This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

# M$^3$D-BFS: a Multi-stage Dynamic Fusion Strategy for Sample-Adaptive Multi-Modal Brain Network Analysis

Rui Dong, Xiaotong Zhang, Jiaxing Li, Yueying Li, Jiayin Wei, Youyong Kong*

School of Computer Science and Engineering, Southeast University

dongrui_0427.seu.edu.cn, zxt_000902@163.com, Jiaxing_Li@seu.edu.cn, yueyli@163.com, 220236174@seu.edu.cn, kongyouyong@seu.edu.cn

## Abstract

*Multi-modal fusion is of great significance in neuroscience which integrates information from different modalities and can achieve better performance than uni-modal methods in downstream tasks. Current multi-modal fusion methods in brain networks, which mainly focus on structural connectivity (SC) and functional connectivity (FC) modalities, are static in nature. They feed different samples into the same model with identical computation, ignoring inherent difference between input samples. This lack of sample adaptation hinders model's further performance. To this end, we innovatively propose a multi-stage dynamic fusion strategy (M$^3$D-BFS) for sample-adaptive multi-modal brain network analysis. Unlike other static fusion methods, we design different mixture-of-experts (MoEs) for uni- and multi-modal representations where modules can adaptively change as input sample changes during inference. To alleviate issue of MoE where training of experts may be collapsed, we divide our method into 3 stages. We first train uni-modal encoders respectively, then pretrain single experts of MoEs before finally finetuning the whole model. A multi-modal disentanglement loss is designed to enhance the final representations. To the best of our knowledge, this is the first work for dynamic fusion for multi-modal brain network analysis. Extensive experiments on different real-world datasets demonstrates the superiority of M$^3$D-BFS. Code can be available at https://github.com/KamonRiderDR/M3D-BFS.*

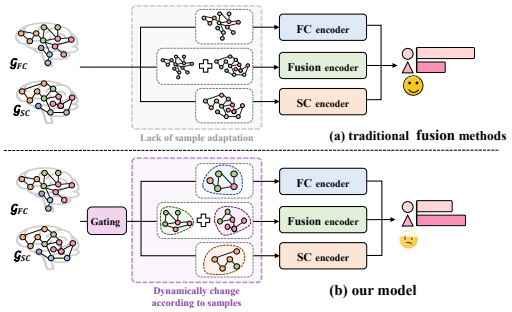



Figure 1. Illustration of M$^3$D-BFS. Compared with traditional methods, M$^3$D-BFS can adaptively adjust for different input samples.

## 1. Introduction

Most current neuroscience studies are based on uni-modal learning, involving only one modality [2, 27, 32]. In brain network analysis, most works focus on functional magnetic resonance imaging (fMRI) [29] and diffusion magnetic resonance imaging (dMRI) [24]. However, in existing real-world medical scenarios, data from SC and FC offer complementary information [8], while single-modality-based methods are insufficient to capture comprehensive useful messages. On the other hand, multi-

*Corresponding author

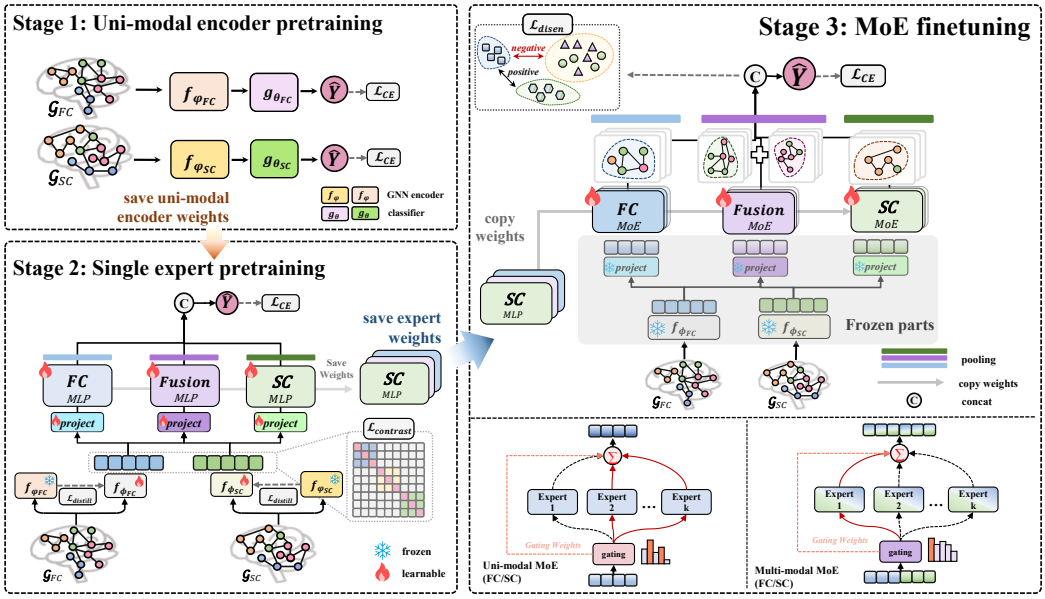



Figure 2. The overall framework of our proposed M$^3$D-BFS. It consists of 3 stages: **(1) stage 1**: training for uni-modal encoders. **(2) stage 2**: pretraining of M$^3$D-BFS by using single expert. **(3) stage 3**: Finetuning M$^3$D-BFS via finetuing MoE blocks. $\mathcal{L}_{disen}$ is designed to enhance final embeddings. Structures of MoEs (uni/multi-modal) are also plotted.

modal fusion, which integrates information from different modalities, can utilize data more efficiently and achieve better performance in downstream tasks, like gender classification and disorder classification.

