ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890789

Hierarchical Spatiotemporal Attention Network for Fine-grained Brain Cognitive State Recognition

Yike Wu
School of Computer Science and Engineering
Southeast University
Nanjing, Jiangsu Province
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications(Southeast University),
Ministry of Education, China
yikewu@seu.edu.cn

Ning An
School of Computer Science and Engineering
Southeast University
Nanjing, Jiangsu Province
ningan@seu.edu.cn

Zixuan Zeng
School of Software Engineering
Southeast University
Nanjing, Jiangsu Province
zengzixuan2209@seu.edu.cn

YouYong Kong*
School of Computer Science and Engineering
Southeast University
Nanjing, Jiangsu Province
kongyouyong@seu.edu.cn

***Abstract*—Brain cognitive state recognition based on functional Magnetic Resonance Imaging(fMRI) can capture brain functional activities under different tasks and help understand the neural mechanisms of the brain, which has always been one of the focuses of neuroscience research. Different from the prediction of the brain cognitive domain, the prediction of brain fine-grained cognitive state is based on each moment in the process of executing the task. Therefore, it is necessary to extract more fine-grained effective information. Existing studies focus on modeling and classifying the complete time series, ignoring the brain activity state at each time. So we propose a hierarchical spatiotemporal attention network(FineBrainNet) to recognize fine-grained brain cognitive state. Guided by coarse-grained cognitive domain labels, we trained different sub-modules for fine-grained states under each cognitive domain to capture relevant cognitive state changes more accurately in specific task. Extensive experiments on the HCP-Task dataset show that FineBrainNet can achieve accurate prediction of fine-grained brain cognitive state.**

***Index Terms*—fMRI, spatiotemporal attention, fine-grained state, brain state recognition.**

I. INTRODUCTION

Identifying the cognitive state of the brain has always been one of the focuses of neuroscience research [1]. Modern imaging techniques, such as functional Magnetic Resonance Imaging(fMRI), provide opportunities to study complex cognitive processes in the human brain [2].

There is evidence that the functional connectivity (FC) provides key information for brain state recognition [3]. Current models for brain state recognition mainly focus on static and dynamic modeling of FC. Static FC mainly extracts the correlation between different brain regions and represents the global data in the form of a graph, thus naturally using graph neural networks to learn brain information [4]–[7]. Although static methods can extract global information, they may lose the temporal information. Therefore, recent studies [8]–[10] have focused on modeling brain information dynamically by using a series of FC segmented according to time. At the same time, some studies [9], [11]–[13] have attempted to combine static and dynamic brain information to extract more comprehensive brain information. However, unlike the prediction of the brain cognitive state domain, the prediction of fine-grained brain cognitive state requires the extraction of brain state information in a shorter time slice, which makes it costly and almost impossible to directly apply the time-varying properties of FC obtained by sliding window approach to the recognition of fine-grained brain cognitive state.

In actual scenarios, subjects will perform different tasks at each moment according to guidance, showing different brain cognitive states. Different cognitive states in a period of time can be summarized into the same brain cognitive state domain. Existing studies focus on modeling and classifying the complete time series, ignoring the brain activity state at each time. BrainNetFormer [12]successfully predicted fine-grained states when predicting the cognitive domains but did not make full use of the prediction information of them. In fact, the cognitive domains of the brain provide important guidance information, which can limit the prediction of fine-grained states to a certain range. This guidance information comes from the organization of cognitive tasks, that is, the types of cognitive states of the brain in a period of time can first be divided into several cognitive domains, and each moment in every cognitive domains can be divided into specific states. Therefore, how to use coarse-grained cognitive domain labels to guide the prediction of fine-grained states under each domain is important.

