked latent gradient are uniquely explained by changes in the group-level gaze location and the dynamics of visual vigilance level approximated by the eye-open duration. Last but not least, we show that the inter-subject variability in brain response is partially explained by the variability in their gaze positions. As validated in fMRI under the movie-watching task, we expect this method to be a highly valuable tool for investigating the origins of brain activity under diverse neural states and disease conditions, due to its superior generalizability.

When we surveyed the data-driven deep learning models to learn useful representations of rsfMRI data, we prioritized to guarantee the generalizability of the model across fMRI data under different neural states. Against other deep generative models such as the generative adversarial network (GAN) and its family models that show the great reconstruction ability, we chose VAE due to its reliable and robust interpretability of latent representations over various types of inputs without possible critical failures such as a mode collapse frequently reported in GAN model (Goodfellow et al., 2014, 2020; Thanh-Tung and Tran, 2018). We find that our VAE model pretrained in 3T rsfMRI data is highly generalizable against various types of fMRI not only recorded under different imaging protocols but also under the brain state receiving and processing rich and dynamical audiovisual sensory information (Figure 4.5). Given the promising result in the movie-watching fMRI data, we believe the pretrained VAE will be able to learn representations of other task-fMRI data and even to fMRI data under different disease conditions. Hence, the VAE can be a valuable and generalizable tool making it possible to directly compare representational geometries between different brain states and/or between different disease conditions.

One defining property of VAE making it distinctive from the conventional blind source separation methods, e.g., ICA and PCA, is its non-linearity originated from convolutional layers that are progressively compressing complex and non-linear input to simple and independent latent variables. Therefore, the non-linear relationship between observations and generative factors can be properly disentangled by VAE while presumably not feasible with linear compressing methods. Such characteristics of VAE enabled us to separate the task-evoked brain activity from the mixture of task-evoked and spontaneous brain activities, which was impossible with the linear dimension reductions method ICA (Figure 4.7 and Figure 4.8). In other words, the linear superposition between task-evoked activity and spontaneous activity is valid only in the latent space non-linearly derived VAE. While the VAE was forced to learn the non-linear mapping of generative factors governing the spontaneous cortical patterns, there was no explicit constraint addressing the task-rest interaction since movie-watching fMRI was not introduced during training the VAE model. Then, why does the negative task-rest interaction diminish in the latent space defined by the VAE? One possible explanation is the negative task-rest interaction observed in the cortical space is magnified by the non-linear and complex hemodynamic relationship between fMRI activity and neural activity, as reported in concurrent fMRI-LFP study (Schölvinck et al., 2010), fMRI-iEEG studies (Lachaux et al., 2007; Murta et al., 2016; Ridley et al., 2017), and fMRI-EEG studies

(Hanslmayr et al., 2011; Haufe et al., 2018). In fact, even after considering the non-linearity of fMRI cortical patterns, there remains dissimilarity ($r^2 = 1 - 0.59$) between task-evoked covariance matrix and movie-watching vs. rest covariance matrix, suggesting this dissimilarity may reflect the true negative task-rest interaction during the movie-watching task. This hypothesis awaits confirmation by future studies.

The human brain dynamically modulates the functional interactions between brain networks accomplished by neural ensembles, to enable the diverse cognitive and perceptual processes (Buschman et al., 2012). In line with the electrophysiology study (Buschman et al., 2012), one fMRI study showed that the functional organization of the brain, characterized by the functional connectivity between brain regions, is dynamically changing its shape over the transitions between intelligent behavioral phases (Gonzalez-Castillo et al., 2015). Given the sluggish nature of fMRI data, measuring the dynamics of brain organizations has been commonly done using the time-varying connectivity analysis that estimates the functional connectivity by sliding the short window. While this time-varying functional connectivity analysis has proved its utility in many applications (Allen et al., 2014; Betzel et al., 2020; Bolton et al., 2020; Calhoun et al., 2014), one critical challenge in the method has been spotted that the window length balancing the trade-off between the temporal dynamics and reliability of measured functional connectivity should be chosen carefully (Leonardi and Van De Ville, 2015). In this study, instead of using the window, we utilized the positional displacement of task-evoked latent representations as a primary tool for understanding the brain dynamics under the movie-watching neural state. We discovered that the gradient magnitude was sensitive to different video contents (Figure 4.10), and the gradient direction, defined by principal basis functions, was specific to changes in various objects (Figure 4.13). Collectively, we believe tracing the trajectory of representations in the latent space can be a useful alternative to characterize the dynamics of brain networks.

Our analysis with VAE opens a new venue for mapping task-evoked functional networks of the brain. Here, we defined basis functions, which explain the dynamics of the latent gradient, as the network interactions, opposed to the inter-subject FC analysis or condition-specific FC analysis (compare FC during movie-watching vs. FC during resting-state) (Betzel et al., 2020; Demirtaş et al., 2019; Hasson et al., 2004; Vanderwal et al., 2017). Among two methods, only condition-specific analysis have reported negative FCs, but those FCs have been overlooked because they have been assumed to originate from the negative task-rest interaction (Demirtaş et al., 2019; Lynch et al., 2018). In the current study, even after effectively diminishing the negative task-rest interaction, we found some of basis functions of task-evoked latent gradient mapped the negative interaction between cortical networks covering not only sensory networks but also cognitive networks e.g., the default mode network and the dorsal attention network (Figure 4.15). Furthermore, those basis functions were dynamically varying according to semantic aspects of movie stimuli (Figure 4.16) and the task-relevant gaze information (Figure 4.17). Collectively, our results support the notion that the negative synchrony between brain networks is risen by the neural bases and the dynamical network interactions form the basis of perceptual processes in daily life.

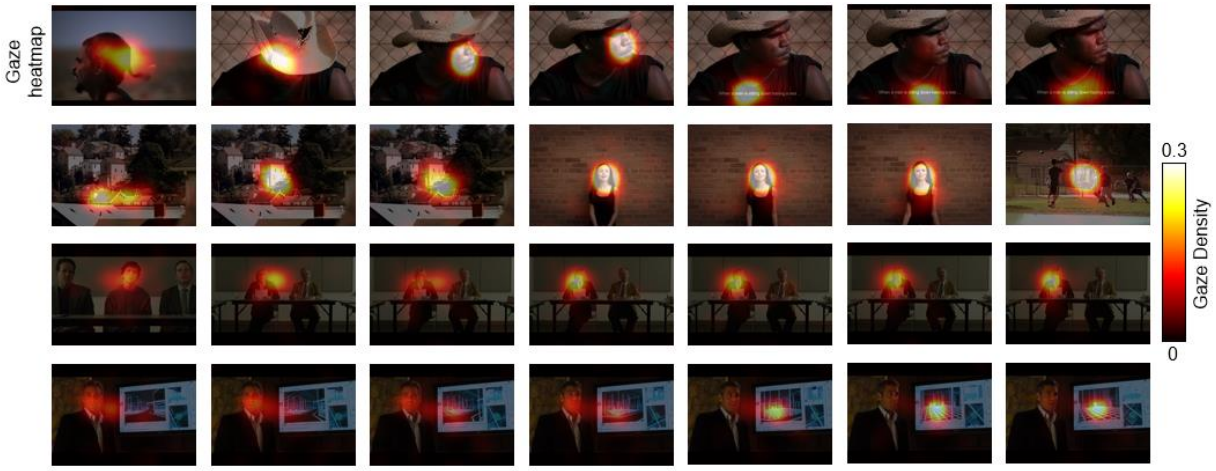


Figure 4.1. Example of group-level eye gaze heat map. Four 7-seconds movie clips with group-level gaze heatmap were presented. Each row stands for each movie clip. The temporal resolution of the movie scene is 1 second.

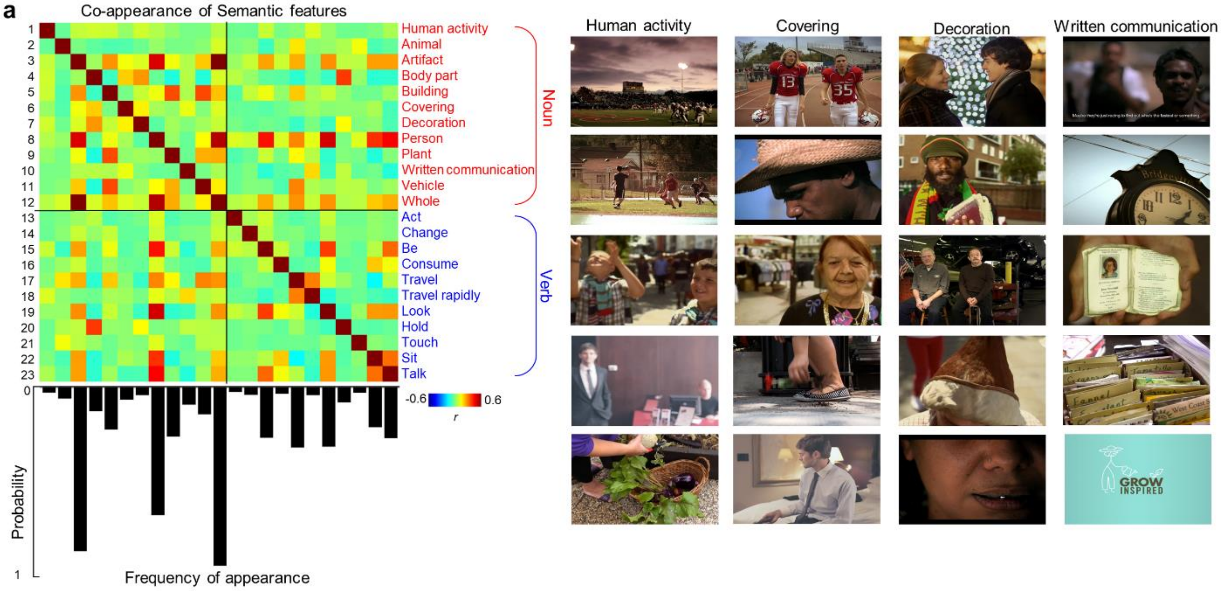


Figure 4.2. Semantic label of movie stimuli. (a) Appearance frequency of semantic label (bottom) and co-appearance frequency between semantic labels (top). (b) Few scenic examples of semantic labels.

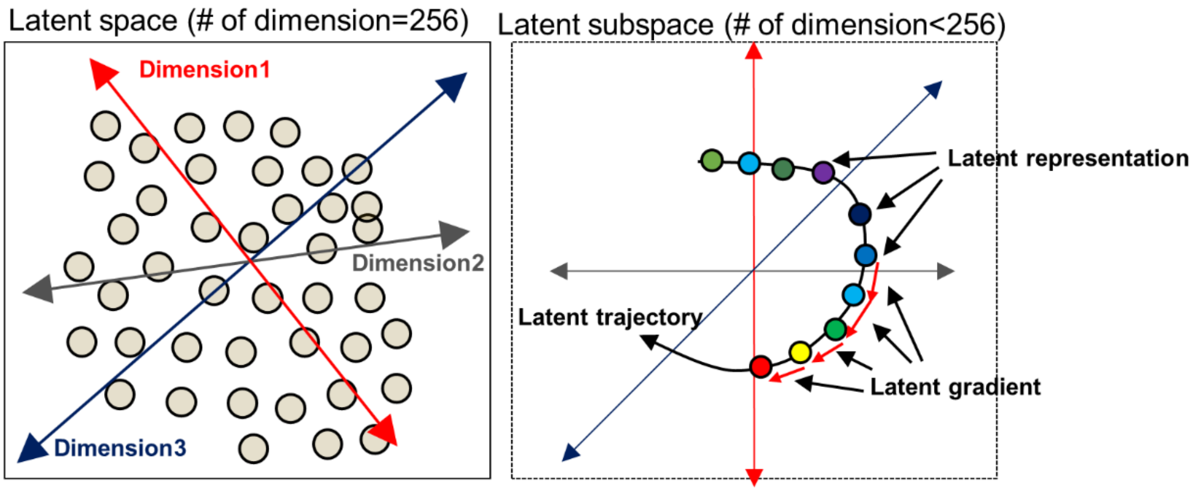


Figure 4.3. Definition of representation, gradient, and trajectory in the latent space. (Left) From the original latent space (# of dimension=256), principal subspaces explaining the task-evoked representations are defined using PCA. (Right) Task-evoked latent representation stands for a latent representation of a single fMRI time point averaged across subjects. Task-evoked latent trajectory is a trend how latent representations are traveling over the progression of movie stimulus. Task-evoked latent gradient, i.e., positional displacement of task-evoked representation, stands for the subtraction between temporally adjunct latent representations.

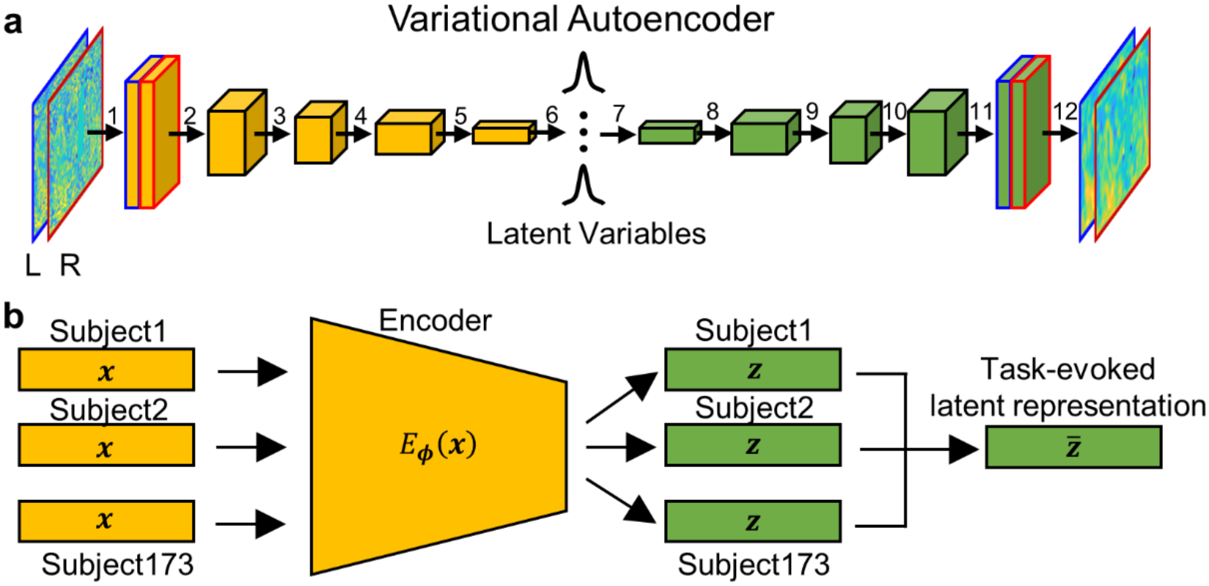


Figure 4.4. Illustration of VAE model and task-evoked latent representation. (a) VAE Model architecture. The VAE model consists of two compartments, the encoder (coded as yellow) and the decoder (coded as green). The convolution operations are defined as: 1: convolution (kernel size=8, stride=2, padding=3) with rectified nonlinearity, 2-5: convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 6: fully-connected layer with re-parametrization, 7: fully-connected layer with rectified nonlinearity, 8-11: transposed convolution (kernel size=4, stride=2, padding=1) with rectified nonlinearity, 12: transposed convolution (kernel size=8, stride=2, padding=3). Blue and red boxes stand for the input images from left and right hemispheres, respectively. (b) VAE encoder with learnt parameters ϕ outputs latent representations ($\boldsymbol{z}$) given individual's fMRI data ($\boldsymbol{x}$). Latent representations ($\bar{\boldsymbol{z}}$) evoked by the movie-watching task are estimated by averaging latent representations across 173 subjects.