Multi-modal brain network fusion typically utilizes data from fMRI (functional) and dMRI (structural) modalities. Different from other image-based medical fusion methods, multi-modal brain network fusion mainly considers fusion between different graphs (FC and SC) to better capture topology features from brains. In general, current fusion stategies can be divided into 3 categories: **(a) Representation-level fusion** which directly concats multi-modal representations for fusion. [19, 26, 33]. **(b) Graph-level fusion** between multi-modal brain networks. ALNEGAT [5] designs an attention-based adjacency matrix for fusion. Cross-GNN [30] proproses a graph-based learning framework to capture finer inter-modal dependencies. **(c) transformer-based** fusion methods. RH-BrainFS [31] considers regional heterogeneity between SC-FC and utilized bottleneck transformer for indirect fusion. NeuroPath [25] adds coupling mechanism between FC and SC to improve FC's performance.

Despite the advancements in multi-modal brain network fusion, current methods are static in nature. They feed all the inputs into the same models with identical computation, ignoring the inherent difference between different input samples. In Computer Vision and Natural Language Processing, this is called multi-modal dynamic fusion, which aims to adaptively fuse multi-modal data and make predictions based on the data-wise models [3, 12, 28]. Such issue also exists in multi-modal brain network analysis, where different samples may have significant differences in strength and connectivity of some brain regions [1, 7, 10]. While current methods treat all the samples equally by using same models to encode and fuse, they fail to avoid the effects for inherent variance among different samples during encoding and fusion. In other word, *model will have better performance if it can adaptively select unimodal-wise parts and*

*multimodal-wise parts for representation and fusion as the input sample dynamically changes.*

To fill the gap mentioned above in neuroscience, in this work we **innovatively propose a dynamic fusion approach for sample-adaptive multi-modal brain network analysis**. As is shown in Fig. 1, for other traditional fusion methods, all the input will be directly sent into specific or fusion modules. Our method instead adds gating module before to adaptively capture modal-specific and multi-modal data for further representation. We realize our gating module based on mixture-of-experts (MoE). However, directly adopting MoE is a challenge as it may face expert collapse in training. Therefore, we propose a **M**ulti-stage, **M**ulti-**M**odal **D**ynamic-**B**rain **F**u**S**ion Strategy (**M$^3$D-BFS**). M$^3$D-BFS consists of 3 stages: We first train uni-modal encoders for SC/FC modality respectively. Then we pretrain our model by training single experts for each MoE respectively. SC-FC contrastive loss and unimodal-guided distillation loss are designed for better performance of pretraining. Finally we finetune our model by only keeping MoE modules trainable. The weights for experts in previous stage are used to initialize experts in corresponding MoEs. To further enhance final representations (uni-modal/multi-modal) and their correlations with downstream tasks, we proposed a multi-mdoal disentangle loss to better capture task-related information. To the best of our knowledge, M$^3$D-BFS is the first work for dynamic fusion in sample-wise multi-modal brain network analysis. Extensive experiments on different real-world datasets demonstrate the effectiveness of our method.

## 2. Related Works

**Dynamic Multi-modal Fusion.** Firstly introduced in Computer Vision, dynamic fusion aims to dynamically select parts from multi-view images for fusion according to different samples [12]. MoGE [3] designed local gating and global gating modules for dynamic fusion. DynMM [28] considered dynamic scene for text-vision-audio modalities and proposed a new gating network for efficient dynamic fusion. [34] identified the correlations between multi-modal fusion and uncertain estimation and comes up with a robust fusion method.

**Multi-modal Brain Network Fusion.** To fully integrate data from different modalities for better downstream tasks, several methods have been developed for multi-modal brain network fusion. Generally, current fusion methods can be divided into three categories: **(1) Representation-level fusion** which directly concats or adds multi-modal representations before final prediction, like MMGNN [19], GBDM [33] and D-mhgcn [26]. **(2) Graph-level fusion** which focuses on fusion graph construction. ALNEGAT [5] designs an Attention-based graph fusion method by learning a fusion adjancency matrix. Cross-GNN [30] proposes a graph-based framework to capture finer cross-modal structure information. **(3) Transformer-based** fusion. RH-BrainFS [31] uses bottleneck transformer to avoid direct interaction between SC and FC. NeuroPath [25] introduces coupling module between SC and FC to improve accuracy of FC-based prediction.

## 3. Methodology

### 3.1. Notations

For a given multi-modal brain network dataset, we define it as $\mathcal{D} = \{(\mathcal{G}^i_{SC}, \mathcal{G}^i_{FC}), y_i\}^n_{i=1}$, where $n$ represents the sample size. For each sample, $\mathcal{G}^i_{SC}, \mathcal{G}^i_{FC}$ stands for structural connectivity graph and functional connectivity graph respectively, and $y_i$ is the corresponding label (depending on different tasks, gender, disease, etc) and $y_i \in \{0, 1\}$. For brain network of each modality: $\mathcal{G}_m = (A_m, X_m, \mathcal{V}_m, \mathcal{E}_m), m \in \{SC, FC\}$, nodes $\mathcal{V}_m$ represents different brain regions in the brain while edges $\mathcal{E}_m$ represents the connectivity between two brain regions. $X_m \in \mathbb{R}^{N\times d}$ denotes node features for each region and $A_m \in \mathbb{R}^{N\times N}$ denotes the adjacency matrix for brain network $\mathcal{G}_m$.

