A *Wasserstein* Recipe for Replicable Machine Learning on Functional Neuroimages

Jiaqi Ding[1,2], Tingting Dan[2], Ziquan Wei[1,2], Paul Laurienti[3], and Guorong Wu[1,2]

[1] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA
jiaqid@cs.unc.edu ziquanw@email.unc.edu

[2] Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA
{grwu, Tingting_Dan}@med.unc.edu

[3] Department of Radiology, Wake Forest University School of Medicine, Winston-Salem, NC 27101, USA
plaurien@wakehealth.edu

Abstract. Advances in neuroimaging have dramatically expanded our ability to probe the neurobiological bases of behavior *in-vivo*. Leveraging a growing repository of publicly available neuroimaging data, there is a surging interest for utilizing machine learning (ML) approaches to explore new questions in neuroscience. Despite the impressive achievements of current deep learning models, there remains an under-acknowledged risk: the variability in cognitive states may undermine the experimental replicability of the ML models, leading to potentially misleading findings in the realm of neuroscience. To address this challenge, we first dissect the critical (but often missed) challenge of ensuring the replicability of predictions despite task-irrelevant functional fluctuations. We then formulate the solution as a domain adaptation, where we design a dual-branch Transformer with minimizing *Wasserstein* distance. We evaluate the cognitive task recognition accuracy and consistency of test and retest functional neuroimages (serial imaging measures of the same cognitive task over a short period of time) of the Human Connectome Project. Our model demonstrates significant improvements in both replicability and accuracy of task recognition, showing the great potential of reliable deep models for solving real-world neuroscience problems.

Keywords: Functional MRI · Model Replicability · Cognitive Task Recognition

1 Introduction

A crucial challenge in the field of neuroscience is to comprehend the emergence of cognition and behavior because of brain activity [18]. The technique of functional magnetic resonance imaging (fMRI) offers a non-invasive method to observe cerebral blood flow changes, which correlate with brain activity [14]. In

addition, the vast repository of publicly available functional neuroimaging data presents opportunities to identify potential imaging biomarkers and uncover new neurobiological pathways through deep learning methodologies [25, 11].

In the wake of the success of Transformers in NLP and CV [24], they also have shown significant advancements in interpreting neural activities from fMRI BOLD signals [5]. This entails recognizing cognition-specific spatiotemporal patterns within extensive neuroimaging datasets [3, 26, 23]. For instance, fused window transformers [3] trained on BOLD time series data can classify seven cognitive tasks from the HCP-task dataset [1].

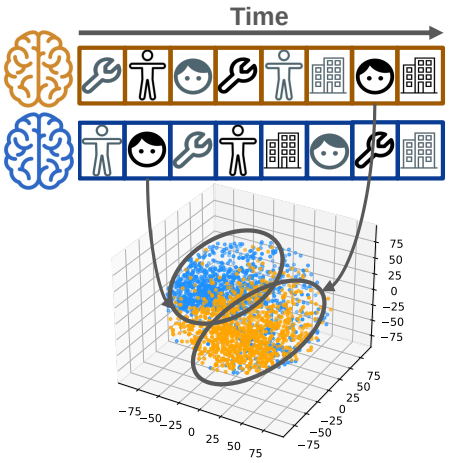

Fig. 1: Challenge of learning replicability in functional neuroimaging studies. *Top:* Test and retest fMRI scans (from HCP). *Bottom:* The feature representations on test (in orange) and retest fMRI (in blue) manifest distinct distributions on one example of the cognitive tasks.

Real-World Challenges of Machine Learning in Cognitive Neuroscience. The human brain is a complex network that undergoes continuous functional fluctuations over time [2]. Increasing evidence suggests that these transient states are fundamental to the dynamic nature of our cognitive functions. In scientific researches, it is crucial to ensure replicable findings across repeated experiments, yielding statistically consistent results for the same subject under identical conditions [27]. However, the concept of *test-retest reliability*—the consistency of successive measurements under unchanged conditions [16]—remains insufficiently integrated into the development and evaluation of new deep learning models in neuroscience. For example, the application of the well-known BERT to fMRI data [23] lacks evaluation of task prediction reliability on retest data, this oversight may undermine the reliability of findings crucial for valid scientific conclusions in neuroimaging studies.

