Cross-Modal Brain Graph Transformer via Function-Structure Connectivity Network for Brain Disease Diagnosis

Jingxi Feng[1], Heming Xu[1], Junhao Cai[2], Yujie Chang[1], Dong Zhang[1], Shaoyi Du[1] ✉, and Juan Wang[3] ✉

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
{jingxifeng8}@gmail.com {hemingxu, shaoyidu}@xjtu.edu.cn
{dongzhangcv}@gmail.com
[2] Shenzhen University, ShenZhen 518060, China
{caijunhao27}@gmail.com
[3] Department of Ultrasound, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710114, China
{wangjuan}@optimism@163.com

Abstract. Multi-modal brain networks represent the complex connectivity between different brain regions from both functional and structural perspectives, which is of great significance for brain disease diagnosis. However, existing methods are limited to information fusion in the feature dimension, failing to fully exploit the complementary information between functional and structural connectivity networks. To address these issues, this paper proposes a cross-modal brain graph transformer (CBGT) method for brain disease diagnosis, which also provides an in-depth analysis of coupled function-structure connectivity networks. Specifically, CBGT consists of two main modules: the cross-modal Transformer module enhances the attention mechanism by utilizing structural connectivity features extracted through machine learning methods, capturing long-range dependencies in the cross-modal brain network. The cross-modal topK pooling module combines information from both functional and structural connectivity networks to select significant regions of interest (ROIs) during the reconstruction of the pooled graph, aiming to retain as much effective information as possible. Experiments conducted on the ABIDE and ADNI datasets demonstrate that the proposed method outperforms state-of-the-art approaches. Interpretation analysis reveals that the proposed method can identify multi-modal biomarkers associated with brain diseases.

Keywords: Cross-modal · Transformer · Brain disease · Pooling.

1 Introduction

The brain network analysis technology based on neuroimaging data facilitates the exploration of the potential associations between brain function and structure, which is of great significance for the brain disease diagnosis. Brain networks

are categorized into functional connectivity networks and structural connectivity networks. Functional connectivity networks are constructed using signals extracted through functional magnetic resonance imaging (fMRI) [20], which reflect the coherence of brain activity across different ROIs. Structural connectivity networks are built using fiber tracts measured by diffusion tensor imaging (DTI) [2], describing the physical neural fiber connections between ROIs.

Previous studies have demonstrated that functional connectivity networks and structural connectivity networks can provide complementary information to each other [18]. However, due to the complex interactions between the two types of brain networks, effectively leveraging complementary information from functional and structural connectivity networks to enhance brain disease diagnostic performance remains a challenge.

The core of brain disease diagnosis based on brain network analysis lies in modeling and learning the relationships between brain regions, where extensive long-range dependencies exist within brain networks [1, 23]. Unlike the local information propagation mechanism of graph neural networks (GNNs) [4, 24], the self-attention mechanism in Transformer can effectively capture long-range dependencies between nodes. Recent studies [13, 7, 8] have introduced Transformers into brain network analysis. For instance, the brain network Transformer (BrainNetTF) [13] utilizes a Transformer encoder to learn node embeddings of functional connectivity networks and employs an orthogonal clustering readout method to obtain graph-level representations, achieving promising performance in brain disease diagnosis tasks. However, such methods overlook the interdependence between functional and structural connectivity networks.

Additionally, existing research has utilized multi-modal information from functional and structural connectivity networks for disease diagnosis. For example, BrainNN [29] treats functional and structural connectivity networks as multi-view data, extracting features from each using GNNs and employing contrastive learning for multi-modal fusion. However, these methods neglect the supporting role of structural connectivity in functional connectivity and the widespread long-range dependencies in brain networks.

To address these issues, we propose a cross-modal brain graph Transformer for brain disease diagnosis tasks. First, we extract important features from the structural connectivity network using machine learning methods to enhance the attention mechanism. Secondly, a cross-modal topK pooling method is employed to retain significant ROIs during dimensionality reduction by considering information from both functional and structural connectivity networks. Finally, a soft-voting strategy is used to integrate the feature representations extracted from each layer, achieving the final brain disease diagnosis. In summary, this paper makes three major contributions:

1) We propose a cross-modal brain graph Transformer method that effectively leverages the complementary information from functional and structural connectivity networks to achieve accurate brain disease diagnosis.

