FROG: A Fine-grained Spatiotemporal Graph Neural Network with Self-supervised Guidance for Early Diagnosis of Alzheimer's Disease

Shuoyan Zhang, Qingmin Wang, Min Wei, Jiayi Zhong, Ying Zhang, Ziyan Song, Chenyang Li, Xiaochen Zhang, Ying Han, Yunxia Li, for the Alzheimer's Disease Neuroimaging Initiative, Han Lv, and Jiehui Jiang, *Senior Member, IEEE*

***Abstract*— Functional magnetic resonance imaging (fMRI) has demonstrated significant potential in the early diagnosis and study of pathological mechanisms of Alzheimer's disease (AD). To fit subtle cross-spatiotemporal interaction and learn pathological feature from fMRI, we proposed a fine-grained spatiotemporal graph neural network with self-supervised learning (SSL) for diagnosis and biomarker extraction of early AD. First, Considering the spatiotemporal interaction of the brain, we designed two masks that leverage the spatial correlation and temporal repeability of the fMRI. Afterwards, temporal gated inception convolution and graph scalable inception convolution were proposed for the spatiotemporal autoencoder to enhance subtle cross-spatiotemporal variation and learn noise-suppressed signals. Furthermore, a spatiotemporal scalable cosine error with high selectivity for signal reconstruction was designed in SSL to guide the autoencoder to fit the fine-grained pathological features in an unsupervised manner. A total of 5687 samples from four across-population cohorts were involved. The accuracy of our model was 5% higher than the state-of-the-art models, which included four AD diagnostic models, four SSL strategies, and three multivariate time series models. The neuroimaging biomarkers were precisely localized to the abnormal brain regions, and correlated significantly with the cognitive scale and biomarkers (P<0.001). Moreover, the AD progression is reflected through the mask reconstruction error of our SSL strategy. The results demonstrate that our model can effectively capture spatiotemporal and pathological features, and our work provides a novel and relevant framework for the early diagnosis of AD based on fMRI.**

I. Introduction

ALZHEIMER'S disease (AD) is an irreversible neurodegenerative disease in which patients exhibit noticeable memory, language, and executive impairment that affect daily life, resulting in significant medical burdens [1], [2]. Mild cognitive impairment (MCI) is often regarded as the early stage of AD and might progress rapidly to dementia [3]. Early detection and intervention at the preclinical AD are crucial for delaying the progression of pathology and cognitive decline [4], [5]. While positron emission tomography (PET) and cerebrospinal fluid (CSF) offer valuable insights into pathology and clinical diagnosis, the search for noninvasive, cost-effective, and readily accessible biomarkers remains crucial [6]. Currently, functional magnetic resonance imaging (fMRI) is performed a valuable tool to explore and characterize the pathological changes, stages, and abnormal activity with high spatiotemporal resolution by measuring the blood oxygen level dependent (BOLD) in a non-invasive way [7], [8]. Since AD is a neurological disconnection disorder, abnormal connections can be observed in functional image as well as locally and globally [9], [10]. According to the anatomical structure of the brain, extracting BOLD signals of specific regions from fMRI to construct brain network can model the brain in the view of neural signal transmission, so as to study the neurodegeneration and spread pattern of AD. The graph method that can represent the spatiotemporal correlation of the brain network has become the main method for the analysis of fMRI [11], [12]. Excitingly, graph neural network (GNN) is widely used in brain networks because of its ability to extract high-level topological features from irregular and unordered graph [13], [14]. As a result, many studies have constructed static and spatiotemporal models based on GNN to study AD through computing of spatial and temporal convolution to learn pathological features [15].

However, due to the complex spatiotemporal variation and low signal-to-noise ratio of fMRI, learning effective representations that can capture robust pathological information

This work was supported by the National Natural Science Foundation of China (No. 62376150), the Science and Technology Innovation 2030 Major Projects (2022ZD0211606), and Shanghai Industrial Collaborative Innovation Project (No. XTCX-KJ-2023-37).

Shuoyan Zhang is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China.

Qingmin Wang, Jiayi Zhong, Ziyan Song, Chenyang Li and Jiehui Jiang are with the Institute of Biomedical Engineering, School of Life Sciences, Shanghai University, Shanghai 200444, China.

Min Wei and Ying Han are with the Department of Neurology, Xuanwu Hospital of Capital Medical University, Beijing 100053, China.

Ying Zhang is with the School of Medicine, Shanghai University, Shanghai 200444, China.

Xiaochen Zhang and Yunxia Li are with the Department of Neurology, Shanghai Pudong Hospital, Fudan University Pudong Medical Center, Shanghai 201399, China.

Han Lv is with the Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China.

Shuoyan Zhang, Qingmin Wang, and Min Wei contributed equally to this work.

Corresponding authors: Jiehui Jiang (e-mail: jiangjiehui@shu.edu.cn); Han Lv (e-mail: chrislvhan@126.com)

and biomarkers becomes difficult [16], [17]. Some researchers introduced self-supervised learning (SSL) into the fMRI data [18], [19]. By setting pseudo-label and pretraining task, SSL helps the model extract rich semantic information from the data itself, so as to achieve robust and generalized results [20]. Although SSL can help GNN models extract effective features from fMRI, there are still two challenges in this issue: **(1) Overlooking subtle cross-spatiotemporal interaction.** Brain activity has many activation cycles, and there is a delay in the transmission of nerve signals between brain regions, which makes fMRI have a short cross-spatiotemporal interaction. Under the noise disturbance, the model will easily ignore the subtle spatiotemporal change. **(2) Underfitting of pathological features with low selectivity constraint.** The BOLD signal of the diseased brain regions and the normal regions showed different modes, but the low selectivity of pretraining constraints could not focus on the abnormal brains with individual heterogeneity, resulting in the underfitting of pathological information in the model.

To overcome above challenges, the present study proposes a **f**ine-g**r**ained spati**o**temporal **G**NN (FROG) with SSL guidance for fMRI analysis of early AD diagnosis. Using the spatial correlation and temporal repeability of fMRI characteristics, we design two specific masks for SSL to enhance cross-spatiotemporal interaction. Afterwards, a spatiotemporal autoencoder is proposed to fit subtle spatiotemporal variation and suppress noise through mask reconstruction, in which temporal gated inception convolution and graph scalable inception convolution are designed to learn deep topological structure. Further, according to the differences in BOLD signals between diseased and normal brain regions, a spatiotemporal scalable cosine error (STSCE) constrained the SSL to focus on the signal pattern of abnormal regions and weaken the influence of normal regions, thus guiding the model to capture fine-grained pathological features in individual level with unsupervised setting. Finally, under the guidance of SSL, a multi-scale readout is used to acquire robust spatiotemporal features and finetune the model to make it suitable for early AD diagnosis and biomarker extraction. By validating the model across multiple cohorts, we intend to develop a more precise and effective tool for the early diagnosis of AD, thereby improving prognosis. Additionally, we also aim to advance previous deep learning researches and explore beyond data-driven methods to expound potential pathological mechanisms responsible for the model's robust diagnostic performance.

The contributions of our studies are as follows:

1) To solve the problem that GNN model ignores subtle cross-spatiotemporal variation in fMRI, a reconstructive SSL strategy is proposed to enhance spatiotemporal interaction and achieve noise-suppressed feature extraction.

2) For underfitting problems with low selectivity constraint, a STSCE loss for spatiotemporal graph convolution is designed to amplify the abnormal BOLD signal and guide the model to focus on fine-grained pathological features.

3) The study intent to propose an optimal model for the diagnosis of MCI and explore AD pathology to further explain the pathological mechanisms responsible for the robust diagnostic performance of this fMRI radiomic model.

II. Related Works

A. AD Diagnosis with GNN

In order to extract the deep topological features of functional brain network, some researches have introduced GNN into fMRI analysis for AD diagnosis. Zhou et al. [21] extracted features based on local attention and global attention. Han et al. [22] used a multi-layer graph convolutional model for Early AD classification. Considering the dynamic variation of fMRI, some studies have introduced the temporal module into GNN to learn the spatiotemporal information. He et al. [23] used the temporal multi-head attention and spatial multi-head attention to capture the spatiotemporal dependence. An et al. [24] proposed a dynamic spatiotemporal GNN which includes temporal block, spatial block and pooling block. Zhang et al. [25] proposed a sparse attention-based node-merging for constructing hierarchical functional brain network to achieve different scale feature extraction. To solve the problem of asynchronous dependence of BOLD signal, Zhang et al. [26] proposed a sliding window method based on attention mechanism to fit the spatiotemporal information of functional connection. However, the spatiotemporal variation of fMRI is implicit and noisy, and the conventional GNN model is not capable of learning robust features and accurately locating abnormal brain regions solely driven by the AD classification task.

B. SSL for fMRI

Due to the advantages of feature effectiveness and robustness brought by SSL, this technology has begun to be introduced into fMRI research. Wang et al. [27] proposed a GNN model with contrastive learning. The brain network views of the same subject are close to each other, and different subject are far away from each other. Peng et al. [28] designed fMRI transformations, two samples are trained by consistency regularization. These models are built on the population graph, it is impossible to extract individual abnormal feature. To solve this problem, some researches apply SSL technology to individual graph. Wen et al. [29] performed the SSL on the reconstruction of nodes and edges on the brain network to enhance of the topology-aware of GNN. Zhang et al. [30] constructed the edge-dropped graph by Bernoulli mask and extracted invariant features by contrastive SSL. Yang et al. [31] randomly droped some timepoints of BOLD signal for constructing pseudo functional connectivity, and enhanced the spatial features using the latent representation alignment. However, the above studies only focused on enhancing functional connectivity without taking into account the spatiotemporal features. Wang et al. [32] proposed a contrastive SSL model based on spatiotemporal GNN, and extracted spatiotemporal robustness features by contrastive learning of fMRI subsequences. The contrastive constraint in the subsequence can only extract multi-view invariant features, which is not in line with the dynamic variations of fMRI.

C. SSL for Time Series

The preprocessed fMRI can be regarded as a time series, where each brain region corresponds to a BOLD signal.

TABLE I
DESCRIPTION OF THE NOTATION.

