심리학석사 학위논문

Exploring spatiotemporal brain dynamics with fMRI Transformers

fMRI 트랜스포머를 활용한 시공간 뇌 역동 탐색

2023년 8월

서울대학교 대학원

심리학과 생물심리학 전공

권 준 범

Exploring spatiotemporal brain dynamics with fMRI Transformers

지도 교수 차 지 욱

이 논문을 심리학석사 학위논문으로 제출함

2023년 6월

서울대학교 대학원

심리학과 생물심리학 전공

권 준 범

권 준 범의 심리학석사 학위논문을 인준함

2023년 7월

위 원 장 안 우 영 (인)

부위원장 문 태 섭 (인)

위 원 차 지 욱 (인)

Abstract

The modeling of spatiotemporal brain dynamics from high-dimensional data, such as functional MRI, is a formidable task in neuroscience. However, existing studies predominantly rely on simplistic heuristic features from functional MRI, which poses the risk of overlooking crucial aspects of brain dynamics. This study addresses the limitations of existing computational approaches by proposing two deep neural networks for functional MRI: Swin fMRI Transformer (SwiFT) and Swin fMRI Transformer with UNET (SwiFUN). These models are designed to directly process 4D resting-state fMRI data and predict cognitive and biological variables and specific task-related brain activity. We evaluate our modules using multiple largest-scale human functional brain imaging datasets, such as the Human Connectome Project (HCP), Adolescent Brain Cognitive Development (ABCD) study, and UK Biobank (UKB). Our experimental outcomes reveal that SwiFT consistently outperforms recent state-of-the-art models in predicting sex, age, and cognitive intelligence. Furthermore, SwiFUN surpasses a commonly used approach, a generalized linear model, for predicting task-related brain activity from resting-state fMRI. Our work holds substantial potential in facilitating scalable learning of functional brain imaging in neuroscience research by reducing the hurdles associated with analyzing complex brain dynamics in high-dimensional fMRI.

Keywords: brain dynamics, functional MRI, Transformers, cognitive and biological variables, task-related activity

Student Number: 2021-23364

Acknowledgements

I would like to express my sincere gratitude and appreciation to everyone who contributed to completing this master's thesis. First and foremost, I am deeply grateful to my thesis advisor, Jiook Cha, for his invaluable guidance, support, and expertise throughout this research journey. His profound expertise, insightful feedback, and unwavering encouragement have played a pivotal role in shaping the content and development of this thesis. I would also like to extend my heartfelt thanks to my thesis committee members, Taesup Moon (Seoul National University) and Woo-Young Ahn (Seoul National University), for their time, thoughtful input, and constructive suggestions. Their expertise and scholarly insights have immensely enriched the quality of this work. I appreciate the Shinjae Yoo (Brookhaven National Laboratory) and NESAP (Exascale Science Applications Program) teams, which provided me with a conducive research environment and access to valuable resources. I want to acknowledge and express my sincere appreciation to the co-first author, Yongho Kim from Seoul National University, as well as co-authors Sangyoon Bae from Seoul National University, Sunghwan Joo from Sungkyunkwan University, Donggyu Lee from Sungkyunkwan University, and Yoonho Jung from Seoul National University, for their significant contributions to the 'SwiFT: Swin 4D fMRI Transformer', the paper included in Chapter 2 of this thesis. Their expertise, collaboration, and dedication have been invaluable in designing the research, pre-processing and analyzing the data, and interpreting the results. Finally, I would like to express my sincere gratitude to my family, friends, and loved ones for their unwavering support, patience, and understanding throughout this challenging academic endeavor.

Thank you all.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

Brain Dynamics and its Relation to Adaptive Human Behavior The human brain is a complex and dynamic system characterized as an extensive network generating complex spatiotemporal dynamics of its activity [3]. A growing body of evidence has indicated that the spatiotemporal dynamics of brain activity are key to shaping adaptive human behaviors such as perception [4], attention [5], and emotional processing [6]. The dynamics exhibit distinct patterns across subjects as they engage with an ever-changing environment, thereby manifesting individual variabilities, including cognitive abilities [7, 8, 9, 10] and personality traits, such as neuroticism and extraversion scores [11, 12] On the other hand, psychiatric disorders emerge when the brain dynamics deviate from their normal trajectory during interactions with the environment. This disruption leads to a compromised cognitive and emotional state, impairing the capacity to engage in adaptive behavior. For instance, studies have shown that anomalous disruptions in brain dynamics are highly correlated with psychiatric or neurological

disorders, such as Alzheimer's disease [13], schizophrenia [14], and attention deficit hyperactivity disorder (ADHD) [15]. These abnormalities in brain dynamics are influenced by an individual's genetic predisposition and inherited vulnerabilities. However, genetic factors do not guarantee the development of a mental illness. Instead, they influence the environment one chooses to be in and determine the individual's sensitivity to that environment. Our brain dynamics can function optimally when individuals engage in appropriate interactions with positive life experiences and a nurturing home environment. The Hebbian rule supports this notion by suggesting that the structure and function of the brain can change through the right interactions [16]. Therefore, it is crucial in the fields of neuroscience and medicine to establish the connection between brain dynamics and adaptive cognition and behaviors in the constantly changing environment, as well as their maladaptive manifestations in the context of disease conditions. This enables us to comprehend the individual genetic and environmental factors that impact adaptive brain function and how they are expressed in intricate behaviors. Furthermore, understanding individual brain dynamics will provide the basis for the early identification of risk factors, thereby facilitating the prevention of mental disorders. This knowledge will also contribute to predicting the trajectory of mental disorders and implementing appropriate treatments to prevent their progression.

Group comparison with functional MRI Functional Magnetic Resonance Imaging (fMRI) is a widely utilized neuroimaging modality that enables the noninvasive exploration of intricate brain dynamics, offering a time resolution ranging from 0.5 to 3 seconds. Functional MRI (fMRI) captures a temporal sequence of 3D images (a stack of 2D slices) of Blood Oxygenation Level Dependent (BOLD) signals, which include the physiological changes in blood flow, blood volume, and oxygenation that occur in response to neural activity in the brain. [17]. In fMRI, specific regions are activated together during rest or when performing a specific task, and this continuous pattern rep-

resents the functional network organization of the human brain [18]. Functional MRI has enabled the rapid exploration of detailed functional brain anatomy, significantly contributing to understanding the relationship between complex brain functioning and adaptive human behaviors. To analyze the brain network, researchers usually reduce the sequences of brain volumes into low-dimensional multivariate time series by aggregating voxel intensities in specific brain regions of interest (ROI) based on standardized brain atlases or statistically-clustered parcellation (e.g., Independent component analysis). Pairwise correlations between the time series of each ROI are widely used for analyzing the functional dynamics over the timepoints, called 'functional connectivity' [19]. To investigate the brain regions related to specific brain functioning and its relation to human behaviors, mass-univariate analysis that focuses on group comparison has been widely used for several decades. Mass-univariate analysis refers to a method that compares individuals with a specific disease to a control group without the disease in order to identify group-level structural and functional differences in brain regions. Since the analysis is intuitive and easily interpretable, it has been a longstanding tool for analyzing brain images and has advanced our understanding of brain mechanisms for psychiatric diseases. However, there are three primary limitations associated with mass-univariate analysis. Firstly, this methodology assumes independence among each brain region and voxel, conducting multiple independent comparisons for every voxel to determine the extent of significant differences between disease and control groups. However, this assumption significantly overlooks the characteristic of the brain as a complex network, where each brain regions consistently interact with each other for adaptive brain functioning [20, 21, 22]. Secondly, mass-univariate analysis for Functional MRI is highly susceptible to various sources of noise, such as scanner effects and head movements, which decreases test-retest reliability within and across subjects [23, 24, 25, 26]. The scanner effect refers to the phenomenon where brain scans exhibit different probabilistic distributions depending on the type of scanner used. This

effect complicates the integration of data from different sources in multi-site studies, which involve using various types of scanners. Lastly, while mass-univariate analysis can identify group differences between disease and control groups, it cannot be used to predict individual-level outcomes from a single MRI scan because of its low signal-to-noise ratio (SNR) [20, 27]. The research on fMRI has been hampered by the lack of predictive power owing to the gap between the complexity of the brain network and the contrasting simplicity of brain imaging analytics [28, 27].

Individual-level prediction with Machine Learning The recent objective in neuroscience and psychology is not only to identify group differences but also to discover individual variability in brain dynamics that determine cognitive and behavioral differences. To provide personalized diagnoses and prognoses of mental disorders, developing integrative biomarkers including brain imaging, genome, and social information is of paramount importance in neuroscience and precision psychiatry. To achieve this objective, machine learning (ML) applications have emerged as a highly influential approach in developing biomarkers by modeling complex non-linear relationships between extracted brain features and behavioral outcomes, expected to surpass the limited predictive power imposed by traditional linear models. Unlike the mass-univariate approach, ML methods assume inter-correlation between brain regions and find the best combinations of brain features for the optimal individual-level prediction, shifting the research objective from group-based comparison to personalized diagnosis and prognosis. The effectiveness of machine learning approaches is determined by preprocessing methods that support the hypothesis. This requires the expertise of domain specialists to determine how to extract features (e.g., connectivity matrices) from large fMRI datasets. There have been attempts to reduce the dimension of brain images using machine learning-based feature extraction and predict diseases from an individual's functional connectivity through predictive ML models such as

Connectome-based predictive model (CPM), support vector machine (SVM), and ensemble methods (e.g., XGBoost). For instance, Galioulline et al. [29] proposed that the future incidence of depression can be forecasted using resting state fMRI from healthy adults by combining ML methods such as regression dynamic causal models (rDCMs) and support vector machine (SVM). Moreover, some research focused on differentiating various subtypes within heterogeneous psychiatric disorders. They established connections between the subtypes and treatment responsiveness, aiming to provide personalized treatments based on biomarkers derived from functional MRI [30]. These studies target psychiatric disorders such as depression and ADHD that encompass a wide range of co-occurring symptoms and exhibit varied responses to treatment. DrynsDale et al. [31] suggested that depressive symptom-related subtypes defined by functional connectivity can enhance the diagnostic accuracy of depression and predict the responses to the anti-depressant. Specifically, the study used canonical correlation to extract informative features from resting-state functional connectivity, discovering four subtypes based on the features with hierarchical clustering. They verified the high clinical utility of the subtypes by classifying the depressive subtypes of unseen subjects and predicting the treatment effect of repetitive transcranial magnetic stimulation (rTMS). However, Dinga et al. [32] criticized the limited reproducibility of the result, pointing out the low statistical significance of canonical correlation and clusters observed in another dataset. [32] Despite the promising predictive power of ML-based approaches in many studies, researchers consistently cast doubt on the reliability of ML-based approaches. The same ML models can exhibit inconsistent performance depending on how the brain images are preprocessed. Inappropriate feature extraction aggravates the biases in fMRI, for example, over- or under-correcting scanner effects [33, 28]. Furthermore, depending on extracted features can overlook vital factors in spatiotemporal dynamics, making the predictive model focus on either spatial or temporal dynamics of functional MRI. Several studies suggest that minimal

preprocessing can enhance the performances of predictive models by maximally utilizing the information in brain MRI. [34] Unfortunately, machine learning is not suitable for dealing with raw images. [35, 36] Given the aforementioned constraints, identifying biomarkers that can aid in psychiatric diagnosis or prognosis at the individual level is exceedingly challenging, and substantial limitations exist when applying these methods in practice.

1.2 Deep Learning for fMRI

Deep learning for personalized approach Deep learning models have emerged as a promising solution for discriminating subtle individual differences in neuroimaging. Unlike the machine learning approach that requires sophisticated feature engineering before the analysis, the deep learning approach automatically extracts essential information (representations) from minimally pre-processed fMRI data. This characteristic alleviates concerns regarding the compromise of critical information through pre-processing methods. By maximally utilizing rich information in fMRI, deep learning algorithms can uncover hidden patterns in intrinsic brain dynamics that may not be apparent through simple linear models and machine learning approaches. Additionally, previous research shows that deep learning models exhibit significantly higher predictive performance than traditional machine learning models when trained on a sufficient amount of fMRI data [28]. The rich representation of the deep learning approach stems from its hierarchical structure also called a deep neural network (DNN). The deep neural network applies multiple non-linear transformations to input data to model various levels of complexity [35]. At each layer, a deep learning model represents distinct features of the input data. Low-level features tend to capture elementary components in the data such as contours, edges, and colors for the image, which is more susceptible to noise. On the contrary, high-level features learned by deep neural networks are more

semantically meaningful and suitable for discerning content within images that are robust to noise. Methods such as multi-layer perceptron (MLP), convolutional neural network (CNN), and Transformer are commonly employed to effectively capture high-level features. High-level features in deep neural networks can provide insightful solutions to detect individual differences in brain dynamics with highly intricate and subtle patterns, which are not easily distinguishable from noise [20]. These attributes allow higher performances in discriminating individual variability than previous approaches. For instance, research on deep learning approaches has demonstrated outstanding performance in identifying psychiatric disorders compared to existing machine learning approaches [3, 37, 38]. Deep learning models also offer remarkable flexibility in processing the brain in diverse ways based on the characteristics of the brain. For instance, a recent deep learning approach such as graph neural networks (GNNs) processes brain images as graphs, which utilizes the spatiotemporal locality of the brain network [39]. The spatiotemporal locality indicates that brain regions in close proximity in both space and time exhibit a higher degree of information exchange [40, 41]. In traditional approaches, the relationship between different brain regions is often determined by calculating correlations between fMRI time series corresponding to regions of interest (ROI), resulting in a fixed and explicit form of functional connectivity. This approach typically processes spatial and temporal information independently or focuses on either of them. On the other hand, graph neural networks (GNNs) allow for a more dynamic and implicit representation of connectivity by learning the ”embedding” of connections between regions. This flexible approach enables the extraction of features from brain networks that are relevant to behavioral characteristics. Some variants of GNNs incorporate spatial and temporal interactions between distant brain regions, showing superior performances than traditional functional connectivity. For example, spatiotemporal graph neural network(ST-GCN) has shown higher predictive performance in biological variables such as sex and age than traditional machine learn-

ing models. It achieves this by learning patterns in the spatiotemporal connectivity between brain regions based on parcellated fMRI timeseries [40]. The idea of simultaneously learning spatial and temporal representation has emerged as a prominent trend in the analysis of fMRI using deep neural networks.

Challenges in deep learning approach Several methods have been proposed to apply deep learning methods to fMRI data, but most of them have been analyzed on multi-variate fMRI timeseries where features are extracted per ROI using statistical clustering or anatomical atlas. While this approach allows the model to learn much more information than traditional functional connectivity, the feature extraction process can also cause information loss. To ensure that the model is extracting the most information possible, it is preferable to run the analysis on minimally processed 4D fMRI. However, 4D fMRI is a sequence of 3D volumes, and processing such a model requires significant computing resources. In recent years, there have been many studies using 3D structural MRI for end-to-end learning, but few researchers have the infrastructure (e.g., several terabytes of storage and GPU resources) to analyze 4D fMRI directly, so it has not received as much attention as ROI-based methods. However, the few studies that have processed four-dimensional brain images suggest that this approach significantly outperforms traditional ROI methods. This task is not only feasible but also holds immense value in terms of exploring uncharted areas of research. However, there are several challenges to developing such a model using 4D fMRI. Deep neural networks typically require a substantial number of samples to perform well and generalize to independent datasets [42]. Insufficient samples can result in unstable training, overfitting to a small number of samples, and poor generalizability. In particular, fMRI data is high-dimensional data with a low signal-to-noise ratio (SNR), which requires a substantial amount of training data to utilize such data without dimensionality reduction [43]. However, performing neuroimaging studies to ac-

quire fMRI data is costly and requires specialized expertise, which makes it hard to increase the sample size. To address these limitations, transfer learning has emerged as a promising technique. Transfer learning is a method to transfer knowledge to solve one task (source domain) to another similar task (target domain). Typically, a source domain includes training a large-scale model with huge amounts of data in supervised or self-supervised ways [44], which is called the pre-training stage. The pre-trained model from the source domain is applied to the target domain with smaller datasets to complement the insufficient sample size and achieve better performances. Sometimes, the stage requires some updates in the weight of the deep neural network, which is called fine-tuning. The 'transfer and finetune' approach has demonstrated its effectiveness in analyzing neuroimaging. Population studies such as Adolescent Brain Cognitive Development (ABCD), Human Connectome Project (HCP), and UK Biobank (UKB) have emerged as valuable resources for large-scale pre-training. By transferring the knowledge of models trained on these large-scale datasets to smaller disease datasets, the limited sample size of traditional disease studies can be compensated. For instance, studies have shown that models pre-trained on large-scale structural MRI datasets can enhance their prediction performance when applied to other disease-related data [45, 46]. Several studies in the field of fMRI have provided evidence of the effectiveness of transfer learning, either in two sources of datasets [47, 48] or different tasks in the same dataset [49, 50]. However, to the best of our knowledge, no models have been developed for 4D fMRI data that can effectively scale to various types of data. Unlike computer vision (CV) and natural language processing (NLP) research, which have seen significant investment and interest in developing large-scale models, limited computing resources (e.g., several terabytes of storage and GPU resources) and the resulting lack of attention from scientists has hindered the development of large-scale foundation models for functional MRI. These limitations in fMRI studies highlight the critical need to develop efficient and scalable models for effective

representation learning.

Another challenge inherent in deep learning is the difficulty in interpreting the reasons behind a model's predictions, regardless of its performance. Deep learning models have long been referred to as "black boxes," and in the context of psychiatric diseases, it is crucial not only to achieve high diagnostic and prognostic performance but also to understand the changes in brain regions that lead to such conclusions. This is particularly important for real-world applications in medical fields and precision psychology. Deep learning models, directly processing high-dimensional fMRI data for prediction tasks, can offer valuable insights into which brain regions contributed to the prediction. The result can be visualized on high-resolution brain images. For example, Nguyen et al. [49] demonstrated that GradCAM, a well-known interpretation method developed for 2D naturalistic images, can be applied to 4D fMRI data to visualize the explanatory regions. They trained an attention-based deep neural network (DNN) to detect the type of task from short sequences of 4D task-state fMRI data. This approach reveals the predominantly activated brain regions over time during specific types of tasks. The findings of this study align with previous research, confirming the consistency of the identified brain regions associated with the specific tasks [49]. This result implies that recently famous interpretation methods in the computer vision domain, such as Integrate Gradients, can be applied to 4D fMRI as well.

1.3 Research Aims and Thesis Outline

This thesis aims to improve our understanding of brain dynamics in functional MRI relating it to cognitive and biological factors, as well as complex human behaviors. Although deep learning has recently been heavily applied to ROI-based parcellated fMRI, few deep learning studies have been developed for minimally preprocessed 4D fMRI. This is due to insufficient sample size, inefficient and inappro-

priate model structures for spatiotemporal information, and the lack of explainable AI (XAI) methods applicable to neuroscience. Currently, fMRI research faces challenges in developing suitable methodologies to process a huge amount of fMRI data from increasing population studies such as Adolescent Brain Cognitive Development (ABCD), Human Connectome Project (HCP), and UK Biobank (UKB).

Transformer, a deep neural network developed to perform natural language processing (NLP), has profoundly impacted society, revolutionizing diverse domains including voice recognition, machine translation, and image recognition. Its unparalleled capacity to comprehend intricate patterns and dependencies within data has been instrumental in transforming these fields. Furthermore, Transformers have also been applied to neuroimaging modalities to target diverse research topics in psychology and neuroscience such as diagnosing psychiatric disorders and detecting brain tumors. The main advantage of Transformers is that they extract the most important information from data based on a multi-head self-attention mechanism, and they exhibit scalable performance improvements as the number of data increases compared to existing deep learning models. Transformer models have demonstrated remarkable transfer learning capabilities, enabling knowledge transfer from large-scale models to smaller ones. This strength allows for the efficient extension of knowledge across different tasks and domains, facilitating effective adaptation and utilization of pre-trained models. Existing studies have shown significantly higher prediction performance using Transformers for minimally preprocessed 4D fMRI, suggesting the promise of this research direction [49, 50]. We demonstrate that Transformers may tackle significant challenges in neuroscience.

Therefore, this study addresses the limitations of existing fMRI-based studies by extending the recently developed Video Swin Transformer [51], designed for efficient video recognition tasks, to the realm of 4D fMRI. This research proposes an efficient, scalable, and interpretable 4D fMRI transformer. This study shows the effectiveness

of the developed 4D fMRI Transformer in predicting human biological and cognitive variables. Furthermore, we demonstrate that the utility of the model can be extended to predict individual differences in brain activity during specific tasks from resting-state fMRI—a more challenging task. This thesis consists of two studies which aim to accomplish two following objectives:

Objective 1: To develop an efficient, scalable, and interpretable fMRI Transformer for biological and cognitive variables prediction.

The key research questions for this objective are:

• Does fMRI Transformer exhibit higher predictive performance and computational efficiency than baseline models?

• How effective is the fMRI Transformer for the transfer learning between different datasets?

• Which brain regions are observed with an explainable AI method and what can be inferred from the result?

Objective 2: To predict performed task-state brain activity from resting-state fMRI

The key research questions for this objective are:

• Does the predicted task activation map generated by the Transformer model exhibit superior overall qualities compared to those produced by the baseline models?

• Does the predicted task activation map exhibit individual variability?