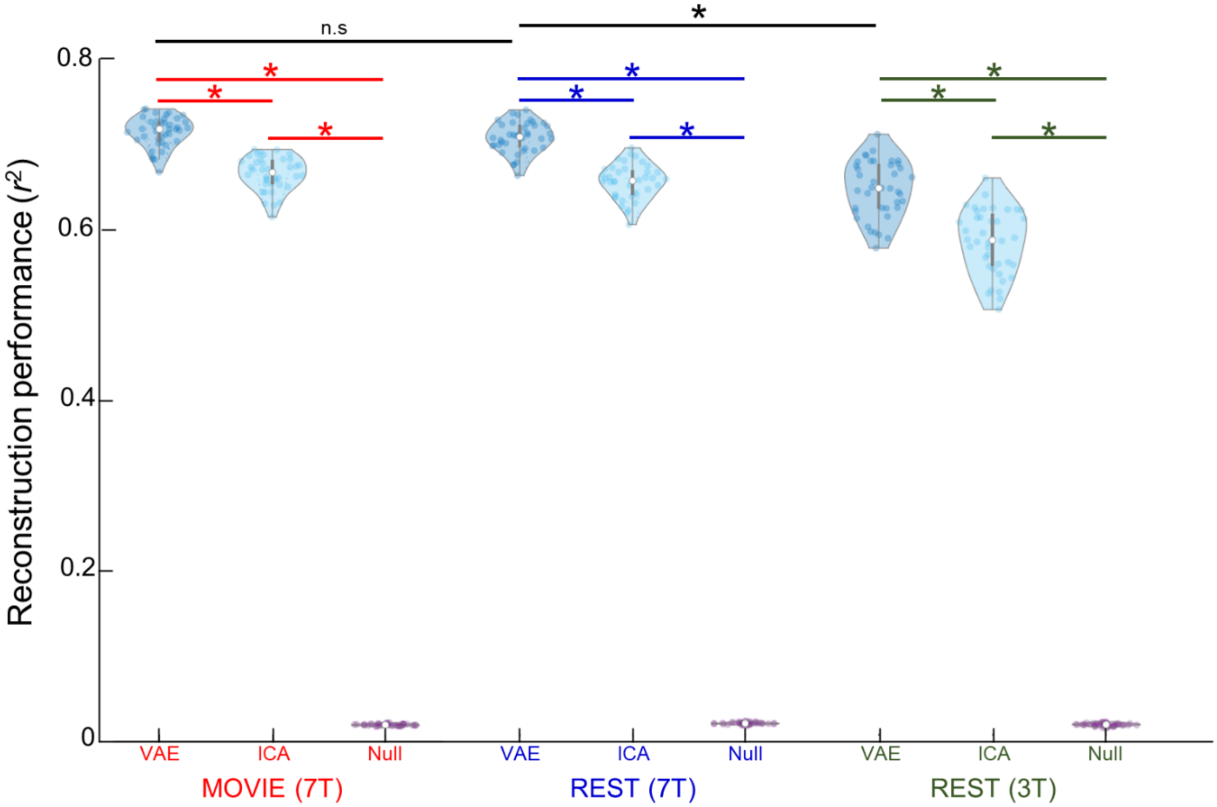


Figure 4.5. Generalizability of pretrained VAE model. Three different datasets (red, 7T movie-watching fMRI data; blue, 7T resting-state fMRI; green, 3T resting-state fMRI) were reconstructed by VAE and independent component analysis (ICA), as a linear counterpart of VAE model. Reconstruction performances of fMRI cortical patterns are measured by squared correlation coefficient. The null distributions were estimated by estimating the similarity between synthesized cortical patterns and smoothed cortical input patterns. *: Bonferroni-corrected $p<0.01$, two-sample t-test for testing across different datasets (black lines) and paired t-test for testing within the datasets (colored lines), n=40 subjects. n.s: not significant, Bonferroni-corrected $p>0.05$.

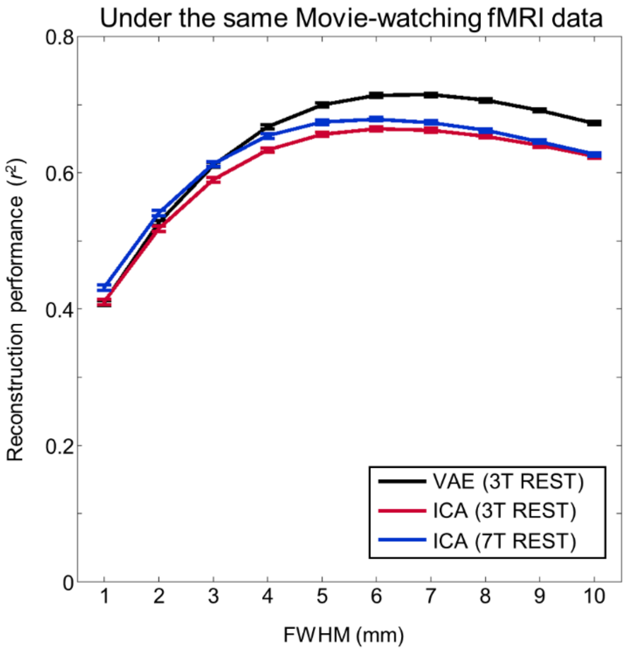


Figure 4.6. Smoothing effect of VAE vs. ICA. The reconstruction performance, in terms of the percentage of variance in the fMRI images as explained by the model reconstruction, is shown for VAE, ICA, and ICA estimated from 7T rsfMRI, as a function of FWHM (from 1 to 10 mm) applied to the spatial smoothing of the fMRI images. The error bar stands for the standard error of mean.

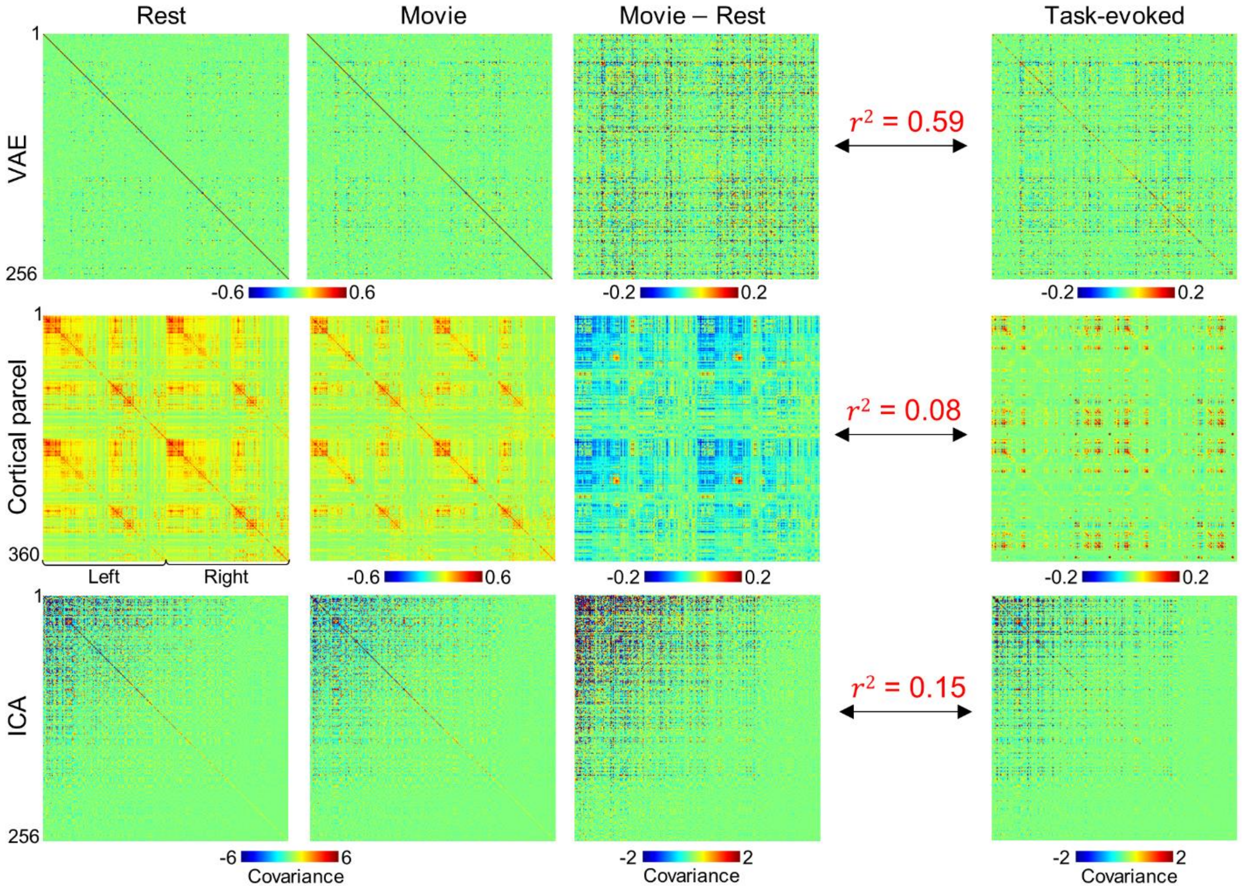


Figure 4.7. Linear superposition between task-evoked activity and spontaneous activity in different latent spaces. fMRI data under movie-watching task (1st column) and resting state (2nd column) are compressed into latent spaces defined by VAE (top), or latent-spaces by cortical parcels (middle) or ICA (bottom). For each latent space, covariance matrix between latent variables is measured. Additionally, task-evoked covariance matrix is acquired by estimating correlation between group-level latent variables. Similarity between condition-specific covariance matrix (3rd column, movie-watching vs. rest) and task-evoked covariance matrix is quantified by measuring squared correlation value between off-diagonal elements of two matrices.

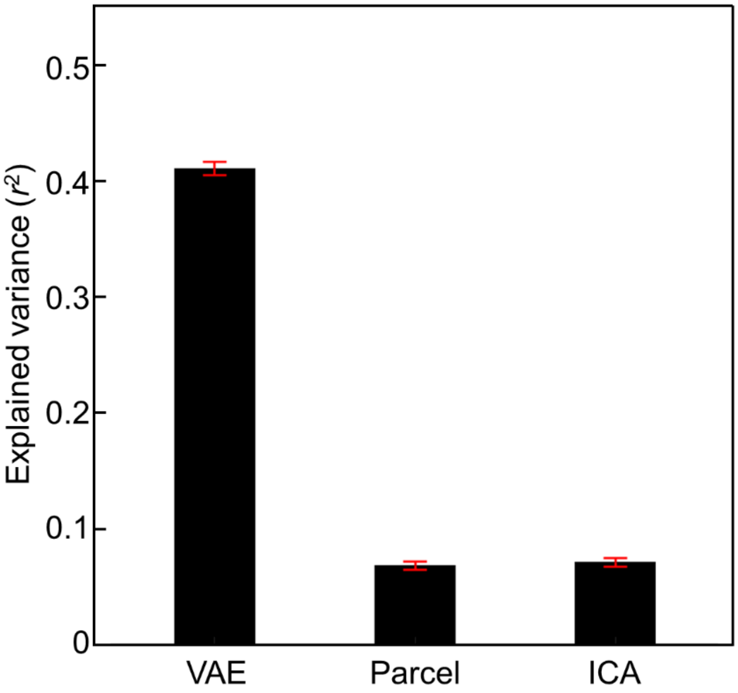


Figure 4.8. Linear superposition between task-evoked activity and spontaneous activity in the partitioned dataset. By partitioning 173 subjects into ten subsamples (n=17), similarity between condition-specific correlation matrix and task-evoked correlation matrix is quantified by measuring squared correlation value between off-diagonal elements of two matrices. Error bar stands for standard deviation of mean. Explained variance by VAE-derived latent space is significantly higher than parcel- or ICA-based latent space ($p<0.01$ for parcel; $p<0.01$ for ICA). There is also significant difference between parcel-based latent space and ICA-based latent space ($p<0.01$). One-way ANOVA, F=863.34, $p<0.01$, post-hoc paired t-test, Bonferroni correction.

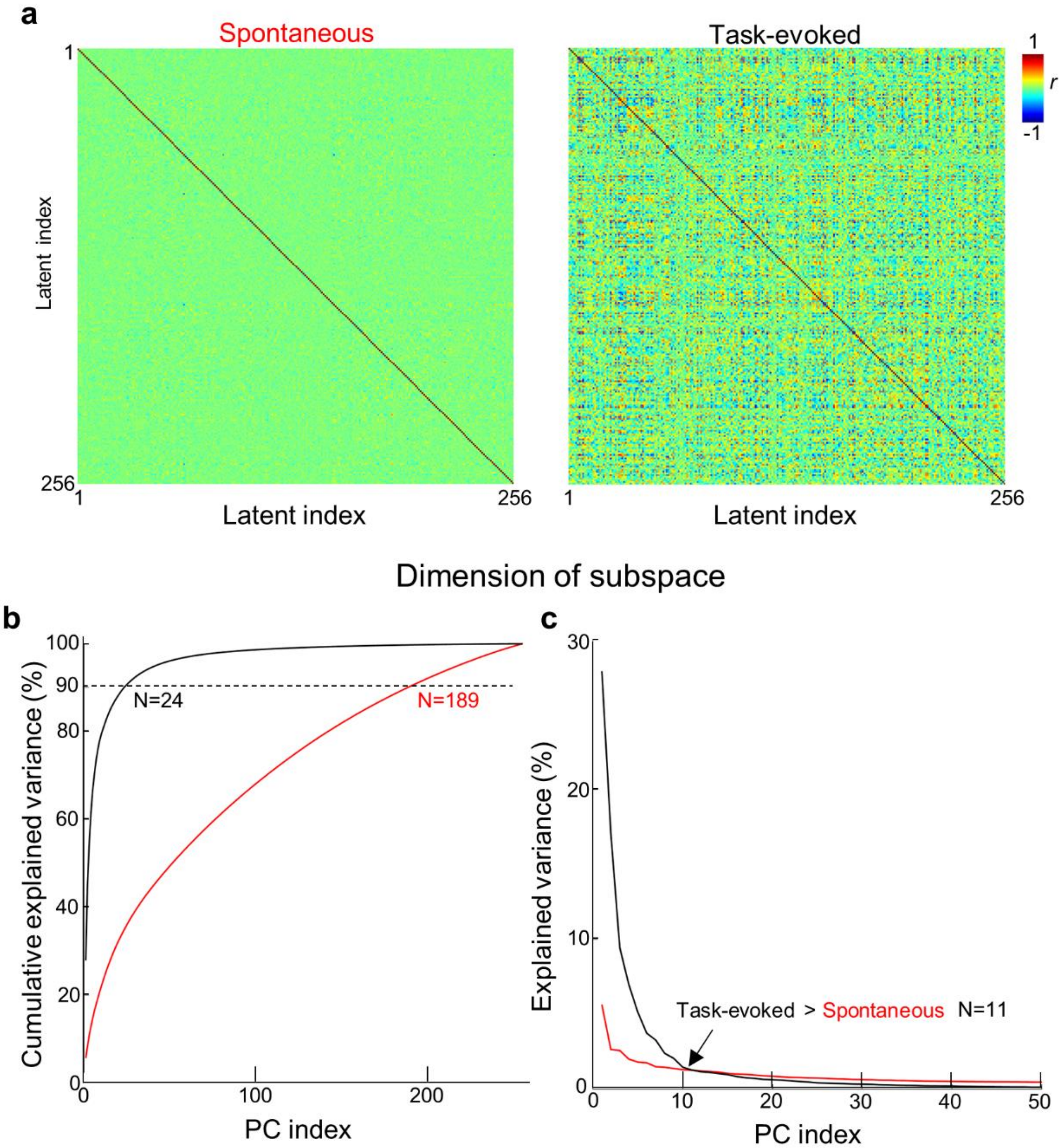


Figure 4.9. Geometry of task-evoked latent representation and spontaneous latent representation. (a) Cross correlation matrix between latent variables, under different brain states. (b) Variance explained by principal components. The number of principal components required to explain 90 % of latent variables are counted (dashed line). (c) Variance explained by each principal component. The black arrow stands for the crossover.