### 3.2. Overview of M$^3$D-BFS

In this section, we will introduce more details about our M$^3$D-BFS. The structure of our framework is shown in Fig. 2 (stage 3), we project unimodal embeddings and send them into following

MoE blocks. Via MoE's gating mechanism, different brain regions can be assigned to following experts dynamically, thus realizing sample-adaptive dynamic fusion. To facilitate better training, we divide our M$^3$D-BFS into 3 stages: **(1) stage 1: Uni-modal encoder pretraining**, where we train SC and FC encoders respectively. **(2) stage 2: Single expert pretraining**, where we train our multi-modal fusion model using single expert network for SC/FC/Fusion. **(3) stage 3: MoE finetuning** for dynamic multi-modal fusion, where we finetune MoE blocks while keeping other modules frozen.

### 3.3. Uni-modal encoder pretraining

Directly training a multi-modal fusion model can lead to sub-optimal performance, sometimes the performance of the fusion model can be worse than uni-modal models. This phenomenon of insufficient learning during multi-modal fusion is called "modality laziness" [9] which hurts fusion model's generalization ability.

Thus, encouraged by [9], we first train uni-modal encoders for FC/SC respectively, which **serve as guidance for multi-modal encoder in stage 2**. Formally, for given data $\{(\mathcal{G}^i_{FC}, \mathcal{G}^i_{SC}), y_i\}$ which stands for $i$-$th$ sample, we denote the uni-modal encoder as:

$$\hat{y_i}_m = g_{\theta_m} \circ f_{\varphi_m}(\mathcal{G}^i_m) = g_{\theta_m}(f_{\varphi_m}(\mathcal{G}^i_m)) \quad (1)$$

Here $m \in \{SC, FC\}$. $f_{\varphi_m}$ denotes the encoder and here we realize it with a two-layer GCN. $g_{\theta_m}$ is for classifier and we use a one-layer MLP (multi-layer perceptron) for implementation. For either SC or FC, we use cross-entropy loss ($\mathcal{L}_{ce}$) to optimize the uni-modal model.

After uni-modal training, we save the weights for SC/FC encoders ($f_{\varphi_{SC}}, f_{\varphi_{FC}}$), which will be used for soft alignment in stage 2.

### 3.4. Single expert pretraining

We train our multi-modal fusion model after we pretrain uni-modal encoders. However, directly training MoE is challenging in practice, which may lead to instability during training process [21].

According to [4, 15, 16], progressive training strategy is an effective way to train powerful MoE models. To this end, we further adopt a **"pretrain then finetune" paradigm** for MoE training [15]. Formally, we first train single expert then finetune the MoEs. M$^3$D-BFS is independent of the specific architecture of MoE experts and here we utilize MLP as our expert network. For representations from SC/FC encoders (denote as $h_{SC}, h_{FC}$), we project and feed them into following experts. Corresponding to our design of MoEs, we consider three types of expert MLPs: FC-specific, SC-specific and Fusion experts. Modal-specific experts take the representation of uni-modal as input while Fusion expert takes the concatenation of SC and FC. **READOUT**$(\cdot)$ will be applied for the output of each expert before concatenation and final prediction (shown in Eq. 2, and **FFN**$(\cdot)$ is the feedforward network layer).

$$\begin{cases} \hat{y_i} & = \textbf{FFN}(\text{Concat}(z^i_{SC},\ z^i_{FC},\ z^i_{Fusion})) \\ z^i_m & = \textbf{READOUT}(h^i_m), m \in \{SC, FC, Fusion\} \end{cases} \quad (2)$$

However, simple cross-entropy loss ($\mathcal{L}_{ce}$) is not enough for multi-modal supervised learning and the reasons lie in label's insufficient supervision for different modalities. Therefore, we add the supervision of pretrained uni-modal encoders from stage 1 [9]. For representations through encoders $(f_{\phi_{SC}}, f_{\phi_{FC}})$ from stage 1, we design a uni-modal guided distillation loss as follows:

$$\mathcal{L}_{distill} = \frac{\tau^2}{2}\left(\mathcal{D}_{KL}(\frac{h_{\varphi_{FC}}}{\tau}, \frac{h_{\phi_{FC}}}{\tau}) + \mathcal{D}_{KL}(\frac{h_{\varphi_{SC}}}{\tau}, \frac{h_{\phi_{SC}}}{\tau})\right) \quad (3)$$

The output of the uni-modal encoders from stage 2 are denoted as $(h_{\phi_{SC}}, h_{\phi_{FC}})$ and we regard stage 1 representations as teacher logits $(h_{\varphi_{SC}}, h_{\varphi_{FC}})$. Through $\mathcal{L}_{distill}$ as soft alignment, our M$^3$D-BFS can keep the performance of uni-modal encoders.

