An Attention-based Mid-level Feature Representation Incorporated 3D Residual Neural Network for Brain Age Estimation

Yu Zhang, Jiaru Yang, Masaaki Omura, Zhenyu Lei,
Hideyuki Hasegawa, and Shangce Gao, *Senior Member, IEEE*

***Abstract*—Brain age is considered an important biomarker of pathology that can be used to detect many common neurodegenerative disorders. One of the most promising ways to effectively estimate brain age is by employing a deep neural network model for 3D volume magnetic resonance imaging (MRI). However, most existing voxel-based prediction models are not able to achieve excellent accuracy in large-scale MRI datasets that span all ages. In this paper, we propose a model based on a residual network architecture for estimating brain age from 3D volumes, called MAR3D. MAR3D uses 3D residual blocks for local representation learning and introduces an attention mechanism-based (AM) Block to refine mid-level features through max-pooling sampling, regional multi-head attention, and spatial restoration. This design enhances cross-regional feature representation while preserving the 3D structure of brain MRI. To evaluate the performance of our proposed MAR3D in predicting brain age, we collected six public datasets comprising 7,680 brain MRIs ranging in age from 6-96 years. Compared to existing brain age estimation methods, MAR3D achieved the best performance with a mean absolute error of 2.298 years. Under the same experimental conditions, MAR3D outperformed state-of-the-art and classical models across six different metrics. Experimental results demonstrate that the proposed MAR3D model yields significant improvements, achieving a 6.8-38.9% improvement in mean absolute error compared to state-of-the-art methods and outperforming the original 34-layer backbone model by 33.8% in terms of overall effectiveness. In addition, we visualize model-emphasized regions from sagittal, coronal, and axial planes to provide a qualitative interpretation of the learned attention patterns. Our code is available at: https://github.com/zy-251026/MAR3D.**

***Index Terms*—brain age estimation, deep learning, magnetic resonance imaging, attention**

I. INTRODUCTION

AGING is an unavoidable natural process, and the human brain accumulates damage at different rates with age due to genes and diseases [1]. Therefore, a definition of brain age has been proposed as an indicator to study the aging process in the human brain. Brain age estimation is not only scientifically important but also of extensive clinical value. Numerous studies have shown that many types of neurological and metabolic diseases are associated with abnormal brain aging [2], such as Alzheimer's disease, type II diabetes mellitus, epilepsy, and so on. The difference between predicted brain age and chronological age is known as the estimated age difference (EAD), which is used to reflect the degree of deviation from the regular aging trajectory. A positive EAD means that the brain is aging at an accelerated rate, while a negative EAD indicates a very healthy brain. By assessing the value of EAD, sub-healthy people at higher risk of disease can be identified and appropriate treatment programs can be developed in advance to avoid specific diseases [3].

Conventional brain age estimation increasingly uses deep learning (DL) to analyze human neuroimages and make a determination of its biological age. MRI is the most commonly used neuroimaging technology for head scans to image the human brain. In comparison with computed tomography (CT), MRI provides a superior contrast in displaying soft tissues and is well-suited for detailed demonstration of neurological structures. In addition to ensuring high-quality imaging and safety, multimodal MRI can capture diverse information to meet various requirements. Variants of MRI include functional MRI (fMRI), structural MRI (sMRI), diffusion tensor imaging (DTI), magnetic resonance angiography, and diffusion-weighted imaging. Currently, MRI technology has been widely applied to research on aging and brain-related diseases [4].

The prediction model of brain age can be classified into slice-based and voxel-based approaches in terms of input data [5]. In the former case, single or multiple 2-dimensional slices extracted from raw MRI scans are fed into the DL model as input images. It is apparent that the slice-based approach can obtain a classification result with low spatio-temporal consumption. Due to the lesser number of parameters for the slice-based approach, the DL model can avoid the partial interference of noise and redundant data. Relatively, the voxel-based approach utilizes the 3-dimensional MRI scans to fully extract the pixel information. In comparison with slices, 3D MRI images contain a far greater number of parameters, which leads to the fact that the voxel-based approach can extract more global features. Theoretically, the DL model with the voxel-based approach is capable of perfectly analyzing brain MRI images and solving brain age estimation problems [6].

A highly accurate mechanism for predicting brain age can serve as a powerful medical tool, promoting a deeper understanding of neurodegenerative diseases. Furthermore, it enhances early diagnosis of the disease, risk assessment, and monitoring of treatment efficacy, ultimately promising

This research was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grants JP25K21298, JP25K24386, JP26K03333, JP26K23843, and JP25K03179, and Japan Science and Technology Agency (JST) Support for Pioneering Research Initiated by the Next Generation (SPRING) under Grant JPMJSP2145. (Corresponding Author: Shangce Gao)

Yu Zhang is with the Department of Laboratory Medicine of The First Affiliated Hospital & Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, 310000, China. (E-mail: 0625716@zju.edu.cn)

Jiaru Yang is with the School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China. (E-mail: yangjr@ntu.edu.cn)

Masaaki Omura, Zhenyu Lei, Hideyuki Hasegawa, and Shangce Gao are with the Faculty of Engineering, University of Toyama, Toyama-shi, 930-8555, Japan. (E-mail: momura@eng.u-toyama.ac.jp; leizg@eng.u-toyama.ac.jp; hasegawa@eng.u-toyama.ac.jp; gaosc@eng.u-toyama.ac.jp)

to improve the quality of life for patients [7]. Nevertheless, the current prediction models are inadequate for accurately estimating brain age due to a lack of comprehensive understanding of brain MRI images. The accuracy of brain age prediction mainly depends on image information within regions such as the medial frontal gyrus, caudate nucleus, and paracentral lobule in brain MRI [8]. These regions are biologically relevant to aging as they exhibit pronounced age-related structural changes. The medial frontal gyrus, part of the prefrontal cortex, undergoes substantial volume reduction associated with executive function decline [9]. The caudate nucleus shows significant age-related atrophy affecting motor control and cognitive processing [10]. The paracentral lobule demonstrates distinctive morphological changes reflecting sensorimotor function deterioration with age [11]. Imaging features including gray matter volume, cortical thickness, and morphological variability within these regions contribute most significantly to model predictions. Therefore, the capacity to highlight the importance of these brain regions constitutes a crucial criterion for evaluating whether a model demonstrates both strong predictive performance and biologically inspired relevance. Conventional prediction models extract information from these localized regions of the brain by processing different levels of features in the image through convolution-like operations and ultimately summarize them into global features to estimate brain age [12]. However, global features built on low-level features usually remain image-level information without advanced structured image descriptions [13].

Within the visual processing framework, the primary visual cortex and extrastriate cortex correspond to the extraction regions for low-level and mid-level features, respectively. In recent years, extensive research has demonstrated that mid-level feature representations play a crucial role in image recognition tasks [14]. Unlike low-level features that capture only local patterns in the form of edges and textures through pixel-level responses [15], mid-level features integrate basic information into structurally meaningful representations, including contour fragments, curvature, symmetry, and junction points [16]. Neurophysiological studies have shown that structured mid-level representations enrich the original low-level signals, enabling higher-level shape abstraction that significantly enhances performance in image recognition and classification tasks. Consequently, incorporating mid-level descriptors is essential for 3D brain age prediction tasks using MRI data, as these features capture the most discriminative anatomical components within brain regions. Therefore, we designed a novel feature extraction module that can better process 3D MRI images based on the characteristics of mid-level feature and regional information in the human brain.

The main contributions of this work are described as follows:

1) In this study, we integrated the attention mechanism-based (AM) block based on the sampling and regional attention encoder into the original 3D ResNet to address the issue of brain age estimation, resulting in the development of a novel model, designated as MAR3D. In the proposed architecture, one AM Block is incorporated after each residual stage to capture and integrate multi-level feature representations across different spatial scales for comprehensive 3D MRI analysis. The MAR3D model demonstrated superior prediction accuracy when tested on six brain MRI datasets comprising individuals of varying age ranges and conditions, outperforming existing state-of-the-art methods.

2) The mid-level feature representation was introduced for the first time into the field of 3D brain age estimation, with an efficient feature extraction method designed based on the characteristics of human brain MRI images. Specifically, we employ dynamic max pooling to selectively compress spatial dimensions while preserving discriminative age-related features, followed by a multi-head attention encoder that transforms these features into independent mid-level representations. In order to transform the downsampled features into independent mid-level features, the multi-head attention mechanism is employed to assign weights to the information within different regions of the brain, thereby enhancing the image representation capability of the entire model.

3) These contour fragments, extracted through mid-level feature analysis, specifically delineate critical anatomical boundaries and morphological characteristics that demonstrate significant age-related variations across different demographic groups. By capturing these discriminative structural patterns, mid-level features provide substantially more expressive and interpretable representations that directly correlate with established neuroanatomical markers of brain aging. Furthermore, we provide attention score-based graphical visualizations to illustrate the spatial localization of the most informative brain regions contributing to age prediction, providing neurobiological insights into aging mechanisms and enhancing model interpretability for potential clinical applications.

II. Related Work

Convolutional neural networks (CNNs) imitate the function of the human visual system and rely on their distinctive network architecture to efficiently analyze image data. The hierarchical feature learning and local perception mechanisms of CNNs are capable of extracting image features progressively from simple to complex. For processing high-resolution MRI images, CNNs demonstrate superior performance compared to other machine learning methods. Hence, CNNs have been widely applied to brain MRI images and have achieved high accuracy in complex tasks, such as brain age estimation, gender classification, and disease diagnosis. As shown in Table I, Ueda *et al.* utilize an 11-layer 3D CNN to predict ages based on 1101 T1-weighted images of healthy Japanese subjects, obtaining a mean absolute error (MAE) of 3.67 years [17]. B. A. Jonsson *et al.* train a deep learning model that incorporates residual blocks into 3D CNN, providing a MSE of 3.63 years in the large dataset [18]. Cole proposes a multi-modal age estimation mechanism, which employs a least absolute shrinkage and selection operator regression to predict the chronological age with MAE = 3.55 years [19]. Bashyam *et al.* train a DeepBrainNet based on the Inception-ResNet-v2, yielding an MAE of 3.70 years [20]. Cheng *et al.* design a novel two-stage-age-network (TSAN) architecture to predict the brain age on the basis of sex labels and MRI

TABLE I
A SUMMARY OF RECENTLY PROPOSED DEEP LEARNING METHODS FOR BRAIN AGE ESTIMATION ON MRI DATASETS.

Research	Publish year	Sample #	Age range	Modality	Type	Model	MAE(years)
Ueda et al.	2019	1101	20-80	T1-weighted	3D-voxel	3DCNN	3.67
Jonsson et al.	2019	12499	20-86	T1-weighted	3D-voxel	CNN+3DResNet	3.63
Cole	2020	17426	45-80	Multimodality	2D-slice	LASSO	3.51
Bashyam et al.	2020	11729	3-95	T1-weighted	2D-slice	Inception-ResNet	3.7
Cheng et al.	2021	6586	17-98	T1-weighted	3D-voxel	TSAN	2.42
Dinsdale et al.	2021	19687	44-80	T1-weighted	3D-voxel	3DCNN+VGG-16	2.97
He et al.	2021	16705	0-97	Multimodality	3D-voxel	FiA-Net	3
He et al.	2022	8379	0-97	T1-weighted	2D-slice	GL-Transformer	2.7
He et al.	2022	6049	0-97	T1-weighted	3D-voxel	SFCN+Transformer	2.38
Rajashekar et al.	2022	2074	21-81	Multimodality	3D-voxel	3DCNN	3.85
Cai et al.	2023	16458	46-81	Multimodality	3D-voxel	Graph Transformer	2.71
Our proposed	-	7680	6-96	T1-weighted	3D-voxel	MAR3D	**2.298**

images, obtaining an MAE of 2.42 years [21]. Dinsdale *et al.* develops a 3D VGG-16 model and achieves a prediction accuracy with MAE = 2.97 on 19687 subjects [7]. Peng *et al.* propose a lightweight deep learning model, Simple Fully Convolutional Network (SFCN), for brain age prediction using T1-weighted MRI. SFCN achieves a state-of-the-art MAE of 2.14 years on the UK Biobank dataset [22]. Mouches *et al.* train two common 3D CNN to simultaneously analyse T1-weighted MRI and time-of-flight magnetic resonance angiography. An MAE of 3.85 years is obtained after the final linear regression model [23]. Recently, alternative approaches have explored novel techniques to capture structured brain features more effectively. For instance, 3D-TDR incorporates a tensor regression layer into a 3D CNN framework, enabling the extraction of more discriminative structural changes from high-dimensional whole-brain MRI data [24]. EquiTrack model utilizes a hybrid architecture that combines a denoising CNN with an E-CNN to effectively decouple the processing of anatomically irrelevant intensity features from equivariant spatial feature extraction in 3D medical images In parallel with CNN-based methods, transformer architecture demonstrates high efficiency and adaptability to large datasets by capturing features globally through a self-attention mechanism [25].

Recent studies have further emphasized the importance of robust representation learning, attention-guided feature aggregation, and cross-dataset generalization for MRI-based brain age estimation and broader attention-based medical image analysis [27], [28], [29]. These works indicate that performance gains depend not only on deeper backbones, but also on how anatomical structure, regional dependencies, and scanner heterogeneity are handled during feature learning. In 2021, He *et al.* propose a fusion-with-attention neural network (FiA-Net) to fuse contrast and morphometry image information of raw T1-weighted MRI. FiA-Net provides an MAE of 3.85 years [30]. Then, He *et al.* propose a novel global-local transformer and relation learning mechanism to further reduce the MAE of predicted age to 2.70 and 2.38 [31], [32], respectively. However, Transformer-only architectures often underperform on high-dimensional MRI data due to their substantial parameter requirements and training complexity. Consequently, hybrid designs combining CNN with self-attention have proven more effective for MRI imaging tasks. Cai *et al.* mix the two-stream convolutional autoencoder and a graph transformer to estimate the chronological age on a large dataset, yielding an MAE of 2.71 years [33]. Wu *et al.* propose an attention-based 3D multi-scale CNN that integrates multi-scale spatial features for Alzheimer's disease classification from sMRI, achieving high classification accuracy with fewer parameters [34]. Pei *et al.* introduce a multi-scale attention-based pseudo-3D convolutional network for Alzheimer's diagnosis using structural MRI, combining pseudo-3D convolutions with attention modules to capture hierarchical spatial features [35]. Zhang *et al.* develop an anatomical feature attention-enhanced 3D-CNN that fuses engineered anatomical features with deep convolutional representations to improve brain age prediction, yielding an MAE of 2.20 years in a small dataset[36]. Although these studies demonstrate the usefulness of attention in MRI analysis, most of them use attention to recalibrate features within the same convolutional level, to aggregate global tokens in transformer branches, or to fuse manually designed anatomical descriptors with learned representations. MAR3D differs from these designs in that the AM Block first compresses 3D residual features through stage-wise max-pooling, then performs regional multi-head attention on the compacted spatial tokens, and finally restores the attended features through max-unpooling before residual-stage integration. Thus, the proposed module is intended to construct mid-level structured representations rather than only reweighting existing low-level convolutional responses. This mid-level focus enables both enhanced predictive accuracy and improved biological interpretability in 3D brain age estimation tasks.

In general, diverse deep learning approaches have achieved notable progress in brain age estimation, but each method presents inherent limitations that motivate our proposed solution. The traditional 3D CNN extracts local features by 3D convolution operators and summarizes global features by catching image information within each patch based on multiple convolutional layers. However, these so-called global features are always based on low-level features without considering structured image description. Therefore, we modified the feature extraction module as shown in Fig. 1. In this figure, we employ the 3D convolution operation to extract the low-level features of the human brain and invoke max-pooling to convert the low-level features into mid-level features with compactness. Each mid-level feature after pooling contains

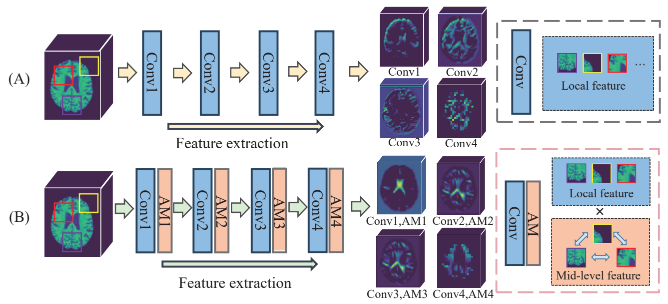


Fig. 1. The feature extraction processing. (A) Traditional convolutional networks extract only local features per layer. (B) Proposed AM Block extracts local features while preserving mid-level features by Capturing image information of different brain regions.

some overall information in different regions of the human brain, so we further utilize the weighted importance and long-distance dependencies of the self-attention mechanism to enhance the global representation ability of the model.

III. Method

A. Backbone for Feature Extraction

We select the 34-layer 3D residual neural network as the backbone to predict the chronological age of brain MRI, which is the classical and efficient deep learning model for solving computer vision tasks. The backbone network mainly contains an initial module, four residual blocks, a global average pooling, and a fully-connected (FC) layer.

1) Initial Module: This module contains a convolutional layer with a kernel size of $7\times7\times7$ and a max-pooling layer. In comparison with the common convolution operation, the initial convolutional layer employs a larger kernel size in order to cover a wider input area, thereby facilitating the capture of more local spatial information. The 3D convolution operation can be formulated as $\sum W \odot X$, where W is the convolution kernel and X is a local image extracted from the sliding window with the same scale. The max-pooling layer, immediately following the initial convolutional layer further downsamples the global feature spatially. This layer aids in reducing the spatial dimensions of the input data while retaining crucial features by identifying the maximum value as the key within a region.

2) Residual Block: Residual block is the core component of the ResNet, which is designed to solve the problem of gradient vanishing and gradient explosion in deep neural networks. The residual block introduces a concept of residual mapping, which allows the network to learn deeper feature representations more efficiently, thus improving the performance and training efficiency of the whole model. Given an input X, the function that a traditional neural network wishes to fit can be expressed as $\phi(X)$, whereas residual mapping is concerned with learning the $\phi(X)-X$. Consequently, residual mapping is more capable of capturing information on minor changes and details than traditional methods. The process of residual block focuses on the task of feature extraction from the input data, employing a series of convolution, batch normalization and activation functions. In this work, we select a 34-layers ResNet as the backbone, which has excellent feature extraction capability with lower number of parameters and resource consumption compared to the deeper network architecture.

3) Regression Model: Based on the extracted features, we call the final regression model to estimate brain age. Because directly using a large fully connected layer would introduce a substantial parameter and computational burden, global average pooling is first applied for dimensionality reduction. Finally, the fully-connected layer maps the high-level features to a scalar age prediction through a linear regression output. No bounded activation or post-hoc clipping is applied during training or evaluation. Therefore, predictions outside the physiological age range can occur in principle, although they are rare for the proposed MAR3D model.

B. AM Block

We propose a novel module, named AM Block, which aims to enhance the global representation of the backbone model on 3D MRI images. It is evident that networks which rely solely on convolutional operations are susceptible to the issue of global feature singularity. Furthermore, the global features constructed on low-level features often result in the loss of a significant amount of high-level information. To address these limitations, we incorporated one AM Block after each residual stage of the 3D ResNet backbone, rather than after every basic residual unit. In MAR3D-34, the four AM Blocks are placed after the residual stages with 64, 128, 256, and 512 channels, respectively. For an input feature tensor, the AM Block applies 3D max-pooling with stage-specific window size L. The pooled features are flattened into regional tokens, encoded by multi-head attention, restored by max-unpooling with the recorded indices, and then fused with the residual-stage output by residual addition.

The AM Block extracts mid-level features through a hierarchical process. First, a max pooling reduces dimensionality and highlights key regions. Then, the mid-level feature abstraction is achieved by an attention encoder, which transforms these compacted features into independent and semantically enriched representations, followed by unpooling and residual integration to preserve both local and global feature extraction patterns. We designed the AM Block as an additional module integrated into the original ResNet architecture and conducted a comparative analysis under a 34-layer configuration. Compared to the original ResNet model, our proposed MAR3D model increases the number of parameters by 29%, the model size by 11.2%, and the training time by 4.82%.

