# Estimating Brain Age with Global and Local Dependencies

Yanwu Yang*†, Xutao Guo*†, Zhikai Chang*, Chenfei Ye* Yang Xiang†, Haiyan Lv¶ and Ting Ma*†‡§
*Harbin Institute of Technology at Shenzhen, China
†Peng Cheng Laboratory, Shenzhen, China
‡Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China
§Xuanwu Hospital Capital Medical University, Beijing, China
¶MindsGo Life Science Co.Ltd, Shenzhen, China

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897454

***Abstract*—The brain age has been proven to be a phenotype of relevance to cognitive performance and brain disease. Achieving accurate brain age prediction is an essential prerequisite for optimizing the predicted brain-age difference as a biomarker. As a comprehensive biological characteristic, the brain age is hard to be exploited accurately with models using feature engineering and local processing such as local convolution and recurrent operations that process one local neighborhood at a time. Instead, Vision Transformers learn global attentive interaction of patch tokens, introducing less inductive bias and modeling long-range dependencies. In terms of this, we proposed a novel network for learning brain age interpreting with global and local dependencies, where the corresponding representations are captured by Successive Permuted Transformer (SPT) and convolution blocks. The SPT brings computation efficiency and locates the 3D spatial information indirectly via continuously encoding 2D slices from different views. Finally, we collect a large cohort of 22645 subjects with ages ranging from 14 to 97 and our network performed the best among a series of deep learning methods, yielding a mean absolute error (MAE) of 2.855 in validation set, and 2.911 in an independent test set.**

***Index Terms*—Long-range dependencies, Brain age estimation, Transformer, CNN**

## I. Introduction

Recently, researches have demonstrated that MRIs could be used to predict chronological age and show that the brain age, derived purely from neuroimaging data is vital to help improve detection of early-age neurodegeneration and predict age-related cognitive decline [3]. Meanwhile, predicted age difference (PAD), the difference between predicted brain age and chronological age, correlates with the measures of mental and physical changes [11]. For example, positive PAD introduces that the brain is older than the actual age and the subject is experiencing accelerated aging. Furthermore, it is also shown to be associated with cognitive impairments [16, 4], brain injuries, and other brain diseases [14, 15]. Therefore, it is an essential prerequisite to achieve accurate brain age estimating for quantifying the PAD as a biomarker.

The brain age estimation is a fine-grained recognition task, and the actual brain characteristics of the T1w image and structural changes could be hardly sensed explicitly. Recently, deep learning methods like 3D Convolution Neural Network (CNN) have been used to predict brain age and achieve promising results [3, 21, 18] without a prior bias or hypothesis. However, CNN methods are limited by only processing local neighborhood features and propagating signals progressively [25], where the hidden features might be lost and more inductive bias would be introduced. Recently vision Transformer models retain global image information and could relate long-range relationships between patches using self-attentions, achieving state-of-the-art performance on image classification, object detection, and semantic segmentation. The success of these models demonstrates the potential for Transformer to be used in the vision domain with the essential characteristic of encoding long-range dependencies and retaining global information. Nevertheless, studies have pointed out that long-range dependencies would also fail to work well, where a position is often less correlated far away from it, compared with those that are nearer [8]. Therefore, there is still a gap between encoding representations with short-range and long-range dependencies, which restricts the models' flexibility in diverse spatial scales and relationships in images [8].

To address this problem, we propose a novel network for brain age estimation, called the Global and Local Dependency Network (GLDN). The GLDN sufficiently utilizes the CNNs for encoding densely-distributed local features and strengthening locality with local dependencies, and Transformer for encoding sparsely-distributed semantic concepts and establishing global dependencies. Especially, the fusion block is the basic module in our model and allows locating and aggregating the local and global concepts from CNNs and Transformer of each stage. Besides, we propose a new vision Transformer block, called Successive Permuted Transformer (SPT) for locating long-range dependencies of 3D medical images. The SPT leverages the spatial information of 3D images to be encoded by 2D operations by different views, which indirectly locates the spatial relationships and brings computation efficiency. Finally, we compared our model with

This research has been conducted using the UK Biobank Resource under Application Number 56113. The study is supported by grants from the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (NO:ZYYCXTD-C-202004), Shenzhen Longgang District Science and Technology Development Fund Project (LGKCXGZX2020002), Basic Research Foundation of Shenzhen Science and Technology Stable Support Program (GXWD20201230155427003-20200822115709001), the National Key Research and Development Program of China (2021YFC2501202), and the National Natural Science Foundation of China (62106113).