Notation	Description
G	The brain network of a sample
V	Node set of the G, v is element in V
E	Edge set of the G, e is element in E
$\mathbf{X} \in \mathbb{R}^{\lvert V\rvert \times T}$	Preprocessed BOLD signal
T	Sequence length of the $\mathbf{X}$, t is index of T
$\mathbf{A} \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Adjacency matrix
V^M	Node set covered by mask, $\lvert V^M\rvert = 20$
T^M	Interval length covered by mask, $T^M = 30$
$\mathbf{M} \in \{0,1\}^{\lvert V\rvert \times T}$	Mask for $\mathbf{X}$ transformation
$\mathbf{X}_{mask} \in \mathbb{R}^{\lvert V\rvert \times T}$	Masked BOLD signal
U	Masked times of two types, u is index
$\mathbf{X}_{v,:} \in \mathbb{R}^{\lvert V^M\rvert \times T}$	Signal of all time steps of v-th node
$\mathbf{M}^u_{v,:} \in \mathbb{R}^{\lvert V^M\rvert \times T}$	The u-th mask of all time steps of v-th node
$\mathbf{Y}^u_{v,:} \in \mathbb{R}^{\lvert V^M\rvert \times T}$	Reconstruction of all time steps of v-th node
$\mathbf{X}_{v,t:t+T^M} \in \mathbb{R}^{\lvert V\rvert \times T^M}$	Signal of interval steps of v-th node
$\mathbf{M}^u_{v,t:t+T^M} \in \mathbb{R}^{\lvert V\rvert \times T^M}$	The u-th mask of interval steps of v-th node
$\mathbf{Y}^u_{v,t:t+T^M} \in \mathbb{R}^{\lvert V\rvert \times T^M}$	Reconstruction of interval steps of v-th node
L	Number of spatiotemporal units, l is index
C	Number of feature channel, c is index
$\mathbf{H}^l \in \mathbb{R}^{C \times \lvert V\rvert \times T}$	Feature map in the l-th spatiotemporal unit
$\mathbf{W}^{m,l}_k \in \mathbb{R}^{C \times 1 \times k}$	Convolutional kernel of memory gate
$\mathbf{W}^{f,l}_k \in \mathbb{R}^{C \times 1 \times k}$	Convolutional kernel of forgetting gate
$k \in \{3,5,7,9\}$	Convolutional kernel size
$\mathbf{W}^l_b \in \mathbb{R}^{C \times C}$	Parameter matrix of graph convolution
$\breve{\mathbf{A}}_b \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Normalized adjacency matrix
$b \in \{0,1,2\}$	Adjacency relationship
$\mathbf{Y} \in \mathbb{R}^{\lvert V\rvert \times T}$	The reconstructed signal
$\mathbf{W}^l_o \in \mathbb{R}^{C \times 1 \times 1}$	Convolutional kernel of readout
$\hat{\mathbf{Y}} \in \mathbb{R}^{1 \times \{2,3\}}$	Prediction probability

TABLE II
DEMOGRAPHIC AND CLINICAL INFORMATION OF MULTI-COHORTS.

Cohort	Group	Gender	Age	Education	MMSE
CoRR	CN (3585)	1816/1769	27.3±16.9	-	-
ADNI	CN (579)	345/234	71.0±6.4	16.8±2.2	29.1±1.3
	MCI (476)	211/265^a	71.9±7.3	15.9±2.7^a	27.6±2.1^a
	AD (222)	95/127^b	73.8±7.4^b	15.6±2.7^b	21.8±3.7^b
Xuanwu	CN (389)	217/172	65.1±8.8	12.2±5.7	28.5±1.7
	MCI (167)	95/72	69.0±9.2^a	10.6±4.5^a	24.3±3.6^a
	AD (56)	34/22	72.2±9.7^b	10.3±5.0	17.7±5.0^b
Tongji	CN (71)	31/40	71.0±8.0	12.0±3.7	27.0±2.2
	MCI (85)	47/38	71.6±8.1	10.9±4.3	23.8±3.3^a
	AD (57)	21/36	73.7±8.3	8.3±5.4^b	14.7±6.2^b

Therefore, SSL of time series can provide some inspiration. These studies can roughly be divided into contrastive-based and reconstruction-based pretraining strategies. Eldele et al. [33] obtained samples from two views through strong augmentation and weak augmentation, and learned discriminative representations through temporal contrasting and contextual contrasting. Yue et al. [34] employed hierarchical contrastive learning and contextual consistency to extract multi-scale and diverse representations. Cheng et al. [35] adopted a sliding window to segment sequences and added masks to generate augmented samples, optimizing the contrastive loss through masked codeword classification and masked representations regression. Zerveas et al. [36] proposed a reconstructive strategy, which covers the input signal with a mask, and predicts the masked signal through the encoder. Zhang et al. [37] utilized the fast Fourier transform to get the frequency domain of the time series, and then dropped some patches in both the time domain and the frequency domain. Representation is learned through cross-domain reconstruction. However, the idea of contrastive-based method is to extract multi-view invariant features, which does not align with the dynamic changes of fMRI. While reconstruction-based methods can avoid this issue, existing strategies have not designed pretraining tasks that capture spatiotemporal interactions. Therefore, it is still difficult to effectively learn the representation of fMRI.

III. MATERIALS AND METHODS

A. Data and Preprocessing

We used data from multi-cohorts to validate models, from the Alzheimer's Disease Neuroimaging Initiative dataset (ADNI) (http://adni.loni.usc.edu/), Sino Longitudinal Study on Cognitive Decline (SILCODE) project in Xuanwu Hospital of Capital Medical University (Xuanwu) [38], Tongji Hospital of Tongji University (Tongji), and Consortium for Reliability and Reproducibility (CoRR) [39]. There was a total of 5687 samples in the four cohorts, including 4624 with cognitive normal (CN), 728 with MCI, and 335 with AD. Demographic information includes gender, age, and education. Neuropsychological scale is minimum mental state examination (MMSE).

For ADNI and Xuanwu cohorts, the inclusion criteria refer to previous works [40] and [41]. In Tongji cohort, the diagnostic criteria refer to [42], [43]. The CoRR cohort consisted of neurotypical subjects, which were used for pretraining of FROG model. Since the pretraining did not require strict quality control of the data, we only excluded the missing data, leaving 3585 samples. Demographic and clinical information were shown in Table II. In the Group column, the (.) indicates the number of samples. In the Age, Education, and MMSE columns, we calculated the mean±standard deviation. In the Gender column, we used female/male to represent the number of samples respectively. We used Chi-square test for gender, and T-test for age, education, and MMSE. Superscript a indicates significant differences between CN and MCI groups ($P < 0.01$), and superscript b denotes significant differences between CN and AD groups ($P < 0.01$). We found that the control group (CN) and the disease groups (MCI, AD) differed in several variables, particularly the cognitive scale (MMSE), suggesting that our data were statistically significant.

T1-weighted imaging (T1WI) and fMRI were included for each subject. The Data Processing Assistant for Resting-State fMRI (DPARSF) [44], [45] was used to preprocess fMRI. The first ten frames of fMRI were removed, and the remaining frames carry out slicing time correction and head motion correction. Subjects with head motion exceeding the threshold of 3mm and 1° were excluded. fMRI was registered

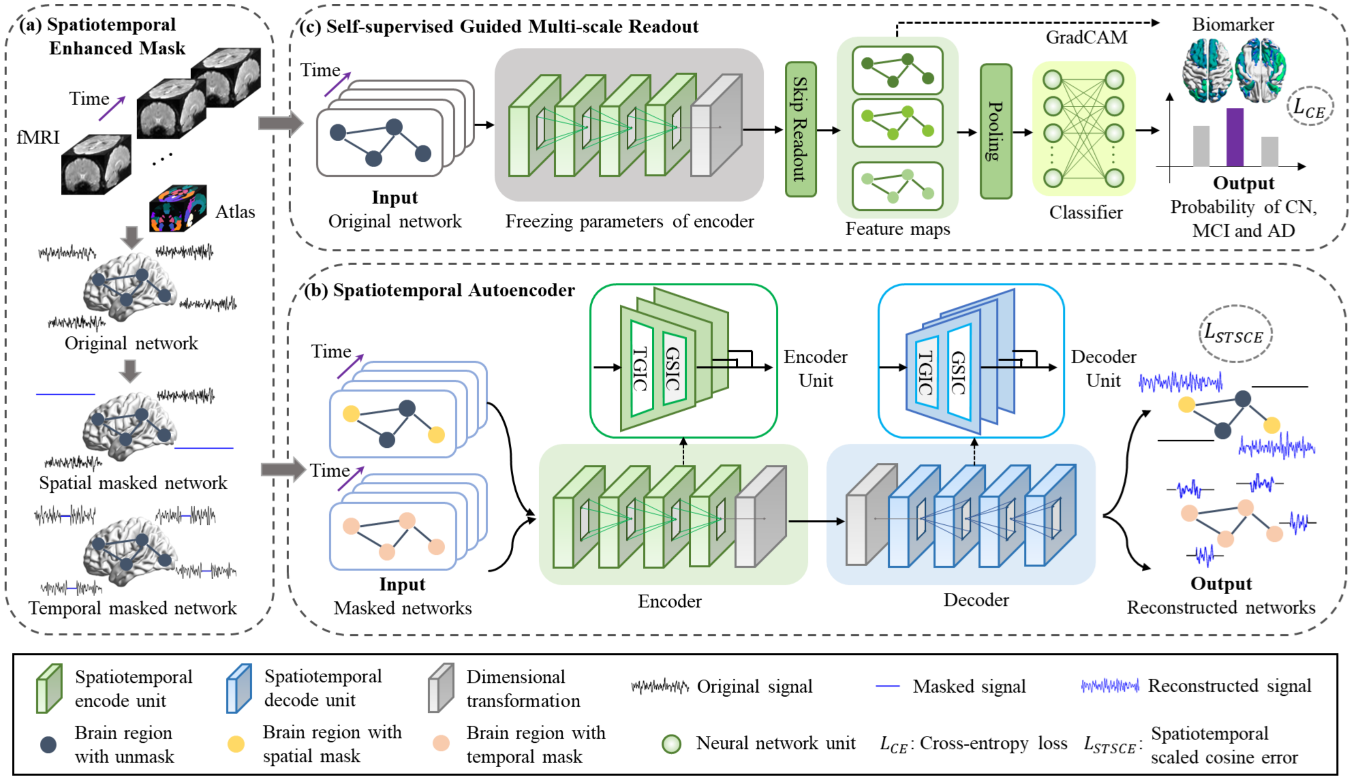


Fig. 1. The framework of the proposed FROG. The model consists of three parts, which are spatiotemporal enhanced mask, spatiotemporal autoencoder, and self-supervised guided multi-scale readout. TGIC and GSIC denote temporal gated inception convolution and graph scalable inception convolution respectively.

to T1WI and then normalized to the Montreal Neurological Institute (MNI) space with 3mm×3mm×3mm. Temporal filtering (0.01-0.08Hz) and spatial smoothing (6mm Gaussian kernel) were performed on the images. Finally, The AAL atlas [46] is used to extract the BOLD signal and exclude the cerebellar regions, so the number of brain region is 90. The preprocessed BOLD signal can be regarded as a multivariable sequence. The sequence length of Xuanwu and Tongji cohorts is 229 and 230, respectively. For ADNI cohort, which includes ADNI2 and ADNI3, the sequence length is 130 and 187, respectively. If the experiment is carried out in a multi-cohort combination, we pad the sequence length to 230 in a repetitive manner.