Exploring the issue further, we delve into the origins of variability, observing the different feature distributions that stem from differences between subjects' brain functions, which are influenced by biological factors [22] like neural oscillation, and fluctuations within the same subject due to previous cognitive activities (the order of multiple pre-designed cognitive tasks varies between test and retest scans). The issue of model replicability is illustrated in Fig.1, where we attempt

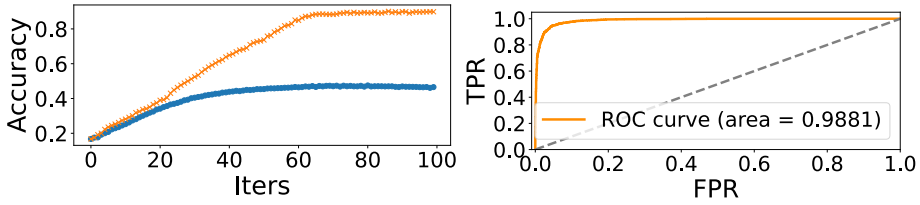

Fig. 2: *Left:* The ACC curves of task recognition on scan 2 by model exclusively trained on scan 1 (in blue) and model trained on a mixed dataset including both scan 1 and scan 2 (in orange). *Right:* The ROC curve of adversarial validation between scan 1 and scan 2.

to account for external factors that could sway brain activity readings, such as the state of the participant when applied to the real-world neuroscience studies.

Why did the current deep models fail on test/retest replicability? Our results (Fig.2 left) display the task recognition results on real fMRI data, analyzed by a Transformer model trained solely on the test data, yet it underperforms when predicting on the retest data, as illustrated by the blue curve.

We surmise that this limited generalizability might result from overlooking the variability in functional brain dynamics within subjects across diverse cognitive tasks. To investigate it, we employ adversarial validation technique [9] to discern differences between test and retest data. We use a binary classifier to determine if the sample belongs to test or retest data. The high performance of adversarial validation (AUC=0.9881, ACC=0.9129 in Fig.2 right) indicates the disparate distributions between the two datasets. One potential solution for that involves training the models on a mixed dataset including test and retest to capture more diverse features. This approach, as shown by the orange curve in Fig.2 left, notably enhances prediction accuracy for both datasets. However, supervised learning requires corresponding labels, limiting applicability to new, unlabeled datasets.

Building upon the concept of domain adaptation [17, 4, 8], we propose a novel dual-branch transformer with minimizing Wasserstein distance tailored for functional neuroimaging. This model integrates the alignment of distributions with the feature representations learning within an end-to-end framework. This strategy is designed to bolster the model's replicability of cognitive task recognition agnostic for biological (external) factors across different fMRI scans.

2 Method

Model Overview. Within the framework of domain adaptation, the goal is to craft a model that, once trained on the *source domain* (S), demonstrates accurate predictions for the *target domain* (T). The focus is therefore on model's ability of learning task-specific and domain-invariant feature representations through the lens of unsupervised classification on retest data.

Suppose we partition the brain into N regions. Then we extract the average BOLD signal in each region over time t, denoted by $x_n^t (n = 1, ..., N)$. Following the definition in [3], we consider the snapshot $x^t = [x_1^t, x_2^t, ..., x_N^t]$ as a vector that expresses the whole-brain neural activity at time t. Since each task is scheduled for a certain time period, we assign a label $l_m (m = 1, ..., M)$ of cognitive task for each time segment of BOLD snapshots $x = [x^t | t = 1, ..., \Gamma]$[4]. Thus, BOLD segments dissembled from the test fMRI data form the source domain X_S where each time segment of BOLD is associated with cognitive task labels. On the other hand, BOLD segments extracted from the retest fMRI data form the target domain X_T, where the task label for each segment x is unknown.

The network architecture of our approach is depicted in Fig.3, showcasing an end-to-end domain adaptive transformer.

Token embedding for BOLD snapshots. First, we project each vector of BOLD snapshot x^t to a latent subspace with a manageable dimension d through a fully connected (FC) layer f. Then, we tokenize each projected task segment by:

$$Embedding(x) = concat(x_{cls}(m), f[x^t | t = 1, ..., \Gamma]), \tag{1}$$

where the BOLD snapshot segment x is associated with a task label l_m. We further incorporate positional encoding (PE) [24] as: $\overrightarrow{PE}_t^i = \begin{cases} sin(\omega_k, t), & i = 2k \\ cos(\omega_k, t), & i = 2k+1 \end{cases}$ into the embedding process by element-wise adding ($\bigoplus$) as shown in the bottom of Fig.3, where t is the position of each time point, and $\omega_k = \frac{1}{10000^{2k/d}}$ depends on the k and the output dimension d of the FC layer f in Eq.1. It allowed us to encode the relative position of the temporal heuristics, effectively enabling the Transformer to grasp essential positional information from the BOLD time series. This encoding method eliminates the need for learning extra parameters, speeding up calculations and improving efficiency.