Chapter 2

EFFICIENT 4D FUNCTIONAL MRI TRANSFORMERS

2.1 Introduction

Recently, deep neural networks have been applied to fMRI to investigate the nonlinear relationship of brain dynamics with human cognition and behaviors [37, 38, 20, 40, 52]. Researchers have broadly pursued two distinct lines of work. The first approach is the so-called *ROI-based method*, in which the high-dimensional fMRI data (with around 300,000 voxels) is clustered into the temporal sequence of hundreds of pre-defined brain regions (ROIs) using anatomical segmentation [53] or statistical clustering (e.g., Independent component analysis) [54]. The choice of method for extracting dynamics depends on the hypothesis being tested. By averaging the voxel intensities within each brain region, this approach reduces volumes into low-dimensional features, approximating the number of regions of interest (ROIs) assuming that voxels within each ROI will demonstrate similar activities across multiple time points. However, manually extracting features from volumes may be prone to losing information

important to capture subtle variability across the individual brains [55]. Additionally, feature extraction can be time-consuming and, depending on how it is performed, can lead to inconsistent results even for the same data. If the preprocessing is not done in a proper way, the extracted features can be sensitive to the effects of scanning devices or parameters, which can decrease the true effect of the extracted features [33]. The second DNN-based approach is the *two-step deep learning approaches*, in which the fMRI data is used as input, with specialized architectures used for spatial and temporal domains separately for better computation and memory efficiency. Namely, for learning spatial features, convolutional neural networks (CNNs) are used, and for temporal, long short-term memory (LSTM) [56] or Transformers [50, 49] are used. By making deep neural networks learn spatial and temporal representation from raw fMRI data with four dimensions, deep learning can fully utilize the information from fMRI essential for the given task [55, 34]. Previous studies have reported that 4D models show superior performances in fMRI classification or regression tasks compared to models using hand-crafted features [56, 50, 49, 50]. However, separating the spatial and temporal domains may limit the capability of capturing comprehensive information among brain regions across time points. Therefore, a critical unresolved issue is whether an efficient, end-to-end DNN that utilizes 4D fMRI input can be formulated to better model and learn the brain dynamics compared to previous approaches.

To incorporate spatial and temporal features with attention, research on video recognition has proposed pure transformer-based models computing attention over whole image patches in the video. However, the approaches adopted factorization methods that two separate attention-based encoders sequentially process spatial and temporal interactions [57, 58]. Furthermore, video Swin Transformer, which features hierarchical transformer architecture, was proposed to replace the separate spatial and temporal attention with local attention focusing on the local relationship between nearby patches, showing considerably higher efficiency and performances in video

recognition tasks [51]. However, to our best knowledge, such a method has yet to be applied to fMRI.

To that end, we propose **Swi**n 4D **f**MRI **T**ransformer **(SwiFT)**, a 4D extension of the Swin Transformer [59] architecture, which can jointly learn the spatiotemporal representations of the brain's intrinsic activity directly from high-dimensional fMRI in an end-to-end fashion. The main gist of our method is to employ the 4D local window attention structure, which makes SwiFT readily applicable to process large-scale, high-dimensional 4D data with low computational complexity. We note that while 3D variants of Swin Transformers have been proposed before [60, 57, 58, 51] to take video or medical image inputs, to the best of our knowledge, this is the first work to extend Swin Transformer to take 4D data input and to apply it to the fMRI data. Our experimental results show that the end-to-end learning capability of SwiFT unlocks its potential to learn complex spatiotemporal patterns in fMRI effectively. Specifically, we evaluate SwiFT's performance on three representative fMRI benchmarks: the Human Connectome Project (HCP) [61], the Adolescent Brain Cognitive Development (ABCD) [62], and the UK Biobank (UKB) [63, 64]. Across various classification and regression tasks, including sex classification and age/intelligence prediction, SwiFT significantly outperforms the recent baselines of the above three kinds: *i.e,* those based on simple feature-based ML, ROI-based DNNs, or DNNs with separate architecture for spatial and temporal signals. Furthermore, we also demonstrate that it would be feasible to apply the widely successful "pre-train and fine-tune" framework to SwiFT. Namely, by pre-training SwiFT using contrastive loss-based self-supervised learning, we show that fine-tuning the pre-trained model for each specific task yields superior results compared to models trained from scratch. We believe this capability has the potential to empower researchers to construct large-scale foundation models for fMRI akin to those utilized in several other application domains. Finally, to provide a comprehensive analysis, we present the interpretation results using Integrated gradient

with Smoothgrad sQuare (IG-SQ), a recent explainable artificial intelligence (XAI) technique, for SwiFT's predictions and conduct ablation studies to substantiate our modeling choices.

2.2 Related Work

ROI-based Models To analyze the brain network, researchers typically reduce the sequences of brain volumes into low-dimensional multivariate time series by aggregating voxel intensities in specific regions of interest (ROI) based on standardized brain atlases, considering pairwise correlations between the time series of each ROI as functional connectivity[19, 65]. Most DNNs, such as graph neural networks (GNN), were designed to treat the brain network as a graph, considering each ROI as nodes and pairwise correlation between them as edges. For instance, BrainNetCNN, consisting of multiple graph convolutional filters, was proposed to model various levels of topological interactions in structural [66] and functional connectivity [52]. Kan et al. [52] proposed Transformer for analyzing brain networks, which employs attention weights for learning individual connectivity strengths between each ROI and applies an orthonormal clustering readout operation for functional connectivity to locate functional clusters related to specific human behaviors, acquiring informative embeddings for predicting psychiatric and biological outcomes. Some recent studies proposed methods to capture spatiotemporal dynamics directly from extracted fMRI timeseries, utilizing Transformer by separating spatial and temporal attention units [67], introducing masked sequence modeling [68], and focusing on local representations with fused window multi-head self-attention (FW-MSA) [69].

4D fMRI-based Models Existing DNNs for 4D fMRI typically process spatial and temporal information separately. The C3d-LSTM [56] integrates 3D convolutional neural networks (CNNs) to extract spatial embeddings in each 3D volume of a 4D

fMRI and then feeds the spatial embeddings to LSTM for temporal encoding. TFF [50] replaces LSTM with Transformer for extracting temporal features and proposes reconstruction-based pre-training steps. To learn spatial features before the downstream task, TFF concatenates decoder layers after Transformer and minimizes three reconstruction-based losses; L1 loss, perceptual loss, and intensity loss. Brain Attend and Decode (BAnD) [49] suggests a pre-training method to predict the types of cognitive tasks performed during an fMRI scanning. A 3D CNN encoder is then trained to predict the target variable from fMRI volumes. Furthermore, the pre-trained encoder learns temporal features by attaching multi-head self-attention layers. Of note, the aforementioned models may have issues of unstable training of spatial and temporal data, which involves multiple training steps and large memory usage. These limitations may result in sub-optimal model computation and learning capability.

Transformers for Computer Vision Tasks Following the success of Transformers in natural language processing [70, 71], many works apply the multi-head self-attention mechanism of Transformers for computer vision tasks, such as image classification, image segmentation, and object detection. One of the major challenges here is the computation complexity increasing quadratically with respect to the number of tokens, an elementary unit for Transformers typically amounting to several hundred. As images have a much larger number of unique tokens (pixels) than word tokens in natural languages, using image pixels for the input token to Transformers was infeasible in most cases. Vision transformer (VIT) [72] tackles this issue by introducing a unique token unit for vision tasks, a patch consisting of several image pixels, significantly decreasing the number of tokens compared to using image pixels. However, VIT does not solve the quadratic increase of computational costs in self-attention layers, limiting the wider application of VIT for vision tasks such as semantic segmentation and object detection. Swin Transformer [59] reduces the computational complexity to be linear to

the number of tokens by applying self-attention only within a local window, consisting of several patches, instead of running it over whole image patches. Along with the local windowed attention, Swin Transformer also introduces shifted window attention to allow cross-window connections and patch merging (downsampling) steps to produce hierarchical representations. This approach proved successful in many vision tasks, such as image classification, object detection, and semantic segmentation. Swin UNETR [1] demonstrated Swin Transformer's utility in brain tumor segmentation tasks using 3D structural MRI by coupling a Swin Transformer encoder with CNN-based decoders. Volumetric Aggregation with Transformers [73], a 4D Convolutional Swin Transformer, extended the Swin Transformer model to accept a 4D correlation map of two CNN-extracted 2D image features, utilizing the model for cost aggregation. Liu et al. [51] suggested the utility of Swin Transformer for capturing spatiotemporal dynamics, applying it to general video recognition tasks including human action recognition, showing considerably higher efficiency and performances compared to previous video recognition models [57, 58]. Overall, these works present the feasibility of applying Swin Transformers to higher spatiotemporal dimensions, and to the best of our knowledge, such a method has yet to be applied to functional brain imaging.

2.3 Method Description

2.3.1 Swin 4D fMRI Transformer (SwiFT)

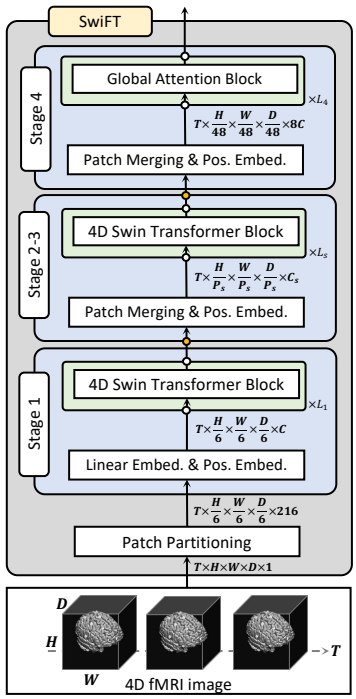
Overall architecture In line with the recent studies [51, 60], which introduce 3D extensions to enhance the capabilities of the Swin Transformer [59], we propose an advancement in the architecture to incorporate an additional temporal dimension, thereby enabling its application to 4D data. It is worth noting that no previous attempts have been made to extend the Swin Transformer in this manner.

The overall architecture of our model is depicted in Figure 2.1a. The SwiFT archi-

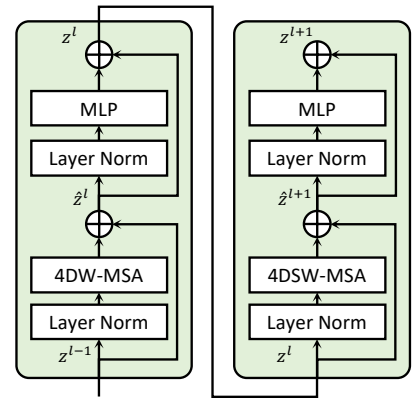
tecture utilized in our study consists of four distinct stages. Each stage is constructed through the implementation of patch merging, with linear embedding employed in the case of Stage 1. Additionally, positional embedding is incorporated, and multiple (Swin) Transformer blocks are applied repeatedly within the stages. The model processes an input fMRI data of size $T \times H \times W \times D \times 1$, which consists of a length T sequence of fMRI volumes ($H \times W \times D$) with a single channel. During the initial patch partitioning step, the input fMRI data is partitioned into $T \times \frac{H}{P} \times \frac{W}{P} \times \frac{D}{P}$ patches with P^3 voxels. In this study, H, W, and D are 96, and the initial patch size P is 6. During the linear embedding process, patches with size P^3 are transformed into C-dimensional tokens. This transformation effectively maps the spatially-neighboring pixels within a patch onto a token.

Next, following an absolute positional embedding layer, multiple layers of 4D Swin Transformer blocks are applied on the embedded patches, forming Stage 1 of SwiFT. Starting from Stage 2, a patch merging layer at the beginning of each stage reduces the number of tokens by merging 8 spatially-neighboring patches. After the patch merging layer, an absolute positional embedding layer followed by multiple layers of 4D Swin Transformer blocks is applied, together forming Stage 2 and onward. In the final stage, Stage 4, the 4D Swin Transformer blocks are replaced by global attention Transformer blocks which carry out global attention instead of local window attention. This computationally expensive global attention is made possible by the significant reduction in the number of tokens achieved through the patch merging steps executed in the preceding stages. Global attention Transformer blocks allow each token to globally attend to all other tokens rather than being restricted to the tokens within the local window.

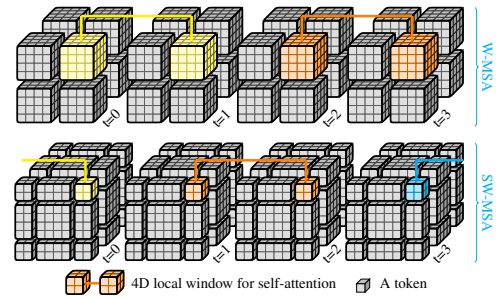
Patch merging Following prior works [51, 60, 74, 75], the patch merging step is only performed for the three spatial dimensions (H, W, D) and not for the temporal



(a) Overall architecture of SwiFT



(b) Successive 4D Swin Transformer Blocks



(c) 4DW-MSA and 4DSW-MSA

Figure 2.1: Figures depicting the structure of SwiFT and its components.

dimension (T), thereby merging a group of $8 = 2 \times 2 \times 2$ neighboring patches into a single patch for each time frame. During the patch merging operation, the spatial dimensions are reduced by half, and the channel size (C) is doubled as compensation. Thus, in figure 2.1a, the numbers P_2, P_3, C_2, and C_3 are $12, 24, 2C$, and $4C$, respectively.

As a general example, consider a tensor with arbitrary dimensions $T \times H' \times W' \times D' \times C'$ before passing through the patch merging layer. During patch merging, this tensor is reshaped into a new tensor with dimensions $T \times \frac{H'}{2} \times \frac{W'}{2} \times \frac{D'}{2} \times 8C'$, where $2 \times 2 \times 2$ spatially neighboring patches are concatenated along the channel dimension. Then, each $8C'$ channel in the resulting tensor is projected onto a $2C'$ dimensional space by apply-

ing a single fully connected layer, resulting $T \times \frac{H'}{2} \times \frac{W'}{2} \times \frac{D'}{2} \times 2C'$ in total. The process of patch merging facilitates the hierarchical feature-extraction structure of SwiFT and reduces the computational complexity of the subsequent layers. We clarify that while the patch merging is operated only on the spatial dimensions, the temporal information is still well-incorporated via the windowed attention.

4D window multi-head self-attention The core of the Swin Transformer model is the window multi-head self-attention (W-MSA) layer, which allows the model to process a larger number of tokens while limiting self-attention only within a local window. In SwiFT, the 3D window mechanism is extended to 4D windows; given input tokens with a size of $T \times H' \times W' \times D'$, the tokens are partitioned by a predetermined window size of $P \times M \times M \times M$, resulting in $\lceil \frac{T}{P} \rceil \times \lceil \frac{H'}{M} \rceil \times \lceil \frac{W'}{M} \rceil \times \lceil \frac{D'}{M} \rceil$ non-overlapping local windows.

However, simply stacking multiple window self-attention layers would be undesirable since there would be no crosstalk across different windows. To that end, a shifted window multi-head self-attention (SW-MSA) layer enables cross-window connections. Namely, we extend the 3D shifted window mechanism to 4D shifted windows as well; in $P \times M \times M \times M$ windows obtained from the W-MSA layer, we shift the windows of the successive layer by $(\frac{P}{2}, \frac{M}{2}, \frac{M}{2}, \frac{M}{2})$ tokens.

The detailed operations of our 4D W-MSA and SW-MSA are shown in Figure 2.1c. In this example, the applied size of input tokens and the windows are $T \times H' \times W' \times D' = 4 \times 8 \times 8 \times 8$ and $P \times M \times M \times M = 2 \times 4 \times 4 \times 4$, respectively. Then, by following the window partitioning methods described above, the numbers of grouped windows in W-MSA and SW-MSA become $2 \times 2 \times 2 \times 2 = 16$ and $3 \times 3 \times 3 \times 3 = 81$, respectively. Such separately applied window self-attention plays a key role in effectively extracting spatiotemporal feature representation from the 4D fMRI data. Note that although the number of windows increases in SW-MSA, the actual computation

cost is maintained to be similar by leveraging the cyclic-shifting batch computation proposed in [59].

Combining the W-MSA layer and the SW-MSA layer, two successive 4D Swin Transformer blocks, as shown in Figure 2.1b, are computed as the following:

$$\hat{z}^l = \text{4DW-MSA}(\text{LN}(z^{l-1})) + z^{l-1}, \quad z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l$$
$$\hat{z}^{l+1} = \text{4DSW-MSA}(\text{LN}(z^l)) + z^l, \quad z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1},$$

in which 4D(S)W-MSA, LN, and MLP denote the 4D (Shifted) Window Multi-head Self-Attention, Layer Norm, and Multi-Layer Perceptron module, respectively. Moreover, $\hat{z}^l$ and z^l denote the output features of the 4D(S)W-MSA module and the following MLP module for block l, respectively.

4D absolute positional embedding Even though previous models utilize relative position biases to encode positional information, we have opted instead for an absolute position embedding scheme for SwiFT. While absolute positional embeddings are more computationally expensive for low-dimensional data [59], since we are dealing with much larger scale 4D data, the absolute positional embeddings become more cost-effective than the relative positional bias. We compare the effectiveness of the two positional embedding methods in A.3.

To that end, we add a learnable embedding at the beginning of each stage of the Transformer right after the patch merging step. In line with [58], we separately add positional embeddings for the spatial and temporal dimensions. Specifically, given an input tensor with dimensions of $T \times H' \times W' \times D' \times C'$, we define spatial and temporal positional embedding tensors with dimensions of $1 \times H' \times W' \times D' \times C'$ and $T \times 1 \times 1 \times 1 \times C'$, respectively. These tensors are then added to the input tensor using broadcasting.

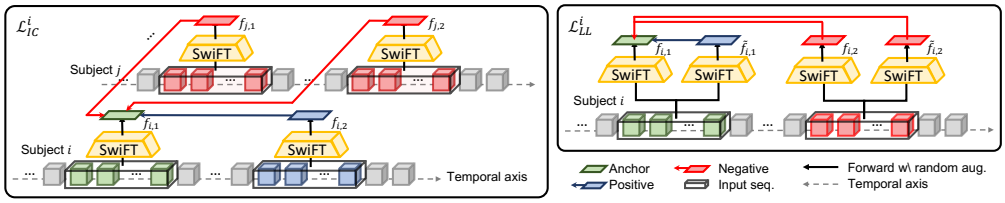


Figure 2.2: Illustration of two different contrastive losses for the pre-training of SwiFT. (left: Instance contrastive loss, right: Local-local temporal contrastive loss)

2.3.2 Self-supervised Pre-trainining

Our proposed end-to-end model structure allows efficient self-supervised pre-training of SwiFT, which can then be fine-tuned for specific tasks. This unique capability sets our method apart from other approaches in Section 2.2 relying on ROI-based brain data or a two-step learning approach for 4D fMRI. We achieve this by using two different contrastive loss-based pre-training objectives adapted from [76]. Figure 2.2 depicts the positive and negative pairs for the two loss functions, where the InfoNCE [77] loss of the pairs is calculated for the final loss function.

Instance contrastive loss The instance contrastive loss ($\mathcal{L}_{IC}$) is a type of contrastive-based loss function that considers a representation to be positive if two distinct fMRI sub-sequences come from the same subject and negative if they come from different subjects. The feature representation passes through three layers: SwiFT, global average pooling, and a multi-layer perceptron (MLP) head. To clearly define this loss function, we denote the feature representation as $f_{i,p}$, where $i \in \{1, ..., B\}$ refers to the subject index for a given batch size B, and p refers to the fMRI sub-sequence index (either 1 or 2). Since we are sampling two sub-sequences for each of the B subjects, this amounts to a total of $2B$ representations. For each subject i, the anchor, positive, and negative representations are set as $f_{i,1}$, $f_{i,2}$, and the remaining $2B - 2$ feature representations, respectively. Using this setup, the instance contrastive loss for subject i is denoted and

defined as

$$\mathcal{L}_{IC}^{i} = -\log \frac{h(f_{i,1}, f_{i,2})}{\sum_{j=1}^{B}[\mathbb{1}_{[j\neq i]}\left(h(f_{i,1}, f_{j,1}) + h(f_{i,1}, f_{j,2})\right]\right)},$$

where h denotes the exponential of the cosine similarity between two vectors, and $\mathbb{1}$ denotes an indicator function that equals one if the condition is true and equals zero otherwise.

Local-local temporal contrastive loss The local-local temporal contrastive loss ($\mathcal{L}_{LL}$), in contrast to instance contrastive loss, determines both positive and negative pairs within a single subject. Namely, a positive representation is derived from the same fMRI sub-sequence, but they are applied with different random augmentations. On the other hand, negative representations come from distinct fMRI sub-sequences from the same subjects. The feature representations are obtained in the same manner as the instance contrastive loss. To clearly define this loss function, we use the same notation of feature representation as $f_{i,p}$, but the range of p is changed to $\{1, 2, ..., N\}$, where N is the number of fMRI sub-sequences from a single subject. We denote an fMRI sub-sequence of the same subject i and fMRI sub-sequence p but different random augmentation as $\tilde{f}_{i,p}$. Since we are sampling two differently augmented versions for each of the N sub-sequences, this amounts to a total of $2N$ representations. Using this setup, the local-local temporal contrastive loss for subject i is denoted as $\mathcal{L}_{LL}^{i}$ and defined as

$$\mathcal{L}_{LL}^{i} = -\sum_{p=1}^{N} \log \frac{h(f_{i,p}, \tilde{f}_{i,p})}{\sum_{q=1}^{N}[\mathbb{1}_{[q\neq p]}(h(f_{i,p}, f_{i,q}) + h(f_{i,p}, \tilde{f}_{i,q}))]},$$

where h denotes the exponential of the cosine similarity between two vectors.

2.4 Experiments

2.4.1 Experimental Setting

Datasets The Adolescent Brain Cognitive Development (ABCD) study is a longitudinal, multi-site investigation of brain development and related behavioral outcomes in children [62]. The dataset is open to the scientific community but requires authorized access. After quality control, we used the resting state fMRI of 9,128 children (age $= 118.95 \pm 7.46$ months, 52.4% female) from release 2. For fMRI preprocessing, we used a well-established pipeline, fMRIprep [78, 79], which includes reducing the bias field, skull-stripping, alignment to structural image, and spatial normalization to standard space for a pediatric brain [80]. After fMRIprep, we applied low pass filtering to smooth signal, head movement correction, and artifact removal regressing out signals from non-grey matters (aCompcor) [81].

