Learning 3D Medical Image Models From Brain Functional Connectivity Network Supervision For Mental Disorder Diagnosis

Xingcan Hu[1,2], Wei Wang[1], and Li Xiao[1,2(✉)]

[1] MoE Key Laboratory of Brain-Inspired Intelligence Perception and Cognition,
University of Science and Technology of China, Hefei 230052, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center,
Hefei 230088, China
xiaoli11@ustc.edu.cn

Abstract. In MRI-based mental disorder diagnosis, most previous studies focus on functional connectivity network (FCN) derived from functional MRI (fMRI). However, the small size of annotated fMRI datasets restricts its wide application. Meanwhile, structural MRIs (sMRIs), such as 3D T1-weighted (T1w) MRI, which are commonly used and readily accessible in clinical settings, are often overlooked. To integrate the complementary information from both function and structure for improved diagnostic accuracy, we propose CINP (**C**ontrastive **I**mage-**N**etwork **P**re-training), a framework that employs contrastive learning between sMRI and FCN. During pre-training, we incorporate masked image modeling and network-image matching to enhance visual representation learning and modality alignment. Since the CINP facilitates knowledge transfer from FCN to sMRI, we introduce network prompting. It utilizes only sMRI from suspected patients and a small amount of FCNs from different patient classes for diagnosing mental disorders, which is practical in real-world clinical scenario. The competitive performance on three mental disorder diagnosis tasks demonstrate the effectiveness of the CINP in integrating multimodal MRI information, as well as the potential of incorporating sMRI into clinical diagnosis using network prompting.

Keywords: structural MRI · functional connectivity network · mental disorder diagnosis.

1 Introduction

By detecting the blood-oxygen-level-dependent (BOLD) responses to neural activity throughout the brain, functional MRI (fMRI) has become the leading neuroimaging technique for non-invasive study of human brain functions relevant to various behavioral and cognitive traits [20,13]. Recently, fMRI-derived functional connectivity network (FCN), as a graph architecture with nodes being brain regions-of-interest (ROIs) and each edge being functional connectivity (FC) between paired ROIs, has received considerable attention in diagnosis of mental disorders [30,3], where FC is in general measured as statistical dependence between BOLD signals of paired ROIs. So far a significant amount of work has been dedicated to learning deep and differentiable representations of FCN for improving diagnostic accuracy, such as graph neural networks (GNNs) [19,28,14], convolutional neural networks (CNNs) [16], and graph transformer [15].

Despite significant progress, such FCN-based deep learning methods for mental disorder diagnosis has yet to be widely adopted in real-world clinical practice. There are two pervasive challenges, i.e., the limited generalizability due to the insufficient annotated fMRI data volume, and the lack of integration with anatomical information from the easily obtainable structural MRI (sMRI), such as 3D T1w MRI. Since the anatomical structure of the brain inherently constrains its function [21], 3D T1w MRI, which assesses brain anatomy, holds potential for diagnosing mental disorders. An efficient integration of structural and functional perspectives can provide a more comprehensive view of neurobiological abnormalities in mental disorders, leading to better diagnostic precision.

Motivated by the success of contrastive learning on large-scale image-text pairs [22], efforts in the biomedical domain have focused on pre-training vision-language models using medical images and their corresponding radiology reports [2]. Beyond images and texts, it is noteworthy that various MRI modalities, such as sMRI and fMRI, inherently provide contrasting perspectives by offering structural and functional information about the human brain, respectively. For example, a bidrectional mapping scheme [31] performed contrastive learning between diffusion MRI-derived structural connectivity networks and BOLD signals. F2TNet [12] transferred knowledge from fMRI to sMRI using ROI-level contrastive learning, so as to enable accurate phenotypic predictions with sMRI alone. For Alzheimer's Disease prediction, Fedorov et al. [10] applied both inter- and intra-modal contrastive learning between sMRI and fALFF features. However, the scalability of these studies is somewhat limited by the specific model architecture, the small amount of data, and/or the coarse representation of features.