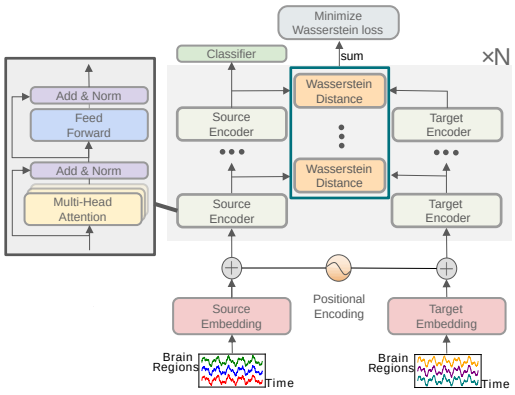

Fig. 3: Network architecture of our replicable deep model. It extends the traditional transformer framework to take the intra-subject functional fluctuations into account by (1) a fMRI-specific token embedding mechanism, (2) the dual-branch framework, and (3) wasserstein loss function for domain adaptation between test and retest fMRI scans.

[4] To simply the problem, we assume each multi-session fMRI scan has been clipped into a collection of task-specific segments based on the pre-defined task schedules. Also, we assume that each segment of BOLD snapshots has equal length of time Γ.

Dual-branch Transformer by Minimizing Wasserstein Distance between Two Domains. In the schematic presented in Fig.3, depicted within a large grey area, it introduces a dual-branch framework that alternates between self-attention [24] encoders of the source domain and target domain within several iterations. This architecture aimed at fostering a robust exchange of information and developing a nuanced inter-domain contextual understanding while reducing dependency across two domains.

To align the distributions of two domains, we seek to minimize the Wasserstein distance between the feature representations of the source and target domains. In the context of our model, the Wasserstein distance is employed to evaluate the alignment between output features—referred to as classification token of each sample—generated by each layer from both domains. Since linear ground distance is usually more robust to outliers and noise than a quadratic cost, we use l_1-norm Wasserstein distance as:

$$\mathcal{W}_1(f_1, f_2) = \sum_{i=1}^{dim} |\frac{1}{n}\sum_{j=1}^{n} f_1[j,i] - \frac{1}{n}\sum_{j=1}^{n} f_2[j,i]|, \tag{2}$$

where dim is the number of dimension of the inputs, and n is the number of samples. This process involves initially calculating the mean values of the distributions across each feature dimension. Subsequently, the differences between the mean values in the corresponding dimensions of the respective distributions are calculated. The aggregation of these differences serves as a quantitative representation of the Wasserstein distance between the two sets of features. To integrate this measure into the learning process of the model, we sum the Wasserstein distances computed across all layers as shown in the green box of Fig.3, formulating this aggregate as the Wasserstein distance loss. This strategy is instrumental in bridging the gap between the source and target domain features, thereby facilitating domain adaptation.

This approach assumes that the Wasserstein distance across multiple dimensions can be approximated by averaging the distances of the projections onto individual dimensions. While it simplifies the true multidimensional Wasserstein distance, which would require solving an optimization problem to find the optimal transport plan, the method is computationally efficient and offers an informative approximation for training purposes as a loss function measuring the distance between distributions.

Overall, the network architecture detailed in this study consists of two distinct phases, as depicted in Fig.3. Initially, a dual-branch structure is integrated into the model to concurrently process data from both source and target domains. This dual-branch configuration employs the source domain branch to steer the learning trajectory of the target branch, facilitating an exchange of information between the two. Subsequently, the model computes the Wasserstein distance to quantify the divergence in feature distributions between the two domains. By leveraging the Wasserstein loss, the model actively minimizes this domain discrepancy, thus aligning the target domain features with those of the source. The source encoders are trained in a supervised manner to ensure its

classification ability, which in turn is transferred to the target domain branch in an unsupervised fashion to imbue it with analogous classification ability. The key characteristic of this architecture is the iterative process of domain distance minimization, through which the model extrapolates the source domain's knowledge and augments its prediction performance within the target domain.