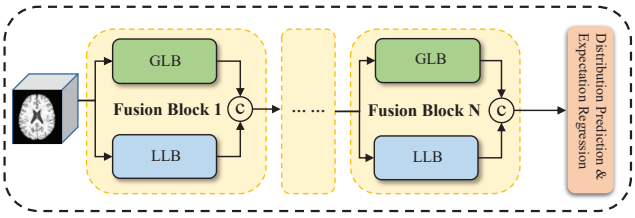


Fig. 1. The framework of the GLDN network, including several fusion blocks, and a distribution prediction and expectation regression module. Each fusion block(in yellow) is composed of a global learning block (GLB, in green), a local learning block (LLB, in blue), and an aggregation operation (shown as a symbol C).

a series of models including CNNs and vision Transformer models on a large cohort of five datasets, where the results are improved compraed with other well-estimated models.

## II. Method

### A. Framework

We illustrate the sketch map of the GLDN network in Fig. 1, which is composed of several fusion blocks, and a classifier with a label distribution prediction and expectation regression module. Each fusion block is built with a global learning block, a local learning block, and an aggregation operation.

*1) **Local Learning Block**:* In detail, two CNN blocks are embedded in the local learning block. Each CNN block consists of a 3D convolution layer with a kernel size of 3, padding size of 1, and batch normalization, a ReLU activation, and a max-pooling layer with a pooling size of 2. With two CNN blocks, the feature size of input would be reduced into $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The local Learning Block could be denoted as:

$$x^l_{local} = LLB(x^l) = CB(CB(x^l)) \tag{1}$$

$$CB(x^l) = MaxPool(BN(ReLU(Conv(x^l)))) \tag{2}$$

where $x^l$ is the input of the $l-th$ fusion block, $x^l_{local}$ is the output of a local learning block, $LLB$ denotes the local learning block, and $CB$ denotes a CNN block.

*2) **Global Learning Block**:* The global learning block is built with successive permute Transformer blocks to capture global information with full-range dependencies. Apart from the original ViT structure, the SPT, shown in Fig. 2, is designed to better suit 3D medical images with three sequential Transformer parts, where each part locates relationships of slices separated from different views (Sagittal, Axial, Coronal). In detail, each part consists of a permute operation, a patch splitting layer, Transformer layers, and a patch merging layer.

To reduce numerous parameters in modeling 3D medical images using Transfromer, the 3D input is permuted along each axis and cropped into 2D slices. For example, a permute operation would transform an input in the size of $96 \times 114 \times 96$ into 96 slices with each size of $114 \times 96$ along the first axis. The slice would be continued to be segmented into non-overlapping patches by patch splitting.

The Transformer encoders consist of multi-head self-attention blocks (MSA), layer norms, and fully connected feed-forward blocks. The input tokens $x_t \in R^{N \times d}$ are linear projected into $qkv$ spaces, where queries (Q), keys (K), values (V), and the output, a weighted sum of the values are computed as

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Finally, the multi-head attention layer is defined by combining multiple attentions. The outputs of several self-attention blocks are concatenated and sent into the patch merging layers. Through computing dot-product, the similarity between different tokens is calculated, resulting in long-range and global attention.

The patch merging layers are implemented for feature map compression via concatenating the features of each group of $2 \times 2$ neighboring patches and linear projection, at the same time producing hierarchical representations, by reference to the design in [17]. The patch merging layer is implemented followed by the Transformer layer. Within a SPT, the input would be reduced into $(\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4})$ with each patch merging layer downsampling the feature size of one view into $\frac{1}{2} \times \frac{1}{2}$. The design of successive Transformer indirectly realizes the spatial position relationship learning through permute operations. Finally, the proposed GLDN network can be noted as:

$$x^l_{local} = LLB(x^l) \tag{4}$$

$$x^l_{global} = GLB(x^l) \tag{5}$$

$$x^l_{fusion} = Aggregate(\{x^l_{local}, x^l_{global}\}) \tag{6}$$

where, $l$ denotes the $l$-th layer, and $GLB$ denotes the global learning block.