In this regard, this paper focuses on scaling contrastive pre-training on suject-level sMRI and fMRI for mental disorder diagnosis. We collect a large cohort of 4619 paired 3D T1w MRI images and fMRI-derived FCNs for pre-training. The Contrastive Image-Network Pre-training (CINP) framework (see Fig. 1) is proposed to learn representations of 3D T1w MRI images through FCN supervision for mental disorder diagnosis tasks. Specifically, paired 3D T1w MRI images and FCNs are fed into a visual encoder and a network encoder to extract embeddings, respectively. The cosine similarity matrix between image embeddings and FCN embeddings is generated to compute image-network contrastive loss. Masked image modeling and image-network matching are used for better representation learning and modality alignment. In particular, we develop a network prompting protocol, which leverages only 3D T1w MRI images from suspected patients and a small amount of FCNs from different patient classes for diagnosis of mental disorders. The similarities between the embedding of the 3D T1 MRI image of a suspected patient and the embeddings of FCNs are calculated. The patient is assigned to the class where the corresponding FCNs hold the highest similar-

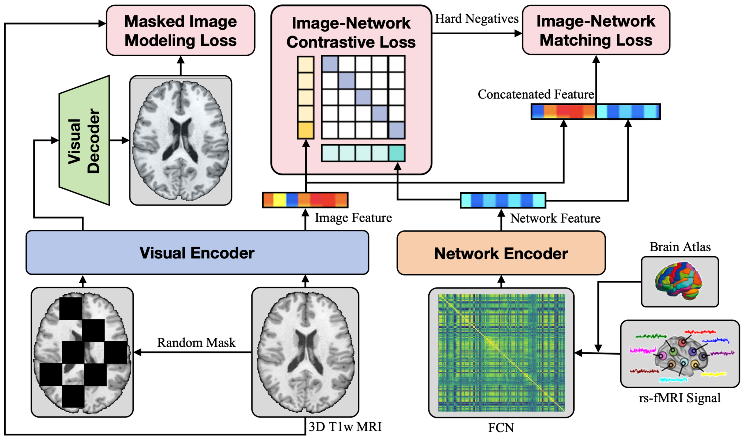

Fig. 1. The framework of CINP. It primarily consists of a visual encoder, a visual decoder, and a network encoder.

ity with the image. The effectiveness of CINP is finally demonstrated on three mental disorder datasets by comparing CINP with FCN-based, sMRI-based, and multimodal models.

2 Methods

2.1 Contrastive Image-Network Pre-training

As illustrated in Fig. 1, CINP is mainly composed of a visual encoder, a visual decoder, and a network encoder. For the visual encoder, an 8-layer 3D swin transformer is constructed and initialized with the weights from [25]. Given an input 3D T1w MRI image I, we randomly mask 30% of the voxels, resulting in a masked MRI image I^*. Through the visual encoder, the normalized image embedding $\boldsymbol{v}_I$ and the normalized masked image embedding $\boldsymbol{v}_I^*$, both with a dimension of 768, are obtained. The visual decoder, based on the transpose convolution layer, follows the architecture in [25], utilizing the masked image embedding $\boldsymbol{v}_I^*$ to reconstruct a volumetric MRI image $\hat{I}$. We adopt brain network transformer (BNT) [15] as the backbone for the network encoder, where an FCN N is encoded into a 768-dimensional normalized network embedding $\boldsymbol{w}_N$.