3 Experiments

Data Description. **Working Memory dataset from HCP.** Our working memory dataset [21] is a subset of the HCP-Task fMRI dataset [1]. The HCP-Task aims to capture brain activity during various cognitive tasks, each designed to elicit specific patterns of brain activity. It focuses on working memory tasks and comprises two distinct scans. A total of 1081 subjects were selected, with each fMRI scan consisting of 405 time points. These subjects were collected involving tasks under the 2-back and 0-back conditions across various stimuli (body, place, face and tool) intermixed with periods of rest, and thus there are 8 classes per subject. The sequence of tasks varied between the two scan sessions, enabling an analysis of the replicability and consistency of brain activation patterns. **Simulated Data Generation.** We use the SimTB toolbox [7] to simulate 2000 fMRI sequences with each consisting of 405 time points that mimic three distinct brain states. A depiction (Fig.4, left) showcases three states, each characterized by three interconnected modules, along the diagonal of their respective functional connectivity matrices [6]. It also involves two scans with different State orders for assessing the methods' replicability.

Experimental Settings. In our experiments, the samples in one scan are used for training, and the other dataset is randomly divided into a validation set and a test set with a ratio of 0.4/0.6, and we compare it with four counterpart methods: BolT [3], LSTM [12], SwinTransformer (SiwnT) [13], HATNet [15]. Our model has 6 layers and 4 heads in each encoder, and the hidden dimension is 1024. We used Adam optimizer and the batch size of the training is 64.

3.1 Evaluation on Simulated data

To evaluate the effectiveness of the method we've developed, we conducted tests across a spectrum of noise conditions by adding uncorrelated Gaussian noise to the simulated BOLD signals, setting the signal-to-noise ratios (SNRs) at levels ranging from 60dB down to 0.5dB. The SNR is used here to quantify the proportion of signal power to noise power. By adjusting the SNR values in our tests, we can examine how our method is robust to varying degrees of noise, providing insights into its performance in real-world scenarios with varying levels of noise interference. The recognition results on these simulated data, depicted in Fig.4 (right), demonstrate that our approach maintains relative stability across a broad range of noise conditions.

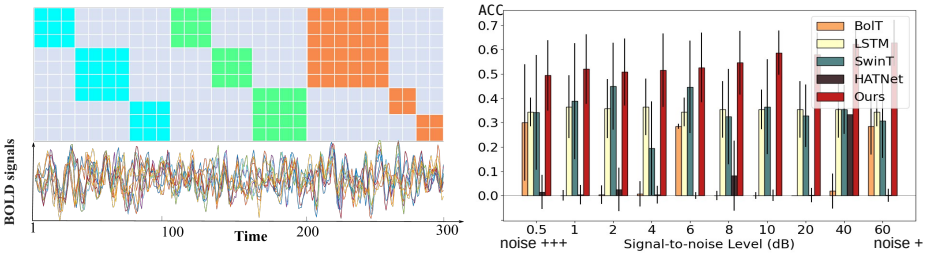

Fig. 4: **Left:** Simulated data generation. **Right:** The recognition accuracy on simulated data with different levels of noise.

3.2 Application on Task-based fMRI Data

Evaluating the Accuracy of Task Recognition. For all comparative methods, including BolT [3], LSTM [12], SwinT [13], and HATNet [15], they did not show the model replicability across different scans. In light of this limitation, we retrained and evaluated all these models on the working memory data with two scans. This enabled us to unveil their replicability across distinct scans.

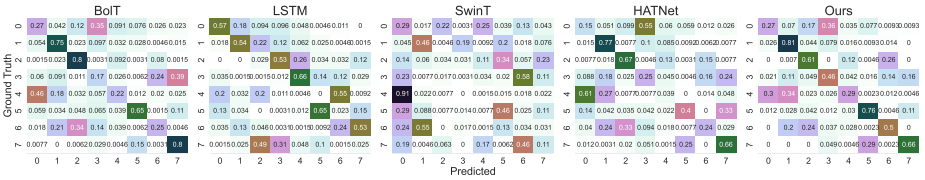

Fig. 5: The confusion matrices of all methods that are trained on scan 1 and tested on scan 2 fMRI data. 0: 2back tool, 1: 0back body, 2: 2back face, 3: 0back tool, 4: 2back body, 5: 2back place, 6: 0back face, 7: 0back place.