B. Model

1) Overview of FROG: GNN is a deep learning model computed on a graph, we need to first build a brain network based on fMRI data. A brain network of subject could be represented by $G = (V, E)$, where V denotes a node set and E denotes an edge set, $|V| = 90$, $|E| = 8100$. The i-th node is represented as v_i. The relationship between the i-th node and the j-th node is denoted as $e_{ij} = (v_i, v_j)$. The BOLD signal in the brain regions is regarded as a nodal feature and denoted as $\mathbf{X} \in \mathbb{R}^{|V|\times T}$. T denotes sequence length of the BOLD signal. The weight of edge e_{ij} obtained by Pearson correlation coefficient. The edges of a subject's brain network are represented by an adjacency matrix of the form $\mathbf{A} \in \mathbb{R}^{|V|\times|V|}$. Parameters are defined in Table I.

We introduce the overall calculation flow of the FROG model. As shown in Fig. 1, the model consists of three parts: **(a) spatiotemporal enhanced mask.** The original brain network is constructed from the preprocessed fMRI, and the spatial masked network and temporal masked network are obtained by mask transformation. **(b) spatiotemporal autoencoder.** It is used to reconstruct spatial and temporal masks to fit the spatiotemporal interaction of fMRI, extract noise suppression features, and focus on abnormal regions. **(c) self-supervised guided multi-scale readout.** A pretrained autoencoder with the encoder parameters frozen is finetuned to MCI classification. Below, we describe the details of each part separately.

2) Spatiotemporal Enhanced Mask: Different from traditional image reconstruction, BOLD signal can be regarded as a multivariate time series, there are many methods for time series data transformation, such as adding Gaussian noise, time warping, adding drift, etc., but these methods will destroy the spatiotemporal interaction of the BOLD signal.

To solve this problem, we propose two specific masks for BOLD signal transformation, the spatial mask and temporal mask. The spatial mask is to randomly select some nodes and cover the whole sequence of BOLD signal. The temporal mask is a BOLD signal that is randomly selected for a continuous period of time and covered with a mask. V^M represents the node set covered by the spatial mask and T^M is the interval length of continuous time point covered by the temporal mask. To simplify notation, we uniformly denote the mask as $\mathbf{M} \in \{0,1\}^{|V|\times T}$. Within the mask range, the element value is 0, and vice versa is 1. The calculation of the spatial and temporal mask transformations of the BOLD signal is shown in (1). The elements in $\mathbf{M}_{v\in V^M,:}$ and $\mathbf{M}_{v\in V,t:t+T^M}$ are 0. $\odot$ denotes Hadamard product. If the BOLD signals in brain network are not covered by the mask, we call it an **original network**. If the BOLD signals are covered by the spatial mask or the temporal mask, we call it a **spatial masked network** or a **temporal masked network**.

$$\begin{aligned}\mathbf{X}_{mask} =&\{\mathbf{M}_{v\in V^M,:}\odot\mathbf{X},\mathbf{M}_{v\in V,t:t+T^M}\odot\mathbf{X}\},\\ &0<|V^M|\le|V|,0<t+T^M\le T\end{aligned} \tag{1}$$

For the spatial mask, we assume that the model can recover the masked signal according to the topological relationship of the brain network. For temporal mask, because fMRI is periodic, the masked time period can be reconstructed according to the time series before and after. Conversely, the signals that are covered, forming the reconstructed networks. The spatiotemporal relationship of BOLD signal can be enhanced in fMRI after mask processing.

3) Spatiotemporal Autoencoder: In SSL, the reconstruction pretraining task can achieve signal denoising and fit subtle variation [47]. Thus, we design a spatiotemporal autoencoder based on reconstruction task to learn fine-grained and robust representation. The masked network as the input, spatiotemporal autoencoder outputs reconstructed networks. Inspired by [48], spatiotemporal autoencoder consists of spatiotemporal units which include temporal gated inception convolution and graph scalable inception convolution.

Temporal gated inception convolution is proposed to extract multi-temporal dependency relationship in time points. We use four convolution kernels, 1×3, 1×5, 1×7, 1×9 in the time dimension. As shown in (2), represents the convolution operator, $\mathbf{H}^{l-1}$ represents the feature map, in particular $\mathbf{H}^0=\mathbf{X}_{mask}$. And $\mathbf{W}_k^l$ denotes the parameters of the temporal convolution, and $K=\{3,5,7,9\}$ is the set of kernel size. In order to suppress the temporal noise, we utilize a gating mechanism to filter the features. The feature maps of the multi-temporal kernel are concatenated to obtain the temporal embedding $\mathbf{H}_{TE}^l$.

$$\mathbf{H}_{TE}^l=||_{k\in K}tanh(\mathbf{W}_k^{m,l}*\mathbf{H}^{l-1})\odot\sigma(\mathbf{W}_k^{f,l}*\mathbf{H}^{l-1}) \tag{2}$$

$$\mathbf{H}_{SE}^l=||_{b\in B}relu(\breve{\mathbf{A}}_b\mathbf{H}^{l-1}\mathbf{W}_b^l) \tag{3}$$

Graph scalable inception convolution is designed to learn multi-spatial topology feature in brain regions. We choose three spatial relationships, which are self-loop, 1-hop, 2-hop, because too deep neighboring nodes will lead to too smooth node embeddings. Correspondingly, $B=\{0,1,2\}$ represents the hops in the spatial relationships, and $\breve{\mathbf{A}}_0=\mathbf{I}$, $\breve{\mathbf{A}}_1=\hat{\mathbf{A}}$, $\breve{\mathbf{A}}_2=\hat{\mathbf{A}}\hat{\mathbf{A}}$. Where $\hat{\mathbf{A}}=\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$, $\tilde{\mathbf{A}}=(\mathbf{A}+\mathbf{I})$, $\tilde{\mathbf{D}}_{ii}=\sum_j\tilde{\mathbf{A}}_{ij}$. The calculation of graph scalable inception convolution is shown in (3), $\bar{\mathbf{A}}_b$ and $\mathbf{W}_b^l$ form the graph convolution kernel, and the topological features can be extracted by matrix multiplication. The spatial embedding of the three relationships are concatenated to get $\mathbf{H}_{SE}^l$ [49]. Based on temporal gated inception convolution and graph scalable inception convolution, our spatiotemporal unit as shown in (4). l indicates the index of the spatiotemporal unit. Unlike the encoder, $*$ in the decoder is the transposed convolution for (2).

$$\begin{aligned}\mathbf{H}^l=&||_{b\in B}relu(\breve{\mathbf{A}}_b||_{k\in K}tanh(\mathbf{W}_k^{m,l}*\mathbf{H}^{l-1})\\ &\odot\sigma(\mathbf{W}_k^{f,l}*\mathbf{H}^{l-1})\mathbf{W}_b^l)\end{aligned} \tag{4}$$

Traditional reconstruction task often uses mean squared error (MSE) or mean absolute error (MAE) as a loss function, but MSE and MAE have problems with sensitivity and low selectivity [50]. To solve these difficulties, we propose STSCE, as shown in (5)(6)(7). The '$-$' on the $\mathbf{M}$ indicates the bitwise inversion. $\mathbf{Y}$ represents the reconstructed signals. γ is the scaling factor. The spatial mask and temporal mask are reconstructed. U is the transformation times of two type masks. The loss function only calculates the range covered by the mask, because the signal without mask coverage is easy to learn, resulting in the model is not focused into the mask range, and the abnormal signal is not amplified.

$$\begin{aligned}L_{SSCE}=&\sum_{u=1}^{U}\sum_{v\in V^M}(1-\frac{(\overline{\mathbf{M}}_{v,:}^u\odot\mathbf{Y}_{v,:}^u)^\mathsf{T}(\overline{\mathbf{M}}_{v,:}^u\odot\mathbf{X}_{v,:})}{|\overline{\mathbf{M}}_{v,:}^u\odot\mathbf{Y}_{v,:}^u|\cdot|\overline{\mathbf{M}}_{v,:}^u\odot\mathbf{X}_{v,:}|})^\gamma,\\ &\gamma\ge 1\end{aligned} \tag{5}$$

$$\begin{aligned}L_{TSCE}=&\sum_{u=1}^{U}\sum_{v\in V}(1-(\overline{\mathbf{M}}_{v,t:t+T^M}^u\odot\mathbf{Y}_{v,t:t+T^M}^u)^\mathsf{T}\\ &(\overline{\mathbf{M}}_{v,t:t+T^M}^u\odot\mathbf{X}_{v,t:t+T^M})\\ &/(|\overline{\mathbf{M}}_{v,t:t+T^M}^u\odot\mathbf{Y}_{v,t:t+T^M}^u|\\ &\cdot|\overline{\mathbf{M}}_{v,t:t+T^M}^u\odot\mathbf{X}_{v,t:t+T^M}|))^\gamma,\gamma\ge 1\end{aligned} \tag{6}$$

$$L_{STSCE}=\frac{L_{SSCE}+L_{TSCE}}{2U} \tag{7}$$

STSCE gives us the freedom to set the range of masks, such as spatial masks and temporal masks, which get rid of the influence of mask dimension. Furthermore, STSCE can adaptively paying more attention to abnormal brain regions that are harder to learn through scaling factor. Therefore, the STSCE guides the model to capture the pathological features in an unsupervised manner.

4) Self-supervised Guided Muti-scale Readout: After the reconstruction task is completed, we freeze the parameters of the encoder, input the original network, and the model could still retain fine-grained pathological features in pretraining phase. Meanwhile, each spatiotemporal unit represents a feature at a scale, and multi-scale features can further improve the generalization of the model. From this, we propose a self-supervised guided multi-scale readout for model finetuning.

$$\mathbf{Z}=||_{v=1}^{|V|}pool(\sum_{l=1}^{L}relu(\mathbf{W}_{v,o}^l*\mathbf{H}_v^l)) \tag{8}$$

Multi-scale feature extraction and dimension transformation are shown in (8), $\mathbf{H}_v^l$ is multi-scale feature map of spatiotemporal unit. $2L$ represents the number of total spatiotemporal units, and we take the feature map of the previous L units for readout. The feature map of nodes is operated by pooling layer and concatenation. $\mathbf{Z}$ is the whole graph representation of each individual, which is used for MCI classification. And multilayer perceptron as classifier, cross-entropy loss is used for optimization.