### B. LDL & expectation regression and loss function

In this paper, we leverage the label distribution learning combined with the expectation regression as the loss function. This strategy forces the deep learning regression model to take care of the ambiguity among labels [5]. In this paper, all the ages ($y \in R$) of healthy subjects range from 14 to 97. And we define the label set as $L = (l_k | k = 14, 15, ..., 97)$, and $\Delta l = 1$ as the discrete step size. The probability density function of normal distribution is chosen to generate the ground-truth $(q_k | k = 14, 15, ..., 97)$ with a hyper-parameter $\theta$:

$$y = \sum_k l_k q_k, q_k \in (0, 1) \tag{7}$$

$$q_k = \frac{p_k}{\sum_k p_k} \tag{8}$$

$$p_k = \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{(l_k - y)^2}{2\theta^2}} \tag{9}$$

The goal of the label distribution prediction is to maximize the similarity between $q_k$ and the predicted distribution $\hat{q_k}$. The Kullback-Leibler divergence is employed as the measurement of the dissimilarity between ground-truth label distribution and prediction distribution:

$$L_{kl} = \sum_k q_k log \frac{q_k}{\hat{q_k}} \tag{10}$$

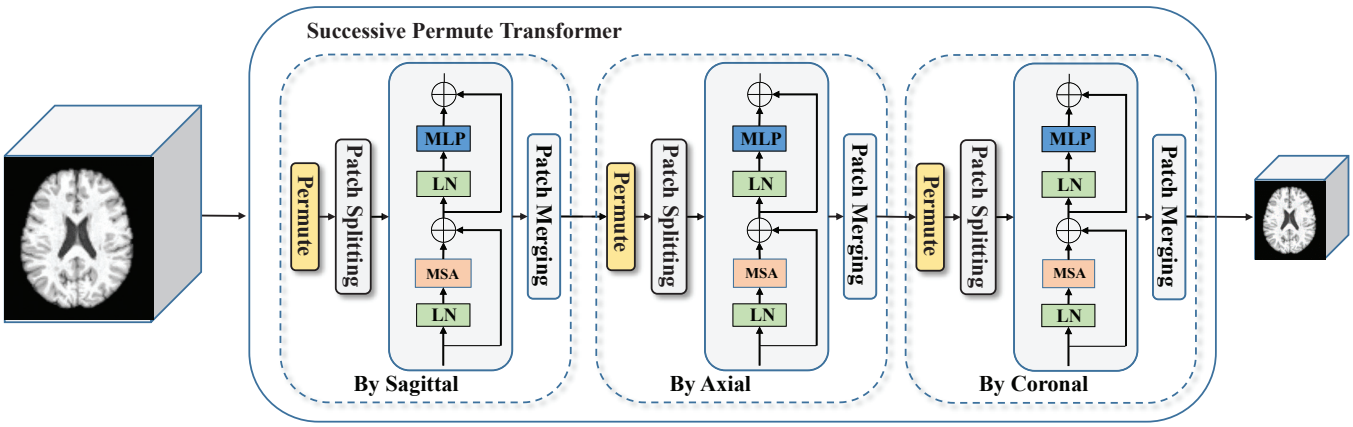


Fig. 2. The design of a SPT block, where the 3D information is learned by three successive Transformer. Permutation operations leverages the framework to relate spatial context from one view at one time.

