S3F: A MULTI-VIEW SLOW-FAST NETWORK FOR ALZHEIMER'S DISEASE DIAGNOSIS

Ziqiao Weng[1], *Jingjing Meng*[1], *Zhaohua Ding*[2], *Junsong Yuan*[1]

[1]University at Buffalo, Buffalo, NY, USA
[2]Vanderbilt University, Nashville, TN, USA

alexziqiaoweng@gmail.com, jmeng2@buffalo.edu, zhaohua.ding@vanderbilt.edu, jsyuan@buffalo.edu

ABSTRACT

Alzheimer's disease (AD) is the most common form of dementia in the elderly. As early detection and diagnosis is imperative for the intervention and prevention of its progression into more detrimental stages, pioneering works have been proposed that use the resting-state functional MRI (rs-fMRI) to identify early mild cognitive impairment (EMCI) based on various convolutional neural networks (CNNs). However the accuracy is not satisfactory. In this paper, we propose a multi-view model based on the SlowFast network, a recently proposed model for video recognition. The rs-fMRI data are treated as videos from three perspectives (i.e. coronal, horizontal and sagittal, corresponding to three anatomical planes in human body) and the jointly learned hierarchical representations are fused in the fully connected layer. We examine our model on a publicly accessible Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Our method significantly outperforms other competing methods and achieves state-of-the-art accuracy. Besides, we also provide a baseline on the classification task over all clinical phases of AD.

Index Terms— Alzheimer's disease, resting-state functional MRI, convolutional neural networks

1. INTRODUCTION

Alzheimer's Disease (AD) is a severe neurological disease and is also a type of dementia, a slowly progressive mental deterioration caused by generalized degeneration of the brain, which can occur in middle or old age, especially for the elderly. Although existing methods can temporarily slow down the degeneration process, currently there is no cure for this disease. The property of AD makes the detection of it at its early stage (as early as possible) is of great importance for early intervention [1].

Recently, the rapid development of deep learning techniques allows an increasing body of researches to use deep learning methods combined with biological marks to diagnose AD [2] more precisely.

Most of deep learning assisted AD diagnosis methods leverage structural Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) modal data [3] [4] [5]. MRI exquisitely pictures the anatomy and physiological processes of the body [6] [7]. PET biochemically analyzes changing metabolic patterns of body tissues of brain development.

However, when diagnosing AD at an early stage, MRI or PET can provide little visible structural and metabolic changes within the brain. Thus, early AD identification studies opt to utilize the resting-state functional MRI (rs-fMRI), which is collected from subjects at resting state without being engaged in a given task. rs-fMRI reveals more informative metabolic blood-oxygen-level dependent (BOLD) signals in dynamic dementia procedure and cerebral functional information [8]. Besides, the brain functional networks (BFNs), generated from functional connectivity (FC) between various brain regions based on temporal BOLD signals, have been widely applied for MCI detection [9], [10]. Moreover, pioneering studies have been focusing on a preclinical stage even earlier than MCI, called Early MCI (EMCI), in order to prevent potential memory and thinking deterioration progress [11], [12], [13], [1] [14]. Kam *et al.* exploit group information guided ICA derived BFNs to construct basic 3D CNN to do EMCI diagnosis and fuse the different BFNs to jointly learn embedded representations [11]. This study was further extended to take the consideration of both static BFNs (sBFNs) and dynamic BFNs (dBFNs), assuming the functional connectivity (FC) will not change during scanning period [12].

Although prior studies succeed in making full use of extensive leading and complicated brain imaging analytical methods (e.g., BFNs) (which require a lot of clinical or cognitive domain knowledge and expert's real-world experience), these approaches adopt mostly 2D convolutional neural networks (CNNs) (e.g. DenseNet [15]), which results in unsatisfactory performance.

Therefore, to overcome these difficulties, in this work, we propose a novel generic multi-view model to learn highly-complex representations hidden in spatiotemporal rs-fMRI data for AD diagnosis. Our work is also motivated by algorithms in video analysis [16], [17], [18], [19], [20] and multi-view learning [21], [22], [23].