2) We propose a cross-modal Transformer method that utilizes key features filtered through structural connectivity network as the enhanced mask for atten-

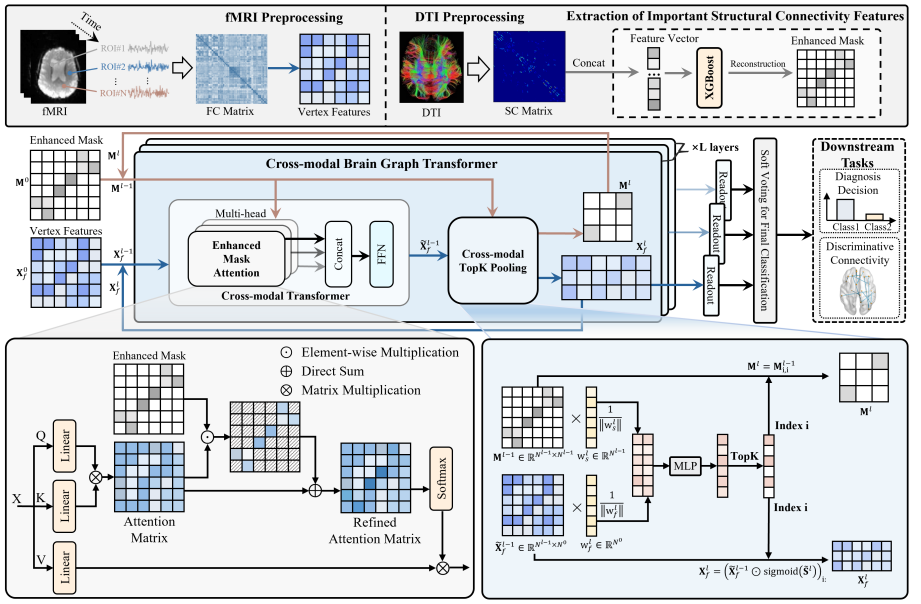

Fig. 1. The proposed cross-modal brain graph Transformer framework.

tion computation based on functional connectivity network, thereby capturing the long-range dependencies in cross-modal brain networks.

3) We propose a cross-modal topK pooling method that integrates information from functional and structural connectivity networks to retain ROIs critical for disease prediction. This method reduces graph size while preserving as much effective information as possible.

2 Method

2.1 Overview

As shown in Fig. 1, the overall framework consists of three components: 1) Extraction of important structural connectivity features, where XGBoost is used to extract key structural connectivity features as the enhanced mask for the initial input to the cross-modal brain graph Transformer. 2) The cross-modal brain graph Transformer consists of multiple iterative layers, each including the cross-modal Transformer layer and the cross-modal topK pooling layer. 3) At the decision layer, a soft voting strategy integrates feature representations from each layer of the cross-modal brain graph Transformer to achieve a final decision.

2.2 Extraction of Important Structural Connectivity Features

The structural connectivity matrix $\mathbf{X}_s \in \mathbb{R}^{N^0 \times N^0}$ can be obtained from DTI image, where $\mathbf{X}_{s,(i,j)}$ denotes the number of fiber connections between ROIs i

and j. Previous studies have shown that $\mathbf{X}_{s,(i,j)}$ can range from zero to several thousand and has a skewed distribution. To make the sample means conform to a normal distribution, we first use the logarithmic transformation [5] to narrow down the range of fibre numbers between the two ROIs:

$$\mathbf{X}_{s,(i,j)} = \log_{10}(\mathbf{X}_{s,(i,j)} + 1). \tag{1}$$

Then we normalize $\mathbf{X}_{s,(i,j)}$:

$$\mathbf{X}_{s,(i,j)} = \frac{\mathbf{X}_{s,(i,j)} - (\mathbf{X}_s)_m}{(\mathbf{X}_s)_s}, \tag{2}$$

where $(\mathbf{X}_s)_m$ and $(\mathbf{X}_s)_s$ are the mean and standard deviation of $\mathbf{X}_s$. To extract influential feature information from structural connectivity network, we vectorise the normalised structural connectivity matrix $\mathbf{X}_s$. The vectorized network samples are input into the XGBoost, which calculates the weight of each feature. By setting a threshold p, unimportant feature information is filtered out. Then, we reconstruct these filtered features, and the reconstruction matrix is used as the initial enhanced mask $\mathbf{M}^0 \in \mathbb{R}^{N^0 \times N^0}$ in the cross-modal Transformer.