We also used the resting-state fMRI of 1,084 healthy young adults (age $= 28.80 \pm 3.70$ years, 54.4% female) from the Human Connectome Project (HCP) (S1200 data) [82, 83], and 5,935 middle and old aged adults (age $= 54.971 \pm 7.53$ years, 52.7% female) from the UK Biobank (UKB) [84]. We used preprocessed data provided by Human Connectome Projects [85] and UK Biobank [63, 64], which follows the fMRI volume pipeline, including reducing the bias field, skull-stripping, cross-modality registration, and spatial normalization to standard space.

For each of the 4D fMRI volumes, we globally normalized brain images over the four dimensions except for the background regions. Then we filled the background with the minimal voxel intensity value. To easily divide the volume into patches for SwiFT, we changed the 3D volume into a shape of $96 \times 96 \times 96$ by cropping and padding on the background. To evaluate the performances of ROI-based models, following the preprocessing steps of [52] as closely as possible, we applied the HCP MMP1 atlas [86] to each fMRI volume to obtain the time series data for each ROI. Sub-

sequently, we processed this ROI series to generate functional connectivities, which involves computing the Pearson correlation coefficient to construct a correlation matrix. The correlation matrix is then Fisher Transformed, serving as the input for the ROI-based models in Section 2.4.1.

To evaluate our models, we constructed three random splits with a ratio of (train: validation: test) = $(0.7 : 0.15 : 0.15)$ and reported the average performances across the three splits.

Targets We chose the sex [87], age [88], and cognitive intelligence (NIH Toolbox [89] for ABCD, HCP datasets, and "fluid intelligence" for UKB dataset) of each subject as the prediction target for our models. These targets are significant since the relationship between the brain and these targets represents a fundamental brain-biology and brain-cognition association. Also, the capability to predict these outcomes can prove the model's capability to process fMRI volumes, possibly leading to the prediction of clinical outcomes of debilitating brain disorders, such as Alzheimer's disease, schizophrenia, autism, and bipolar disorder [90, 91, 92, 93]. For these reasons, predicting these outcomes from brain imaging has been an important benchmark task in recent computational neuroscience [52, 37, 94].

The regression targets (age, intelligence) were z-normalized to bring stable training regardless of the range of the target variable. Since the age has a unit (e.g., years or months), we transformed the z-scaled age back to its original scale of months or years when reporting the performance metrics.

Balanced accuracy and AUC (Area Under ROC Curve) were used to evaluate model performances for the binary classification task. Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used to evaluate model performances for the regression tasks.

Implementation Details

For SwiFT, we use the same architecture across all of our experiments, using the architecture corresponding to the Swin-T variant from [59, 51] with a channel number of $C = 36$. The numbers of layers are fixed to $\{L_1, L_2, L_3, L_4\} = \{2, 2, 6, 2\}$ which corresponds to a model with three stages with 2, 2, 6 consecutive 4D Swin Transformer blocks for each stage and a final stage with two consecutive global attention Transformer blocks. In the 4D(S)W-MSA cases, we set $P = M = 4$. The final output of the model is obtained by applying a global average pooling layer on the output feature map of Stage 4, followed by an MLP head. For training, the Binary Cross Entropy (BCE) loss was used for the binary classification task, and the Mean Squared Error (MSE) loss was used for regression tasks. For the ABCD dataset, input training images were randomly augmented to prevent the model from overfitting. Random augmentations include affine transformation, Gaussian noise, and Gaussian smoothing. The same augmentations were applied for contrastive pre-training in section 2.4.3.

Due to memory constraints, instead of inputting the entire fMRI volume of a subject, we divided the volume into 20-frame sub-sequences and used them as the input. Each sub-sequence was treated as a data point for training, meaning the appropriate loss function was calculated and backpropagated for each sub-sequence. For inference, the logits from the sub-sequences of each particular subject were averaged, yielding a single output for each subject.

Computational complexity The computational complexities of a single global attention Transformer block (denoted as MSA & MLP) and a 4D Swin Transformer block (denoted as W-MSA & MLP) for input with a dimension of $T \times H' \times W' \times D' \times C'$ can be calculated as

$$\Omega(\text{MSA \& MLP}) = 12NC'^2 + 2N^2C' \quad \Omega(\text{W-MSA \& MLP}) = 12NC'^2 + 2PM^3NC',$$

where the number of tokens $N = TH'W'D'$. In practice, on Stage 1 of SwiFT, setting the values used for the experiments $C' = 36, T = 20, H' = W' = D' = 16, P = M = 4$, the two terms $12NC'^2$ and $2PM^3NC'$ are balanced with the second term only being 1.19 times the first term. Compared to this, with global attention the $2N^2C'$ term becomes 379 times larger than the $12NC'^2$ term, taking up most of the computation budget and creating a bottleneck. For successive stages, N is reduced by a factor of 8, and C' is increased by a factor of 2, resulting in Stage 1 being the most computationally expensive.

Baselines

ROI-based models We used ROI-based deep learning methods as baseline models, which are listed as BrainNetCNN [66], VanillaTF [52], and Brain Network Transformer (BNT) [52]. These models utilize functional connectivity data as input, which is computed using temporal correlations (Pearson correlation) of every pair of brain regions. To evaluate such methods, we followed the hyper-parameter and implementations of these three models from [52]. In addition, we also employed XGBoost (eXtreme Gradient Boosting) [95] in conjunction with the features described in [53] to compare a traditional machine learning model with that of deep learning-based models. We used the flattened upper triangular correlation matrix as the input for XGBoost.

TFF The Transformer Framework for fMRI (TFF) [50] consists of 3D CNNs to reduce the dimensionality of fMRI volumes, which are then passed to a transformer encoder layer. It has been reported that the model achieves SOTA (State-Of-The-Art) performances in sex classification and age regression in HCP datasets compared to other deep neural networks [40, 96]. The original model requires two reconstruction-based pre-training steps to stabilize training and enhance performance in downstream tasks. In this study, we adopted the original architecture and added the following strategies to

Table 2.1: Performance comparison to baselines on classification and regression tasks

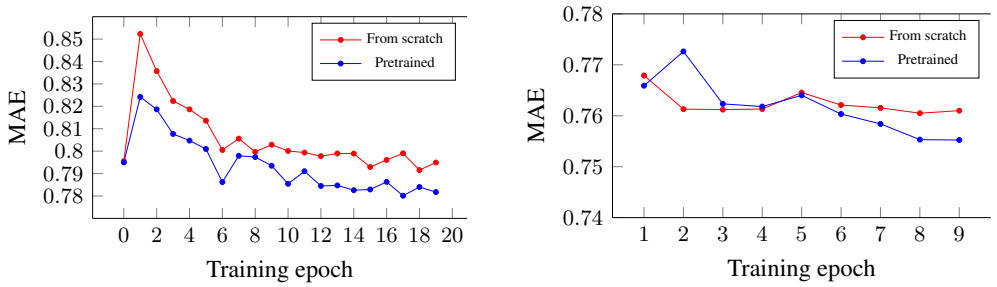
Method	ABCD				HCP						UKB					
	Sex		Intelligence		Sex		Age (year)		Intelligence		Sex		Age (year)		Intelligence	
	ACC	AUC	MSE	MAE	ACC	AUC	MSE	MAE	MSE	MAE	ACC	AUC	MSE	MAE	MSE	MAE
XGBoost	69.5	76.7	0.977	0.770	68.5	75.5	14.3	3.12	0.991	0.813	79.5	87.6	48.8	5.85	1.055	0.816
BrainNetCNN[66]	**80.1**	87.9	0.969	0.767	77.1	84.9	12.6	2.97	0.984	0.805	86.8	93.8	42.7	5.36	1.001	0.800
VanillaTF[52]	77.4	85.1	0.961	0.764	77.9	85.2	12.5	2.95	0.987	0.812	87.0	95.1	41.4	5.26	0.999	0.799
BNT[52]	79.1	**88.9**	0.955	0.767	81.0	88.0	12.8	2.98	1.001	0.830	87.0	94.8	39.6	5.17	0.998	0.798
TFF[50]	73.8	80.2	0.968	0.768	92.5	97.5	13.8	3.11	0.953	0.795	96.8	99.5	42.1	5.10	0.997	**0.783**
SwiFT (ours)	79.3	87.8	**0.932**	**0.756**	**92.9**	**98.0**	**8.6**	**2.36**	**0.903**	**0.786**	**97.7**	**99.8**	**18.2**	**3.40**	**0.992**	0.796

enhance the efficiency and prediction performance: we added initial CNN layers with $2 \times 2 \times 2$ strides to reduce the intermediate caches, which allows significant reductions in inefficient memory usage and increased batch size in our experiments. Additionally, we used learning techniques such as automatic mixed precisions, gradient accumulation, and Stochastic Gradient Descent with Warm Restarts (SGDR) to stabilize training and enhance prediction performance. Since TFF also accepts the fMRI volume as its input, due to memory constraints, the technique of dividing the volume into 20-frame sub-sequences described in Section 2.4.1 is also implemented. Note that this could potentially lead to a loss of important information during feature extraction because it processes each fMRI frame independently using 3D CNN, which means that temporal features do not collaborate.

2.4.2 Classification and Regression Results

In Table 2.1, we compared the performance of SwiFT against various baselines on sex classification and age, intelligence regression tasks. The 4 ROI-based baselines include XGBoost, BNT [52], BrainNetCNN [66], and VanillaTF [52]. TFF [50] was also included as a Transformer-based baseline with a CNN encoder. The SwiFT model was trained from scratch as in Section 2.4.1.

On the sex classification task, SwiFT outperforms all of the baseline models on



(a) Results for HCP

(b) Results for ABCD

Figure 2.3: Effect of UKB pre-training evaluated on (A) HCP and (B) ABCD intelligence prediction tasks.

the HCP and UKB dataset while showing competitive results against the best ROI-based models (BrainNetCNN) on the ABCD dataset. On the regression tasks, SwiFT outperforms all baselines, especially for the age prediction tasks, although it is to be noted that all of the models still have a large room for improvement on the UKB intelligence prediction task.

2.4.3 Effects of Pre-training on Downstream Tasks

To demonstrate the effectiveness of contrastive pre-training described in Section 2.3.2, SwiFT pre-trained on a larger dataset (UKB) was fine-tuned on a smaller dataset (HCP), and a comparable-sized dataset (ABCD) for the intelligence prediction task, which has room for improvement compared to sex and age prediction tasks. The model was pre-trained using the combination of the instance contrastive loss function ($\mathcal{L}_{IC}$) and the local-local temporal contrastive loss function ($\mathcal{L}_{LL}$) such that the training objective is to minimize the sum ($\mathcal{L}_{IC}+\mathcal{L}_{LL}$). The feature representations used in the loss calculation were obtained in the same manner as other tasks; by applying a global average pooling layer on the output feature map of Stage 4, followed by an MLP head. Figure 2.3a depicts the average performance of the model for each training epoch during

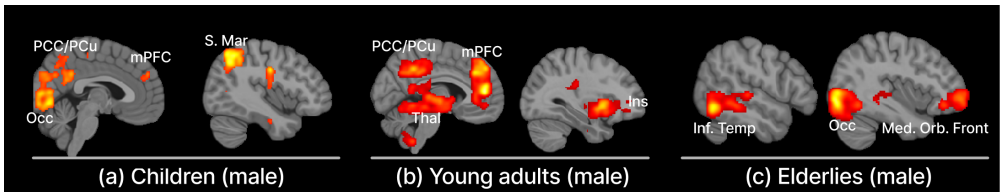


Figure 2.4: Interpretation maps with Integrated Gradients (IG) for sex classification. (Sagittal plane) **(a)** ABCD **(b)** HCP **(c)** UKB

the fine-tuning process. The results from a model trained from scratch (Section 2.4.2) are also shown for comparison. In HCP intelligence prediction, the pre-trained model consistently performs better during the early training stage than the model trained from scratch. In contrast, in ABCD intelligence prediction, the pre-trained model exhibits dramatic drops in performance at the early stage of training and gradually attains a better performance at the later stage of fine-tuning. This initial worse performance might result from the sub-optimal training hyper-parameters for fine-tuning, such as the sub-optimal initial learning rate. The result suggests that contrastive pre-training on a larger dataset shows promise toward improving downstream performance.

2.4.4 Interpretation Results

Using an Integrated gradient with Smoothgrad sQuare (IG-SQ) implemented in Captum framework [55, 97, 98], we identified the brain regions showing high explanatory power on the sex classification task. We acquired 4D IG-SQ maps from test sets and filtered out incorrectly predicted samples. To find the spatial patterns of the brain showing explanatory power across subjects, we normalized the IG-SQ maps, smoothing the maps with a Gaussian filter. Then we averaged the maps over time dimensions and across subjects.

Figure 2.4 shows the brain regions contributing to successful sex classification. Each image was thresholded with a z-value of 1.5 for visualization purposes. The brain

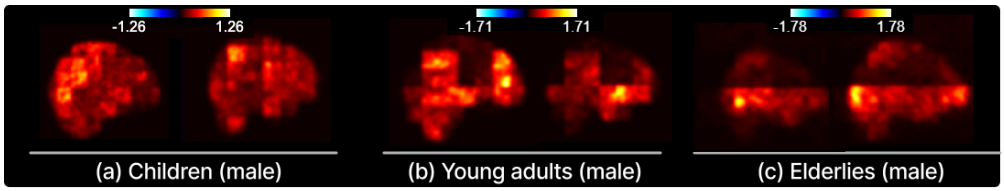


Figure 2.5: Standard deviation over time dimension in Interpretation maps with Integrated Gradients (IG) for sex classification. (The same sagittal plane as Figure 2.4) **(a)** ABCD **(b)** HCP **(c)** UKB

regions include, in children (ABCD), the medial prefrontal gyrus (mPFC), posterior cingulate cortex (PCC), precuneus (PCu), and parietal gyrus (the default mode network). In young adults (HCP), similar brain regions were observed with a broader and higher intensity in mPFC, while showing unique brain regions such as the thalamus and insular cortex. In middle and old-aged adults (UKB), we acquired the highest IG values in the inferior temporal gyrus and medial orbitofrontal cortex. The results confirm those regions implicated in prior brain sex difference literature [99, 100, 101, 102].

To test whether the IG-SQ values are consistent over several time points, we acquired standard deviation over time dimensions within each sub-sequence, acquiring 3D standard deviation maps. The 3D standard deviation maps were averaged across all subjects. Figure 2.5 represents the resulting map, where each voxel intensity means how much variability exists within the voxel over time. We found that each brain region had a different degree of change in sex explainability over time. In particular, we found that the standard deviation was larger in brain regions with higher average intensity in Figure 2.4. We checked the Spearman rank-order correlation coefficients between the two 3D maps used in Figure 2.5 and Figure 2.4. The two maps were significantly and highly correlated in ABCD (r = 0.8448, $p < 0.001$), HCP (r = 0.9982, $p < 0.001$), and UKB (r = 0.9994, $p < 0.001$). These correlations suggest that the brain regions with the higher explanatory power for predicting gender on average also have

Table 2.2: Efficiency of 4D fMRI Transformers

Method	# Param.	FLOPs	Throughput (samples/sec)
TFF	729M	40.72G	53.60
SwiFT	4.64M	2.62G	104.46

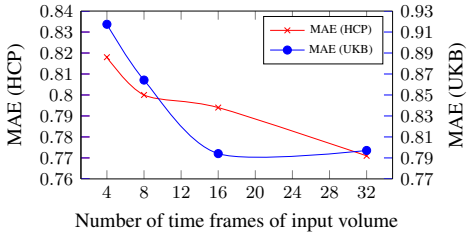
a higher variation in explanatory power over time.

2.4.5 Model Efficiency

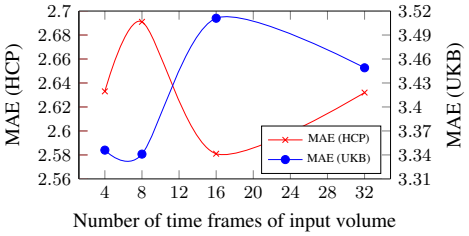
In Table 2.2, we compared the efficiency of SwiFT against TFF [50], the previous 4D fMRI Transformer comparable to SwiFT in their computational costs. Dummy data with random numbers were used for the experiment. Each sample consists of 20 volumes with the shape of $96 \times 96 \times 96$. Floating point operations (FLOPs) were used to estimate the multiply-add computations required to process the 4D fMRI volumes. The throughput (samples/s) denotes the number of 20-volume samples processed per second, calculated using a single NVIDIA A100 GPU. For an accurate measurement, the throughput was measured using synchronized timing with an initial GPU warmup step and was repeated 100 times. The results show that SwiFT has 158.4 times fewer parameters, requires 15.5 times fewer multi-add operations, and processes input data 1.94 times faster than TFF. From here, it can be seen that SwiFT is much more efficient than TFF while also attaining better performance as in Table 2.1.

2.4.6 Effect of Input Sequence Length and Time Window Analysis

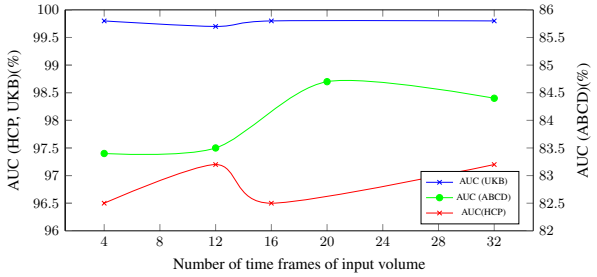
To justify using a 4D model, we investigated the effect of the time sequence length (number of input fMRI volumes). We kept the model architecture and hyper-parameters constant, such as the local window size, to ensure a fair comparison. We adjusted the



(a) Results for intelligence prediction.



(b) Results for age prediction.



(c) Results for sex prediction.

Figure 2.6: Effect of the number of time frames of the input fMRI volume on (A) intelligence (HCP, UKB), (B) age (HCP, UKB), and (C) intelligence prediction tasks (HCP, UKB, ABCD).

mini-batch size to keep the total number of training iterations per epoch constant. We note that the training data augmentation on the ABCD dataset was not applied in this experiment to keep the training environment consistent compared to other datasets, and thus the model has a lower performance compared to the results posted in 2.4.2.

In Figure 2.6, we compared the model performance with a varying number of input fMRI volumes (4, 8, 16, 32 time frames) on intelligence, age, and sex prediction tasks, respectively. All the performances in the figure are averaged performances from three repetitions. In the intelligence prediction task–which is a challenging task considering the MAE of about 0.8 (z-score) was only slightly better than the variance of one–the longer sequence lengths (16 and 32) of fMRI led to better results in both young adults (HCP) and elders (UKB). In age prediction too, in young adults (HCP), we also found

the positive effect of longer sequence length on predictive performance, observing peak performance at 16 time frames. However, in predicting the age of elders (UKB), the longer sequence lengths (16, 32) resulted in poorer performance in age prediction. Namely, the findings of the former three cases showed that the longer sequence lengths enabled better learning of temporal dynamics needed to predict intelligence in young adults and elders and age in young adults. But the last case of the age prediction task in elders showed that the benefit might not be generalizable to all the cases. The performance of the sex classification task on the HCP and UKB dataset is already saturated near 100% AUC, and thus we observed that changing the number of input time frames has a small or negligible effect on the performance, ranging within one standard deviation. In contrast, the performance peaked at around 20 input frames for the ABCD dataset, the default number of input volumes used for the other experiments.

SwiFT model was trained by sequentially processing individual 20-frame subsequences of fMRI data to encompass the entire fMRI dataset. For inference purposes, the predictions obtained from the fMRI sequences were aggregated by averaging the logit values of each subject. This averaged value was then utilized as the final prediction for the respective subject. To ensure that the predictions of time windows from the same subjects are homogeneous and a few noisy time windows do not decide the final predictions, we verified how many subjects exhibited distinct predictions among their time windows using the sex classification task of ABCD, HCP, and UKB datasets. Each ABCD, HCP, and UKB subject has 18, 60, and 24 time windows.

As seen in the histogram in Figure 2.7, over 90% of subjects showed identical predictions among the time windows (0.992 for UKB, 0.907 for HCP, and 0.927 for ABCD). Of note, 1.0 in the x-axis denotes that the predictions from the time windows of the subjects are perfectly correct, while 0.0 means none of the time windows exhibited correct predictions. This suggests that the predictions of time windows are homogeneous and that the final predictions of each subject are not biased toward a few

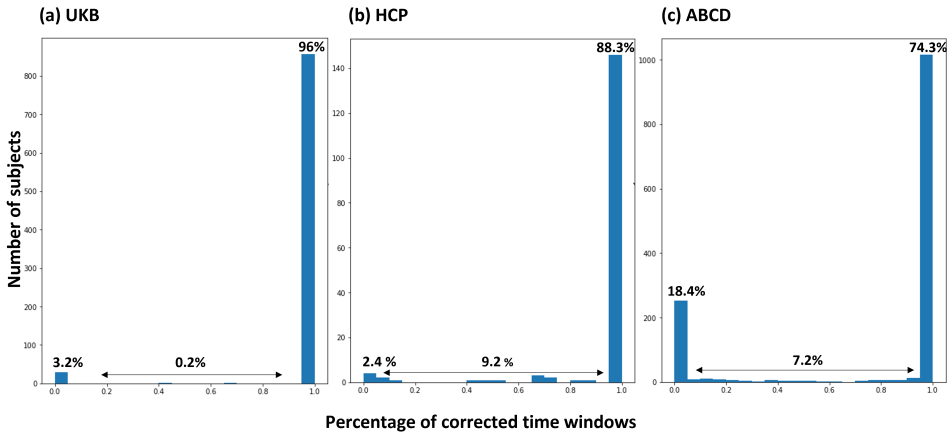


Figure 2.7: Inner-subject accuracy of sex classification.

time windows.