To identify the brain regions with abnormal activity for MCI, we adopt gradient-weighted class activation mapping (Grad-CAM) [51], [52] to find them. We set the Grad-CAM gradient calculation after the multi-scale feature transformation and before the mean pooling layer. The calculation steps of Grad-CAM suitable for the spatiotemporal GNN are shown in (9)(10)(11).

$$\boldsymbol{\alpha}^l = \frac{1}{|V|T}\sum_{v=1}^{|V|}\sum_{t=1}^{T}\frac{\partial\hat{\mathbf{Y}}}{\partial\mathbf{H}^l_{v,t}} \tag{9}$$

$$\mathbf{F} = relu(\sum_{c=1}^{C}\boldsymbol{\alpha}^{l,c}\mathbf{H}^{l,c}) \tag{10}$$

$$\varphi_v = \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_{v,t} \tag{11}$$

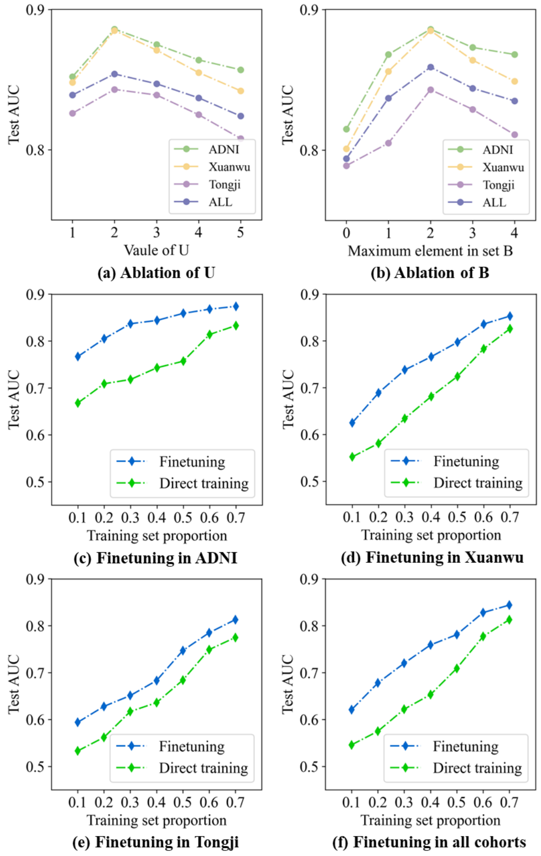
Firstly, the gradient value $\boldsymbol{\alpha}$ on each feature map of the specified class is calculated. Secondly, the weighted sum of the gradient values and the feature maps is filtered by $relu$ to remove the negative values, and obtain the activation map $\mathbf{F}$. Thirdly, the average value of the time dimension is taken to get the weight value of the node level, which is used as the basis for locating the abnormal brain region. φ_v^c is the brain region weight, which is regard as neuroimaging biomarker.

IV. Experiments and Results

A. Experimental Settings

This study pretraining phase using the CoRR cohort. The remaining ADNI, Xuanwu, and Tongji cohorts were used to finetune and verify the performance of the models on AD. First, we verified that SSL improved the classification of the model on multi-cohorts. Then, ablation experiments were carried out to test the effectiveness of each module of the FROG model. Next, we conducted statistical analysis based on biomarkers. Finally, the semantic information learned from fMRI data by SSL was validated under unsupervised settings. The FROG model was implemented using PyTorch [53], the devices were NVIDIA GeForce RTX 4060 Ti and Intel(R) Core(TM) i5-14600K. The operating system was Ubuntu. The code is publicly accessible, the link is https://github.com/syzhangbme/FROG.

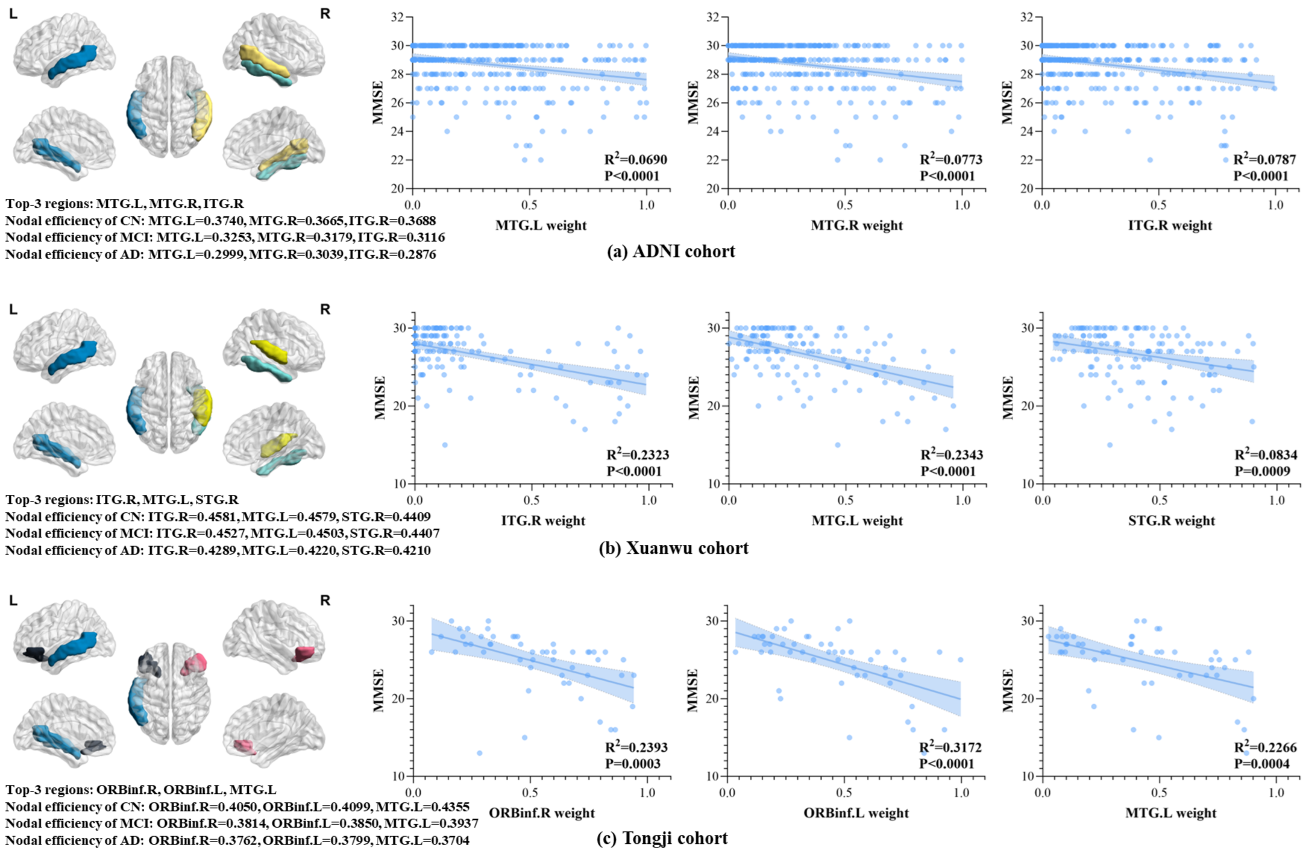
To verify the superiority of the FROG, the state-of-the-art (SOTA) models and SSLs were compared in this study. We selected the AD diagnostic model based on spatiotemporal graph convolution, which are spatiotemporal graph transformer network (STGTN) [23], dynamic spatiotemporal graph neural network (DSGNN) [24], hierarchical functional brain network (HFBN) [25], asynchronous common and individual functional brain networks (ACI-FBN) [26] respectively. And four latest fMRI SSL models, namely brain graph SSL (BrainGSL) [29], adversarial graph contrastive learning (A-GCL) [30], brain network analysis with mask modeling and representation alignment via SSL (BrainMass) [31], and unsupervised contrastive graph learning (UCGL) [32]. Meanwhile, we used SSL strategies that perform well, namely time-series representation learning framework via temporal and contextual contrasting (TS-TCC) [33], time masked autoencoder (TimeMAE) [35]. The advanced feature extractors, they are decomposition Transformers with auto-correlation (Autoformer) [54], inverted Transformer (iTransformer) [55], and modern temporal convolutional network (ModernTCN) [56]. In addition, the methods of adding and deleting modules were used to carry out ablation experiments to verify the effectiveness of each module. For the neuroimaging biomarkers extracted by the FROG, the correlation and mediation effects with the cognitive scale, plasma and PET was analyzed using linear regression.



(a) Ablation of U
(b) Ablation of B
(c) Finetuning in ADNI
(d) Finetuning in Xuanwu
(e) Finetuning in Tongji
(f) Finetuning in all cohorts

Fig. 2. Ablation of U and B, and training set proportion for finetuning.

It is worth noting that TS-TCC and TimeMAE require special settings before being employed in our model. For TS-TCC, we replace 3-block convolutional architecture with our encoder to obtain the latent representation of the BOLD signal. The positive and negative samples are in accordance with TS-TCC. Specifically, in the contextual contrasting module, within one batch, the two augmented views of the same sample are considered as positive samples, while others are regarded as negative samples. The two transformation strategies are strong augmentation and weak augmentation (jitter scale ratio is 1.1, jitter ratio is 0.8, max seg is 8). For TimeMAE, we divide the BOLD signals along the sliding window into multiple non-overlapping subsequences (wave length is 8), and then apply the mask strategy to these subsequences for contrastive learning (mask ratio is set to 0.6). Moreover, in the architecture



(a) ADNI cohort

(b) Xuanwu cohort

(c) Tongji cohort

Fig. 3. Linear correlation between Grad-CAM scores in brain regions and cognitive scales. And changes of nodal efficiency of top 3 brain regions in CN, MCI, and AD. $\boldsymbol{R^2}$ index represented the degree of linear fitness.

TABLE III
PERFORMANCE OF MODELS IN CN AND MCI CLASSIFICATION IN ADNI, XUANWU, AND TONGJI COHORTS.

