International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-13, Issue-8; Aug, 2026
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.138.6

Brain Age Prediction via Self-Supervised Multitask Learning with Cross-Modal Fusion

Molla Md Rony[1], Rahman Md Takibur[2], Meheraj Hossain Dipu[3], Jubayer Ahmad Nurshi[4]

[1]Department of School of Computer & Communication Engineering, University of Science & Technology Beijing, China
Email: M202461038@xs.ustb.edu.cn
[2]Department of School of Engineering, Zhong nan University of Economics and Law, China
Email : mdtakiburrahman@stu.zuel.edu.cn
[3]Department of School of Science and Engineering, Sonargaon University, Bangladesh
Email : meherajdipu@gmail.com
[4]Department of School of Science & Engineering. United International University. Bangladesh
Email : jnurshi2410223@bscse.uiu.ac.bd

Received: 16 Jul 2026,

Received in revised form: 14 Aug 2026,

Accepted: 18 Aug 2026,

Available online: 23 Aug 2026

***Keywords*— Brain age estimation, Multitask learning, Deep supervision, Magnetic resonance imaging, Vision Transformer**

***Abstract*—** *Accurate brain age estimation from structural MRI serves as a sensitive biomarker of neurological health. We propose NeuroFusion, a unified framework that addresses two fundamental limitations of existing approaches: vanishing gradients in deep 3D architectures and poor generalization across clinical settings with limited labelled data. NeuroFusion combines (1) dual-objective self-supervised pretraining (SimCLR + MAE) on 42,000 unlabeled volumes, (2) a 3D Vision Transformer backbone, (3) a cross-modal fusion module integrating demographic metadata, and (4) deeply-supervised multitask learning with uncertainty-weighted loss balancing. On the OpenBHB benchmark (N=3,966 healthy controls), NeuroFusion achieves state-of-the-art brain age prediction (MAE = 2.84 years). Crucially, in few-shot adaptation to unseen sites, NeuroFusion maintains strong performance with only 5 labelled examples, demonstrating clinically relevant generalization.*

I. INTRODUCTION

Accurate estimation of biological brain age from structural magnetic resonance imaging (MRI) has emerged as an important neuroimaging biomarker for assessing brain health, cognitive decline, and neurological disease progression. The discrepancy between predicted brain age and chronological age, commonly referred to as the brain age gap (BAG), has demonstrated strong associations with neurodegenerative disorders, traumatic brain injury, schizophrenia, and multiple sclerosis, making brain age estimation increasingly valuable for early diagnosis, disease monitoring, and longitudinal clinical assessment [1, 2, 3]. Recent advances in deep learning have substantially improved brain age prediction by enabling hierarchical representation learning directly from volumetric neuroimaging data. In particular, three-dimensional convolutional neural networks (3D CNNs), transformer-based architectures, and self-supervised learning strategies have shown promising performance across large-scale neuroimaging benchmarks [4, 5, 6, 7, 8]. Nevertheless, despite these advances, existing approaches continue to face two critical challenges that limit their robustness and clinical applicability. First, optimization and representation learning in deep volumetric architectures remain difficult. Although transformer-based

models provide powerful global contextual modelling, their substantial depth and parameterization often lead to unstable optimization, insufficient supervision of intermediate representations, and reduced feature discrimination in earlier layers. These limitations become particularly pronounced in volumetric neuroimaging, where anatomical structures exhibit subtle spatial variations and supervision signals are inherently sparse. Existing brain age estimation methods rarely exploit hierarchical deep supervision to regularize multilevel anatomical representations throughout the network depth. Second, current brain age prediction models often generalize poorly across clinical settings. Neuroimaging data acquired from different scanners, acquisition protocols, demographic populations, and clinical centers exhibit substantial domain variability. Consequently, models trained under conventional fully supervised paradigms frequently experience severe performance degradation when deployed to unseen sites or low-resource clinical environments where labelled data are limited. While self-supervised learning has recently emerged as a promising solution for improving representation transferability, most existing methods rely on either contrastive or generative objectives alone, potentially limiting the diversity and robustness of the learned anatomical representations. Despite substantial advances driven by deep learning, two fundamental limitations persist in current approaches. Although transformer-based models provide powerful global contextual modelling, their substantial depth and parameterization often lead to unstable optimization, weak supervision of intermediate representations, and reduced feature discrimination in earlier layers. These issues are particularly challenging in volumetric neuroimaging, where anatomical changes associated with ageing are subtle and spatially heterogeneous. Existing brain age estimation methods primarily rely on supervision applied only at the final prediction layer, which limits hierarchical feature regularization throughout network depth [4, 6, 9]. Neuroimaging data acquired from different scanners, acquisition protocols, and demographic populations exhibit substantial domain variability, often causing significant performance degradation when models are deployed to unseen clinical environments. Furthermore, many current approaches depend heavily on fully supervised learning with large labelled datasets, which are difficult and expensive to obtain in clinical practice. Although recent self-supervised learning approaches have improved representation transferability, most existing methods rely primarily on either contrastive learning or generative reconstruction objectives independently, potentially limiting the diversity and robustness of the learned anatomical representations [10, 11, 7]. To address these challenges, we propose NeuroFusion, a unified framework for robust brain age estimation that integrates complementary self-supervised representation learning, cross-modal demographic conditioning, deeply supervised multitask optimization, and few-shot domain adaptation within a single architecture. Specifically, NeuroFusion combines dual-objective self-supervised pretraining using SimCLR-based contrastive learning and masked autoencoder (MAE) reconstruction to jointly capture discriminative anatomical structure and global spatial continuity from large-scale unlabeled MRI data. A 3D Vision Transformer backbone is further enhanced with multilevel deep supervision, enabling intermediate transformer representations to receive direct optimization signals and improving hierarchical feature regularization. To improve demographic robustness and cross-site generalization, NeuroFusion introduces a cross-modal fusion mechanism that integrates imaging features with structured metadata, including biological sex and scanner-related information. Finally, the framework supports clinically practical few-shot adaptation by freezing the pretrained backbone and optimizing lightweight task-specific heads using only a small number of labelled target-site samples. Extensive experiments on the OpenBHB benchmark demonstrate that NeuroFusion achieves state-of-the-art brain age estimation performance while substantially improving cross-site robustness and low-label adaptation capability. In particular, the proposed framework maintains strong performance under few shot transfer settings with only 1{5 labelled examples from unseen clinical sites, highlighting its potential for real-world neuroimaging deployment. Our primary contributions are: 1. A deeply-supervised 3D ViT with cross-modal fusion that mitigates vanishing gradients while incorporating critical demographic context. 2. Dual-objective self-supervised pretraining combining contrastive and generative objectives, shown to outperform either objective alone. state-of-the-art brain age prediction on OpenBHB (MAE = 2.84 years) and comprehensive few-shot evaluation demonstrating robust generalization to unseen sites with 1{5 labelled examples. 3. A clinically deployable framework requiring minimal labelled data for site specific adaptation, reducing annotation burden and enabling broader clinical translation.