In this paper, we propose a new model **FineBrainNet** that integrates dynamic information and static information to predict the fine-grained cognitive state of the brain hierarchically. Our main contributions are as follows: (1) We design a static-dynamic encoding structure to extract brain information, and use spatiotemporal cross-attention to achieve information

supported by "the Fundamental Research Funds for the Central Universities2242024k30035"

*Corresponding author

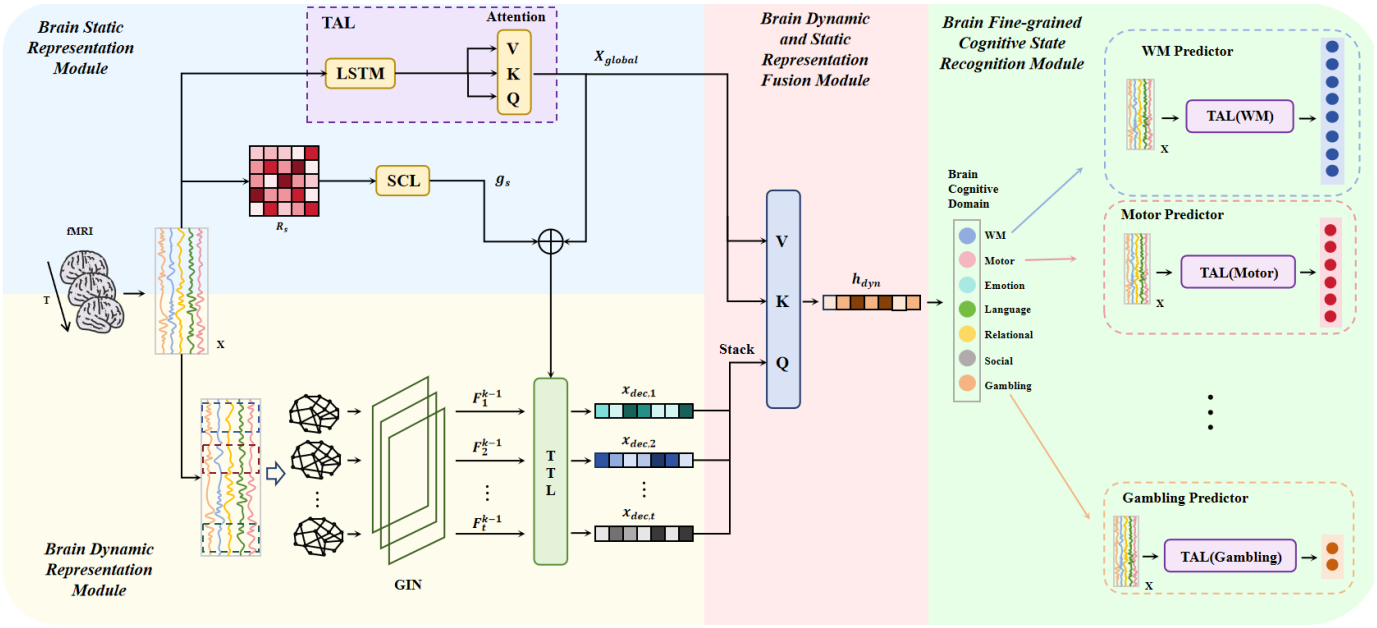

Fig. 1. The framework of FineBrainNet. First, the BOLD signal is sent to Brain Static Representation Module to get a global static representation of the brain in time and space. Then, dynamic graph data is obtained through sliding window approach and sent to Brain Dynamic Representation Module. After GIN and TTL, we get the graph readout vector. After Brain Static-Dynamic Representation Fusion Module, the static and dynamic representations are fused and the cognitive domain state of the brain is determined. Brain Fine-grained Cognitive State Recognition Module reuses the previous structure and finally identifies the fine-grained cognitive state of the brain.

fusion. (2) We propose a hierarchical guided brain cognitive state recognition model. The model uses coarse-grained cognitive domain labels for guidance, first classifies the brain cognitive state domain, and then trains the specific models under each domain to predict the fine-grained brain state. (3) We conducted experiments on a large Human Connectome Project (HCP) dataset containing 7 main tasks and 25 subtasks, verifying the superiority of the model proposed in this paper.