Model	ADNI				Xuanwu				Tongji				ALL			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
Autoformer	74.2±3.6	74.8±3.3	73.8±7.5	83.4±2.3	76.4±6.5	71.2±8.4	78.7±6.3	83.0±4.5	64.8±4.4	58.8±15.3	71.6±22.7	73.4±2.0	71.4±2.9	71.9±4.6	71.1±7.4	80.2±1.7
iTransformer	74.9±2.4	70.4±4.1	78.6±4.8	83.2±2.4	76.6±6.4	73.6±9.2	77.9±5.8	83.4±4.5	63.4±5.8	65.9±8.6	60.2±18.7	71.5±4.7	72.6±1.1	71.6±6.6	73.2±4.8	80.6±1.6
ModernTCN	75.8±2.3	72.5±3.2	78.6±2.0	83.8±2.3	75.9±6.6	73.6±8.4	76.9±6.2	83.2±4.5	64.7±6.6	67.1±7.1	61.7±17.0	72.0±4.6	72.8±1.0	69.6±6.0	75.1±5.3	80.4±1.7
STGTN	75.4±0.8	75.8±5.0	75.0±3.0	83.6±2.4	73.9±4.8	72.4±11.8	74.6±2.8	82.7±4.5	66.0±8.6	60.0±13.1	73.1±8.5	71.9±4.9	71.9±2.1	72.8±8.6	71.3±8.3	80.9±1.6
DSGNN	77.3±1.9	74.6±2.8	79.5±2.5	85.3±2.3	76.1±5.3	74.8±11.4	76.6±3.5	83.5±4.4	66.6±7.4	67.1±6.0	66.0±16.1	72.1±4.6	72.3±3.8	70.7±3.8	73.4±8.4	80.9±1.6
HFBN	76.8±1.7	73.3±4.5	79.6±1.0	84.4±2.3	74.8±8.5	72.4±7.5	75.9±9.3	83.9±4.4	66.7±8.4	64.7±9.8	68.8±20.2	73.2±3.8	72.7±1.4	73.9±4.9	71.9±4.7	81.1±1.8
ACI-FBN	76.6±1.8	72.9±5.0	79.6±4.6	84.2±2.2	77.2±6.6	75.4±8.6	77.9±6.0	84.1±4.6	67.3±5.3	**77.6±10.1**	54.5±18.7	72.5±3.7	72.2±3.4	74.7±5.2	70.5±8.5	81.5±1.5
BrainGSL	77.5±1.8	76.0±4.7	78.7±3.5	85.8±2.4	76.6±6.8	73.6±9.5	77.9±6.9	84.3±4.7	68.0±4.4	63.5±8.6	73.0±10.7	72.9±4.0	73.7±1.7	69.9±4.4	76.4±5.2	81.5±1.6
A-GCL	78.7±2.3	76.9±6.9	80.1±3.7	86.3±2.4	77.3±6.4	73.6±8.4	78.9±6.1	84.9±4.6	68.6±4.9	67.1±10.3	70.2±21.0	73.5±3.5	73.6±1.3	67.8±6.2	77.7±5.9	80.7±1.5
BrainMass	78.9±2.0	75.6±4.1	81.5±3.4	86.5±2.4	77.7±4.9	74.2±9.1	79.2±5.6	86.7±4.4	69.2±5.7	71.8±11.4	66.3±10.1	73.7±2.7	74.0±3.2	73.0±6.4	74.8±9.7	82.5±1.6
UCGL	76.2±2.6	77.9±3.8	74.8±5.7	84.6±2.3	77.0±4.6	76.0±10.6	77.4±4.4	86.9±4.1	68.6±6.2	69.4±12.0	67.3±21.2	73.4±4.4	73.1±3.3	71.6±8.1	74.2±10.1	80.9±1.6
FROG	**81.1±2.3**	**78.4±5.5**	**83.4±2.2**	**88.6±2.2**	**79.3±4.0**	**79.0±6.5**	**79.4±3.3**	**88.5±3.1**	**76.2±5.8**	70.6±12.3	**83.0±7.4**	**86.9±4.6**	**79.1±2.2**	**76.1±5.2**	**81.1±2.3**	**87.1±2.1**

of TimeMAE, we replace the 1D-CNN with our encoder, which is used to extract the spatiotemporal embeddings of the subsequences, thereby ensuring the integrity of TimeMAE.

In order to obtain the generalized results, we adopt 5-fold cross-validation. 80% of the data were used as the training set, the remaining 20% of the data were regard as the test set, which were performed 5 times in turn. In pretraining and finetuning phases, the learning rate was 0.00001, batch size was 30, L2 coefficient was 0.0001. The scaling factor γ was set to 2 for STSCE loss. The accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC) were utilized as measures of classification. The mean±standard deviation of these measures under the 5-fold cross-validation was taken as the test result.

B. Classification Performance

We used CoRR cohort samples for SSL pretraining, and then froze encoder parameters for MCI classification. We finetuned the model in ADNI, Xuanwu, and Tongji cohorts respectively, as shown in Table III. It can be seen that FROG achieves the highest results in ACC, SEN, SPE, AUC. The results of the three time series models (Autoformer, iTransformer, ModernTCN) were not satisfactory. Although they learned temporal relationships well, the spatial relationships were learned in adaptive ways, such as attention coefficients. However, adaptive spatial relationships might be inconsistent with domain knowledge in brain science. The four spatiotemporal GNN models (STGTN, DSGNN, HFBN, ACI-FBN)

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3552638

TABLE IV
PERFORMANCE OF MCI CLASSIFICATION MODELS TEST IN CN AND AD CLASSIFICATION IN ADNI, XUANWU, AND TONGJI COHORTS.

Model	ADNI				Xuanwu				Tongji				ALL			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
Autoformer	80.5±2.9	76.6±7.5	82.0±4.2	87.4±2.6	84.9±3.9	64.5±15.3	87.9±5.0	83.7±7.2	71.9±5.8	72.3±21.2	71.9±14.8	82.8±4.8	74.3±4.3	66.6±4.2	76.8±5.7	78.5±2.1
iTransformer	80.5±2.9	81.5±4.8	80.1±5.0	87.5±2.7	82.2±4.3	63.0±20.8	85.1±6.4	83.9±7.2	71.9±7.5	68.6±20.0	74.8±15.8	83.8±4.7	74.5±3.6	68.7±5.4	76.4±4.2	79.1±2.2
ModernTCN	80.8±3.3	73.9±9.0	83.4±3.9	87.6±2.8	81.8±3.7	71.7±17.0	83.3±5.4	84.6±7.3	71.8±6.9	72.0±14.4	72.0±14.5	83.6±4.6	74.1±2.9	68.1±5.9	76.0±4.0	79.3±2.1
STGTN	81.2±2.7	79.3±4.9	81.9±3.8	87.9±2.8	81.6±4.2	70.0±18.5	83.3±6.4	84.9±7.4	72.6±6.6	70.5±19.8	74.8±14.4	83.4±5.0	75.8±3.3	66.9±5.7	78.6±4.7	79.8±2.2
DSGNN	82.7±3.6	76.6±1.1	85.0±4.7	87.9±2.7	81.3±3.6	69.8±18.2	83.0±5.7	85.1±7.5	73.4±5.4	73.8±13.8	73.3±13.6	83.4±4.8	75.0±3.1	68.7±8.4	77.1±4.3	80.3±2.2
HFBN	81.0±2.6	79.8±6.2	81.5±4.0	88.3±2.7	83.8±5.3	66.5±17.9	86.4±6.7	85.0±7.6	72.6±7.0	66.7±16.2	77.6±16.2	83.9±4.6	76.6±3.0	66.6±6.1	79.9±3.7	81.6±2.1
ACI-FBN	81.9±3.3	80.2±4.8	82.6±4.5	89.0±2.7	82.7±3.9	71.7±17.0	84.3±6.0	85.2±7.6	73.5±6.5	68.6±20.0	77.6±15.5	84.9±4.6	76.6±2.6	68.1±2.4	79.4±3.0	81.9±1.8
BrainGSL	82.3±2.9	78.4±4.8	83.8±4.2	88.6±2.6	80.4±3.8	71.7±17.0	81.8±5.9	85.3±7.8	73.5±6.1	73.9±17.2	73.4±16.0	85.2±4.8	77.0±1.9	69.9±5.2	79.3±3.2	81.9±1.9
A-GCL	82.5±3.2	78.4±3.6	84.1±5.1	89.1±2.6	82.9±4.0	70.0±18.5	84.8±6.1	85.3±7.8	74.2±6.7	72.3±20.4	76.3±17.3	84.3±4.8	77.7±2.6	67.2±6.2	81.0±4.0	82.1±1.7
BrainMass	82.5±3.5	82.0±6.2	82.7±4.7	90.1±2.9	83.1±3.2	68.3±20.5	85.4±6.7	86.2±7.9	73.4±7.0	73.8±15.0	73.3±15.0	81.9±5.3	77.6±2.7	66.6±3.2	81.1±3.7	82.2±1.7
UCGL	82.2±3.5	83.3±3.7	81.7±5.1	89.7±2.8	82.2±4.5	71.8±17.3	83.8±6.1	85.8±7.9	74.3±6.5	68.6±20.0	79.0±11.7	85.4±4.4	76.6±3.0	71.0±3.2	78.3±3.6	81.9±2.0
FROG	**85.6±3.3**	**85.6±4.2**	**85.7±5.5**	**92.8±2.5**	**86.1±4.4**	**73.3±11.1**	**87.9±5.0**	**86.3±8.0**	**78.2±4.4**	**75.6±14.1**	**80.5±10.8**	**87.3±4.1**	**82.8±3.1**	**78.2±11.6**	**84.3±4.8**	**89.3±2.2**

had a similar structure to ours, but the spatiotemporal feature extraction did not take into account the cross-spatiotemporal dependence of BOLD signal, and multiple stacking of single-kernel temporal and graph convolutions would cause excessive smoothing of spatiotemporal graphs, which is not good for fitting subtle dynamic changes. The SSL GNN models (BrainGSL, A-GCL, BrainMass, UCGL) introduced structure augmentation or multi-view invariant, these self-supervised strategies were designed according to the cofluctuation of BOLD signal in brain regions, which results in incomplete dynamic learning. The fine-grained spatiotemporal features were still not captured. But data augmentation of SSL would also bring some benefits, so compared with spatiotemporal GNN models, they were slightly improved.

Further, we tested the model finetuned by MCI classification in AD classification. We found that our model can still achieve the best results in the AD classification, and the AD classification were higher than MCI. This indicated that there was common pathological information in the MCI group and the AD group, and common pathological information will change with the course of the disease. Detailed results of AD classification can be found in Table IV. Besides, From the tables, we can see that the lower bound of the confidence interval is still better than the results of the SOTA models, indicating that our model is stable.

C. Ablation Study

In order to verify the effectiveness of each part of the FROG model, we conducted ablation experiments. Among them, after removing the mask, the performance of the model will significantly decline. For the two masked methods, temporal mask was more important than spatial mask, which may be because some information about spatial relationship of brain has been reflected in the adjacency matrix. Compared to without mask, local spatial interactions, cross spatiotemporal interactions, and other such information need to be learned through spatial masked network. After the multi-scale readout was removed and only the output features of the last spatiotemporal unit were used, the advantages of multi-scale topological features of the brain network disappeared. This may be because the features extracted from the later spatiotemporal units were not comprehensive. We also introduced common reconstruction loss functions (e.g., MSE and MAE) for comparison, which were not as effective as our STSCE. In addition, we compared two self-supervised strategies (TS-TCC and TimeMAE) for time series, and they got not bad results, but the spatial capture ability is not strong enough to achieve the highest performance. See Table V for detailed results. 'w/o' denoted without, 'w/' denoted with, 'S-mask' meant spatial mask, 'T-mask' was temporal mask.