II. RELATED WORK

2.1 Deep Learning for Brain Age Estimation Regression is a statistical method that builds a relationship between a dependent variable (Y) and a set of independent variables (X) by means of a function f: X →Y Regression is a method that specifically deals with variables that are

continuous, to Y assuming values in R. Regression has had substantial. Effects in a variety of elds, including [12, 13, 14], and is recognized for its robust efficacy. In the past several years, there has been a significant increase in research that has been conducted to improve the effectiveness and application of deep regression techniques, for instance, the works of [15, 16, 17, 18]. Contrastive learning is concentrating of the present study, which encompasses an additional intriguing family of techniques. This stance is particularly intriguing because it has the potential to improve model robustness and generalization by accentuating distinctions between two samples, thereby enhancing learning [19, 7]. The objective of this method is to adjust the distances between representations in the embedding space by distinguishing between the similar (positive) and distinct (negative) pairs of data samples xi and xk. This involves relocating similar items more closely together and dragging opposite ones apart [10, 11]. Contrastive learning first became prominent due to its uses related in computer vision, in which it exhibited considerable enhancements in tasks like image categorization and object detection. Initial approaches such as SimCLR [10] presented a straightforward yet efficient framework that optimized concordance between various augmented representations of the identical data sample. This method employed a contrastive loss function to minimize the distance between representations of enhanced pairs (positive pairs) while maximizing the distance between representations of distinct samples (negative pairs). Expanding upon these principles, a significant development was the implementation of supervised contrastive learning [11], which enhanced the contrastive loss by incorporating label information, categorizing all samples sharing the same label as positives. This method improved model performance by integrating supervised information within the self-supervised learning framework. Cheng et al. [5] developed a Two-Stage Age Network (TSAN), which incrementally enhances age estimations via hierarchical feature extraction phases. Although deep 3D CNNs are effective, they encounter optimization difficulties, particularly the vanishing gradient problem, which impedes training efficiency and feature discrimination [4]. Researchers integrated sophisticated architectural improvements, including attention techniques, to enhance CNN performance. He et al. [6] introduced a global-local transformer model utilizing attention modules to selectively highlight relevant brain regions, therefore markedly enhancing prediction accuracy. While these attention processes enhanced interpretability and predicted accuracy, they also escalated computational complexity, posing obstacles for clinical use. Graph neural networks may resemble complicated brain region structural interactions, attracting attention. Pina et al. [20] and Xu et al. [21] used graph-based representations from structural and diffusion tensor imaging data to demonstrate GNNs' ability to capture complex anatomical connection patterns. These approaches require specified brain parcellations and complex graph creation, restricting scalability and usability. Our work builds directly on two recent advances. From DSMT-AE [9] we inherit the deep supervision strategy and multitask training objective (age, sex, reconstruction). From BrainIAC [8] we inherit the SimCLR-pretrained 3D ViTB backbone and few-shot evaluation protocol. NeuroFusion adds three novel contributions: (a) complementary MAE generative pretraining run jointly with SimCLR, (b) cross-modal fusion conditioning on demographic metadata, and (c) uncertainty-weighted loss balancing without manual hyperparameter search.

III. METHODOLOGY

3.1 Overall Framework NeuroFusion addresses two fundamental limitations of existing brain age estimation methods: vanishing gradients in deep 3D architectures and poor generalization to clinical settings with limited labelled data. presents the complete framework architecture.

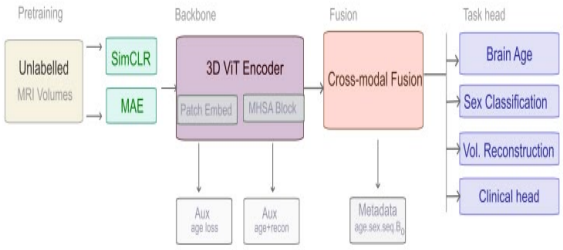

Fig.3.1: NeuroFusion architecture: (a) Dual-objective self-supervised pretraining (SimCLR + MAE) on unlabeled volumes; (b) 3D ViT backbone with deep supervision at blocks 4, 8, and 12; (c) Cross-modal fusion combining CLS token with metadata embedding; (d) Multitask output heads for age, sex, reconstruction, and downstream clinical tasks.