Image-Network Contrastive Learning. Studies have shown that contrastive learning using image-text pairs can construct a joint semantic space of vision and language [22]. Therefore, we aim to enhance the representations of 3D T1w MRI images with FCN supervision through contrastive learning between image-network pairs. Specifically, given an image-network pair (i.e., a 3D T1w MRI image and an FCN), we aim to learn a similarity score $s(I,N) = \boldsymbol{v}_I^\top \boldsymbol{w}_N$, such that positive pairs (image and network from the same subject) have higher similarity scores, while negative pairs (image and network from different subjects) have lower similarity scores. For each image and network in a batch, the softmax-normalized image-to-network and network-to-image similarities are calculated as

$$p_k^{\text{in}}(I) = \frac{\exp(s(I, N_k)/\tau)}{\sum_{k=1}^{K} \exp(s(I, N_k)/\tau)} \quad \text{and} \quad p_k^{\text{ni}}(N) = \frac{\exp(s(N, I_k)/\tau)}{\sum_{k=1}^{K} \exp(s(N, I_k)/\tau)}, \tag{1}$$

where the temperature factor τ is a learnable parameter and K denotes the batch size. Let $\boldsymbol{y}^{\text{in}}(I)$ and $\boldsymbol{y}^{\text{ni}}(N)$ represent the ground-truth similarities of all images and networks, where positive pairs having a similarity of 1 and negative pairs having a similarity of 0. Based on the cross entropy $\text{H}(\cdot, \cdot)$, the image-network contrastive (INC) loss is defined as

$$\mathcal{L}_{\text{INC}} = \frac{1}{2} \mathbb{E}_{(I,N)\sim D} \left[\text{H}\left(\boldsymbol{y}^{\text{ni}}(I), \boldsymbol{p}^{\text{ni}}(I)\right) + \text{H}\left(\boldsymbol{y}^{\text{in}}(N), \boldsymbol{p}^{\text{in}}(N)\right) \right]. \tag{2}$$

Masked Image Modeling. Masked image modeling (MIM) aims to learn robust representations of MRI images. The MIM loss is defined by an L_1 loss between the raw MRI image I and the reconstructed MRI image $\hat{I}$, i.e.,

$$\mathcal{L}_{\text{MIM}} = \mathbb{E}_{(I,\hat{I})\sim D} \| I - \hat{I} \|_1. \tag{3}$$

Image-Network Matching. Image-network matching (INM) is a binary classification task, which predicts whether a given image-network pair is from the same subject. Specifically, the concatenation of the image embedding $\boldsymbol{v}_I$ and the network embedding $\boldsymbol{w}_N$, denoted as $[\boldsymbol{v}_I, \boldsymbol{w}_N]$, is passed through a fully-connected layer to output a binary classification probability q. The INM loss is defined as

$$\mathcal{L}_{\text{INM}} = \mathbb{E}_{(I,N)\sim D} \left[H\left(\boldsymbol{z}_{\text{INM}}, \boldsymbol{q}(I, N)\right) \right], \tag{4}$$

where $\boldsymbol{z}_{\text{INM}}$ is the ground-truth label represented by a 2-dimensional one-hot vector. Moreover, we sample hard negatives as indicated in [18]. To be specific, in each data batch, the image-network similarity in (1) is used to sample image-network pairs which do not come from the same subject but own a high similarity for the INM loss. Finally, the complete pre-training loss of CINP is

$$\mathcal{L} = \mathcal{L}_{\text{INC}} + \alpha \mathcal{L}_{\text{MIM}} + \beta \mathcal{L}_{\text{INM}}. \tag{5}$$

2.2 Network prompting

Previous methods employ linear probe or fine-tuning protocols to apply the pre-trained model to downstream tasks, which require a substantial number of annotated data and fall short when encountering the prevalent clinical scenario where fMRI are not routinely collected. To address this challenge, as shown in Fig. 2, we propose network prompting, inspired by the insight that FCNs from different subject classes (e.g., health controls and autism spectrum disorder