1) Sampling Method: Spatial pooling has been widely applied in image processing as a classical downsampling method and has shown partial correlations with mid-level representations. [37]. Considering that the main contribution of ResNet is to simplify the fit function to a residual mapping, we should ensure data sparsity when adding mid-level features as additional learning units. However, the substantial parameter dimensionality of 3D MRI features extracted by residual blocks poses significant computational challenges when directly integrated into multi-head attention mechanisms. This direct integration would not only result in degraded

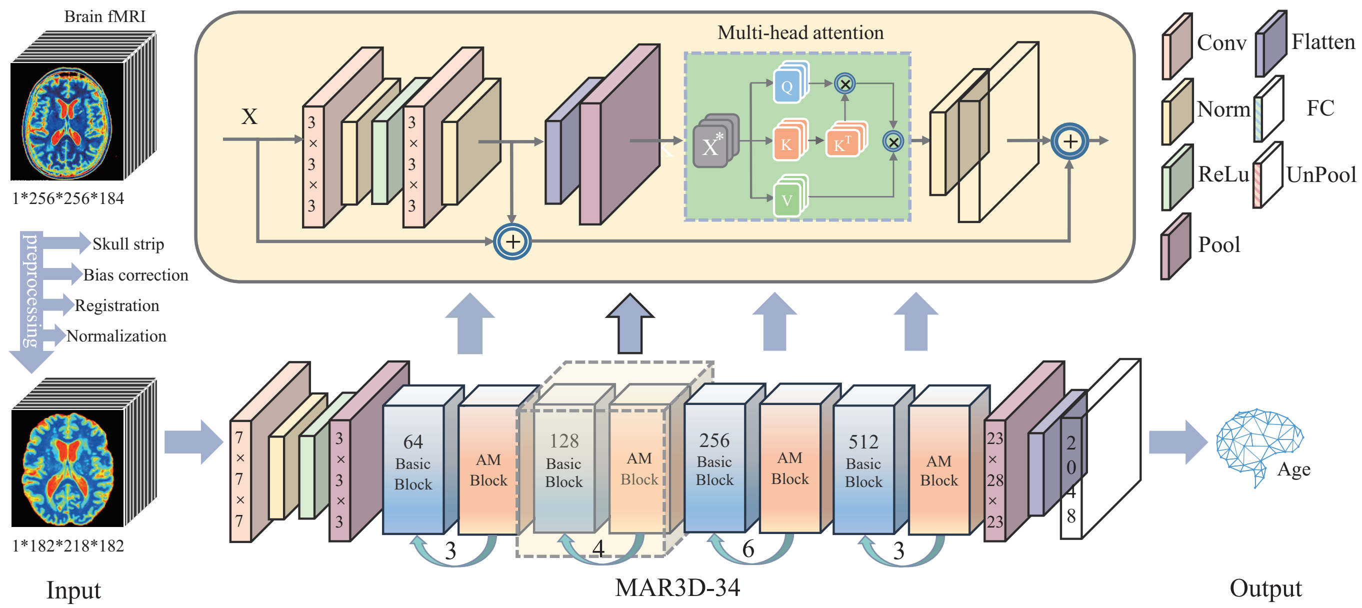


Fig. 2. The architecture of MAR3D.

performance due to excessive memory requirements, but also lead to substantial increases in computational complexity and parameter count. To address this issue, we employ 3D max pooling operations, which effectively reduce parameter dimensionality while preserving the most salient attention regions within brain structures. Furthermore, to optimize the effectiveness of the multi-head attention mechanism, we adjust the pooling window size L across different stages to ensure that the amount of pixel data within each channel aligns with the optimal parameter settings for attention computation. For a local feature $\mathcal{F}_l$ with input size (B, C, D, W, H) obtained by the residual block, the output can be pooled as a tensor $\mathcal{F}_m \in \mathbb{R}^{B \times C \times D_L \times W_L \times H_L}$, where $D_L = D/L$, $W_L = W/L$, and $H_L = H/L$ when the spatial dimensions are divisible by the pooling window size L. Since the output vector needs to utilize unpooling to restore its dimensions after the attention encoder, the indices mask of the maximum value in each region is recorded. The final unpooling operation is performed by utilizing the recorded indices from the corresponding max pooling layer to restore the original spatial positions of the encoded features.

However, mid-level features possess other properties besides data compactness that cannot be brought by pooling operations, the most important of which is data independence. Therefore, we utilize an attention encoder after max pooling to fully transform low-level features into mid-level features. Attention encoder is a data processing method based on matrix operations and fully connected networks, utilizing a multi-head attention mechanism to extract different mid-level features with data independence and multiple focal points [38], [39], [40]. Multi-head attention enables our model to simultaneously focus on different aspects of brain aging across regions and scales, where each attention head can specialize in capturing distinct meso-level features such as cortical-subcortical interactions, hemispheric asymmetries, or regional volumetric relationships, while the multi-head architecture allows parallel processing of these diverse spatial dependencies.

Algorithm 1 Pseudocode of AM Block

Require: Input $\mathbf{X} \in \mathbb{R}^{B \times C \times D \times W \times H}$
Ensure: Output $\mathbf{Y} \in \mathbb{R}^{B \times C \times D \times W \times H}$

$\mathbf{X}_{pool}, \mathcal{I} \leftarrow \text{MaxPool}(\mathbf{X})$ $\{\mathbf{X}_{pool} \in \mathbb{R}^{B \times C \times \frac{D}{L} \times \frac{W}{L} \times \frac{H}{L}}\}$
$\mathbf{X}' \leftarrow \mathcal{LN}(\text{Flatten}(\mathbf{X}_{pool}))$
$\mathbf{QKV} \leftarrow \text{Linear}(\mathbf{X}')$ $\{\mathbf{QKV} \in \mathbb{R}^{3 \times B \times C \times \frac{DWH}{L^3}}\}$
$\mathbf{Q}, \mathbf{K}, \mathbf{V} \leftarrow \text{MultiHead}(\mathbf{QKV})$ $\{\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B \times head \times \frac{DWH}{L^3} \times d_h}\}$
$\mathbf{Attn} \leftarrow \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_h}}\right)$
$\mathbf{Out} \leftarrow \mathbf{AttnV}$ $\{\mathbf{Out} \in \mathbb{R}^{B \times head \times \frac{DWH}{L^3} \times d_h}\}$
$\mathbf{Out} \leftarrow \mathcal{LN}(\text{Linear}(\text{Concat}(\mathbf{Out})))$
$\mathbf{Y} \leftarrow \text{Reshape}(\text{MaxUnpool}(\mathbf{Out}, \mathcal{I}))$
return $\mathbf{Y}$

2) *Regional Attention Encoder:* Considering that the input of 3D images does not fulfill the conditions of the attention encoder itself, we will flatten the sampled data $\mathcal{F}_m$ to $\mathcal{F}'_m$ by a flatten operation, and the multi-head attention encoding output can be formulated as:

$$Multi\_head = w^o(AT_1, AT_2, ..., AT_{head}) + b^o \tag{1}$$

where Q, K, V denote the query, keys, values matrices, which are respectively obtained through linear transformation operations $\mathcal{W}^Q \odot \mathcal{F}'_m$, $\mathcal{W}^k \odot \mathcal{F}'_m$, and $\mathcal{W}^v \odot \mathcal{F}'_m$. The parameter matrices $\mathcal{W}^Q \in \mathbb{R}^{d_q}$, $\mathcal{W}^k \in \mathbb{R}^{d_k}$, and $\mathcal{W}^v \in \mathbb{R}^{d_v}$, where the number of dimensions d_q=d_k=d_v=$\frac{D*W*H}{L^3}$. w^o and b^o are the weights and thresholds in fully-connected layer after concatenating the $head$ attentions. AT_i represents an attention,

TABLE II
THE DETAILS OF DATASETS USED FOR BRAIN AGE ESTIMATION.

Dataset	Samples #	Range of ages	Average age
ADNI	2692	55.0-93.0	76.92
CoRR	3200	6.0-88.0	26.31
SALD	494	19.0-80.0	45.18
DLBS	315	20.6-89.1	54.61
IXI	563	19.9-86.3	48.65
OASIS-1	416	18.0-96.0	52.70
Overall	7680	6.0-96.0	49.28

it is formulated as:

$$\begin{aligned} AT_i &= Attention(\mathcal{W}^Q \odot Q, \mathcal{W}^K \odot K, \mathcal{W}^V \odot V) \\ Attention(Q,K,V) &= softmax(\frac{QK^T}{\sqrt{d}})V \end{aligned} \quad (2)$$

For each different head, the $Attention$ mechanism first calculates the matrix product of Q and K, then performs a normalization operation to reduce the value of the product, and finally multiplies by V after applying the softmax function to obtain the encoding result. The $head$ attentions form a multi-head attention mechanism to represent different aspects of features at multiple scales, thereby extracting deeper mid-level features. The re-encoded sequence data $\mathcal{F}'_m$ assigns new weights to each region of the original input, rendering it completely independent of the previous image information. This provides richer mid-level features for the residual model. Following the unflattening operation, $\mathcal{F}'_m$ restores the 3D shape and proceeds to the next round of residual block computation. The full pseudocode of the AM Block is presented in Algorithm 1, where $\mathcal{W}^Q \in \mathbb{R}^{d_h \times d_q}$, $\mathcal{W}^K \in \mathbb{R}^{d_h \times d_k}$, and $\mathcal{W}^V \in \mathbb{R}^{d_h \times d_v}$ are the projection matrices for each attention head. For clarity, we denote the pooled spatial size as $\frac{DWH}{L^3}$ and the per-head embedding dimension as d_h. After flattening, each channel produces regional tokens. The query, key, and value tensors are then projected and reshaped into $B \times head \times \frac{DWH}{L^3} \times d_h$. The proposed AM Block provides a unified framework for mid-level feature extraction in 3D brain MRI analysis. By sequentially combining max-pooling for dimensionality reduction, attention encoding for semantic abstraction, and unpooling for spatial restoration, the AM Block transforms original convolutional outputs into richer and more independent representations. These mid-level features extend beyond simple spatial compactness, capturing cross-regional dependencies, interactions between cortical and subcortical, and intermediate structural patterns that are critical for brain age estimation.

C. Training Method

Our proposed network is completely constructed by the torch framework based on PyTorch platform, and we utilize the Adaptive Moment Estimation (Adam) optimizer to update the parameter with a learning rate of 1e-3. The loss function we selected is the mean absolute error, which is formulated as follows:

$$\mathcal{L} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \quad (3)$$

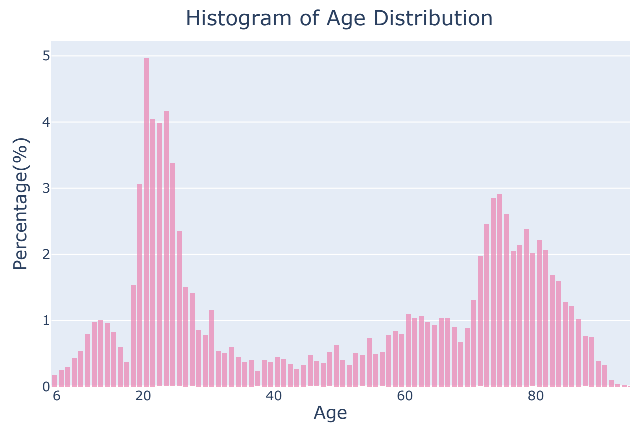


Fig. 3. The age distribution histogram of the entire dataset.

where y_i and $\hat{y}_i$ respectively denote the chronological and predicted age of ith instance, N is the number of instances in each batch. In this work, we set the learning rate α as 1e-3 with training epochs of 50, and the batch size is defined as 6 due to the limitation of video card. The model was trained for 24 hours on a single NVIDIA RTX A6000 GPU with 48G GPU memory.

IV. EXPERIMENT SETUP

In this section, we introduce the selected public brain MRI image datasets, including the source and details of the data. Subsequently, we outline the preprocessing process employed for all MRI images. The experimental part of this work consists of two components: the first component gives the performance comparison of MAR3D with state-of-the-art models on the same dataset, while the second component involves an ablation experiment, leveraging comprehensive tables and visualizations to validate the effectiveness of our proposed AM Block. Our code is available at: https://github.com/zy-251026/MAR3D. The repository will include the model implementation, training scripts, main preprocessing instructions, random seed settings, and split information permitted by the licenses of the public datasets.

A. Datasets

In this paper, we aggregate six publicly available datasets to validate the evaluation performance of network model, the details can be shown in Table II. The Alzheimer's Disease Neuroimaging Initiative (ADNI) collects the MRI and PET images to develop biomarkers of the disease, and the image data can be obtained from the websites https://adni.loni.usc.edu/. Consortium for Reliability and Reproducibility (CoRR) [41], Southwest University Adult Lifespan Dataset (SALD) [42], and Dallas Lifespan Brain Study (DLBS) integrate brain fMRI images from labs around the world, which can be found for download at the website https://fcon_1000.projects.nitrc.org/. Information eXtraction from images (IXI) dataset collects images from three hospitals in London, it is available for download at https://brain-development.org/ixi-dataset/. The Open Access Series of Imaging Studies (OASIS) [43] is

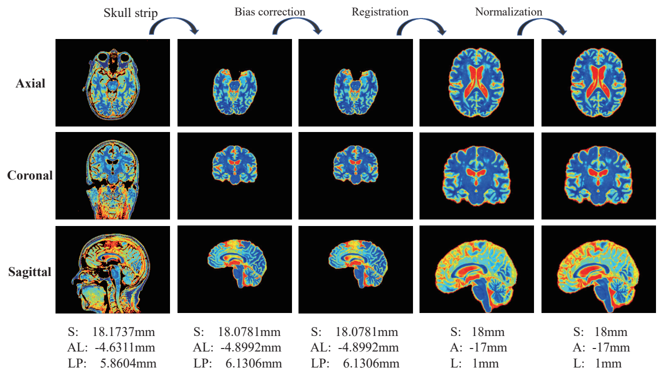


Fig. 4. The illustration of preprocessing method for an example, dataset: ADNI1, ID: I109517, age: 76. (S: Superior, A: Anterior, L: Left, AL: Anterior-Left, LP: Left-Posterior.)

a freely available neuroimaging community with the website https://sites.wustl.edu/oasisbrains/. It should be noted that ADNI contains subjects with different clinical statuses, including cognitively normal controls, mild cognitive impairment, and Alzheimer's disease, while OASIS-1 includes nondemented and demented older adults. In the present study, these diagnostic labels were not used as model inputs. Instead, the model was trained to estimate chronological age from T1-weighted MRI. Sex and diagnostic metadata were not used as input features in the current model. Because the availability and format of these demographic variables differ across the six public datasets, the present experiments focus on chronological age prediction from T1-weighted MRI alone. Incorporating harmonized demographic and clinical variables will be considered in future work.

When selecting the dataset, two primary considerations were taken into account. Firstly, the dataset must encompass MRI images across a wide range of age groups. Secondly, mean age of the total dataset should approximate the median of the age range. The former takes into account current theoretical results from aging-related research, which indicate that as age increases, brain grey matter gradually decreases, while white matter increases until around age 40 and subsequently decreases [44]. Therefore, while white matter volumes are comparable at 10 and 75 years of age, the total dataset spanning ages 6 to 96 years poses a significant challenge for deep learning models to accurately estimate age based solely on image data. Additionally, the sample set with a mean age of 49.28 years ensures confidence in evaluating the performance of our proposed models, which avoids misleading accuracy caused by data imbalance. As illustrated in Fig. 3, the age distribution of the dataset is highly imbalanced, with a notably smaller number of samples in the 30–60 age range compared to the 20-30 and 70-85 range. This imbalance introduces considerable challenges to the robustness of model and its capacity to generalize across different age groups. It is worth noting that the images in ADNI dataset come not only from cognitively normal healthy individuals but also from those with mild cognitive impairment and Alzheimer's disease. This further highlights the high-level performance requirements of feature extraction model for our chosen datasets.

TABLE III
THE AVERAGE PERFORMANCE OF DIFFERENT ESTIMATION MODELS IN THE TRAINING PROCESSING.

Model	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_b⇑	CS(α=3)⇑	$p(r)$-value
MAR3D-34	**2.298**±0.060	4.140	**0.988**	**0.979**	**0.896**	**79.50%**	-
3DCNN-AFAC	3.414±0.161	4.745	0.983	0.957	0.827	54.84%	0.0019(0.97)
EfficientNet-b4	2.898±0.181	4.718	0.983	0.968	0.865	67.44%	0.0273(0.69)
CL-3DResNet	2.801±**0.057**	4.328	0.986	0.969	0.866	70.39%	0.0019(0.97)
SFCN	3.198±0.269	4.369	0.987	0.966	0.849	59.97%	0.0019(0.97)
M3T	3.767±0.111	5.368	0.978	0.956	0.824	55.26%	0.0019(0.97)
PCRLV2	2.465±0.101	**3.980**	0.987	0.978	0.893	76.30%	0.0625(0.58)

Addition: 95% confidence intervals of MAE: MAR3D-34: [2.224, 2.375], 3DCNN-AFAC: [3.323, 3.780], EfficientNet-b4: [2.673, 3.122], CL-3DResNet: [2.742, 2.879], SFCN: [2.862, 3.531], M3T: [3.627, 3.904], PCRLV2: [2.325, 2.572].

B. Preprocessing

Since the equipment and settings used for the collected MRI images are different, each image has different attributes such as resolution and orientation. Therefore, it is necessary to employ a series of preprocessing techniques for standardization, the process includes the following steps: (1) Skull stripping (2) N4 bias field correction (3) Template registration (4) Voxel normalization. Skull stripping is used to separate the brain and skull to reduce the interference during feature extraction. Here, we choose the trained UNet model for segmentation. After extraction of the brain, the N4 bias field correction is executed to correct intensity inhomogeneities or biases within the images [45]. Template registration provides a unified reference framework, which utilize the FMRIB software library (FSL) to reconstruct each MRI image based on an MNI152 standard-space T1-weighted average structural template [46]. Finally, the Gaussian mixture modeling approach is used to calculate the normalization values. After the preprocessing method, each MRI image can be resized to 182*218*182 with isotropic spatial resolution of 1 mm^3, an example has been shown as Fig. 4. The preprocessing methods are explained in detail in the appendix of the paper.

C. Performance metric

We employ six performance metrics to quantify evaluation of brain age estimation, containing MAE, root mean square error ($RMSE$), Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (R_s), Kendall rank correlation coefficient (τ_b), and cumulative score with threshold A (CS_A). MAE is defined as (3), and the other five indicators are formulated as:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y_i})^2}{N}} \tag{4}$$

$$PCC = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y_i} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y_i} - \bar{\hat{y}})^2}} \tag{5}$$

$$R_s = 1 - \frac{6\sum_{i=1}^{N}(R(y_i) - R(\hat{y_i}))^2}{N(N^2 - 1)} \tag{6}$$

$$\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d - T_y)(N_c + N_d - T_{\hat{y}})}} \tag{7}$$

$$CS_A = \frac{N_{|y_i - \hat{y_i}| \le A}}{N} \times 100\% \tag{8}$$

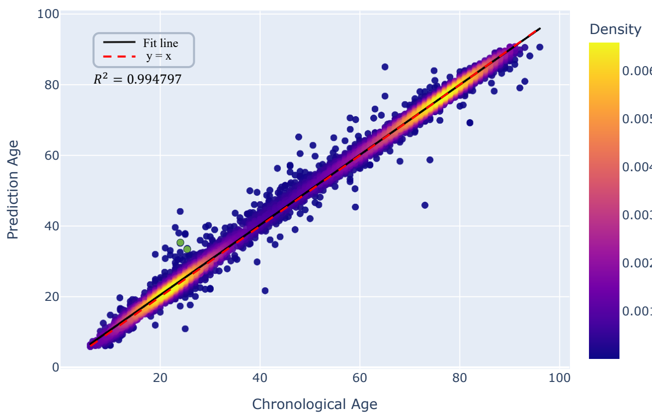


Fig. 5. The scatter diagram of prediction and chronological age obtained by MAR3D-34.

where R denotes the relative position label of the observation, N_c and N_d represent the number of concordant pairs and discordant pairs, and T is the number of ties only in the variant. We employ several performance metrics to provide a comprehensive performance analysis of model from different aspects, adding consideration of linear correlation and robustness evaluation except for error.

D. Competitors

We select several representative deep learning models as competitors to our proposed MAR3D, including a CNN model enhanced by anatomical feature attention (3DCNN-AFAC) [36], a 117-layer CNN model with a highly effective compound coefficient (EfficientNet-b4) [47], and a contrastive learning-based brain age estimation model (CL-3DResNet) trained with the age-aware L_{exp} objective [26]. In addition to these CNN-based models, we consider two state-of-the-art voxel-based neural networks as competitors, named M3T [48] and PCRLV2 [49]. M3T is a 3D MRI classifier combining the characteristics of both CNNs and Transformers, effectively recognizing MRI images with multi-plane and multi-slice information. PCRLV2 incorporates pixel-level information into high-level features to improve the representation ability of the model. All comparison models were trained on the same preprocessed MRI volumes and evaluated under the same repeated random split protocol. For architectures originally designed for classification, the final classification layer was replaced by a single linear regression output, and the models were trained using the same MAE loss unless otherwise specified. For the contrastive-learning-based brain age model, we followed its original L_{exp}-based training strategy and used the resulting age prediction output for comparison.

V. Experimental result

The experimental results are divided into three main sections. First, we compare our proposed model against classical CNN models and the SOTA 3D transformers across multiple performance metrics. Then for MAR3D, we compare with several different original backbone models, proving the effectiveness of AM Block. Finally, we visualize which regions

TABLE IV
The optimal performance of trained estimation models on the entire dataset.

Model	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_b⇑	CS(α=3)⇑
MAR3D-34	**1.250**	**1.893**	**0.997**	0.993	0.942	**93.19%**
3DCNN-AFAC	3.185	16.76	0.831	0.966	0.853	61.84%
EfficientNet-b4	1.638	2.486	0.995	0.989	0.925	87.02%
CL-3DResNet	1.541	6.59	0.96	0.991	0.936	91.07%
SFCN	2.413	5.015	0.982	0.980	0.894	72.27%
M3T	2.878	3.986	0.988	0.968	0.859	63.31%
PCRLV2	1.599	2.532	0.996	**0.994**	**0.950**	91.42%

of the human brain are key for predicting age based on the attention scores.