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

To summarize, the contributions of the paper are the following:

- Our approach provides an innovative way to model the spatiotemporal rs-fMRI as 3-dimensional videos with different spatial and temporal capacity. Thus the highly-complex and hierarchical features of rs-fMRI can be captured for better AD diagnosis.
- We propose a multiview network that models the anatomical planes of human brain in rs-fMRI data and fuses the jointly learned representations in the fully connected layer for final prediction.
- We enhance the backbone network [24] with a powerful channel-wise self-attention module [25] and exploit a variant of focal loss to alleviate the disequilibrium of data distribution.
- We provide a baseline for all-phase AD classification task, including clinically normal (CN), subjective memory concerns (SMC), early mild cognitive impairment (EMCI), mild cognitive impairme (MCI), late mild cognitive impairment (LMCI) and Alzheimer's disease (AD). Experiments conducted on challenging benchmarks show our proposed approach outperforms the state-of-the-art techniques.

2. METHOD

This section introduces our proposed pipeline. We propose to improve AD diagnosis in terms of classification and dementia detection by applying 3D CNN to operate well-normalized spatiotemporal rs-fMRI data from three anatomical directions (see Fig. 1).

2.1. Data normalization

The raw resting-state functional MRI (rs-fMRI) data in DICOM format are downloaded from the publicly accessible Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Besides, the rs-fMRIs are obtained by 3T Philips scanners at multiple sites with scrupulous quality control (QC). The parameters of rs-fMRIs are: slice number = 48, volume number = 140, repetition time (TR) = 3,000 ms, echo time (TE) = 30 ms, voxel size = 3×3×3.3 mm^3, flip Angle=80.0 degree.

For normalization of the rs-fMRIs, we use the user-friendly Data Processing Assistant for Resting-State fMRI (DPARSF) toolbox in Matlab platform, which handles the rs-fMRIs in follwing steps [26]: **DICOM to NIfTI**: converting the DICOM data into Neuroimaging Informatics Technology Initiative (NIfTI) data, which contains crucial spatial information. **Slice timing**: correcting the time differences of 2D image acquisition between slices. **Head motion correction**: excluding head motion disturbance and adjusting the brain image in each slices to the same area. **Normalization**: normalizing the rs-fMRIs into standard spatial space (i.e. the Montreal Neurosciences Institute space) for objective inter-subject comparisons. **Smoothing and Filtering**: applying Gaussian filter to average image intensity in voxel level to reduce noise or oscillation.

The spatiotemporal size of rs-fMRI is denoted as $[W, H, S, T]$, where W and H are width and height of each slice, F is the slice number, T is the temporal length (i.e. volumes).

2.2. Framework Overview

2.2.1. Spatial compression and temporal decomposition

The rs-fMRI data goes through above normalization pipeline still encompasses redundant information, where the voxel intensity equal to or close to zero. Thus, we condense the rs-fMRI data by discarding first 6 slices and last 6 slices of each volume in coronal orientation (*i.e.*, the second dimension of 4D rs-fMRI), which also makes rs-fMRI more suitable for our designed framework. Furthermore, the size of rs-fMRI data is compressed to $[61, 61, 61, T]$.

Then we decompose 4D rs-fMRI to multiple 3D volumes as the input data by using Statistical Parametric Mapping (SPM) software package. Finally, we adopt Z-Score standardization to further normalize them into 0-1 voxel space.

2.2.2. S3F Network

The CNN used as the backbone of our framework is the SlowFast network [24], which is a dual-stream model for video recognition, incorporating a Slow stream and a Fast stream.

Slow stream: The Slow stream is designed to capture spatial semantic information by analyzing a video at low frame rates and slow refreshing speed. It has a large temporal stride τ (*i.e.* samples only one out of every τ frames). We denote the total sampled frames as F, where it is set to 16 in our experiments.

Fast stream: The Fast stream is good at catching rapid temporal motions in high frame rates and a refreshing speed. It holds a comparatively small temporal stride τ/α, where we assign 1/8 to α in our experiments. It indicates that the temporal capacity of the Fast stream is 8 times as large as that of the Slow stream.