II. METHODOLOGY

As is shown in Fig. 1. , our model includes four modules. We get brain spatiotemporal static representations through Brain Static Representation Module and brain dynamic representations through Brain Dynamic Representation Module. After that we fuse the static and dynamic information via Brain Static-Dynamic Representation Fusion Module. Guided by the predicted domain labels, we send the processed information to Brain Fine-grained Cognitive State Recognition Module and train different predictors for fine-grained states, getting the prediction of fine-grained brain state finally.

A. Brain Static Representation Module

In this module, the Blood Oxygen Level Dependent(BOLD) signal, as input, will be modeled statically to get the brain spatiotemporal static representation.

1) Temporal Attention Layer: In the brain static representation processing module, the first task is to encode the brain signal into a vector and get a static representation of the whole brain signal, which is then sent to the Brain Static-Dynamic Representation Fusion Module for spatiotemporal attention fusion. So in Temporal Attention Layer(TAL), we use the Long Short-Term Memory(LSTM) model to positionally encode the time series.

$$X_e = XW_e + \text{LSTM}(XW_e) \tag{1}$$

where $X \in \mathbb{R}^{T\times N}$ is the original BOLD signal, $W_e \in \mathbb{R}^{N\times C}$ is a learnable matrix, T is the total length of the input signal, N is the number of brain regions, C is the number of hidden dimension and $X_e \in \mathbb{R}^{T\times C}$ is the encoded BOLD signal. Afterwards, we use a single-head attention layer to extract global information:

$$Q = X_eW_Q; K = X_eW_K; V = X_eW_V \tag{2}$$

$$X_{\text{global}} = \text{softmax}(\frac{QK^T}{\sqrt{C}})V \tag{3}$$

Through the operation, we get $X_{global} \in \mathbb{R}^{T\times C}$, which is the global static representation of the brain, containing the spatial information of the brain.

2) Spatial Coding Layer: Spatial Coding Layer(SCL) is applied to embed the whole BOLD signals into a global vector to represent the static characteristics of the whole brain. FC is able to capture brain functional patterns, so we encode it as a static embedding. In order to capture the spatial characteristics between different brain regions, we use the correlation matrix of the brain region in the entire period as an input of uniform length. We flatten the correlation matrix R_s into a vector, and then send the vector to SCL to get the global spatial static representation g_s of the brain:

$$g_s = SCL(\text{Flatten}(R_s)) \tag{4}$$

where $R_s \in \mathbb{R}^{N\times N}$, $g_s \in \mathbb{R}^C$. The SCL is a neural network consisting of a linear layer, a normalization layer, and an

activation layer. At this point, we get spatiotemporal static representations of the brain over the entire time period.

B. Brain Dynamic Representation Module

This module will get the dynamic graph data constructed by a sliding window approach and extract the dynamic information.

1) Graph Convolutional Layer: We use graph isomorphism networks (GIN) [14] to extract graph structure information for processing. The construction of dynamic graph data $G_{dyn} = (V, F, A)$ is based on the sliding window approach, which generates $N_w = [T - L]/S$ continuous sliding windows by setting the time window length to L and moving it with a step size of S. $V = \{v_1, v_2, ..., v_N\}$ represents the set of nodes in the brain region. $F \in \mathbb{R}^{N_w \times N \times L}$ contains the dynamic node features obtained by the sliding window approach. $F_t \in \mathbb{R}^{N \times L}$ represents the node feature set at time point t. $A \in \mathbb{R}^{N_w \times N \times N}$ is the set of dynamic adjacency matrix, where $A_t \in \mathbb{R}^{N \times N}$ is the brain adjacency matrix at time point t.

In the graph convolution module, we multiply the adjacency matrix with the node feature matrix to aggregate the neighbor information of each node. At the same time, self-loops are introduced and adjusted by a learnable parameter ϵ. Finally, the aggregated features are nonlinearly transformed through a multi-layer perceptron to learn more complex feature representations. The $i - th$ layer is calculated as follows:

$$F_t^i = \text{MLP}((A_t + \epsilon^i I) \times W^i F_t^{i-1}) \tag{5}$$

where A_t is the adjacency matrix at time t, ϵ is the self-loop weight, I is the identity matrix, $W^i \in \mathbb{R}^{C \times C}$ is a learnable matrix of the $i - th$ layer and F_t^{k-1} is the last layer of the graph convolutional network output at time t. After passing through graph convolutional layer, the features of different brain regions at each moment are spatially learned.