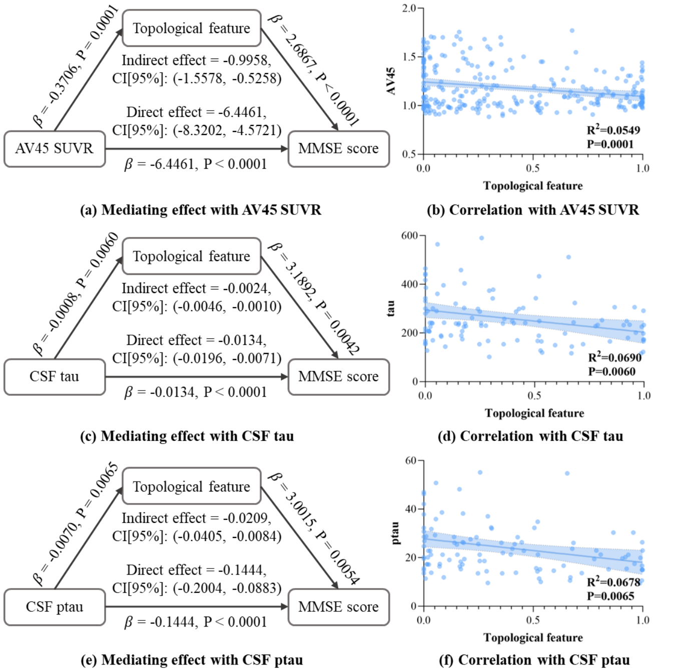
The ablation of U and B were shown in Fig. 2(a)(b). When they were both 2, the results were the best. Too small U will result in insufficient data augmentation, and too large U will also cause overfitting. When the maximum element of B set was 0, the model was equivalent to the combination of temporal convolution and multilayer perceptron, and the ability of spatiotemporal learning was lost, so the results were the worst. When the maximum element of set B was a larger value, the model parameters will increase, and the performance will decrease to some extent. Through SSL, the pretrained model can be corrected with only a part of data to achieve good results in downstream tasks. Therefore, we took a proportional number of samples with group labels from the training set to finetune the MCI classification of the model. As shown in Fig. 2(c)(d)(e)(f), the classification result of the proportion of the training set, we could see that the finetuned model can outperform the model trained directly with labeled data. Further, we observed the changing trend of the curve, the beginning part of the rise was fast, the end part was close to flat. This showed that SSL could be adapted to small sample datasets.

D. Correlation and Mediating Analysis of Biomarker

For the validation of biomarkers, we analyzed the correlation between brain region weights and cognitive scale, as well as the changes of graph theory indicators of brain networks over the course of the AD. Next, we further verified the effectiveness of these brain regions. We took the top 3 brain regions from each cohort, and got the 6 regions, namely left of middle temporal gyrus (MTG.L), right of inferior temporal gyrus (ITG.R), right of middle temporal gyrus (MTG.R), right of superior temporal gyrus (STG.R), right of inferior frontal gyrus, orbital part (ORBinf.R), left of inferior frontal gyrus, orbital part (ORBinf.L). We calculated the correlation between brain region weights and MMSE score, and calculated the changes in nodal efficiency of these brain regions in CN, MCI, and AD groups. As shown in Fig. 3, these brain regions showed a significant negative correlation with MMSE score,

TABLE V
THE MODEL ABLATION OF CN AND MCI CLASSIFICATION IN ADNI, XUANWU, AND TONGJI COHORTS.

Model	ADNI				Xuanwu				Tongji				ALL			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
w/ TS-TCC	79.5±2.0	78.1±4.8	80.7±3.4	87.0±2.4	78.6±3.2	77.2±6.7	79.2±2.8	87.7±3.6	74.4±6.1	72.9±9.6	76.0±10.8	82.0±6.5	74.3±2.8	73.4±6.6	75.0±8.2	83.2±1.6
w/ TimeMAE	79.1±2.3	79.6±5.5	78.8±2.7	87.7±2.3	77.5±3.7	81.4±8.1	75.8±2.4	87.5±3.6	74.3±7.5	74.1±8.8	74.6±13.4	82.1±6.8	74.8±1.0	71.7±5.8	77.0±4.6	82.8±1.4
w/o masks	78.3±1.1	79.4±6.3	77.4±4.8	87.4±2.2	77.0±6.6	76.5±10.9	77.1±6.7	87.0±4.2	72.4±8.7	72.9±10.9	71.7±16.4	81.5±7.4	74.0±3.2	73.0±6.4	74.8±9.7	82.5±1.6
w/ T-mask	80.3±2.0	78.2±6.2	82.0±3.2	88.2±2.3	79.0±4.4	73.6±9.4	81.2±2.6	85.5±4.5	75.0±7.4	72.9±8.8	77.4±10.6	82.2±7.0	77.3±1.9	75.8±4.6	78.3±3.1	85.4±1.8
w/ S-mask	79.7±2.3	80.5±6.5	79.1±1.9	88.0±2.3	77.5±3.0	76.0±9.7	78.2±1.3	86.3±4.2	75.0±7.5	75.3±6.9	74.7±13.2	81.4±6.8	75.3±2.0	75.1±7.1	75.4±8.0	84.8±1.7
w/o readout	79.4±1.6	75.6±5.6	82.6±3.1	87.0±2.3	78.4±5.0	76.0±7.7	79.4±4.5	85.8±4.5	74.3±6.9	76.5±3.7	71.7±14.4	82.5±6.4	76.5±1.7	75.8±3.9	76.9±1.8	85.2±1.8
w/ MSE	79.2±2.0	76.2±6.2	81.7±2.3	86.8±2.4	79.0±3.2	75.4±9.8	80.5±1.4	87.1±4.1	75.6±6.4	74.1±12.1	77.4±8.4	83.1±6.2	76.8±2.3	75.3±8.2	77.9±8.7	85.7±1.9
w/ MAE	79.1±1.7	75.2±5.0	82.4±2.8	86.6±2.4	78.4±4.5	74.2±7.9	80.2±4.8	86.8±4.2	75.0±9.1	76.5±5.3	73.1±15.4	82.1±7.0	75.6±3.5	72.2±7.4	77.9±9.8	84.0±1.6



(a) Mediating effect with AV45 SUVR **(b) Correlation with AV45 SUVR**

(c) Mediating effect with CSF tau **(d) Correlation with CSF tau**

(e) Mediating effect with CSF ptau **(f) Correlation with CSF ptau**

Fig. 4. Mediating effect and linear correlation of topological feature for AV45 SUVR, CSF tau, and CSF ptau. β was the coefficient in the mediation effect, and CI denoted the confidence interval.

and the nodal efficiency gradually decreased with the course of the AD.

Since the multi-scale features can be regarded as topological features of deep functional connections, reflecting the pathological loss of the global brain connection, we adopted t-SNE [57] to reduce the dimensionality of the topological features and conducted mediation analysis. The results of topological feature with 18F-florbetapir standardized uptake value ratio (AV45 SUVR), CSF tau, CSF phosphorylated tau (CSF ptau), and MMSE score were shown in Fig. 4. We observed that there were significant correlations AV45 SUVR (β=-6.4461, P<0.0001), CSF tau (β=-0.0134, P<0.0001), CSF ptau (β=-0.1444, P<0.0001), and the results of MMSE. Interestingly, these correlations were mediated by topological features (Fig. 4(a),(c),(e)), indicating partial mediation effects. Topological feature was correlated with AV45 SUVR (R^2=0.0549, P=0.0001, Fig. 4(b)), CSF tau (R^2=0.0690, P=0.006, Fig. 4(d)), and CSF ptau (R^2=0.0678, P=0.0065, Fig. 4(f)).

E. Analysis of Self-supervised Learning

Diseased brain regions have different signal mode than normal brain regions, the error could reflect differences in signals for the reconstruction task of SSL. We carried out mask reconstruction experiments on multi-cohorts respectively, and the results were shown in Table VI.

TABLE VI
RESULTS OF BOLD SIGNAL RECONSTRUCTION OF CN, MCI, AND AD.

Cohort	Error	CN	MCI	AD
ADNI	MSE	0.1824±0.0200	0.2071±0.0225	0.2421±0.0210
	MAE	0.1229±0.0089	0.1338±0.0080	0.1491±0.0064
Xuanwu	MSE	0.1629±0.0161	0.1893±0.0180	0.2123±0.0213
	MAE	0.1186±0.0062	0.1307±0.0063	0.1413±0.0072
Tongji	MSE	0.1636±0.0185	0.1839±0.0197	0.2208±0.0173
	MAE	0.1165±0.0070	0.1276±0.0068	0.1439±0.0056

We could see that with the progression of the disease, the reconstruction difficulty of the BOLD signal gradually increases. In the same way, we calculated the correlation between reconstruction error and MMSE, it could still see a certain degree of relationship as shown in Fig. 5(a)(c)(e). Based on the differences in the reconstruction errors of CN, MCI, and AD groups found above, we took out the node embeddings from the spatiotemporal autoencoder, and then used t-SNE for dimensionality reduction, and drew the scatter plots of three classes in multi-cohorts. From the Fig. 5(b)(d)(f), we observed that the embeddings of the 90 brain regions in each class can be clearly clustered. This also showed that the spatiotemporal autoencoder can perceive the AD process information.

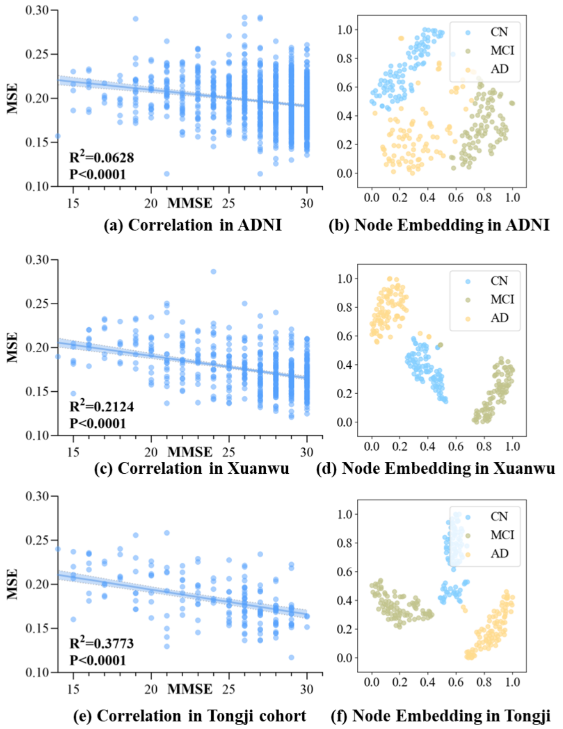
V. DISCUSSION

Given the highly heterogeneous and disabling of AD, exploring an excellent neuroimaging biomarker for accurately recognizing prodromal AD populations exactly is crucial. In this study, we proposed a fine-grained spatiotemporal GNN model with SSL for robust feature learning in fMRI data, applying it to identification and biomarker extraction of early AD. Our work encompassed three aspects. The first highlight of the model, was that on the basis of the fMRI spatial correlation and temporal repeability, two masks were designed to enhance spatiotemporal interaction. Secondly, TGIC and GSIC were proposed to capture subtle fMRI variation and suppress noise in the spatiotemporal autoencoder. In addition, the STSCE to amplify the abnormal BOLD signal and reduced the influence of the normal BOLD signal. Excitingly, with this design, our FROG model could automatically focus on abnormal brain regions in an unsupervised learning manner according to the spatiotemporal differences of signal patterns between diseased and normal brain regions, thus guiding the model to accurately perceive pathological information without relying on any prior knowledge, and adapted well to the individual heterogeneity of AD. After SSL, the model retained the advanced features in the downstream transfer learning. Therefore, more robust neuroimaging biomarker extraction could be achieved in multiple cohorts.