2.3 Cross-modal Brain Graph Transformer

The cross-modal brain graph Transformer consists of two components in each layer: the cross-modal Transformer layer and the cross-modal topK pooling layer, which will be introduced in the following two subsections.

Cross-modal Transformer To capture the long-range dependencies based on structural connectivity in the cross-modal brain network, the proposed cross-modal Transformer layer applies the enhanced mask during the attention score computation process. Formally, for the l-th layer of the cross-modal brain graph Transformer, we utilise the multi-head self-attention module with the enhanced mask to generate more expressive node features and the output $\widetilde{\mathbf{X}}_f^{l-1}$ is obtained by the following equation:

$$\widetilde{\mathbf{X}}_f^{l-1} = \left(\Big\|_{z=1}^{Z} \mathbf{h}^{l,z}\right) \mathbf{W}_O^l \,, \tag{3}$$

$$\mathbf{T}^{l,z} = \text{softmax}\left(\frac{\mathbf{W}_Q^{l,z}\mathbf{X}_f^{l-1}\left(\mathbf{W}_K^{l,z}\mathbf{X}_f^{l-1}\right)^\top}{\sqrt{d_K^{l,z}}} \odot \left(1 + \mathbf{M}^{l-1}\right)\right), \tag{4}$$

$$\mathbf{h}^{l,z} = \mathbf{T}^{l,z}\mathbf{W}_V^{l,z}\mathbf{X}_f^{l-1}, \tag{5}$$

where $\mathbf{X}_f^{l-1} \in \mathbb{R}^{N^{l-1} \times N^0}$ denotes the input node feature matrix in the l-th layer, the initial node feature matrix is the functional connectivity (FC) matrix $\mathbf{X}_f^0 \in \mathbb{R}^{N^0 \times N^0}$ obtained by calculating the Pearson correlation coefficient between ROIs. $\|$ denotes the concatenation operator, Z is the number of heads,

$\mathbf{W}_O^l$, $\mathbf{W}_Q^{l,z}$, $\mathbf{W}_K^{l,z}$, and $\mathbf{W}_V^{l,z}$ are learnable model parameters, and $d_K^{l,z}$ is the first dimension of $\mathbf{W}_K^{l,z}$. The matrix $\mathbf{M}^{l-1} \in \mathbb{R}^{N^{l-1} \times N^{l-1}}$ is the enhanced mask in the l-th layer, where the initial mask is $\mathbf{M}^0$. The $\mathbf{M}^{l-1}$ is used in the attention computation process to enhance the attention of each ROI to its structurally connected neighboring brain regions. The output node features $\widetilde{\mathbf{X}}_f^{l-1}$ and the enhanced mask $\mathbf{M}^{l-1}$ will be input into the cross-modal topK pooling layer.

Cross-modal TopK Pooling Recent findings have shown that some ROIs are more important than the others in the prediction of brain disease [12, 3]. Thus, it is crucial to use a node pooling layer to reduce the size of the graph and only preserve some important nodes. To this end, we propose a cross-modal topK pooling method that comprehensively evaluates node importance by integrating cross-modal brain network information, retaining the top k most important nodes. Specifically, in the cross-modal topK pooling layer of the l-th layer in the cross-modal brain graph Transformer, we project the enhanced mask $\mathbf{M}^{l-1}$, which reflects the filtered structural connectivity information, and the output node features $\widetilde{\mathbf{X}}_f^{l-1}$, onto learnable vectors $\mathbf{w}_s^l \in \mathbb{R}^{N^{l-1}}$ and $\mathbf{w}_f^l \in \mathbb{R}^{N^0}$, respectively. The node importance from the structural and functional perspectives, $\mathbf{s}_1^l$ and $\mathbf{s}_2^l$, can be obtained through the following calculations:

where $\|\cdot\|$ is the L_2 norm. $\mathbf{s}_1^l$ and $\mathbf{s}_2^l$ are concatenated to form $\mathbf{S}^l \in \mathbb{R}^{N^{l-1} \times 2}$, which is then passed through a single-layer multi-layer perceptron (MLP) to learn the final score vector $\widetilde{\mathbf{S}}^l$ that reflects the cross-modal importance.