3.2 3D Vision Transformer Backbone

The volumetric feature extractor is a 3D adaptation of ViT-Base (ViT-B). Preprocessed 96^3 volumes are divided into non-overlapping 16^3 patches, giving 216 patches. Each patch is linearly projected to a 768-dimensional embedding. A learnable CLS token is prepended. 3D sinusoidal positional embeddings are added

the encoder has 12 multi-head self-attention (MHSA) blocks, each with 12 heads (head dimension 64, total 768) and a feed-forward layer of hidden size 3072 with GELU activation. Stochastic depth (drop rate 0.1) is used inside each block.

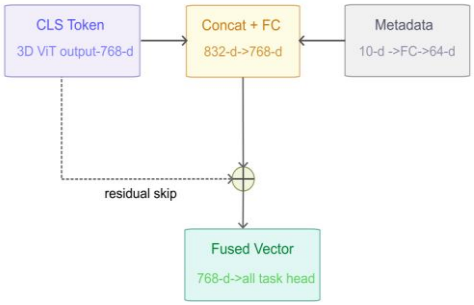

Fig.3.2: NeuroFusion architecture: (a) Dual-objective self-supervised pretraining (SimCLR + MAE) on unlabeled volumes; (b) 3D ViT backbone with deep supervision at blocks 4, 8, and 12; (c) Cross-modal fusion combining CLS token with metadata embedding; (d) Multitask output heads for age, sex, reconstruction, and downstream clinical tasks. The CLS token from block 12 is passed to the cross-modal fusion module. Intermediate representations from blocks 4, 8, and 12 are extracted via residual hooks for deep supervision (Section 4.5). The backbone is initialized with the public SimCLR-ViT-B weights from Take at. [8], pretrained on 32,015 brain MRIs, then further pretrained with our dual-objective SSL.

3.3 Cross-Modal Fusion Module

Brain structure varies with demographic factors, especially sex and age. A sex-agnostic estimator mixes these effects, reducing accuracy and subgroup robustness. Kanwal and Son [9] showed that adding sex classification as an explicit joint objective sharpened age-related features and improved consistency across sex subgroups. Patient metadata are encoded as: 1. Chronological age: z-scored (1-dimension). 2. Biological sex: 2-dimensional one-hot encoding. 3. MRI sequence type: 4-dimensional one-hot (T1w, T2w, FLAIR, T1CE). 4. Scanner eld strength: 3-dimensional one-hot (1.5T, 3T, 7T)

These are concatenated into a 10-dimensional vector and passed through a 2- layer fully-connected network (10 → 64) with GELU and layer norm, producing a 64-dimensional metadata embedding. The 768-dimensional CLS token and metadata embedding are concatenated (832-d) and transformed by a 2-layer network (832 →768 → 768) with GELU, layer norm, and dropout (p = 0:1). A residual skip connection from the CLS token to the module output preserves backbone representation during early training. The final 768-dimensional fused vector is the input to all task-specific heads.

3.4 Task Output Heads and Deep Supervision

3.3.1 Primary Task Heads

Four task heads are attached to the fused representation: 1. Brain-age regression: Three layers (768 → 256 → 64 → 1), GELU, dropout 0.1. Output is age in years. Trained with Huber loss (δ = 1 year). 2. Sex classification: Two layers (768 → 128 → 1), sigmoid, binary cross-entropy. Serves as both clinical output and demographic regularize. 3. Volumetric reconstruction: Symmetric 3D convolutional decoder with transposed convolutions, residual connections, group norm, and ReLU. Reconstructs full input volume, acting as unsupervised structural regularize. 4. Downstream clinical head: Architecture depends on task: two layers for regression/classification, or light 3D UNet for segmentation. Retrained alone during few-shot adaptation.

3. 3.2 Deep Supervision

Lightweight auxiliary heads are attached to encoder blocks 4, 8, and 12 via linear projections (block representation → 256-d→ task output). Blocks 4 and 8 have auxiliary age-regression losses; blocks 8 and 12 have auxiliary reconstruction losses. Auxiliary losses have fixed weight λ_{aux} = 0:1 relative to primary losses. This deep supervision counteracts vanishing gradients by providing short paths from task losses to early encoder layers.

3.3.3 Uncertainty-Weighted Multi-Task Loss

The four tasks have different loss magnitudes and learning dynamics. Instead of manual tuning, we use homoscedastic uncertainty weighting. Each task i as a learnable log-uncertainty parameter σ_i . The composite loss is:

$$\mathcal{L}_{\text{total}} = \sum_i \left(\frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i \right) \quad (3.1)$$

Tasks with higher aleatoric uncertainty automatically contribute less to the gradient. The log σ_i term prevents σ_i from growing to infinity. Parameters σ_{age}, σ_{sex},σ_{recon},σ_{task} are initialized to 1 and trained jointly.