2) Timing Transfer Layer: In the Timing Transfer Layer(TTL), we introduce the global spatiotemporal static vector of the brain from the Brain Static Representation Module to calculate the weights of brain regions. The g_s obtained through the Brain Static Representation Module is added to X_{global} to get the query vector. Then we perform dot product attention calculation on F_t^{k-1} to get the weights of different brain regions. At the same time, superimpose the graph readout vectors at each sampling time to form $X_{\mathbf{dec}} = [x_{dec,1}, ..., x_{dec,N_w}] \in \mathbb{R}^{N_w \times C}$.

$$SA_t = \text{softmax}(F_t^{k-1}(g_s + X_{\text{global},t})) \tag{6}$$

$$x_{dec,t} = (SA_t^T + 1)F_t^{k-1} \tag{7}$$

where $SA_t \in \mathbb{R}^N$, which measures the attention distribution of brain regions at sample point t. F_t^{k-1} is the output of the last layer of network at sample point t. $x_{dec,t} \in \mathbb{R}^C$ is the final graph readout vector of sample point t.

C. Brain Static-Dynamic Representation Fusion Module

This module is specifically used to analyze and decode the time series data of brain regions. It uses a multi-head self-attention mechanism to effectively capture the dynamic interactions between brain regions. The input of the attention layer is the time series data of brain regions. The input data is processed in parallel by num_heads attention heads, and each head independently calculates the attention weight:

$$Q_T = X_{dec}W_{QT}; K_T = X_{dec}W_{KT}; V_T = X_{dec}W_{VT} \tag{8}$$

$$H_{embed} = \text{softmax}(\frac{Q_T K_T^T}{\sqrt{d_k}})V_T \tag{9}$$

where $H_{\text{embed}} \in \mathbb{R}^{L \times C}$ is the fused feature matrix, which contains rich information. After attention calculation, the feature matrix passes through fully connected network and layer-normalized residual connection. At this time, the global information and local information are fused, and then the embedding vector with aggregated information is summed in the time dimension to obtain the final brain state representation vector $h_{dyn} \in \mathbb{R}^C$, which is finally sent to the linear feedforward layer to obtain the recognition result of the brain cognitive domain, which will then be used for fine-grained brain state recognition.

D. Brain Fine-grained Cognitive State Recognition Module

After identifying the brain's cognitive domains, we have obtained a rough recognition of the brain's fine-grained cognitive states. We then designed a Brain Fine-grained Cognitive State Recognition Module to predict the fine-grained states under each cognitive domain. Our design is inspired by the way cognitive tasks are organized, which reflects the unique characteristics of the subjects' brain cognitive states, that is, the types of brain cognitive states over a period of time can first be divided into several cognitive domains, and each shorter time slice in each cognitive domain can be divided into specific fine-grained cognitive states. In TAL, we encode the global information of the brain as X_{global}, making it possible to predict the fine-grained cognitive state of the brain.

In the fine-grained cognitive state prediction layer, we establish a submodule for each predicted cognitive state domain, which reuses the structure of the TAL as the fine-grained brain cognitive state prediction submodule of the predicted cognitive state domain. After the original BOLD signal $X \in \mathbb{R}^{T \times N}$ corresponding to the state domain is input into the corresponding prediction submodule, the fine-grained state feature information of the period is extracted. Then, the extracted vector is passed through a fully connected layer to get the fine-grained state probability vector $l \in \mathbb{R}^{N_d}$, where N_d is the number of possible fine-grained cognitive states contained in the cognitive state domain. Finally, l is dimensionally expanded to get the final fine-grained state probability vector $l_{final} \in \mathbb{R}^M$ for cross entropy loss classification, where M