A. Comparison with classical and SOTA models

1) Training process: Each model was run 10 times independently with 50 epochs, and the final predicted age was recorded to calculate the mean performance. During each training session, the entire dataset (7680) was randomly split at the subject level into the training set (90%, 6912) and the test set (10%, 768). When repeated scans or follow-up acquisitions were available for the same subject, they were kept within the same subset to avoid subject-level data leakage. The batch size was set between 1 and 8, depending on the GPU's memory limit. Table III demonstrates the performance of MAR3D and other competitor models for the total dataset on six performance metrics. As Table III shown, MAR3D-34 performs superior than other competitors across most metrics. Specifically, compared to three CNN-based models, our proposed method improves MAE by 17.9%-32.6%, RMSE by 4.3%-12.8%, and CS (α=3) by 11.45%-31.01%. In comparison to the transformer variant model M3T, MAR3D-34 performs significantly higher CS, and the MAE and RMSE of MAR3D-34 are lower by 1.469 years and 1.228 years, respectively. Compared with PCRLV2, MAR3D-34 achieves a 6.8% improvement in MAE and a 3.2% increase in CS, whereas PCRLV2 obtains a slightly lower RMSE by 0.16 years. This indicates that MAR3D provides stronger average absolute-error performance and threshold-based accuracy, while PCRLV2 remains competitive in squared-error sensitivity. In Table III, we present the p-values and effect sizes obtained from the paired Wilcoxon signed-rank test, using the MAE values from the 10 repeated random splits as paired observations. MAR3D-34 was used as the reference method in each comparison. The effect size was calculated as $r = |Z|/\sqrt{n}$, where Z is the standardized Wilcoxon statistic and n is the number of paired observations. Except for PCRLV2, all p-values are below 0.05, demonstrating that MAR3D-34 achieves statistically significant improvements. In contrast, although the difference between MAR3D-34 and PCRLV2 is not statistically significant at the 0.05 level, MAR3D-34 still performs favorably on MAE, rank correlations, and CS. Therefore, PCRLV2 should be regarded as the closest competing method in our comparison. In addition, MAR3D significantly outperformed most competitors with large effect sizes (r-value>0.5) and stable confidence intervals, indicating substantial practical differences rather than marginal statistical

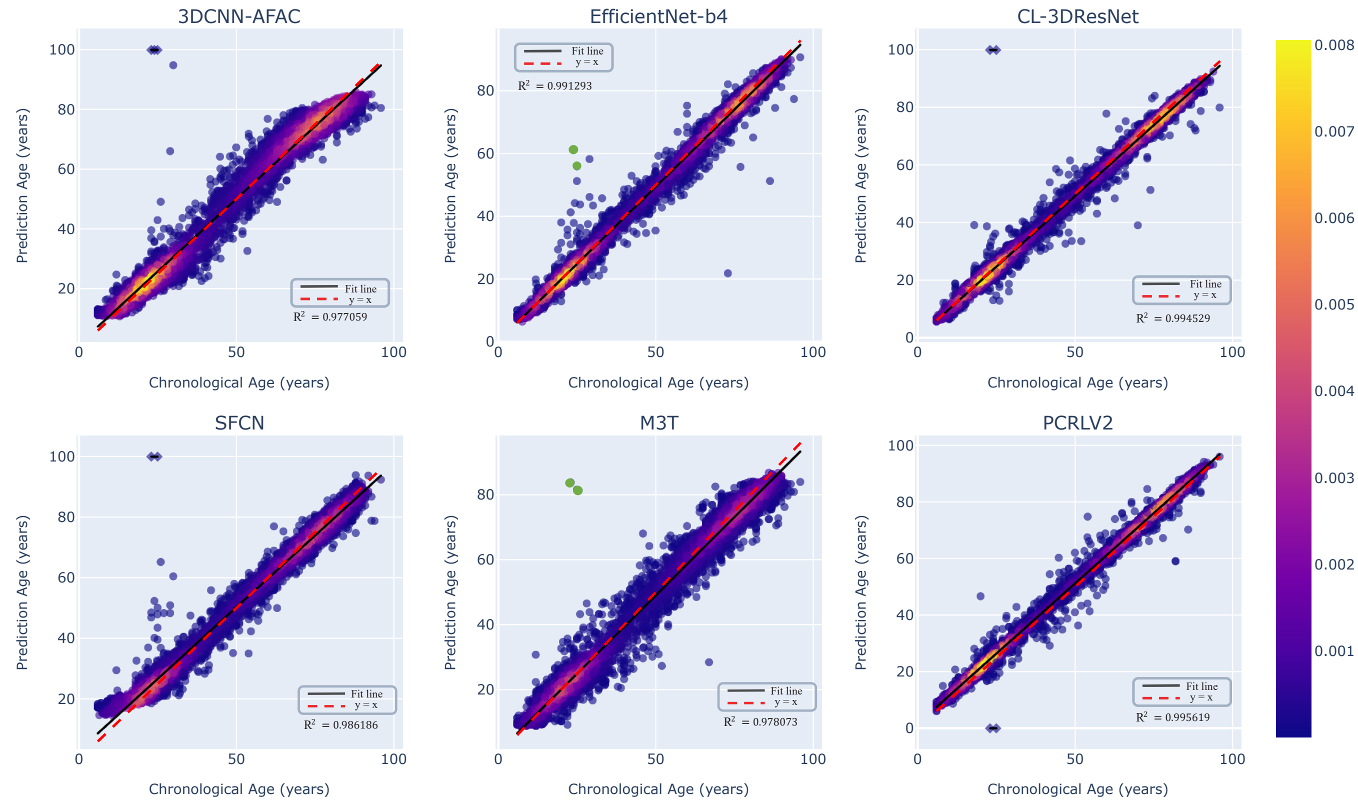


Fig. 6. The scatter diagram of prediction and chronological age obtained by other competitors.

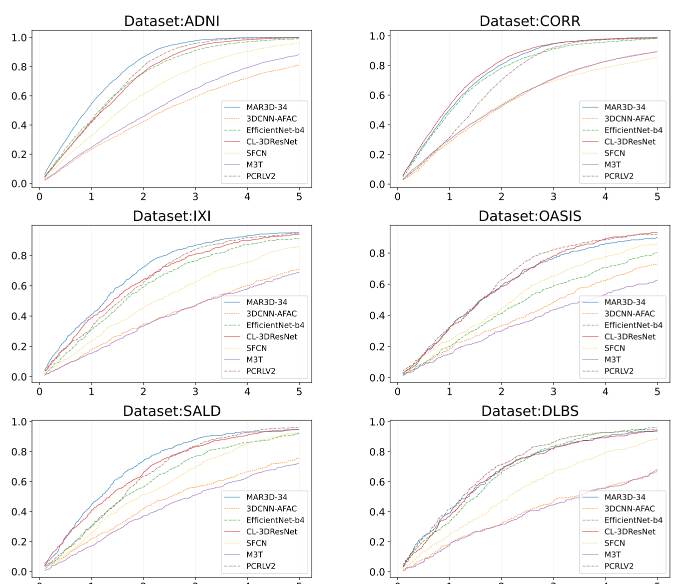


Fig. 7. The CS curve of error threshold α from 0.1 to 5 years of different estimation models on six sub-datasets, each curve is taken at intervals of 0.001.

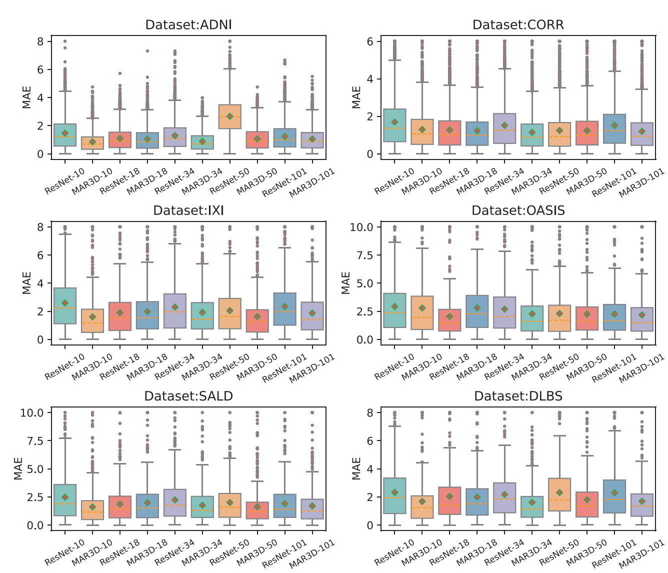


Fig. 8. The box-and-whisker plots with overlaid scatters of different-layer architectures on six sub-datasets, for each MAR3D and ResNet model, we select the most commonly used architecture of 10, 18, 34, 50, and 101 layers.

differences. Fig. 5 and Fig. 6 illustrate the scatter plots of our proposed MAR3D-34 and other competitors with chronological age as the x-axis and predicted age as the y-axis, the color of scatter points represents the density of a region, while the diamond-shaped points are marked as outliers. From this figure, we can find that the predicted ages obtained by MAR3D maintain a low error compared to the chronological ages, without outliers. In contrast, 3DCNN-AFAC, CL-3DResNet, SFCN, and PCRLV2 generate several outliers when predicting two special samples with chronolog-

TABLE V
THE EXPERIMENTAL RESULT OF MAR3D COMPARED BY BACKBONE MODEL ON THE WHOLE DATASET.

Model	Layer	MAE$\Downarrow$	RMSE$\Downarrow$	PCC$\Uparrow$	$R_s\Uparrow$	$\tau_b\Uparrow$	CS(α=3)$\Uparrow$	p-value(E_s)
ResNet	10	3.310±0.392	17.83	0.811	0.969	0.871	70.74%	0.0019(0.97)
	18	2.734±0.193	4.233	0.986	0.975	0.878	71.60%	0.0039(0.91)
	34	3.475±0.136	4.766	0.983	0.954	0.825	71.71%	0.0097(0.81)
	50	4.319±0.269	27.17	0.655	0.961	0.846	55.75%	0.0019(0.97)
	101	3.530±0.264	5.970	0.974	0.957	0.832	59.30%	0.0019(0.97)
MAR3D	10	2.602±0.188	7.967	0.955	0.975	0.887	76.22%	0.0273(0.69)
	18	2.668±0.120	4.273	0.986	0.977	0.884	73.25%	0.0039(0.91)
	34	**2.298±0.060**	4.140	**0.987**	**0.979**	**0.896**	**79.50%**	-
	50	2.558±0.159	**4.007**	**0.987**	0.976	0.881	74.24%	0.0371(0.55)
	101	2.728±0.145	4.127	**0.987**	0.972	0.871	69.91%	0.0039(0.91)

Addition: 95% confidence intervals of MAE, ResNet, 10: [2.822, 3.797], 18: [2.491, 2.972], 34: [3.278, 3.672], 50: [3.984, 4.655], 101: [2.976, 3.631], MAR3D, 10: [2.355, 2.825], 18: [2.520, 2.819], 34: [2.224, 2.375], 50: [2.339, 2.736], 101: [2.595, 2.956].

TABLE VI
THE EXPERIMENTAL RESULT OF MAR3D WITH DIFFERENT PARAMETER OF AM BLOCK.

Model	head	d_h	MAE$\Downarrow$	RMSE$\Downarrow$	PCC$\Uparrow$	$R_s\Uparrow$	$\tau_b\Uparrow$	CS(α=3)$\Uparrow$	p-value(E_s)
MAR3D-34	4	132	3.142±0.153	5.906	0.975	0.966	0.857	67.92%	0.0019(0.97)
	6	88	2.952±**0.046**	4.281	0.987	0.965	0.851	65.55%	0.0019(0.97)
	8	66	2.945±0.091	5.828	0.975	0.971	0.867	67.84%	0.0019(0.97)
	12	44	2.873±0.205	4.633	0.984	0.973	0.869	71.07%	0.0019(0.97)
	16	33	2.670±0.118	4.185	0.987	0.970	0.868	72.50%	0.0039(0.91)
	22	**24**	**2.298**±0.060	4.140	**0.988**	**0.979**	**0.896**	**79.50%**	-
	24	22	2.462±0.091	**4.061**	0.988	0.977	0.888	76.56%	0.0429(0.64)
	48	11	2.929±0.162	10.503	0.928	0.977	0.886	72.66%	0.0019(0.97)
	66	8	4.631±0.208	32.822	0.583	0.962	0.854	64.38%	0.0019(0.97)

Addition: 95% confidence intervals of MAE, (4,132): [2.955, 3.336], (6, 88): [2.829, 3.058], (8, 66): [2.801, 3.095], (12, 44): [2.616, 3.127], (16, 33): [2.522, 2.817], (22, 24): [2.224, 2.375], (24, 22): [2.346, 2.573], (48, 11): [2.728, 3.131], (66, 8): [4.371, 4.888].

ical ages of 24 and 25, resulting in predictions greater than 100 or less than 0. EfficientNet-b4, M3T, and MAR3D-34 did not predict outliers in these two special samples. However, EfficientNet-b4 and M3T exhibited severe prediction errors, with age estimates deviating from the chronological ages by 35 years and 57 years, respectively. In contrast, our proposed MAR3D achieved substantially smaller errors of only 12 years and 10 years for these two samples. This indicates that MAR3D effectively addresses both overfitting and underfitting, providing accurate age estimations even for challenging brain instances. Additionally, the black line represents the ordinary least squares (OLS) regression line for all the scatter points, and the red dotted line indicates the ideal estimation $y = x$ (when the predicted age equals the chronological age). R^2 is the fit of linear regression to describe the correlation between two fitted lines, with a positive correlation near 1, a negative correlation near -1, and no correlation near 0. It is obvious that the fit line of MAR3D closely aligns with the ideal $y = x$ line, and a R-value of 0.994 demonstrates the effectiveness of our proposed model.

2) Model performance: Since the image data during the training process are different caused by differences in dataset split, resulting in unstable performance of the model. Therefore, we test the optimal trained model on the entire dataset and analyzed the prediction results of brain age for each sub-dataset. Table IV shows the best prediction accuracy of models for the total dataset, where the bold values represent the best performance among all models. It is worth noting that only MAE and RMSE are preferable when lower, while the remaining performance metrics are opposite. As Table IV shown, MAR3D-34 achieves dominant performance in each metric. In comparison with other competitor models, MAR3D-34 reduces MAE by 0.343-1.935 years, RMSE by 0.593-14.86, and increase CS (α=3) by 1.77%-31.35%. In terms of three correlation, MAR3D-34 achieved the best result in PCC and ranked only below PCRLV2 in R_s and τ_b.

Fig. 7 displays the cumulative score curve of each model for six sub-datasets with different age thresholds A. It is obvious that MAR3D-34 achieves dominant performance in all ranges of α in the ADNI, IXI, and SALD datasets. In the remaining datasets, MAR3D-34 also shows performance equivalent to SOTA models. In particular, ADNI includes MRI images with labels for normal, mild cognitive impairment, and Alzheimer's disease. Despite this complexity, MAR3D-34 can still accurately estimate the age of each brain with a low error.

B. Ablation experiment

In this paper, we incorporate AM Block into ResNet as an additional module to extract the mid-level features of input images, resulting in MAR3D. Without the AM Block, MAR3D is equivalent to the original ResNet. To evaluate the performance of proposed Block, we give the experimental results of MAR3D and ResNet with commonly used architecture. Table V lists the prediction metrics of MAR3D and ResNet with different numbers of layer on the entire datasets, where the statistical column reports the paired Wilcoxon signed-rank test against MAR3D-34 and the corresponding effect size.

According to the experimental results in Table V, it is obvious that MAR3D-34 has a significant advantage in comparison with other architectures except for RMSE and PCC, and it ranked first based on the Wilcoxon test. The better performance of MAR3D-34 compared with deeper MAR3D-50 and MAR3D-101 may be related to the balance between representation capacity and optimization stability. Since 3D MRI inputs require small batch-size training and contain many scanner- and subject-specific variations, deeper networks may introduce redundant parameters and become more sensitive to overfitting. The AM Block already enhances regional representation in the 34-layer backbone, so further depth does not necessarily provide additional benefit. Furthermore, MAR3D outperforms the original ResNet model at each layer with all p-value<0.05 and r-value>0.5, demonstrating that the mid-level features extracted from AM Block effectively provide more information and knowledge in the MRI images to the regression model, resulting in a superior brain age prediction accuracy. Notably, ResNet exhibits an especially high RMSE at 10 and 50 layers, indicating that it produced exceedingly large or small predictions for brain age, leading to a poor performance when squared. In contrast, the MAR3D model rarely produces outliers, proving that incorporating AM Block can further enhance the robustness of model.

To demonstrate the performance of different model architectures on various sub-datasets, Fig. 8 illustrates box-and-whisker plots of the MAE for each model when estimating distinct datasets. It can be observed that the boxes for MAR3D are generally better than those of ResNet under the same conditions, indicating that our proposed AM Block is an effective mechanism to reduce errors between prediction ages

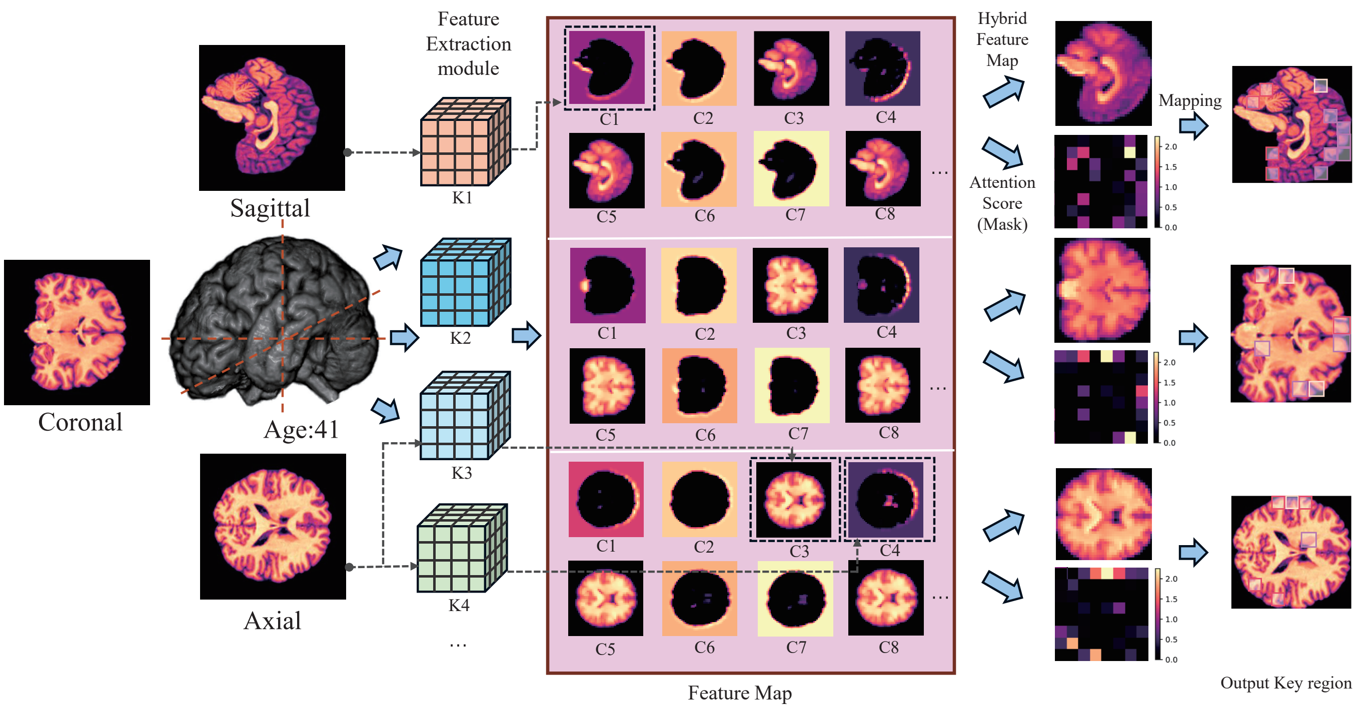


Fig. 9. Representative visualization of model-derived attention weights in one MRI brain from three anatomical planes: axial, coronal, and sagittal.

and chronological ages. Additionally, the median line for MAR3D is lower than that for ResNet at each layer, further proving the superior performance of the former compared to the latter. It is worth noting that the median position is more representative of the model performance than the mean value in the case of more outliers.

To further investigate the optimal hyperparameter configuration of the AM Block, we conduct additional ablation studies on the multi-head attention mechanism by varying the number of attention heads and embedding dimensions. Table VI presents the performance indicators of MAR3D-34 with different numbers of $head$ and d_h. In this ablation experiment, since the flattened dimension for the attention computation is 528, we select different factors of 528 as the number of heads. The final configuration was selected according to the average MAE over the same repeated-split evaluation protocol used for the other ablation settings, all candidate head numbers were evaluated under the same training and testing procedure in Table VI. From the evaluation metrics and p-values shown in this table, it can be observed that the best performance is achieved when the number of heads and embedding dimensions are set to ($head$=22, d_h=24), and the performance gradually improves as $head$ approaches 22. This observation indicates that brain age prediction from MRI data requires a relatively large number of attention heads to ensure sufficient feature diversity and capture the complex spatial relationships inherent in neuroimaging data. Conversely, configurations with either insufficient or excessive heads result in suboptimal performance due to inadequate representational capacity or attention dilution. These findings confirm that the multi-head attention mechanism in AM Block requires careful tuning to achieve optimal brain age prediction performance, and our chosen configuration represents an effective balance between model expressiveness and generalization capability.

C. Attention Score

Attention scores are a set of weight coefficients used in attention mechanisms to measure the importance of different positions in the input data. They indicate how much attention the model pays to each position in the input when calculating each output. For brain MRI images, attention weights provide a model-derived indication of which regions are emphasized during prediction. These maps are useful for qualitative interpretation, but they should not be regarded as direct causal or biological evidence without further atlas-based and neurobiological validation. Attention weights are typically calculated based on the similarity between Q and K, and then multiplied by V to obtain a weighted summation. This calculation method can be illustrated by (2).