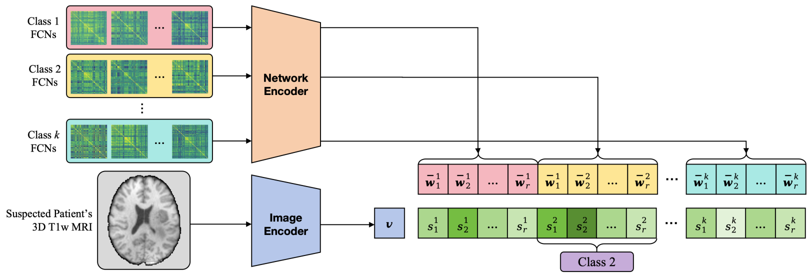
Class 1
FCNs
Class 2
FCNs
Class k
FCNs
Suspected Patient's
3D T1w MRI
Network
Encoder
Image
Encoder
$\boldsymbol{v}$
$\overline{\boldsymbol{w}}_1^1$
$\overline{\boldsymbol{w}}_2^1$
$\overline{\boldsymbol{w}}_r^1$
$\overline{\boldsymbol{w}}_1^1$
$\overline{\boldsymbol{w}}_2^2$
$\overline{\boldsymbol{w}}_r^2$
$\overline{\boldsymbol{w}}_1^k$
$\overline{\boldsymbol{w}}_2^k$
$\overline{\boldsymbol{w}}_r^k$
s_1^1
s_2^1
s_r^1
s_1^2
s_2^2
s_r^2
s_1^k
s_2^k
s_r^k
Class 2

Fig. 2. Workflow of network prompting. In this case, the 3D T1w MRI image is classified as class 2, since the mean of $s_1^2, s_2^2, \ldots, s_r^2$ is the highest average similarity.

patients) exhibit significant group-level differences [32,11]. We also hypothesize that the pre-trained CINP can measure the similarity between 3D T1w MRI image embeddings and FCN embeddings in the learned joint semantic space.

Specifically, we obtain a set of network embeddings $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ of k subject classes, each containing n FCNs, i.e, $\mathcal{C}_l = \{\boldsymbol{w}_1^l, \boldsymbol{w}_2^l, \ldots, \boldsymbol{w}_n^l\}$ for $l = 1, 2, \ldots, k$. For each class, we partition the FCNs into r disjoint subsets of equal size, i.e., $\mathcal{C}_l = \cup_{i=1}^r \mathcal{C}_l^i$, where $\mathcal{C}_l^i \cap \mathcal{C}_l^j = \varnothing$ for $i \neq j$, and $|\mathcal{C}_l^i| = \dfrac{|\mathcal{C}_l|}{r}$ for $i = 1, 2, \ldots, r$. The nework embeddings within the same subset are averaged to form r group-level reference network embeddings: $\overline{\mathcal{U}} = \{\{\overline{\boldsymbol{w}}_1^l, \overline{\boldsymbol{w}}_2^l, \ldots, \overline{\boldsymbol{w}}_r^l, \}| l = 1, 2, \ldots, k\}$, where $\overline{\boldsymbol{w}}_i^l = \dfrac{1}{|\mathcal{C}_l^i|} \sum_{\boldsymbol{w} \in \mathcal{C}_l^i} \boldsymbol{w}$ for $i = 1, 2, \ldots, r$, which help eliminate subject-level biases. Then, we calculate the similarities between the image embedding $\boldsymbol{v}$ of the suspected patient's 3D T1w MRI and the reference network embeddings, denoted as $\mathcal{S} = \{\{s_1^l, s_2^l, \ldots, s_r^l\}| l = 1, 2, \cdots, k\}$, where $s_i^l = \boldsymbol{v}^\top \overline{\boldsymbol{w}}_i^l$ for $i = 1, 2, \ldots, r$. Based on the average similarities of the image embedding with each class's FCNs $\overline{\mathcal{S}} = \{\overline{s}^1, \overline{s}^2, ..., \overline{s}^k\}$, where $\overline{s}^l = \dfrac{1}{r} \sum_{i=1}^r s_i^l$ for $l = 1, 2, \ldots, k$, we assign the patient to the class k' with the highest average similarity, i.e., $\overline{s}^{k'} = \max(\overline{s}^1, \overline{s}^2, \ldots, \overline{s}^k)$.