Meanwhile, the Fast stream maintains a low channel capacity (fewer kernel channels). The number of output channels of the Slow stream is C and its Fast to Slow ratio is β, which is set to 1/8 in our experiments.

High to low frequency connection: As the captured information of Fast stream is high temporal frequency and Slow stream is lower temporal frequency, this lateral stream from Fast to Slow will share and fuse the learned features of Fast to

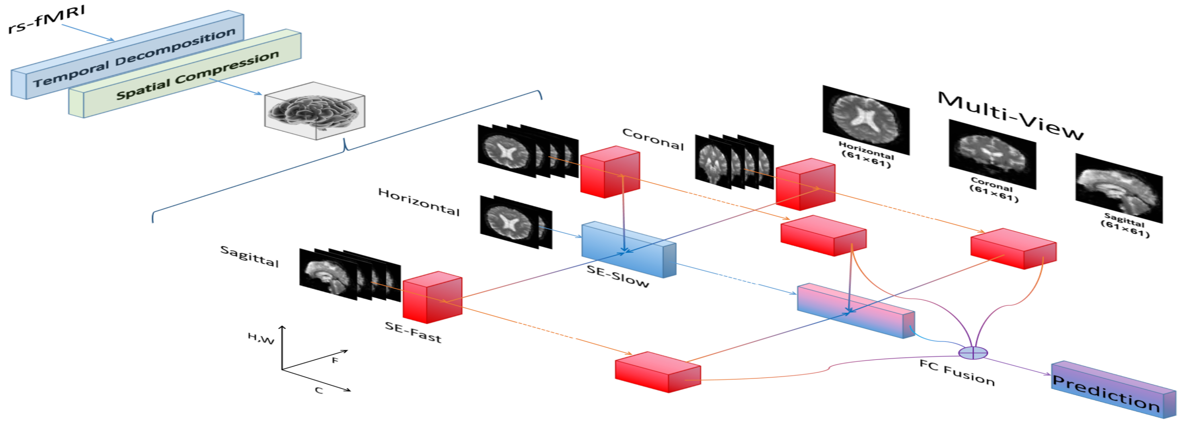

Fig. 1. Overview of Multi-view S3F. The red and blue blocks represent the Fast stream and Slow stream respectively, both of which are equipped with the SE-layer. The Fast stream samples more frames but outputs fewer channels, comparing with the Slow stream. Lateral information connections from the Fast stream to the Slow stream repeats 4 times. Best viewed in color.

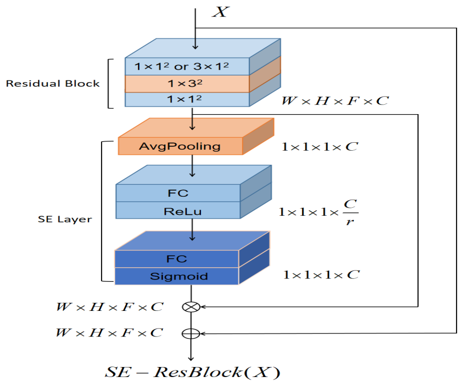

Fig. 2. The stack of Residual module and SE layer.

Slow by concatenation in channel level. Additionally, to address the problem of unmatched feature shapes of Slow and Fast stream ($[F, W\times H, C]$ for Slow and $[\alpha F, W\times H, \beta C]$ for Fast), a time-stride 3D convolution layer with 5×1^2 kernel, $\alpha\times 1^2$ stride as well as $2\beta C$ output channels is performed.

SE-ResNet: We integrate the lightweight SE-Layer [25] to our 3D Residual Block. Typically, SE-Layer has two steps: squeeze and excitation, which are designed to learn channel-wise weights to explore more important information (similar to attention mechanism). The squeeze operation squeezes the learned spatial representations into a channel descriptor by using adaptive average pooling. The excitation operation opt to catch dependencies across multiple channels by firstly introducing a dimensionality-reduction fully-connected (FC) layer followed by ReLU function to learn nonlinear cooperation between channels, where the reduction factor r is used to reduce model complexity and overall number of parameters. Secondly, a fully connected layer returns the channel dimension back and a final sigmoid layer scales the output into 0-1 space. In the end, an emphasized representation will be obtained by multiplying channel scale factors and original input feature maps. A diagram exhibiting the structure of our SE-ResNet is shown in Fig. 2.