Besides, for better SC/FC representations, we adopt practice from LIMoE [17] pretraining which utilized CLIP-based loss [18] between batch data. For given SC/FC representations $(h_{\phi_{SC}}, h_{\phi_{FC}}) \in \mathbb{R}^{N \times d}$, we utilize **C**ontrastive **M**ultimodal **B**rain **P**re-training loss (**CMBP**, denote as $\mathcal{L}_{contrast}$) for better align representations between SC and FC. We use pooling for $(h_{\phi_{SC}}, h_{\phi_{FC}})$, and get $d$-dimension representations for SC and FC modality. $\mathcal{L}_{contrast}$ minimizes the distance between paired SC-FC data within one batch during training: $\mathcal{L}_{contrast} = \frac{1}{2}(\mathcal{L}_{S2F} + \mathcal{L}_{F2S})$. Here we use

**InfoNCE**$(\cdot)$ for implementation.

During this stage, classification loss $\mathcal{L}_{ce}$, distillation loss $\mathcal{L}_{distill}$ and $\mathcal{L}_{contrast}$ together make the loss function (Eq. 4, where $\beta$ is a hyper-parameter ranging $(0, 1)$).

$$\mathcal{L}_{stage2} = \mathcal{L}_{ce} + \beta \cdot (\mathcal{L}_{distill} + \mathcal{L}_{contrast}) \quad (4)$$

After training, we save the weights of single MLP experts (SC, FC, Fusion), which will be used in stage 3.

### 3.5. MoE finetuning

**MoE finetuning for dynamic multi-modal fusion.** Based on the pretrained model in stage 2, We finally finetune our M$^3$D-BFS in stage 3 to make our training process more stable. In this stage we replace the MLP experts with MoEs, whose experts have the same structure as MLP experts. As is shown in Fig. 2, We only train MoEs during this stage while keeping other modules frozen. We consider three types of MoEs: Uni-modal MoEs (SC-specific, FC-specific) which capture SC-only or FC-only features and Multi-modal MoE which dynamically fuses SC and FC features. Moreover, to avoid insufficient learning for some experts in MoE training, we apply the MLP weights saved in stage 2 to initialize all the expert networks in corresponding MoEs.

Each MoE consists of two parts: expert networks and gating network. Here we follow the classical practice from [21] and use MLP to realize both modules. For MoE $m$ which denotes its type ($m \in \{ SC, FC, Fusion\}$), the output can be written as follows:

$$y_m = \sum_{i=1}^{N_m} G_m(h_m)_i \cdot E_m(h_m)_i \quad (5)$$

Where the input $h_m$ takes the dimension of $\mathbb{R}^{N\times d}$ for $m \in \{SC, FC\}$ ($N_m = N$) and $\mathbb{R}^{2N\times d}$ for $m \in \{Fusion\}$ ($N_m = 2N$). $G_m$ is gating network for corresponding MoE. Following [21], we apply **Softmax** function for gating output. Before taking softmax, we add Gaussian noise and only keep top k values. Here we set k as 1, which means each brain region selects only one expert during training and inference. As the output of $G_m$ depends on input $X$, it can dynamically assign brain regions to different experts according to different data samples.

However, direct training process can get collapsed easily, where all input can be assigned to only one expert, and distributions of gating network can be imbalanced, where favored experts are trained more rapidly and are preferred during inference while other experts will get insufficient training [6, 21]. Thus MoE will degenerate into a single expert. To alleviate this issue of imbalanced training between experts, we apply the Importance loss and Load loss for each MoE ($\mathcal{L}_{moe}$). For given MoE, the loss is the combination of $\mathcal{L}_{Importance}$ and $\mathcal{L}_{load}$, where the importance loss and load loss are defined as:

$$\begin{cases} \mathcal{L}_{Importance}(X) = CV(Importance(X))^2 = CV(\sum_{x\in X} G(x))^2 \\ \mathcal{L}_{Load}(X) = CV(Load(X))^2 = CV(\sum_{x\in X} P(x))^2 \end{cases} \quad (6)$$

Here $CV(\cdot)$ means the coefficient of variation of the values. The initialization of experts' weights, as well as the MoE loss function together alleviate imbalanced training issue of MoE which makes the training more stable and more efficient.

**Representation-enhanced multi-modal disentanglement loss.** In multi-modal learning, insufficient label information for supervision may influence the performance for modal-specific and fusion embeddings. One solution is through disentangled representation learning which reduces redundancy between modal-specific and modal-fusion information while keeping their correlations with label. Here, derived from [14]'s method, we design our multi-mdoal disentanglement loss to learn better uni-modal and multi-modal representations ($\mathcal{L}_{disen}$). We apply pooling function after the output of MoE and for each data, the final embeddings are $\{z_{SC}, z_{FC}, z_{Fusion}\} \in \mathbb{R}^d$. Our $\mathcal{L}_{disen}$ is based on contrastive loss and we hope that the final embeddings can be as orthogonal to each other as possible for better representation. We use the concatenation of three representations and get the anchor sample for $i$-$th$ data $z^i_{anchor}$ by concating $z^i_{SC}, z^i_{FC}$ and $z^i_{Fusion}$. $z^i_{anchor} \in \mathbb{R}^d$ is used for prediction and for each $z^i_m$, we regard other types as negative samples and $z^i_{anchor}$ as positive sample.

Table 1. Comparison results on two different datasets (HCP, ZDXX). We run 10 times and record average acc ± std (%). The best results are marked **bold**, the second best underline and the third best gray.