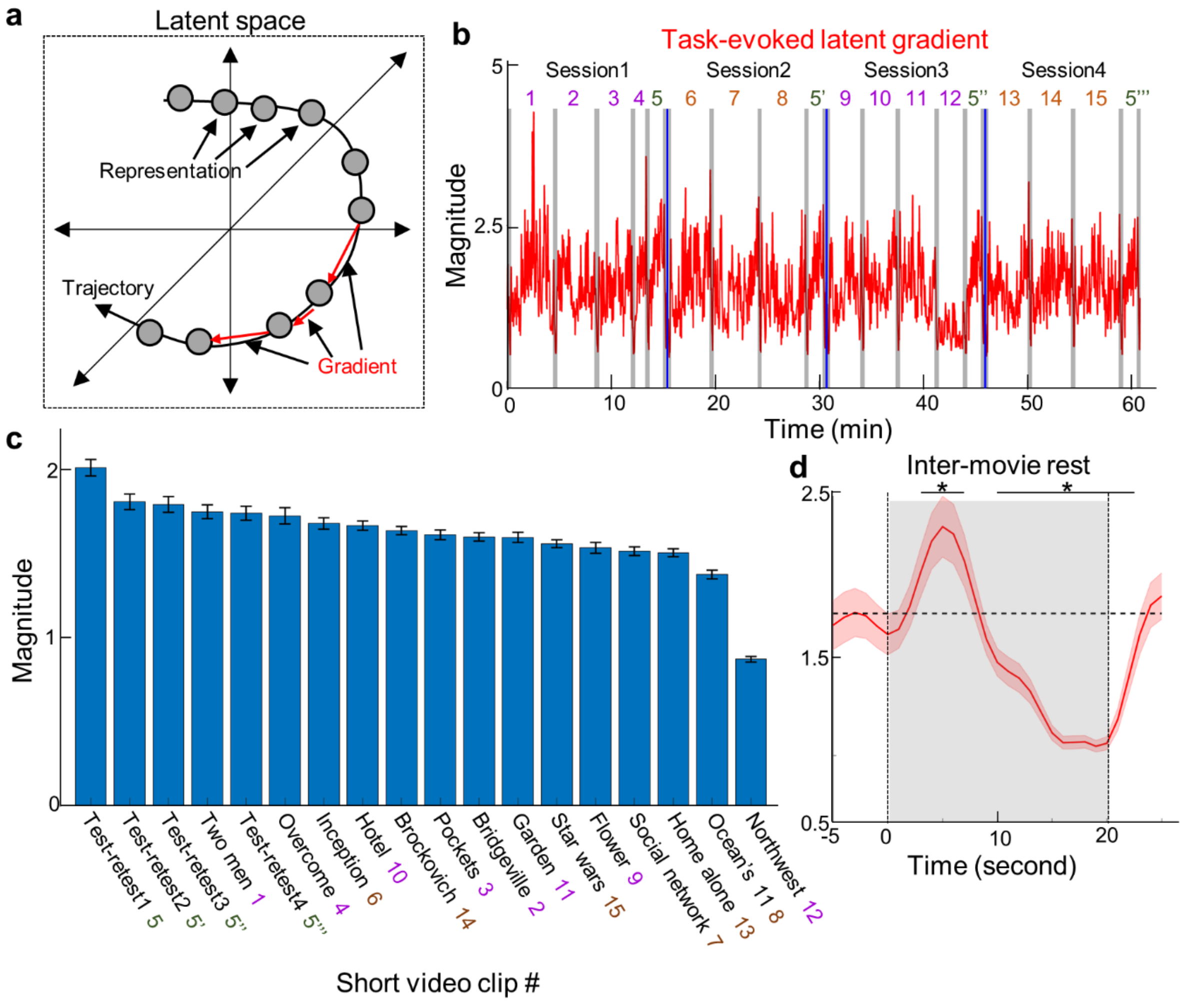


Figure 4.10. Displacement magnitude of latent representation is specific to movie contents. (a) Representation, gradient, and trajectory defined in the latent space. (b) Magnitude of task-evoked latent gradient as a function of time. Gray box stands for inter-movie resting period (20 seconds). Four recording runs are concatenated. Blue line stands for the boundary between different sessions. The number stands for the numbering of short movie clips (purple: from Vimeo, orange: from Hollywood movie). One movie clip (5, 5', 5'', and 5''') is repeated over recordings. (c) Average magnitude of task-evoked latent gradient for different movie clips. There is a statistically significant difference in the mean magnitude between video clips (one-way ANOVA, F= 39.96, p<0.001). (d) The dynamics of magnitude over the progression of inter-movie rest periods (n=15), including 5 seconds earlier and 10 seconds later of rest period. Dashed line stands for averaged magnitude during movie stimuli present. *: q<0.05, FDR-corrected.

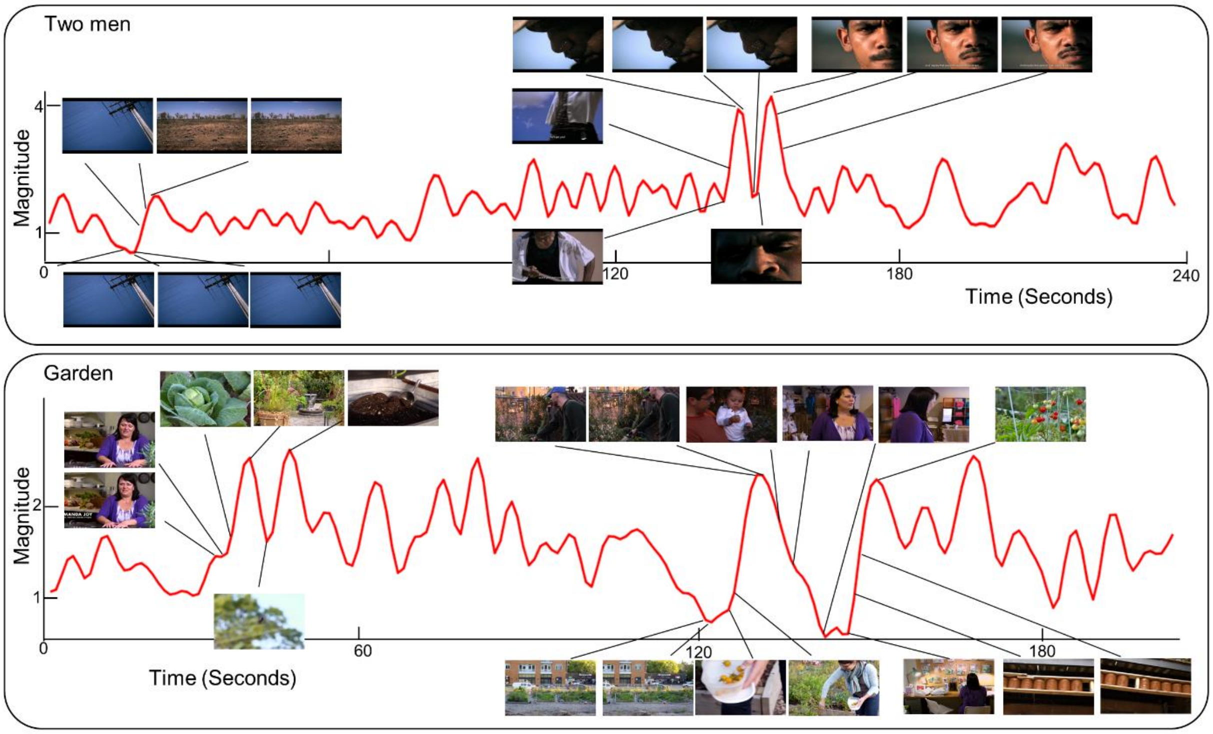


Figure 4.11. Example of scenic changes related to the magnitude of task-evoked latent gradient. From two short video clips (1: Two men; top and 11: Garden; bottom), movie scenes corresponding to lower or higher magnitude of latent gradient are visualized, after considering hemodynamic delay (5 seconds).

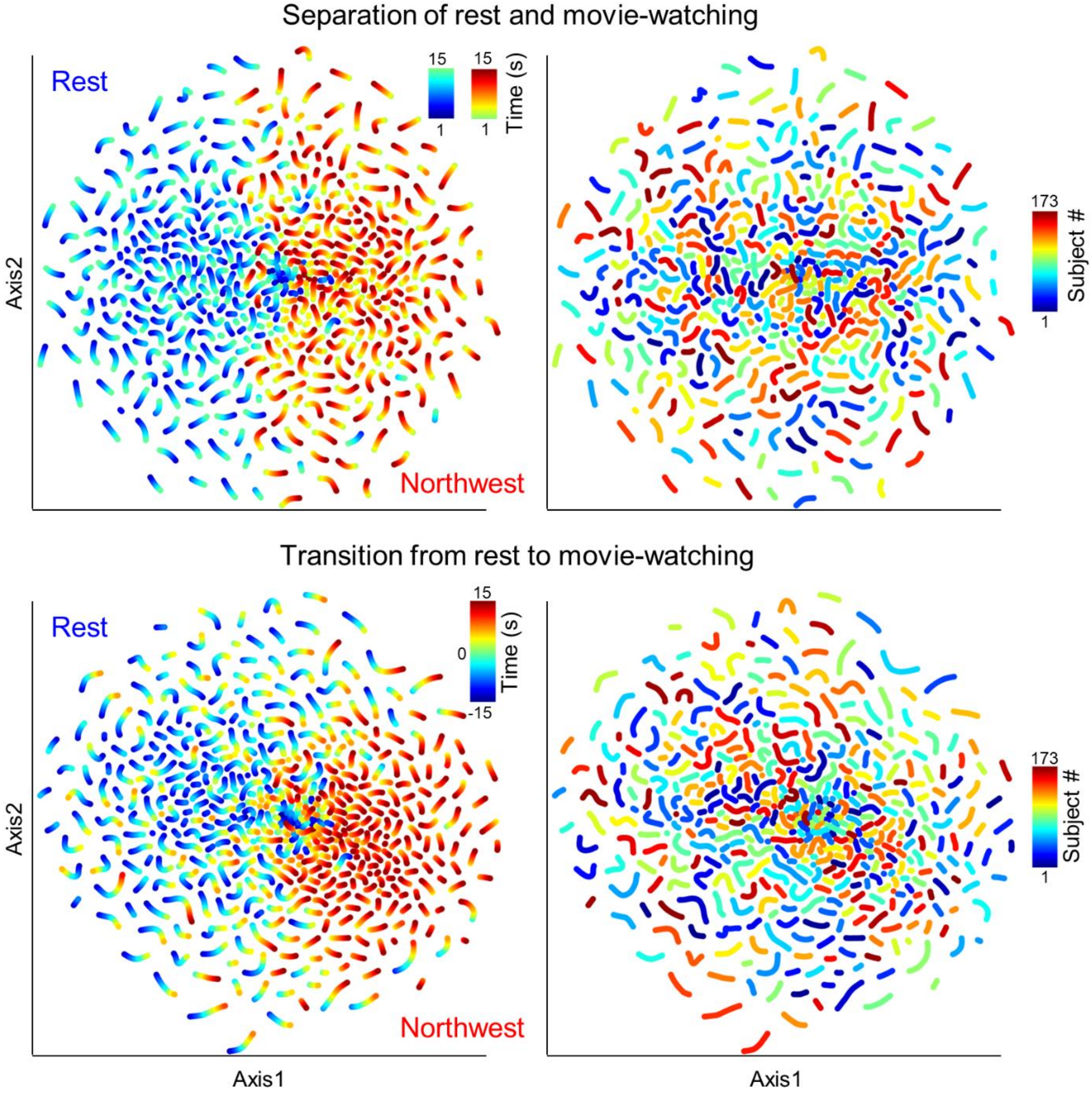


Figure 4.12 The trajectory of latent representations of individuals. t-SNE map of latent representations during watching the movie (Northwest, 104 seconds after the starting of movie), rest, or the transition from rest to movie-watching (bottom). Representations are color-coded by the task conditions (left), and subject indices (right).

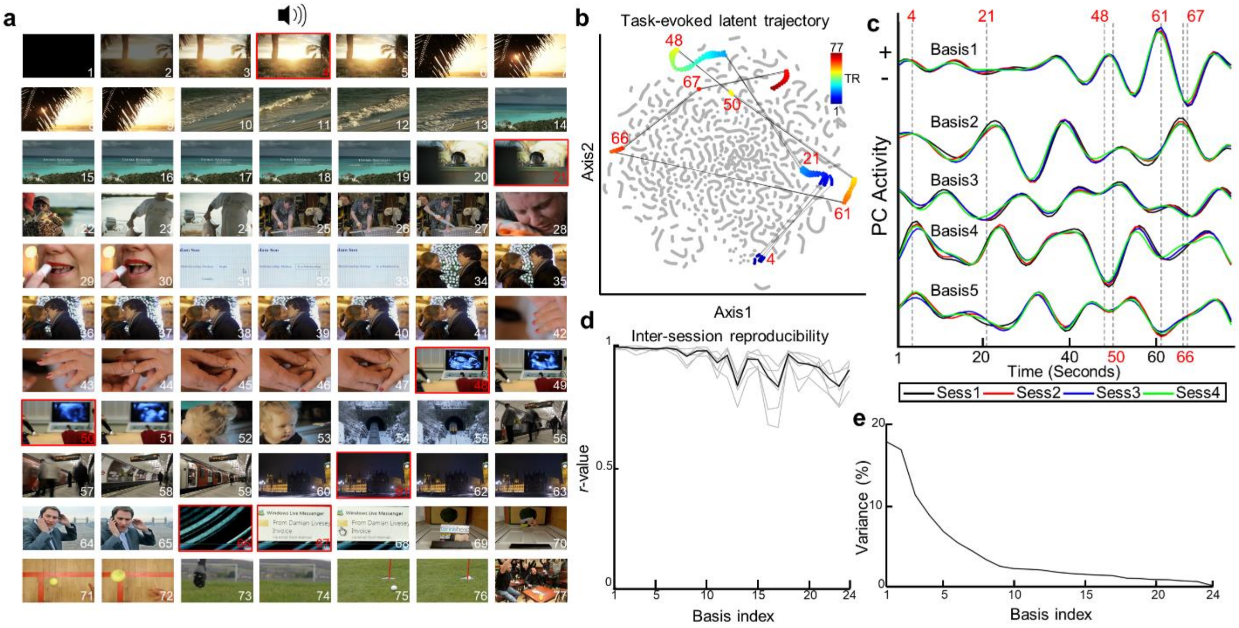


Figure 4.13. Reproducibility of task-evoked latent gradient. (a) scenes of a short video clip (5; test-retest) are visualized at the pace of 1 seconds. Speaker image stands for the starting of audio. Red box and red number stand for the time points that have big jumps in panel (b). Hemodynamic delay (5 seconds) is considered. (b) t-SNE map of task-evoked representations during watching test-rest video clips (color-coded, 4 sessions). Gray dots stand for task-evoked representations from other movie-watching periods. (c) Dynamics of top-5 basis functions explaining the task-evoked latent gradient. Big jumps found in (b) are marked as dashed lines and red numbers. (d) The inter-session reproducibility of dynamics of basis functions. Gray thin line stands for each pair between sessions (n=6). Black thick line stands for the averaged reproducibility across 6 pairs of sessions. (e) The variance of latent gradient explained by each basis.