3 Experiments

3.1 Experimental Settings

Datasets and Preprocessing. We collected several publicly available datasets where subjects have both 3D T1w MRI and resting-state fMRI (rs-fMRI) images. As a result, the large cohort for pre-training included four datasets (i.e., HBN [1], HCP [27], QTIM [23], and CNP [5]), while three mental disorder datasets (i.e., ABIDE [8], ADHD [7], and SRPBS [24]) were used for evaluation. We listed these datasets with basic demographic information in Table 1. Using the fMRIPrep [9] preprocessing pipeline, which is robust to variations in scan acquisition protocols, all MRI images were preprocessed and had a voxel size of $2 \times 2 \times 2\ mm^3$. To derive FCNs from rs-fMRI on the automated anatomical labelling (AAL) atlas [26],

Table 1. Demographic information of 7 datasets used in this study. (HC: Health Control, ASD: Autism Spectrum Disorder, ADHD: Attention Deficit Hyperactivity Disorder, MDD: Major Depression Disorder, SCZ: Schizophrenia)

Name	Usage	Size	Gender (M/F)	Age (mean±sd)	Samples
HBN [1]	Pre-training	2254	1455/799	10.73±3.39	-
HCP [27]		1080	495/585	(20-40)	-
QTIM [23]		1024	388/636	20.71±4.00	-
CNP [5]		261	152/109	33.29±9.29	-
ABIDE [8]	Evaluation	855	719/136	16.92±7.91	395 ASD, 460 HC
ADHD [7]		872	538/334	11.98±3.34	325 ADHD, 547 HC
SRPBS [24]		1397	799/598	38.37±13.65	125 ASD, 255 MDD, 147 SCZ, 783 HC, 87 Others

which contains 116 ROIs, we calculated FC as Pearson's correlation between BOLD signals of paired ROIs. The corresponding row for each node in the FCN matrix was treated as the node features.

Implementation Details. The CINP pre-training model was implemented with PyTorch 1.13.1 and MONAI 1.2.0. For pre-training, we utilized the Adam optimizer with an initial learning rate of 10^{-5} and a weight decay of 10^{-5}. The cosine annealing schedule was applied for the learning rate decreasing to 10^{-6}. The batch size was set to 256. The pre-training was performed on 8 NVIDIA A800 GPUs for 400 epochs, taking approximately 100 hours. Since the importance and scale of three losses are comparable, we set $\alpha = \beta = 1$ in (5) based on the validation set performance.

During the pre-training, the input 3D T1w MRI images were randomly augmented (i.e., Gaussian noise addition, flipping, intensity scaling and shifting) to learn robust representations. After augmentation, the volumes of all 3D T1 MRI images were resized to $96 \times 96 \times 96$. We evaluated CINP using the linear probe protocol, where embeddings from the pre-trained model were used as input to a support vector machine (SVM) classifier. Notably, only the embeddings of 3D T1-weighted MRI images generated by CINP were provided to the SVM in our experiments.

Performance Evaluation. The diagnosis of ASD and ADHD were treated as binary classification tasks on the ABIDE and ADHD datasets, respectively, and were evaluated using accuracy (ACC), area under the ROC curve (AUC), and Matthews correlation coefficient (MCC). For the SRPBS dataset, HC, ASD, MDD, SCZ, and other mental disorders were classified, where ACC and MCC were used for evaluation. For linear probe and fine-tuning protocols, we randomly divided the evaluation dataset into training (70%), validation (10%), and testing (20%) sets. Note that the network prompting protocol used the same data partitioning, but sampled only 10% of FCNs from the training set to simulate a low-data scenario, while 3D T1w MRI images in the testing set were evaluated.

Table 2. Classification results of different models on three mental disorder datasets.