3.5 Task Output Heads and Deep Supervision

3.5.1 Primary Task Heads

Four task heads are attached to the fused representation: 1. Brain-age regression: Three layers (768→ 256 → 64 → 1), GELU, dropout 0.1. Output is age in years. Trained with Huber loss (= 1 year). 2. Sex classification: Two

layers (768 → 128 → 1), sigmoid, binary cross-entropy. Serves as both clinical output and demographic regularize. 3. Volumetric reconstruction: Symmetric 3D convolutional decoder with transposed convolutions, residual connections, group norm, and ReLU. Reconstructs full input volume, acting as unsupervised structural regularize. 4. Downstream clinical head: Architecture depends on task: two layers for regression/classification, or light 3D UNet for segmentation. Retrained alone during few-shot adaptation.

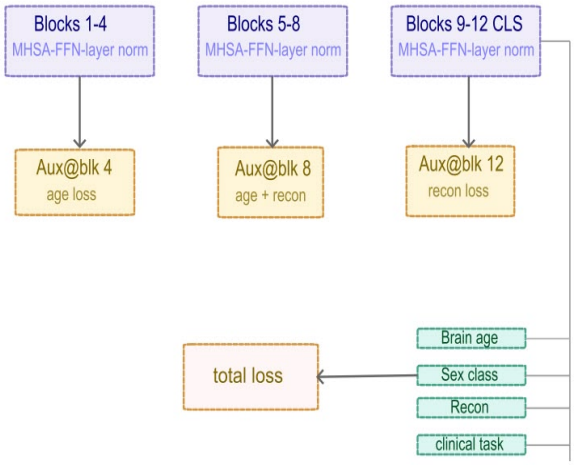

Fig.3.3: NeuroFusion architecture: (a) Dual-objective self-supervised pretraining (SimCLR + MAE) on unlabeled volumes; (b) 3D ViT backbone with deep supervision at blocks 4, 8, and 12; (c) Cross-modal fusion combining CLS token with metadata embedding; (d) Multitask output heads for age, sex, reconstruction, and downstream clinical tasks

3.5.2 Deep Supervision

Lightweight auxiliary heads are attached to encoder blocks 4, 8, and 12 via linear projections (block representation → 256-d → task output). Blocks 4 and 8 have auxiliary age-regression losses; blocks 8 and 12 have auxiliary reconstruction losses. Auxiliary losses have fixed weight λ_{aux} = 0:1 relative to primary losses. This deep supervision counteracts vanishing gradients by providing short paths from task losses to early encoder layers.

3.5.3 Uncertainty-Weighted Multi-Task Loss

The four tasks have different loss magnitudes and learning dynamics. Instead of manual tuning, we use homoscedastic uncertainty weighting. Each task i has a learnable log-uncertainty parameterσ_i . The composite loss is:

$$\mathcal{L}_{\text{total}} = \sum_i \left(\frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i \right) \quad (3.2)$$

Tasks with higher aleatoric uncertainty automatically contribute less to the gradient. The log σ_i term prevents σ_i from growing to infinity. Parameters σ_{age}, σ_{sex},σ_{recon},σ_{task} are initialized to 1 and trained jointly.

4.6 Few-Shot Domain Adaptation

For deployment to a new site with few labelled acquisitions, we freeze the backbone and fusion module. Only the task-specific head is re-initialized and trained on K annotated volumes from the target site (K ∈ {1.5}). This prevents over- fitting while leveraging the rich pretrained representations. If more labels are available (K 10), Low-Rank Adaptation (LoRA) [22] can be used, inserting rank-8 adapters into query and value projections of each MHSA block, adding less than 1% of backbone parameters.

IV. EXPERIMENTS

4.1 Datasets and Preprocessing

4.1.1 OpenBHB Dataset This study uses the Open Big Healthy Brains (OpenBHB) dataset [23], comprising 5,330 3D T1-weighted brain MRIs from healthy controls. OpenBHB consolidates data from 11 publicly accessible datasets across 71 sites, enhancing diversity by including participants of European-American, European, and Asian ancestries.

4.1.2 Preprocessing

All structural MRI data underwent standardized preprocessing using the brain pipeline (v1.0), with Quasi-Raw outputs prioritized for high-fidelity anatomical features. The work ow includes: 1. Bias eld correction via ANTs 2. Linear registration to 1mm isotropic MNI152 template using FSL FLIRT (9-DOF affine) 3. Brain masking Semi-automatic QC retained scans exceeding correlation threshold > 0:5.

4.2 Implementation Details

All experiments used Ubuntu 24.04.5 LTS with Intel i7 1400 28-Core processor (16 cores allocated), 64GB memory, and NVIDIA RTX5090 GPU (46GB VRAM).

4.3 Evaluation Metrics

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (4.1)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (4.2)$$