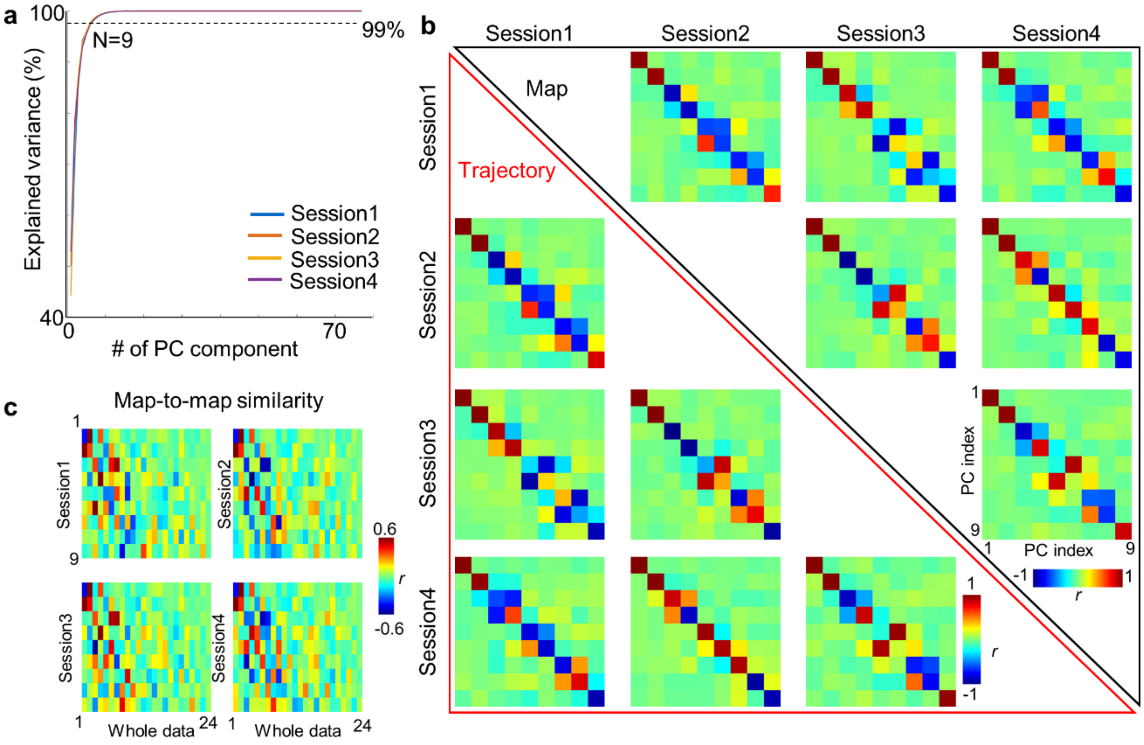


Figure 4.14. Reproducibility of basis functions defined by repeated datasets. The cumulative variance of latent gradient explained by basis functions. For four sessions, 9 basis functions explained 99% variance (dashed line). (b) Inter-session similarity of basis functions (black upper triangle) and their dynamics (red lower triangle). (c) Similarity between basis functions estimated from the whole data and basis functions defined from repeated dataset.

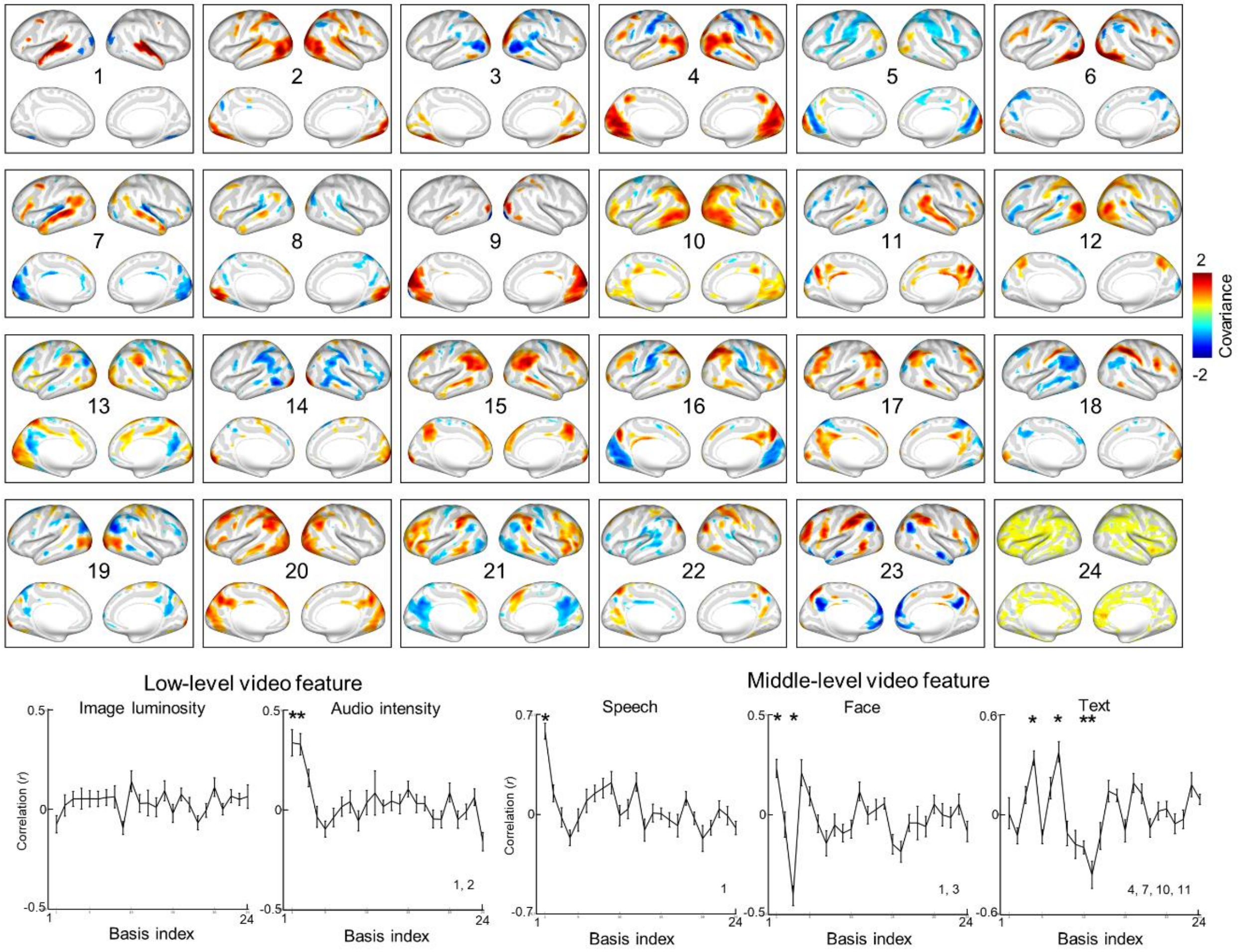


Figure 4.15. Cortical mapping of task-evoked gradient in the latent space and its predictability by video features. (Top) Cortical mapping of 24 basis functions defining the task-evoked latent gradient. (Bottom) Per short video clip, the dynamics of basis functions are correlated with low- (left panel) and middle-level video features (right panel). Error bar stands for the standard error of mean. The number in bottom right stands for basis functions having Fisher's z-values different from zero. *: q<0.05, FDR-correction, one-sample t-test.

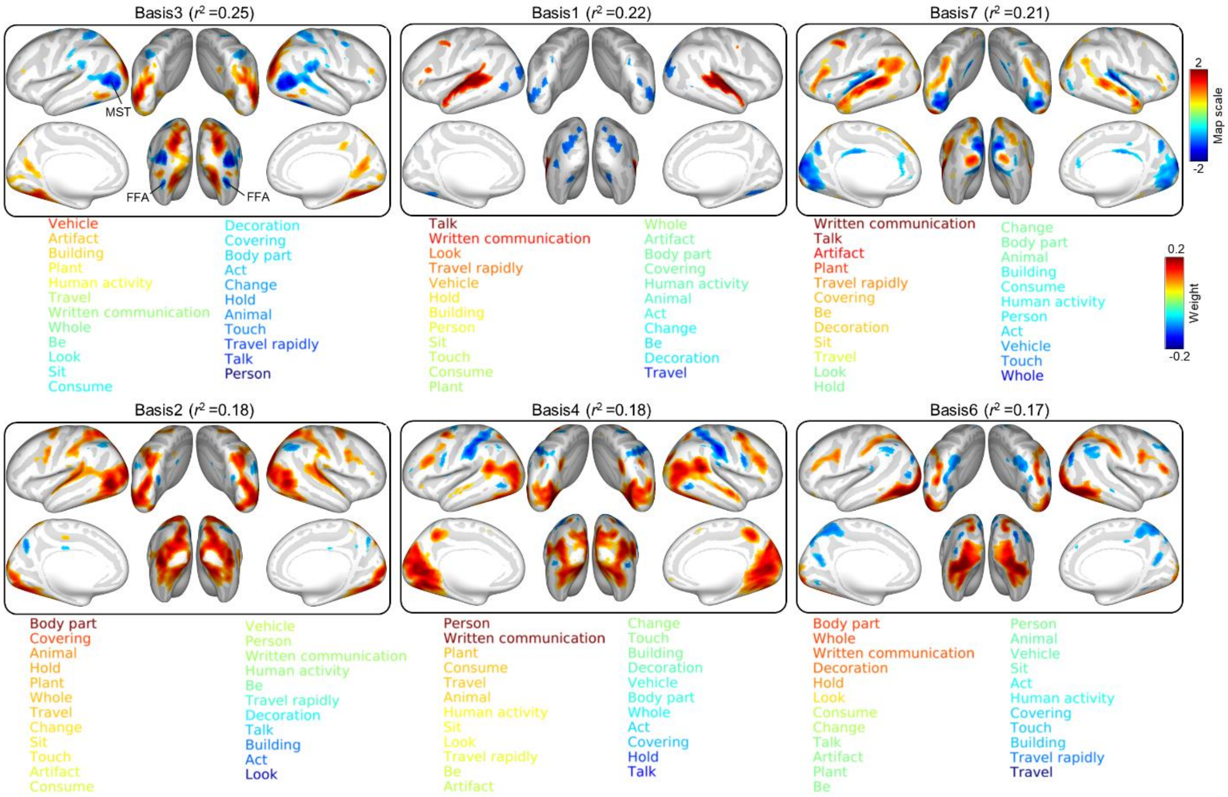


Figure 4.16. Semantic meaning of task-evoked latent gradient. Top-6 basis functions having highest explained variance by regression model consisting of HRF-convolved semantic features (from left top to right bottom). Per basis function, each semantic label is color-coded by its coefficient weight in the fitted ordinary regression model.

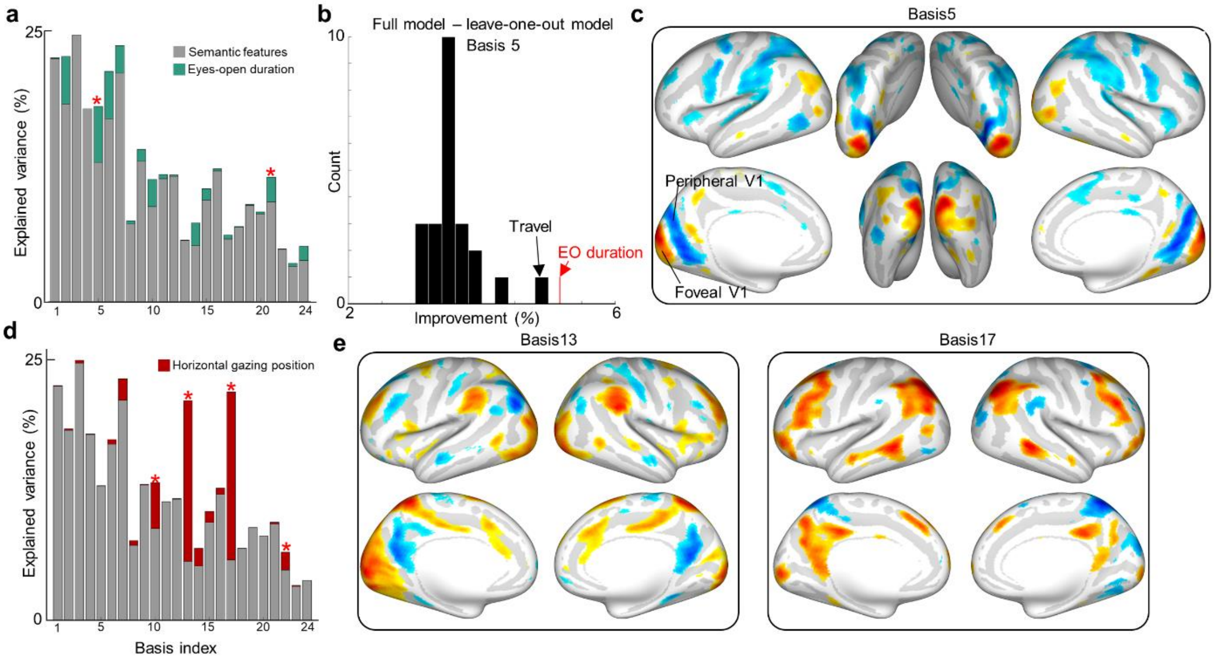


Figure 4.17. Task-specific gaze position is related to task-evoked latent gradient. (a) and (d) The variance of each basis function explained by regression model only with semantic features (gray), with semantic features and changes in eye-open duration feature (green), or with semantic features and changes in horizontal gazing position feature (dark red). The red star stands for the basis function that shows highest explained variance by gaze-informed feature, as estimated in panel (b). (b) The improvement of explained variance by each semantic feature or eye-open duration feature, in basis function 5. (c) and (e) The cortical mapping of basis functions showing great improvement by gaze-informed features.