| Datasets | Methods | Modality | ACC | SEN | SPE | F1 | AUC |
|---|---|---|---|---|---|---|---|
| **HCP** | GCN | FC | 64.03 ± 1.21 | 55.01 ± 2.88 | 71.75 ± 2.37 | 57.15 ± 2.49 | 63.38 ± 1.25 |
| | GCN | SC | 68.23 ± 1.76 | 66.22 ± 4.79 | 72.27 ± 4.53 | 67.91 ± 2.71 | 69.61 ± 1.75 |
| | BrainNPT | FC | 67.78 ± 1.53 | 60.67 ± 1.34 | 73.98 ± 2.83 | 59.15 ± 3.70 | 66.97 ± 2.44 |
| | BrainNPT | SC | 70.74 ± 1.49 | 68.95 ± 4.82 | 70.02 ± 5.67 | 68.16 ± 2.13 | 70.11 ± 1.73 |
| | SVM | FC, SC | 74.49 ± 2.97 | 71.18 ± 4.95 | 77.32 ± 3.30 | 71.96 ± 3.49 | 74.25 ± 3.05 |
| | Random Forest | FC, SC | 68.24 ± 2.94 | 54.88 ± 5.78 | 79.64 ± 4.16 | 61.31 ± 4.44 | 67.26 ± 3.04 |
| | MMGNN | FC, SC | 73.33 ± 2.82 | 71.17 ± 4.52 | 75.17 ± 5.67 | 71.10 ± 2.88 | 73.17 ± 2.72 |
| | AL-NEGAT | FC, SC | 75.12 ± 3.66 | 72.86 ± 7.74 | 84.46 ± 5.05 | 76.13 ± 4.70 | 78.66 ± 3.81 |
| | Cross-GNN | FC, SC | 78.93 ± 1.05 | 73.53 ± 1.84 | **85.04 ± 2.02** | **76.81 ± 0.86** | 79.82 ± 0.72 |
| | RH-BrainFS | FC, SC | 78.63 ± 4.36 | 75.59 ± 6.75 | 81.25 ± 6.04 | 76.49 ± 6.91 | 78.42 ± 4.38 |
| | NeuroPath | FC, SC | 77.77 ± 1.87 | 75.24 ± 2.61 | 80.38 ± 1.50 | 75.78 ± 1.84 | 77.81 ± 1.67 |
| | **M$^3$D-BFS (Ours)** | FC, SC | **81.21 ± 0.85** | **76.14 ± 3.32** | 82.50 ± 3.07 | 76.75 ± 1.25 | **80.80 ± 1.07** |
| **zhongdaxinxiang** | GCN | FC | 68.63 ± 1.28 | 64.61 ± 4.67 | 72.68 ± 4.52 | 64.68 ± 3.13 | 68.64 ± 1.18 |
| | GCN | SC | 72.15 ± 2.49 | 72.79 ± 5.14 | 73.24 ± 4.80 | 71.90 ± 4.25 | 72.02 ± 2.52 |
| | BrainNPT | FC | 70.65 ± 1.85 | 73.00 ± 4.59 | 69.01 ± 4.30 | 69.27 ± 2.64 | 71.00 ± 2.03 |
| | BrainNPT | SC | 72.83 ± 1.73 | 74.57 ± 3.80 | 69.89 ± 3.87 | 73.63 ± 4.07 | 72.78 ± 2.65 |
| | SVM | FC, SC | 66.06 ± 1.56 | 58.01 ± 2.49 | 74.04 ± 1.77 | 62.03 ± 2.27 | 66.03 ± 1.56 |
| | Random Forest | FC, SC | 62.43 ± 2.19 | 56.17 ± 2.87 | 68.60 ± 3.48 | 59.04 ± 2.73 | 62.38 ± 2.20 |
| | MMGNN | FC, SC | 59.72 ± 3.18 | 65.10 ± 4.83 | 54.42 ± 4.42 | 59.96 ± 4.02 | 59.76 ± 3.15 |
| | AL-NEGAT | FC, SC | 71.86 ± 2.49 | 75.26 ± 3.62 | 68.12 ± 6.19 | 72.16 ± 2.24 | 71.69 ± 2.56 |
| | Cross-GNN | FC, SC | 74.01 ± 2.33 | **77.39 ± 5.71** | 70.46 ± 5.65 | 74.38 ± 3.15 | 73.92 ± 2.33 |
| | RH-BrainFS | FC, SC | 78.48 ± 1.43 | 76.20 ± 4.06 | 80.72 ± 3.60 | 77.35 ± 1.97 | 78.46 ± 1.43 |
| | NeuroPath | FC, SC | 78.01 ± 1.17 | 77.03 ± 2.23 | 79.17 ± 3.47 | 77.54 ± 2.28 | 78.50 ± 1.24 |
| | **M$^3$D-BFS (Ours)** | FC, SC | **80.85 ± 2.44** | 77.22 ± 3.04 | **83.91 ± 4.95** | **77.82 ± 3.62** | **80.74 ± 2.57** |

$\mathcal{L}_{disen}$ can be written as Eq. 7:

$$\mathcal{L}_{disen} = -\frac{1}{n \cdot M} \sum_{i=1}^{n} \sum_{m=1}^{M} \log \frac{\exp\left(\frac{z_m^i \cdot z_{anchor}^i}{\tau_d}\right)}{\exp\left(\frac{z_m^i \cdot z_{anchor}^i}{\tau_d}\right) + \sum_{\substack{q=1, \\ q \neq m}}^{M} \exp\left(\frac{z_m^i \cdot z_q^i}{\tau_d}\right)} \tag{7}$$

Here $M$ defers to different types ($\{SC, FC, Fusion\}$). According to Eq. 7, through $\mathcal{L}_{disen}$, we can disentangle different representations while increasing their relevance to downstream tasks, thus enhancing their representations. Finally, the loss function in stage 3 is in Eq. 8 with $\alpha \in (0, 1)$:

$$\mathcal{L}_{stage3} = \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{moe} + (1 - \alpha) \cdot \mathcal{L}_{disen} \tag{8}$$

# 4. Experiments

## 4.1. Experimental settings

**Datasets.** We evaluate our M$^3$D-BFS on two different datasets: **(1) Human Connectsome Project (HCP)** [23], a public dataset on gender classification which consists of 560 female and 479 male samples. **(2) Major Depressive Disorder dataset from hospitals (zhongdaxinxiang, also short as ZDXX)**, including the Affiliated Zhongda Hospital of Southeast University and the Second Affiliated Hospital of Xinxiang Medical University, which consists of 94 controls and 93 MDD patients.