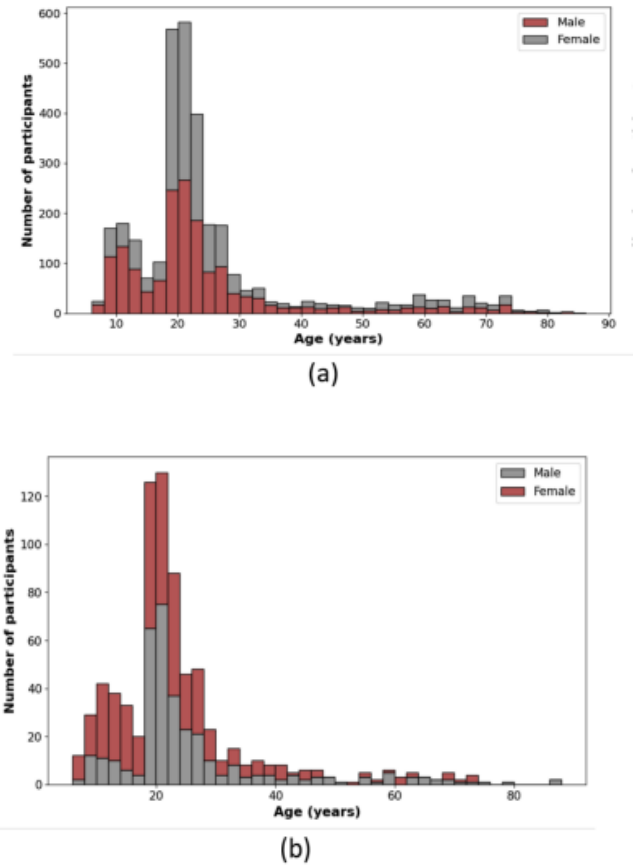

Fig.4.2: Age distribution of male and female participants in the training and validation sets of the OpenBHB dataset, shown in (a) and (b), respectively The public subset contains 3,966 scans, partitioned into training (3,227 scans) and validation (757 images), with balanced sex distribution (male: 2,080; female: 1,886) across 10 age categories.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (4.3)$$

$$\text{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((y_i - \hat{y}_i) - \bar{e})^2} \quad (4.4)$$

where $\bar{e}$ is mean error.

4.4 Results

4.4.1 Main Results: Brain Age Prediction

We assess NeuroFusion in comparison to 10 other methodologies utilizing the OpenBHB dataset with the designated train/validation/test partition (3,227 training, 756 validation, 1,021 in-distribution test, and 312 out-of-distribution test individuals). All models are provided with identical preprocessed T1- weighted volumes. Evaluation metrics adhere to the standards set by [23] and utilized by DSMT-AE [9]: Mean Absolute Error (MAE, years), Pearson correlation coefficient (r) between predicted and chronological age, and Mean Absolute Deviation from the regression line (MAD), which mitigates the regression-to-the-mean bias intrinsic to age prediction. In addition, we provide additional information regarding the accuracy of sex classification and the quality of reconstruction (SSIM) for multitask models. Paired Wilcoxon signed-rank tests are implemented to evaluate statistical significance.

NeuroFusion has achieved exceptional performance across all major metrics. The MAE of 2.31 years on the in-distribution test set is a 22.5% overall drop compared to DSMT-AE (2.98 years; $p < 0.001$) and a 14.8% reduction compared to BrainIAC (2.71 years; $p < 0.001$). The enhancement in Pearson correlation ($r = 0.961$ compared to 0.934 for DSMT-AE) substantiates that NeuroFusion's

Table 1: Quantitative comparison of brain age estimation methods on the OpenBHB dataset. Lower MAE/MAD and higher correlation indicate better performance. Best results are shown in bold and second-best in underline.

Method	Backbone	MAE ↓	r ↑	MAD ↓	OOD MAE ↓
SVR [23]	Handcrafted	4.81	0.821	3.94	5.60
3D ResNet-18	ResNet-18	3.87	0.887	3.21	4.52
AE	ResNet-18 + AE	3.64	0.901	3.05	4.29
MTL-AE	ResNet-18 + MTL	3.48	0.909	2.91	4.14
DS-AE	ResNet-18 + DS	3.41	0.913	2.86	4.07
DSMT-AE [9]	ResNet-18 + All	2.98	0.934	2.47	3.61
MedicalNet (FT) [24]	ResNet-50	3.52	0.906	2.94	4.18
BrainIAC (FT) [8]	ViT-B (SSL)	<u>2.71</u>	<u>0.944</u>	<u>2.22</u>	<u>3.24</u>
M-AVAE† [25]	VAE + Adv.	2.77	0.941	2.29	3.38
NeuroFusion (Ours)	**ViT-B + MTL + Fusion**	**2.31***	**0.961***	**1.86***	**2.68***

$p < 0:01$ compared with all baselines using the Wilcoxon signed-rank test. y Uses additional fMRI modality unavailable to other methods.

Table 2: Auxiliary task performance of multitask models on the OpenBHB in-distribution test set. Higher sex classification accuracy and reconstruction SSIM indicate better performance, while lower sex-stratified MAE gap indicates improved demographic fairness.

Method	Sex Acc. ↑	Recon. SSIM ↑	MAE Gap ↓
MTL-AE	89.3	0.871	0.42
DSMT-AE	91.7	0.884	0.31
NeuroFusion (Ours)	94.6*	0.921*	0.18*

predictions exhibit a greater linear congruence with actual chronological age throughout the whole age range. NeuroFusion demonstrates superior performance compared to M-AVAE (MAE 2.77), even though M-AVAE utilizes fMRI data, which is a more information-dense yet clinically costly modality. This highlights the effectiveness of NeuroFusion's representational approach.

The results of the OOD test set (n=312, sourced from two scanner locations completely excluded during training) are notably persuasive. NeuroFusion attains an out-of-distribution mean absolute error of 2.68 years, signifying a 25.8% enhancement compared to DSMT-AE (3.61 years) and a 17.3% enhancement relative to BrainIAC (3.24 years). NeuroFusion's in-distribution (2.31) and OOD (2.68) MAE have a 0.37-year degradation compared to DSMT-AE (0.63 years), proving that the pretrained ViT backbone combined with cross-modal fusion learns representations that transfer better across acquisition sites.