Fig. 9 presents a representative qualitative example using a brain MRI with a chronological age of 41. This visualization is intended to illustrate how the trained model distributes attention across anatomical planes, rather than to establish a population-level neurobiological pattern. After feeding the whole 3D image into our proposed MAR3D-34, we obtain multiple 2D feature maps from different channels during the process of feature extraction. For a better observation of the brain, only the sagittal, coronal, and axial directions are shown in this figure. After channel-wise averaging, we derive a representative feature map and an attention scoring mask. By integrating the acquired results, we can analyze the brain's regions with varying degrees of informativeness for age estimation from three perspectives.

From the sagittal view, high attention weights are observed near the superior sagittal sinus region. Because this structure is not brain parenchyma, the corresponding response may reflect vascular, boundary, registration, or preprocessing-related information rather than direct age-related tissue change. Furthermore, the straight gyrus and cerebellum are situated within a slightly lower weights area, thereby providing valuable data for age estimation. Additionally, the medial frontal gyrus contributes a relatively smaller but still observable model response. When observed from the coronal view, the postcentral gyrus and superior temporal gyrus are the most informative, followed by the middle temporal gyrus. It is noteworthy that the medial frontal gyrus is identified as significant regions from the sagittal view, also hold substantial weights from the coronal view. From the axial view, the region of postcentral gyrus provides the most information, with the optic radiation and lateral sulcus also having high weights. Lastly, the caudate nucleus offer some additional information.

VI. Discussion and Conclusion

In this work, we propose a novel network architecture for identifying the age of brain MRI images, called MAR3D. This model uses a 3D ResNet as its backbone and incorporates an attention mechanism-based AM Block to provide mid-level features to the residual blocks, thereby constructing more comprehensive global features for the final regression model. Our proposed MAR3D demonstrates excellent performance on six public MRI datasets, including evaluation metrics such as MAE, RMSE and correlation coefficients. According to the results in Table I and Table III, MAR3D outperforms current state-of-the-art methods in terms of error range for both 3D volume and 2D slice approaches, and it outperforms other 3D SOTA network models in most evaluation metrics. In ablation experiments, we compared the performance differences between MAR3D architectures with different layer depths and the backbone network, showing that our proposed AM Block improves accuracy for ResNets of any depth.

In addition to improving the accuracy of brain age prediction, we also visualized the regions of focus of the AM Block in the trained MAR3D model when extracting features from different brain areas. The slices from three perspectives highlight some key regions that provide critical information for age estimation in Fig. 9. The regions identified by our proposed model as containing important information for brain age estimation are consistent with those mentioned in existing research like caudate nucleus, medial frontal gyrus and so on [50]. These observations suggest that MAR3D tends to emphasize regions that are broadly consistent with reported age-related anatomical patterns, although further quantitative and biological validation is required. At the same time, we believe that MAR3D holds great promise for solving other 3D image recognition tasks as well.

Several limitations should be noted. First, although the combined cohort covers a broad age range, its age distribution is imbalanced, particularly for middle-aged subjects. Second, the six datasets were acquired from different sites and scanners, and residual site-related effects may remain even after preprocessing and registration. Third, the attention maps presented in this study provide only preliminary model-based interpretability. Therefore, atlas-level quantitative validation and independent neurobiological analyses are required before strong anatomical conclusions can be drawn.

Acknowledgment

The ADNI dataset is funded by the National Institutes of Health Grant U19 AG024904, OASIS is supported by P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Thank you to the researchers from the funding projects mentioned above for providing the MRI data used in this work.

References

[1] D. Harman, "The aging process.," *Proc. Natl. Acad. Sci. U.S.A. (PNAS)*, Nov. 1981, vol. 78, no. 11, pp. 7124–7128.
[2] Y. Liu *et al.*, "The interaction between ageing and Alzheimer's disease: insights from the hallmarks of ageing," *Transl. Neurodegener.*, vol. 13, no. 1, p. 7, 2024.
[3] J. H. Cole and K. Franke, "Predicting age using neuroimaging: innovative brain ageing biomarkers," *Trends Neurosci.*, vol. 40, no. 12, pp. 681–690, 2017.
[4] F. Haugg *et al.*, "Imaging biomarkers of ageing: a review of artificial intelligence-based approaches for age estimation," *Lancet Healthy Longev.*, vol. 6, no. 7, p. 100728, 2025.
[5] M. Tanveer *et al.*, "Deep learning for brain age estimation: A systematic review," *Inf. Fusion*, vol. 96, pp. 130–143, 2023.
[6] L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza, and A. Mechelli, "Machine learning for brain age prediction: Introduction to methods and clinical applications," *EBioMedicine*, vol. 72, 2021.
[7] N. K. Dinsdale *et al.*, "Learning patterns of the ageing brain in mri using deep convolutional networks," *NeuroImage*, vol. 224, p. 117401, 2021.
[8] Y. Zhao, A. Klein, F. X. Castellanos, and M. P. Milham, "Brain age prediction: Cortical and subcortical shape covariation in the developing human brain," *NeuroImage*, vol. 202, p. 116149, 2019.
[9] N. Raz *et al.*, "Regional Brain Changes in Aging Healthy Adults: General Trends, Individual Differences and Modifiers," *Cereb. Cortex*, vol. 15, no. 11, pp. 1676–1689, 2005.
[10] H. Song *et al.*, "Subcortical brain volumetric differences related to white matter lesion volume and cognition in healthy aging," *npj Aging*, vol. 11, no. 1, p. 44, 2025.
[11] Y. Liu *et al.*, "Structural and iron content changes in subcortical vascular mild cognitive impairment: a combined voxel-based morphometry and quantitative susceptibility mapping study," *Brain Res. Bull.*, vol. 220, p. 111160, 2025.
[12] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Netw.*, vol. 33, no. 12, pp. 6999–7019, 2022.
[13] T. Xu *et al.*, "ATMNet: Adaptive Two-Stage Modular Network for Accurate Video Captioning," *IEEE Trans. Syst., Man, Cybern.*, vol. 55, no. 4, pp. 2821–2833, 2025.
[14] Z. Zheng, G. An, S. Cao, D. Wu, Q. Ruan, "Collaborative and Multilevel Feature Selection Network for Action Recognition," *IEEE Trans. Neural Netw.*, vol. 34, no. 3, pp. 1304–1318, 2023.
[15] W. E. Vinje and J. L. Gallant, "Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.
[16] G. L. Malcolm, I. I. Grien and C. I. Baker, "Making Sense of Real-World Scenes," *Trends Cognit. Sci.*, vol. 20, no. 11, pp. 843–856, 2016.
[17] M. Ueda *et al.*, "An age estimation method using 3d-cnn from brain mri images," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 380–383.
[18] B. A. Jónsson *et al.*, "Brain age prediction using deep learning uncovers associated sequence variants," *Nat Commun*, vol. 10, no. 1, p. 5409, 2019.
[19] J. H. Cole, "Multimodality neuroimaging brain-age in uk biobank: relationship to biomedical, lifestyle, and cognitive factors," *Neurobiol. Aging*, vol. 92, pp. 34–42, 2020.

[20] V. M. Bashyam *et al.*, "Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide," *Brain*, vol. 143, no. 7, pp. 2312–2324, 2020.

[21] J. Cheng *et al.*, "Brain age estimation from mri using cascade networks with ranking loss," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3400–3412, 2021.

[22] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith *et al.*, "Accurate brain age prediction with lightweight deep neural networks," *Med. Image Anal.*, vol. 68, p. 101871, 2021.

[23] P. Mouches, M. Wilms, D. Rajashekar, S. Langner, and N. D. Forkert, "Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions," *Hum. Brain Mapping*, vol. 43, no. 8, pp. 2554–2566, 2022.

[24] L. Chen, and X. Luo, "Tensor Distribution Regression Based on the 3D Conventional Neural Networks," *IEEE/CAA J. Autom. Sin.*, vol. 10, no. 7, pp. 1628–1630, 2023.

[25] B. Billot *et al.*, "SE(3)-Equivariant and Noise-Invariant 3D Rigid Motion Tracking in Brain MRI," *IEEE Trans. Med. Imag.*, vol. 43, no. 11, pp. 4029–4040, 2024.

[26] C. Barbano, B. Dufumier, E. Duchesnay, M. Grangetto, and P. Gori, "Robust brain age estimation from structural MRI with contrastive learning," *Pattern Recognit. Lett.*, vol. 203, pp. 78–84, 2026.

[27] H. Kim *et al.*, "A novel deep learning-based brain age prediction framework for routine clinical MRI scans," *npj Aging*, vol. 11, no. 1, p. 70, 2025.

[28] R. Cheng *et al.*, "Robust Brain Tumor Segmentation With Incomplete MRI Modalities Using Hölder Divergence and Mutual Information-Enhanced Knowledge Transfer," *IEEE/CAA J. Autom. Sinica*, vol. 13, no. 4, pp. 939–954, 2026.

[29] P. Wu, H. Li, L. Hu, J. Ge, and N. Zeng, "A Local-Global Attention Fusion Framework With Tensor Decomposition for Medical Diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 6, pp. 1536–1538, 2024.

[30] S. He *et al.*, "Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy mris across lifespan," *Med. Image Anal.*, vol. 72, p. 102091, 2021.

[31] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 213–224, 2022.

[32] S. He, Y. Feng, P. E. Grant, and Y. Ou, "Deep relation learning for regression and its application to brain age estimation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2304–2317, 2022.

[33] H. Cai, Y. Gao, and M. Liu, "Graph transformer geometric learning of brain networks using multimodal mr images for brain age estimation," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 456–466, 2023.

[34] Y. Wu, Y. Zhou, W. Zeng, Q. Qian, and M. Song, "An Attention-Based 3D CNN With Multi-Scale Integration Block for Alzheimer's Disease Classification," *IEEE J. Biomed. Health. Inf.*, vol. 26, no. 11, pp. 5665–5673, 2022.

[35] Z. Pei *et al.*, "Multi-scale attention-based pseudo-3D convolution neural network for Alzheimer's disease diagnosis using structural MRI," *Pattern Recognit.*, vol. 131, p. 108825, 2022.

[36] Y. Zhang *et al.*, "Improving brain age prediction with anatomical feature attention-enhanced 3D-CNN," *Comput. Biol. Med.*, vol. 169, p. 107873, 2024.

[37] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1717–1724.

[38] X. Zhu, Y. Zhou, D. Wang, W. Ouyang, and R. Su, "Mlst-former: Multi-level spatial-temporal transformer for group activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3383–3397, 2022.

[39] T. Li, Y. Han, Y. Wang, and Z. Geng, "A Self-Attention Mechanism Integrating Adaptive Double Subspace for Fault Detection in Industrial Processes," *IEEE Trans. Syst., Man, Cybern.*, vol. 55, no. 1, pp. 540–549, 2025.

[40] P. Chen, Z. Gao, M. Yin, J. Wu, K. Ma, and C. Grebogi, "Multiattention Adaptation Network for Motor Imagery Recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 52, no. 8, pp. 5127–5139, 2022.

[41] X.-N. Zuo *et al.*, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Sci. Data*, vol. 1, no. 1, pp. 1–13, 2014.

[42] D. Wei *et al.*, "Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset," *Sci. Data*, vol. 5, no. 1, pp. 1–10, 2018.

[43] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *J. Cogn. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, 2007.

[44] S. Safaiyan *et al.*, "White matter aging drives microglial diversity," *Neuron*, vol. 109, no. 7, pp. 1100–1117, 2021.

[45] N. J. Tustison *et al.*, "N4itk: improved n3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, 2010.

[46] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

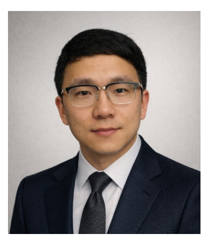
[47] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 6105–6114.

[48] J. Jang and D. Hwang, "M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20718–20729.

[49] H.-Y. Zhou, C. Lu, C. Chen, S. Yang, and Y. Yu, "A unified visual information preservation framework for self-supervised pre-training in medical image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8020–8035, 2023.

[50] L. Bellantuono *et al.*, "Predicting brain age with complex networks: From adolescence to adulthood," *NeuroImage*, vol. 225, p. 117458, 2021.

Yu Zhang received the Ph.D. degree in Science and Engineering from the University of Toyama, Toyama, Japan, in 2025. He is currently a post-doctoral researcher with the Department of Laboratory Medicine, The First Affiliated Hospital and Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, China. His current research interests include computer vision, medical image analysis, and interdisciplinary applications of AI in biomedical sciences.



Jiaru Yang received the Ph.D. degree in Science and Engineering from the University of Toyama, Toyama, Japan, in 2025. He is current a lecturer with the School of Artificial Intelligence and Computer Science, Nantong University, China. His current interests include optimization theory, computational intelligence and machine learning.

Masaaki Omura received the B.E., M.E., and Ph.D. degrees in engineering from Chiba University, Chiba, Japan, in 2015, 2017, and 2020, respectively. While he was a Ph.D. Student, he was supported by the Research Fellowship for Young Scientists (DC1) of the Japan Society for the Promotion of Science (JSPS). From 2020 to 2021, he was a Post-doctoral Researcher with the University of Toyama, Toyama, Japan, granted by the Research Fellowship for Young Scientists (PD) of the JSPS. Since 2024, he has been an Associate Professor at the University of Toyama. His research interests include ultrasonic tissue characterization and functional imaging in biological soft tissues. Dr. Omura is a member of the Acoustical Society of Japan and a fellow of the Japan Society of Ultrasound in Medicine.

Zhenyu Lei received the Ph.D. degree in Science and Engineering from the University of Toyama, Toyama, Japan, in 2023. He is currently an Assistant Professor with the Faculty of Engineering, University of Toyama, Japan. His current research interests include evolutionary computation, machine learning, and neural network for real-world applications and optimization problems.

Hideyuki Hasegawa received the B.E. degree from Tohoku University, Sendai, Japan, in 1996. He received the Ph.D. degree from Tohoku University in 2001. He received the research fellowships for young scientists (DC1) and (PD) from the Japan Society for the Promotion of Science in 1998 and 2001, respectively. He became a research associate in 2002, lecturer in 2005, and associate professor in 2007 at the Graduate School of Engineering, Tohoku University. From 2008, he became an associate professor at the Graduate School of Biomedical Engineering, Tohoku University. He has been a professor at the Graduate School of Science and Engineering for Research, University of Toyama from 2015. His research interest is medical ultrasonics. He serves as a member of the technical program committee of the IEEE International Ultrasonics Symposium and an associate editor for IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control.

Shangce Gao received his Ph.D. degree in Innovative Life Science from University of Toyama, Toyama, Japan in 2011. He is currently a Professor with the Faculty of Engineering, University of Toyama, Japan. His current research interests include nature-inspired technologies, machine learning, and neural networks for real-world applications. He serves as an Associate Editor for many international journals such as IEEE Transactions on Neural Networks and Learning Systems, and IEEE/CAA Journal of Automatica Sinica.

Supplementary File for "An Attention-based Mid-level Feature Representation Incorporated 3D Residual Neural Network for Brain Age Estimation"

Yu Zhang, Jiaru Yang, Masaaki Omura, Zhenyu Lei,
Hideyuki Hasegawa, and Shangce Gao, *Senior Member, IEEE*

I. Preprocessing

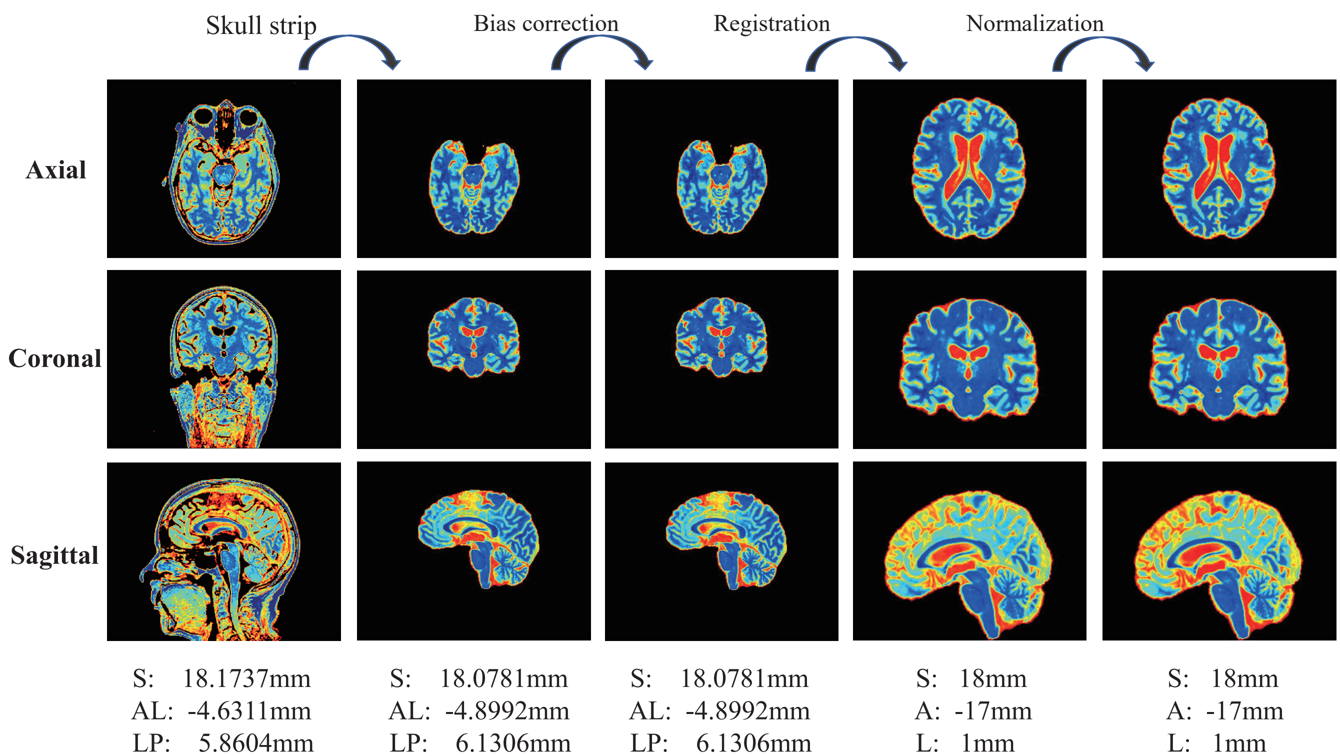


Fig. S.I. The illustration of preprocessing method for an example, dataset: ADNI1, ID: I109517, age: 76. (S: Superior, A: Anterior, L: Left, AL: Anterior-Left, LP: Left-Posterior.)

In order to achieve optimal results in brain age estimation tests across various public datasets, it is essential to implement effective image preprocessing techniques. This section uses an example from the ADNI dataset to demonstrate the role of each preprocessing step. To better illustrate the effect of each preprocessing method, we use the slices in three directions for observation. As shown in Fig. S.I, Skull stripping involves using a trained segmentation model UNet to separate the entire brain from the head, thereby reducing the impact of unnecessary regions on age estimation. Another easily observable process is template registration, which uses a template file to standardize various MRI images into a uniform format. The remaining steps, N4 bias field correction and Voxel normalization, are difficult to distinguish directly in the image, yet they play crucial roles in data preprocessing. N4 bias field correction can repair distortions in MRI images caused by machine imperfections,

This research was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grants JP25K21298, JP25K24386, JP26K03333, JP26K23843, and JP25K03179, and Japan Science and Technology Agency (JST) Support for Pioneering Research Initiated by the Next Generation (SPRING) under Grant JPMJSP2145. (Corresponding Author: Shangce Gao)

Yu Zhang is with the Department of Laboratory Medicine of The First Affiliated Hospital & Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, 310000, China. (E-mail: 0625716@zju.edu.cn)

Jiaru Yang is with the School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China. (E-mail: yangjr@ntu.edu.cn)

Masaaki Omura, Zhenyu Lei, Hideyuki Hasegawa, and Shangce Gao are with the Faculty of Engineering, University of Toyama, Toyama-shi, 930-8555, Japan. (E-mail: momura@eng.u-toyama.ac.jp; leizg@eng.u-toyama.ac.jp; hasegawa@eng.u-toyama.ac.jp; gaosc@eng.u-toyama.ac.jp)

which can lead to decreased prediction accuracy due to varying pixel intensities in white or gray matter. In this work, we employ a 3-dimensional B-spline transformation formula to register all the MRI images. Assuming that the coordinates of a 3D image are (x,y,z), the (x',y',z') after B-spline transformation can be formulated as:

$$\begin{aligned} x' &= \sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l} N_{i,p}(x)\cdot N_{j,q}(y)\cdot N_{k,r}(z)\cdot P_{i,j,k}^{x} \\ y' &= \sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l} N_{i,p}(x)\cdot N_{j,q}(y)\cdot N_{k,r}(z)\cdot P_{i,j,k}^{y} \\ z' &= \sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l} N_{i,p}(x)\cdot N_{j,q}(y)\cdot N_{k,r}(z)\cdot P_{i,j,k}^{z} \end{aligned} \tag{1}$$

where $P_{i,j,k}^{*}$ represents the coordinates in a direction, $N(*)$ is the basic function of B-spline, which can be expressed as:

$$\begin{aligned} N_{i,0}(x) &= \begin{cases} 1, if x_i \le x < x_{i+1} \\ 0, otherwise \end{cases} \\ N_{i,p}(x) &= \frac{x - x_i}{x_{i+p} - x_i}\cdot N_{i,p-1}(x) + \frac{x_{i+p+1} - x}{x_{i+p+1} - x_{i+1}}\cdot N_{i+1,p-1}(x) \end{aligned} \tag{2}$$

and

$$\begin{aligned} N_{j,0}(y) &= \begin{cases} 1, \ if y_j \le y < y_{j+1} \\ 0, \ otherwise \end{cases} \\ N_{j,q}(y) &= \frac{y - y_j}{y_{j+q} - y_j}\cdot N_{j,q-1}(y) + \frac{y_{j+q+1} - y}{y_{j+q+1} - y_{j+1}}\cdot N_{j+1,q-1}(y) \end{aligned} \tag{3}$$

and

$$\begin{aligned} N_{k,0}(z) &= \begin{cases} 1, \ if z_k \le z < z_{k+1} \\ 0, \ otherwise \end{cases} \\ N_{k,r}(z) &= \frac{z - z_k}{z_{k+r} - z_k}\cdot N_{k,r-1}(z) + \frac{z_{k+r+1} - z}{z_{k+r+1} - z_{k+1}}\cdot N_{k+1,r-1}(z) \end{aligned} \tag{4}$$

Finally, we use a Gaussian Mixture Model based approach to perform voxel normalization. For given a voxel value $V_{x,y,z}$, the normalized process can be represented as:

$$G(x,y,z) = \frac{cV_{x,y,z}}{\mu} \tag{5}$$

where c is a constant and μ is the mean of the Gaussian distribution with the largest mean.