And the expectation regression takes the predicted distribution and the label set $L$ as inputs. The expectation regression module minimizes the error between the expected value $\hat{y}$ and ground-truth $y$. The $L_1$ loss is used as the error measurement:

$$L_{mae} = |\hat{y} - y| = |\sum_k \hat{q_k} l_k - y| \tag{11}$$

Totally, the weighted combination of $L_{kl}$ and $L_{mae}$ with the weight $\lambda$ is employed as our loss function:

$$L = L_{kl} + \lambda L_{mae} = \sum_k q_k log \frac{q_k}{\hat{q_k}} + \lambda |\sum_k \hat{q_k} l_k - y| \tag{12}$$

## III. Experiments

### A. Datasets

The methods were evaluated on T1-weighted MR images from a large cohort consisting of IXI database (http://brain-development.org), the Alzheimer's Disease Neuroimaging Initiative (ADNI) [9], UK Biobank[24, 20], the Open Access Series of Imaging Studies (OASIS)[19], and 1000 Functional Connectomes Project (1000-FCP, http://www.nitrc.org/projects/fcon_1000). Only healthy subjects were selected in our experiments. A total of 22645 T1-weighted MRI images of subjects aging between 14 and 97 years old are selected to form our cohort. All data were acquired at either 1.5T or 3T T1-weighed MRI.

### B. Data Preprocessing

All data were all processed with the same pipeline, including image FOV truncation [10], AC-PC align, brain skull stripping, bias field correction [23], and linear-registration into the standard MNI space. Additionally, z-score normalization is employed to narrow the gap between different data centers. It is proved to improve the synthesis results and is vital for successful deep learning-based MR image synthesis [22]. After preprocessing, all images are down-sampled by linear-registering into the standard $2mm^3$ MNI space and padded into the size of $96 \times 112 \times 96$ for successive non-overlapping downsampling.

### C. Experimental Setting

For comparison, we divided all the data samples into three subsets including training set (80%), validation set (10%), and a test set (10%). The test set is fixed and a cross-validation with 4 fold was performed on the rest samples. The performance is evaluated by the mean absolute error (MAE), root mean squared error (RMSE), Pearson correlation coefficient (PCC), and Spearman's rank correlation coefficient (SRCC).

Our proposed GLDN is embedded with two fusion blocks to ensure the integrity of slicing and downsampling. The first fusion block is built with an SPT block with a patch size of $[8 \times 8]$, and two CNN blocks with the channel number of [16, 32]. The second fusion block receives the concatenation of the first fusion block with a channel of 40. Within the second fusion block, the patch size of the SPT is set as $[2 \times 2]$, and the convolution channel is set as [64, 128].

For better comparison, machine learning methods with feature engineering and deep learning methods with CNNs, and Transformer networks are both carried out. An ensemble model of XGBoost [2] and LightGBM [12] methods was carried out, using relative volume fraction of the brain regions segmented by FastSurfer [7]. A series of end-to-end CNN based methods including 3D-ResNet, SFCN [21], and TSAN [18] are trained and implemented. For ResNet, the original 2D operations were replaced by 3D and a dropout with drop rate = 0.5 was applied before the final fully connected layer. Here we implement the first stage of TSAN for fair comparison. Besides, Transformer methods like ViT, DETR [1], Nested Transformer are also compared. The depth of layers and number of the attention heads are tuned according to different architectures with a grid search of [4, 6, 8, 12]. These models receive 2D images as input for default and are modified to suit 3D medical images by replacing 2D operations into 3D.

In addition, we implement the DeTR design with different numbers of layers, where 1/2/3 CNN blocks are compared in DeTR-1/2/3 respectively. Especially, the nested Transformer is modified by reference to the [6] and is carried out using an

TABLE I
PERFORMANCE OF BRAIN ESTIMATIONS USING DIFFERENT MODELS.