2.5 Discussion

Investigating the spatiotemporal dynamics of the human brain poses a formidable challenge owing to the lack of powerful analytics permitting individual-level prediction of cognitive or clinical outcomes such as psychiatric or neurological diseases. Thus, we present an efficient Transformer model designed for high-dimensional 4D brain functional MRI to learn these spatiotemporal dynamics and predict biological and cognitive outcomes. Throughout various tasks, our Transformer models consistently outperform the state-of-art Transformer models for four-dimensional fMRI and feature-based (i.e., functional connectivity) machine learning models commonly used in the domain. SwiFT performs better with significantly less memory and training time than the existing Transformer model for 4D fMRI [50]. In addition, the Integrated Gradient with Smoothgrad shows the feasibility of interpreting the spatial patterns of the functional representations from the Transformer that contribute to a given task. Lastly, we demonstrated the feasibility of self-supervised learning for transfer learning between different

data sources.

In Figure 2.4, we found that explanatory brain regions for classifying sex varied with age. Across all age groups, we observed brain regions associated with the default mode network, such as the medial prefrontal gyrus (mPFC), posterior cingulate cortex (PCC), precuneus (PCu), and parietal gyrus. The default mode network is a set of regions known to be consistently activated when we are at rest and not performing a specific cognitive task. Previous studies have frequently reported pronounced sex-based differences in the activation patterns of these regions [103, 104, 102, 100, 99]. Furthermore, unique regions were observed in young adults (HCP) and middle or old-aged adults (UKB). The thalamus and insular cortex, where various sensory modalities are integrated, were observed in young adults (HCP). In middle or older-aged adults (UKB), the inferior temporal gyrus (ITG) and medial orbitofrontal cortex (mOFC) indicated higher contribution to sex classification, where diverse sensory information is integrated for higher-level visual processing (ITG) and decision-making (mOFC). It matches the previous findings reporting the sex differences in those regions [105, 106, 102]. The observed sex differences in these brain regions suggest that the integrative hubs for functional networks develop and diverge between males and females as the human brain ages.

In Figure 2.5, our observations indicate that the integrated gradient value for sex classification exhibits variation along the time dimensions rather than being uniform. Further, we discovered that the regions providing better explanations for sex prediction demonstrate significant variability over time. Considering that the observed regions tend to be integrative hubs for functional networks, these explanatory power fluctuations for sex may stem from ongoing interactions (functional connections) between these regions. For instance, the connectivity between default mode networks is known to represent sex differences [103]. These findings shed light on the real-time manifestation of individual differences in each brain region at the single volume scale.

Unlike correlation-based approaches that merely yield time averages of brain dynamics, this new approach provides a more comprehensive understanding. Future research should address intriguing questions, such as investigating the correlation between the interpretation maps for sex and the local functional connectivity computed within that window. Moreover, exploring whether this relationship can account for variations in other variables could be crucial for advancing our understanding of brain connectivity and individual differences.

The findings from the window-based analysis in Figure 2.7 suggest that the traditional research paradigm of fMRI, which involves collecting data over multiple trials, might transform. The consistent outcomes observed across sequences indicate that the model does not necessitate the utilization of all resting-state fMRI sequences to infer variables in a given subject; only a small subset may suffice. While conventional functional connectivity-based methods require a large number of time points across multiple trials to enhance the statistical power of a study, these results suggest that with an effectively pre-trained deep neural network, it may be feasible to detect the status of a subject by collecting merely several seconds of fMRI.

Previously, some studies proposed the feasibility of self-supervised learning for transfer learning between different data sources based on multivariate fMRI time series [36], grayordinate fMRI data [47], or 4D fMRI [49, 50]. To evaluate the feasibility of using 4D fMRI to transfer knowledge from large-scale datasets to others, we investigated the effect of transfer learning with self-supervised contrastive learning for fMRI. The pre-training fostered early convergence during fine-tuning on some downstream tasks, such as HCP and ABCD intelligence prediction. However, the advantages of self-supervised pre-training were not evident in other tasks. The predictive performance in sex or age tasks, after fine-tuning, was not significantly better than training the model from scratch, and in some cases, it was even worse. This result can be attributed to the distinct age range between UK Biobank used for pre-training and the

other two datasets for downstream tasks, ABCD and HCP. In addition, the scanner effect of fMRI, which confounds the biological and cognitive features, can also hamper knowledge transfer. Otherwise, the performance increase might be limited because we have already reached the upper bounds for the sex and age prediction tasks.

To enhance the effect of pre-training, we suggest that various datasets should be included during pre-training. By training the model on datasets from multiple sources and learning a shared representation, we anticipate obtaining a more resilient representation that can mitigate noise, such as the scanner effect. Furthermore, optimizing pre-training methodology for fMRI with various sources of datasets are promising future research topic. We intend to employ other widely-used pre-training strategies that train deep neural networks in supervised ways, such as continual learning [107] and multi-task learning [108]. In these methods, establishing suitable training objectives, which prevent catastrophic forgetting during continual training and overcome unstable training during multi-task learning, is important for adopting those techniques to develop foundation models. Lastly, the potential of the 'pre-train and transfer' approach is expected to be even more impactful when applied to small-scale disease data. Researchers have recently been actively recruiting large cohorts datasets such as the Autism Brain Imaging Data Exchange (ABIDE) and the Healthy Brain Network (HBN). However, most studies still compare a limited number of disease groups with a control group. In this study, relatively straightforward variables such as sex, age, and intelligence were predicted in the downstream task. In addition, the downstream task included a substantial number of subjects, where the influence of pre-training may be constrained. If challenging variables such as ASD, ADHD, and depression need to be predicted with a limited amount of data, the effect of transfer learning can be significantly magnified. Thus, exploring the additional effects of transfer learning on such small-scale data will be a prominent research topic in the future. In conclusion, the strong outcomes of this study may stimulate future exploration of scalable spatiotem-

poral learning in computational and clinical neuroscience.

2.6 Limitations

Although our model has demonstrated high performance and efficiency compared to existing models, it still presents certain limitations for neuroscientists aiming to apply it to their specific subjects. The fMRI data utilized in this experiment ranges from a minimum volume of 383 (586 megabytes) to 1200 (1.3 gigabytes) per subject. While SwiFT significantly reduces the number of parameters and enhances computational speed compared to the existing fMRI Transformer, training a model on such data necessitates more than 24 gigabytes of GPU resources and storage space for thousands of fMRI images. This can pose a significant challenge for researchers with limited computing resources. We have demonstrated that the computational cost resulting from data size can be partially mitigated by partitioning the large volume into smaller volumes (patches) using patch partitioning techniques. To enable broader adoption of our model among researchers, it becomes imperative to develop an optimal patch partitioning method that can effectively reduce the computational cost associated with data size while preserving the essential characteristics of fMRI data. Furthermore, it is essential to consider model efficiency aspects, such as gradient compression, to ensure its widespread applicability.

Our study is based on a sliding window approach, and learning is performed on sub-sequences, which only offers a limited description of its capability to handle long-term dynamics. Processing entire fMRI volumes, which can amount to several gigabytes for multiple subjects, is unfeasible considering limitations in GPU resources. When using a sliding window, the model primarily focuses on the local temporal dynamics of the fMRI, which restricts its ability to learn long-term temporal patterns. Figure 2.6 indicates that longer sequences do not necessarily improve the performance

of some tasks. Considering that long fMRIs exhibit both local and global dynamics, the performance decrease with larger numbers of fMRI volumes is counter-intuitive. This raises concerns about the model's ability to handle long-term temporal dynamics. Additionally, comparing performance differences across a range of up to 32 time points is a small interval to observe the effect of sequence length. Previous connectivity-based deep neural networks have compared hundreds of time points, and the performance differences between dozens of time points were not significant [40].

Several studies have examined the minimum sequence length (window size) necessary to capture individual differences. Window-based correlation approaches indicate that excessively long windows smooth out genuine dynamics, while overly short sequences are susceptible to noise and may erroneously focus on spurious connectivity fluctuations [109]. Leonardi and Van De Ville [110] recommend using fMRI recordings of approximately 100 seconds to capture non-stationary fluctuations when using sliding window correlation. Other research provides statistical support for the hypothesis suggesting that a shorter sequence length of 40 seconds is sufficient for sliding window correlation [109]. Additionally, a study employing a graph-based deep neural network for a sex classification task on HCP fMRI data suggests that moderately large window size is optimal for performance, but excessively large window sizes severely degrade performance [40]. These findings align with ours, indicating that providing models with excessively long fMRIs does not necessarily improve performance. The window-based analysis in Figure 2.7 demonstrates that sub-sequences from the same subject are homogeneous in predicting sex, suggesting that longer time points may provide redundant information. If the information within a long fMRI sequence is redundant, increasing the number of model parameters to accommodate the longer sequence becomes wasteful and may lead to overfitting, resulting in poor prediction performance. In the future, it is crucial to explore the relationship between local and global brain dynamics during resting state to validate this possibility. This can be achieved by

enhancing the model architecture to accommodate longer data sequences beyond 32 volumes. Alternatively, increasing the sampling rate of fMRI data can cover wider ranges of fMRI sequences without extending the sequence length. Both approaches may help in comprehending extended temporal patterns in brain activity.

Lastly, there are limitations associated with implementing Integrated Gradient for SwiFT. As depicted in (b) and (c) of Figure 2.5, it appears unnatural that the integrated gradient values are manifested as square-shaped regions. This phenomenon of discontinuous boundaries can be attributed to the utilization of 4D W-MSA and 4D SW-MSA in SwiFT, wherein the input fMRI volume is divided into patches, and attention is applied solely between patches that fall within a certain window. These discontinuous boundaries are an inherent limitation of the current window-based attention mechanism and should be considered when interpreting future results from this model.

Chapter 3

PREDICTING TASK ACTIVATION MAP FROM RESTING-STATE FMRI

3.1 Introduction

The biological and cognitive variables examined in Chapter 2 are stable individual differences that do not change significantly over time. However, the primary objective of this study is to comprehend how these distinct variances in brain dynamics manifest as a neural activity when individuals engage in dynamic interactions with their surroundings. One notable advantage of fMRI is its ability to capture real-time brain activity changes as humans engage in complex environments. Previous studies have examined the relationship between brain dynamics and human behavior by requiring participants to perform multiple tasks in fMRI, identifying brain regions associated with specific functions and behaviors. However, the successful execution of task-based fMRI necessitates participant cooperation, rigorous experimental control, and a laborious fMRI process, limiting its wider application. Multiple studies have demonstrated that the brain dynamics observed during resting-state fMRI can reliably predict an individual's

brain activities while engaging in diverse tasks [2, 111, 112]. These findings highlight the predictive power of resting-state fMRI in capturing the underlying neural processes associated with task performance. Decoding brain activity within a dynamic environment from resting-state fMRI is more challenging than predicting consistent individual differences from fMRI. Successfully predicting individual differences in brain activity during task performance would provide valuable insights into the relationship between brain dynamics and human behavior. From a practical perspective, if these predictions prove to be accurate, it could potentially imply the ability to anticipate human brain activity during a wide range of tasks using only a short resting-state fMRI scan. This would eliminate the need for extensive resources or controlled experiments conducted by experts.

The task of predicting task activity from resting-state fMRI is based on the understanding that the individual differences observed during task performance are not arbitrary but rather grounded in the consistent characteristics of each individual. The prediction of fMRI Task-related changes in the blood-oxygen-level-dependent (BOLD) signal are typically very small, approximately 2% for cognitive tasks, compared to the signal observed when no task is performed [27]. Group comparisons are commonly employed to enhance the signal-to-noise ratio (SNR) and statistical power when investigating brain function or dysfunction concerning specific situations and behaviors [113]. These studies often recruit disease and control groups to identify brain regions that show the most pronounced differences in group means during task performance. In this context, individual variability is viewed as an inconsistent and fluctuating factor, encompassing aspects like arbitrary task strategy, and is considered a form of noise that needs to be controlled through experimental manipulation [2]. On the contrary, recent studies have argued that individual differences in brain activation while performing the same task are not just noise but inherent to the brain and are related to multiple cognitive and behavioral functions. For instance, research has demonstrated

that brain activation measured during cognitively demanding tasks can effectively predict individual intelligence, surpassing the predictive power of resting-state functional connectivity [114, 115, 116]. This suggests that individual differences in task performance hold meaningful information and can be directly exploited for greater clinical value.

Similarities between task-related activation and resting-state functional connectivity have been proposed in several studies [117, 118, 111, 119]. Based on the same functional architecture of resting state functional connectivity and task activation, several studies have confirmed that the task-state brain activation map of unseen subjects can be predicted from task-independent connectivity at rest using a generalized linear model [2, 120, 121, 116, 122]. Tavor et al. [2] conducted an influential study that revealed the potential of resting-state functional connectivity in accurately predicting task activation in diverse domains, such as language, relational processing, and working memory. Despite some performance variations within and across tasks, it is remarkable that resting-state fMRI combined with a simple generalized linear model can accurately forecast brain activity across different tasks. These findings suggest that subtle individual differences observed during tasks are not mere noise but reflect consistent personal traits, indicating the presence of a unique individual brain fingerprint that can be extracted from resting-state fMRI data. Building upon Tavor's pioneering work, recent studies have further advanced task activation prediction by leveraging advanced techniques such as machine learning and deep learning, surpassing the performance of Tavor's GLM-based model (ConnTask) [123, 113, 47].

Predicting task activation maps accurately from resting-state fMRI offers numerous practical advantages. Beyond predicting task activation maps, these activation maps can be used for many clinical purposes. The predicted task activation maps are more informative of the biological and cognitive variables than resting-state functional connectivity [116, 120, 124]. In a study that predicted brain activity during working

memory tasks using multi-site fMRI data, the predicted task activation map exhibited even better performance than the actual task activation map in predicting intelligence scores [120]. This suggests that, as the activation maps are obtained from task-independent states, the maps may be more resilient against confounding factors like scanner effects, head motion, and arbitrary task strategy than actual activation maps. As a result, they enable a more precise capture of genuine brain activity. Furthermore, these models can be utilized to generate task activation maps in people who have difficulty undergoing task fMRI [125, 126]. This approach enables researchers to overcome the limitations posed by difficulty in obtaining task-based fMRI data and extends the applicability of resting-state fMRI to a wider range of individuals.

However, existing studies have several limitations. First, the question remains whether the previous input features of resting state fMRI are optimal for this prediction. The previous studies mainly use grayordinate fMRI data from the Human Connectome Project (HCP) to focus on and analyze cortical regions [121, 120, 2, 113, 123]. The grayordinate data represents the cortical surface based on the functional structure of cortical gyri and sulci. The advantage of grayordinate fMRI data is that a sparse and comparable representation can be extracted by projecting the fMRI signals of multiple people onto a high spatial and temporal resolution map. To effectively process such surface data, surface-based deep learning models have been developed, showing new possibilities for fMRI research[127, 128, 47]. However, the loss of information may accompany the projection of three-dimensional brain images into one dimension. Furthermore, while the grayordinate fMRI has been used to extract functional connectivity using group ICA and dual regression [129, 130], the optimal preprocessing processes or the number of features for predicting task activation maps from resting state fMRI has not yet been systematically analyzed, which requires more thorough analysis for the best practice [125, 123]. In addition, ConnTask, a widely used glm-based method for the task, predicts activities in each region separately with several

linear models based on predetermined parcellation and combines the predicted maps from each model, limiting the transferability between the brain regions. To accurately predict the subtle variations observed in individuals during task performance, it is crucial to understand the interconnectedness of all voxels rather than considering each region of resting-state fMRI in isolation. This integrated approach allows for a comprehensive exploration of the relationships between different brain regions, enabling a more accurate prediction of task activity.

In this study, we propose **Swi**n **f**MRI Transformer with **UN**ET (SwiFUN), a 4D fMRI Transformer that can generate better-quality task-activation maps using 4D resting-state fMRI. This is the first work to apply Transformer to generate 3D task activation maps from 4D fMRI data. Adapting the basic architecture of Swin UNEt Transformer (Swin UNETR) [1], which is a variant of SwiFT in Chapter 2 and has a UNET-based decoder, we show that the end-to-end learning capability of 4D fMRI Transformer unlocks its potential to predict a high-resolution task-related brain activity by capturing complex spatiotemporal patterns in 4D resting-state fMRI. During the brain activation map prediction for an emotional matching task in the UK Biobank (UKB) study [63, 64], we found that the overall concordance, representing the average level of similarity between the predicted and actual maps, surpassed that of the conventional method. Additionally, to identify the subtle individual variations in each person's task activation map, we utilized contrastive learning to maximize the difference between predicted task activation maps, adapting previously proposed reconstruction-contrastive loss [47]. Our experiments show that learning a richer representation from resting-state fMRI may better predict human brain activity associated with specific tasks, suggesting that this approach could be a promising way to create realistic task activation maps.

3.2 Method

3.2.1 Experimental Setting

Task definition

We examine task activation maps, three-dimensional volumes representing the active regions during a specific task. While previous studies often projected volumes into surface space (CIFTI) [61] due to resource limitations and fMRI characteristics, our study predicts the activity of whole-brain activation maps in volumetric form. However, to facilitate comparison with existing baseline models, we masked some regions using a template atlas image to restrict the analysis to the comparable regions as the Conntask, a glm-based model proposed by Tavor et al. [2], which requires parcellation for the prediction (refer to 3.2). Specifically, we employed 100 cortical parcels, each assigned to one of the seven brain networks provided by Schaefer et al. [131]. As a result, 132,032 valid voxels were selected for analysis. For the baseline models that predict one-dimensional task activation maps, we excluded the masked voxels from the volume and flattened the remaining 132,032 voxels for further analysis.

Datasets

UK Biobank data UK Biobank (UKB) is a large biomedical database that contains health-related information from half a million UK participants. To evaluate the model's ability to generate task activation maps, we ran the analysis on the preprocessed resting-state and task-state fMRI of 7,038 individuals (age $= 54.971 \pm 7.53$ years, 52.7% female) from UK Biobank release 2. Resting-state fMRI has a $2.4 \times 2.4 \times 2.4mm$ resolution, TR of 0.735 s, and 6 minutes (490 time points). The initial preprocessing includes motion correction, group-mean intensity normalization, high-pass temporal filtering, and EPI warping. Structured artifacts were further removed

with ICA+FIX cleaning [132, 133]. The resting-state fMRI was then registered to standard MNI space [134]. The detailed acquisition protocol and preprocessing process are described in [63] and [64].

The task-state fMRI has a similar acquisition protocol and preprocessing processes as resting-state fMRI, except that it has a shorter duration of 4 minutes without ICA+FIX cleansing. Hariri faces/shapes "emotion" task was executed from UK Biobank, which has a relatively shorter overall duration and fewer total stimulus block repeats than those used in the Human Connectome Project (HCP) [135, 136]. During the experiment, participants are sequentially presented with faces or shapes in each block of trials. In the face-matching task, participants are instructed to indicate which of the two faces at the bottom of the screen matches the face at the top. The faces used in the experiment express either anger or fear. In contrast, the second task involves identifying the matching shape from two options displayed at the bottom of the screen, with the shape to be matched shown at the top of the screen. This experiment aims to discover brain regions related to emotion and face processing, focusing on the amygdala which reveals the preferential response to emotional stimuli, such as the angry face. To evaluate the performances, we used three contrasts derived from the emotion-matching experiments; Contrast 1 (Shapes), 2 (Faces), and 5 (Faces-Shapes). Investigating amygdala activation in the last contrast (Faces-Shapes) is particularly important.

Feature extraction for connectivity-based task activation (connTask) maps Previous methods have considered resting-state functional modes useful for predicting task activation maps. Functional modes refer to consistent spatial patterns of brain activity observed across different individuals. Group-Independent Component Analysis (ICA) is widely used to find the group-level functional modes from several resting-state fMRI. The spatial maps from group-level ICA are often considered data-driven

parcellation that separates fMRI data into independent components (IC), each representing distinct brain networks involved in various cognitive processes or functional systems. UK Biobank executed the group ICA and provided two versions of group-level ICs (25 and 100) (refer to [63] for the detailed process). We downloaded the group-level parcellations from the URL: `http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=9028`. We filtered out components thought to be artefactual from the initial 25 and 100 group-ICA components, leaving 21 and 55 independent components for the dual regression.

We executed dual regression using the spatial group IC maps as templates to derive functional modes for each subject. Before the dual regression, we masked the group IC maps and 4D fMRI with a whole-brain mask to exclude unnecessary background values and flattened the volumes, leaving one-dimensional vectors including valid voxels (237,969). The dual regression consisted of two steps. The fMRI data were regressed onto these spatial IC maps in the first step to estimate the subject-specific time courses associated with each IC component. This step aims to extract the network-specific signal for each individual. In the second step, the previous subject-specific time courses were regressed into the previous fMRI data, creating individual network-specific spatial IC maps. The individual network-specific spatial IC maps were then used for weighted seed-to-voxel analysis. The individual IC maps were used to regress against the individual resting-state fMRI time series, resulting in a single time series for each spatial map. Subsequently, each time series was correlated with the original fMRI data to generate connectivity maps for each spatial IC. The resulting connectivity features have dimensions of voxels by the number of independent components (ICs).

Metrics

In this study, we evaluated the predictive performance of task activation maps by assessing the overall agreement of the predicted maps with the true maps and how well the predicted maps identify individual variability. Since there is a high degree of agreement between activation maps of subjects performing the same task, the overall agreement can be significantly improved by simply predicting the group mean of the actual task activation map. This can pose a challenge when attempting to incorporate subtle distinctions in the task activation map among individuals performing the identical task. Therefore, we utilized individual identification metrics to verify that the model was not just predicting the group average.