To be more specific, we tested the model on multiple fMRI cohorts of AD, and conducted the classification, model ablation, neuroimaging biomarker extraction, and validated the effectiveness of SSL. To verify the robustness of the model, we pretrained the model using the CoRR cohort with the most data, and finetuned it on ADNI, Xuanwu, and Tongji cohorts. The result showed that our model got the highest result on MCI identification. Further, we tested the model with AD classification, and still achieved good classification results, showing that our model could extract generalized pathological features across cohorts and disease progression.

It is worth noting that even though CoRR is younger than the other cohorts and all subjects are CN, pretraining can still provide a better result. There are two reasons for this. On the one hand, the model can learn the normal feature distribution only through normal samples in the pretraining phase. In the finetuning phase, when abnormal samples are input, the model perceives the difference in feature distribution [58], [59]. Therefore, using normal samples for pretraining can introduce prior knowledge of normal features into the model, so that the model could perceive the abnormal features of the AD, and guide the downstream diagnostic model to accurately focus on the pathological features. On the other hand, SSL has demonstrated the effectiveness of using other datasets for pretraining, even if the data modality is different. The diversity of samples is more important, because SSL improves the performance by providing richer semantic information [60], [61]. Although the age of the CoRR dataset is younger than other AD datasets, finetuning enables the model to eliminate age bias. Like other SSL studies of medical imaging [60], [62], the CoRR dataset comes from multiple centers, the sample diversity fully meets our work needs.

Next, we used Grad-CAM to map gradient values to find salient brain regions, and extracted the weights of the top 3 regions for each cohort and found that they correlated with the MMSE and their nodal efficiency changed significantly with the disease course. These results validated that our neuroimaging biomarker extraction was robust and stable. Moreover, most of these brain regions belong to the sensorimotor network, default mode network, subcortical network [63], [64], which were consistency with prior studies [65]–[67]. Meanwhile, AV45 SUVR, CSF levels of total tau, and ptau play essential roles in both the prodromal of AD stages and disease progression [6], [68]–[70]. We observed that topological features mediate the relationship between CSF biomarkers (tau and ptau) and MMSE score, implying that the effects pathway might encompass various aspects and



(a) Correlation in ADNI (b) Node Embedding in ADNI
(c) Correlation in Xuanwu (d) Node Embedding in Xuanwu
(e) Correlation in Tongji cohort (f) Node Embedding in Tongji

Fig. 5. Linear correlation between MSE and MMSE, and node embeddings of dimension reduction.

dimensions.

To verify the effectiveness of the model design, we performed ablation experiments on each module of the model, and we found that of the two masks, the temporal mask was more important than the spatial mask, because some spatial information was already contained in the adjacency matrix. At the same time, when the model removed the two masks, multi-scale readout, and STSCE, the performance of the model decreased. This showed that every module in our model design was valid. For SSL, we calculated the error of mask reconstruction. As we hypothesized, the difficulty of BOLD signal reconstruction gradually increases as the disease progression. Based on the characteristics of the BOLD signal, we calculated the correlation between the reconstruction error and MMSE. Although the correlation was not as strong as the Grad-CAM, the model has perceived the spatiotemporal variation of the BOLD signal in the unsupervised setting. At the same time, the spatiotemporal autoencoder could also learn the difference features between groups. Therefore, we took out the node embedding from the autoencoder and visualized them, and we could see that the nodes in the same group can gather into a cluster, potentially characterizing differences in pathological states during AD progression to some extent.

Although our proposed FROG model achieved robust and stable classification and imaging biomarker extraction, it still has some shortcomings. On the one hand, we have carried out studies in gray matter, but in recent years, the AD in white matter have been mined, which is the medium of signal transmission between gray matter, and including white matter can more precisely study the spatiotemporal variation of BOLD signal. On the other hand, we only used fMRI single modal data. For a subject, multi-modal data can obtain more accurate and comprehensive diagnosis and analysis results. Especially for AD, fMRI combined with molecular imaging is a valuable and hot research direction. In the future, we will incorporate molecular images and diffusion weighted images to study brain networks, and reveal and understand prodromal AD from different perspectives.

VI. Conclusion

In this study, we proposed the fine-grained spatiotemporal GNN with SSL for diagnosis of early AD. An autoencoder consisting of temporal gated inception convolution and graph scalable inception convolution was used for spatiotemporal feature extraction and signal reconstruction, and an SSL strategy was designed to enhance subtle spatiotemporal interaction and learn pathological features. A total of 5687 samples from four across-population cohorts were involved, along with 11 SOTA models for comparison. Through pretraining in normal samples, our model could perceive the abnormal feature distribution of AD, which guides the classification task to accurately focus on the pathological features and achieve the best performance in multiple cohorts. Next, we carried out a statistical analysis of the neuroimaging biomarkers extracted from the model, including the correlation of cognitive scales and the mediating effects of cognitive biomarkers, and these analyses achieved significant results. This shows that our model indeed captures the pathological features. Further, we verified the effectiveness of SSL, and the model could perceive the AD progression only through mask reconstruction. This work provided a novel and relevant framework for the early diagnosis of AD based on fMRI, and contributed to the precision intervention and personalized diagnosis.

Acknowledgment

We would like to thank the Ying Han's and Yunxia Li's teams for their help in collection and management of neuroimaging and clinic data.

References

[1] P. Scheltens, K. Blennow, M. M. Breteler, B. De Strooper, G. B. Frisoni, S. Salloway, and W. M. Van der Flier, "Alzheimer's disease," *The Lancet*, vol. 388, no. 10043, pp. 505–517, 2016, publisher: Elsevier.

[2] Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, and V. A. Bohr, "Ageing as a risk factor for neurodegenerative disease," *Nature Reviews Neurology*, vol. 15, no. 10, pp. 565–581, 2019, publisher: Nature Publishing Group UK London.

[3] A. Gustavsson, N. Norton, T. Fast, L. Frölich, J. Georges, D. Holzapfel, T. Kirabali, P. Krolak-Salmon, P. M. Rossini, M. T. Ferretti, and others, "Global estimates on the number of persons across the Alzheimer's disease continuum," *Alzheimer's & Dementia*, vol. 19, no. 2, pp. 658–670, 2023, publisher: Wiley Online Library.

[4] S. J. Vos, F. Verhey, L. Frölich, J. Kornhuber, J. Wiltfang, W. Maier, O. Peters, E. Rüther, F. Nobili, S. Morbelli, and others, "Prevalence and prognosis of Alzheimer's disease at the mild cognitive impairment stage," *Brain*, vol. 138, no. 5, pp. 1327–1338, 2015, publisher: Oxford University Press.

[5] L. Parnetti, E. Chipi, N. Salvadori, K. D'Andrea, and P. Eusebi, "Prevalence and risk of progression of preclinical Alzheimer's disease stages: a systematic review and meta-analysis," *Alzheimer's Research & Therapy*, vol. 11, pp. 1–13, 2019, publisher: Springer.

[6] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, and others, "NIA-AA research framework: toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018, publisher: Elsevier.

[7] F. de Vos, M. Koini, T. M. Schouten, S. Seiler, J. van der Grond, A. Lechner, R. Schmidt, M. de Rooij, and S. A. Rombouts, "A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease," *NeuroImage*, vol. 167, pp. 62–72, 2018, publisher: Elsevier.

[8] E. L. Dennis and P. M. Thompson, "Functional brain connectivity using fMRI in aging and Alzheimer's disease," *Neuropsychology Review*, vol. 24, pp. 49–62, 2014, publisher: Springer.

[9] W. W. Seeley, R. K. Crawford, J. Zhou, B. L. Miller, and M. D. Greicius, "Neurodegenerative diseases target large-scale human brain networks," *Neuron*, vol. 62, no. 1, pp. 42–52, 2009, publisher: Elsevier.

[10] J. J. Palop, J. Chin, and L. Mucke, "A network dysfunction perspective on neurodegenerative diseases," *Nature*, vol. 443, no. 7113, pp. 768–773, 2006, publisher: Nature Publishing Group UK London.

[11] 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, publisher: Nature Publishing Group UK London.

[12] H.-J. Park and K. Friston, "Structural and functional brain networks: from connections to cognition," *Science*, vol. 342, no. 6158, p. 1238411, 2013, publisher: American Association for the Advancement of Science.

[13] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: a review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020, publisher: Elsevier.

[14] 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, publisher: IEEE.

[15] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "Graph-based deep learning for medical diagnosis and analysis: past, present and future," *Sensors*, vol. 21, no. 14, p. 4758, 2021, publisher: MDPI.

[16] F. I. Karahanoğlu and D. Van De Ville, "Transient brain activity disentangles fMRI resting-state dynamics in terms of spatially and temporally overlapping networks," *Nature Communications*, vol. 6, no. 1, p. 7751, 2015, publisher: Nature Publishing Group UK London.

[17] T. B. Parrish, D. R. Gitelman, K. S. LaBar, and M.-M. Mesulam, "Impact of signal-to-noise on functional MRI," *Magnetic Resonance in Medicine*, vol. 44, no. 6, pp. 925–932, 2000, publisher: Wiley Online Library.

[18] H. Li, D. Srinivasan, C. Zhuo, Z. Cui, R. E. Gur, R. C. Gur, D. J. Oathes, C. Davatzikos, T. D. Satterthwaite, and Y. Fan, "Computing personalized brain functional networks from fMRI using self-supervised deep learning," *Medical Image Analysis*, vol. 85, p. 102756, 2023, publisher: Elsevier.

[19] A. Thomas, C. Ré, and R. Poldrack, "Self-supervised learning of brain dynamics from broad neuroimaging data," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, vol. 35, New Orleans, USA, Dec. 2022, pp. 21 255–21 269.

[20] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 6, no. 1, p. 74, 2023, publisher: Nature Publishing Group UK London.

[21] Z. Zhou, Q. Wang, X. An, S. Chen, Y. Sun, G. Wang, and G. Yan, "A novel graph neural network method for Alzheimer's disease classification," *Computers in Biology and Medicine*, vol. 180, p. 108869, 2024, publisher: Elsevier.

[22] S. Han, Z. Sun, K. Zhao, F. Duan, C. F. Caiafa, Y. Zhang, and J. Solé-Casals, "Early prediction of dementia using fMRI data with a graph convolutional network approach," *Journal of Neural Engineering*, vol. 21, p. 016013, 2024, publisher: IOP Publishing.