| | | Validation Set | | | | Independent Test Set | | | |
|---|---|---|---|---|---|---|---|---|---|
| Type | Models | MAE | RMSE | PCC | SRCC | MAE | RMSE | PCC | SRCC |
| Machine learning | XGBoost+LightGBM | 4.290 | 7.130 | 0.815 | 0.818 | 4.298 | 7.150 | 0.808 | 0.814 |
| CNNs | Resnet18 | 3.265 | 4.386 | 0.871 | 0.867 | 3.383 | 4.612 | 0.858 | 0.854 |
| | Resnet34 | 3.204 | 4.314 | 0.873 | 0.867 | 3.334 | 4.640 | 0.861 | 0.851 |
| | Resnet50 | 3.226 | 4.553 | 0.871 | 0.867 | 3.349 | 4.817 | 0.854 | 0.850 |
| | SFCN | 2.993 | 4.097 | 0.875 | 0.869 | 3.093 | 4.228 | 0.872 | 0.856 |
| | TSAN (first-stage) | 2.948 | 4.150 | 0.874 | 0.868 | 3.076 | 4.350 | 0.861 | 0.866 |
| Transformer Models | ViT | 3.419 | 4.613 | 0.863 | 0.865 | 3.536 | 4.859 | 0.848 | 0.851 |
| | DeTR-1 | 3.335 | 4.509 | 0.866 | 0.865 | 3.344 | 4.517 | 0.863 | 0.854 |
| | DeTR-2 | 2.920 | 4.101 | 0.876 | 0.866 | 2.997 | 4.260 | 0.863 | 0.854 |
| | DeTR-3 | 2.933 | 4.122 | 0.876 | 0.868 | 3.047 | 4.286 | 0.867 | 0.855 |
| | Nested Transformer | 3.112 | 4.392 | 0.873 | 0.866 | 3.234 | 4.630 | 0.844 | 0.850 |
| | Ours (w/o CNN) | 3.041 | 4.153 | 0.873 | 0.868 | 3.167 | 4.335 | 0.866 | 0.854 |
| | Ours | **2.855** | **3.960** | **0.881** | **0.871** | **2.911** | **4.010** | **0.879** | **0.869** |

inner Transformer (depth: 6, heads: 8) encoding 2D slices, and an outer Transformer encoding spatial information (depth: 8, heads: 12).

All the models were trained from scratch with a initialized learning rate of 1e-4, and a batch size of 128. The learning rate is increased to 1e-4 in 200 warmup epochs. The $\lambda$ was initialized with 0 and set to 1 when the validation loss had not been decreased for 50 epochs. Models were trained using a stable adaptive optimizer, Adam [13], with a L2 weight decay coefficient = 0.00005, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The best model was obtained based on the validation loss and an early stopping criterion was imposed when the validation loss did not improve for 80 epochs. To reduce the risk of overfitting, two data argumentation methods were applied during training, consisting of random rotation and random image shifting. The rotation angles were between $-10^{\circ}$ and $10^{\circ}$ and the input was random shifted by between -5 and 5 voxels along every axis with equal probability. All the experiments are carried out using Pytorch on 8 NVIDIA-Tesla V100 GPU devices.

## IV. RESULTS

The detailed evaluation results are shown in Table. I. The machine learning method using feature engineering with an ensemble of XGBoost and LightGBM did not perform as well as deep learning methods. With the layers getting deep in ResNet, the performance increases first and remains or even decreases a bit, where the ResNet-34 achieves the best among all the ResNet families, with the MAE of 3.204, RMSE of 4.314, PCC of 0.873, and SRCC of 0.869 in the validation set, and the MAE of 3.334, RMSE of 4.640, PCC of 0.861, and SRCC of 0.851 in the test set. The design of light fully convolution model with large number of channels and dense model with asymmetric convolution achieve promising and comparable results in CNN models, better than most Transformer based models. Transformer based model also achieves promising results, where the DeTR with 2 CNN blocks performs the best with the MAE of 2.920, RMSE of 4.101, PCC of 0.876, and SRCC of 0.868 in the validation set, and the MAE of 2.997, RMSE of 4.260, PCC of 0.863, and SRCC of 0.854 in the test set. With the number of CNN blocks increasing, the performance reached a plateau with excessive abstract low-level features.

We show ablation experiments on our proposed GLDN network without CNN stressing on locality. Although we excluded the CNN for discarding the localized dependencies, our design of SPT is better than methods with pure vision Transformers (ViT, Nested Transformer) for 3D medical image learning with continuously encoding relationship along different axes and achieves the MAE of 3.041, RMSE of 4.153, PCC of 0.881, and SRCC of 0.871 in the validation set, and the MAE of 3.167, RMSE of 4.335, PCC of 0.866, and SRCC of 0.854 in the test set. Compared with CNN models and Transformer models, the methods (DeTR, GLDN) of using the CNN and Transformer together for encoding features achieve the best. Finally, our proposed GLDN generally obtains the best results with the lowest MAE and the highest PCC (MAE: 2.855, RMSE: 3.960, PCC: 0.881, SRCC: 0.871 in the validation set, and MAE: 2.911, RMSE: 4.010, PCC: 0.879, SRCC: 0.869 in the test set).