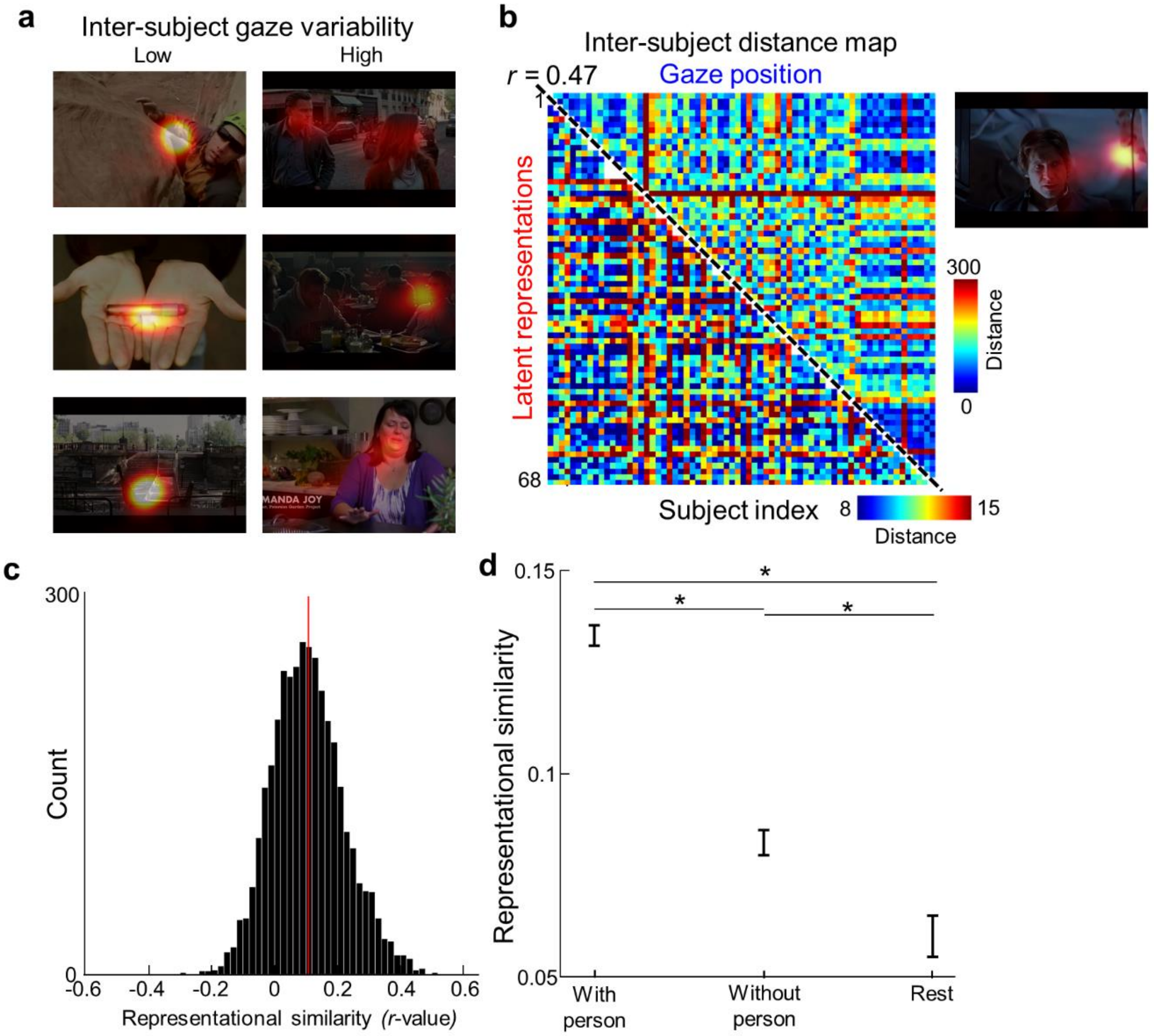


Figure 4.18. Individual variability in gaze position explains individual variability in brain response. (a) Exemplified movie scenes having lower (left) or higher (right) inter-subject variability in gaze position. (b) Per fMRI data point, Euclidean distance between individual's gaze position (upper triangle) or between individual's brain response in the latent space (lower triangle). These distance matrices are acquired when the specific scene (right) presents. (c) Per time point of fMRI data, the representational similarity between two distance matrices is measured and plotted as a histogram. Red line stands for the average representational similarity. The center of distribution (red line) is not 0 (p<0.01, one-sample t-test). (d) The representational similarity is grouped based on the presence of face or during inter-session rest period. One-way ANOVA, F= 144.43, p<0.01, post-hoc two-sample t-test, FDR-corrected.

Table 4.1 Description of movie stimuli

# (Sess)	Short title	Original title (year)	Source	TR	Features	Short description	Note
1 (1)	Two men	Two Men (2009)	Vimeo	245	Speech; Face; Text	A man see two men running and talk about possible motives	With subtitle and narration
2 (1)	Bridgeville	Welcome to Bridgeville (2011)	Vimeo	221	Speech; Face; Text	People explain why they love their small town	
3 (1)	Pockets	Pockets (2008)	Vimeo	189	Speech; Face; Text	People explain their items in pocket and significance of items	
4 (1)	Overcome	Inside the Human Body (2011)	Vimeo	65	Speech; Face	Short documentary about people overcome physical disabilities	With narration
5	Test-retest	23 Degrees South, LXIV (2011)	Vimeo	84	Speech; Face; Text	Concatenation of short videos (2-6 s) covering various objects	
6 (2)	Inception	Inception (2010)	Hollywood	228	Speech; Face	A man introduce how one can manipulate another person’s dream	Metaphysical e.g., Moving building
7 (2)	Social Network	The social network (2010)	Hollywood	260	Speech; Face; Text	Mark Zuckerberg’s disciplinary hearing and its aftermath	
8 (2)	Ocean’s 11	Ocean’s Eleven (2001)	Hollywood	251	Speech; Face	A man explains the plot heisting a casino with visual aids	
5’	Test-retest						
9 (3)	Flower	Off The Shelf (2008)	Vimeo	181		A journey of a flower from the house to the field (with music)	Only music
10 (3)	Hotel	1212	Vimeo	185	Face	Metaphysical encounter between a man and woman in a hotel room	
11 (3)	Garden	Mrs. Meyer’s Clean (2013)	Vimeo	205	Face; Text	Documentary about a woman who runs an urban garden community	
12 (3)	Northwest	Northwest Passage	Vimeo	143	Text	Portray of wet landscapes and abandoned buildings	No human presented
5’’	Test-retest						
13 (4)	Home alone	Home alone (1990)	Hollywood	234	Speech; Face	A kid is searching his home, realizing he is alone at home	Poor video quality
14 (4)	Brockovich	Erin Brockovich (2000)	Hollywood	231	Speech; Face	A woman talks with a woman and visits a legal office with her children	
15 (4)	Star wars	The Empire Strikes Back (1980)	Hollywood	257	Speech; Face	A scene on an icy planet and Han Solo visit headquarter of rebellion	
5’’’	Test-retest						

Table 4.2 Description of semantic labels

Numbering	Label used in the study	Wordnet	Numbering	Label used in the study	Wordnet
1	Human activity	Act.n.02	13	Act	Act.v.02
2	Animal	Animal.n.01	14	Change	Change.v.02
3	Artifact	Artifact.n.01	15	Be	Be.v.01
4	Body part	Body_part.n.01	16	Consume	Consume.v.02
5	Building	Building.n.01	17	Travel	Travel.v.01
6	Covering	Covering.n.02	18	Travel rapidly	Travel_rapidly.v.01
7	Decoration	Decoration.n.01	19	Look	Look.v.01
8	Person	Person.n.01	20	Hold	Hold.v.02
9	Plant	Plant.n.01	21	Touch	Touch.v.01
10	Written communication	written_communication.n.01	22	Sit	Sit.v.01
11	Vehicle	Vehicle.n.01	23	Talk	Talk.v.02
12	Whole	Whole.n.02			

5. CONCLUSION

5.1 Conclusion

The work presented in the thesis has established an unsupervised deep generative model to learn representations embedded in the observed fMRI data under different brain conditions, the resting state and the movie-watching state. By combining the VAE model, originally introduced by (Higgins et al., 2017; Kingma and Welling, 2013), with a new reformatting strategy of input fMRI data, the VAE model proposed here was highly efficient in terms of model training without losing useful information of input data. We found this model was highly generalizable to various fMRI datasets, requiring no further fine-tuning step. Therefore, I expect this model will be generalizable to other fMRI data with different brain conditions, disease phenotypes.

In Chapter 3, I showed that the VAE model has learned to represent and to generate patterns of cortical activity and connectivity using latent variables. Furthermore, I observed that the latent representation and its trajectory represented the spatiotemporal characteristics of rsfMRI activity, and the latent variables reflected the principal gradients of the latent trajectory and drove activity changes in cortical networks. Latent representations were clustered by both individuals and brain states. Interestingly, I found that representational geometry captured as covariance or correlation between latent variables, rather than cortical connectivity, could be used as a more reliable feature to accurately identify subjects from a large group, even if only a short period of data was available per subject. Ultimately, my results presented in Chapter 3 suggested that representational learning of rsfMRI done by the VAE model can be an alternative to conventional cortical mapping analysis.

In Chapter 4, by applying the pretrained VAE model to unseen fMRI data while watching naturalistic movies, I further validated that my VAE model was highly generalizable to fMRI data under different brain conditions and different recording parameters. One interesting finding under movie-watching fMRI data was that negative task-rest interaction observed in the cortical space was largely diminished in the latent space. Task-evoked latent representations and its trajectory were utilized to understand the dynamics of brain networks throughout the movie stimulus. I found principal bases defining the latent trajectory evoked by the task were predicted by many aspects of video: low-, middle-, high-level video features and by exogenous eye movement. Principal bases had unique interaction patterns between brain networks spanning from low-level sensory to high-level cognitive. Finally, inter-subject variability of brain activity was explained by the endogenous eye movement of individuals. Overall, I expect this application can be a good example of how one can employ VAE to excel our understanding of the brain system under different conditions such as tasks or disease.

The VAE model proposed in the thesis has a big potential as a beneficial analytical tool generalizable to various types of brain signal, given the applications and results I presented. However, there are several factors that can be further improved. First, the VAE model did not incorporate temporal information of fMRI data. Instead, I analyzed the temporal dynamics of fMRI data using the trajectory of representations in the latent space. While desirable, designing a model to learn the spatial and temporal information of fMRI data simultaneously requires a clever model design, abundant computational resources, and enough data. One possible solution can be marrying the VAE model and the RNN-type architecture. As storing temporal information with enough non-linearity may cost lots of computational burden, I believe my geometric reformatting trick (Figure 2.1) can be a remedy addressing this computational issue. Another factor that should be improved is the inclusion of subcortical fMRI data in the analysis. The cortico-subcortical communication plays an important role in many cognitive functions such as memory consolidation and attention (Censor et al., 2014; Heller et al., 2016; Heyder et al., 2004). One possible way is transforming the subcortical image into the 2D image, as done in the cortical activity. Then, three images will be fed to the VAE model, and merged features of images after passing through few convolutional layers. However, different from two hemispheres, there is no solid rationale how and where the images from two hemispheres will be functionally matched to the images of subcortical activity. Thus, more sophisticated input design that can include subcortical regions without adding computational burden to the model will be needed.

5.2 Future Applications

The model, analytical methods, and scientific findings that I have shown in the thesis can propose some scientifically significant opportunities for future works. One immediate application is the geometry of fMRI data under different task conditions. As the representations of rsfMRI were distributed spanning the whole latent space (Figure 3.5 and Figure 4.9), we can consider the rsfMRI as a null distribution in the latent space. Therefore, engaging in any specific task will change the geometry of representations in the latent space, occupying the subset of latent spaces as shown with movie-watching fMRI data (Figure 4.9). Thus, I speculate engaging in simpler tasks will occupy a smaller subspace than more complex tasks. Perhaps, the size of the subspace occupied by engaging tasks can be a quantitative surrogate reflecting the objective difficulty of tasks. Except for several experiments such as the N-back working memory task, it is practically hard to compare the difficulty of one experiment to the others. If my hypothesis turned out to be true, the size of the subspace can be a proxy measurement of task difficulty, and the direction of subspace can be another useful measurement in which cognitive or sensory aspects of the brain were required by the task.

Another application can be a hyper-alignment of an individual's functional map. A seminar paper done by (Haxby et al., 2011) has shown the inter-subject functional displacement can be aligned using relatively simple geometric transformation. As the VAE is forced to learn the common representations among the population during resting state, the learned latent variables can be utilized as a non-linear, unsupervised geometric transformation matrix between subjects' brain maps. This idea sounds particularly interesting since this geometric transformation can be generalizable and easily extended. For example, if one wants to project a functional map of an infant's brain to the standard map of healthy adults (e.g., MNI space), she/he can build another VAE model trained to reconstruct activity patterns of the infant's brain. Possibly, instead of training from scratch, one can finetune the original VAE model. Then, instead of wrapping one brain into another brain in the convolved cortical space, we can simply estimate the linear transformation between latent spaces of two VAE models. Once the linear transformation matrix is reasonably established, functional converting of the infant's brain to MNI space will be simple. This idea can be further extended to multi-modal source imaging such as EEG-fMRI source imaging. We can build a VAE model that can reconstruct the EEG signal. Then, the linear wrapping between two latent spaces (one from fMRI and another from EEG) will be utilized to project EEG data as one fMRI pattern and vice versa. Overall, given the insights and limitations of the VAE model, I hope this VAE model can be a good reference model that inspires other neuroscientists and neuroscience engineers.

REFERENCES

Ahmed, M.R., Zhang, Y., Liu, Y., and Liao, H. (2020). Single Volume Image Generator and Deep Learning-based ASD Classification. IEEE Journal of Biomedical and Health Informatics.

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., and Calhoun, V.D. (2014). Tracking whole-brain connectivity dynamics in the resting state. Cerebral cortex 24, 663-676.

Amico, E., and Goñi, J. (2018). The quest for identifiability in human functional connectomes. Scientific reports 8, 1-14.

Arthurs, O.J., and Boniface, S. (2002). How well do we understand the neural origins of the fMRI BOLD signal? Trends Neurosci 25, 27-31.

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron 85, 390-401.

Beckmann, C.F., and Smith, S.M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE transactions on medical imaging 23, 137-152.

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M.A., Logothetis, N.K., and Panzeri, S. (2008). Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. Journal of Neuroscience 28, 5696-5709.

Bengs, M., Gessert, N., and Schlaefer, A. (2020). 4d spatio-temporal deep learning with 4d fmri data for autism spectrum disorder classification. arXiv preprint arXiv:200410165.

Benson, N.C., Jamison, K.W., Arcaro, M.J., Vu, A.T., Glasser, M.F., Coalson, T.S., Van Essen, D.C., Yacoub, E., Ugurbil, K., and Winawer, J. (2018). The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. Journal of vision 18, 23-23.

Betzel, R.F., Byrge, L., Esfahlani, F.Z., and Kennedy, D.P. (2020). Temporal fluctuations in the brain's modular architecture during movie-watching. NeuroImage, 116687.

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S.G., de Zwart, J.A., and Duyn, J.H. (2009a). Modulation of spontaneous fMRI activity in human visual cortex by behavioral state. Neuroimage 45, 160-168.

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S.G., de Zwart, J.A., Shmueli, K., and Duyn, J.H. (2009b). Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. Magnetic resonance imaging 27, 1019-1029.

Birn, R.M., Smith, M.A., Jones, T.B., and Bandettini, P.A. (2008). The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. Neuroimage 40, 644-654.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., and Hyde, J.S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnetic resonance in medicine 34, 537-541.

Bolton, T.A., Freitas, L.G., Jochaut, D., Giraud, A.-L., and Van De Ville, D. (2020). Neural responses in autism during movie watching: Inter-individual response variability co-varies with symptomatology. NeuroImage, 116571.

Brass, M., Derrfuss, J., Forstmann, B., and von Cramon, D.Y. (2005). The role of the inferior frontal junction area in cognitive control. Trends in cognitive sciences 9, 314-316.

Brown, J.A., Lee, A.J., Pasquini, L., and Seeley, W.W. (2020). Intrinsic brain activity gradients dynamically coordinate functional connectivity states. bioRxiv.

Bubic, A., Von Cramon, D.Y., and Schubotz, R.I. (2010). Prediction, cognition and the brain. Frontiers in human neuroscience 4, 25.