We employed the diagonal median based on the Pearson correlation to assess the similarity between the predicted and actual activation maps. An N by N matrix represented the pairwise correlation between N individuals' actual activation maps and N individuals' predicted activation maps. The mean and median of the diagonal elements were used to evaluate overall concordance, providing an overview of how well the predicted task activation maps correlated with the actual subject's maps. Additionally, Mean Square Error, a training objective, was used as an additional metric to measure overall concordance.

To compare how well the predicted activation maps represent individual differences, we utilized several metrics based on Pearson correlation: the diagonality index, top-1 accuracy, and diagonal percentile mean. The diagonality index is calculated by averaging the off-diagonal elements within the N by N correlation matrix and subtracting it from the previously obtained diagonal mean. While this metric effectively captures individual differences, it may be sensitive to outliers in the off-diagonal elements. Top-1 accuracy measures the percentage of cases where the predicted map for a specific subject exhibits the highest correlation with the subject's actual map among

all predicted maps. This metric evaluates how many predicted maps are optimal for each subject. The top-1 accuracy metric only considers the case where the predicted map exhibits the highest correlation in a binary manner. As a result, it does not consider the cases where the predicted map shows a higher correlation with the actual map than most predicted maps but does not exhibit the highest correlation. To address this limitation, we also utilized the diagonal percentile mean, which averages the percentile of correlation between the predicted maps and the actual map compared to other predicted maps for each subject. If the predicted map for a specific subject demonstrates lower agreement with the actual map compared to other predicted maps, the diagonal percentile would be closer to 0.5. Conversely, if the predicted activation map for a subject is the most similar to the real activation map of the subject compared to other predicted maps, the diagonal percentile for the subject would approach a value of 1. Furthermore, we conducted a Kolmogorov-Smirnov (K-S) test to determine whether there is a significant distributional difference between the cumulative density functions (CDF) of off-diagonal and diagonal correlations. These metrics comprehensively evaluate the predicted activation maps regarding overall concordance and individual identification.

The metrics used for evaluation are as follows:

Overall concordance

- *Diagonal Mean*
- *Diagonal Median*
- *Mean Square Error*

Individual Identification

- *Diagonality index*
- *Top-1 accuracy*
- *Diagonal percentile mean*
- *K-S test statistics*

3.2.2 Swin fMRI UNetr (SwiFUN)

While SwiFT has shown strong performance in processing resting-state fMRI data, it inherently follows a classification model structure. To generate a 3D activation map that captures subtle individual differences, a decoder-like structure such as U-Net [137, 138] is necessary. Additionally, the model needs to have a deep layer structure to generate high-dimensional images. However, the 4D SW-MSA used in SwiFT has limitations for such tasks. Therefore, to address these challenges, we propose Swin fMRI UNet Transformer (SwiFUN), capable of generating task activation maps by processing the spatiotemporal dynamics of 4D fMRI data. SwiFUN is based on the architecture of Swin UNet TRansformer (UNETR) model proposed for brain tumor segmentation tasks in 3D structural MRI. SwiFUN takes multiple time points of fMRI volumes as input to predict a single 3D task activation map. We implemented based on SwinUNETR module provided by MONAI framework [1].

As shown in Figure 3.1, the intermediate outputs of each Swin Transformer layer are fed into the UNET decoder through skip connections. This UNET structure enhances training stability and facilitates the generation of higher-resolution image information. Structurally, SwiFUN differs from SwiFT in Figure 2.1 regarding how it handles temporal information. While SwiFT incorporates temporal order information using 4D shifted window multi-head self-attention and temporal positional embedding, SwiFUN utilizes the temporal axis as a channel. This approach of integrating temporal information from the initial layers, rather than averaging output activations along the temporal axis, facilitates the generation of a single task activation map by considering all the relevant information from resting-state fMRI.

We trained SwiFUN with an AdamW optimizer and Cosine Annealing Warmup Restart scheduler for ten epochs. The model was trained to minimize mean squared error (MSE) or Reconstruction-Contrastive loss. Due to memory constraints, utilizing

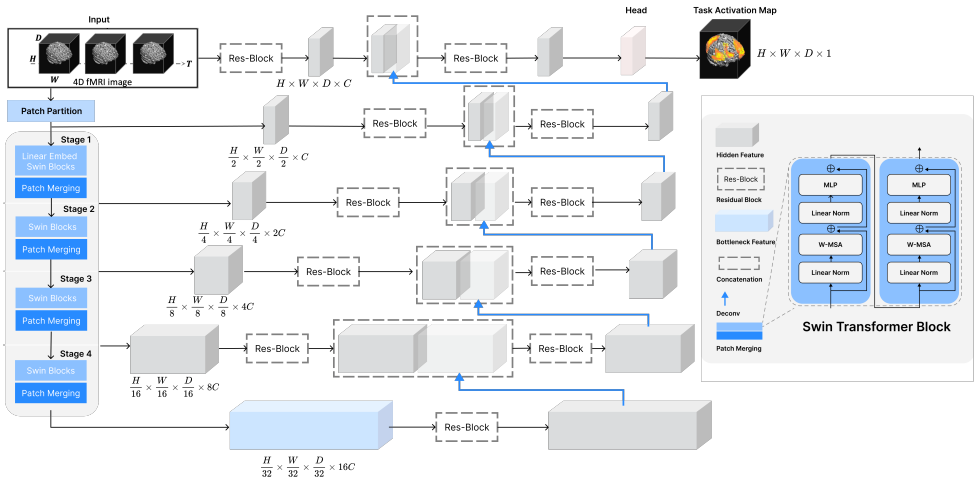


Figure 3.1: Overall architecture of SwiFun. Unlike SwiFT, the time dimension (T) is considered as channel dimension at the first stage. Figure is adapted from Swin UNETR [1].

the entire sequence simultaneously is not feasible. To accommodate the memory limitations, we partitioned the 490 volumes of each subject into sub-sequences consisting of 30 volumes. In addition, we used a mini-batch size of 4 during the experiments (refer to A.2 to find the effect of sequence length and mini-batch size on the performances) Subsequently, we trained the model to predict the task activation map for each subject based on these sub-sequences. During inference for a specific subject's activation map, we calculated task activation maps for each sub-sequence and then averaged them to determine the subject's final task activation map. Furthermore, we conducted an evaluation to assess whether the performance varied depending on the length of the sequence.

3.2.3 Reconstruction-Contrastive loss

This loss is a contrastive loss proposed by Ngo et al [47] for predicting grayordinate contrast images. This loss makes the predicted map and the actual map of a particular subject similar, while the predicted map and another person's map are far apart. This loss is developed to complement the trade-off between two contrasting objectives, overall concordance and individual identification. In A.5, the observed trade-off between overall concordance (diagonal mean) and individual identification (diagonality index) during the training of SwiFUN with MSE Loss supports for the utilization of RC loss. This trade-off highlights the need to incorporate the RC loss to balance capturing group-level patterns and preserving individual differences. Specifically, we intend to maximize the predicted task activation map's specificity without decreasing overall concordance.

Our loss differs from the loss proposed in the previous studies in several ways. In the previous study, the contrastive loss (L_C) was limited to accommodating only two samples. In contrast, our study extends the contrastive loss to involve multiple subjects, enabling the computation of pair-wise mean square errors. Unlike previous approaches that utilized features extracted from one resting state fMRI per subject, we utilize sub-sequences of 4D fMRI data as input in this study. To treat sub-sequences from the same subjects differently from those from different subjects, we excluded sub-sequences cropped from the same subject from the L_C calculation. Moreover, our method differs from previous approaches in that we no longer rely on a two-step training process, where the same-subject error L_R is first trained to converge, and then L_C is applied from a certain point onwards. Instead, we introduced a parameter λ that allows us to consider the relative weight of the two loss terms. These changes facilitate end-to-end training of our model, resulting in improved efficiency and performance.

$$L_R = \frac{1}{N}\sum_{i=0}^{n} d(\hat{x}_i, x_i), L_C = \frac{1}{N^2 - N}\sum_{x_j \in B_i, j \neq i} d(\hat{x}_j, x_i)$$

$$L_{RC} = \lambda L_R - (1 - \lambda) L_C$$

The reconstructive-contrastive loss L_{RC} is defined as follows: Given a mini-batch of N samples B, where each sample x_i represents the target 3D task activation image of subject i, and $\hat{x}$ represents the corresponding prediction. $N^2 - N$ in L_C loss denotes the number of all possible pairs between predicted maps and actual maps from different samples in a batch. $d()$ denotes the distance function, mean square error (MSE).

3.2.4 Baseline

ConnTask

Several previous studies have utilized GLM-based ConnTask to predict task activation maps from resting state functional modes [2, 121, 120]. While they mainly utilized 1D grayordinate fMRI data by projecting 4D fMRI data into 1D grayordinate space (CIFTI), in this study, the volume data was masked and flattened to be a vector for a fair comparison with SwiFUN, which utilizes volumetric fMRI data in voxel space. As seen in 3.2, one-hundred generalized linear models, corresponding to 100 cortical parcels in Schaefer's atlas [131], are trained to predict task activation maps from connectivity features (Independent Components). Each region of interest (ROI) in the task activation map is predicted from the connectivity features in the corresponding ROI. Note that only independent components are used as input features, considering voxels in connectivity features as independent training samples. After β_k is trained, β_k is averaged over all subjects during inference. Five-fold Cross-validation was executed to iteratively train the models with 80% of subjects and predict the rest 20% of the task activation map. We validated the effect of training samples and the number of independent components on the predictive performances, changing the number of samples

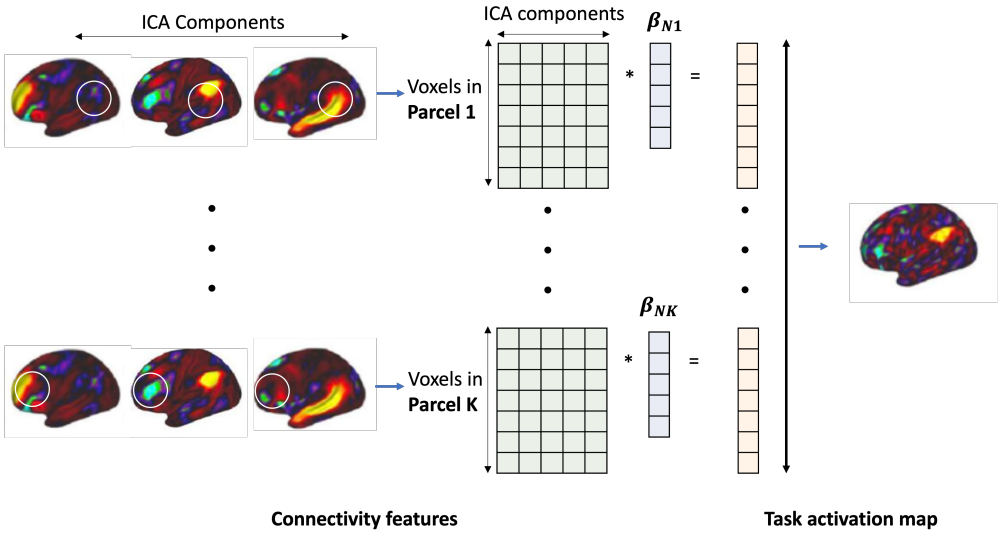


Figure 3.2: ConnTask Pipeline proposed by Tavor et al. [2]. N represents the number of training subjects, and K denotes the number of brain regions in an atlas image. After the $\beta_k s$ are acquired for each training subject, they are averaged over all training subjects to predict unseen subjects during inference.

(100, 500, and 1000) and independent components (21, 55). Due to the occurrence of memory errors when using a larger number of samples, we maximized the utilization of available samples within the available resources(512 GB of DDR4 memory). The through experiments with varying hyper-parameters can be found in A.6.

Swin fMRI Transformer (SwiFT)

Since SwiFT, developed for outcome prediction using 4D resting-state fMRI in Study 2, lacks a decoder for reconstructing a high-dimensional 3D task activation map, we extended the original SwiFT model by incorporating MLP layers. This allowed us to predict a long one-dimensional task activation map comprising valid voxels (132,032) from the 4D resting-state fMRI data. The hyper-parameters, including the embedding

size and training parameters, remained consistent with those used in Study 2. After obtaining the intermediate activations from SwiFT with dimensions of (batch, output channel, width, height, depth, time), we employed adaptive average pooling to reduce the dimensions in width, height, depth, and time into 1. Subsequently, we utilized MLP layers to project the resulting embeddings with the dimension of (batch, output channel) into a one-dimensional task activation map.

Test-Retest Contrasts

The UKB dataset contains revisited data. From the release2 data, which includes 7,038 subjects used in the experiment, we identified 577 subjects who also revisited in release 3. To assess the test-retest reliability of the contrast maps and the performance of our models in terms of overall correspondence and subject identification, we compared the task activation maps from the first and second visits. This allowed us to calculate overall concordance and individual identification metrics based on the initial and revisited data correlations.

3.3 Result

3.3.1 Performances Comparison

In Figure 3.3, we compared the performance of SwiFUN against two models, ConnTask and SwiFT, on predicting task activation maps of shape, faces, and faces-shapes. After comparing the performance of ConnTasks with different numbers of samples and independent components (ICs), we found that the best overall performance was achieved by utilizing 1000 samples and 55 ICs (refer to A.6 to see the effect of a varying number of samples and IC). The test-retest contrasts from 577 subjects were also presented for comparison. In addition, we compared the performances of SwiFUN with two different kinds of losses, mean square error (MSE) and reconstruction-

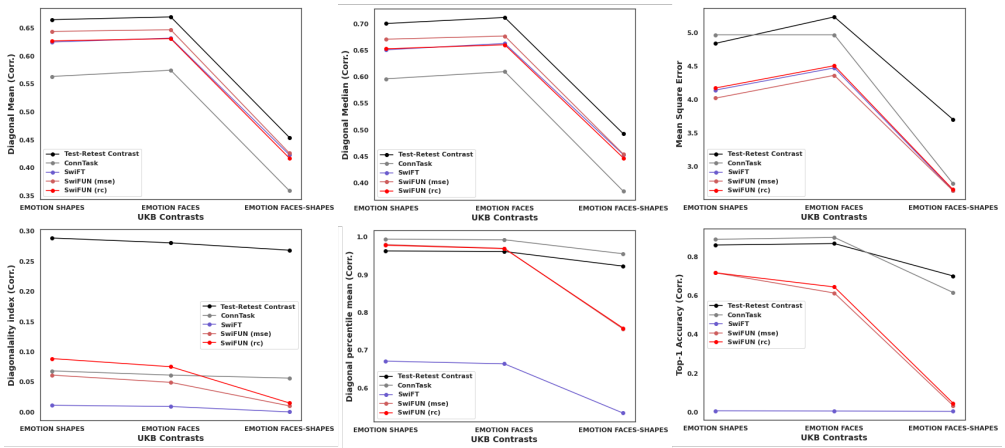


Figure 3.3: Overall performances of SwiFun and its baseline models. Each row represents the overall concordance and individual identification.

contrastive (RC) losses. The weight of the contrastive loss term $(1 - \lambda)$ in RC loss was specified as 0.66 (refer to Figure 3.6 to find the effect of $1 - \lambda$). SwiFT and SwiFUN were trained with the same train, validation, and test dataset with a ratio of (train: validation: test) = $(0.7 : 0.15 : 0.15)$. The number of SwiFT and SwiFUN test subjects amounted to 1057, which was used for comparing the performances. The upper row in Figure 3.3 represents metrics regarding overall concordance, including diagonal mean, median, and mean square error. The lower row in Figure 3.3 indicates metrics related to individual identification, including the diagonality index, diagonal percentile mean, and top-1 accuracy.

Overall Concordance

SwiFUN outperforms ConnTask in all three contrasts and metrics, showing more than 10% increases in diagonality mean and median. Although SwiFUN did not reach the test-retest reliability, SwiFUN with MSE loss showed the highest performance compared to its baseline models. SwiFT showed comparable performances with SwiFUN

with RC loss and slightly worse performances than SwiFUN with MSE loss. The comparison of MSE revealed that SwiFUN with MSE Loss yielded a significantly lower error than the test-retest task activation maps, indicating that the model was well-trained according to the training objective. Minimizing the mean square error was not necessarily aligned with improving other correlation-based metrics.

Individual identification

In the diagonality index, SwiFUN with RC loss showed a higher diagonality index in FACES and SHAPES contrasts compared to ConnTask and SwiFUN with MSE loss. In the FACES-SHAPES task, SwiFUN with RC loss outperformed SwiFUN with MSE loss, but both models exhibited significantly lower performance than the ConnTask. SwiFT exhibited very poor performance in all three contrast maps, revealing the limitations of relying solely on the structure of the classification model for generating high-resolution 3D images. The test-rest reliability of the diagonality index is much higher than predictive models, which means that the activation map from repeated task-state fMRI preserved the individual variability from the initial task activation maps.

In the diagonal percentile mean, SwiFUN with MSE and RC loss performed better than retested task activation maps in FACES and SHAPES contrasts. However, they exhibited lower performances than retested task activation maps in the FACES-SHAPES contrast. ConnTask outperformed the other models in all three contrasts, which shows an overall high diagonal percentile mean. SwiFT showed remarkably lower performances compared to other models in all three contrasts. Similarly, in top-1 accuracy, SwiFUN showed higher performances than SwiFT, but lower performances than retested contrasts and ConnTask.

Overall, SwiFUN with RC loss showed higher performances of individual identification than SwiFUN with MSE loss. While ConnTask showed higher individual identification in top1-accuracy and FACES-SHAPES contrasts, SwiFUN with RC Loss

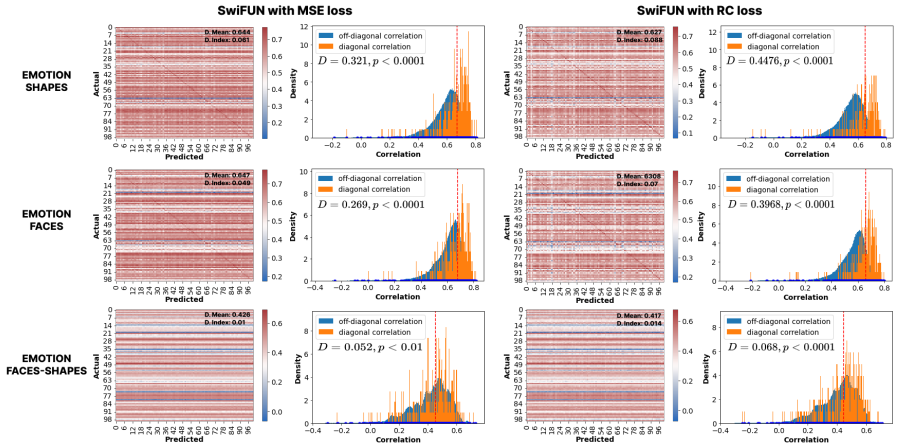


Figure 3.4: The correlation matrix and histograms of diagonal and off-diagonal correlations of SwiFUN with two different losses (left: MSE, right: RC).

showed comparable or higher diagonality index and diagonal percentile mean in FACES and SHAPES contrasts.

Kolmogorov-Smirov Test

We conducted a Kolmogorov-Smirnov test to assess the statistical significance of the disparity between the cumulative distribution functions (CDF) of diagonal and off-diagonal correlations. In Figure 3.4, We examined how the diagonal and off-diagonal correlations differ when training SwiFUN with two different types of loss. The first and third column represents the correlation maps between predicted and actual maps. The diagonal elements of the correlation matrix represent the correlations between predicted and actual maps from the same individuals. The salient trend in diagonal elements represents high prediction specificity. Only 100 of 1057 test subjects were presented for visualization. The second and fourth column shows the histogram of diagonal and off-diagonal correlations with the effect size and p-values for the disparity between the two distribution. The histogram's red dotted line indicates the diagonal median.

The results revealed that regardless of the loss type, all SwiFUN models exhibited a significant distinction between diagonal and off-diagonal correlations in all task contrast maps ($p < 0.0001$). In both FACES and SHAPES tasks, it can be observed that the histogram of diagonal correlations is shifted towards the right compared to the histogram of off-diagonal correlations. However, in the FACES-SHAPES contrast map, the effect size (D) from the Kolmogorov-Smirnov test was smaller, and on the histogram, the distributions of diagonal and off-diagonal correlations overlap, indicating poorer individual identification. Using the RC loss resulted in lower overall correlation values than the MSE loss. However, the diagonal elements became more prominent than the off-diagonal components. This observation suggests a trade-off between overall concordance and individual identification. For all task contrast maps, RC loss led to higher diagonality indexes and larger effect sizes (D) from the Kolmogorov-Smirnov test than the MSE loss.

3.3.2 Prediction of volumetric task activation map

Figure 3.5 displays the top three subjects with the highest correlation between the actual and predicted maps using SwiFUN with MSE loss for each task contrast (SHAPES, FACES, and SHAPES-FACES). While the predictions were made on volume data, the volumetric activation maps were projected onto a surface for visualization. Nilearn, a brain imaging analysis tool, was utilized for this purpose. Each result underwent thresholding for the top 3% activated voxels. The three subjects, '1651120', '1716181', and '1354872', exhibited correlations of 0.806, 0.816, and 0.732 across the tasks. In the SHAPES and FACES contrasts, the activation in the posterior medial cortex and visual cortex observed in the actual maps was also evident in the predicted maps. In the FACES-SHAPE contrast, subtle activations in the prefrontal cortex, superior temporal gyrus, and visual cortex were accurately predicted.