## V. CONCLUSION

In this paper, the GLDN was proposed to predict individual brain age based on brain MRI images with Transformer encoding global representations and establishing global dependencies and CNN stressing on locality with local dependencies. The architecture improves feature diversity and aggregates the multi-scale information. In our experiments, the combination of convolutions and Transformer would achieve promising results among all the models, where our proposed model achieves the optimal, yielding an MAE of 2.911, RMSE of 4.010, PCC of 0.879, and SRCC of 0.869 on the independent test set. Overall, we suspect that the coordination between Transformer and convolution has a great potential for analyzing neuroimaging-based individualized prediction of the clinical or behavioral phenotype.

## REFERENCES


[1] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European Conference on Computer Vision*. Springer. 2020, pp. 213–229.

[2] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015).

[3] James H Cole et al. "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker". In: *NeuroImage* 163 (2017), pp. 115–124.

[4] Katja Franke et al. "Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI". In: *Neuroimage* 63.3 (2012), pp. 1305–1312.

[5] Bin-Bin Gao et al. "Deep label distribution learning with label ambiguity". In: *IEEE Transactions on Image Processing* 26.6 (2017), pp. 2825–2838.

[6] Kai Han et al. "Transformer in transformer". In: *arXiv preprint arXiv:2103.00112* (2021).

[7] Leonie Henschel et al. "Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline". In: *NeuroImage* 219 (2020), p. 117012.

[8] Shaofei Huang et al. "ORDNet: Capturing omni-range dependencies for scene parsing". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8251–8263.

[9] Clifford R Jack Jr et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27.4 (2008), pp. 685–691.

[10] Mark Jenkinson et al. "Fsl". In: *Neuroimage* 62.2 (2012), pp. 782–790.

[11] Benedikt Atli Jónsson et al. "Brain age prediction using deep learning uncovers associated sequence variants". In: *Nature communications* 10.1 (2019), pp. 1–10.

[12] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[13] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[14] Nikolaos Koutsouleris et al. "Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders". In: *Schizophrenia bulletin* 40.5 (2014), pp. 1140–1153.

[15] Anil Kuchinad et al. "Accelerated brain gray matter loss in fibromyalgia patients: premature aging of the brain?" In: *Journal of Neuroscience* 27.15 (2007), pp. 4004–4007.

[16] Franziskus Liem et al. "Predicting brain-age from multimodal imaging data captures cognitive impairment". In: *Neuroimage* 148 (2017), pp. 179–188.

[17] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *arXiv preprint arXiv:2103.14030* (2021).

[18] Ziyang Liu et al. "Brain Age Estimation from MRI Using a Two-Stage Cascade Network with Ranking Loss". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 198–207.

[19] Daniel S Marcus et al. "Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults". In: *Journal of cognitive neuroscience* 22.12 (2010), pp. 2677–2684.

[20] Karla L Miller et al. "Multimodal population brain imaging in the UK Biobank prospective epidemiological study". In: *Nature neuroscience* 19.11 (2016), pp. 1523–1536.

[21] Han Peng et al. "Accurate brain age prediction with lightweight deep neural networks". In: *Medical Image Analysis* 68 (2019), p. 101871.

[22] Jacob C Reinhold et al. "Evaluating the impact of intensity normalization on MR image synthesis". In: *Medical Imaging 2019: Image Processing*. Vol. 10949. International Society for Optics and Photonics. 2019, 109493H.

[23] John G Sled, Alex P Zijdenbos, and Alan C Evans. "A nonparametric method for automatic correction of intensity nonuniformity in MRI data". In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 87–97.

[24] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *Plos med* 12.3 (2015), e1001779.

[25] Xiaolong Wang et al. "Non-local neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.