The overall performance is shown in Table 1, and the confusion matrices of test results for all methods are depicted in Fig.5. Bolt demonstrates superior accuracy in predicting classes 2 and 7, but its efficacy diminishes significantly for the remaining classes, indicating an imbalanced prediction accuracy across classes. LSTM and HATNet exhibit enhanced predictive capabilities for some of the classes but display sharply declined accuracy for samples from other classes. And SwinT's overall performance is subpar. In contrast, our proposed method demonstrates relatively accurate predictions across nearly all classes, especially excelling in classes 1, 2, 5, and 7, where it achieves an accuracy of approximately 80%. Notably, our approach achieves an overall balanced and accurate prediction, highlighting the efficacy of its features.

Furthermore, our investigation extended to the evaluation of methodologies for addressing the domain adaptation problems, including DANN [10], MCD

Table 1: Benchmark tests w.r.t. current comparison methods.

(%)	BolT	LSTM	SwinT	HATNet	DANN	MCD	DIRT-T	**Ours**
Pre	46.38	26.78	14.71	37.04	41.25	30.31	14.13	**54.54**
F1	46.43	26.31	15.29	37.22	41.06	32.76	12.21	**53.21**
Acc	48.79	25.89	17.45	39.78	44.03	30.89	16.89	**54.77**

[19], and DIRT-T [20]. A comparative analysis of these specialized techniques is shown in Table 1.

Our method stands out by using two distinct feature extractors for the two domains, minimizing the distance at the feature level. Specifically, we observed that samples of different classes exhibit similar shift scales, indicating a relatively neat shifting pattern. This shift arises from (i)the different phase encoding directions for scan1 and scan2 during the HCP-task data acquisition; (ii) in the continuous fMRI acquisition, changes in the order of tasks will lead to different fluctuations in the processed BOLD. These factors suggest group-level differences rather than individual variability. Thus using the same feature extractor for two domains (used in DA models) might confuse the model. Meanwhile, DANN uses domain discriminators to close two domains, but with strong feature extraction (as we use Transformer), feature distribution matching becomes weak, leading to class mismatch. MCD relies heavily on the discrepancy discrimination by two classifiers. The neat shift makes it difficult for the model to identify challenging samples in the target domain, limiting MCD's effectivenesses. Overall, despite their commendable efficacy in the realm of vision tasks, their performance on fMRI data exhibited insufficiency.

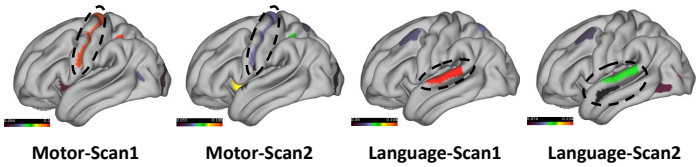

Fig. 6: Brain mapping for the tasks of Motor and Language based on test (scan 1) and retest data (scan 2).

Brain Mapping for Task Recognition. To explore how different tasks activate specific brain regions and to examine if these activations remain consistent across test/retest scans, we adopt the evaluation approach introduced by BolT [3] on the entire HCP-task dataset rather than just focusing on working memory. To accurately depict the relationship between tasks and brain regions, we choose tasks that primarily activate single-function brain areas, like Motor and Language tasks. Initially, we train a Transformer for binary classification, using classification tokens from the last layer as representatives of the samples in both scan 1 and scan 2 data for Motor and Language tasks. Then, we employ a

logistic regression model to classify these features, obtaining model coefficients, which offer insights into the importance of features for tasks belonging to the positive class, as attention maps for brain mapping.

In illustrating attention maps for each task, we focus on the top 20 brain regions with the highest weights. As depicted in Fig.6, for the Motor task, the area circled in black corresponds to the Primary Motor Cortex, a component of the Somatosensory and Motor region. Remarkably, our identification remained precise and consistent across both test and retest datasets. Regarding the Language task, we accurately identified the Early Auditory area and Auditory Association area in the test data and identified the Auditory Association area in the retest data, as highlighted within the black circle. Notably, the brain regions identified in both test and retest data exhibit considerable overlap, affirming the high consistency of the features extracted from both datasets.