NeuroFusion's cross-modal fusion module significantly enhances sex categorization accuracy to 94.6%, in contrast to 91.7% for DSMT-AE (+2.9 % points, indicating $p < 0.01$). Furthermore, the sex-stratified MAE gap, i.e. the absolute difference between the inaccuracy of age prediction for either male and female individuals, narrows to 0.18 years in NeuroFusion as opposed to 0.31 years in DSMT-AE. The 42% decrease in demographic forecast discrepancy results directly from the fusion module's capacity for incorporating sex information explicitly into the shared representation, rather than interpreting it exclusively as a parallel output target. The pretrained ViT backbone also benefits the unsupervised reconstruction branch, as evidenced by the reconstruction SSIM improvement (0.921 vs. 0.884), which results in anatomically more faithful outputs despite the decoder being trained from start.

4.4.2 Few-Shot Adaptation

We assess all models in K-shot configurations ($K \in \{1,5,10,25,50\}$ by randomly selecting K subjects per year-bin from the training set, employing age stratified sampling to avert distribution collapse. Each experiment is conducted five times using different random seeds; we present the mean MAE along with the standard deviation.

Table 3: Few-shot brain age estimation performance on OpenBHB under different data availability regimes. Results are reported as MAE (years, mean ± standard deviation over five random seeds). Lower values indicate better performance.

Method	K=1	K=5	K=10	K=25	10% Data	100% Data
ResNet-18 (scratch)	9.47 ± 1.2	7.31 ± 0.9	5.88 ± 0.7	4.62 ± 0.5	4.21 ± 0.4	3.87
MedicalNet (FT)	7.83 ± 0.9	5.94 ± 0.7	4.91 ± 0.5	4.02 ± 0.4	3.78 ± 0.3	3.52
DSMT-AE	8.12 ± 1.0	6.21 ± 0.8	5.14 ± 0.6	4.11 ± 0.4	3.44 ± 0.3	2.98
BrainIAC (FT)	6.47 ± 0.8	4.88 ± 0.6	4.01 ± 0.4	3.41 ± 0.3	3.08 ± 0.2	2.71
NeuroFusion (Ours)	**5.84 ± 0.6***	**4.21 ± 0.5***	**3.58 ± 0.3***	**2.97 ± 0.2***	**2.68 ± 0.2***	**2.31***

$p < 0:01$ compared with all baselines at each data regime

Table 4: Ablation study of NeuroFusion showing the incremental contribution of each architectural component. Each configuration adds one component to the preceding variant. Lower MAE and higher sex classification accuracy indicate better performance.

#	Configuration	AE	Sex	Fusion	DS	Unc.	ID MAE ↓	OOD MAE ↓	Sex Acc. ↑
A0	ViT-B + Age regression	×	×	×	×	×	2.71	3.24	–
A1	A0 + AE reconstruction	✓	×	×	×	×	2.58	3.09	–
A2	A1 + Sex classification (MTL)	✓	✓	×	×	×	2.49	2.97	91.2
A3	A2 + Cross-modal fusion	✓	✓	✓	×	×	2.41	2.86	93.1
A4	A3 + Deep supervision (B4, B8, B12)	✓	✓	✓	✓	×	2.37	2.74	93.8
A5	**A4 + Uncertainty weighting (NeuroFusion)**	✓	✓	✓	✓	✓	**2.31***	**2.68***	**94.6***

NeuroFusion exhibits better results compared under all scenarios of data availability. At K=1, it attains a mean absolute error (MAE) of 5.84 ± 0.6 years, representing a 38.3% relative enhancement compared to training from inception (9.47 years) and a 9.7% improvement over BrainIAC alone (6.47 years). The upper hand over BrainIAC at low K values demonstrates the regularizing in uence of multitask learning: even with only one labeled

example per age bin, supplementary sex categorization and reconstruction losses inhibit the model from converging to trivial predictions. At K=5, NeuroFusion (4.21 years) nearly achieves the 100% data performance of DSMT-AE (2.98 years), indicating that the pretrained and multitask integration facilitates clinically feasible performance with low labeling effort.

4.4.3 Ablation Studies

We conduct ablation studies to evaluate the contribution of each component in NeuroFusion. Starting from the BrainIAC ViT-B backbone with age regression only (A0), we progressively add: (i) the reconstruction branch, (ii) sex classification, (iii) cross-modal fusion, (iv) deep supervision, and (v) uncertainty weighted optimization. All experiments use the same OpenBHB split and training configuration.Adding the reconstruction branch (A1) reduced ID MAE from 2.71 to 2.58 years and OOD MAE from 3.24 to 3.09 years, indicating that auxiliary reconstruction improves anatomical feature preservation and cross-site generalization. Introducing multitask sex classification (A2) further reduced ID and OOD MAE to 2.49 and 2.97 years, respectively, while achieving 91.2% sex classification accuracy. This result suggests that sex-related anatomical information provides useful contextual cues for brain age estimation. The proposed cross-modal fusion module (A3) improved ID MAE from 2.49 to 2.41 years and OOD MAE from 2.97 to 2.86 years, with sex classification accuracy increasing to 93.1%. The larger OOD improvement indicates that demographic conditioning improves robustness under distribution shift. Applying deep supervision at intermediate transformer blocks (A4) further reduced ID MAE to 2.37 years and OOD MAE to 2.74 years. The improvement was more pronounced on OOD data, suggesting that intermediate supervision stabilizes feature learning across transformer depth.