$$\widetilde{\mathbf{S}}^l = \sigma(\mathbf{S}^l \cdot \mathbf{W}^l + \mathbf{b}^l)\,, \tag{6}$$

where $\mathbf{W}^l \in \mathbb{R}^2$ represents the learnable weights and $\mathbf{b}^l \in \mathbb{R}^{N^{l-1}}$ represents the bias. Based on $\widetilde{\mathbf{S}}^l$, the index vector $\mathbf{i}$ for the top k nodes can be selected. Utilizing these indices, the pooled node features $\mathbf{X}_f^l$ and the enhanced mask $\mathbf{M}^l$ for the next layer can be obtained, which can be represented as follows:

$$\mathbf{i} = \text{topk}(\widetilde{\mathbf{S}}^l, k) \tag{7}$$

$$\mathbf{X}_f^l = (\widetilde{\mathbf{X}}_f^{l-1} \odot \text{sigmoid}(\widetilde{\mathbf{S}}^l))_{\mathbf{i}:}\,, \quad \mathbf{M}^l = \mathbf{M}_{\mathbf{i},\mathbf{i}}^{l-1}\,, \tag{8}$$

where $\odot$ denotes the elementwise product and $(\cdot)_{i,j}$ is an indexing operation which takes elements at row indices specified by i and column indices specified by j (colon denotes all indices).

2.4 Soft Voting for Final Classification Decision

To make a final label prediction, we employ the soft voting strategy at the decision level. The output node features from each brain graph Transformer layer are concatenated to obtain the graph-level representation $\hat{\mathbf{x}}^l$ for each layer. These are then input into a MLP, where the classification probabilities $p^l(c)$ from each

layer are combined using a soft voting strategy for the final classification. The calculation formula for soft voting can be expressed with the following equation:

$$P(c) = \frac{\sum_{l=1}^{L} p^l(c)}{L}, \tag{9}$$

The probability for each class c, $P(c)$, is calculated by averaging the predicted probabilities from each layer and the predicted class is determined by $\arg\max_c P(c)$. The whole process is supervised with cross-entropy loss.

3 Experiments

3.1 Dataset

The proposed method is evaluated on two public datasets, Alzheimer's Disease Neuroimaging Initiative (ADNI) [11] and Autism Brain Imaging Data Exchange (ABIDE) [10]. This study uses a subset of the ADNI with a total of 330 cases, including 66 AD patients, 125 mild cognitive impairment (MCI) patients, and 139 NCs. Only 95 subjects in ABIDE dataset have both fMRI and DTI data, including 53 Autism Spectrum Disorder (ASD) patients and 42 normal controls (NCs). We preprocess the fMRI via DPARSF[4], and the DTI via PANDA. The brain space is parcellated into 90 ROIs based on the AAL atlas [19].

3.2 Experimental Settings

Implement and Evaluation Metrics. Models are implemented in PyTorch and trained on NVIDIA 3090. Adam optimizer [15] is used with an initial learning rate of 1×10^{-4}. Using grid search, the structural feature selection threshold p is set to 3, and the optimal number of CBGT layers L is determined to be 2. Five-fold cross-validation is employed on the dataset, and the proposed method is evaluated in terms of four metrics: accuracy, sensitivity, specificity, and AUC.
Compared Methods. To verify the effectiveness of our proposed method, we compare it with BrainGNN [16], BrainIB [28], BrainNetTF [13], ALTER [26], SVM [9], GAT [21], BrainNN [29], MME-GCN [17], Cross-GNN [25].

3.3 Experimental Results and Analysis

Table 1 shows the performance metrics of all methods. The results demonstrate that our method achieves the highest accuracy across all classification tasks. In the MCI vs. NC classification task, our method increases the accuracy by more than 3.9% compared to other cross-modal methods. The primary reason for these gains is that our approach effectively captures long-range dependencies in the cross-modal brain network. Meanwhile, the cross-modal topK pooling, considering node cross-modal importance, preserves essential graph information and improves diagnostic accuracy.

[4] http://rfmri.org/DPARSF

Table 1. Comparative Experiments of Classification Tasks on Different Datasets.