Buckner, R.L., Andrews-Hanna, J.R., and Schacter, D.L. (2008). The brain's default network: anatomy, function, and relevance to disease.

Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in $\beta $-VAE. arXiv preprint arXiv:180403599.

Burgess, C.P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A. (2019). Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:190111390.

Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., and Miller, E.K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron 76, 838-846.

Buxton, R.B. (2009). Introduction to functional magnetic resonance imaging: principles and techniques (Cambridge university press).

Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nature reviews neuroscience 13, 407.

Byrge, L., and Kennedy, D.P. (2019). High-accuracy individual identification using a “thin slice” of the functional connectome. Network Neuroscience 3, 363-383.

Calhoun, V.D., Adali, T., Pearlson, G., and Pekar, J.J. (2001). Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Human brain mapping 13, 43-53.

Calhoun, V.D., Miller, R., Pearlson, G., and Adalı, T. (2014). The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. Neuron 84, 262-274.

Censor, N., Dayan, E., and Cohen, L.G. (2014). Cortico-subcortical neuronal circuitry associated with reconsolidation of human procedural memories. Cortex 58, 281-288.

Chang, C., and Glover, G.H. (2010). Time–frequency dynamics of resting-state brain connectivity measured with fMRI. Neuroimage 50, 81-98.

Chang, C., Leopold, D.A., Schölvinck, M.L., Mandelkow, H., Picchioni, D., Liu, X., Frank, Q.Y., Turchi, J.N., and Duyn, J.H. (2016). Tracking brain arousal fluctuations with fMRI. Proceedings of the National Academy of Sciences 113, 4518-4523.

Chen, G., Chen, G., Xie, C., and Li, S.-J. (2011). Negative functional connectivity and its dependence on the shortest path length of positive network in the resting-state human brain. Brain connectivity 1, 195-206.

Chen, J., Li, X., Calhoun, V.D., Turner, J.A., van Erp, T.G., Wang, L., Andreassen, O.A., Agartz, I., Westlye, L.T., and Jonsson, E. (2020). Sparse Deep Neural Networks on Imaging Genetics for Schizophrenia Case-Control Classification. medRxiv.

Chen, S., and Hu, X. (2018). Individual identification using the functional brain fingerprint detected by the recurrent neural network. Brain connectivity 8, 197-204.

Chou, Y.-h., Sundman, M., Whitson, H.E., Gaur, P., Chu, M.-L., Weingarten, C.P., Madden, D.J., Wang, L., Kirste, I., and Joliot, M. (2017). Maintenance and representation of mind wandering during Resting-State fMRI. Scientific reports 7, 40722.

Churchland, M.M., Byron, M.Y., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., and Scott, B.B. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nature neuroscience 13, 369-378.

Cole, M.W., and Schneider, W. (2007). The cognitive control network: integrated cortical regions with dissociable functions. Neuroimage 37, 343-360.

Cui, Y., Zhao, S., Chen, Y., Han, J., Guo, L., Xie, L., and Liu, T. (2019). Modeling brain diverse and complex hemodynamic response patterns via deep recurrent autoencoder. IEEE Transactions on Cognitive and Developmental Systems.

Curtis, B.J., Williams, P.G., Jones, C.R., and Anderson, J.S. (2016). Sleep duration and resting fMRI functional connectivity: examination of short sleepers with and without perceived daytime dysfunction. Brain and behavior 6, e00576.

D’Souza, N.S., Nebel, M.B., Wymbs, N., Mostofsky, S., and Venkataraman, A. (2019). Integrating Neural Networks and Dictionary Learning for Multidimensional Clinical Characterizations from Functional Connectomics Data. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 709-717.

Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., and Beckmann, C.F. (2006). Consistent resting-state networks across healthy subjects. Proceedings of the National Academy of Sciences of the United States of America 103, 13848-13853.

Decety, J., and Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. The neuroscientist 13, 580-593.

Demirtaş, M., Ponce-Alvarez, A., Gilson, M., Hagmann, P., Mantini, D., Betti, V., Romani, G.L., Friston, K., Corbetta, M., and Deco, G. (2019). Distinct modes of functional connectivity induced by movie-watching. NeuroImage 184, 335-348.

Dinges, D.F., Mallis, M.M., Maislin, G., and Powell, J.W. (1998). Evaluation of techniques for ocular measurement as an index of fatigue and as the basis for alertness management. (United States. National Highway Traffic Safety Administration).

Dixon, M.L., De La Vega, A., Mills, C., Andrews-Hanna, J., Spreng, R.N., Cole, M.W., and Christoff, K. (2018). Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention networks. Proceedings of the National Academy of Sciences 115, E1598-E1607.

Dosenbach, N.U., Fair, D.A., Miezin, F.M., Cohen, A.L., Wenger, K.K., Dosenbach, R.A., Fox, M.D., Snyder, A.Z., Vincent, J.L., and Raichle, M.E. (2007). Distinct brain networks for adaptive and stable task control in humans. Proceedings of the National Academy of Sciences 104, 11073-11078.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., *et al.* (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine 23, 28-38.

Dubreuil-Vall, L., Ruffini, G., and Camprodon, J.A. (2020). Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG. Frontiers in neuroscience 14.

Dvornek, N.C., Ventola, P., and Duncan, J.S. (2018a). Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (IEEE), pp. 725-728.

Dvornek, N.C., Yang, D., Ventola, P., and Duncan, J.S. (2018b). Learning generalizable recurrent neural networks from small task-fmri datasets. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 329-337.

Eavani, H., Satterthwaite, T.D., Gur, R.E., Gur, R.C., and Davatzikos, C. (2013). Unsupervised learning of functional network dynamics in resting state fMRI. In International conference on information processing in medical imaging (Springer), pp. 426-437.

Ebrahimi-Ghahnavieh, A., Luo, S., and Chiong, R. (2019). Transfer Learning for Alzheimer's Disease Detection on MRI Images. In 2019 IEEE International Conference on Industry 40, Artificial Intelligence, and Communications Technology (IAICT) (IEEE), pp. 133-138.

El-Gazzar, A., Quaak, M., Cerliani, L., Bloem, P., van Wingen, G., and Thomas, R.M. (2019). A Hybrid 3DCNN and 3DC-LSTM based model for 4D Spatio-temporal fMRI data: An ABIDE Autism Classification study. In OR 20 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging (Springer), pp. 95-102.

Elliott, M.L., Knodt, A.R., Cooke, M., Kim, M.J., Melzer, T.R., Keenan, R., Ireland, D., Ramrakha, S., Poulton, R., and Caspi, A. (2019). General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. NeuroImage 189, 516-532.

Fan, L., Su, J., Qin, J., Hu, D., and Shen, H. (2020). A Deep Network Model on Dynamic Functional Connectivity With Applications to Gender Classification and Intelligence Prediction. Frontiers in neuroscience 14, 881.

Fang, F., and He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. Nature neuroscience 8, 1380-1385.

Feng, W., Halm-Lutterodt, N.V., Tang, H., Mecum, A., Mesregah, M.K., Ma, Y., Li, H., Zhang, F., Wu, Z., and Yao, E. (2020). Automated MRI-Based Deep Learning Model for Detection of Alzheimer’s Disease Process. International Journal of Neural Systems 30, 2050032.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., and Constable, R.T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature neuroscience 18, 1664.

Fischl, B. (2012). FreeSurfer. Neuroimage 62, 774-781.

Fox, M.D., Buckner, R.L., Liu, H., Chakravarty, M.M., Lozano, A.M., and Pascual-Leone, A. (2014). Resting-state networks link invasive and noninvasive brain stimulation across diverse psychiatric and neurological diseases. Proceedings of the National Academy of Sciences 111, E4367-E4375.

Fox, M.D., Corbetta, M., Snyder, A.Z., Vincent, J.L., and Raichle, M.E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. Proceedings of the National Academy of Sciences 103, 10046-10051.

Fox, M.D., and Raichle, M.E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature reviews neuroscience 8, 700-711.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., and Raichle, M.E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proceedings of the National Academy of Sciences 102, 9673-9678.

Frackowiak, R.S. (2004). Human brain function (Elsevier).

Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. Neuron 88, 220-235.

Friston, K.J., Josephs, O., Rees, G., and Turner, R. (1998). Nonlinear event-related responses in fMRI. Magnetic resonance in medicine 39, 41-52.

Friston, K.J., Mechelli, A., Turner, R., and Price, C.J. (2000). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage 12, 466-477.

Fukunaga, M., Horovitz, S.G., van Gelderen, P., de Zwart, J.A., Jansma, J.M., Ikonomidou, V.N., Chu, R., Deckers, R.H.R., Leopold, D.A., and Duyn, J.H. (2006). Large-amplitude, spatially correlated fluctuations in BOLD fMRI signals during extended rest and early sleep stages. Magnetic Resonance Imaging 24, 979-992.

Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Adeli, E., and Pohl, K.M. (2020). Spatio-Temporal Graph Convolution for Resting-State fMRI Analysis. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 528-538.

Gao, M.-S., Tsai, F.-S., and Lee, C.-C. (2020a). Learning a Phenotypic-Attribute Attentional Brain Connectivity Embedding for ADHD Classification using rs-fMRI. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (IEEE), pp. 5472-5475.

Gao, Z., Wang, X., Yang, Y., Li, Y., Ma, K., and Chen, G. (2020b). A Channel-fused Dense Convolutional Network for EEG-based Emotion Recognition. IEEE Transactions on Cognitive and Developmental Systems.

Garrity, A.G., Pearlson, G.D., McKiernan, K., Lloyd, D., Kiehl, K.A., and Calhoun, V.D. (2007). Aberrant “default mode” functional connectivity in schizophrenia. American journal of psychiatry 164, 450-457.

Gaxiola-Valdez, I., and Goodyear, B.G. (2012). Origins of intersubject variability of blood oxygenation level dependent and arterial spin labeling fMRI: implications for quantification of brain activity. Magnetic Resonance Imaging 30, 1394-1400.

Gilbert, C.D., and Li, W. (2013). Top-down influences on visual processing. Nature Reviews Neuroscience 14, 350.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., and 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., and Polimeni, J.R. (2013). The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105-124.

Glover, G.H. (2011). Overview of functional magnetic resonance imaging. Neurosurgery Clinics 22, 133-139.

Goceri, E. (2019). Diagnosis of Alzheimer's disease with Sobolev gradient-based optimization and 3D convolutional neural network. International journal for numerical methods in biomedical engineering 35, e3225.

Goense, J.B., and Logothetis, N.K. (2008). Neurophysiology of the BOLD fMRI signal in awake monkeys. Current Biology 18, 631-640.

Goldman, R.I., Stern, J.M., Engel Jr, J., and Cohen, M.S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. Neuroreport 13, 2487.

Goncalves, S., De Munck, J., Pouwels, P., Schoonhoven, R., Kuijer, J., Maurits, N., Hoogduin, J., Van Someren, E., Heethaar, R., and Da Silva, F.L. (2006). Correlating the alpha rhythm to BOLD using simultaneous EEG/fMRI: inter-subject variability. Neuroimage 30, 203-213.

Gonzalez-Castillo, J., Hoy, C.W., Handwerker, D.A., Robinson, M.E., Buchanan, L.C., Saad, Z.S., and Bandettini, P.A. (2015). Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. Proceedings of the National Academy of Sciences 112, 8762-8767.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems 27, 2672-2680.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM 63, 139-144.

Gopinath, K., Krishnamurthy, V., Cabanban, R., and Crosson, B.A. (2015). Hubs of anticorrelation in high-resolution resting-state functional connectivity network architecture. Brain Connectivity 5, 267-275.

Greicius, M.D., Krasnow, B., Reiss, A.L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences 100, 253-258.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., and Mackay, C.E. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage 95, 232-247.

Güçlü, U., and van Gerven, M.A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. Frontiers in computational neuroscience 11, 7.

Guo, X., Dominick, K.C., Minai, A.A., Li, H., Erickson, C.A., and Lu, L.J. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. Frontiers in neuroscience 11, 460.

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., and Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. NeuroImage 198, 125-136.

Han, K., Wen, H., Zhang, Y., Fu, D., Culurciello, E., and Liu, Z. (2018). Deep predictive coding network with local recurrent processing for object recognition. In Advances in neural information processing systems, pp. 9201-9213.

Hanslmayr, S., Volberg, G., Wimber, M., Raabe, M., Greenlee, M.W., and Bäuml, K.-H.T. (2011). The relationship between brain oscillations and BOLD signal during memory formation: a combined EEG–fMRI study. Journal of Neuroscience 31, 15674-15680.

Hardy, N.F., and Buonomano, D.V. (2018). Encoding time in feedforward trajectories of a recurrent neural network model. Neural computation 30, 378-396.

Hasson, U., Malach, R., and Heeger, D.J. (2010). Reliability of cortical activity during natural stimulation. Trends in cognitive sciences 14, 40-48.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. science 303, 1634-1640.

Haufe, S., DeGuzman, P., Henin, S., Arcaro, M., Honey, C.J., Hasson, U., and Parra, L.C. (2018). Elucidating relations between fMRI, ECoG, and EEG through a common natural stimulus. NeuroImage 179, 79-91.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., and Ramadge, P.J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404-416.

He, B.J. (2013). Spontaneous and task-evoked brain activity negatively interact. Journal of Neuroscience 33, 4672-4682.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., and Yeo, B.T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage 206, 116276.

Hebart, M.N., and Hesselmann, G. (2012). What visual information is processed in the human dorsal stream? Journal of Neuroscience 32, 8107-8109.

Heller, A.S., Cohen, A.O., Dreyfuss, M.F., and Casey, B. (2016). Changes in cortico-subcortical and subcortico-subcortical connectivity impact cognitive control to emotional cues across development. Social Cognitive and Affective Neuroscience 11, 1910-1918.

Heyder, K., Suchan, B., and Daum, I. (2004). Cortico-subcortical contributions to executive control. Acta psychologica 115, 271-289.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. Iclr 2, 6.

Hirabayashi, T., Takeuchi, D., Tamura, K., and Miyashita, Y. (2013). Functional microcircuit recruited during retrieval of object association memory in monkey perirhinal cortex. Neuron 77, 192-203.

Ho, T.K.K., Gwak, J., Park, C.M., and Song, J.-I. (2019). Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches. IEEE Access 7, 24392-24403.

Honey, C.J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., and Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. Proceedings of the National Academy of Sciences 106, 2035-2040.

Horovitz, S.G., Braun, A.R., Carr, W.S., Picchioni, D., Balkin, T.J., Fukunaga, M., and Duyn, J.H. (2009). Decoupling of the brain's default mode network during deep sleep. Proceedings of the National Academy of Sciences of the United States of America 106, 11376-11381.

Horovitz, S.G., Fukunaga, M., De Zwart, J.A., Van Gelderen, P., Fulton, S.C., Balkin, T.J., and Duyn, J.H. (2008). Low frequency BOLD fluctuations during resting wakefulness and light sleep: A simultaneous EEG-fMRI study. Human Brain Mapping 29, 671-682.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., and Liu, T. (2017). Modeling task fMRI data via deep convolutional autoencoder. IEEE transactions on medical imaging 37, 1551-1561.

Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., and Gonzalez-Castillo, J. (2013a). Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage 80, 360-378.

Hutchison, R.M., Womelsdorf, T., Gati, J.S., Everling, S., and Menon, R.S. (2013b). Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. Human Brain Mapping 34, 2154-2177.