We also introduce preprocess of our dataset. **(1) SC**: The brain's diffusion toolbox of FMRIB is utilized dMRI preprocessing [20]. Then the brain regions are obtained ($\mathcal{V}_{SC}$) through Anatomical Automatic labeling (AAL) template. After that, we use DSI Studio [11] for fiber tracking before we finally get the features of SC graph ($X_{SC}$) by counting the number of structural connective fibres between different regions of AAL. The construction of $A_{SC}$ is based on $X_{SC}$ by adding a certain threshold. **(2) FC**: The Data Preprocessing Assistant for Resting-State Function (DPARSF) MRI tookit [22] is utilized for fMRI preprocessing. Then the average time series are computed for each brain region

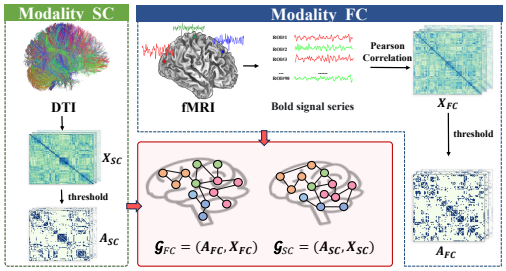



Figure 3. Multi-modal brain networks construction.

with AAL template. Pearson correlation is then calculated as function matrix, which denotes the feature matrix for FC ($X_{FC}$). Its $A_{FC}$ is obtained by thresholding a certain proportional quantization.

**Baselines.** For uni-modal methods, as we choose **GCN** as our uni-modal encoders, we record its results in uni-modal scenes. Besides, we select **BrainNPT** [13], a transformer-based method for brain network analysis. For multi-modal methods, we select (1) 2 methods which concatenate SC/FC features as input: **SVM**, **Random Forest**; (2) 1 method concatenates uni-modal embeddings: **MMGNN** [19]; (3) 2 methods which construct a graph between different modalities for fusion: **AL-NEGAT** [5], **Cross-GNN** [30] and (4) 2 transformer-based fusion methods: **RH-BrainFS** [31] and **NeuroPath** [25]. Code implementations of all baseline methods are taken from their respective original papers.

**Experimental settings.** We choose 3-layer GCN with its hidden dimension 128 for uni-modal encoders. For fusion parts, we utilize a two-layer MoE. Top-k value of each MoE block is set as 1. We adopt Adam as optimizer and training epochs is 500 in total with 300 epochs as early stopping. For hyper-parameters settings, $\alpha \in \{0.2, 0.4, 0.6, 0.8\}, \beta \in \{0.3, 0.5, 0.7, 0.9\}$. The number of experts in MoE $k \in [2, 6]$. The learning rate is 0.005 with weight dacay as 0.0001. Our model is implemented in PYG and trained on RTX Titans with 24GB memory.

We evaluate model's performance on five metrics: accuracy (**ACC**), sensitivity (**SEN**), specificity (**SPE**), f1 score (**F1**) and ROC-AUC (**AUC**), where higher value means better performance. We use 10-fold cross validation on 10 random runs and record mean value and standard deviation.

Table 2. Ablation studies on sub-stages and sub-modules.

| Ablation types | Stage 1 | Stage 2 | | Stage 3 | Datasets | |
|---|---|---|---|---|---|---|
| | *KD* | *pretrain* | $\mathcal{L}_{contrast}$ | $\mathcal{L}_{disen}$ | HCP | zhongdaxinxiang |
| $S_1.A_1$ | ✗ | ✗ | ✗ | ✗ | 76.84 ± 1.92 | 77.28 ± 4.71 |
| $S_2.A_1$ | ✓ | ✗ | ✗ | ✗ | 78.16 ± 1.57 | 79.67 ± 4.42 |
| $S_2.A_2$ | ✓ | ✓ | ✗ | ✗ | 78.57 ± 1.06 | 77.00 ± 2.52 |
| $S_3.A_1$ | ✓ | ✓ | ✓ | ✗ | 77.80 ± 1.22 | 78.68 ± 4.06 |
| **M$^3$D-BFS** | ✓ | ✓ | ✓ | ✓ | **81.21** ± 0.85 | **80.85** ± 2.44 |

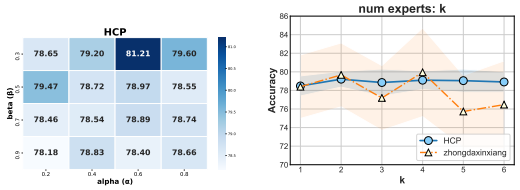



Figure 4. Sensitivity analysis experiments.

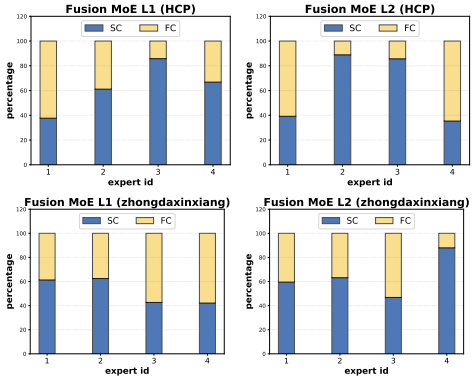



Figure 5. Visualization for uni-modal percentage (SC:FC) in each fusion MoE block.