Multi-View: The Multi-View concept is inspired by human anatomy structure that human body has three anatomical planes, we extend the single view input to three anatomical view inputs, which we call Multi-vie:

Horizontal view: divides the brain into upper to lower slices. Coronal view: divides the brain into front to back slices. Sagittal view: divides the brain into left to right slices.

The results in [24] prove that the Slow stream plays a major role and the Fast stream brings an auxiliary enhancement in the SlowFast network. Thus, in our framework, one Slow stream is used to process the rs-fMRI from a given view (this view choice will be discussed in ablation study) and three auxiliary Fast streams are utilized for the three distinctive perspectives. We denote this network as S3F.

At the end of our pipeline, four fully connected layers through the S3F network are fused by concatenation for final prediction.

3. EXPERIMENTS: EMCI CLASSIFICATION

We first evaluate our framework on the EMCI classification task using standard evaluation protocols.

3.1. Dataset

We use the raw resting-state functional MRI (rs-fMRI) data from 54 EMCI subjects and 50 CN subjects from the ADNI database. And for each subject we download all available valid records longitudinally. Thus, following [12],we finally collect a total of 351 samples (172 and 179 for CN and EMCI, respectively). Then, by splitting every 4D data to multiple 3D data, including 140 brain scans, the dataset expand to a total of 49,140 3D brain scans, consisting of 24,080 and 25,060 samples for CN and EMCI respectively.

3.2. Experimental Settings

To validate the effectiveness of our proposed Multi-View model and provide fair comparisons with recent methods, we follow [12] to adopt 5-fold cross-validation. Notice that, since most subjects have multiple scans and to simulate the real-world application scenario, we only consider the baseline scan (*i.e.*, earliest scan) of each subject for testing whereas the training dataset includes all the scans of all subjects. Our model is implemented on Pytorch. The Multi-View network is optimized by the Stochastic Gradient Descent (SGD) algorithm with a learning rate of 1×10^{-3} and a momentum of 0.9. As for loss, we adopt the cross-entropy loss. The batch size for training is 30, whereas we use the rs-fMRI data without the temporal decomposition for testing.

3.3. Metrics Evaluation

In this paper, CN and EMCI represent positive and negative instances, respectively. We use the following criteria for quantitative evaluation:

- Accuracy (ACC) = (TP + TN)/(TP + TN + FP + FN)
- Sensitivity (SEN) = TP/(TP + FN)
- Specificity (SPC) = TN/(TN + FP)
- Positive Predictive Value (PPV) = TP/(TP + FP)
- Negative Predictive Value (NPV) = TN/(TN + FN)

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

3.4. Baselines

For a comprehensive evaluation, we perform comparative studies of our method against the state-of-the-art deep learning based EMCI diagnosis methods: simple brain functional networks (BFN) based 3D CNN (SB-CNN) and multiple-BFN-based 3D CNN (MB-CNN) [11]; static and dynamic SB-CNN (sdSB-CNN) and static and dynamic multiple SB-CNN (sdMB-CNN) [12]. Six BFNs have been tested in their methods, including default-mode network (DMN); two fronto-parietal networks (FPN 1&2); two attention networks (AN 1&2); executive control network (ECN). MB-CNN is a unified model that fuses several or all SB-CNNs to fully connected layers for joint feature learning.

We also compare our model with SlowFast (SF) and Multiview-SlowFast (Multiview-SF), where each anatomical view simply corresponds to a SlowFast network.

3.5. Results

For each method, We evaluate its performance on using entire brain rs-fMRI data and only white matter region extracted from corresponding rs-fMRI data by using a unified white mask. We also report the subject-wise (*i.e.*, majority vote) results, which are more suitable in real clinical scenarios.

The results compared with state-of-the-art methods are reported in Table 1. In comparison with the previous state-of-the-art (sdMB-CNN) [9] methods, our best model (SE-S3F-H) provides about **10%** higher accuracy. Notably, almost all our results are better than existing results.