[23] P. He, Z. Shi, Y. Cui, R. Wang, D. Wu, A. D. N. Initiative, and others, "A spatiotemporal graph transformer approach for Alzheimer's disease diagnosis with rs-fMRI," *Computers in Biology and Medicine*, vol. 178, p. 108762, 2024, publisher: Elsevier.

[24] X. An, Y. Zhou, Y. Di, Y. Han, and D. Ming, "A novel method to identify mild cognitive impairment using dynamic spatio-temporal graph neural network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 3328–3337, 2024, publisher: IEEE.

[25] J. Zhang, Y. Guo, L. Zhou, L. Wang, W. Wu, and D. Shen, "Constructing hierarchical attentive functional brain networks for early AD diagnosis," *Medical Image Analysis*, vol. 94, p. 103137, 2024, publisher: Elsevier.

[26] J. Zhang, X. Wu, X. Tang, L. Zhou, L. Wang, W. Wu, and D. Shen, "Asynchronous functional brain network construction with spatiotemporal transformer for MCI classification," *IEEE Transactions on Medical Imaging*, 2024, publisher: IEEE.

[27] X. Wang, L. Yao, I. Rekik, and Y. Zhang, "Contrastive functional connectivity graph learning for population-based fMRI classification," in *Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention*, Singapore, Sep. 2022, pp. 221–230, publish: Springer.

[28] L. Peng, N. Wang, J. Xu, X. Zhu, and X. Li, "GATE: graph CCA for temporal SElf-supervised learning for label-efficient fMRI analysis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 391–402, 2023, publisher: IEEE.

[29] G. Wen, P. Cao, L. Liu, J. Yang, X. Zhang, F. Wang, and O. R. Zaiane, "Graph self-supervised learning with application to brain networks analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 4154–4165, 2023, publisher: IEEE.

[30] S. Zhang, X. Chen, X. Shen, B. Ren, Z. Yu, H. Yang, X. Jiang, D. Shen, Y. Zhou, and X.-Y. Zhang, "A-GCL: adversarial graph contrastive learning for fMRI analysis to diagnose neurodevelopmental disorders," *Medical Image Analysis*, vol. 90, p. 102932, 2023, publisher: Elsevier.

[31] Y. Yang, C. Ye, G. Su, Z. Zhang, Z. Chang, H. Chen, P. Chan, Y. Yu, and T. Ma, "BrainMass: advancing brain network analysis for diagnosis with large-scale self-supervised learning," *IEEE Transactions on Medical Imaging*, vol. 43, no. 11, pp. 4004–4016, 2024, publisher: IEEE.

[32] X. Wang, Y. Chu, Q. Wang, L. Cao, L. Qiao, L. Zhang, and M. Liu, "Unsupervised contrastive graph learning for resting-state functional MRI analysis and brain disorder detection," *Human Brain Mapping*, vol. 44, no. 17, pp. 5672–5692, 2023, publisher: Wiley Online Library.

[33] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, Montreal, Canada, Aug. 2021.

[34] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu, "TS2Vec: towards universal representation of time series," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Virtual Event, Mar. 2022, pp. 8980–8987.

[35] M. Cheng, Q. Liu, Z. Liu, H. Zhang, R. Zhang, and E. Chen, "TimeMAE: self-supervised representations of time series with decoupled masked autoencoders," *arXiv preprint arXiv:2303.00320*, 2023.

[36] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A Transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Virtual Event, Aug. 2021, pp. 2114–2124.

[37] W. Zhang, L. Yang, S. Geng, and S. Hong, "Self-supervised time series representation learning via cross reconstruction Transformer," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 16 129–16 138, 2024, publisher: IEEE.

[38] X. Li, X. Wang, L. Su, X. Hu, and Y. Han, "Sino Longitudinal Study on Cognitive Decline (SILCODE): protocol for a Chinese longitudinal observational study to develop risk prediction models of conversion to mild cognitive impairment in individuals with subjective cognitive decline," *BMJ Open*, vol. 9, no. 7, p. e028188, 2019, publisher: British Medical Journal Publishing Group.

[39] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos, and others, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Scientific Data*, vol. 1, no. 1, pp. 1–13, 2014, publisher: Nature Publishing Group.

[40] J. Jiang, M. Wang, I. Alberts, X. Sun, T. Li, A. Rominger, C. Zuo, Y. Han, K. Shi, and f. t. A. D. N. Initiative, "Using radiomics-based modelling to predict individual progression from mild cognitive impairment to Alzheimer's disease," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, no. 7, pp. 2163–2173, 2022, publisher: Springer.

[41] T.-R. Li, Y. Wu, J.-J. Jiang, H. Lin, C.-L. Han, J.-H. Jiang, and Y. Han, "Radiomics analysis of magnetic resonance imaging facilitates the identification of preclinical Alzheimer's disease: an exploratory study," *Frontiers in Cell and Developmental Biology*, vol. 8, p. 605734, 2020, publisher: Frontiers Media SA.

[42] R. C. Petersen, R. Doody, A. Kurz, R. C. Mohs, J. C. Morris, P. V. Rabins, K. Ritchie, M. Rossor, L. Thal, and B. Winblad, "Current concepts in mild cognitive impairment," *Archives of Neurology*, vol. 58, no. 12, pp. 1985–1992, 2001, publisher: American Medical Association.

[43] B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. DeKosky, S. Gauthier, D. Selkoe, R. Bateman, and others, "Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria," *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014, publisher: Elsevier.

[44] C. Yan and Y. Zang, "DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI," *Frontiers in Systems Neuroscience*, vol. 4, p. 1377, 2010, publisher: Frontiers.

[45] C.-G. Yan, X.-D. Wang, X.-N. Zuo, and Y.-F. Zang, "DPABI: data processing & analysis for (resting-state) brain imaging," *Neuroinformatics*, vol. 14, pp. 339–351, 2016, publisher: Springer.

[46] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002, publisher: Elsevier.

[47] C. Niu, M. Li, F. Fan, W. Wu, X. Guo, Q. Lyu, and G. Wang, "Noise suppression with similarity-based self-supervised deep learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1590–1602, 2022, publisher: IEEE.

[48] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, vol. 31, California USA, Feb. 2017, issue: 1.

[49] F. Frasca, E. Rossi, D. Eynard, B. Chamberlain, M. Bronstein, and F. Monti, "SIGN: scalable inception graph neural networks," in *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, Jul. 2020, graph Representation Learning and Beyond (GRL+) Workshop at the 37th International Conference on Machine Learning.

[50] Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang, "GraphMAE: self-supervised masked graph autoencoders," in *Proceedings of the 28th International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, Aug. 2022, pp. 594–604.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the 16th International Conference on Computer Vision*, Venice, Italy, Oct. 2017, pp. 618–626.

[52] P. Das and A. Ortega, "Gradient-weighted class activation mapping for spatio temporal graph convolutional network," in *Proceedings of the 47th International Conference on Acoustics, Speech, and Signal Processing*, Singapore, May 2022, pp. 4043–4047.

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and others, "Pytorch: an imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, Dec. 2019, pp. 8026–8037.

[54] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: decomposition transformers with auto-correlation for long-term series forecasting," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, vol. 34, Virtual Event, Dec. 2021, pp. 22 419–22 430.

[55] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "Itransformer: inverted transformers are effective for time series forecasting," in *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, May 2024.

[56] D. Luo and X. Wang, "ModernTCN: a modern pure convolution structure for general time series analysis," in *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, May 2024.

[57] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[58] A. Kascenas, P. Sanchez, P. Schrempf, C. Wang, W. Clackett, S. S. Mikhael, J. P. Voisey, K. Goatman, A. Weir, N. Pugeault, and others, "The role of noise in denoising models for anomaly detection in medical images," *Medical Image Analysis*, vol. 90, p. 102963, 2023, publisher: Elsevier.

[59] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "Encoder-decoder contrast for unsupervised anomaly detection in medical images," *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 1102–1112, 2024, publisher: IEEE.

[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 39th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, Jun. 2022, pp. 20 730–20 740.

[61] H. A. Al Kader Hammoud, T. Das, F. Pizzati, P. H. Torr, A. Bibi, and B. Ghanem, "On pretraining data diversity for self-supervised learning,"

in *Proceedings of the 18th European Conference on Computer Vision*. Milan, Italy: Springer, Oct. 2024, pp. 54–71.

[62] H.-Y. Zhou, C. Lu, C. Chen, S. Yang, and Y. Yu, "A unified visual information preservation framework for self-supervised pre-training in medical image analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8020–8035, 2023, publisher: IEEE.

[63] Y. He, J. Wang, L. Wang, Z. J. Chen, C. Yan, H. Yang, H. Tang, C. Zhu, Q. Gong, Y. Zang, and others, "Uncovering intrinsic modular organization of spontaneous brain activity in humans," *PloS One*, vol. 4, no. 4, p. e5226, 2009, publisher: Public Library of Science San Francisco, USA.

[64] M. Xia, J. Wang, and Y. He, "BrainNet Viewer: a network visualization tool for human brain connectomics," *PloS One*, vol. 8, no. 7, p. e68910, 2013, publisher: Public Library of Science San Francisco, USA.

[65] F. Agosta, M. A. Rocca, E. Pagani, M. Absinta, G. Magnani, A. Marcone, M. Falautano, G. Comi, M. L. Gorno-Tempini, and M. Filippi, "Sensorimotor network rewiring in mild cognitive impairment and Alzheimer's disease," *Human Brain Mapping*, vol. 31, no. 4, pp. 515–525, 2010, publisher: Wiley Online Library.

[66] F. Agosta, M. Pievani, C. Geroldi, M. Copetti, G. B. Frisoni, and M. Filippi, "Resting state fMRI in Alzheimer's disease: beyond the default mode network," *Neurobiology of Aging*, vol. 33, no. 8, pp. 1564–1578, 2012, publisher: Elsevier.

[67] J. Li, Y. Gong, and X. Tang, "Hierarchical subcortical sub-regional shape network analysis in Alzheimer's disease," *Neuroscience*, vol. 366, pp. 70–83, 2017, publisher: Elsevier.

[68] V. L. Villemagne, V. Doré, S. C. Burnham, C. L. Masters, and C. C. Rowe, "Imaging tau and amyloid-β proteinopathies in Alzheimer disease and other conditions," *Nature Reviews Neurology*, vol. 14, no. 4, pp. 225–236, 2018, publisher: Nature Publishing Group UK London.

[69] C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010, publisher: Elsevier.

[70] R. Ossenkoppele, B. N. van Berckel, and N. D. Prins, "Amyloid imaging in prodromal Alzheimer's disease," *Alzheimer's Research & Therapy*, vol. 3, pp. 1–3, 2011, publisher: Springer.