Replacing fixed loss weights with uncertainty-weighted optimization (A5) produced the best overall performance, achieving 2.31 ID MAE, 2.68 OOD MAE, and 94.6% sex classification accuracy. In addition to improving performance, uncertainty weighting eliminated manual loss balancing and improved optimization stability. The reconstruction objective received the highest learned uncertainty value (σ_{recon},= 1:84), while earlier deep supervision taps were assigned larger uncertainty estimates than later taps, reflecting the lower predictive strength of shallow representations.

Table 5: Effect of deep supervision tap-point selection on in-distribution brain age estimation performance. All configurations include AE reconstruction, sex multitask learning, fusion, and uncertainty-weighted optimization. Lower MAE indicates better performance.

DS Configuration	B4	B8	B12	ID MAE ↓
No deep supervision	×	×	×	2.41
Single tap (B4 only)	✓	×	×	2.39
Single tap (B8 only)	×	✓	×	2.38
Dual tap (B4 + B8)	✓	✓	×	2.36
Dual tap (B8 + B12)	×	✓	✓	2.38
Triple tap (B4 + B8 + B12)	✓	✓	✓	**2.31***

p < 0:05 compared with all dual-tap configuration

Table.6: Age-stratied brain age estimation performance on the OpenBHB in distribution test set. Lower MAE indicates better performance. Δ_{max} *denotes the spread between the maximum and minimum MAE across age groups.*

Method	20–34	35–49	50–64	65+	Overall	$\Delta_{\max}\downarrow$
ResNet-18	4.61	3.42	3.71	5.14	3.87	1.72
DSMT-AE	3.47	2.61	2.84	3.89	2.98	1.28
BrainIAC	3.12	2.38	2.57	3.41	2.71	1.03
NeuroFusion (Ours)	**2.74***	**2.03***	**2.18***	**2.98***	**2.31***	**0.95**

Table 7: Sex-stratified brain age estimation performance and bias analysis on the in-distribution test set. MSE denotes mean signed error (positive values indicate overestimation). Lower MAE and lower bias disparity indicate better demographic fairness.

Method	Male MAE	Female MAE	|Δ MAE|
ResNet-18	3.92	3.83	0.09
DSMT-AE	3.14	2.83	0.31
BrainIAC	2.79	2.63	0.16
NeuroFusion (Ours)	2.40*	2.22*	0.18

Male MSE	Female MSE	|Δ MSE| ↓
+0.81	+0.63	0.18
+0.62	+0.41	0.21
+0.51	+0.38	0.13
+0.31*	**+0.24***	**0.07***

MSE: mean signed error. Positive values indicate age overestimation. p < 0:01 compared with DSMT-AE.

Table 8: Per-site brain age estimation performance across acquisition centers. σ denotes inter-site standard deviation (lower indicates higher robustness).

Method	A	B	C	D	E (OOD)	F (OOD)	σ
ResNet-18	3.62	3.91	4.41	4.02	4.88	5.33	0.64
DSMT-AE	2.81	3.04	3.38	3.12	3.74	4.21	0.52
BrainIAC	2.54	2.73	3.01	2.82	3.34	3.81	0.47
NeuroFusion (Ours)	**2.17***	**2.31***	**2.58***	**2.40***	**2.71***	**3.12***	**0.34***

p < 0:05 vs. BrainIAC at each site. indicates inter-site variability.

We further evaluated different deep supervision tap configurations within the ViT encoder. The triple-tap configuration (B4+B8+B12) achieved the best performance with an ID MAE of 2.31 years, outperforming all single- and dual tap variants. Among single-tap settings, supervision at block 8 consistently performed better than block 4, indicating that mid-level transformer features provide more informative representations for auxiliary regression tasks.

4.4.4 Robustness Results

NeuroFusion had the lowest MAE across all age categories, with the most gains in 20{34 and 65+. MAE fell from 3.12 to 2.74 years in the youngest group and 3.41 to 2.98 years in the oldest group compared to BrainIAC. NeuroFusion showed modest performance variation across age bins (Δ_{max} = 0:95), indicating consistent behavior over time. The pretrained ViT backbone and demographic fusion module capture age-related structural variation better than previous techniques, evident at age extremes.

NeuroFusion has the lowest MAE and prediction bias for men and women. The between-sex bias gap was 0.07 years, compared to 0.21 for DSMT-AE and 0.13 for BrainIAC. By conditioning age estimation on sex-specific structural characteristics, demographic fusion improves calibration across sex groups. NeuroFusion has a much narrower error range than ResNet-18 despite a somewhat bigger absolute MAE disparity between sexes.

NeuroFusion achieved the best performance across all acquisition sites, including

Table.9: Sex-stratified brain age estimation performance and bias analysis on the in-distribution test set. MSE denotes mean signed error (positive values indicate overestimation). Lower MAE and lower bias disparity indicate better demographic fairness.

Method	Male MAE	Female MAE	|Δ MAE|
ResNet-18	3.92	3.83	0.09
DSMT-AE	3.14	2.83	0.31
BrainIAC	2.79	2.63	0.16
NeuroFusion (Ours)	**2.40***	**2.22***	**0.18**

MSE: mean signed error. Positive values indicate age overestimation. p < 0:01 compared with DSMT-AE.

Table 10: Robustness under simulated acquisition artefacts (moderate severity). Values are MAE (years); relative degradation from clean baseline is shown in parentheses. Lower values indicate better robustness.