II. Supplementary Experimental results

This section primarily supplements the experimental results not mentioned in the main text. It includes the performance metrics and scatter plots of our proposed MAR3D-34, as well as the competitor models 3DCNN-AFAC, EfficientNet-b4, CL-3DResNet, SFCN, M3T, and PCRLV2 on six datasets.

A. MAR3D-34

This sub-section lists the six performance metrics for MAR3D-34 on each dataset. From Table S.I, it can be seen that it has an excellent performance on the ADNI dataset, with an NAE of less than 1 year and a CS of over 99%, proving that our proposed MAR3D-34 is effective in handling anomalous datasets (ADNI includes individuals with three different AD statuses). Additionally, it also performs well on the other five datasets.

TABLE S.I
Detailed experimental results of MAR3D-34 on six datasets.

DataSet	MAR3D-34					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	0.878	1.116	0.986	0.984	0.908	99.146%
CoRR	1.179	2.038	0.991	0.962	0.870	94.719%
IXI	2.005	2.905	0.987	0.986	0.907	81.705%
OASIS-1	2.444	3.911	0.988	0.977	0.890	75.721%
SALD	1.766	2.376	0.992	0.982	0.903	83.603%
DLBS	1.636	2.349	0.994	0.993	0.934	85.397%

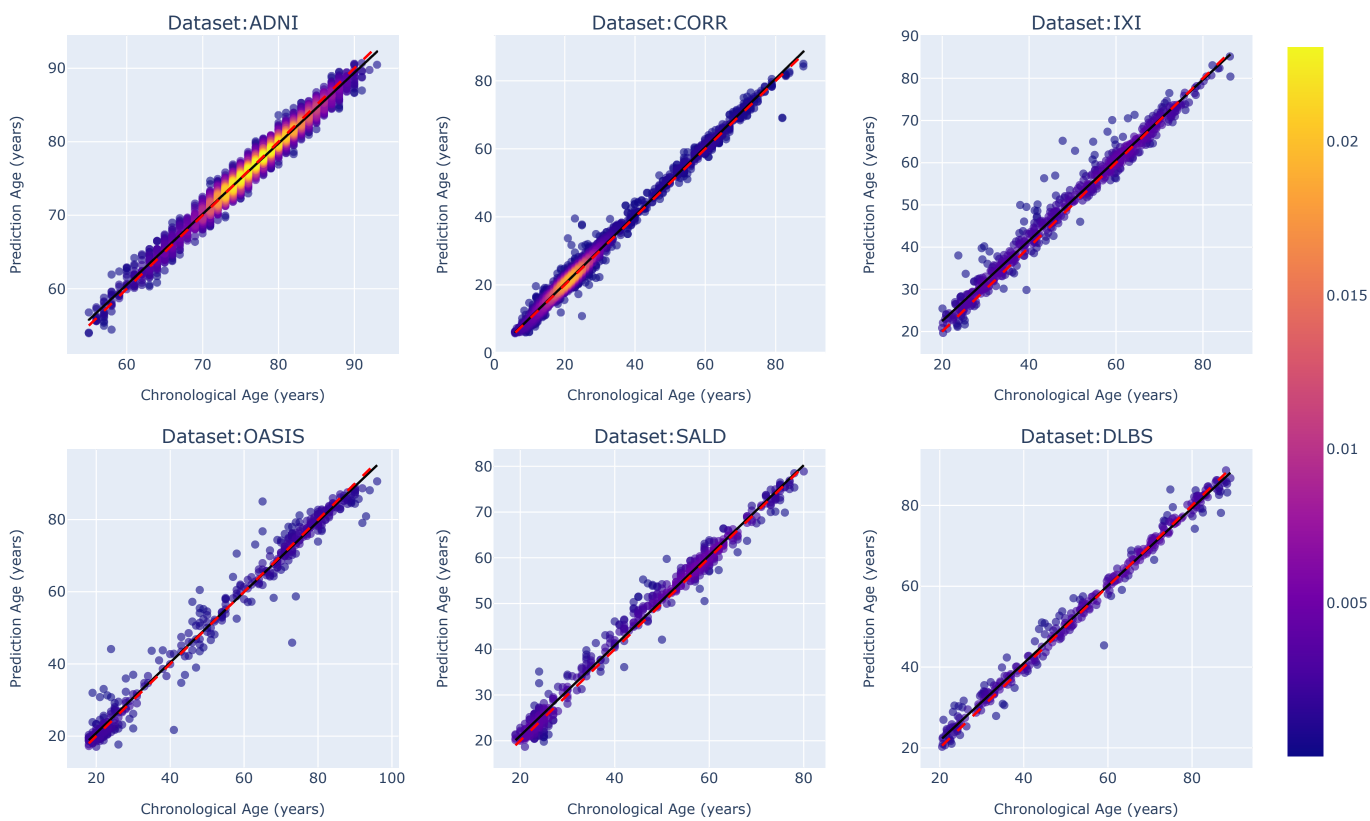


Fig. S.II. The scatter diagram of MAR3D-34 on six datasets.

B. 3DCNN-AFAC

This sub-section lists the performance metrics of 3DCNN-AFAC. From Table S.II, we can clearly see that it performs well on most metrics. However, when handling the CoRR dataset, the RMSE reaches 25.483, indicating the presence of some outliers. Fig. S.III shows that while estimating two brains with chronological ages around 25 years, 3DCNN-AFAC predicted two outlier values over than 100 years, demonstrating the poor robustness of model.

TABLE S.II
DETAILED EXPERIMENTAL RESULTS OF 3DCNN-AFAC ON SIX DATASETS.

DataSet	3DCNN-AFAC					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	2.935	3.717	0.852	0.815	0.640	59.138%
CoRR	3.071	25.483	0.481	0.881	0.731	70.781%
IXI	3.874	5.000	0.958	0.956	0.812	47.247%
OASIS-1	3.847	5.011	0.981	0.949	0.814	48.558%
SALD	3.434	4.566	0.970	0.947	0.806	56.275%
DLBS	4.006	5.121	0.969	0.970	0.846	46.667%

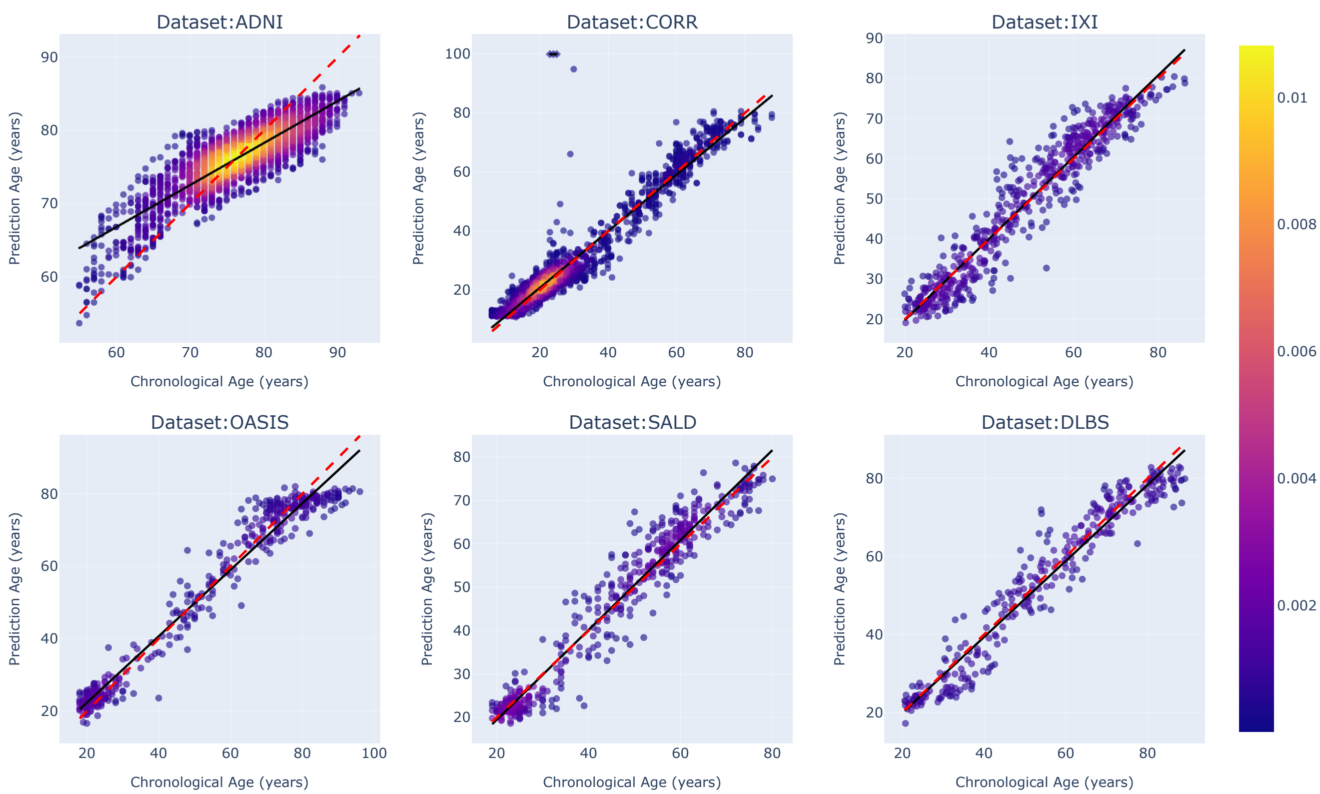


Fig. S.III. The scatter diagram of 3DCNN-AFAC on six datasets.

C. *EfficientNet-b4*

This sub-section provides the experimental data for EfficientNet-b4. From Table S.III, it is obvious that its overall evaluation results are relatively stable, but the errors are significant. According to Fig. S.IV, the prediction performance on the ADNI and OASIS datasets is poor, with the fit line showing low similarity to the ideal estimation.

TABLE S.III
DETAILED EXPERIMENTAL RESULTS OF EFFICIENTNET-B4 ON SIX DATASETS.

DataSet	EfficientNet-b4					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.393	1.781	0.968	0.967	0.860	91.085%
CoRR	1.409	2.217	0.989	0.957	0.853	91.031%
IXI	2.173	2.971	0.984	0.985	0.899	76.377%
OASIS-1	3.413	5.137	0.982	0.969	0.872	59.135%
SALD	2.196	2.965	0.988	0.977	0.884	76.721%
DLBS	1.877	3.046	0.989	0.988	0.925	83.492%

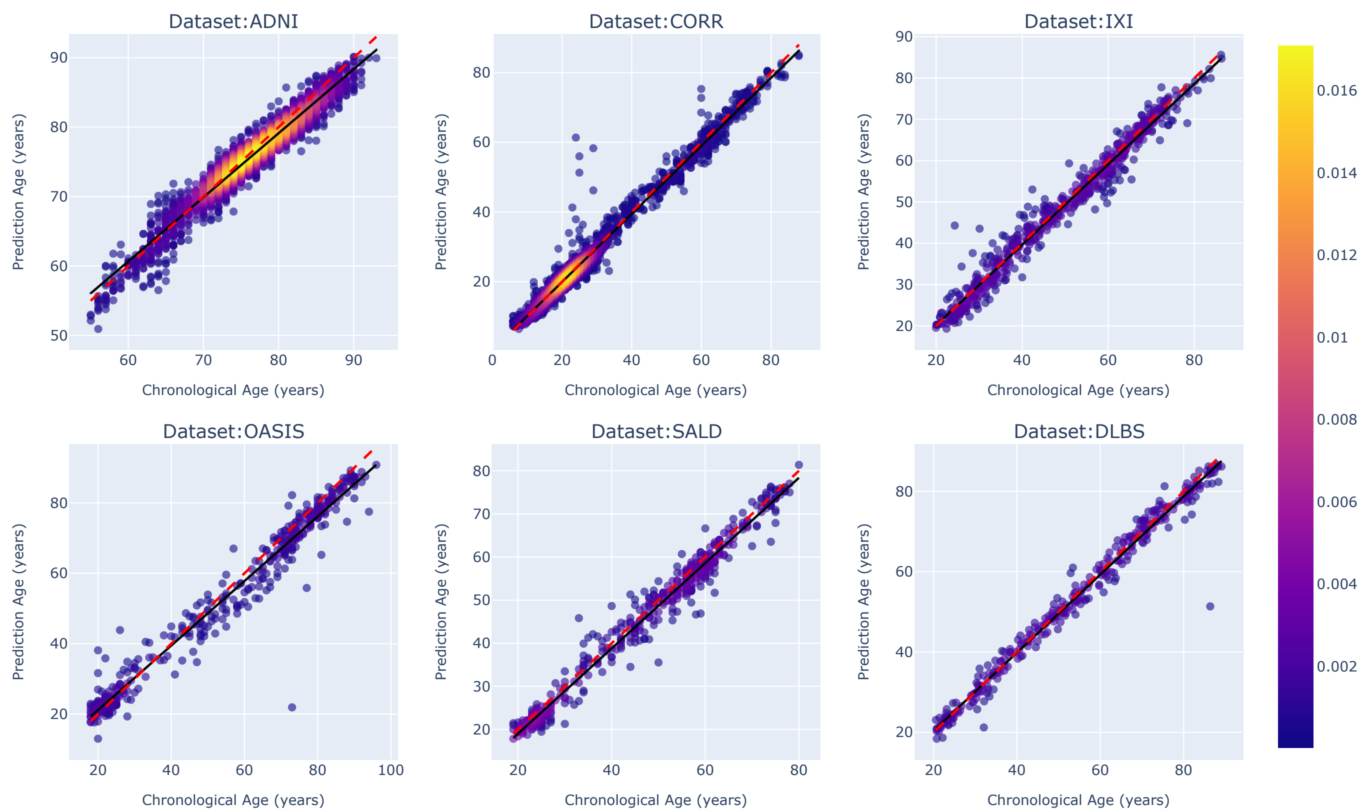


Fig. S.IV. The scatter diagram of EfficientNet-b4 on six datasets.

D. CL-3DResNet

This sub-section shows the performance metrics of CL-3DResNet. From Table S.IV, we can clearly see that it performs well on most metrics. However, when handling the CoRR dataset, the RMSE reaches 9.843, indicating the presence of some outliers. Fig. S.V shows that while estimating two brains with chronological ages around 25 years, CL-3DResNet predicted two outlier values over than 100 years, demonstrating the poor robustness of model.

TABLE S.IV
DETAILED EXPERIMENTAL RESULTS OF CL-3DRESNET ON SIX DATASETS.

DataSet	CL-3DResNet					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.371	1.688	0.982	0.978	0.892	93.351%
CoRR	1.436	9.843	0.828	0.955	0.858	94.719%
IXI	1.866	2.511	0.989	0.989	0.909	80.995%
OASIS-1	2.325	3.744	0.990	0.982	0.904	78.125%
SALD	1.857	2.729	0.989	0.979	0.893	83.198%
DLBS	1.968	2.979	0.989	0.989	0.918	81.905%

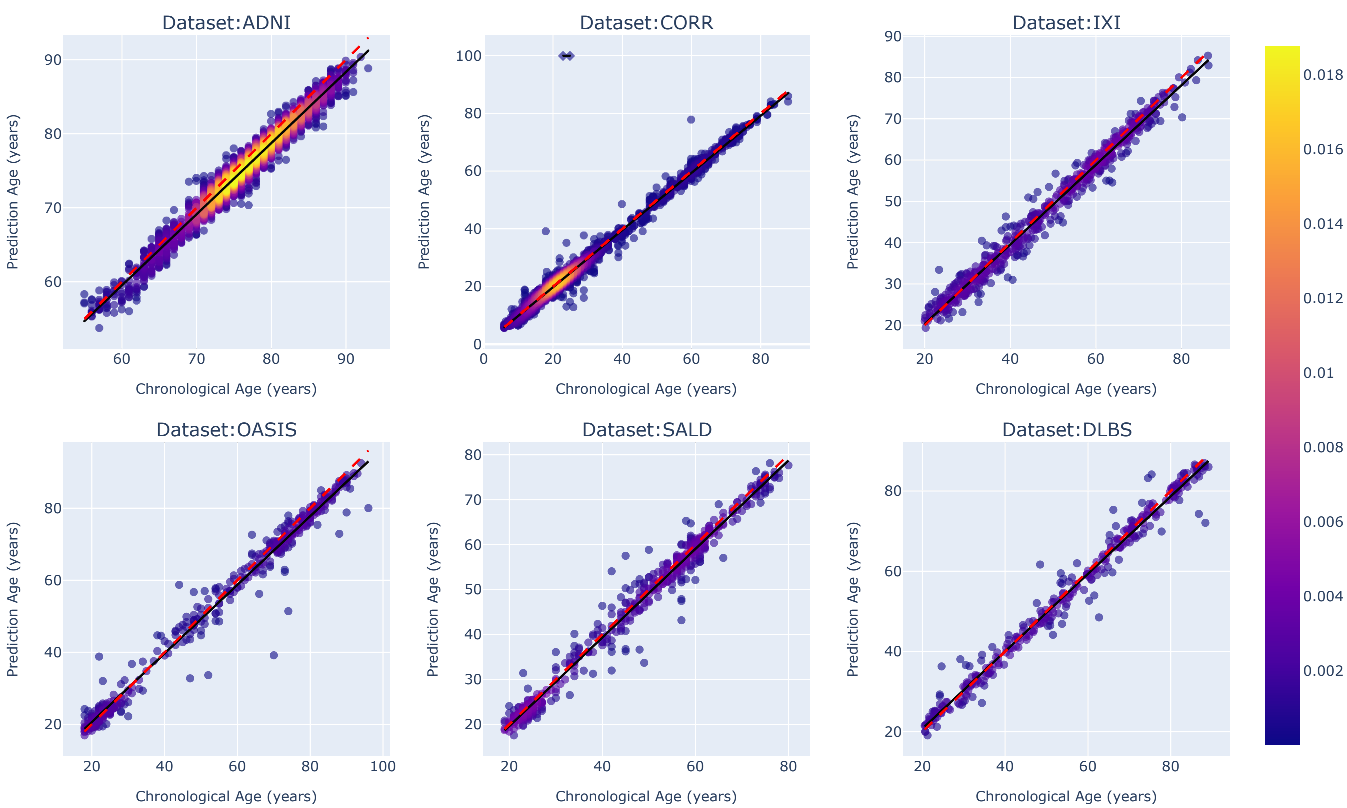


Fig. S.V. The scatter diagram of CL-3DResNet on six datasets.

E. SFCN

This sub-section shows the performance metrics of SFCN. SFCN is a classical 3d cnn-based network model for MRI recognitipon. However, it can been find that its RMSE on the CoRR dataset is as high as 7.018 in Table S.V. Similar to the other model, SFCN also obtains two outlier values greater than 100 when estimating brains with chronological ages around 25 years in CoRR dataset.

TABLE S.V
DETAILED EXPERIMENTAL RESULTS OF SFCN ON SIX DATASETS.