Datasets (Tasks)		ADNI(AD vs. NC)				ADNI(AD vs. MCI)			
Modal	Method	ACC (%)	SPE (%)	SEN (%)	AUC (%)	ACC (%)	SPE (%)	SEN (%)	AUC (%)
fMRI	BrainGNN	$74.6_{\pm 9.5}$	$78.4_{\pm 9.8}$	$66.0_{\pm 9.8}$	$78.0_{\pm 6.7}$	$65.5_{\pm 5.2}$	$68.7_{\pm 6.1}$	$57.5_{\pm 9.9}$	$61.6_{\pm 5.4}$
	BrainIB	$74.1_{\pm 9.3}$	$76.3_{\pm 8.6}$	$69.5_{\pm 12.6}$	$77.3_{\pm 6.4}$	$66.5_{\pm 5.2}$	$71.2_{\pm 7.6}$	$55.0_{\pm 3.4}$	$65.4_{\pm 6.0}$
	BrainNetTF	$75.1_{\pm 6.1}$	$79.1_{\pm 7.6}$	$67.8_{\pm 13.0}$	$76.3_{\pm 5.8}$	$67.0_{\pm 4.4}$	$71.2_{\pm 6.9}$	$59.0_{\pm 4.3}$	$66.3_{\pm 5.2}$
	ALTER	$75.6_{\pm 4.4}$	$79.0_{\pm 7.3}$	$68.8_{\pm 14.1}$	$77.1_{\pm 6.6}$	$67.0_{\pm 4.4}$	$72.8_{\pm 7.8}$	$55.9_{\pm 6.4}$	$67.1_{\pm 4.2}$
DTI	SVM	$69.3_{\pm 5.7}$	$73.8_{\pm 10.3}$	$55.7_{\pm 13.2}$	$68.7_{\pm 6.0}$	$60.2_{\pm 5.7}$	$65.5_{\pm 11.5}$	$50.3_{\pm 10.7}$	$61.0_{\pm 7.4}$
	GAT	$70.7_{\pm 4.4}$	$77.4_{\pm 8.2}$	$57.0_{\pm 17.5}$	$69.1_{\pm 10.6}$	$63.4_{\pm 7.6}$	$68.3_{\pm 9.7}$	$51.2_{\pm 7.1}$	$62.3_{\pm 7.3}$
Both	BrainNN	$77.6_{\pm 2.8}$	$83.8_{\pm 8.3}$	$63.4_{\pm 12.5}$	$77.5_{\pm 3.0}$	$65.0_{\pm 6.5}$	$70.9_{\pm 8.4}$	$51.3_{\pm 6.0}$	$61.5_{\pm 4.0}$
	MME-GCN	$77.6_{\pm 7.8}$	$81.5_{\pm 10.0}$	$69.7_{\pm 13.4}$	$74.9_{\pm 5.5}$	$69.6_{\pm 2.8}$	$76.3_{\pm 4.9}$	$56.7_{\pm 8.5}$	$64.1_{\pm 6.5}$
	Cross-GNN	$78.5_{\pm 7.9}$	$82.8_{\pm 9.5}$	$70.1_{\pm 13.7}$	$77.5_{\pm 6.1}$	$70.1_{\pm 5.3}$	$71.9_{\pm 6.3}$	$\mathbf{63.4}_{\pm 5.1}$	$70.4_{\pm 1.9}$
	Ours	$\mathbf{81.5}_{\pm 4.5}$	$\mathbf{86.4}_{\pm 4.1}$	$\mathbf{71.5}_{\pm 13.7}$	$\mathbf{79.4}_{\pm 6.7}$	$\mathbf{74.3}_{\pm 3.1}$	$\mathbf{80.1}_{\pm 4.4}$	$62.4_{\pm 14.3}$	$\mathbf{71.1}_{\pm 7.9}$