4 Conclusion

In this work, we highlight a challenge of limited generalizability in current deep learning models utilized in fMRI research, we propose an effective solution employing a new dual-branch transformer with minimizing Wasserstein distance specifically designed for analyzing multi-session fMRI data from test-retest scans. Our deep model outperforms established benchmarks, demonstrating superior accuracy and replicability in identifying cognitive tasks. This underscores the versatility and potential of our data-driven methodology across diverse neuroscience studies.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., et al.: Function in the human connectome: task-fmri and individual differences in behavior. Neuroimage **80**, 169–189 (2013)
2. Bassett, D.S., Sporns, O.: Network neuroscience. Nature neuroscience **20**(3), 353–364 (2017)
3. Bedel, H.A., Sivgin, I., Dalmaz, O., Dar, S.U., Çukur, T.: Bolt: Fused window transformers for fmri time series analysis. Medical Image Analysis **88**, 102841 (2023)
4. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. Advances in Neural Information Processing Systems **19** (2007)
5. Caballero-Gaudes, C., Reynolds, R.C.: Methods for cleaning the bold fmri signal. Neuroimage **154**, 128–149 (2017)
6. Dan, T., Huang, Z., Cai, H., Lyday, R.G., Laurienti, P.J., Wu, G.: Uncovering shape signatures of resting-state functional connectivity by geometric deep learning on riemannian manifold. Human Brain Mapping **43**(13), 3970–3986 (2022)
7. Erhardt, E.B., Allen, E.A., Wei, Y., Eichele, T., Calhoun, V.D.: Simtb, a simulation toolbox for fmri data under a model of spatiotemporal separability. Neuroimage **59**(4), 4160–4167 (2012)

8. Fang, Y., Wang, M., Potter, G.G., Liu, M.: Unsupervised cross-domain functional mri adaptation for automated major depressive disorder identification. Medical image analysis **84**, 102707 (2023)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. International Conference on Machine Learning (ICML) (2015)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
11. Guo, H., Zhang, Y.: Resting state fmri and improved deep learning algorithm for earlier detection of alzheimer's disease. IEEE Access **8**, 115383–115392 (2020)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
14. Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A.: Neurophysiological investigation of the basis of the fmri signal. nature **412**(6843), 150–157 (2001)
15. Mehta, S., Lu, X., Weaver, D., Elmore, J.G., Hajishirzi, H., Shapiro, L.: Hatnet: an end-to-end holistic attention network for diagnosis of breast biopsy images. arXiv preprint arXiv:2007.13007 (2020)
16. Noble, S., Scheinost, D., Constable, R.T.: A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. Neuroimage **203**, 116157 (2019)
17. Pan, S.J., Yang, Q.: A Survey of Transfer Learning. IEEE Press (2010)
18. Price, C.J.: The evolution of cognitive models: From neuropsychology to neuroimaging and back. Cortex **107**, 37–49 (2018)
19. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3723–3732 (2018)
20. Shu, R., Bui, H., Narui, H., Ermon, S.: A dirt-t approach to unsupervised domain adaptation. In: International Conference on Learning Representations (2018)
21. Smith, J., Johnson, E., Williams, S.: Neural correlates of working memory: An fmri study. Cognitive Neuroscience Journal **25**(3), 123–135 (2020). https://doi.org/10.1234/cogneuro.2020.12345
22. Tanabe, J., Miller, D., Tregellas, J., Freedman, R., Meyer, F.G.: Comparison of detrending methods for optimal fmri preprocessing. NeuroImage **15**(4), 902–907 (2002)
23. Thomas, A., Ré, C., Poldrack, R.: Self-supervised learning of brain dynamics from broad neuroimaging data. Advances in Neural Information Processing Systems **35**, 21255–21269 (2022)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
25. Wen, D., Wei, Z., Zhou, Y., Li, G., Zhang, X., Han, W.: Deep learning methods to process fmri data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion. Frontiers in neuroinformatics **12**, 23 (2018)
26. Yu, X., Zhang, L., Zhao, L., Lyu, Y., Liu, T., Zhu, D.: Disentangling spatial-temporal functional brain networks via twin-transformers. arXiv preprint arXiv:2204.09225 (2022)

27. Zuo, X.N., Xing, X.X.: Test-retest reliabilities of resting-state fmri measurements in human brain functional connectomics: a systems neuroscience perspective. Neuroscience & Biobehavioral Reviews **45**, 100–118 (2014)