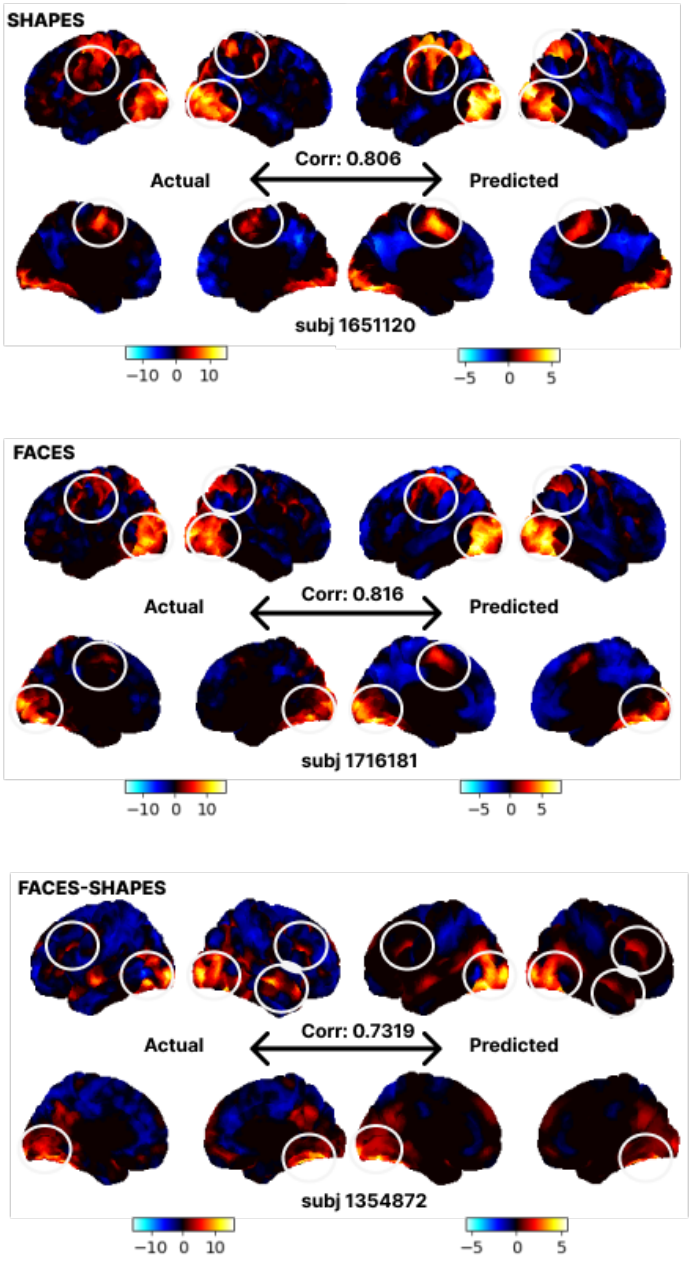


Figure 3.5: The actual and predicted activation maps with the highest correlations from SwiFUN with MSE loss. (lateral and medial view)

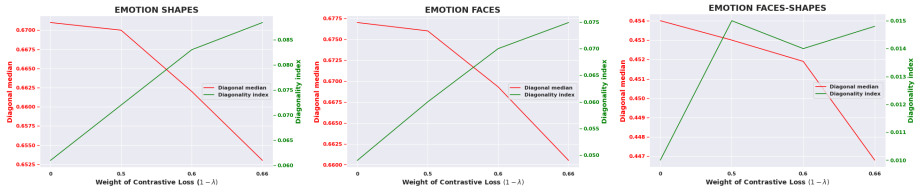


Figure 3.6: The effect of Contrastive Loss on diagonal mean (overall concordance) and diagonality index (individual identification)

3.3.3 Increasing weight of L_C improves individual identification

In Figure 3.6, we investigated how the diagonal median and diagonality index vary with the adjustment of the weight of the contrastive loss in the RC loss. We conducted experiments with three settings for $1 - \lambda$: 0 (L_R only), 0.5 ($L_R : L_C = 1 : 1$), 0.6 ($L_R : L_C = 1 : 1.5$), and 0.66 ($L_R : L_C = 1 : 2$). The results showed that as we increased the weight of the contrastive loss term in all contrast maps, the diagonal median decreased while the diagonality index increased. However, compared with the results of predicting the SHAPES contrast map in Figure A.5, where no contrastive loss term was used, we can observe that the increase in the diagonality index is much more significant compared to the relatively small decrease in the diagonal median. For instance, in the case of predicting the SHAPES contrast in Figure A.5, after the diagonal median converged at 0.671, it decreased by 0.04, while the diagonality index increased by only 0.001. On the other hand, while $1 - \lambda$ increased from 0 to 0.66, the diagonality index significantly improved by 0.027, with a similar decrease in the diagonal median by 0.05. This indicates that by using the RC loss, it is possible to effectively enhance individual identification performance while sacrificing overall concordance to a lesser extent.

3.4 Discussion

Generating various task activation maps from resting-state functional connectivity is a well-established task that has received much attention [2, 111, 116, 120]. The task holds a substantial potential value, particularly for patients or children who encounter challenges performing complicated tasks within fMRI. However, the predictive performance of existing approaches was limited due to their simplicity. To this end, we propose Swin fMRI Transformer with UNet architecture for predicting a 3D task activation map to show that brain dynamics from 4D resting-state fMRI can be trained end-to-end to predict task-state brain activity effectively. SwiFUN outperformed ConnTask on diagonal median and mean absolute error, representing better overall concordance in every contrast. The performance was comparable to the test-retest reliability. While original SwiFUN performed worse than ConnTask on individual identification, adopting reconstruction-contrastive (RC) boosted the performances of SwiFUN over ConnTask in SHAPES and FACES contrasts. This result indicates that 4D fMRI analysis models based on deep learning can effectively predict individual task-related brain activity.

In Figure A.1, we discovered a delicate balance between overall concordance and individual identification. We employed a mean square error as a training objective to refine the model's performance. This objective enhances the diagonal median (representing overall concordance) and the diagonality index (representing individual identification) up to a certain threshold. However, there is a notable reduction in the diagonal median after a few epochs. This phenomenon arises because the model initially learns the distinct characteristics of each task but subsequently makes trade-offs, prioritizing maximal individual variability at the cost of overall concordance. During this phase, the diagonality index tends to be exceedingly low, typically falling below 0.1. Consequently, in metrics concerning individual identification, SwiFUN generally exhibited

lower performance than ConnTask. To mitigate this decline in overall concordance, we introduced a reconstruction-contrastive loss. This loss function ensures that the predicted task activation map remains highly distinctive among subjects through the training objectives. As a result, individual identification performance was enhanced with a minimal drop in overall concordance. Thus, considering the trade-off between overall concordance and individual identification is fundamental to further improve task activation map prediction with deep learning models.

The value of the models trained in this study may be greater when applied to individuals who have difficulty performing tasks within fMRI. Transfer learning can make this possible. A previous study has suggested transfer learning may improve the prediction performance of task activation maps on datasets with small samples [47]. This transfer learning is also valid from healthy to diseased groups [121] and across different sites, MRI vendors, and age groups [112]. Another finding suggests that there are shared characteristics between different tasks. Multiple task activation maps can be predicted from a single model and resting-state fMRI by adopting multi-task learning [47]. Therefore, the next direction of our research is to verify the effect of transfer learning using SwiFUN. Transfer learning is possible for Swin UNETR in segmentation tasks using self-supervised contrastive learning [139]. SwiFUN, which has similar architecture as Swin UNETR, also has the feasibility for transfer learning. Therefore, the model trained on the UKB emotion-matching task may foster generating task activation maps for smaller datasets. This can be validated with the same task from another dataset, HCP.

What information the predicted task activation maps contain is a major future research question. After all, the purpose of predicting task activation maps is to derive clinical value from them that does not exist in resting state fMRI. Therefore, beyond checking how similar the predicted task activation map is to the actual task activation map, we need to evaluate how it relates to other individual difference variables and

whether it has predictive power for other variables. Previous studies have shown that task activation maps predicted by resting-state fMRI better predict cognitive variables such as intelligence [116, 120] than original resting-state fMRI. Accurate predictions for task activation maps led to better intelligence predictions [120]. In addition, Tik et al. [121] demonstrated that the predicted task activation maps of schizophrenia patients have lower overall qualities than those of healthy individuals. These results suggest that the predictive performance of the task activation map itself may be a clue to psychiatric disease. Therefore, our model can be used as a feature extraction method for resting-state fMRI by amplifying individual differences within resting-state fMRI.

Recently, it has been shown that fMRI can reveal more about human cognition and behavior when presented with naturalistic stimuli compared to the resting state. This can be implemented by watching movies in fMRI, which are more real-life situations with complex and dynamic stimuli [140]. Previous studies utilizing movie-watching fMRI data have shown that these data represent unique patterns in functional connectivity compared to resting-state functional connectivity [141]. In addition, this fMRI with naturalistic stimuli exhibits higher predictive performance for individual differences such as intelligence compared to resting-state fMRI, as well as higher performance in generating activation maps for working memory tasks [142, 116, 143]. This may be because environments with richer audiovisual stimuli will likely reveal more individual differences in human thought processes [141, 125]. The enhanced performance in generating task activation maps may be attributed to the similarity between watching a movie and actively performing a task, as opposed to not engaging in any task at all [116]. It would be intriguing to observe whether the impacts of this movie-watching fMRI dataset can be replicated in the proposed 4D fMRI Transformer and to identify the specific areas where it differs from resting-state fMRI. This may give us more insight into what type of fMRI data is best suited to capture brain dynamics.

In conclusion, this study validates, for the first time, how much information can be

extracted from raw resting-state fMRI through various metrics. In addition, we propose training deep neural networks may improve task activation prediction performance in an end-to-end manner. In the future, we anticipate achieving the capability to predict various task-related brain activity from just a few minutes of resting-state fMRI data, significantly reducing the scanning time and effort required to capture task-based fMRI.

3.5 Limitations

While reconstruction-contrastive loss worked for the FACES and SHAPES task activation maps, it was limited in capturing subtle individual differences in a relatively small area, such as the amygdala in FACES-SHAPES contrast. Further analysis of the association between each metric and training objectives is required to identify individuals better. Recently, several research has focused on residualized datasets, subtracting the group-mean activation map from the original task activation maps to leave only individual variability [111, 113]. This initial preprocessing allows the model to focus on the unique properties of each activation map than the commonality of whole activation maps. Thus, the effect of residualization is worth exploring in future work.

Another limitation is that the model's capacity is constrained by its inability to simultaneously address all resting-state fMRI volumes. Given that SwiFUN necessitates significantly deeper layers than classification models (SwiFT), a limitation exists in increasing sequence length. The resting-state fMRI data is segmented into sub-sequences to address this constraint by employing a sliding window approach across the entire fMRI sequence. The task activation predictions from each sub-sequence are subsequently averaged for each subject. However, averaging the 3D task activation maps may potentially obscure subtle individual differences. The reduction in the diagonality index with an increasing batch size supports this hypothesis, as illustrated in Fig-

ure A.2. Thus, it is crucial to determine the distinctive information contained within the task activation maps from each sub-sequence. These maps might show substantial uniformity or notable diversity in their response to changes in resting-state fMRI. Performing such validation can provide invaluable insights into how the changing brain dynamics during the resting state contribute to subtle differences in task-related brain activity.

Chapter 4

CONCLUSIONS AND FUTURE WORK

4.1 Summary

Human brain dynamics is the activity of neurons resulting from human interaction in a changing environment, allowing for adaptive behavior in an ever-changing environment. Moving beyond the traditional methods of analyzing fMRI based on handcrafted features, this study developed an efficient and scalable Transformer that can directly process 4D-shaped fMRI to maximize the learning of spatiotemporal brain dynamics. In Chapter 2, this study showed that brain dynamics extracted from resting-state fMRI using 4D fMRI Transformer could be used to predict individual biological and cognitive variables accurately. In addition, by pre-training the model with self-supervised learning techniques for 4D fMRI, this research found that transfer learning was partially effective in predicting the intelligence scores of ABCD and HCP during fine-tuning. In addition, by utilizing Integrated Gradient, this study confirmed that regions related to the default mode network contribute significantly to sex classifica-

tion across all kinds of datasets and identified data-dependent brain regions integrating sensory information. The biological and cognitive variables studied in Chapter 2 are variables of individual differences that are relatively consistent across individuals. However, the main interest is to uncover how individual differences in these brain dynamics translate into brain activity when humans interact with and act upon their environment. Therefore, in Chapter 3, we adapted the previous 3D Swin Transformer with UNET (Swin UNETR) to predict a 3D task activation map from 4D fMRI. As a result, the study found that models trained directly from 4D fMRI could predict task activation maps more accurately than the traditional glm-based model. The study shows that contrastive learning could amplify the subtle differences within task activation maps. Examining the subjects with the highest prediction accuracy for each task, we verified that the deep learning model could accurately predict activated regions as the actual task activation maps.

4.2 Limitations

The fMRI transformers introduced in this study share a common limitation: they only utilize a portion of the fMRI data within a window. Our fMRI Transformers employ 20 to 30 scans within a sliding window to capture volumetric information. In contrast, existing connectivity-based methods utilize the entire fMRI sequence. This raises questions about whether the model adequately captures long-term temporal dynamics. The results indicate that the optimal cycle for capturing individual differences depends on the specific task. However, due to the relatively short duration of the employed time window (ranging from 4 volumes to 32 volumes), the distinction of performances between the tested windows was quite subtle. To establish the reliability of these findings, future studies should assess model performance across a broader range of sequence lengths.

Furthermore, the methodology employed in this study remains inaccessible to most researchers due to its resource-intensive nature. The scarcity of storage space and computational resources among researchers limits their capacity to accommodate the data encompassing over 15,000 subjects, a pivotal facet of this study. Most researchers will wonder if the model will perform as well when trained on a more limited subset of subjects. Thus, validating the model's efficacy in smaller sample sizes is imperative. This exploration would provide insights into the applicability of deep learning models in functional MRI research and offer guidance to researchers endeavoring to leverage such methodologies.

4.3 Future Directions

This study has demonstrated that by analyzing 4D fMRI with deep learning in an end-to-end learning approach, a richer representation can be obtained compared to existing methods, and higher prediction performance can be achieved in various tasks. The methods proposed in this study can be utilized to solve major problems in psychology and neuroscience. Beyond predicting sex, age, and intelligence scores, which have been validated on various data in this study, the next major challenge will be to predict variables associated with mental disorders. Since many psychiatric disorders are strongly related to sex, age, and cognitive ability, the model is expected to discover individual patterns associated with disease, which can then predict the onset of psychiatric disorders and prevent their progression. In addition, by examining biomarkers strongly related to existing treatments for mental disorders, this approach may provide personalized treatments beyond the existing generic treatments.

Chapter 2 showed that deep learning models for 4D fMRI, combined with explainable AI methods, can identify brain regions associated with specific biological traits at high resolution. It would be valuable to see if this approach could also be used to

help verify brain regions that contributed to the successful prediction of task activation maps in Chapter 3. For instance, in generating the activation maps associated with the emotional matching task, we might expect to see brain regions, such as the amygdala or ventromedial prefrontal cortex (vmPFC), in the resting state associated with the prediction. Previous studies have shown this by representing which independent components from resting-state fMRI had a relatively large impact on task activation map predictions [121]. However, since each independent component covers multiple regions, it is difficult to see how each voxel contributes to the prediction. Using explainable AI methods, evaluating which regions contributed to task activation map predictions for every voxel in 4D fMRI volumes is possible. If it turns out that unexpected regions in resting-state fMRI are explainable for the task activation map, the findings will help us understand the complex interactions between the human brain and behaviors.

4.4 Conclusions

The 4D fMRI transformers presented in this study, along with their promising results, are anticipated to capture the interest of numerous researchers. SwiFT, a transformer model developed within this study, exhibits greater efficiency than existing 4D fMRI transformers while consistently delivering high performance across multiple tasks. Furthermore, SwiFUN shows that the fMRI Transformer can be expanded to investigate complicated human behaviors from brain dynamics. The model might be the first option for researchers studying end-to-end learning on fMRI. Therefore, we believe that this study can guide many people by expanding the scope of existing research and demonstrating the feasibility of such research through multiple experiments.

Appendix A

A.1 Performance Comparison with Standard Deviation

Here, we detail the standard deviation of the results posted in Table 2.1 (manuscript), which compares the performance of SwiFT against other baseline models on various downstream tasks. All of the experiments were performed using three pre-determined random data splits for each dataset, which was shared across all models for a fair comparison. The following Tables (A.1, A.2, A.3) show the performance posted in Table 2.1 (manuscript) along with the standard deviation among the three splits. From the tables, we can conclude that SwiFT outperforms all baseline models above the margin of variability at the following tasks: ABCD intelligence, HCP sex, HCP age, HCP intelligence, UKB sex, and UKB age prediction tasks. Note that the age prediction task was not carried out on the ABCD dataset due to the narrow age range of 9 to 11 years, making it hard to obtain meaningful results from the experiments.

Table A.1: Performance of various models on the ABCD dataset with standard deviation

Method	Dataset: ABCD			
	Sex		Intelligence	
	ACC	AUC	MSE	MAE
XGBoost	$69.5_{\pm 0.59}$	$76.7_{\pm 0.86}$	$0.977_{\pm 0.037}$	$0.770_{\pm 0.016}$
BrainNetCNN [66]	$\mathbf{80.1}_{\pm 0.69}$	$87.9_{\pm 0.37}$	$0.969_{\pm 0.042}$	$0.767_{\pm 0.016}$
VanillaTF [52]	$77.4_{\pm 2.47}$	$85.1_{\pm 3.21}$	$0.961_{\pm 0.050}$	$0.764_{\pm 0.025}$
BNT [52]	$79.1_{\pm 0.80}$	$\mathbf{88.9}_{\pm 0.64}$	$0.955_{\pm 0.058}$	$0.767_{\pm 0.025}$
TFF [50]	$73.8_{\pm 1.13}$	$80.2_{\pm 1.06}$	$0.968_{\pm 0.024}$	$0.768_{\pm 0.009}$
SwiFT (ours)	$79.3_{\pm 1.29}$	$87.8_{\pm 1.31}$	$\mathbf{0.932}_{\pm 0.017}$	$\mathbf{0.756}_{\pm 0.009}$

A.2 Implementation Details

SwiFT To obtain the results detailed in Section 2.4.2 (manuscript), we trained SwiFT from scratch using the following configuration:

- *Optimizer*: AdamW using a cosine decay learning rate scheduler with a linear warm-up (around 5% of total iterations)
- *Learning rate*: After the warm-up, an initial learning rate of 10^{-5} for classification tasks and 5×10^{-5} for regression tasks on the HCP dataset and 10^{-5} for regression tasks on the ABCD dataset.
- *Mini-batch size*: Mini-batch of size 8
- *Epochs*: 10 epochs of training for the sex classification and intelligence prediction tasks, and a maximum of 30 epochs for the age prediction task.
- *Data Augmentation*: Random augmentations including affine transformation, Gaussian noise, and Gaussian smoothing were applied to the input data during training for the ABCD sex and intelligence prediction tasks.

After training, we selected the model instances with the highest validation AUC or lowest validation MSE to report the performance on the test dataset. The model was

Table A.2: Performance of various models on the HCP dataset with standard deviation

Method	Dataset: HCP					
	Sex		Age		Intelligence	
	ACC	AUC	MSE	MAE	MSE	MAE
XGBoost	$68.5_{\pm 3.03}$	$75.5_{\pm 3.14}$	$14.3_{\pm 1.61}$	$3.12_{\pm 0.165}$	$0.991_{\pm 0.084}$	$0.813_{\pm 0.032}$
BrainNetCNN [66]	$77.1_{\pm 2.33}$	$84.9_{\pm 2.28}$	$12.6_{\pm 0.74}$	$2.97_{\pm 0.153}$	$0.984_{\pm 0.034}$	$0.805_{\pm 0.016}$
VanillaTF [52]	$77.9_{\pm 2.08}$	$85.2_{\pm 0.89}$	$12.5_{\pm 1.15}$	$2.95_{\pm 0.182}$	$0.987_{\pm 0.039}$	$0.812_{\pm 0.014}$
BNT [52]	$81.0_{\pm 3.11}$	$88.0_{\pm 3.10}$	$12.8_{\pm 0.89}$	$2.98_{\pm 0.155}$	$1.001_{\pm 0.009}$	$0.830_{\pm 0.014}$
TFF [50]	$92.5_{\pm 1.12}$	$97.5_{\pm 1.77}$	$13.8_{\pm 1.58}$	$3.11_{\pm 0.200}$	$0.953_{\pm 0.074}$	$0.795_{\pm 0.028}$
SwiFT (ours)	$\mathbf{92.9}_{\pm 1.51}$	$\mathbf{98.0}_{\pm 1.79}$	$\mathbf{8.6}_{\pm 0.57}$	$\mathbf{2.36}_{\pm 0.114}$	$\mathbf{0.903}_{\pm 0.077}$	$\mathbf{0.786}_{\pm 0.030}$

Table A.3: Performance of various models on the UKB dataset with standard deviation

Method	Dataset: UKB					
	Sex		Age		Intelligence	
	ACC	AUC	MSE	MAE	MSE	MAE
XGBoost	$79.5_{\pm 1.28}$	$87.6_{\pm 0.94}$	$48.8_{\pm 1.01}$	$5.85_{\pm 0.046}$	$1.055_{\pm 0.199}$	$0.816_{\pm 0.078}$
BrainNetCNN [66]	$86.8_{\pm 0.19}$	$93.8_{\pm 0.31}$	$42.7_{\pm 0.17}$	$5.36_{\pm 0.113}$	$1.001_{\pm 0.141}$	$0.800_{\pm 0.060}$
VanillaTF [52]	$87.0_{\pm 1.31}$	$95.1_{\pm 0.37}$	$41.4_{\pm 1.16}$	$5.26_{\pm 0.142}$	$0.999_{\pm 0.144}$	$0.799_{\pm 0.059}$
BNT [52]	$87.0_{\pm 1.10}$	$94.8_{\pm 0.46}$	$39.6_{\pm 1.07}$	$5.17_{\pm 0.092}$	$0.998_{\pm 0.139}$	$0.798_{\pm 0.058}$
TFF [50]	$96.8_{\pm 0.25}$	$99.5_{\pm 0.06}$	$42.1_{\pm 4.80}$	$5.10_{\pm 0.331}$	$0.997_{\pm 0.123}$	$\mathbf{0.783}_{\pm 0.046}$
SwiFT (ours)	$\mathbf{97.7}_{\pm 0.31}$	$\mathbf{99.8}_{\pm 0.04}$	$\mathbf{18.2}_{\pm 0.94}$	$\mathbf{3.40}_{\pm 0.083}$	$\mathbf{0.992}_{\pm 0.105}$	$0.796_{\pm 0.044}$

trained using four NVIDIA A100 GPUs using the distributed data-parallel (DDP) strategy provided by Pytorch Lightning, with a single training session typically lasting from 4 to 30 hours depending on the dataset and whether data augmentation was used during training.