Datasets (Tasks)		ADNI(MCI vs. NC)				ABIDE(ASD vs. NC)			
Modal	Method	ACC (%)	SPE (%)	SEN (%)	AUC (%)	ACC (%)	SPE (%)	SEN (%)	AUC (%)
fMRI	BrainGNN	$62.8_{\pm 2.4}$	$62.3_{\pm 8.9}$	$65.7_{\pm 8.9}$	$65.3_{\pm 6.1}$	$65.3_{\pm 8.6}$	$69.4_{\pm 12.3}$	$61.0_{\pm 9.2}$	$65.5_{\pm 13.1}$
	BrainIB	$64.7_{\pm 2.9}$	$66.4_{\pm 10.0}$	$63.6_{\pm 13.1}$	$63.8_{\pm 3.9}$	$64.2_{\pm 11.2}$	$60.0_{\pm 12.7}$	$72.8_{\pm 14.5}$	$67.0_{\pm 5.1}$
	BrainNetTF	$65.4_{\pm 6.8}$	$63.3_{\pm 8.5}$	$67.7_{\pm 7.1}$	$62.8_{\pm 11.5}$	$65.3_{\pm 11.3}$	$65.3_{\pm 14.8}$	$68.5_{\pm 14.8}$	$67.0_{\pm 11.2}$
	ALTER	$65.5_{\pm 3.0}$	$\mathbf{72.4}_{\pm 11.7}$	$56.2_{\pm 13.6}$	$62.0_{\pm 9.1}$	$67.4_{\pm 6.1}$	$58.4_{\pm 7.9}$	$\mathbf{78.6}_{\pm 8.4}$	$64.4_{\pm 13.0}$
DTI	SVM	$57.8_{\pm 6.1}$	$58.7_{\pm 12.6}$	$57.9_{\pm 8.6}$	$59.2_{\pm 10.9}$	$59.0_{\pm 9.1}$	$58.9_{\pm 14.1}$	$59.4_{\pm 9.9}$	$66.2_{\pm 11.5}$
	GAT	$60.8_{\pm 5.2}$	$62.1_{\pm 11.4}$	$61.4_{\pm 10.5}$	$63.1_{\pm 8.6}$	$60.0_{\pm 2.6}$	$61.1_{\pm 6.8}$	$58.3_{\pm 13.2}$	$62.4_{\pm 13.0}$
Both	BrainNN	$66.4_{\pm 6.0}$	$64.5_{\pm 7.5}$	$68.8_{\pm 6.8}$	$66.7_{\pm 8.6}$	$67.4_{\pm 7.0}$	$60.7_{\pm 9.9}$	$76.1_{\pm 8.0}$	$67.5_{\pm 11.4}$
	MME-GCN	$66.2_{\pm 3.2}$	$65.5_{\pm 5.9}$	$66.7_{\pm 13.8}$	$64.0_{\pm 7.6}$	$67.4_{\pm 5.2}$	$64.7_{\pm 12.5}$	$71.7_{\pm 4.1}$	$65.3_{\pm 10.2}$
	Cross-GNN	$66.6_{\pm 5.2}$	$66.0_{\pm 8.0}$	$66.1_{\pm 6.3}$	$64.4_{\pm 5.8}$	$68.4_{\pm 3.3}$	$67.9_{\pm 11.7}$	$70.2_{\pm 6.3}$	$67.8_{\pm 5.7}$
	Ours	$\mathbf{70.5}_{\pm 6.7}$	$70.7_{\pm 14.5}$	$\mathbf{71.6}_{\pm 11.5}$	$\mathbf{69.2}_{\pm 4.6}$	$\mathbf{74.7}_{\pm 5.2}$	$\mathbf{74.4}_{\pm 9.5}$	$74.2_{\pm 7.6}$	$\mathbf{76.0}_{\pm 3.5}$

3.4 Ablation Study

We conduct ablation experiments across two datasets and the results are shown in Table 2. Using a vanilla unimodal Transformer results in a 4.9% decrease in accuracy compared to the cross-modal Transformer in the AD vs. NC classification task, indicating that the cross-modal Transformer utilizes structural information to enhance the model's ability to capture long-range dependencies in cross-modal brain networks. Ablation of cross-modal topK pooling or using unimodal topK pooling leads to a performance decline, indicating that cross-modal topK pooling, by considering both functional and structural connectivity, can more accurately identify ROIs crucial for disease prediction.

Table 2. Ablation Study Results on Different Datasets.

Transformer		topK pooling		ADNI (AD vs. NC)		ABIDE (ASD vs. NC)	
unimodal	cross-modal	unimodal	cross-modal	ACC(%)	AUC(%)	ACC(%)	AUC(%)
✓			✓	$76.6_{\pm 2.0}$	$76.9_{\pm 6.9}$	$69.5_{\pm 6.1}$	$64.6_{\pm 9.2}$
	✓			$79.5_{\pm 3.3}$	$77.1_{\pm 6.8}$	$70.5_{\pm 6.3}$	$67.5_{\pm 10.4}$
	✓	✓		$80.5_{\pm 5.1}$	$77.9_{\pm 7.6}$	$71.6_{\pm 5.4}$	$66.2_{\pm 12.0}$
	✓		✓	$\mathbf{81.5}_{\pm 4.5}$	$\mathbf{79.4}_{\pm 6.7}$	$\mathbf{74.7}_{\pm 5.2}$	$\mathbf{76.0}_{\pm 3.5}$