Jang, H., Plis, S.M., Calhoun, V.D., and Lee, J.-H. (2017). Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks. NeuroImage 145, 314-328.

Jang, S., Moon, S.-E., and Lee, J.-S. (2018). EEG-based video identification using graph signal modeling and graph convolutional neural network. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp. 3066-3070.

Jarrett, C. (2009). The restless brain. The Psychologist.

Johns, M.W., Tucker, A., Chapman, R., Crowley, K., and Michael, N. (2007). Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. Somnologie-Schlafforschung und Schlafmedizin 11, 234-242.

Kam, T.-E., Zhang, H., Jiao, Z., and Shen, D. (2019). Deep learning of static and dynamic brain functional networks for early mci detection. IEEE transactions on medical imaging 39, 478-487.

Kashyap, A., and Keilholz, S. (2020). Brain network constraints and recurrent neural networks reproduce unique trajectories and state transitions seen over the span of minutes in resting-state fMRI. Network Neuroscience 4, 448-466.

Kaufmann, C., Wehrle, R., Wetter, T., Holsboer, F., Auer, D., Pollmächer, T., and Czisch, M. (2006). Brain activation and hypothalamic functional connectivity during human non-rapid eye movement sleep: an EEG/fMRI study. Brain 129, 655-667.

Kauppi, J.P., Jääskeläinen, I.P., Sams, M., and Tohka, J. (2010). Inter-subject correlation of brain hemodynamic responses during watching a movie: Localization in space and frequency. Frontiers in Neuroinformatics 4.

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., and Hamarneh, G. (2017). BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage 146, 1038-1049.

Kawato, M. (1999). Internal models for motor control and trajectory planning. Current opinion in neurobiology 9, 718-727.

Kell, A.J., Yamins, D.L., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, 630-644. e616.

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

Khemakhem, I., Kingma, D.P., and Hyvärinen, A. (2019). Variational autoencoders and nonlinear ica: A unifying framework. arXiv preprint arXiv:190704809.

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M.R. (2019a). Detecting abnormalities in resting-state dynamics: An unsupervised learning approach. In International Workshop on Machine Learning in Medical Imaging (Springer), pp. 301-309.

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M.R. (2019b). Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. Neuroimage 199, 651-662.

Khosla, M., Jamison, K., Ngo, G.H., Kuceyeski, A., and Sabuncu, M.R. (2019c). Machine learning in resting-state fMRI analysis. Magnetic resonance imaging.

Kim, H.-C., Bandettini, P.A., and Lee, J.-H. (2019). Deep neural network predicts emotional responses of the human brain from functional magnetic resonance imaging. NeuroImage 186, 607-627.

Kim, H.-c., and Lee, J.-h. (2016). Evaluation of weight sparsity control during autoencoder training of resting-state fMRI using non-zero ratio and Hoyer's sparseness. In 2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI) (IEEE), pp. 1-4.

Kim, J.-H., Zhang, Y., Han, K., Choi, M., and Liu, Z. (2020). Representation Learning of Resting State fMRI with Variational Autoencoder. bioRxiv.

Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980.

Kingma, D.P., and Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:13126114.

Kiviniemi, V., Jauhiainen, J., Tervonen, O., Pääkkö, E., Oikarinen, J., Vainionpää, V., Rantala, H., and Biswal, B. (2000). Slow vasomotor fluctuation in fMRI of anesthetized child brain. Magnetic Resonance in Medicine 44, 373-378.

Klink, P.C., Dagnino, B., Gariel-Mathis, M.A., and Roelfsema, P.R. (2017). Distinct Feedforward and Feedback Effects of Microstimulation in Visual Cortex Reveal Neural Mechanisms of Texture Segregation. Neuron 95, 209-220 e203.

Koppe, G., Toutounji, H., Kirsch, P., Lis, S., and Durstewitz, D. (2019). Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. PLoS computational biology 15, e1007263.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends in cognitive sciences 17, 401-412.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM 60, 84-90.

Lachaux, J.P., Fonlupt, P., Kahane, P., Minotti, L., Hoffmann, D., Bertrand, O., and Baciu, M. (2007). Relationship between task-related gamma oscillations and BOLD signal: New insights from combined fMRI and intracranial EEG. Human brain mapping 28, 1368-1375.

Larson-Prior, L.J., Zempel, J.M., Nolan, T.S., Prior, F.W., Snyder, A., and Raichle, M.E. (2009). Cortical network functional connectivity in the descent to sleep. Proceedings of the National Academy of Sciences of the United States of America 106, 4489-4494.

Lartillot, O., Toiviainen, P., and Eerola, T. (2008). A matlab toolbox for music information retrieval. In Data analysis, machine learning and applications (Springer), pp. 261-268.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. nature 521, 436-444.

León, J., Escobar, J.J., Ortiz, A., Ortega, J., González, J., Martín-Smith, P., Gan, J.Q., and Damas, M. (2020). Deep learning for EEG-based Motor Imagery classification: Accuracy-cost trade-off. Plos one 15, e0234178.

Leonardi, N., and Van De Ville, D. (2015). On spurious and real fluctuations of dynamic functional connectivity during rest. Neuroimage 104, 430-436.

Leopold, D.A., and Maier, A. (2012). Ongoing physiological processes in the cerebral cortex. Neuroimage 62, 2190-2200.

Li, H., and Fan, Y. (2018). Brain decoding from functional MRI using long short-term memory recurrent neural networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 320-328.

Liang, Z., King, J., and Zhang, N. (2012). Anticorrelated resting-state functional connectivity in awake rat brain. Neuroimage 59, 1190-1199.

Liégeois, R., Li, J., Kong, R., Orban, C., Van De Ville, D., Ge, T., Sabuncu, M.R., and Yeo, B.T. (2019). Resting brain dynamics at different timescales capture distinct aspects of human behavior. Nature communications 10, 1-9.

Lindquist, M.A., Loh, J.M., Atlas, L.Y., and Wager, T.D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. Neuroimage 45, S187-S198.

Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., Xu, M., and Initiative, A.s.D.N. (2020a). A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. NeuroImage 208, 116459.

Liu, S., Zhao, L., Wang, X., Xin, Q., Zhao, J., Guttery, D.S., and Zhang, Y.-D. (2020b). Deep Spatio-Temporal Representation and Ensemble Classification for Attention deficit/Hyperactivity disorder. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Liu, X., Chang, C., and Duyn, J.H. (2013). Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns. Frontiers in systems neuroscience 7, 101.

Liu, X., and Duyn, J.H. (2013). Time-varying functional network information extracted from brief instances of spontaneous brain activity. Proceedings of the National Academy of Sciences 110, 4392-4397.

Liu, Z., Rios, C., Zhang, N., Yang, L., Chen, W., and He, B. (2010). Linear and nonlinear relationships between visual stimuli, EEG and BOLD fMRI signals. Neuroimage 50, 1054-1066.

Logothetis, N.K. (2002). The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 357, 1003-1037.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150-157.

Lostar, M., and Rekik, I. (2020). Deep Hypergraph U-Net for Brain Graph Embedding and Classification. arXiv preprint arXiv:200813118.

Lu, H., Jaime, S., and Yang, Y. (2019). Origins of the resting-state functional MRI signal: potential limitations of the "neurocentric" model. Frontiers in neuroscience 13.

Ludwig, C.J., Davies, J.R., and Eckstein, M.P. (2014). Foveal analysis and peripheral selection during active visual sampling. Proceedings of the National Academy of Sciences 111, E291-E299.

Lund, T.E., Nørgaard, M.D., Rostrup, E., Rowe, J.B., and Paulson, O.B. (2005). Motion or activity: their role in intra-and inter-subject variation in fMRI. Neuroimage 26, 960-964.

Lynch, L.K., Lu, K.H., Wen, H., Zhang, Y., Saykin, A.J., and Liu, Z. (2018). Task-evoked functional connectivity does not explain functional connectivity differences between rest and task conditions. Human brain mapping 39, 4939-4948.

Makkie, M., Huang, H., Zhao, Y., Vasilakos, A.V., and Liu, T. (2019). Fast and scalable distributed deep convolutional autoencoder for fMRI big data analytics. Neurocomputing 325, 20-30.

Mandelkow, H., de Zwart, J.A., and Duyn, J.H. (2016). Linear discriminant analysis achieves high classification accuracy for the BOLD fMRI response to naturalistic movie stimuli. Frontiers in human neuroscience 10, 128.

Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L., and Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. Proceedings of the National Academy of Sciences 104, 13170-13175.

Martin, C., Martindale, J., Berwick, J., and Mayhew, J. (2006). Investigating neural–hemodynamic coupling and the hemodynamic response function in the awake rat. Neuroimage 32, 33-48.

Martuzzi, R., Ramani, R., Qiu, M., Rajeevan, N., and Constable, R.T. (2010). Functional connectivity and alterations in baseline brain state in humans. NeuroImage 49, 823-834.

Matsubara, T., Tashiro, T., and Uehara, K. (2019). Deep neural generative model of functional MRI images for psychiatric disorder diagnosis. IEEE Transactions on Biomedical Engineering 66, 2768-2779.

Matsuhashi, M., Ikeda, A., Ohara, S., Matsumoto, R., Yamamoto, J., Takayama, M., Satow, T., Begum, T., Usui, K., and Nagamine, T. (2004). Multisensory convergence at human temporo-parietal junction–epicortical recording of evoked responses. Clinical Neurophysiology 115, 1145-1160.

McIntire, L.K., McKinley, R.A., Goodyear, C., and McIntire, J.P. (2014). Detection of vigilance performance using eye blinks. Applied ergonomics 45, 354-362.

Mechler, F., Victor, J.D., Purpura, K.P., and Shapley, R. (1998). Robust temporal coding of contrast by V1 neurons for transient but not for steady-state stimuli. Journal of Neuroscience 18, 6583-6598.

Mejia, A.F., Nebel, M.B., Barber, A.D., Choe, A.S., Pekar, J.J., Caffo, B.S., and Lindquist, M.A. (2018). Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. NeuroImage 172, 478-491.

Mejias, J.F., Murray, J.D., Kennedy, H., and Wang, X.J. (2016). Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. Sci Adv 2, e1601335.

Meszlényi, R.J., Buza, K., and Vidnyánszky, Z. (2017). Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture. Frontiers in neuroinformatics 11, 61.

Metea, M.R., and Newman, E.A. (2006). Glial cells dilate and constrict blood vessels: a mechanism of neurovascular coupling. Journal of Neuroscience 26, 2862-2870.

Michalareas, G., Vezoli, J., Van Pelt, S., Schoffelen, J.-M., Kennedy, H., and Fries, P. (2016). Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. Neuron 89, 384-397.

Milner, A., and Goodale, M. (1995). Oxford psychology series, No. 27. The visual brain in action. (Oxford University Press New York).

Monier, C., Chavane, F., Baudot, P., Graham, L.J., and Frégnac, Y. (2003). Orientation and direction selectivity of synaptic inputs in visual cortical neurons: a diversity of combinations produces spike tuning. Neuron 37, 663-680.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., and Initiative, A.s.D.N. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398-412.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science 309, 951-954.

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., and Bandettini, P.A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? Neuroimage 44, 893-905.

Murta, T., Hu, L., Tierney, T.M., Chaudhary, U.J., Walker, M.C., Carmichael, D.W., Figueiredo, P., and Lemieux, L. (2016). A study of the electro-haemodynamic coupling using simultaneously acquired intracranial EEG and fMRI data in humans. Neuroimage 142, 371-380.

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

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage 56, 400-410.

Niessing, J., Ebisch, B., Schmidt, K.E., Niessing, M., Singer, W., and Galuske, R.A. (2005). Hemodynamic signals correlate tightly with synchronized gamma oscillations. science 309, 948-951.

Oh, K., Chung, Y.-C., Kim, K.W., Kim, W.-S., and Oh, I.-S. (2019). Classification and visualization of Alzheimer’s disease using volumetric convolutional neural network and transfer learning. Scientific Reports 9, 1-16.

Ong, J.L., Asplund, C.L., Chia, T.T., and Chee, M.W. (2013). Now you hear me, now you don't: eyelid closures as an indicator of auditory task disengagement. Sleep 36, 1867-1874.

Oota, S.R., Rowtula, V., Gupta, M., and Bapi, R.S. (2019). StepEncog: A Convolutional LSTM Autoencoder for Near-Perfect fMRI Encoding. In 2019 International Joint Conference on Neural Networks (IJCNN) (IEEE), pp. 1-8.

Pelphrey, K.A., Singerman, J.D., Allison, T., and McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: the influence of context. Neuropsychologia 41, 156-170.

Pickering, M.J., and Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. Trends in cognitive sciences 18, 451-456.

Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., and Calhoun, V.D. (2014). Deep learning for neuroimaging: a validation study. Frontiers in neuroscience 8, 229.

Ponce-Alvarez, A., Thiele, A., Albright, T.D., Stoner, G.R., and Deco, G. (2013). Stimulus-dependent variability and noise correlations in cortical MT neurons. Proceedings of the National Academy of Sciences 110, 13162-13167.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84, 320-341.

Qiao, K., Chen, J., Wang, L., Zhang, C., Zeng, L., Tong, L., and Yan, B. (2019). Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices. Frontiers in neuroscience 13.

Qureshi, M.N.I., Ryu, S., Song, J., Lee, K.H., and Lee, B. (2019). Evaluation of functional decline in alzheimer’s dementia using 3d deep learning and group ica for rs-fmri measurements. Frontiers in aging neuroscience 11, 8.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., and Shulman, G.L. (2001). A default mode of brain function. Proceedings of the National Academy of Sciences 98, 676-682.

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. Nature Reviews Neuroscience 10, 724-735.

Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature neuroscience 12, 718.

Ravi, D., Blumberg, S.B., Mengoudi, K., Xu, M., Alexander, D.C., and Oxtoby, N.P. (2019). Degenerative Adversarial NeuroImage Nets for 4D Simulations: Application in Longitudinal MRI. arXiv preprint arXiv:191201526.

Riaz, A., Asad, M., Alonso, E., and Slabaugh, G. (2020). DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. Journal of Neuroscience Methods 335, 108506.

Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., and Ganguli, S. (2019). A deep learning framework for neuroscience. Nature neuroscience 22, 1761-1770.

Ridley, B., Wirsich, J., Bettus, G., Rodionov, R., Murta, T., Chaudhary, U., Carmichael, D., Thornton, R., Vulliemoz, S., and McEvoy, A. (2017). Simultaneous intracranial EEG-fMRI shows inter-modality correlation in time-resolved connectivity within normal areas but not within epileptic regions. Brain topography 30, 639-655.

Rogers, B.P., Katwal, S.B., Morgan, V.L., Asplund, C.L., and Gore, J.C. (2010). Functional MRI and multivariate autoregressive models. Magnetic resonance imaging 28, 1058-1065.

Saeed, F., Eslami, T., Mirjalili, V., Fong, A., and Laird, A. (2019). ASD-DiagNet: A hybrid learning approach for detection of Autism Spectrum Disorder using fMRI data. Frontiers in Neuroinformatics 13, 70.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., and Smith, S.M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449-468.

Sämann, P.G., Wehrle, R., Hoehn, D., Spoormaker, V.I., Peters, H., Tully, C., Holsboer, F., and Czisch, M. (2011). Development of the brain's default mode network from wakefulness to slow wave sleep. Cerebral Cortex 21, 2082-2093.

Scheeringa, R., Koopmans, P.J., van Mourik, T., Jensen, O., and Norris, D.G. (2016). The relationship between oscillatory EEG activity and the laminar-specific BOLD signal. Proc Natl Acad Sci U S A 113, 6761-6766.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks 61, 85-117.