To obtain the results detailed in Section 2.4.3. (manuscript), the model was pretrained using the following:

- *Optimizer*: AdamW optimizer using a cosine decay learning rate scheduler with 2000 steps of linear warm-up

- *Learning rate*: After warm-up, an initial learning rate of 10^{-5}
- *Mini-batch size*: 3
- *Epochs*: six epochs of training

After pre-training, we fine-tuned the model without a learning rate scheduler, using 1/10 of the initial learning rate used for from-scratch training.

TFF We used the same data splits to compare our baseline 4D Transformer, TFF, with SwiFT. We followed the number of attention heads (16) and embedding size (2,640) proposed by [50]. To alleviate over-fitting, we applied data augmentation methods to the brain images, such as Gaussian blur and additive Gaussian noise implemented by imgaug[144]. Since TFF requires more computational resources than SwiFT to run the codes, at least 8 hours of training were required using 2 nodes with 4 A100 GPUs.

We trained TFF with the following training setup:

- *Optimizer*: Adam optimizer using a cosine decay learning rate scheduler with a linear warm-up by 5% of total iterations
- *Learning rate*: After warm-up, an initial learning rate of 10^{-4}
- *Mini-batch size*: 32
- *Epochs*: 10 epochs of training

ROI-based models The four ROI-based model baselines, XGBoost [95], BrainNetCNN [66], VanillaTF [52], and Brain Network Transformer [52] were reproduced for our experiments. The reproductions were based on the official code of [52]. However, since the preprocessing codes for the ABCD dataset were not provided by [52], we followed the preprocessing steps described in [52], potentially causing some differences. This obscurity in the preprocessing step is suspected to be one of the reasons for the slight performance gap of the BNT model between our experiments and the

results posted in the original paper [52], despite utilizing the same ABCD dataset.

We trained BrainNetCNN, VanillaTF, and Brain Network Transformer models with the following setup:

- *Optimizer*: Adam optimizer using a cosine decay learning rate scheduler
- *Learning rate*: Learning rate of 5×10^{-5}
- *Mini-batch size*: Mini-batch of size 16
- *Epochs*: 200 epochs of training

We used grid search for hyper-parameter tuning of XGBoost, adjusting the maximum depth and minimal child weight, gamma, learning rate, and colsample by tree. In addition, we conducted 5-fold cross-validation. Hyperparameters are tuned with the following setup:

- *Maximum depth*: Chosen between 3 and 6
- *Minimal child weight*: Chosen between 1 and 7
- *Gamma*: Chosen between 0.0 and 0.4
- *Learning rate*: Chosen between 0.05 and 0.3
- *Colsample by tree*: Chosen between 0.6 and 0.9

Software Version The major software used for our experiments are as the following:

- python 3.10.4
- pytorch 1.12.1
- pytorch-lightning 1.6.5
- monai 1.1.0
- neptune-client 0.16.4
- scipy 1.8.1
- torchvision 0.13.1
- torchaudio 0.12.1

Table A.4: Performance comparison of SwiFT for different positional embedding methods.

Dataset	Method	Sex		Age		Intelligence	
		ACC	AUC	MSE	MAE	MSE	MAE
HCP	Relative	$89.8_{\pm 1.87}$	$95.9_{\pm 1.21}$	$9.0_{\pm 0.56}$	$2.44_{\pm 0.116}$	$0.908_{\pm 0.009}$	$0.775_{\pm 0.011}$
	Absolute	$\mathbf{92.9}_{\pm 1.51}$	$\mathbf{98.0}_{\pm 1.79}$	$\mathbf{8.6}_{\pm 0.57}$	$\mathbf{2.36}_{\pm 0.114}$	$\mathbf{0.903}_{\pm 0.077}$	$\mathbf{0.786}_{\pm 0.030}$
ABCD	Relative	$\mathbf{80.2}_{\pm 1.65}$	$\mathbf{88.9}_{\pm 0.26}$	N/A		$0.936_{\pm 0.029}$	$0.761_{\pm 0.013}$
	Absolute	$79.3_{\pm 1.29}$	$87.8_{\pm 1.31}$			$\mathbf{0.932}_{\pm 0.017}$	$\mathbf{0.756}_{\pm 0.009}$
UKB	Relative	$97.5_{\pm 0.10}$	$99.8_{\pm 0.05}$	$19.4_{\pm 0.53}$	$3.53_{\pm 0.071}$	$1.019_{\pm 0.083}$	$0.807_{\pm 0.035}$
	Absolute	$\mathbf{97.7}_{\pm 0.31}$	$\mathbf{99.8}_{\pm 0.04}$	$\mathbf{18.2}_{\pm 0.94}$	$\mathbf{3.40}_{\pm 0.083}$	$\mathbf{0.992}_{\pm 0.105}$	$\mathbf{0.796}_{\pm 0.044}$

Table A.5: Efficiency comparison of SwiFT for different positional embedding methods.

Method	# Param.	FLOPs	Throughput
Relative	4.66M	2.62G	94.17
Absolute	4.64M	2.62G	104.16

A.3 Comparison of Positional Embedding Methods

We discuss the impact of the switch from a relative positional bias scheme used in most Swin Transformer variants [59, 51, 60] to an absolute positional embedding scheme, as detailed in the paragraph "4D absolute positional embedding" of Section 3.1 (manuscript). We implemented the 4D relative positional bias scheme by extending the 3D relative positional bias described in [51]. Given a window with dimensions of $P \times M \times M \times M$, the 4D relative positional bias $B \in R^{P^2 \times M^2 \times M^2 \times M^2}$ for each self-attention head is integrated as

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V, \tag{A.1}$$

where $Q, K, V \in R^{PM^3 \times C'}$ are the query, key, value matrices and C' is the channel number. A parameterized bias matrix $\hat{B} \in R^{(2P-1)\times(2M-1)\times(2M-1)}$ is used to calculate the values in B since there are only $2P-1$ or $2M-1$ possible position differences for every axis. A separate parameterized bias matrix is used for each attention head at each layer of the Transformer.

Table A.4 shows the overall performance comparison of SwiFT between the absolute positional embedding scheme and the relative positional bias scheme for the HCP, ABCD, and UKB datasets. Comparing the performance of both methods, we found that the absolute positional embedding scheme performs better in most cases. Note that the age prediction task on the ABCD dataset was not tested due to the same reasons detailed in Section A.1.

Additionally, we compared the efficiency of the two methods in Table A.5 through the number of parameters, the number of FLOPs per forward pass, and the throughput. Throughput measures how many fMRI sub-sequences of length 20 the model processes per second during inference on a single A100 GPU. The absolute positional embedding scheme is more memory efficient as it only requires parameters at the beginning of each stage. In addition, the absolute positional embedding scheme is also more computationally efficient as it does not require the 4D relative positional bias B to be reconstructed during each self-attention computation, resulting in a 9.6% throughput improvement. Overall, we conclude that the absolute positional embedding scheme is appropriate for our tasks compared to the relative positional bias scheme.

A.4 Performance of ConnTask with varying number of samples, independent components, and contrast types

In A.6, we compared the performances of ConnTask with different training samples and independent components(IC) over three emotion contrasts. We observed that more

Table A.6: Performance of ConnTask

Task	IC	Sample	D. Mean	D. Median	D. Index	KS_D	MSE	Acc.	D. rank
s	21	100	0.552	0.566	0.064	0.374	4.422	0.88	0.99
s	21	500	0.553	0.583	0.062	0.3	4.645	0.874	0.991
s	21	**1000**	0.556	0.587	0.063	0.317	4.457	0.864	0.991
s	**55**	100	0.553	0.566	0.07	0.379	4.419	0.88	0.992
s	**55**	500	0.56	0.587	0.068	0.327	4.415	0.884	0.994
s	**55**	**1000**	0.563	0.596	0.069	0.346	4.598	0.888	0.994
f	21	100	0.556	0.587	0.057	0.342	4.965	0.9	0.992
f	21	500	0.563	0.597	0.056	0.284	4.868	0.888	0.992
f	21	**1000**	0.568	0.603	0.056	0.305	5.014	0.886	0.989
f	**55**	100	0.557	0.581	0.064	0.357	4.957	0.9	0.996
f	**55**	500	0.57	0.604	0.061	0.31	4.825	0.898	0.993
f	**55**	**1000**	0.575	0.61	0.062	0.333	4.965	0.898	0.992
f-s	21	100	0.336	0.356	0.046	0.202	2.587	0.55	0.897
f-s	21	500	0.341	0.363	0.051	0.208	2.745	0.584	0.95
f-s	21	**1000**	0.351	0.377	0.05	0.213	2.756	0.579	0.952
f-s	**55**	100	0.337	0.346	0.054	0.22	2.591	0.58	0.915
f-s	**55**	500	0.348	0.374	0.057	0.23	2.733	0.61	0.948
f-s	**55**	**1000**	0.359	0.384	0.056	0.236	2.742	0.615	0.955

numbers of components have a positive effect on both overall concordance and individual identification. Including more training samples has a positive effect on overall concordance but seems to have a slightly negative effect on individual identification, except for the FACES-SHAPES contrast. This result is aligned with the previous findings [123].

Figure A.1: Training Curve of Diagonal Mean and Diagonality Index

A.5 Trade-off between overall concordance and individual identification

As shown in Figure A.1, during the training process of SwiFUN with MSE loss, the diagonal median and diagonality index initially increase together, but at some point, the diagonal mean starts decreasing while the diagonality index continues to increase. This indicates that initially, the model is trained to increase overall concordance, similar to the group mean activity. However, at a certain point, the model shifts its focus towards reflecting subtle individual differences at the expense of overall concordance. However, there is a drawback regarding the sharp decrease in the diagonal median compared to the increase in the diagonality index. Therefore, in this study, we addressed this issue by incorporating the Reconstruction-Contrastive loss, where the weights for individual differences are determined directly by the researchers, allowing the model to train in a direction that avoids excessive convergence towards group means and instead reveals individual differences.

A.6 Effect of Input sequence length for SwiFUN

In Figure A.2, we confirmed whether the sequence length impacts the overall concordance (diagonal median) and individual identification (diagonality) of SwiFUN. SwiFUN was trained on the SHAPES contrast map using MSE loss. The experiments

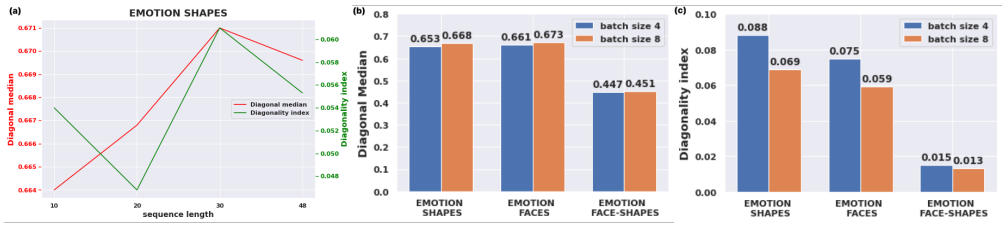


Figure A.2: Effect of sequence length and batch size on the performance of SwiFUN

were conducted with a fixed mini-batch size of 4, and four different settings were tested with lengths of 10, 20, 30, and 48 considering the limitations of GPU memory. The overall performance differences were not substantial, but the diagonal median showed improvement as the sequence length increased, reaching its peak at length 30. On the other hand, the diagonality index displayed the best performance at length 30 but did not exhibit consistent improvement with increasing length. Considering these results and the training speed associated with sequence length, all experiments were conducted with a length of 30.

A.7 Effect of batch size on Reconstruction-Contrastive Loss

In Figure A.2, we confirmed whether SwiFUN trained with RC loss are impacted by the mini-batch size. Considering that the contrastive loss compares samples within the mini-batch, we hypothesized that the mini-batch size would impact performance. In the RC loss, we set the value of $1 - \lambda$ (the weight of the contrastive loss term) to 0.33. Our analysis revealed that in all tasks, the diagonal median exhibited a small increase as the batch size increased. On the other hand, the diagonality index showed a significant decrease with larger batch sizes. This suggests that increasing the number of samples compared can weaken the effect of contrastive loss.

Bibliography

[1] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pp. 272–284, Springer, 2022.

[2] I. Tavor, O. P. Jones, R. B. Mars, S. Smith, T. Behrens, and S. Jbabdi, "Task-free mri predicts individual differences in brain activity during task performance," *Science*, vol. 352, no. 6282, pp. 216–220, 2016.

[3] V. D. Calhoun, R. Miller, G. Pearlson, and T. Adalı, "The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery," *Neuron*, vol. 84, pp. 262–274, Oct. 2014.

[4] Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nature neuroscience*, vol. 8, no. 5, pp. 679–685, 2005.

[5] M. D. Rosenberg, E. S. Finn, D. Scheinost, X. Papademetris, X. Shen, R. T. Constable, and M. M. Chun, "A neuromarker of sustained attention from whole-

brain functional connectivity," *Nature neuroscience*, vol. 19, no. 1, pp. 165–171, 2016.

[6] C.-H. Kao, A. N. Khambhati, D. S. Bassett, M. R. Nassar, J. T. McGuire, J. I. Gold, and J. W. Kable, "Functional brain network reconfiguration during learning in a dynamic environment," *Nature communications*, vol. 11, no. 1, p. 1682, 2020.

[7] M. Song, Y. Zhou, J. Li, Y. Liu, L. Tian, C. Yu, and T. Jiang, "Brain spontaneous functional connectivity and intelligence," *Neuroimage*, vol. 41, pp. 1168–1176, July 2008.

[8] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nat. Neurosci.*, vol. 18, pp. 1664–1671, Nov. 2015.

[9] J. Dubois, P. Galdi, L. K. Paul, and R. Adolphs, "A distributed brain network predicts general intelligence from resting-state human neuroimaging data," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 373, Sept. 2018.

[10] E. Dhamala, K. W. Jamison, A. Jaywant, S. Dennis, and A. Kuceyeski, "Distinct functional and structural connections predict crystallised and fluid cognition in healthy adults," *Human brain mapping*, vol. 42, no. 10, pp. 3102–3118, 2021.

[11] A. D. Nostro, V. I. Müller, D. P. Varikuti, R. N. Pläschke, F. Hoffstaedter, R. Langner, K. R. Patil, and S. B. Eickhoff, "Predicting personality from network-based resting-state functional connectivity," *Brain Struct. Funct.*, vol. 223, pp. 2699–2719, July 2018.

[12] W.-T. Hsu, M. D. Rosenberg, D. Scheinost, R. T. Constable, and M. M. Chun, "Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals," *Social cognitive and affective neuroscience*, vol. 13, no. 2, pp. 224–232, 2018.

[13] N. Franzmeier, J. Neitzel, A. Rubinski, R. Smith, O. Strandberg, R. Ossenkoppele, O. Hansson, and M. Ewers, "Functional brain architecture is associated with the rate of tau accumulation in alzheimer's disease," *Nature communications*, vol. 11, no. 1, p. 347, 2020.

[14] M.-E. Lynall, D. S. Bassett, R. Kerwin, P. J. McKenna, M. Kitzbichler, U. Muller, and E. Bullmore, "Functional connectivity and brain networks in schizophrenia," *Journal of Neuroscience*, vol. 30, no. 28, pp. 9477–9487, 2010.

[15] J. C. Mostert, E. Shumskaya, M. Mennes, A. M. H. Onnink, M. Hoogman, C. C. Kan, A. A. Vasquez, J. Buitelaar, B. Franke, and D. G. Norris, "Characterising resting-state functional connectivity in a large sample of adults with adhd," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 67, pp. 82–91, 2016.

[16] D. Hebb, "The organization of behavior. emphnew york," 1949.

[17] D. J. Heeger and D. Ress, "What does fMRI tell us about neuronal activity?," *Nat. Rev. Neurosci.*, vol. 3, pp. 142–151, Feb. 2002.

[18] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: a review on resting-state fmri functional connectivity," *European neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[19] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fmri," *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.

[20] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neurosci. Biobehav. Rev.*, vol. 74, pp. 58–75, Mar. 2017.

[21] P. C. Mulders, P. F. van Eijndhoven, A. H. Schene, C. F. Beckmann, and I. Tendolkar, "Resting-state functional connectivity in major depressive disorder: a review," *Neuroscience & Biobehavioral Reviews*, vol. 56, pp. 330–344, 2015.

[22] J. M. Sheffield and D. M. Barch, "Cognition and resting-state functional connectivity in schizophrenia," *Neuroscience & Biobehavioral Reviews*, vol. 61, pp. 108–120, 2016.

[23] J. T. Kennedy, M. P. Harms, O. Korucuoglu, S. V. Astafiev, D. M. Barch, W. K. Thompson, J. M. Bjork, and A. P. Anokhin, "Reliability and stability challenges in ABCD task fMRI data," *Neuroimage*, vol. 252, p. 119046, May 2022.

[24] S. Noble, D. Scheinost, and R. T. Constable, "A guide to the measurement and interpretation of fMRI test-retest reliability," *Curr Opin Behav Sci*, vol. 40, pp. 27–32, Aug. 2021.

[25] M. M. Herting, P. Gautam, Z. Chen, A. Mezher, and N. C. Vetter, "Test-retest reliability of longitudinal task-based fMRI: Implications for developmental studies," *Dev. Cogn. Neurosci.*, vol. 33, pp. 17–26, Oct. 2018.

[26] M. L. Elliott, A. R. Knodt, D. Ireland, M. L. Morris, R. Poulton, S. Ramrakha, M. L. Sison, T. E. Moffitt, A. Caspi, and A. R. Hariri, "What is the Test-Retest

reliability of common Task-Functional MRI measures? new empirical evidence and a Meta-Analysis," *Psychol. Sci.*, vol. 31, pp. 792–806, July 2020.

[27] E. S. Finn and R. T. Constable, "Individual variation in functional brain connectivity: implications for personalized approaches to psychiatric disease," *Dialogues in clinical neuroscience*, 2022.

[28] N. K. Dinsdale, E. Bluemke, V. Sundaresan, M. Jenkinson, S. M. Smith, and A. I. Namburete, "Challenges for machine learning in clinical translation of big data imaging studies," *Neuron*, 2022.

[29] H. Galioulline, S. Frässle, S. J. Harrison, I. Pereira, J. Heinzle, and K. E. Stephan, "Predicting future depressive episodes from resting-state fMRI with generative embedding," *Neuroimage*, vol. 273, p. 119986, June 2023.

[30] M. H. Trivedi, P. J. McGrath, M. Fava, R. V. Parsey, B. T. Kurian, M. L. Phillips, M. A. Oquendo, G. Bruder, D. Pizzagalli, M. Toups, *et al.*, "Establishing moderators and biosignatures of antidepressant response in clinical care (embarc): Rationale and design," *Journal of psychiatric research*, vol. 78, pp. 11–23, 2016.

[31] A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, A. F. Schatzberg, K. Sudheimer, J. Keller, H. S. Mayberg, F. M. Gunning, G. S. Alexopoulos, M. D. Fox, A. Pascual-Leone, H. U. Voss, B. J. Casey, M. J. Dubin, and C. Liston, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nat. Med.*, vol. 23, pp. 28–38, Jan. 2017.

[32] R. Dinga, L. Schmaal, B. W. J. H. Penninx, M. J. van Tol, D. J. Veltman, L. van Velzen, M. Mennes, N. J. A. van der Wee, and A. F. Marquand, "Evaluating the evidence for biotypes of depression: Methodological replication and extension of drysdale et al. (2017)," *NeuroImage: Clinical*, vol. 22, p. 101796, Jan. 2019.

[33] Z. Mao, Y. Su, G. Xu, X. Wang, Y. Huang, W. Yue, L. Sun, and N. Xiong, "Spatio-temporal deep learning method for adhd fmri classification," *Information Sciences*, vol. 499, pp. 1–11, 2019.

[34] A. Abrol, Z. Fu, M. Salman, R. Silva, Y. Du, S. Plis, and V. Calhoun, "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning," *Nature communications*, vol. 12, no. 1, p. 353, 2021.

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

[36] M. M. Rahman, U. Mahmood, N. Lewis, H. Gazula, A. Fedorov, Z. Fu, V. D. Calhoun, and S. M. Plis, "Interpreting models interpreting brain dynamics," *Scientific Reports*, vol. 12, no. 1, p. 12023, 2022.

[37] B.-H. Kim, J. C. Ye, and J.-J. Kim, "Learning dynamic graph representation of brain connectome with spatio-temporal attention," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4314–4327, 2021.