TABLE I
COMPARISION EXPERIMENT RESULTS

Methods	Fine-grained Cognitive Brain States ACC(%)
SVM-RBF [16]	57.9(±1.8)
MLP-Mixer [16]	81.7(±1.4)
ST-GCN [16]	74.9(±1.8)
BrainNetFormer [12]	79.2(±0.6)
FineBrainNet	**85.6(±3.2)**

is the total number of all possible fine-grained cognitive state categories. The regularization loss function is as follows:

$$L = -\frac{1}{\sum_{i=1}^{Q} T_i} \sum_{i=1}^{Q} \sum_{t=1}^{T_i} \sum_{j=0}^{M-1} y_{ijt} \log(p_{ijt}) \qquad (10)$$

where y_{ijt} is the true label of the $i-th$ subject at the $t-th$ time, and is 1 only when the $j-th$ cognitive task is performed. p_{ijt} is the predicted probability that the $i-th$ subject performs the $j-th$ cognitive task at the $t-th$ time. M is the number of fine-grained cognitive states, Q is the number of subjects, and T_i is the length of the brain signal data of the $i-th$ subject.

III. EXPERIMENTS

A. Dataset and Experimental Details

The HCP dataset includes brain imaging and behavioral data of more than 1,200 healthy young adults. The task fMRI data includes 7 cognitive state domains and each corresponds to a few different experimental conditions. There are a total of 25 different fine-grained cognitive states (including resting state) under different experimental conditions. We use the S1200 version of the fMRI dataset, which contains a total of 7450 scans from 594 female and 501 male subjects, after excluding incomplete scans. In the preprocessing stage, the brain is divided into 90 brain regions using the atlas - Anatomical Automatic Labeling (AAL) [15]. Our goal is to identify the specific tasks that the subjects are performing at each moment from 25 different experimental conditions. We conducted experiments on the Pytorch platform, using the Adam optimizer. We set the batch size to 16, epoch to 40, learning rate to 0.0005, hidden dimension to 128, window length to 50 and sliding window step to 3, used 10-fold cross validation and trained our model on a single NVIDIA TITAN RTX GPU.

B. Experiment Results

1) Comparison Experiments: We compared our model with several recent models for identifying fine-grained cognitive states in the brain. Currently, most studies focus on the recognition of brain cognitive state domains, and there are few articles on the recognition of fine-grained brain states. Since Ye et al. [16] did not make their code public, we use the best results in their paper for comparison experiments. The results in Table I show that FineBrainNet has higher accuracy and is better than previous models. Although BrainNetFormer uses spatial and temporal attention mechanisms, it does not consider the correlation between the seven cognitive state domains and the fine-grained cognitive states.

TABLE II
ABLATION EXPERIMENT RESULTS

Methods	ACC(%)
FineBrainNet	85.6(±3.2)
w/o Brain Fine-grained Cognitive State Recognition Module	79.2(±0.6)
w/o Brain Static-Dynamic Representation Fusion Module	81.2(±5.1)

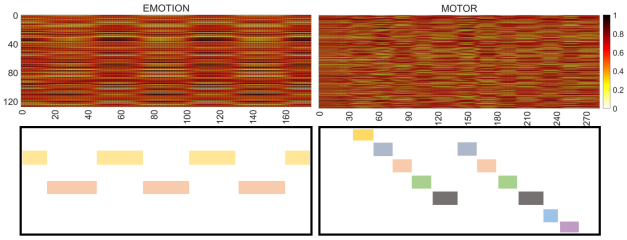

Fig. 2. Temporal attention of EMOTION and MOTOR domain

2) Ablation Experiments: In order to study the role of the Brain Static-Dynamic Representation Fusion Module and the hierarchical architecture, we eliminated Brain Static-Dynamic Representation Fusion Module and Brain Fine-grained Cognitive State Recognition Module respectively, and performed the recognition task again. The results in Table II show that Brain Static-Dynamic Representation Fusion Module and the hierarchical recognition architecture guided by coarse labels are both important for recognizing fine-grained brain states. This may be because that the Brain Static-Dynamic Representation Fusion Module combines the static information and the dynamic features from the previous modules, containing richer brain state information. The hierarchical recognition architecture makes full use of the correlation between the brain cognitive state domain and the fine-grained state and considers the uniqueness of fine-grained states under different cognitive state domains.