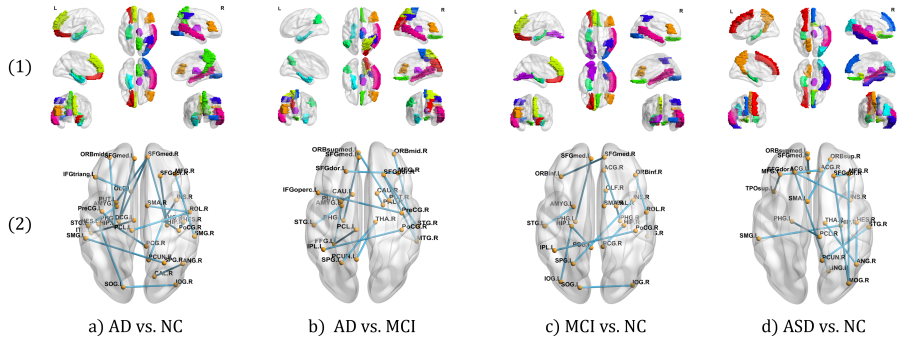

Fig. 2. Visualization of discriminative ROIs and connectivity.

3.5 Interpretation Analysis

Discriminative ROIs Analysis. We use the t-test to measure the scores of each node obtained from cross-modal pooling, in order to identify discriminative brain regions. The important ROIs for AD vs. NC are superior frontal gyrus, medial (SFGmed), superior frontal gyrus, orbital part (ORBsup), parahippocampal gyrus (PHG), etc. The important ROIs for AD vs. MCI are REC, PHG, inferior parietal lobule (IPL), etc. The important ROIs for MCI vs. NC are IPL, amygdala (AMYG), hippocampus (HIP), etc. which is consistent with previous studies [27]. In the ASD vs. NC classification, the important ROIs are precuneus (PCUN), rectus (REC), etc. As mentioned in the literature [14], when ASD patients infer their mental states, the activation of PCUN and REC decreases.
Discriminative Connectivity Analysis. We use a standard t-test to measure the attention maps based on the enhanced mask and display significantly different connections (p-value< 0.01) in Fig.2. By counting the important connections of each ROI in the ASD diagnostic task, it is found that the connections and ROIs are mainly located in the frontal areas such as superior frontal gyrus (SFG) and middle frontal gyrus (MFG). The maturation process in these ROIs is disrupted in ASD patients, leading to social and language dysfunctions [6]. In the diagnostic tasks using the ADNI dataset, the important connections and brain regions are PHG, HIP, etc., which is consistent with previous studies [22].

4 Conclusion

This paper proposes a cross-modal brain graph Transformer (CBGT) method, which effectively leverages the complementary information of functional and structural connectivity networks for brain disease diagnosis. CBGT consists of two primary modules: the cross-modal Transformer module, which enhances the attention mechanism by utilizing filtered key structural connectivity features. The cross-modal topK pooling module, which selects significant ROIs by integrating information from cross-modal brain networks, thereby minimizing infor-