DataSet	SFCN					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.886	2.389	0.945	0.936	0.797	79.903%
CoRR	2.751	7.018	0.894	0.897	0.758	69.469%
IXI	2.696	3.405	0.979	0.979	0.869	62.522%
OASIS-1	2.737	3.631	0.990	0.971	0.868	65.144%
SALD	2.387	3.094	0.985	0.968	0.859	69.636%
DLBS	2.597	3.327	0.987	0.987	0.899	66.349%

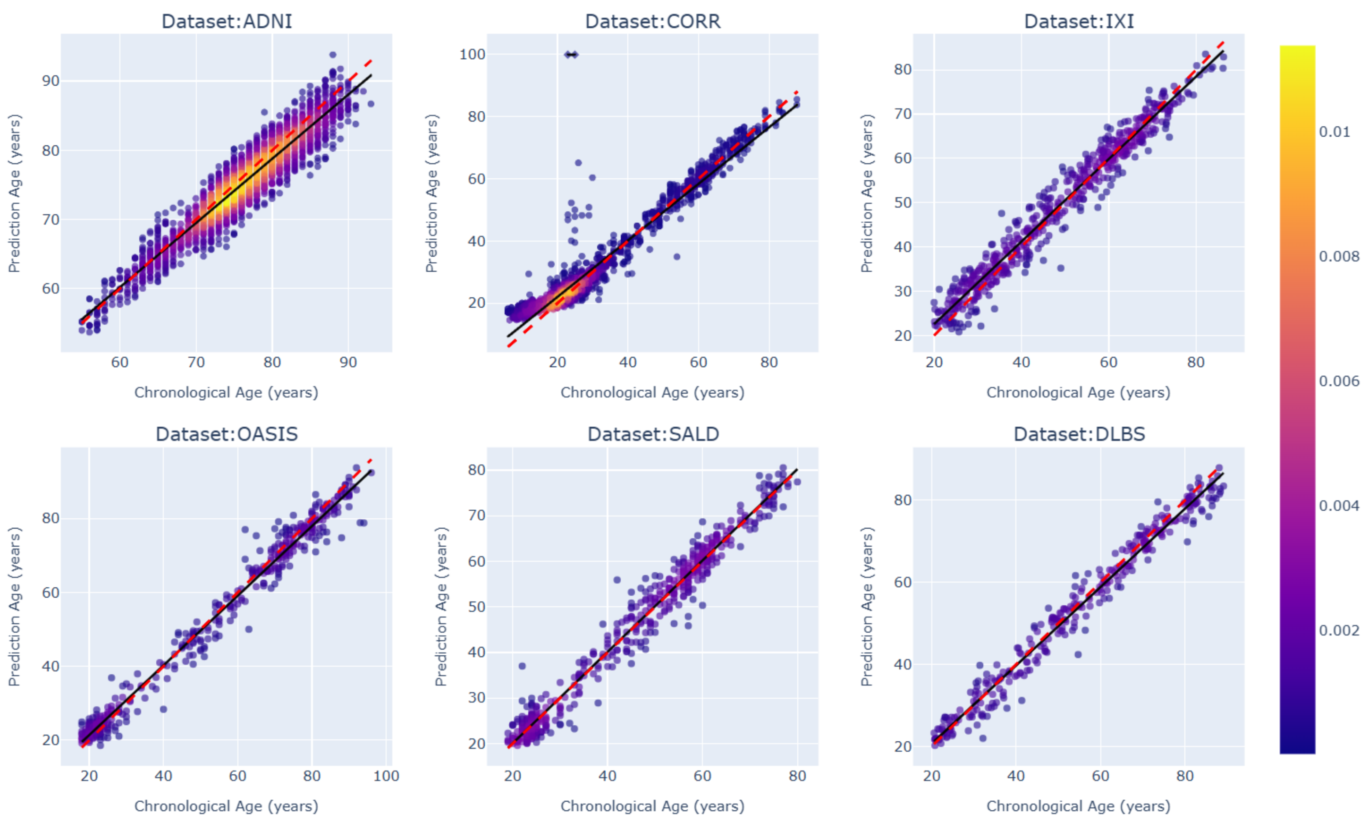


Fig. S.VI. The scatter diagram of SFCN on six datasets.

F. M3T

The aforementioned competitor models are all classical and effective convolution network, whereas M3T is a 3D recognition model based on vision transformers. An analysis of the results presented in Table S.VI and Fig. S.VII reveals that M3T exhibits high stability in performance, but its recognition accuracy is low-level. his is attributed to the inherent limitations of the sequence-based computation method employed by transformers in processing large-scale 3D images.

TABLE S.VI
DETAILED EXPERIMENTAL RESULTS OF M3T ON SIX DATASETS.

DataSet	M3T					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	2.551	3.191	0.904	0.888	0.727	64.822%
CoRR	2.439	3.596	0.970	0.862	0.713	71.250%
IXI	3.946	4.987	0.958	0.954	0.809	46.714%
OASIS-1	4.833	6.517	0.970	0.942	0.792	43.510%
SALD	3.781	4.975	0.967	0.947	0.804	51.012%
DLBS	4.235	5.443	0.967	0.964	0.832	44.762%

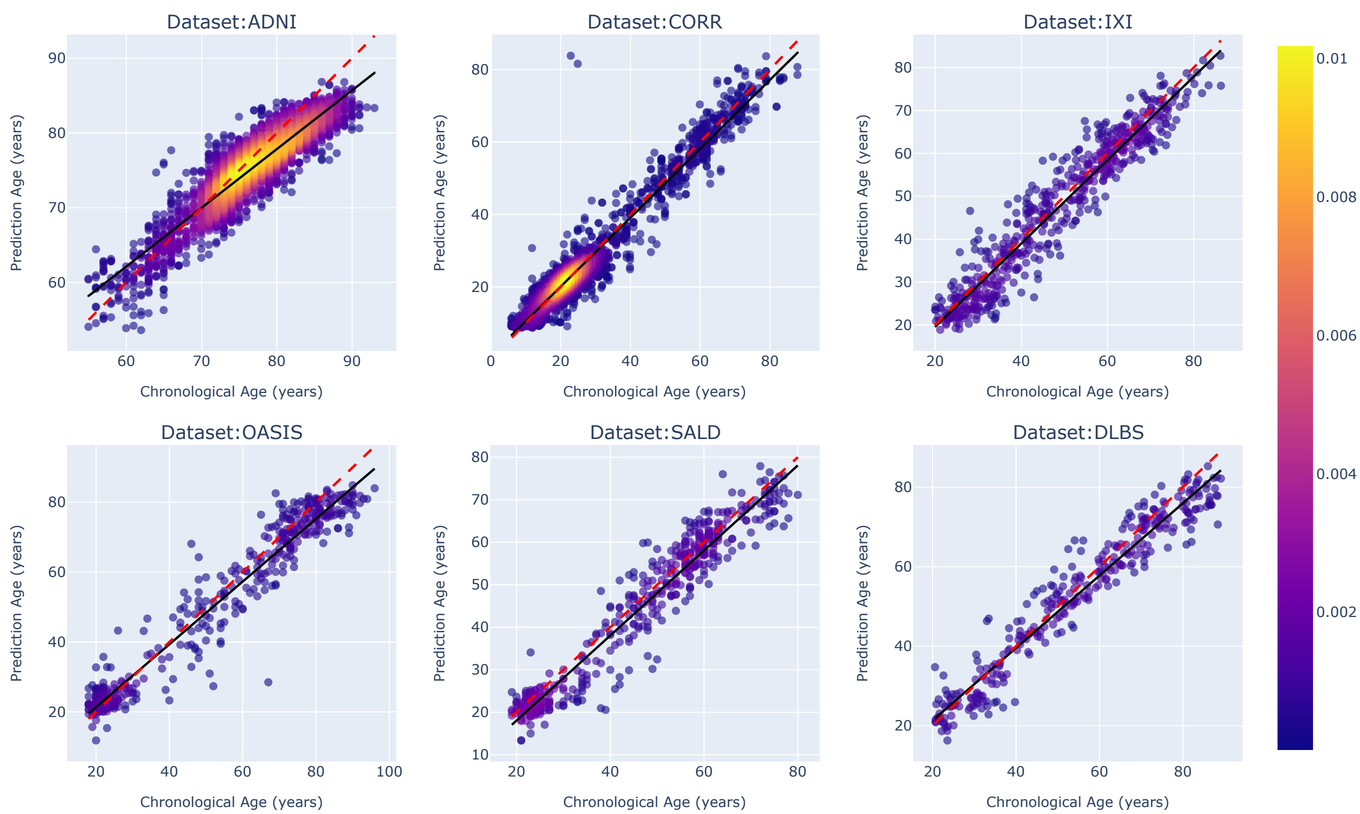


Fig. S.VII. The scatter diagram of M3T on six datasets.

G. PCRLV2

Finally, the most recent and advanced neural network model, named PCRLV2, demonstrates outstanding performance across all datasets enumerated in Table S.VII, exhibiting marginal superiority over our proposed MAR3D-34 on two datasets. However, as illustrated in Fig. S.VIII, the PCRLV2 model incorrectly predicts the chronological ages of 22 and 23 on two MRI images from the CoRR dataset as less than 0, which is an unacceptable result for a prediction model.

TABLE S.VII
DETAILED EXPERIMENTAL RESULTS OF PCRLV2 ON SIX DATASETS.

DataSet	PCRLV2					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.292	1.559	0.986	0.984	0.909	95.877%
CoRR	1.658	2.980	0.984	0.976	0.903	91.906%
IXI	1.933	2.672	0.990	0.989	0.921	84.014%
OASIS-1	2.155	3.283	0.992	0.980	0.899	81.971%
SALD	1.934	2.601	0.992	0.985	0.917	83.806%
DLBS	1.766	2.792	0.993	0.992	0.936	86.032%

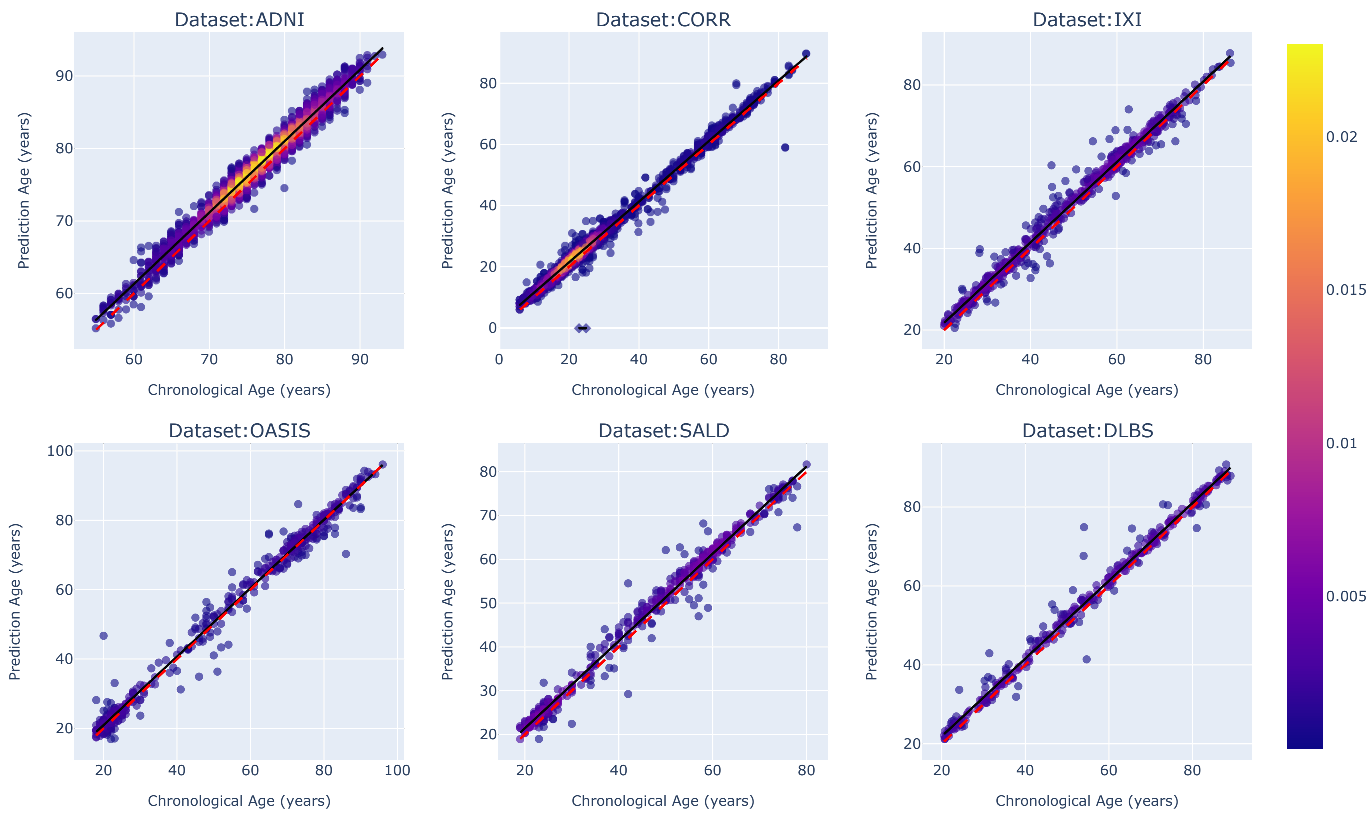


Fig. S.VIII. The scatter diagram of PCRLV2 on six datasets.

III. Supplementary Ablation Experiments

This section mainly presents the performance of the original 3D ResNet and MAR3D models, including the 10, 18, 34, 50, and 101-layer structures. Tables S.VIII - S.XII list the performance differences between the two models at different layers, Figs. S.IX - S.XIII display the scatter plots of the models for each layer across all datasets, and Figs. S.XIV - S.XXII shows the scatter plots of each model architecture on six datasets. From these tables, we can find that MAR3D exhibits superior prediction accuracy at every layer compared to the original ResNet model, demonstrating that the AM Block is an effective feature extraction module that enhances prediction accuracy in any scenario. From the scatter plot comparisons in Figs. S.IX - S.XIII, we can observe that the fit line of MAR3D is better than that of ResNet at the same number of layers, and MAR3D has significantly fewer outliers. This proves the effectiveness and robustness of our proposed model.

TABLE S.VIII
Detailed experimental results of ResNet10 and MAR3D-10 on six datasets.

DataSet	Without AM Block (ResNet10)						With AM Block (MAR3D-10)					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.466	1.867	0.974	0.969	0.868	89.450%	0.854	1.099	0.988	0.985	0.914	98.848%
CoRR	2.007	13.505	0.711	0.936	0.819	83.969%	2.765	7.716	0.240	0.969	0.876	93.250%
IXI	2.643	3.377	0.983	0.980	0.878	64.298%	1.624	2.317	0.991	0.990	0.920	87.389%
OASIS-1	3.012	4.111	0.988	0.974	0.873	60.817%	2.890	4.076	0.990	0.978	0.891	65.385%
SALD	2.502	3.327	0.986	0.971	0.867	68.016%	1.634	2.405	0.991	0.980	0.899	86.032%
DLBS	2.336	3.001	0.990	0.989	0.910	70.794%	1.818	2.982	0.990	0.990	0.928	85.714%

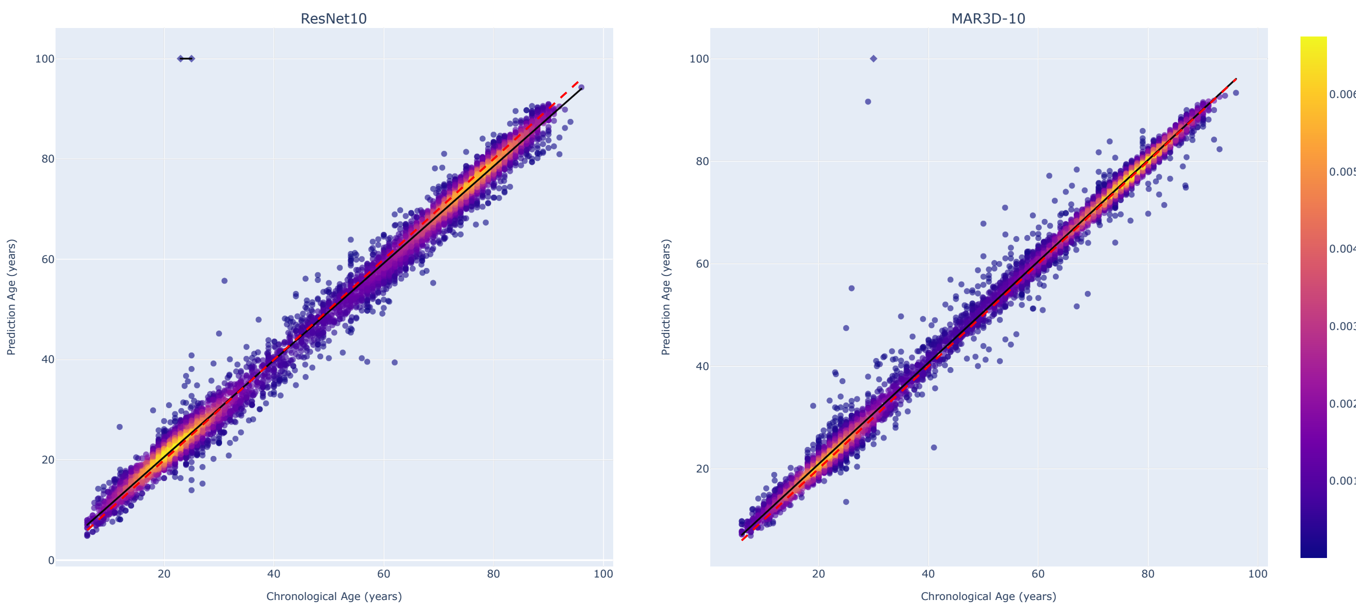


Fig. S.IX. The scatter diagram of 10-layer ResNet and MAR3D on the whole datasets.

TABLE S.IX
Detailed experimental results of ResNet18 and MAR3D-18 on six datasets.

DataSet	Without AM Block (ResNet18)						With AM Block (MAR3D-18)					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.099	1.390	0.985	0.981	0.902	96.322%	1.048	1.353	0.980	0.977	0.888	96.471%
CoRR	2.361	41.994	0.319	0.955	0.852	92.906%	1.261	1.845	0.992	0.954	0.856	94.188%
IXI	1.966	2.786	0.986	0.985	0.899	81.705%	2.037	2.846	0.985	0.984	0.896	81.350%
OASIS-1	2.178	3.530	0.991	0.983	0.904	80.048%	2.900	3.942	0.990	0.979	0.889	62.500%
SALD	1.863	2.525	0.990	0.978	0.891	81.377%	2.048	3.032	0.985	0.973	0.875	80.769%
DLBS	2.180	3.248	0.988	0.987	0.910	79.365%	2.212	3.603	0.984	0.983	0.903	80.000%

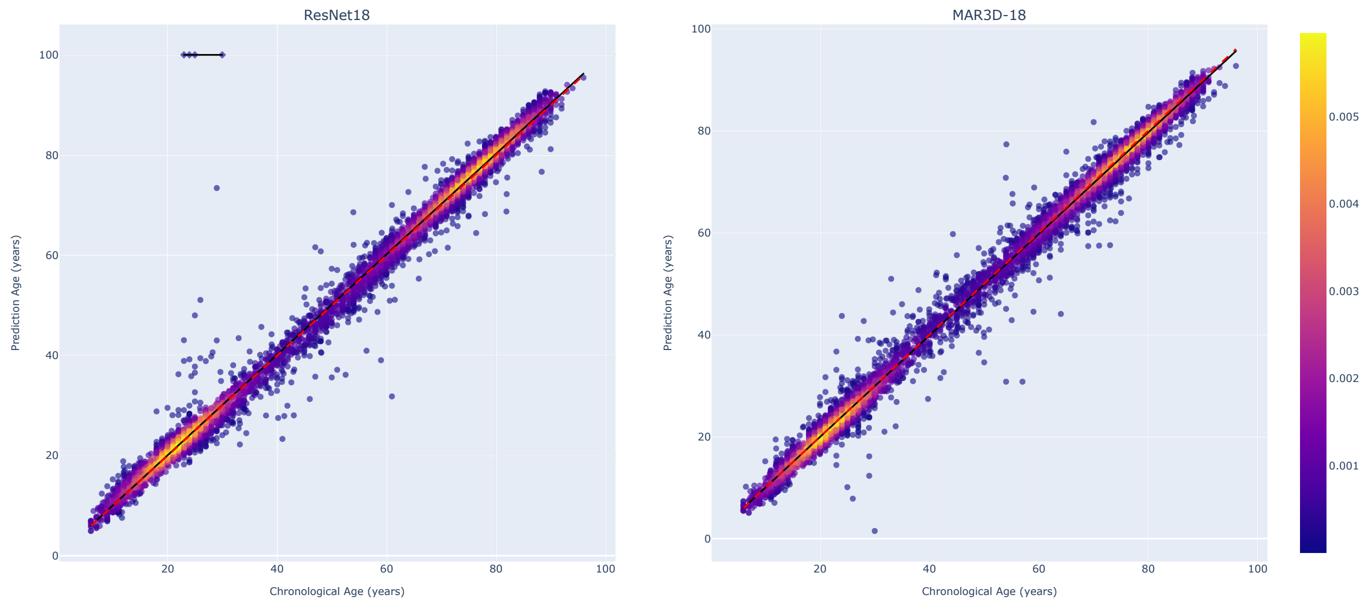


Fig. S.X. The scatter diagram of 18-layer ResNet and MAR3D on the whole datasets.

TABLE S.X

DETAILED EXPERIMENTAL RESULTS OF RESNET34 AND MAR3D-34 ON SIX DATASETS.