The results from using only white matter consistently has a significant overall performance improvement, compared with that from using entire brain rs-fMRI data, indicating that the white matter contains more useful clinical information.

The S3F-H (with 37.39M parameters) slightly outperforms the straightforward Multiview SF (with 100.67M parameters) but holds much less parameters. And it is substantially better than the SlowFast (with 33.55M parameters). The performance of S3F-H is further enhanced by the SE-S3F-H (embedded with SE-Layer), which leads to 2% improvements over ACC and about 10% improvements over SPC and NPV, confirming the strength of the SE layer in detecting and classifying EMCI subjects.

Table 1. Results

Method		ACC (%)	SEN (%)	SPC (%)	PPV (%)	NPV (%)
SB-CNN [11]	(DMN)	70.11	71.96	68.36	70.03	70.86
	(FPN1)	69.45	69.73	69.07	69.75	69.11
	(FPN2)	69.04	69.33	68.67	69.33	68.67
	(AN1)	68.02	69.42	66.53	68.33	67.94
	(AN2)	66.96	68.58	65.29	67.06	67.03
	(ECN)	67.19	67.64	66.62	67.65	66.96
MB-CNN [11]	(ALL)	73.85	73.91	73.69	74.38	73.79
sdSB-CNN [12]	(DMN)	73.17	73.91	72.44	73.70	73.03
	(FPN1)	72.80	73.07	72.44	73.30	72.86
	(FPN2)	71.53	71.07	72.13	72.51	71.62
	(AN1)	70.27	70.98	69.42	70.59	70.64
	(AN2)	69.08	68.27	70.00	70.69	68.48
sdMB-CNN [12]	(ALL)	76.07	76.27	75.87	76.55	75.93
Ours (SF)	(ALL)	**76.03**	**68.71**	**82.61**	**76.81**	**81.51**
	(ALL-VOTE)	**76.91**	**70.00**	**83.09**	**77.99**	**82.21**
	(WM)	**77.92**	**79.51**	**75.93**	**85.92**	**77.02**
	(WM-VOTE)	**77.81**	**80.00**	**75.27**	**86.13**	**76.53**
Ours (Multiview-SF)	(ALL)	**77.79**	**79.23**	**76.13**	**83.24**	**75.92**
	(ALL-VOTE)	**77.91**	**80.00**	**75.64**	**83.47**	**75.65**
	(WM)	**84.05**	**79.66**	**87.86**	**85.87**	**86.76**
	(WM-VOTE)	**84.67**	**80.00**	**88.73**	**86.56**	**89.39**
Ours (S3F-H)	(ALL)	**78.24**	**73.5**	**82.46**	**78.96**	**81.51**
	(ALL-VOTE)	**78.81**	**74.00**	**83.09**	**79.45**	**82.43**
	(WM)	**84.51**	**79.61**	**88.75**	**85.57**	**87.05**
	(WM-VOTE)	**86.52**	**82.00**	**90.36**	**88.89**	**91.34**
Ours (SE-S3F-H)	(ALL)	**80.70**	**67.83**	**92.47**	**77.26**	**91.13**
	(ALL-VOTE)	**80.81**	**68.00**	**92.55**	**77.16**	**90.81**
	(WM)	**86.91**	**88.36**	**85.33**	**91.97**	**86.92**
	(WM-VOTE)	**87.48**	**90.00**	**84.91**	**93.75**	**87.13**

3.6. Ablation Study

We also provide an ablation study on varying the Slow stream of the SE-S3F network and the results are shown in Table 2. When using the whole brain region or white matter region as input, the H (i.e. Horizontal) view is better than the other two views, especially in terms of using only white matter region. However, SE-S3F-S achieves more balanced results across five metrics (i.e. ACC, SEN, SPC, PPV and NPV), where each of them is around 80%. Moreover, all variants of our SE-S3F network are better than the state-of-the-art methods.