[38] N. Asadi, I. R. Olson, and Z. Obradovic, "A transformer model for learning spatio-temporal contextual representation in fmri data," *Network Neuroscience*, pp. 1–41, 2022.

[39] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, "Braingnn: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.

[40] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-temporal graph convolution for resting-state fmri analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, pp. 528–538, Springer, 2020.

[41] A. Bessadok, M. A. Mahjoub, and I. Rekik, "Graph neural networks in network neuroscience," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5833–5848, 2022.

[42] J. Smucny, G. Shi, and I. Davidson, "Deep learning in neuroimaging: Overcoming challenges with emerging approaches," *Frontiers in Psychiatry*, vol. 13, 2022.

[43] J. D. Power, B. L. Schlaggar, and S. E. Petersen, "Recent progress and outstanding issues in motion correction in resting state fmri," *Neuroimage*, vol. 105, pp. 536–551, 2015.

[44] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, pp. 1–7, 2022.

[45] B. Dufumier, P. Gori, J. Victor, A. Grigis, M. Wessa, P. Brambilla, P. Favre, M. Polosan, C. Mcdonald, C. M. Piguet, *et al.*, "Contrastive learning with continuous proxy meta-data for 3d mri classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pp. 58–68, Springer, 2021.

[46] M. Hon and N. M. Khan, "Towards alzheimer's disease classification through transfer learning," in *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pp. 1166–1169, IEEE, 2017.

[47] G. H. Ngo, M. Khosla, K. Jamison, A. Kuceyeski, and M. R. Sabuncu, "Predicting individual task contrasts from resting-state functional connectivity using a surface-based convolutional network," *NeuroImage*, vol. 248, p. 118849, 2022.

[48] L. E. Ismaila, P. Rasti, F. Bernard, M. Labriffe, P. Menei, A. T. Minassian, D. Rousseau, and J.-M. Lemée, "Transfer learning from healthy to unhealthy patients for the automated classification of functional brain networks in fMRI," *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, vol. 12, p. 6925, July 2022.

[49] S. Nguyen, B. Ng, A. D. Kaplan, and P. Ray, "Attend and decode: 4d fmri task state decoding using attention models," in *Machine Learning for Health*, pp. 267–279, PMLR, 2020.

[50] I. Malkiel, G. Rosenman, L. Wolf, and T. Hendler, "Self-supervised transformers for fmri representation," in *International Conference on Medical Imaging with Deep Learning*, pp. 895–913, PMLR, 2022.

[51] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

[52] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, and C. Yang, "Brain network transformer," in *Advances in Neural Information Processing Systems*, 2022.

[53] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, *et al.*, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.

[54] L. D. Nickerson, S. M. Smith, D. Öngür, and C. F. Beckmann, "Using dual regression to investigate network shape and amplitude in functional connectivity analyses," *Frontiers in neuroscience*, vol. 11, p. 115, 2017.

[55] M. M. Rahman, U. Mahmood, N. Lewis, H. Gazula, A. Fedorov, Z. Fu, V. D. Calhoun, and S. M. Plis, "Interpreting models interpreting brain dynamics," *Sci. Rep.*, vol. 12, p. 12023, July 2022.

[56] W. Li, X. Lin, and X. Chen, "Detecting alzheimer's disease based on 4d fmri: An exploration under deep learning framework," *Neurocomputing*, vol. 388, pp. 280–287, 2020.

[57] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

[58] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, vol. 2, p. 4, 2021.

[59] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[60] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

[61] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[62] B. J. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, *et al.*, "The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites," *Developmental cognitive neuroscience*, vol. 32, pp. 43–54, 2018.

[63] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. R. Andersson, L. Griffanti, G. Douaud, T. W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. M. Matthews, and S. M. Smith, "Multimodal population brain imaging in the UK biobank prospective epidemiological study," *Nat. Neurosci.*, vol. 19, pp. 1523–1536, Nov. 2016.

[64] F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. R. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, D. Vidaurre, M. Webster, P. McCarthy, C. Rorden, A. Daducci, D. C. Alexander, H. Zhang, I. Dragonu, P. M. Matthews, K. L. Miller, and S. M. Smith, "Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank," *Neuroimage*, vol. 166, pp. 400–424, Feb. 2018.

[65] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature reviews neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[66] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh, "Brainnetcnn: Convolutional neural networks for

brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.

[67] X. Deng, J. Zhang, R. Liu, and K. Liu, "Classifying asd based on time-series fmri using spatial–temporal transformer," *Computers in Biology and Medicine*, vol. 151, p. 106320, 2022.

[68] M. He, X. Hou, Z. Wang, Z. Kang, X. Zhang, N. Qiang, and B. Ge, "Multi-head attention-based masked sequence model for mapping functional brain networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pp. 295–304, Springer, 2022.

[69] H. A. Bedel, I. Şıvgın, O. Dalmaz, S. U. H. Dar, and T. Çukur, "Bolt: Fused window transformers for fmri time series analysis," *arXiv preprint arXiv:2205.11578*, 2022.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[72] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[73] S. Hong, S. Cho, J. Nam, S. Lin, and S. Kim, "Cost aggregation with 4d convolutional swin transformer for few-shot segmentation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pp. 108–126, Springer, 2022.

[74] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213, 2020.

[75] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, 2017.

[76] I. Dave, R. Gupta, M. N. Rizve, and M. Shah, "Tclr: Temporal contrastive learning for video representation," *Computer Vision and Image Understanding*, vol. 219, p. 103406, 2022.

[77] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304, JMLR Workshop and Conference Proceedings, 2010.

[78] O. Esteban, C. Markiewicz, R. Blair, C. Moodie, A. Isik, A. Aliaga, and K. Gorgolewski, "Fmriprep: a robust preprocessing pipeline for functional mri. biorxiv, 306951," 2018.

[79] O. Esteban, R. Ciric, K. Finc, R. W. Blair, C. J. Markiewicz, C. A. Moodie, J. D. Kent, M. Goncalves, E. DuPre, D. E. Gomez, *et al.*, "Analysis of task-based functional mri data preprocessed with fmriprep," *Nature protocols*, vol. 15, no. 7, pp. 2186–2202, 2020.

[80] V. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, D. L. Collins, B. D. C. Group, *et al.*, "Unbiased average age-appropriate atlases for pediatric studies," *Neuroimage*, vol. 54, no. 1, pp. 313–327, 2011.

[81] Y. Behzadi, K. Restom, J. Liau, and T. T. Liu, "A component based noise correction method (compcor) for bold and perfusion based fmri," *Neuroimage*, vol. 37, no. 1, pp. 90–101, 2007.

[82] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the human connectome project," *Neuroimage*, vol. 80, pp. 105–124, Oct. 2013.

[83] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, M. Kelly, T. Laumann, K. L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A. Z. Snyder, A. T. Vu, M. W. Woolrich, J. Xu, E. Yacoub, K. Uğurbil, D. C. Van Essen, M. F. Glasser, and WU-Minn HCP Consortium, "Resting-state fMRI in the human connectome project," *Neuroimage*, vol. 80, pp. 144–168, Oct. 2013.

[84] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, p. e1001779, Mar. 2015.

[85] H. WU-Minn, "1200 subjects data release reference manual," *URL https://www. humanconnectome. org*, 2017.

[86] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, *et al.*, "A

multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.

[87] R. M. Shansky and C. S. Woolley, "Considering sex as a biological variable will be valuable for neuroscience research," *Journal of Neuroscience*, vol. 36, no. 47, pp. 11817–11822, 2016.

[88] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary, "Brain age and other bodily 'ages': implications for neuropsychiatry," *Molecular psychiatry*, vol. 24, no. 2, pp. 266–281, 2019.

[89] R. C. Gershon, D. Cella, N. A. Fox, R. J. Havlik, H. C. Hendrie, and M. V. Wagster, "Assessment of neurological and behavioural function: the nih toolbox," *The Lancet Neurology*, vol. 9, no. 2, pp. 138–139, 2010.

[90] P. R. Millar, B. A. Gordon, P. H. Luckett, T. L. Benzinger, C. Cruchaga, A. M. Fagan, J. Hassenstab, R. J. Perrin, S. E. Schindler, R. F. Allegri, *et al.*, "Multimodal brain age estimates relate to alzheimer disease biomarkers and cognition in early stages: a cross-sectional observational study," *Elife*, vol. 12, p. e81869, 2023.

[91] K. Supekar, C. de Los Angeles, S. Ryali, K. Cao, T. Ma, and V. Menon, "Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism," *The British Journal of Psychiatry*, vol. 220, no. 4, pp. 202–209, 2022.

[92] T. M. Karrer, D. S. Bassett, B. Derntl, O. Gruber, A. Aleman, R. Jardri, A. R. Laird, P. T. Fox, S. B. Eickhoff, O. Grisel, *et al.*, "Brain-based ranking of cognitive domains to predict schizophrenia," *Human brain mapping*, vol. 40, no. 15, pp. 4487–4507, 2019.

[93] J. Rokicki, T. Wolfers, W. Nordhøy, N. Tesli, D. S. Quintana, D. Alnæs, G. Richard, A.-M. G. de Lange, M. J. Lund, L. Norbom, *et al.*, "Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders," *Human brain mapping*, vol. 42, no. 6, pp. 1714–1726, 2021.

[94] N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, *et al.*, "Prediction of individual brain maturity using fmri," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.

[95] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[96] A. Riaz, M. Asad, S. M. R. Al Arif, E. Alonso, D. Dima, P. Corr, and G. Slabaugh, "Deep fmri: An end-to-end deep network for classification of fmri data," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 1419–1422, IEEE, 2018.

[97] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Local explanation methods for deep neural networks lack sensitivity to parameter values," *arXiv preprint arXiv:1810.03307*, 2018.

[98] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[99] T. Koscik, D. O'Leary, D. J. Moser, N. C. Andreasen, and P. Nopoulos, "Sex differences in parietal lobe morphology: relationship to mental rotation performance," *Brain and cognition*, vol. 69, no. 3, pp. 451–459, 2009.

[100] J. Salinas, E. D. Mills, A. L. Conrad, T. Koscik, N. C. Andreasen, and P. Nopoulos, "Sex differences in parietal lobe structure and development," *Gender Medicine*, vol. 9, no. 1, pp. 44–55, 2012.

[101] P. M. Macey, N. S. Rieken, R. Kumar, J. A. Ogren, H. R. Middlekauff, P. Wu, M. A. Woo, and R. M. Harper, "Sex differences in insular cortex gyri responses to the valsalva maneuver," *Frontiers in neurology*, p. 87, 2016.

[102] S. Weis, K. R. Patil, F. Hoffstaedter, A. Nostro, B. T. Yeo, and S. B. Eickhoff, "Sex classification by resting state brain connectivity," *Cerebral cortex*, vol. 30, no. 2, pp. 824–835, 2020.

[103] B. Ficek-Tani, C. Horien, S. Ju, W. Xu, N. Li, C. Lacadie, X. Shen, D. Scheinost, T. Constable, and C. Fredericks, "Sex differences in default mode network connectivity in healthy aging adults," *Cerebral Cortex*, vol. 33, no. 10, pp. 6139–6151, 2023.

[104] M. Ernst, B. Benson, E. Artiges, A. X. Gorka, H. Lemaitre, T. Lago, R. Miranda, T. Banaschewski, A. L. Bokde, U. Bromberg, *et al.*, "Pubertal maturation and sex effects on the default-mode network connectivity implicated in mood dysregulation," *Translational psychiatry*, vol. 9, no. 1, p. 103, 2019.

[105] X. Wu, X. Lu, H. Zhang, Y. Bi, R. Gu, Y. Kong, and L. Hu, "Sex difference in trait empathy is encoded in the human anterior insula," *Cereb. Cortex*, vol. 33, pp. 5055–5065, Apr. 2023.

[106] M. Leming and J. Suckling, "Deep learning for sex classification in resting-state and task functional brain networks from the uk biobank," *NeuroImage*, vol. 241, p. 118409, 2021.

[107] X. Li, H. Li, and L. Ma, "Continual learning of medical image classification based on feature replay," in *2022 16th IEEE International Conference on Signal Processing (ICSP)*, vol. 1, pp. 426–430, IEEE, 2022.

[108] M. H. Lee, N. Kim, J. Yoo, H.-K. Kim, Y.-D. Son, Y.-B. Kim, S. M. Oh, S. Kim, H. Lee, J. E. Jeon, *et al.*, "Multitask fmri and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder," *Scientific Reports*, vol. 11, no. 1, p. 9402, 2021.

[109] A. Zalesky and M. Breakspear, "Towards a statistical test for functional connectivity dynamics," *Neuroimage*, vol. 114, pp. 466–470, 2015.

[110] N. Leonardi and D. Van De Ville, "On spurious and real fluctuations of dynamic functional connectivity during rest," *Neuroimage*, vol. 104, pp. 430–436, 2015.

[111] M. W. Cole, T. Ito, D. S. Bassett, and D. H. Schultz, "Activity flow over resting-state networks shapes cognitive task activations," *Nature neuroscience*, vol. 19, no. 12, pp. 1718–1726, 2016.

[112] N. Tik, S. Gal, A. Madar, T. Ben-David, M. Bernstein-Eliav, and I. Tavor, "Generalizing prediction of task-evoked brain activity across datasets and populations," *Neuroimage*, vol. 276, p. 120213, June 2023.

[113] Y.-Q. Zheng, S.-R. Farahibozorg, W. Gong, H. Rafipoor, S. Jbabdi, and S. Smith, "Accurate predictions of individual differences in task-evoked brain activity from resting-state fmri using a sparse ensemble learner," *Neuroimage*, vol. 259, p. 119418, 2022.

[114] A. S. Greene, S. Gao, D. Scheinost, and R. T. Constable, "Task-induced brain state manipulation improves prediction of individual traits," *Nat. Commun.*, vol. 9, p. 2807, July 2018.

[115] C. Sripada, M. Angstadt, S. Rutherford, A. Taxali, and K. Shedden, "Toward a "treadmill test" for cognition: Improved prediction of general cognitive ability from the task activated brain," *Hum. Brain Mapp.*, vol. 41, pp. 3186–3197, Aug. 2020.

[116] S. Gal, Y. Coldham, N. Tik, M. Bernstein-Eliav, and I. Tavor, "Act natural: Functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity," *Neuroimage*, vol. 258, p. 119359, Sept. 2022.

[117] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann, "Correspondence of the brain's functional architecture during activation and rest," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, pp. 13040–13045, Aug. 2009.

[118] M. W. Cole, D. S. Bassett, J. D. Power, T. S. Braver, and S. E. Petersen, "Intrinsic and task-evoked network architectures of the human brain," *Neuron*, vol. 83, pp. 238–251, July 2014.

[119] M. L. Elliott, A. R. Knodt, M. Cooke, M. J. Kim, T. R. Melzer, R. Keenan, D. Ireland, S. Ramrakha, R. Poulton, A. Caspi, T. E. Moffitt, and A. R. Hariri, "General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks," *Neuroimage*, vol. 189, pp. 516–532, Apr. 2019.

[120] S. Gal, N. Tik, M. Bernstein-Eliav, and I. Tavor, "Predicting individual traits from unperformed tasks," *Neuroimage*, vol. 249, p. 118920, Apr. 2022.

[121] N. Tik, A. Livny, S. Gal, K. Gigi, G. Tsarfaty, M. Weiser, and I. Tavor, "Predicting individual variability in task-evoked brain activity in schizophrenia," *Hum. Brain Mapp.*, vol. 42, pp. 3983–3992, Aug. 2021.

[122] O. Parker Jones, N. L. Voets, J. E. Adcock, R. Stacey, and S. Jbabdi, "Resting connectivity predicts task activation in pre-surgical populations," *Neuroimage Clin*, vol. 13, pp. 378–385, 2017.

[123] A. D. Cohen, Z. Chen, O. Parker Jones, C. Niu, and Y. Wang, "Regression-based machine-learning approaches to predict task activation using resting-state fMRI," *Hum. Brain Mapp.*, vol. 41, pp. 815–826, Feb. 2020.

[124] K. Yoo, M. D. Rosenberg, Y. H. Kwon, D. Scheinost, R. T. Constable, and M. M. Chun, "A cognitive state transformation model for task-general and task-specific subsystems of the brain connectome," *Neuroimage*, vol. 257, p. 119279, Aug. 2022.

[125] M. Bernstein-Eliav and I. Tavor, "The prediction of brain activity from connectivity: Advances and applications," *Neuroscientist*, p. 10738584221130974, Oct. 2022.

[126] J. Zhang, A. Kucyi, J. Raya, A. N. Nielsen, J. S. Nomi, J. S. Damoiseaux, D. J. Greene, S. G. Horovitz, L. Q. Uddin, and S. Whitfield-Gabrieli, "What have we really learned from functional connectivity in clinical populations?," *NeuroImage*, vol. 242, p. 118466, 2021.

[127] X. Jiang, X. Li, J. Lv, T. Zhang, S. Zhang, L. Guo, and T. Liu, "Sparse representation of hcp grayordinate data reveals novel functional architecture of cerebral cortex," *Human brain mapping*, vol. 36, no. 12, pp. 5301–5319, 2015.

[128] F. Zhao, Z. Wu, and G. Li, "Deep learning in cortical surface-based neuroimage analysis: a systematic review," *Intelligent Medicine*, vol. 3, pp. 46–58, Feb. 2023.

[129] V. Nozais, P. Boutinaud, V. Verrecchia, M.-F. Gueye, P.-Y. Hervé, C. Tzourio, B. Mazoyer, and M. Joliot, "Deep learning-based classification of resting-state fMRI independent-component analysis," *Neuroinformatics*, vol. 19, pp. 619–637, Oct. 2021.

[130] C. F. Beckmann, C. E. Mackay, N. Filippini, S. M. Smith, *et al.*, "Group comparison of resting-state fmri data using multi-subject ica and dual regression," *Neuroimage*, vol. 47, no. Suppl 1, p. S148, 2009.

[131] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.

[132] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE transactions on medical imaging*, vol. 23, no. 2, pp. 137–152, 2004.

[133] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith, "Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers," *Neuroimage*, vol. 90, pp. 449–468, 2014.

[134] G. Grabner, A. L. Janke, M. M. Budge, D. Smith, J. Pruessner, and D. L. Collins, "Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults," in *Medical Image Computing and Computer-*

Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part II 9, pp. 58–66, Springer, 2006.

[135] A. R. Hariri, A. Tessitore, V. S. Mattay, F. Fera, and D. R. Weinberger, "The amygdala response to emotional stimuli: a comparison of faces and scenes," *Neuroimage*, vol. 17, pp. 317–323, Sept. 2002.

[136] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, *et al.*, "Function in the human connectome: task-fmri and individual differences in behavior," *Neuroimage*, vol. 80, pp. 169–189, 2013.

[137] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pp. 205–218, Springer, 2023.

[138] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[139] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

[140] S. B. Eickhoff, M. Milham, and T. Vanderwal, "Towards clinical applications of movie fmri," *NeuroImage*, vol. 217, p. 116860, 2020.

[141] T. Vanderwal, J. Eilbott, E. S. Finn, R. C. Craddock, A. Turnbull, and F. X. Castellanos, "Individual differences in functional connectivity during naturalistic viewing conditions," *Neuroimage*, vol. 157, pp. 521–530, 2017.

[142] E. S. Finn and P. A. Bandettini, "Movie-watching outperforms rest for functional connectivity-based prediction of behavior," *NeuroImage*, vol. 235, p. 117963, 2021.

[143] E. S. Finn, E. Glerean, A. Y. Khojandi, D. Nielson, P. J. Molfese, D. A. Handwerker, and P. A. Bandettini, "Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging," *NeuroImage*, vol. 215, p. 116828, 2020.

[144] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte, *et al.*, "imgaug." `https://github.com/aleju/imgaug`, 2020. Online; accessed 01-Feb-2020.

국문초록

변화하는 환경에서 적응하는 뇌 기능을 연구하기 위해 기능적 자기공명영상(functional magnetic resonance imaging, fMRI)과 같은 고차원 뇌 이미지에서 뇌 역동을 분석하는 것은 중요한 과제이다. 그러나 기존의 fMRI 연구는 주로 단순한 특징 추출 기반 방법들에 의존하기 때문에 뇌 역동에서 중요한 측면을 간과할 위험이 존재한다. 이러한 기존 접근법의 한계를 극복하기 위해 본 연구는 fMRI 분석을 위한 두 가지 심층 신경망(SwiFT, SwiFUN)을 제안한다. 이 모델들은 4차원 형태의 휴지 상태 fMRI (resting-state fMRI, rs-fMRI) 데이터를 직접 처리함으로써 인지 및 생물학적 변수와 특정 과제 수행 시의 뇌 활동을 효과적으로 예측할 수 있다. 본 연구는 인간 커넥톰 프로젝트(Human Connectome Project, HCP), 청소년 뇌 인지 발달 연구(Adolescent Brain Cognitive Development, ABCD), 영국 바이오뱅크(UK Biobank, UKB)과 같은 대규모 fMRI 데이터를 활용한다. 그 결과, SwiFT는 성별, 연령, 그리고 지능 예측에서 현존하는 최신 방법들보다 뛰어난 성능을 보였다. 또한, SwiFUN은 rs-fMRI에서 특정 과제 수행 시의 뇌 활동을 예측하는 태스크에서 기존 일반화된 선형 모델 (generalized linear model, GLM)에 비해 우수한 성능을 보였다. 본 연구는 고차원 fMRI에서 복잡한 뇌 역동을 효과적으로 분석하는 방법을 제시함으로써, 신경과학 분야에서 대규모 fMRI를 활용할 수 있는 새로운 가능성을 시사한다.

주요어: 뇌 역동, 기능적 자기공명영상, 트랜스포머, 인지 및 생물학적 변수, 특정 과제 수행 시의 뇌 활동

학번: 2021-23364