Type	Method	Training	ABIDE			ADHD			SRPBS	
			ACC	AUC	MCC	ACC	AUC	MCC	ACC	MCC
FCN based	GCN [17]	From scratch	61.64	64.30	21.87	60.78	59.60	13.16	53.21	7.91
	BrainGNN [19]		55.79	58.67	9.57	57.23	55.31	12.26	53.08	18.75
	BrainNetCNN [16]		62.92	68.70	26.12	63.31	63.35	19.57	51.43	19.99
	MHAHGEL [28]		63.51	68.09	29.19	64.11	62.00	18.41	56.58	20.22
	BNT [15]		**65.96**	**72.00**	**31.80**	63.42	64.47	19.86	57.08	**29.07**
sMRI based	MedicalNet [6]	Linear probe	53.92	52.37	3.76	63.54	64.74	10.25	55.69	13.32
	PRCLv2 [33]		55.20	51.30	8.93	66.18	67.47	22.91	56.26	13.76
	Swin-UNETR [25]		55.79	54.17	9.50	66.63	67.58	23.72	56.33	14.39
	MedicalNet [6]	Fine tuning	54.39	51.25	4.56	65.71	68.38	24.09	58.57	14.10
	PRCLv2 [33]		54.60	59.65	14.19	67.82	68.23	25.00	57.50	11.74
	Swin-UNETR [25]		57.31	59.83	14.53	67.39	68.39	**26.33**	57.14	14.91
Multi modal	Cross-GNN [29]	From scratch	61.40	62.43	23.04	65.52	66.14	24.74	60.22	21.09
	MultiViT [4]		59.40	60.60	18.20	64.38	65.75	23.03	63.08	22.18
	CAMF [34]		59.05	59.55	15.44	66.67	**71.56**	24.21	61.29	22.10
-	CINP (Ours)	Linear probe	62.86	62.75	19.22	**69.08**	71.00	25.33	**64.29**	22.26

3.2 Quantitative Results

Comparision with Baseline Models. We compared our CINP using the linear probe protocol with FCN-based, sMRI-based, and multimodal models. Based on the implementation in the corresponding papers, four FCN-based and three multimodal models were deployed and trained from scratch on the three mental disorder datasets separately. For three sMRI-based models using the linear probe protocol, we used the pre-trained weights provided in their respective papers. The feature maps from the last layer of these models were flattened, reshaped, and fed into an SVM classifier for classification. We also fine-tuned sMRI-based models for 10 epochs and presented their performances.

The FCN Supervision Enhanced the Embeddings of 3D T1 MRI Images. As shown in Table 2, our CINP achieved the highest ACC on the ADHD and SRPBS datasets. Compared to the best results from sMRI-based and multimodal models, our CINP improved the ACC by 1.46%, 1.26%, and 1.21% on the ABIDE, ADHD, and SRPBS datasets, respectively. It indicates that by conducting contrastive learning between 3D T1 MRI images and FCNs, their mutually complementary information can be fully captured, benefitting mental disorder diagnosis.

The Diagnostic Efficacy of MRI Modalities Differed by Mental Disorder. Although our CINP outperformed all competing models on the ADHD and SRPBS datasets, it did not achieve state-of-the-art performance on the ABIDE dataset. Meanwhile, FCN-based models performed worse than all other types of models on the ADHD dataset. This may indicates that sMRI is more effective in diagnosing ADHD, while ASD identification may require more information about brain function, even though the knowledge has been transferred from FCN to sMRI during

Table 3. Classification results of the network prompting protocol with different numbers of group-level reference networks on two mental disorder datasets.

Reference Network Number (r)	ABIDE			ADHD		
	ACC	AUC	MCC	ACC	AUC	MCC
1	56.45	58.74	15.09	62.27	61.78	21.70
5	**62.56**	**61.34**	**25.84**	64.15	**65.33**	27.11
10	57.24	56.10	13.97	**64.66**	63.48	**29.16**

Table 4. Classification results of CINP with different combinations of pre-training objective functions on three mental disorder datasets.