Table 2. Ablation Study

Method		ACC (%)	SEN (%)	SPC (%)	PPV (%)	NPV (%)
Ours (SE-S3F-H)	(ALL)	**80.70**	**67.83**	**92.47**	**77.26**	**91.13**
	(ALL-VOTE)	**80.81**	**68.00**	**92.55**	**77.16**	**90.81**
	(WM)	**86.91**	**88.36**	**85.33**	**91.97**	**86.92**
	(WM-VOTE)	**87.48**	**90.00**	**84.91**	**93.75**	**87.13**
Ours (SE-S3F-C)	(ALL)	**76.20**	**64.76**	**86.46**	**73.53**	**85.86**
	(ALL-VOTE)	**76.86**	**66.00**	**86.55**	**74.56**	**86.00**
	(WM)	**82.98**	**76.94**	**88.26**	**86.11**	**88.80**
	(WM-VOTE)	**82.76**	**76.00**	**88.73**	**85.85**	**89.11**
Ours (SE-S3F-S)	(ALL)	**81.20**	**80.33**	**81.97**	**82.16**	**81.79**
	(ALL-VOTE)	**80.76**	**80.00**	**81.46**	**81.76**	**81.35**
	(WM)	**81.35**	**75.67**	**86.22**	**80.68**	**85.44**
	(WM-VOTE)	**81.67**	**76.00**	**86.55**	**80.82**	**86.13**

4. CLASSIFICATION ON ALL PHASES OF AD

In this part, we carry out the experiment on a dataset over all clinical phases of Alzheimer's Disease, with the same evaluation metrics as aforementioned.

4.1. Dataset

This dataset is taken from AD's all clinical phases from ADNI, consisting of clinically normal (CN), subjective memory concerns (SMC), early mild cognitive impairment (EMCI), mild cognitive impairme (MCI), late mild cognitive impairment (LMCI) and Alzheimer's disease (AD). We normalize the raw rs-fMRI data from 119 CN subjects, 43 SMC subjects, 83 EMCI subjects, 35 MCI subjects, 33 LMCI subjects and 34 AD subjects. The total decomposed sample sizes of CN, SMC, EMCI, MCI, LMCI and AD are 23524, 8009, 13588, 6751, 5610 and 5168, respectively.

4.2. Experimental Settings

We follow previous 5-fold cross-validation to evaluate our model. Other settings are the same as in previous experiments.

Here, we embed the GHM-C loss into the cross-entropy loss to address the imbalance problem of data distribution [27]. The mechanism of the GHM-C loss is that it can dynamically harmonize the overall gradient contribution of diverse classes, where the large samples as well as the outliers are down-weighted gradually.

Table 3. Results: all stages of AD

Method	ACC (%)	CN (%)	SMC (%)	EMCI (%)	MCI (%)	LMCI (%)	AD (%)
Baseline (SF)	**50.91**	70.84	20.96	53.64	40.30	8.57	58.02
Multiview-SF	**52.88**	77.47	29.15	49.65	30.50	8.50	63.34
SE-S3F-H	**56.44**	86.75	33.15	53.25	20.67	0.00	70.08
SE-S3F-H + GHMC	**56.98**	83.25	21.05	59.33	24.80	7.69	83.54
SE-S3F-S	**53.57**	82.80	29.45	51.05	25.17	8.90	48.39
SE-S3F-S + GHMC	**57.85**	84.65	33.69	53.16	28.47	9.03	77.69

4.3. Results

We have trained the SlowFast network as the baseline of this task. Table 3 reports the performance of all above-mentioned methods. Note that we only use sagittal and horizontal views for SE-S3F, due to their better performance in previous experiment. It is seen that both SE-S3F-H and SE-S3F-S exhibit better performance compared with the Multiview-SF, which generally outperforms the baseline. Additionally, the GHM-C loss brings mild improvements for both SE-S3F-H and SE-S3F-S networks. The performance of the GHM-C loss SE-S3F-S is better than that of SE-S3F-H, which indicates that GHM-C improves the S3F model with different views of Slow stream to different degrees.

5. CONCLUSION

In this paper, we proposed a multi-view learning (i.e. S3F) model for Alzheimer's disease diagnosis. Motivated by the SlowFast network [24], a recently proposed model for video recognition, our model views spatiotemporal rs-fMRIs from three perspectives (*i.e.*, coronal, horizontal and sagittal, corresponding to three anatomical planes in human body) and fuses the jointly learned deep representations in the fully connected layers for final prediction. Experimental results verify the effectiveness of the proposed model with state-of-the-art accuracy over other competing methods.