## 4.2. Comparison experiments

The results of different methods for two datasets are shown in Tab. 1. Compared to uni-modal GCNs, M$^3$D-BFS greatly outperforms uni-modal encoders. Moreover, compared to other multi-modal brain fusion methods, M$^3$D-BFS also performs the best (2.28% **ACC** improvement on HCP and 2.73% **ACC** improvement on ZDXX).

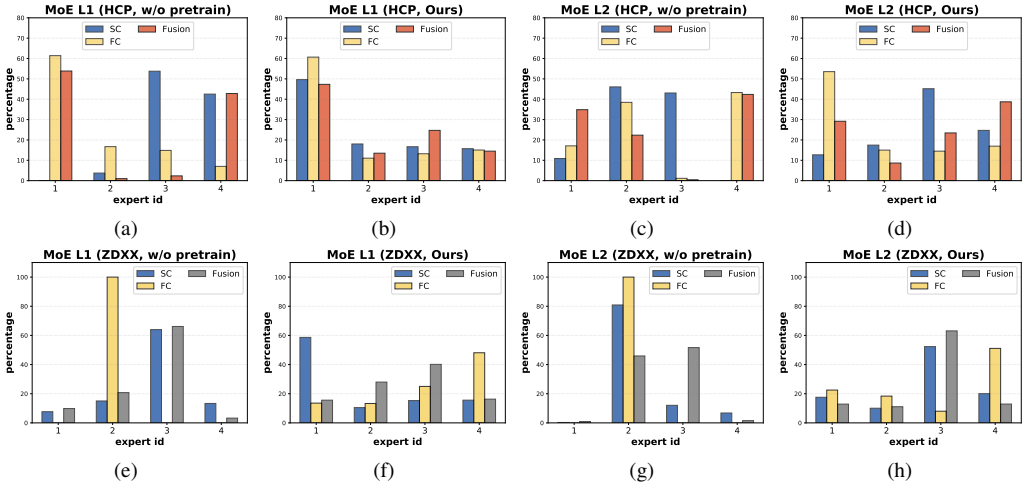



Figure 6. Visualization for percentage of experts in different MoEs. Here each MoE consists of 4 experts.

### 4.3. Ablation studies and sensitivity analyses

We conduct ablation studies to illustrate effectiveness of different sub-modules in M$^3$D-BFS. We validate its effectiveness via stage-by-stage ablation. As is shown in Tab. 2, $S_i.A_j$ means ablation for stage $i$ and $j$-$th$ sub-module in stage $i$. The results in Tab. 2 show that each sub-module works in our M$^3$D-BFS. The distillation from uni-modal encoders ensure the models' representations for uni-modal during fusion ($S_1.A_1$). The "pretrain-finetune" method not only improves accuracy, but also lowers deviation ($S_2.A_1, S_2.A_2$). Finally $\mathcal{L}_{disen}$ leads to better performance ($S_3.A_1$).

We also conduct extensive experiments on different hyper-parameters. For loss function coefficients, here we consider $\beta$ in stage 2 and $\alpha$ in stage 3 and plot the heatmap for them, which is depicted in Fig. 4. For HCP, the accuracy reaches its peak when the values are $(\alpha, \beta) = (0.6, 0.3)$. Finally we conduct experiments on different expert numbers in the MoE block. The value of number k is from 1 to 6, and as is shown in Fig. 4, for HCP, the change of accuracy is subtler than ZDXX dataset.

### 4.4. Visualization for experts

Firstly we plot percentage of different modalities in each fusion MoE block. As is shown in Fig. 5, every fusion expert avoid collapse (SC/FC 100% or 0%) phenomenon, showing that M$^3$D-BFS **gets effective training in dynamic fusion**. Then we plot the distribution of all experts in one MoE block to check if expert collapse occurs in our model. We also plot distributions for models without pretraining (left side) in stage 2 as comparison, which is shown in Fig. 6. For MoE in both layers, model without pretraining may face collapse among experts. While for M$^3$D-BFS (right side), the issue is effectively mitigated, **with more even distributions for different experts**.

## 5. Conclusion

In this work, we propose a multi-stage MoE-based multi-modal brain network fusion method (M$^3$D-BFS) for sample-adaptive dynamic fusion in neuroscience. To the best of our knowledge, this is the first paper discussing dynamic fusion in multi-modal brain fusion. Extensive experiments on different real-world datasets validate the effectiveness of our method. In our future work, we aim to consider more complex cases where different data may have inherent bias in its distribution to achieve better results in different downstream tasks under multi-modal brain network scenarios.

## Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.62471133). This work is also supported by the Big Data Computing Center of Southeast University.