DataSet	Without AM Block (ResNet34)						With AM Block (MAR3D-34)					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.291	1.632	0.973	0.969	0.867	93.945%	0.878	1.116	0.986	0.984	0.908	99.146%
CoRR	2.495	40.016	0.351	0.936	0.821	88.687%	1.179	2.038	0.991	0.962	0.870	94.719%
IXI	2.312	2.962	0.985	0.983	0.885	70.160%	2.005	2.905	0.987	0.986	0.907	81.705%
OASIS-1	2.738	3.669	0.991	0.972	0.871	63.942%	2.444	3.911	0.988	0.977	0.890	75.721%
SALD	2.262	3.033	0.986	0.970	0.864	72.065%	1.766	2.376	0.992	0.982	0.903	83.603%
DLBS	2.256	3.081	0.988	0.988	0.909	74.921%	1.636	2.349	0.994	0.993	0.934	85.397%

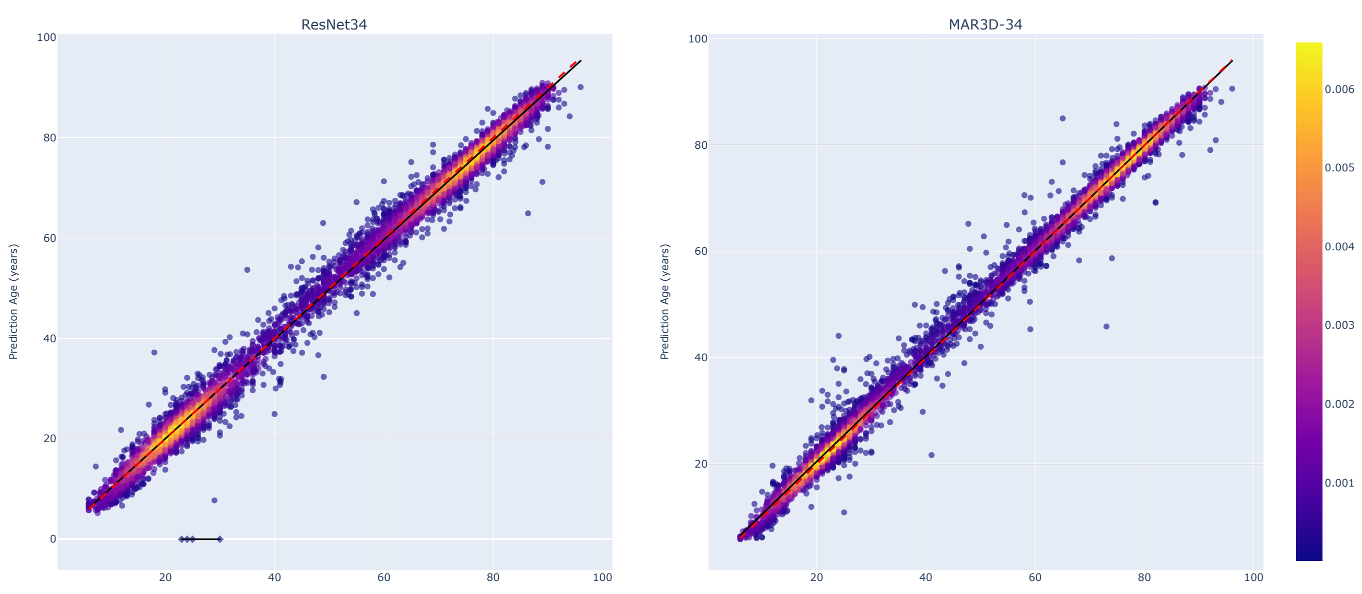


Fig. S.XI. The scatter diagram of 34-layer ResNet and MAR3D on the whole datasets.

TABLE S.XI
DETAILED EXPERIMENTAL RESULTS OF RESNET50 AND MAR3D-50 ON SIX DATASETS.

DataSet	Without AM Block (ResNet101)						With AM Block (MAR3D-101)					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.234	1.568	0.976	0.971	0.873	94.539%	1.056	1.333	0.983	0.980	0.896	97.697%
CoRR	2.816	48.042	0.282	0.928	0.813	87.906%	1.236	1.871	0.992	0.964	0.868	94.312%
IXI	2.428	3.290	0.982	0.980	0.884	69.627%	1.932	2.739	0.986	0.986	0.903	80.995%
OASIS-1	2.410	3.790	0.989	0.977	0.889	73.077%	2.281	3.519	0.990	0.982	0.900	77.404%
SALD	1.915	2.576	0.989	0.976	0.883	78.745%	1.735	2.534	0.990	0.980	0.896	84.008%
DLBS	2.405	3.371	0.987	0.986	0.906	72.381%	1.748	2.527	0.992	0.992	0.925	85.397%

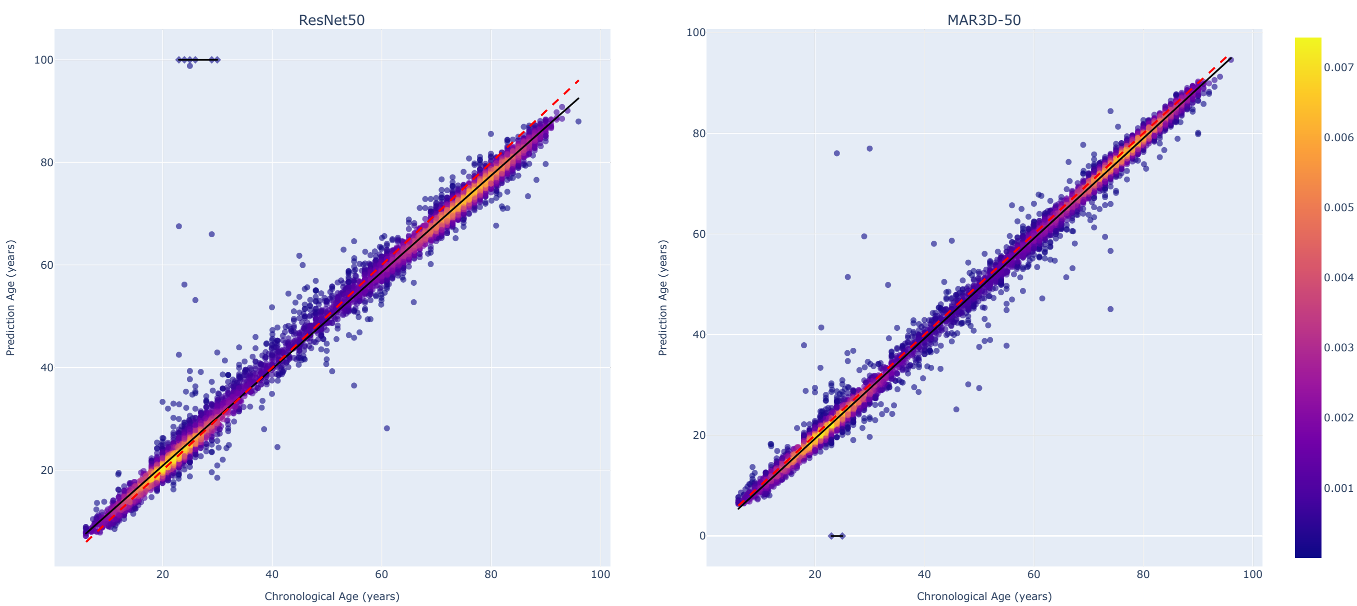


Fig. S.XII. The scatter diagram of 50-layer ResNet and MAR3D on the whole datasets.

TABLE S.XII
DETAILED EXPERIMENTAL RESULTS OF RESNET101 AND MAR3D-101 ON SIX DATASETS.

DataSet	Without AM Block (ResNet101)						With AM Block (MAR3D-101)					
	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑	MAE⇓	RMSE⇓	PCC⇑	R_s⇑	τ_B⇑	CS(α=3)⇑
ADNI	1.234	1.568	0.976	0.971	0.873	94.539%	1.056	1.333	0.983	0.980	0.896	97.697%
CoRR	2.816	48.042	0.282	0.928	0.813	87.906%	1.236	1.871	0.992	0.964	0.868	94.312%
IXI	2.428	3.290	0.982	0.980	0.884	69.627%	1.932	2.739	0.986	0.986	0.903	80.995%
OASIS-1	2.410	3.790	0.989	0.977	0.889	73.077%	2.281	3.519	0.990	0.982	0.900	77.404%
SALD	1.915	2.576	0.989	0.976	0.883	78.745%	1.735	2.534	0.990	0.980	0.896	84.008%
DLBS	2.405	3.371	0.987	0.986	0.906	72.381%	1.748	2.527	0.992	0.992	0.925	85.397%

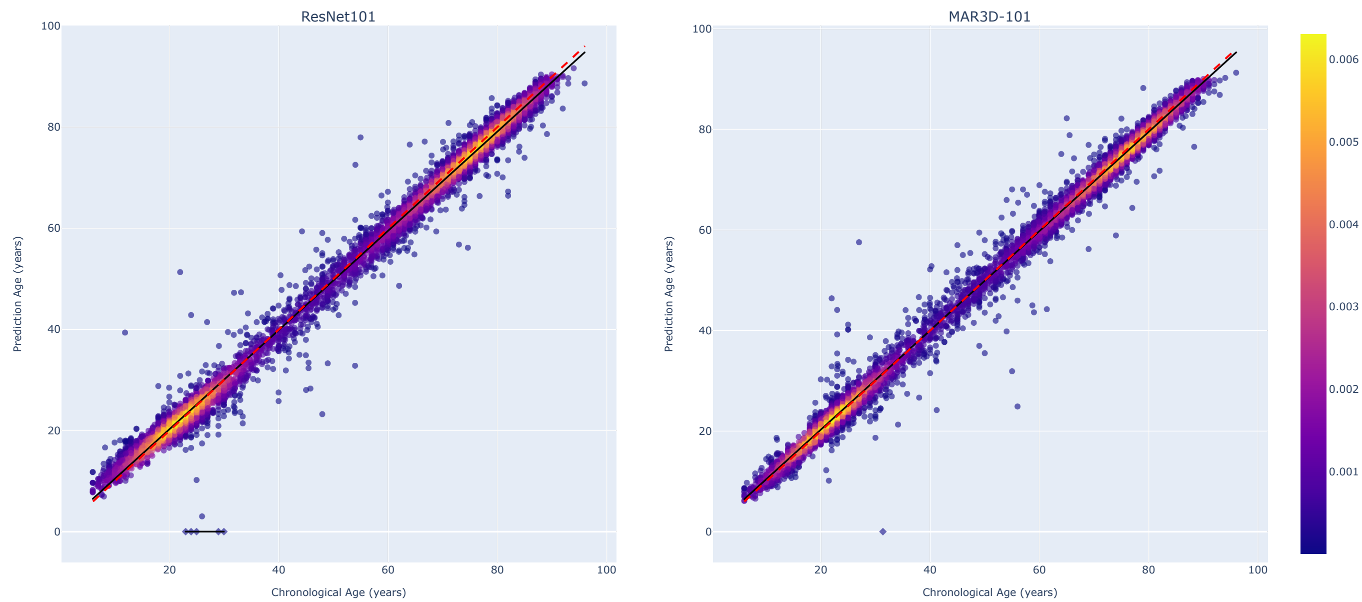


Fig. S.XIII. The scatter diagram of 101-layer ResNet and MAR3D on the whole datasets.

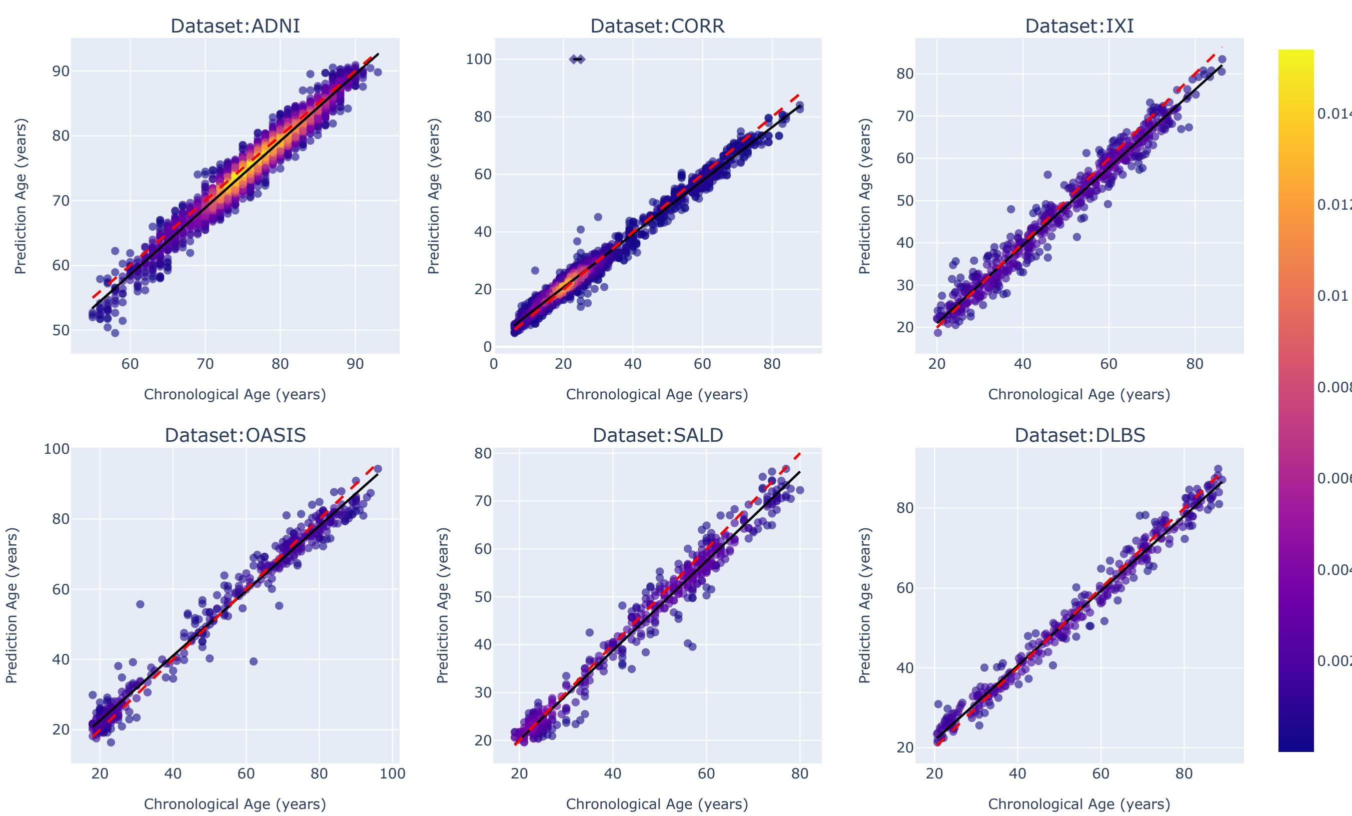


Fig. S.XIV. The scatter diagram of ResNet10 on six datasets.

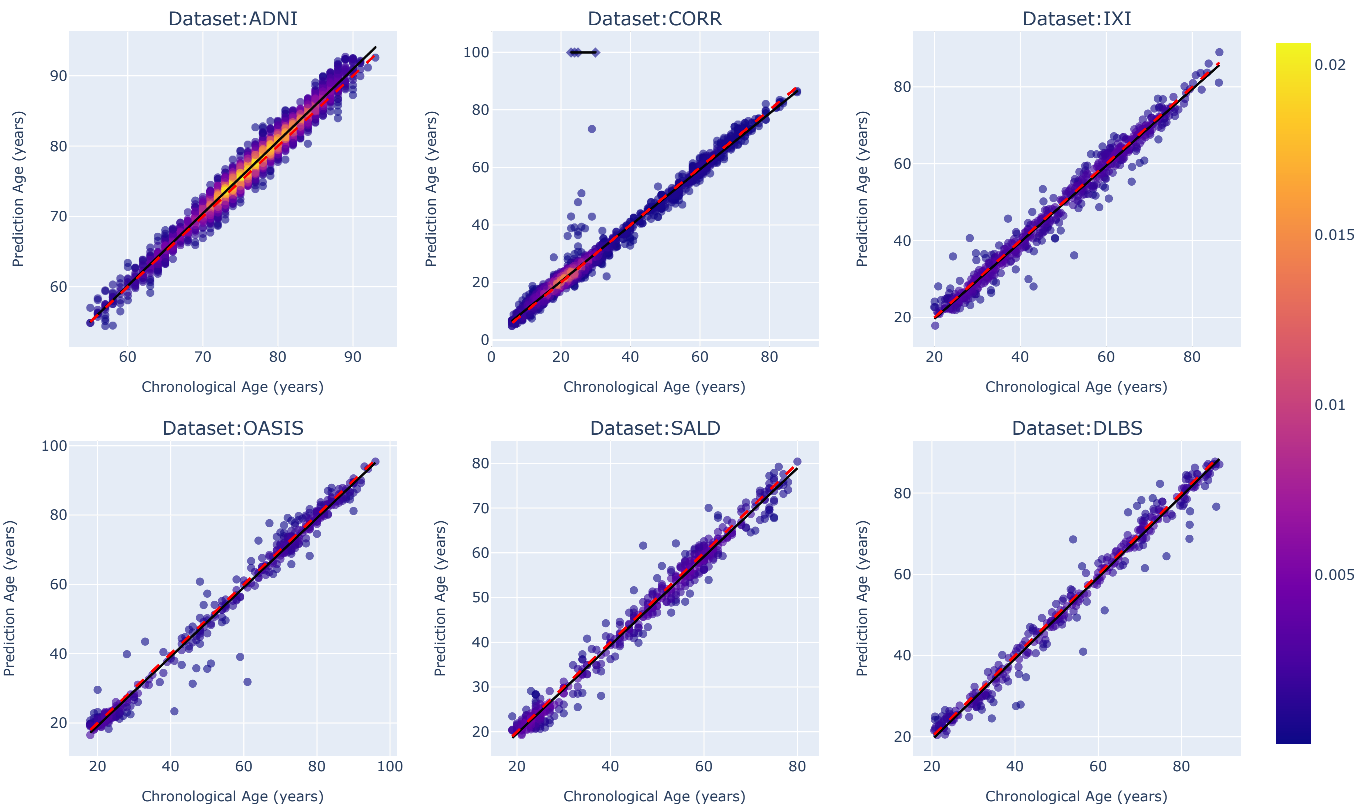


Fig. S.XV. The scatter diagram of ResNet18 on six datasets.

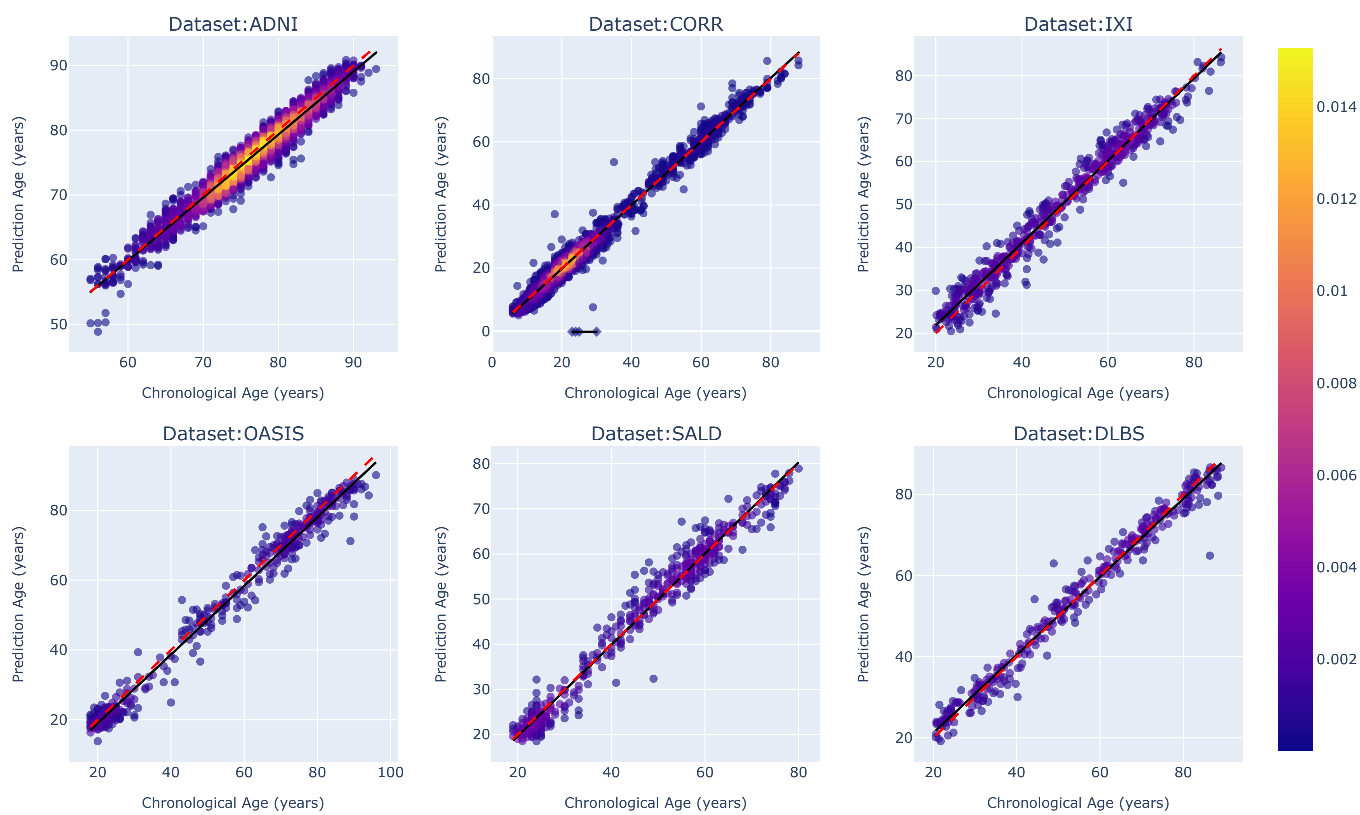


Fig. S.XVI. The scatter diagram of ResNet34 on six datasets.