6. REFERENCES

[1] Alzheimer's Association et al., "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] Dinggang Shen, Guorong Wu, and Heung-Il Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[3] Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu,

et al., "Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment," *Frontiers in neuroscience*, vol. 12, 2018.

[4] Mingxia Liu, Jun Zhang, Dong Nie, Pew-Thian Yap, and Dinggang Shen, "Anatomical landmark based deep feature representation for mr images in brain disease diagnosis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1476–1485, 2018.

[5] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen, "Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[6] Viktor Wegmayr, Sai Aitharaju, and Joachim Buhmann, "Classification of brain mri with big data and deep 3d convolutional neural networks," in *Medical Imaging 2018: Computer-Aided Diagnosis*. International Society for Optics and Photonics, 2018, vol. 10575, p. 105751S.

[7] Manhua Liu, Danni Cheng, Kundong Wang, Yaping Wang, Alzheimer's Disease Neuroimaging Initiative, et al., "Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, no. 3-4, pp. 295–308, 2018.

[8] Daniel A Orringer, "Resting-state fmri for the masses," *Journal of Neurosurgery*, vol. 1, no. aop, pp. 1–2, 2018.

[9] Renping Yu, Lishan Qiao, Mingming Chen, Seong-Whan Lee, Xuan Fei, and Dinggang Shen, "Weighted graph regularized sparse brain network construction for mci identification," *Pattern Recognition*, vol. 90, pp. 220–231, 2019.

[10] Xiaobo Chen, Han Zhang, Lichi Zhang, Celina Shen, Seong-Whan Lee, and Dinggang Shen, "Extraction of dynamic functional connectivity from brain grey matter and white matter for mci classification," *Human brain mapping*, vol. 38, no. 10, pp. 5019–5034, 2017.

[11] Tae-Eui Kam, Han Zhang, and Dinggang Shen, "A novel deep learning framework on brain functional networks for early mci diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 293–301.

[12] Tae-Eui Kam, Han Zhang, Zhicheng Jiao, and Dinggang Shen, "Deep learning of static and dynamic brain functional networks for early mci detection.," *IEEE transactions on medical imaging*, 2019.

[13] Emimal Jabason, M Omair Ahmad, and MN S Swamy, "Deep structural and clinical feature learning for semi-supervised multiclass prediction of alzheimer's disease," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2018, pp. 791–794.

[14] Zhuqing Jiao, Zhengwang Xia, Xuelian Ming, Chun Cheng, and Shui-Hua Wang, "Multi-scale feature combination of brain functional network for emci classification," *IEEE Access*, vol. 7, pp. 74263–74273, 2019.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[16] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[17] Limin Wang, Wei Li, Wen Li, and Luc Van Gool, "Appearance-and-relation networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1430–1439.

[18] Michał Zajac, Konrad Zołna, Negar Rostamzadeh, and Pedro O Pinheiro, "Adversarial framing for image and video classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 10077–10078.

[19] Haisheng Su, Xu Zhao, Tianwei Lin, and Haiping Fei, "Weakly supervised temporal action detection with shot-based temporal pooling network," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 426–436.

[20] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma, "T-c3d: Temporal convolutional 3d network for real-time action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[21] Tan Yu, Jingjing Meng, and Junsong Yuan, "Multi-view harmonized bilinear network for 3d object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 186–194.

[22] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann, "Robust 3d hand pose estimation from single depth images using multi-view cnns," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4422–4436, 2018.

[23] Jingjing Meng, Suchen Wang, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan, "Video summarization via multi-view representative selection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1189–1198.

[24] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6202–6211.

[25] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[26] Chaogan Yan and Yufeng Zang, "Dparsf: a matlab toolbox for" pipeline" data analysis of resting-state fmri," *Frontiers in systems neuroscience*, vol. 4, pp. 13, 2010.

[27] Buyu Li, Yu Liu, and Xiaogang Wang, "Gradient harmonized single-stage detector," in *AAAI Conference on Artificial Intelligence*, 2019.