Schölvinck, M.L., Maier, A., Frank, Q.Y., Duyn, J.H., and Leopold, D.A. (2010). Neural basis of global resting-state fMRI activity. Proceedings of the National Academy of Sciences 107, 10238-10243.

Seo, Y., Morante, M., Kopsinis, Y., and Theodoridis, S. (2019). Unsupervised Pre-training of the Brain Connectivity Dynamic Using Residual D-Net. In International Conference on Neural Information Processing (Springer), pp. 608-620.

Shah, L.M., Cramer, J.A., Ferguson, M.A., Birn, R.M., and Anderson, J.S. (2016). Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. Brain and behavior 6, e00456.

Sharif, H., and Khan, R.A. (2019). A novel machine learning based framework for detection of Autism Spectrum Disorder (ASD). arXiv preprint arXiv:190311323.

Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. Neuroimage 49, 3110-3121.

Sheth, S.A., Nemoto, M., Guiou, M., Walker, M., Pouratian, N., and Toga, A.W. (2004). Linear and nonlinear relationships between neuronal activity, oxygen metabolism, and hemodynamic responses. Neuron 42, 347-355.

Shi, J., Wen, H., Zhang, Y., Han, K., and Liu, Z. (2018). Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. Human brain mapping 39, 2269-2282.

Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., and Laird, A.R. (2009). Correspondence of the brain's functional architecture during activation and rest. Proceedings of the National Academy of Sciences 106, 13040-13045.

Smith, S.M., Miller, K.L., Moeller, S., Xu, J., Auerbach, E.J., Woolrich, M.W., Beckmann, C.F., Jenkinson, M., Andersson, J., and Glasser, M.F. (2012). Temporally-independent functional modes of spontaneous brain activity. Proceedings of the National Academy of Sciences 109, 3131-3136.

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. PLoS computational biology 1.

Stevens, A.A., Tappon, S.C., Garg, A., and Fair, D.A. (2012). Functional brain network modularity captures inter-and intra-individual variation in working memory capacity. PloS one 7, e30468.

Suk, H.-I., Lee, S.-W., Shen, D., and Initiative, A.s.D.N. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Structure and Function 220, 841-859.

Suk, H.-I., Wee, C.-Y., Lee, S.-W., and Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. NeuroImage 129, 292-307.

Sutskever, I., Vinyals, O., and Le, Q.V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pp. 3104-3112.

Thanh-Tung, H., and Tran, T. (2018). On catastrophic forgetting and mode collapse in generative adversarial networks. arXiv, arXiv: 1807.04015.

van den Heuvel, M.P., and Hulshoff Pol, H.E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. European Neuropsychopharmacology 20, 519-534.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., and Consortium, W.-M.H. (2013). The WU-Minn human connectome project: an overview. Neuroimage 80, 62-79.

Van Kerkoerle, T., Self, M.W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., Van Der Togt, C., and Roelfsema, P.R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proceedings of the National Academy of Sciences 111, 14332-14341.

Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., and Castellanos, F.X. (2017). Individual differences in functional connectivity during naturalistic viewing conditions. NeuroImage 157, 521-530.

Venkatesh, M., Jaja, J., and Pessoa, L. (2019). Brain dynamics and temporal trajectories during task and naturalistic processing. Neuroimage 186, 410-423.

Venkatesh, M., Jaja, J., and Pessoa, L. (2020). Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. NeuroImage 207, 116398.

Vincent, J.L., Patel, G.H., Fox, M.D., Snyder, A.Z., Baker, J.T., Van Essen, D.C., Zempel, J.M., Snyder, L.H., Corbetta, M., and Raichle, M.E. (2007). Intrinsic functional architecture in the anaesthetized monkey brain. Nature 447, 83-86.

Viola, P., & Jones, M. (2001). Robust real-time object detection. International journal of computer vision, 4(34-47), 4.

Vu, H., Kim, H.-C., Jung, M., and Lee, J.-H. (2020). fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations. NeuroImage 223, 117328.

Wan, X., Riera, J., Iwata, K., Takahashi, M., Wakabayashi, T., and Kawashima, R. (2006). The neural basis of the hemodynamic response nonlinearity in human primary visual cortex: Implications for neurovascular coupling mechanism. Neuroimage 32, 616-625.

Wang, C., Ong, J.L., Patanaik, A., Zhou, J., and Chee, M.W. (2016). Spontaneous eyelid closures link vigilance fluctuation with fMRI dynamic connectivity states. Proceedings of the National Academy of Sciences 113, 9653-9658.

Wang, L., Li, K., Chen, X., and Hu, X.P. (2019a). Application of convolutional recurrent neural network for individual recognition based on resting state fmri data. Frontiers in Neuroscience 13, 434.

Wang, M., Lian, C., Yao, D., Zhang, D., Liu, M., and Shen, D. (2019b). Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-recurrent network. IEEE Transactions on Biomedical Engineering.

Wang, X., Liang, X., Jiang, Z., Nguchu, B.A., Zhou, Y., Wang, Y., Wang, H., Li, Y., Zhu, Y., and Wu, F. (2020). Decoding and mapping task states of the human brain via deep learning. Human brain mapping 41, 1505-1519.

Wen, H., and Liu, Z. (2016). Broadband electrophysiological dynamics contribute to global resting-state fMRI signal. Journal of Neuroscience 36, 6030-6040.

Wen, H., Shi, J., Chen, W., and Liu, Z. (2018a). Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Scientific reports 8, 1-17.

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

Wen, X., Dong, L., Chen, J., Xiang, J., Yang, J., Li, H., Liu, X., Luo, C., and Yao, D. (2020). Detecting the Information of Functional Connectivity Networks in Normal Aging Using Deep Learning From a Big Data Perspective. Frontiers in neuroscience 13, 1435.

Winder, A.T., Echagarruga, C., Zhang, Q., and Drew, P.J. (2017). Weak correlations between hemodynamic signals and ongoing neural activity during the resting state. Nature neuroscience 20, 1761-1769.

Xia, T., Chartsias, A., Wang, C., and Tsaftaris, S.A. (2019). Learning to synthesise the ageing brain without longitudinal data. arXiv preprint arXiv:191202620.

Xie, H., Calhoun, V.D., Gonzalez-Castillo, J., Damaraju, E., Miller, R., Bandettini, P.A., and Mitra, S. (2018). Whole-brain connectivity dynamics reflect both task-specific and individual-specific modulation: A multitask study. Neuroimage 180, 495-504.

Xiong, J., Rao, S., Jerabek, P., Zamarripa, F., Woldorff, M., Lancaster, J., and Fox, P.T. (2000). Intersubject variability in cortical activations during a complex language task. Neuroimage 12, 326-339.

Xu, L., Liu, Y., Yu, J., Li, X., Yu, X., Cheng, H., and Li, J. (2020). Characterizing autism spectrum disorder by deep learning spontaneous brain activity from functional near-infrared spectroscopy. Journal of Neuroscience Methods 331, 108538.

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

Yan, W., Calhoun, V., Song, M., Cui, Y., Yan, H., Liu, S., Fan, L., Zuo, N., Yang, Z., and Xu, K. (2019). Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data. EBioMedicine 47, 543-552.

Yang, P., Zhou, F., Ni, D., Xu, Y., Chen, S., Wang, T., and Lei, B. (2019). Fused sparse network learning for longitudinal analysis of mild cognitive impairment. IEEE transactions on cybernetics.

Yao, H., Shi, L., Han, F., Gao, H., and Dan, Y. (2007). Rapid learning in cortical coding of visual scenes. Nature neuroscience 10, 772-778.

Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., and Polimeni, J.R. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of neurophysiology.

Yuan, H., Liu, T., Szarkowski, R., Rios, C., Ashe, J., and He, B. (2010). Negative covariation between task-related responses in alpha/beta-band activity and BOLD in human sensorimotor cortex: an EEG and fMRI study of motor imagery and movements. Neuroimage 49, 2596-2606.

Yuan, J., Blumen, H.M., Verghese, J., and Holtzer, R. (2015). Functional connectivity associated with gait velocity during walking and walking-while-talking in aging: A resting-state fMRI study. Human brain mapping 36, 1484-1493.

Zanto, T.P., Rubens, M.T., Bollinger, J., and Gazzaley, A. (2010). Top-down modulation of visual feature processing: the role of the inferior frontal junction. Neuroimage 53, 736-745.

Zeng, D., Huang, K., Xu, C., Shen, H., and Chen, Z. (2020). Hierarchy Graph Convolution Network and Tree Classification for Epileptic Detection on Electroencephalography Signals. IEEE Transactions on Cognitive and Developmental Systems.

Zeng, L.L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., and Hu, D. (2012). Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis. Brain 135, 1498-1507.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and Initiative, A.s.D.N. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856-867.

Zhang, X., Chou, J., and Wang, F. (2018). Integrative analysis of patient health records and neuroimages via memory-based graph convolutional network. In 2018 IEEE International Conference on Data Mining (ICDM) (IEEE), pp. 767-776.

Zhang, Y., Han, K., Worth, R., and Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. Nature communications 11, 1-13.

Zhao, F., Zhao, T., Zhou, L., Wu, Q., and Hu, X. (2008). BOLD study of stimulation-induced neural activity and resting-state connectivity in medetomidine-sedated rat. NeuroImage 39, 248-260.

Zhao, J., Huang, J., Zhi, D., Yan, W., Ma, X., Yang, X., Li, X., Ke, Q., Jiang, T., and Calhoun, V.D. (2020). Functional network connectivity (FNC)-based generative adversarial network (GAN) and its applications in classification of mental disorders. Journal of Neuroscience Methods, 108756.

Zhao, Q., Honnorat, N., Adeli, E., Pfefferbaum, A., Sullivan, E.V., and Pohl, K.M. (2019). Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis. In International Conference on Information Processing in Medical Imaging (Springer), pp. 867-879.

Zhao, Y., Li, X., Zhang, W., Zhao, S., Makkie, M., Zhang, M., Li, Q., and Liu, T. (2018). Modeling 4d fmri data via spatio-temporal convolutional neural networks (st-cnn). In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 181-189.

Zhou, J., Greicius, M.D., Gennatas, E.D., Growdon, M.E., Jang, J.Y., Rabinovici, G.D., Kramer, J.H., Weiner, M., Miller, B.L., and Seeley, W.W. (2010). Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. Brain 133, 1352-1367.

Zou, L., Zheng, J., Miao, C., Mckeown, M.J., and Wang, Z.J. (2017). 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. IEEE Access 5, 23626-23636.

PUBLICATIONS

Jung-Hoon Kim, Yizhen Zhang, Kuan Han, Minkyu Choi, and Zhongming Liu “Representation Learning of Resting State fMRI with Variational Autoencoder”, Under review in NeuroImage.

Woo-Kyoung Yoo, Marine Vernet, **Jung-Hoon Kim**, Anna-Katharine Brem, Shahid Bashir, Fritz Ifert-Miller, Chang-Hwan Im, Mark Eldaief, and Alvaro Pascual-Leone “Interhemispheric and Intrahemispheric Connectivity From the Left Pars Opercularis Within the Language Network Is Modulated by Transcranial Stimulation in Healthy Subjects” Frontiers in Human Neuroscience 14 (2020): 63.

Jung-Hoon Kim, Do-Won Kim, and Chang-Hwan Im. "Brain areas responsible for vigilance: an EEG source imaging study." Brain topography 30, no. 3 (2017): 343-351.

Sunwoo, Jun-Sang, Sanghun Lee, **Jung-Hoon Kim**, Jung-Ah Lim, Tae-Joon Kim, Jung-Ick Byun, Min Hee Jeong et al. "Altered Functional Connectivity in Idiopathic Rapid Eye Movement Sleep Behavior Disorder: A Resting-State EEG Study." Sleep 40.6 (2017).

Minji Lee, Chang-Hyun Park, Chang-Hwan Im, **Jung-Hoon Kim**, Gyu-Hyun Kwon, Laehyun Kim, Won Hyuk Chang, and Yun-Hee Kim, "Motor imagery learning across a sequence of trials in stroke patients", Restorative Neurology and Neuroscience, Restorative neurology and neuroscience 34.4 (2016): 635-645.

Minji Lee, Yun-Hee Kim, Chang-Hwan Im, **Jung-Hoon Kim**, Chang-hyun Park, Won Hyuk Chang, and Ahee Lee, "What is the optimal anodal electrode position for inducing corticomotor excitability changes in transcranial direct current stimulation?", Neuroscience Letters, vol. 584, pp. 347-350, 2015.

Jung-Hoon Kim, Do-Won Kim, Won Hyuk Chang, Yun-Hee Kim, Kiwoong Kim, and Chang-Hwan Im, "Inconsistent outcomes of transcranial direct current stimulation may originate from anatomical differences among individuals: Electric field simulation using individual MRI data", Neuroscience Letters, vol. 564, pp. 6-10, 2014.

Young-Jin Jung, **Jung-Hoon Kim**, Daejeong Kim, and Chang-Hwan Im, "An image-guided transcranial direct current stimulation system: a pilot phantom study", Physiological Measurement, vol. 34, pp. 937-950, 2013.

Young-Jin Jung, **Jung-Hoon Kim**, and Chang-Hwan Im, "COMETS: A MATLAB Toolbox for Simulating Local Electric Fields Generated by Transcranial Direct Current Stimulation (tDCS)," Biomedical Engineering Letters, vol. 3, no. 1, pp. 39-46, 2013.

Patent

Young-Jin Jung, **Jung-Hoon Kim**, and Chang-Hwan Im, "Comets", C-2013-005774, 2013. 03. 21

Conference Presentations

Kim J-H, Kun-Han Lu, Kuan Han, Minkyu Choi, Yizhen Zhang, Zhongming Lui., "Representational Learning of Resting State Functional MRI for Individual Identification", Organization for Human Brain Mapping (OHBM) Annual Meeting, 2020.

Kim J-H, Wen H, Zhang Y, Liu Z., "Mapping Large-Scale Directional Networks with Spectral Features of Electrocorticography", Organization for Human Brain Mapping (OHBM) Annual Meeting, 2019.

Zhang Y, **Kim J-H**, Wen H, Liu Z. (Oral) "High Gamma Electrocorticography in Superior Temporal Gyrus Represents Words during Natural Speech", Organization for Human Brain Mapping (OHBM) Annual Meeting, 2018.

Jung-Hoon Kim, Haiguang Wen, and Zhongming Liu "Development of a EEG-fMRI Fusing Source Imaging Method Based on Neurovascular Coupling Model (FSINC) for Continuous Task Paradigm", Organization of Human Brain Mapping, 2017

Yun-Hee Kim, Minji Lee, Chang-Hwan Im, **Jung-Hoon Kim**, Ahee Lee, Chang-hyun Park, Won Hyuk Chang "What is the Best Position of tDCS Anodal Electrode to Induce the Corticomotor Excitability Change?", Society For Neuroscience, United States, December 16-19, 2014.

Jung-Hoon Kim, Do-Won Kim, Chang-Hwan Im "Inconsistent outcomes of tDCS may originate from anatomical differences among individuals", Organization of Human Brain Mapping, Hamburg, Germany, June 8-12, 2014.

J.-H. Kim, D.-W. Kim, W.-H. Chang, Y.-H. Kim, and C.-H. Im, "Inconsistent outcomes of transcranial direct current stimulation (tDCS) may be originated from the anatomical differences among individuals: A simulation study using individual MRI data", 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Japan, July 03-07, 2013