Loss			ABIDE			ADHD			SRPBS	
$\mathcal{L}_{\text{INC}}$	$\mathcal{L}_{\text{MIM}}$	$\mathcal{L}_{\text{INM}}$	ACC	AUC	MCC	ACC	AUC	MCC	ACC	MCC
✓	✗	✗	58.57	59.94	17.94	62.86	61.20	11.68	58.93	13.64
✓	✓	✗	60.00	57.00	16.95	$\underline{66.23}$	$\underline{68.57}$	$\underline{20.67}$	60.71	$\underline{17.04}$
✓	✗	✓	$\underline{61.42}$	$\underline{60.13}$	**21.52**	63.37	64.69	17.50	$\underline{61.43}$	16.90
✓	✓	✓	**62.86**	**62.75**	$\underline{19.22}$	**69.08**	**71.00**	**25.33**	**64.29**	**22.26**

contrastive pre-training. Notably, on the SRPBS dataset, which included multiple mental disorders, multimodal models and CINP performed better, highlighting the importance of integrating sMRI and FCN for mental disorder subtyping.

Evaluation of Network Prompting. We evaluated the proposed network prompting protocol with different numbers of group-level reference networks ($r = 1, 5, 10$) on the ABIDE and ADHD datasets. As shown in Table 3, using only 10% of the FCNs in the training set of evaluation datasets, the CINP with the network prompting protocol achieved the best MCC (29.16%) on the ADHD dataset and outperformed all the sMRI-based and multimodal models on the ABIDE dataset. This demonstrates the feasibility of pre-training CINP with large-scale image-network pairs through contrastive learning and subsequently leveraging it with the network prompting protocol for mental disorder diagnosis, even when only a small number of FCNs from diagnosed patients are available.

3.3 Ablation Study

We conducted ablation study on the CINP variants, which were pre-trained with different combinations of the three objective functions in Section 2.1. As shown in Table 4, both the MIM and INM losses improved the performance of CINP, with the best performance achieved when all three objective functions were used. Specifically, on the ABIDE dataset, the MIM and INM losses improved the ACC by 1.44% and 2.86%, respectively; on the ADHD dataset, the improvements were 5.71% and 2.85%. Since the MIM loss primarily enhanced the representations of sMRI and the INM loss mainly transferred knowledge from FCNs to sMRI, this further implies that the diagnostic efficacy of MRI modalities varies across mental disorders from the objective function perspective.

4 Conclusion

In this paper, we proposed CINP, a framework that leverages contrastive learning between 3D T1 MRI and FCN. During pre-training, image-network contrastive loss, masked image modeling loss, and network-image matching loss were used to enhance the representations of 3D T1 MRI images with FCN supervision for downstream mental disorder diagnosis tasks. With the pre-trained CINP, we introduced network prompting to utilize only sMRI from suspected patients and a small amount of FCNs from different patient classes for diagnosing mental disorders. Extensive experiments on three mental disorder classification tasks demonstrated the effectiveness of CINP, which sheds light on the potential of incorporating sMRI into clinical diagnosis.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (grant number 62202442).

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

References

1. Alexander, L.M., Escalera, J., et al.: An open resource for transdiagnostic research in pediatric mental health and learning disorders. Scientific data **4**(1), 1–26 (2017)
2. Bannur, S., Hyland, S., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
3. Bastos, A.M., Schoffelen, J.M.: A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Frontiers in systems neuroscience **9**, 175 (2016)
4. Bi, Y., Abrol, A., Fu, Z., Calhoun, V.: Multivit: Multimodal vision transformer for schizophrenia prediction using structural mri and functional network connectivity data. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
5. Bilder, R., Poldrack, R., et al.: "ucla consortium for neuropsychiatric phenomics la5c study" (2020). https://doi.org/10.18112/openneuro.ds000030.v1.0.0
6. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
7. consortium, A..: The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in systems neuroscience **6**, 62 (2012)
8. Di Martino, A., Yan, C.G., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19**(6), 659–667 (2014)
9. Esteban, O., Markiewicz, C.J., et al.: fmriprep: a robust preprocessing pipeline for functional mri. Nature methods **16**(1), 111–116 (2019)
10. Fedorov, A., Geenjaar, E., Wu, L., et al.: Self-supervised multimodal learning for group inferences from mri data: Discovering disorder-relevant brain regions and multimodal links. NeuroImage **285**, 120485 (2024)