mation loss. Experiments conducted on the ABIDE and ADNI datasets demonstrate that the proposed method outperforms state-of-the-art approaches and identifies multi-modal brain network biomarkers.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant Nos. 62088102, 82202183 and 62125305, and Guangdong Major Project of Basic and Applied Basic Research under Grant No. 2023B0303000009.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arnulfo, G., Wang, S.H., Myrov, V., Toselli, B., Hirvonen, J., Fato, M., Nobili, L., Cardinale, F., Rubino, A., Zhigalov, A., et al.: Long-range phase synchronization of high-frequency oscillations in human cortex. Nature communications **11**(1), 5363 (2020)
2. Assaf, Y., Pasternak, O.: Diffusion tensor imaging (dti)-based white matter mapping in brain research: a review. Journal of molecular neuroscience **34**, 51–61 (2008)
3. Baker, J.T., Holmes, A.J., Masters, G.A., Yeo, B.T., Krienen, F., Buckner, R.L., Öngür, D.: Disruption of cortical association networks in schizophrenia and psychotic bipolar disorder. JAMA psychiatry **71**(2), 109–118 (2014)
4. Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., Sun, X.: Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3438–3445 (2020)
5. Curran-Everett, D.: Explorations in statistics: the log transformation. Advances in physiology education **42**(2), 343–347 (2018)
6. Dapretto, M., Davies, M.S., Pfeifer, J.H., Scott, A.A., Sigman, M., Bookheimer, S.Y., Iacoboni, M.: Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. Nature neuroscience **9**(1), 28–30 (2006)
7. Feng, J., Han, X., Xu, H., Wang, J., Jiang, J., Du, S., Gao, Y.: Cross-template-based hypergraph transformer. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
8. Han, X., Feng, J., Xu, H., Du, S., Li, J.: A hypergraph transformer method for brain disease diagnosis. Frontiers in Medicine **11**, 1496573 (2024)
9. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)
10. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F.: Identification of autism spectrum disorder using deep learning and the abide dataset. NeuroImage: Clinical **17**, 16–23 (2018)
11. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
12. Kaiser, M.D., Hudac, C.M., Shultz, S., Lee, S.M., Cheung, C., Berken, A.M., Deen, B., Pitskel, N.B., Sugrue, D.R., Voos, A.C., et al.: Neural signatures of autism. Proceedings of the National Academy of Sciences **107**(49), 21223–21228 (2010)

13. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. Advances in Neural Information Processing Systems **35**, 25586–25599 (2022)
14. Kana, R.K., Maximo, J.O., Williams, D.L., Keller, T.A., Schipul, S.E., Cherkassky, V.L., Minshew, N.J., Just, M.A.: Aberrant functioning of the theory-of-mind network in children and adolescents with autism. Molecular autism **6**, 1–12 (2015)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. Medical Image Analysis **74**, 102233 (2021)
17. Liu, L., Wang, Y.P., Wang, Y., Zhang, P., Xiong, S.: An enhanced multi-modal brain graph network for classifying neuropsychiatric disorders. Medical image analysis **81**, 102550 (2022)
18. Park, H.J., Friston, K.: Structural and functional brain networks: from connections to cognition. Science **342**(6158), 1238411 (2013)
19. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M.: Automated anatomical labelling atlas 3. Neuroimage **206**, 116189 (2020)
20. Van Den Heuvel, M.P., Pol, H.E.H.: Exploring the brain network: a review on resting-state fmri functional connectivity. European neuropsychopharmacology **20**(8), 519–534 (2010)
21. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
22. Vipin, A., Loke, Y.M., Liu, S., Hilal, S., Shim, H.Y., Xu, X., Tan, B.Y., Venketasubramanian, N., Chen, C.L.H., Zhou, J.: Cerebrovascular disease influences functional and structural network connectivity in patients with amnestic mild cognitive impairment and alzheimer's disease. Alzheimer's research & therapy **10**, 1–15 (2018)
23. Wang, Y., Royer, J., Park, B.y., Vos de Wael, R., Larivière, S., Tavakol, S., Rodriguez-Cruces, R., Paquola, C., Hong, S.J., Margulies, D.S., et al.: Long-range functional connections mirror and link microarchitectural and cognitive hierarchies in the human brain. Cerebral Cortex **33**(5), 1782–1798 (2023)
24. Wein, S., Malloni, W.M., Tomé, A.M., Frank, S.M., Henze, G.I., Wüst, S., Greenlee, M.W., Lang, E.W.: A graph neural network framework for causal inference in brain networks. Scientific reports **11**(1), 8061 (2021)
25. Yang, Y., Ye, C., Guo, X., Wu, T., Xiang, Y., Ma, T.: Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning. IEEE Transactions on Medical Imaging **43**(1), 108–121 (2023)
26. Yu, S., Jin, S., Li, M., Sarwar, T., Xia, F.: Long-range brain graph transformer. Advances in Neural Information Processing Systems **37**, 24472–24495 (2024)
27. Zhao, Q., Lu, H., Metmer, H., Li, W.X., Lu, J.: Evaluating functional connectivity of executive control network and frontoparietal network in alzheimer's disease. Brain research **1678**, 262–272 (2018)
28. Zheng, K., Yu, S., Li, B., Jenssen, R., Chen, B.: Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. IEEE Transactions on Neural Networks and Learning Systems (2024)
29. Zhu, Y., Cui, H., He, L., Sun, L., Yang, C.: Joint embedding of structural and functional brain networks with graph neural networks for mental illness diagnosis. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 272–276. IEEE (2022)