## References

[1] Silvia Basaia, Federica Agosta, Elisabetta Sarasso, Roberta Balestrino, Tanja Stojković, Iva Stanković, Aleksandra Tomić, Vladana Marković, Francesca Vignaroli, Elka Stefanova, et al. Brain connectivity networks constructed using mri for predicting patterns of atrophy progression in parkinson disease. *Radiology*, 311(3):e232454, 2024. 2

[2] Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5833–5848, 2022. 1

[3] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23555–23564, 2023. 2, 3

[4] Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1110–1119, 2024. 4

[5] Yuzhong Chen, Jiadong Yan, Mingxin Jiang, Tuo Zhang, Zhongbo Zhao, Weihua Zhao, Jian Zheng, Dezhong Yao, Rong Zhang, Keith M Kendrick, et al. Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 3, 7

[6] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022. 5

[7] Brendon C Chou, Alexander Lerner, Giuseppe Barisano, Daniel Phung, Wilson Xu, Soniya N Pinto, and Nasim Sheikh-Bahaei. Functional mri and diffusion tensor imaging in migraine: a review of migraine functional and white matter microstructural changes. *Journal of Central Nervous System Disease*, 15:11795735231205413, 2023. 2

[8] Shu-Hsien Chu, Keshab K Parhi, and Christophe Lenglet. Function-specific and enhanced brain structural connectivity mapping via joint modeling of diffusion and functional mri. *Scientific reports*, 8(1):4741, 2018. 1

[9] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pages 8632–8656. PMLR, 2023. 4

[10] Mingjun Duan, Yuchao Jiang, Xi Chen, Cheng Luo, and Dezhong Yao. Degree centrality of the functional network in schizophrenia patients. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi Journal of Biomedical Engineering Shengwu Yixue Gongchengxue Zazhi*, 34(6):837–841, 2017. 2

[11] Aina Frau-Pascual, Jean Augustinack, Divya Varadarajan, Anastasia Yendiki, Bruce Fischl, and Iman Aganj. Detecting structural brain connectivity differences in dementia through a conductance model. In *2019 53rd Asilomar conference on signals, systems, and computers*, pages 591–595. IEEE, 2019. 6

[12] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20707–20717, 2022. 2, 3

[13] Jinlong Hu, Yangmin Huang, Nan Wang, and Shoubin Dong. Brainnpt: Pre-training transformer networks for brain network classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024. 7

[14] Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, Farid Razzak, Ji-Rong Wen, and Hui Xiong. Modeling multiple views via implicitly preserving global consistency and local complementarity. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7220–7238, 2022. 5

[15] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *arXiv preprint arXiv:2405.11273*, 2024. 4

[16] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 4

[17] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. 4

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[19] Isaac Sebenius, Alexander Campbell, Sarah E Morgan, Edward T Bullmore, and Pietro Liò. Multimodal graph coarsening for interpretable, mri-based brain graph neural network. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2021. 2, 3, 7

[20] Preya Shah, Arian Ashourvan, Fadi Mikhail, Adam Pines, Lohith Kini, Kelly Oechsel, Sandhitsu R Das, Joel M Stein, Russell T Shinohara, Danielle S Bassett, et al. Characterizing the role of the structural connectome in seizure dynamics. *Brain*, 142 (7):1955–1972, 2019. 6

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 4, 5

[22] Xiao-Wei Song, Zhang-Ye Dong, Xiang-Yu Long, Su-Fang Li, Xi-Nian Zuo, Chao-Zhe Zhu, Yong He, Chao-Gan Yan, and Yu-Feng Zang. Rest: a toolkit for resting-state functional magnetic resonance imaging data processing. *PloS one*, 6(9): e25031, 2011. 6

[23] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013. 6

[24] Laura S Van Velzen, Sinead Kelly, Dmitry Isaev, Andre Aleman, Lyubomir I Aftanas, Jochen Bauer, Bernhard T Baune, Ivan V Brak, Angela Carballedo, Colm G Connolly, et al. White matter disturbances in major depressive disorder: a coordinated analysis across 20 international cohorts in the enigma mdd working group. *Molecular psychiatry*, 25(7):1511–1525, 2020. 1

[25] Ziquan Wei, Tingting Dan, Jiaqi Ding, and Guorong Wu. Neuropath: A neural pathway transformer for joining the dots of human connectomes. *arXiv preprint arXiv:2409.17510*, 2024. 2, 3, 7

[26] Xuyun Wen, Qumei Cao, Yunxi Zhao, Xia Wu, and Daoqiang Zhang. D-mhgcn: An end-to-end individual behavioral prediction model using dual multi-hop graph convolutional network. *IEEE Journal of Biomedical and Health Informatics*, 2024. 2, 3

[27] Nils R Winter, Julian Blanke, Ramona Leenings, Jan Ernsting, Lukas Fisch, Kelvin Sarink, Carlotta Barkhau, Daniel Emden, Katharina Thiel, Kira Flinkenflügel, et al. A systematic evaluation of machine learning–based biomarkers for major depressive disorder. *JAMA psychiatry*, 81(4):386–395, 2024. 1

[28] Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2023. 2, 3

[29] Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019. 1

[30] Yanwu Yang, Chenfei Ye, Xutao Guo, Tao Wu, Yang Xiang, and Ting Ma. Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning. *IEEE transactions on medical imaging*, 43(1):108–121, 2024. 2, 3, 7

[31] Hongting Ye, Yalu Zheng, Yueying Li, Ke Zhang, Youyong Kong, and Yonggui Yuan. Rh-brainfs: regional heterogeneous multimodal brain networks fusion strategy. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7

[32] Usman Zahid, Imran Ashraf, Muhammad Attique Khan, Majed Alhaisoni, Khawaja M Yahya, Hany S Hussein, and Hammam Alshazly. Brainnet: optimal deep learning feature fusion for brain tumor classification. *Computational Intelligence and Neuroscience*, 2022(1):1465173, 2022. 1

[33] Lu Zhang, Li Wang, Jean Gao, Shannon L

Risacher, Jingwen Yan, Gang Li, Tianming Liu, Dajiang Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Deep fusion of brain structure-function in mild cognitive impairment. *Medical image analysis*, 72:102082, 2021. 2, 3

[34] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023. 3