11. Gratton, C., Laumann, T.O., Nielsen, A.N., et al.: Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. Neuron **98**(2), 439–452 (2018)
12. He, Z., Li, W., Jiang, Y., Peng, Z., Wang, P., Li, X., Liu, T., Han, J., Zhang, T., Yuan, Y.: F2tnet: Fmri to t1w mri knowledge transfer network for brain multi-phenotype prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 265–275. Springer (2024)
13. Hu, X., Xiao, L., Sun, X., Wu, F.: Overall survival time prediction of glioblastoma on preoperative mri using lesion network mapping. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 298–307. Springer (2023)
14. Hu, X., Xiao, L., Wang, Y.P.: A graph neural network based fusion of mri-derived brain network and clinical data for glioblastoma survival prediction. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1571–1575. IEEE (2024)
15. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. Advances in Neural Information Processing Systems **35**, 25586–25599 (2022)
16. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., et al.: Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
17. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
18. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
19. Li, X., Zhou, Y., Dvornek, N., et al.: Braingnn: Interpretable brain graph neural network for fmri analysis. Medical Image Analysis **74**, 102233 (2021)
20. Logothetis, N.K.: What we can do and what we cannot do with fmri. Nature **453**(7197), 869–878 (2008)
21. Pang, J.C., Aquino, K.M., Oldehinkel, M., Robinson, P.A., Fulcher, B.D., et al.: Geometric constraints on human brain function. Nature **618**(7965), 566–574 (2023)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
23. Strike, L.T., Blokland, G.A., Hansell, N.K., Martin, N.G., Toga, A.W., Thompson, P.M., de Zubicaray, G.I., McMahon, K.L., Wright, M.J.: "queensland twin imaging (qtim)" (2023). https://doi.org/doi:10.18112/openneuro.ds004169.v1.0.7
24. Tanaka, S.C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., et al.: A multi-site, multi-disorder resting-state magnetic resonance image database. Scientific data **8**(1), 227 (2021)
25. Tang, Y., Yang, D., Li, W., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
26. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., et al.: Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage **15**(1), 273–289 (2002)
27. Van Essen, D.C., Smith, S.M., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
28. Wang, W., Xiao, L., Qu, G., Calhoun, V.D., Wang, Y.P., et al.: Multiview hyperedge-aware hypergraph embedding learning for multisite, multiatlas fmri based functional connectivity network analysis. Medical Image Analysis **94**, 103144 (2024)

29. Yang, Y., Ye, C., Guo, X., Wu, T., Xiang, Y., Ma, T.: Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning. IEEE Transactions on Medical Imaging **43**(1), 108–121 (2023)
30. Yang, Y., Ye, C., Sun, J., Liang, L., Lv, H., Gao, L., Fang, J., Ma, T., Wu, T.: Alteration of brain structural connectivity in progression of parkinson's disease: A connectome-wide network analysis. NeuroImage: Clinical **31**, 102715 (2021)
31. Ye, K., Tang, H., Dai, S., Guo, L., Liu, J.Y., Wang, Y., Leow, A., Thompson, P.M., Huang, H., Zhan, L.: Bidirectional mapping with contrastive learning on multimodal neuroimaging data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 138–148. Springer (2023)
32. Zhang, Y., Zhang, H., Chen, X., Liu, M., Zhu, X., Lee, S.W., Shen, D.: Strength and similarity guided group-level brain functional network construction for mci diagnosis. Pattern Recognition **88**, 421–430 (2019)
33. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
34. Zhou, Z., Orlichenko, A., Qu, G., Fu, Z., Calhoun, V.D., Ding, Z., Wang, Y.P.: An interpretable cross-attentive multi-modal mri fusion framework for schizophrenia diagnosis. arXiv preprint arXiv:2404.00144 (2024)