C. Attention Analysis

The HCP task-fMRI data includes 7 cognitive domains, each of which consists of fine-grained cognitive states. We visualized the temporal attention of the fine-grained cognitive state tasks in the cognitive domains of MOTOR and EMOTION. In Fig. 2. , each row represents a temporal pattern and each column represents a time point of a fine-grained cognitive state. The visualized temporal attention clearly reflects the different patterns of different fine-grained cognitive states, and it can be seen that our hierarchical prediction model can accurately identify the fine-grained cognitive states of the brain.

IV. CONCLUSION

In this work, we propose a novel hierarchical spatiotemporal attention model FineBrainNet, in which the spatiotemporal fusion attention module and the coarse label-guided hierarchical recognition architecture are designed to play an outstanding role in the recognition of fine-grained brain states. The results of our experiments on the HCP-Task dataset also show that FineBrainNet can achieve accurate prediction of fine-grained brain cognitive states.

REFERENCES

[1] Jonas Richiardi, Hamdi Eryilmaz, Sophie Schwartz, Patrik Vuilleumier, and Dimitri Van De Ville. Decoding brain states from fmri connectivity graphs. *Neuroimage*, 56(2):616–626, 2011.

[2] Marek Kubicki, Robert W McCarley, Paul G Nestor, T Huh, Ron Kikinis, Martha Elizabeth Shenton, and Cynthia G Wible. An fmri study of semantic processing in men with schizophrenia. *Neuroimage*, 20(4):1923–1933, 2003.

[3] R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 2013.

[4] Yu Zhang, Loïc Tetrel, Bertrand Thirion, and Pierre Bellec. Functional annotation of human cognitive states using deep graph convolution. *NeuroImage*, 231:117847, 2021.

[5] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

[6] Mahsa Ghorbani, Anees Kazi, Mahdieh Soleymani Baghshah, Hamid R Rabiee, and Nassir Navab. Ra-gcn: Graph convolutional network for disease prediction problems with imbalanced data. *Medical image analysis*, 75:102272, 2022.

[7] Elif Sema Balcioglu, Berkay Doner, Ekansh Sareen, Dimitri Van De Ville, and Hamid Behjat. Joint subject-identification and task-decoding from inferred functional brain graphs via a multi-task neural network. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[8] Soham Gadgil, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Ehsan Adeli, and Kilian 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*, pages 528–538. Springer, 2020.

[9] Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021.

[10] Youyong Kong, Shuwen Gao, Yingying Yue, Zhenhua Hou, Huazhong Shu, Chunming Xie, Zhijun Zhang, and Yonggui Yuan. Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity. *Human brain mapping*, 42(12):3922–3933, 2021.

[11] Jiwon Lee, Eunsong Kang, Junyeong Maeng, and Heung-Il Suk. Eigendecomposition-based spatial-temporal attention for brain cognitive states identification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1921–1925. IEEE, 2024.

[12] Leheng Sheng, Wenhan Wang, Zhiyi Shi, Jichao Zhan, and Youyong Kong. Brainnetformer: Decoding brain cognitive states with spatial-temporal cross attention. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[13] Hasan A Bedel, Irmak Sivgin, Onat Dalmaz, Salman UH Dar, and Tolga Çukur. Bolt: Fused window transformers for fmri time series analysis. *Medical image analysis*, 88:102841, 2023.

[14] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[15] Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

[16] Ziyuan Ye, Youzhi Qu, Zhichao Liang, Mo Wang, and Quanying Liu. Explainable fmri-based brain decoding via spatial temporal-pyramid graph convolutional network. *Human Brain Mapping*, 44(7):2921–2935, 2023.