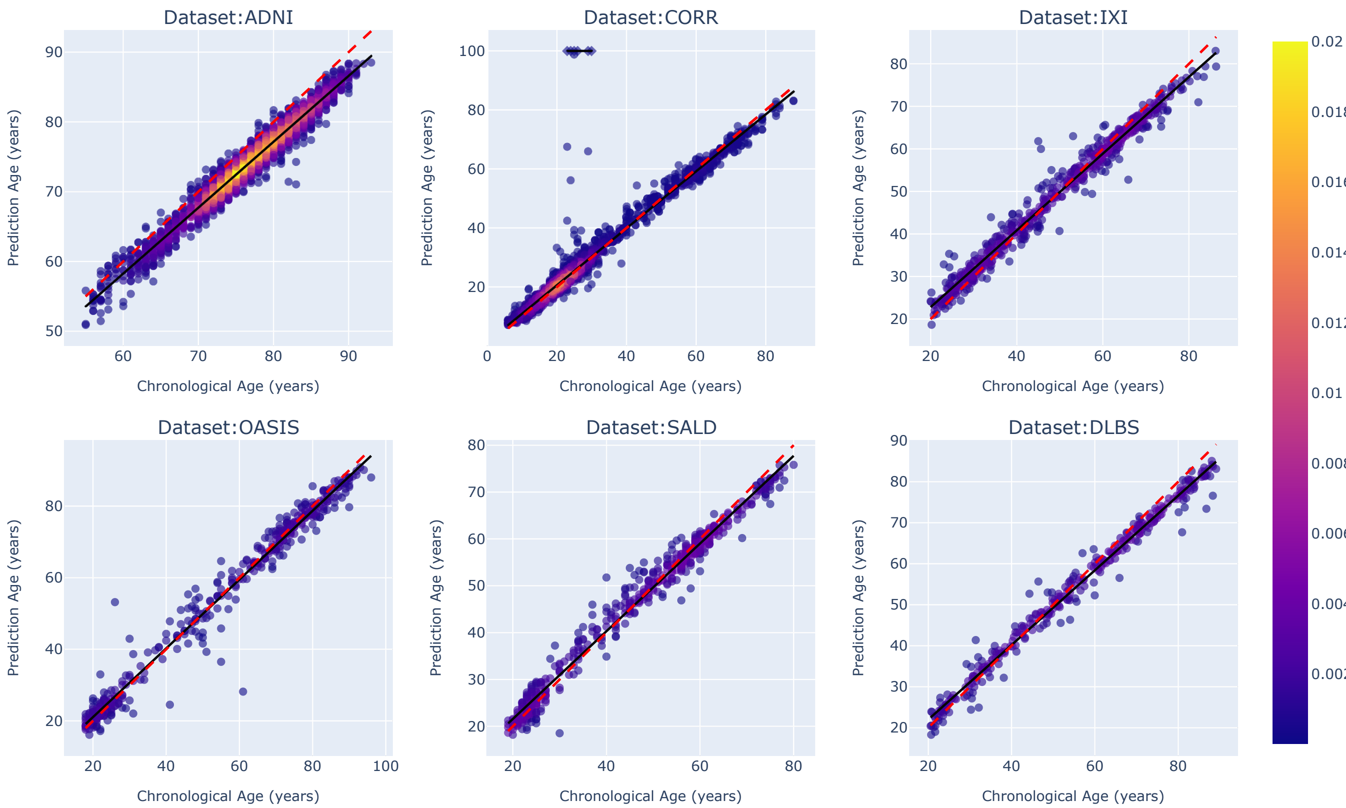


Fig. S.XVII. The scatter diagram of ResNet50 on six datasets.

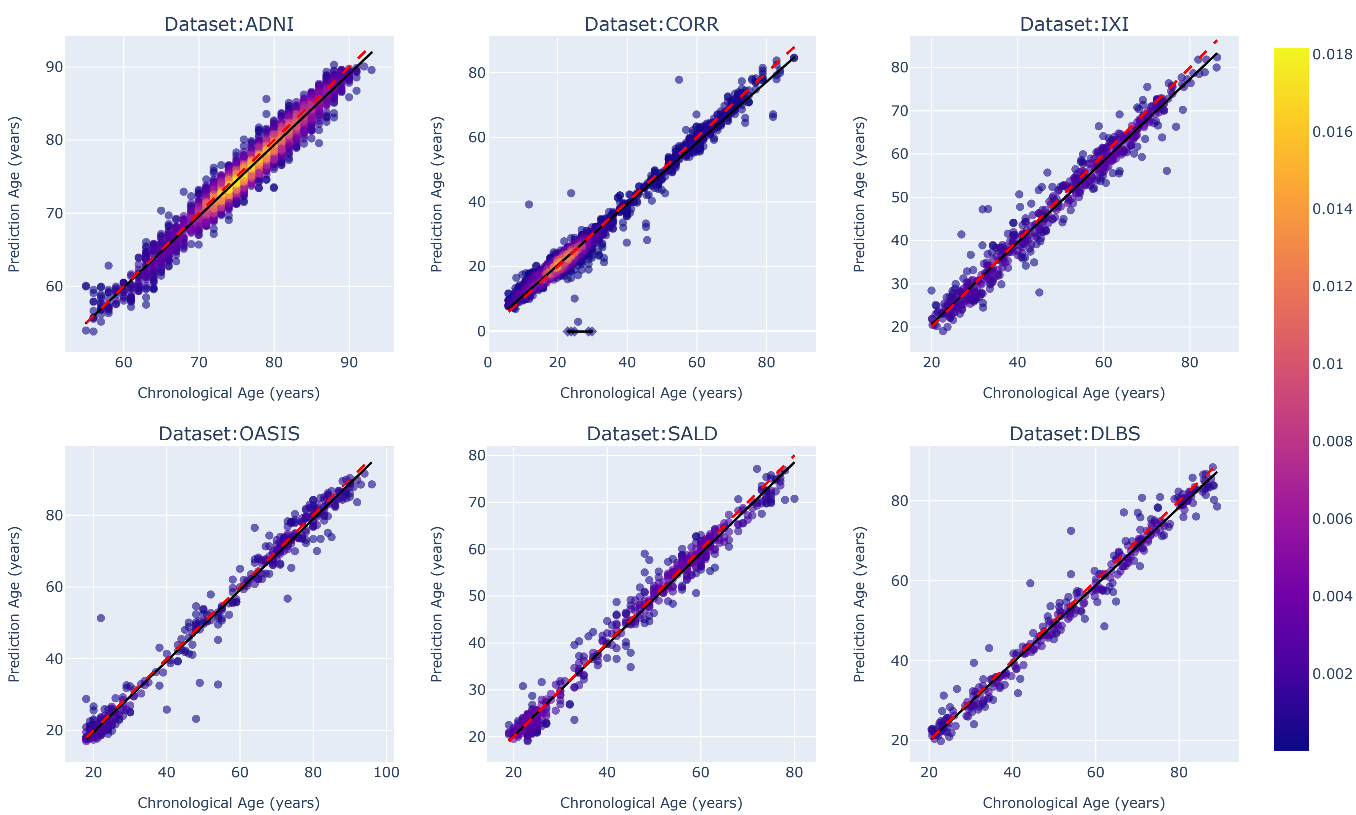


Fig. S.XVIII. The scatter diagram of ResNet101 on six datasets.

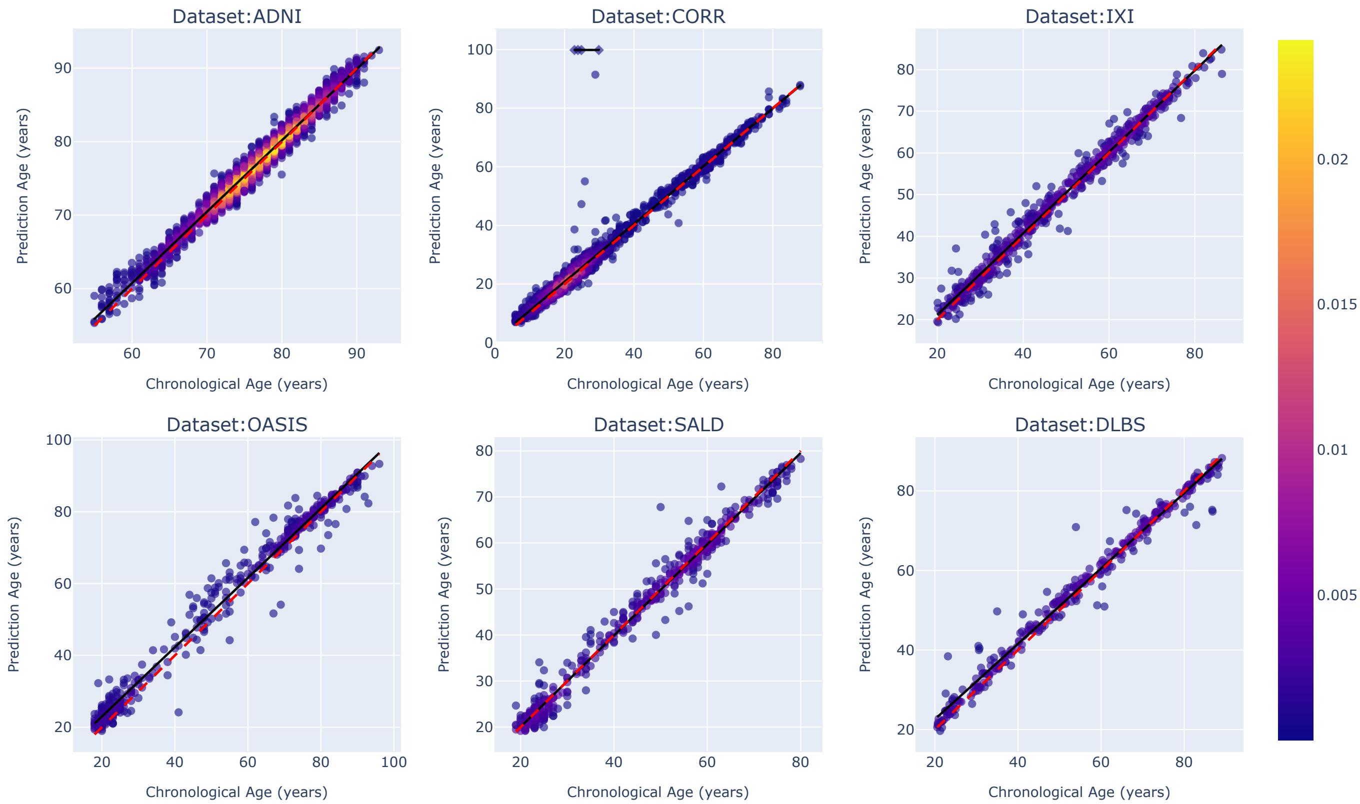


Fig. S.XIX. The scatter diagram of MAR3D-10 on six datasets.

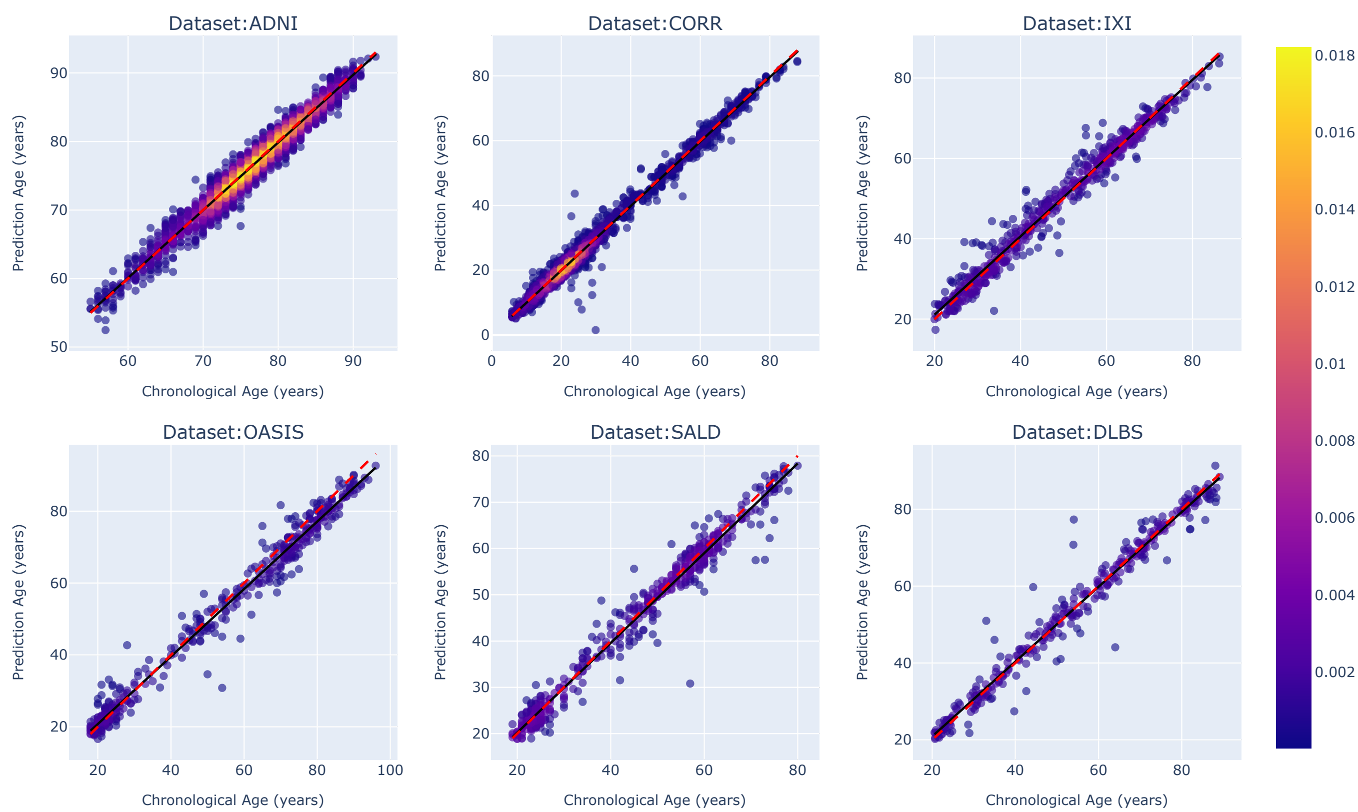


Fig. S.XX. The scatter diagram of MAR3D-18 on six datasets.

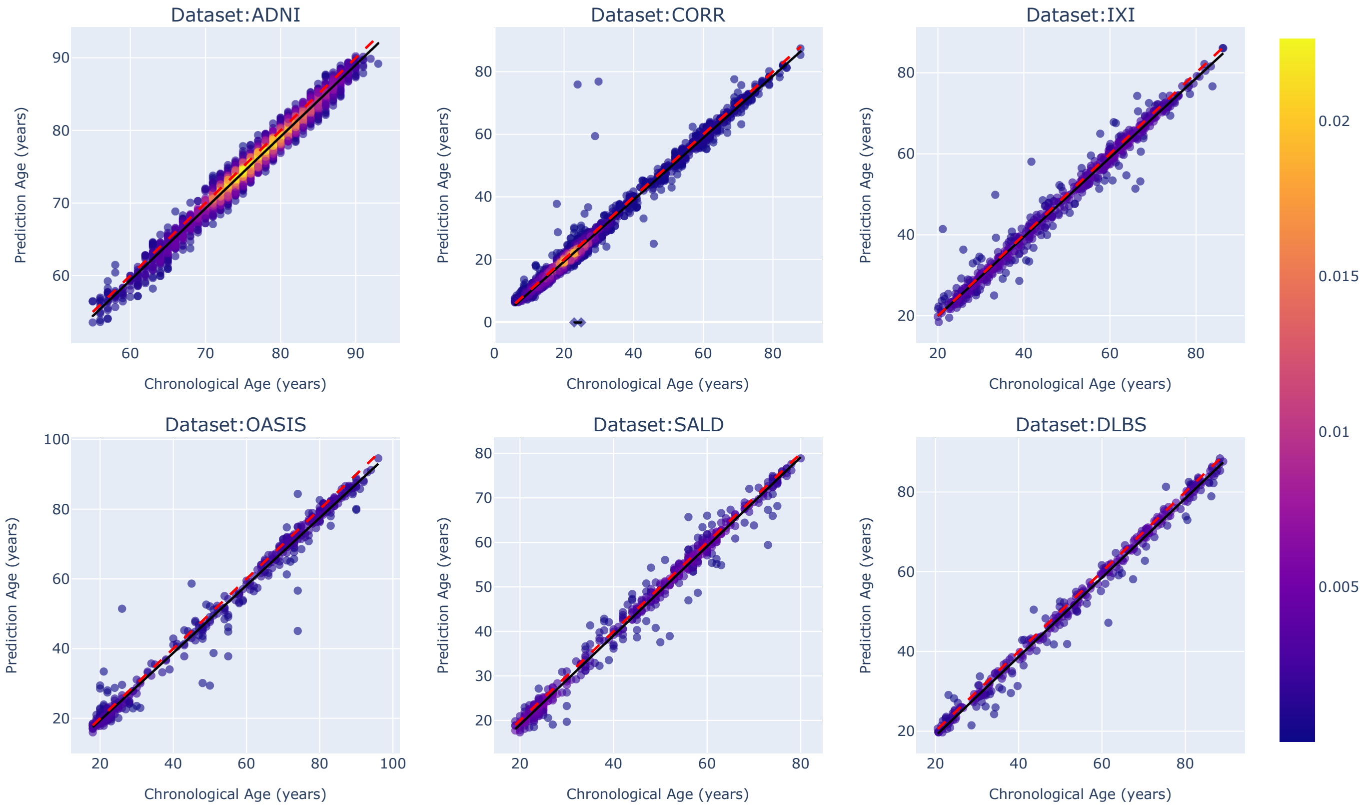


Fig. S.XXI. The scatter diagram of MAR3D-50 on six datasets.

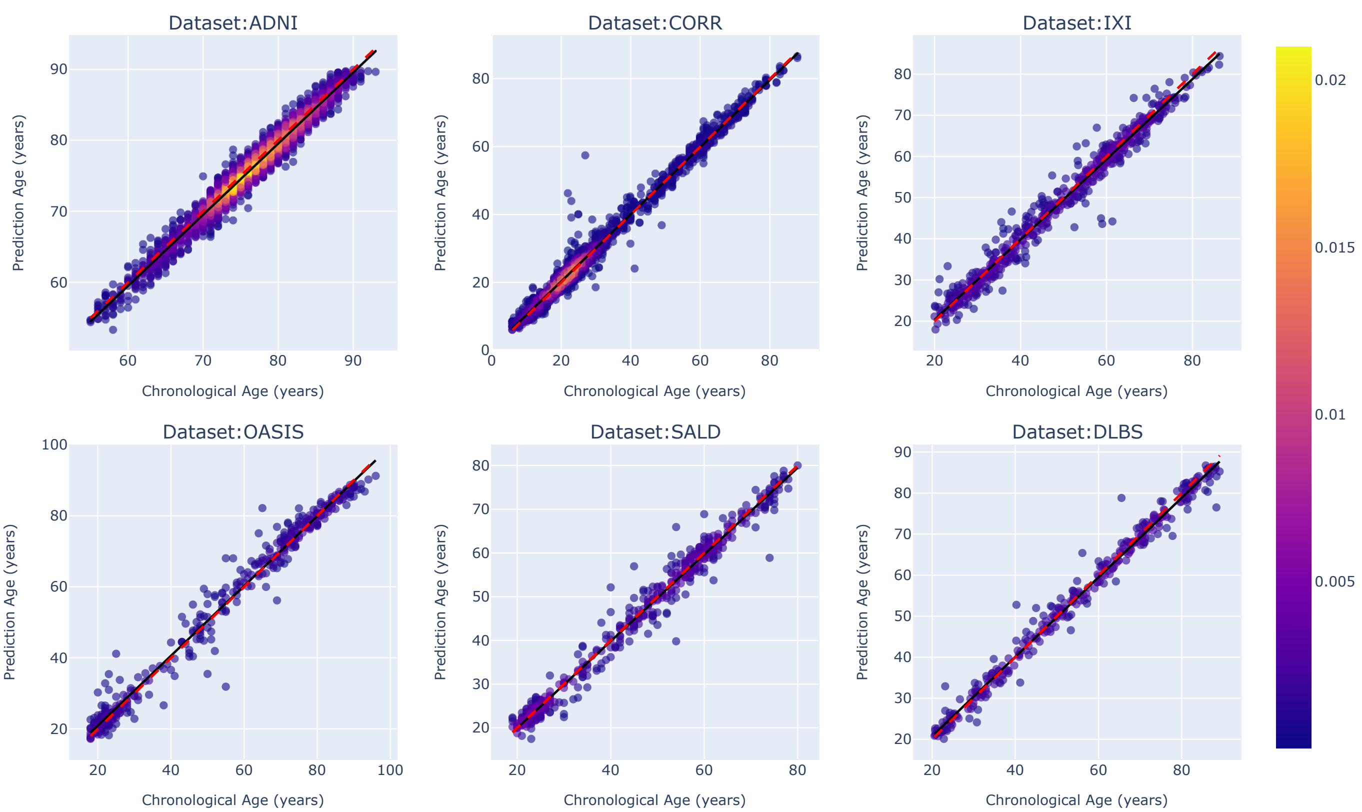


Fig. S.XXII. The scatter diagram of MAR3D-101 on six datasets.