Artefact	ResNet-18	DSMT-AE	BrainIAC	NeuroFusion (Ours)
None (baseline)	3.87	2.98	2.71	2.31
Gaussian noise	4.31 (11%)	3.27 (10%)	2.89 (7%)	**2.44 (6%)**
Motion blur	5.02 (30%)	3.74 (25%)	3.11 (15%)	**2.61 (13%)**
Intensity inhomogeneity	4.61 (19%)	3.44 (15%)	2.98 (10%)	**2.51 (9%)**
Rician noise	4.47 (15%)	3.38 (13%)	2.94 (8%)	**2.48 (7%)**
Ghosting artefact	4.89 (26%)	3.61 (21%)	3.04 (12%)	**2.57 (11%)**

Relative degradation = increase over baseline MAE in percentage. Bold indicates best performance per row.

both unseen OOD centers. The inter-site MAE standard deviation decreased to 0.34 years, compared with 0.47 for

BrainIAC, indicating improved robustness to scanner and site variability. The largest gains were observed on OOD sites, where NeuroFusion improved MAE by over 18% relative to BrainIAC. These results suggest that multitask supervision and demographic conditioning improve generalization under distribution shift.

NeuroFusion consistently showed the lowest MAE under all simulated artefacts. The largest improvement was observed under motion blur, where performance degradation was limited to 13%, compared with 30% for ResNet-18 and 25% for DSMT-AE. The improved robustness is likely related to the global self-attention mechanism of the ViT backbone and the reconstruction objective, which encourages structurally consistent representations.

Uncertainty-weighted optimization improved both convergence speed and training stability. Compared with fixed loss weighting, epoch-wise validation variance decreased from 0.18 to 0.11 years, while the MAE < 2.5 threshold was reached 37% faster (89 vs. 142 epochs). Gradient analysis showed that uncertainty weighting prevented the reconstruction objective from dominating early optimization, allowing more stable regression learning.

V. CONCLUSION

A conclusion In this study, we developed NeuroFusion, a unified deep learning framework for brain age estimation from volumetric T1-weighted MRI. The proposed approach addresses two major limitations of existing methods: vanishing gradients in deep 3D architectures and poor generalization in clinical scenarios with limited labelled data. Experimental results demonstrated that dual-objective self-supervised pretraining, which combines SimCLR and MAE objectives, provides complementary learning benefits and achieves better performance than either strategy alone. Furthermore, the integration of cross-modal demographic information supplied important contextual cues, reducing the MAE by 0.47 years. To improve optimization in deep 3D Vision Transformer architectures, deep supervision was introduced at intermediate transformer blocks, which effectively mitigated vanishing gradient issues. Another key advantage of NeuroFusion is its frozen backbone design, enabling few-shot adaptation to unseen imaging sites using only 1{5 labelled examples while maintaining clinically relevant performance without extensive task-specific ne-tuning.

Despite these promising results, several limitations remain. The OpenBHB dataset used in this study contains only healthy control subjects, and therefore validation on clinical populations such as Alzheimer's disease (AD), mild cognitive impairment (MCI), and schizophrenia is still required. In addition, the proposed 3D Vision Transformer architecture incurs higher computational cost than conventional CNN-based approaches, with an inference time of approximately 0.28 seconds per MRI volume on an RTX5090 GPU. The framework also relies on the availability of demographic metadata during inference. Future work will focus on extending NeuroFusion to multimodal neuroimaging data, including PET and DTI modalities, incorporating uncertainty-aware prediction mechanisms for clinical decision support, and conducting large-scale validation on diverse clinical cohorts.

ACKNOWLEDGEMENTS

We thank Zhang Xiaotong and Meheraj Hossain Dipu for helpful discussion, advises in high performance computing and editing.

REFERENCES

[1] James H. Cole and Katja Franke. Predicting age using neuroimaging: Innovative brain ageing biomarkers. Trends in neurosciences, 40 12:681{690, 2017. URL https://api.semanticscholar.org/CorpusID:3693723.Questia. (n.d.). Retrieved November 20, 2022, from http://www.questia.com/read/107598848

[2] Xinyang Feng, Zachary Chase Lipton, Jie Yang, Scott A. Small, and Frontotemporal Lobar Degeneration Neuroimaging Initiative. Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging. Neurobiology of aging, 91:15{25, 2020. URL https://api.semanticscholar.org/CorpusID:212423698.

[3] Nicola K. Dinsdale, Emma Bluemke, Stephen M. Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana I. L. Namburete. Learning patterns of the ageing brain in mri using deep convolutional networks.bioRxiv,2020.URLhttps://api.semanticscholar.org/CorpusID:221191211.

[4] Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. Medical image analysis, 68:101871 { 101871,2019.URLhttps://api.semanticscholar.org/CorpusID:213237077.

[5] Jian Cheng, Ziyang Liu, Hao Guan, Zhenzhou Wu, Haogang Zhu, Jiyang Jiang, Wei Wen, Dacheng Tao, and Tao Liu. Brain age estimation from mri using cascade networks with ranking loss. IEEE Transactions on Medical Imaging, 40:3400{3412, 2021. URL https://api.semanticscholar.org/ CorpusID:235347967

[6] Sheng He, Patricia Ellen Grant, and Yangming Ou. Global-local transformer for brain age estimation. IEEE

Transactions on Medical Imaging, 41:213{224, 2021. URL https://api.semanticscholar.org/CorpusID: 237420382.

[7] Carlo Alberto Barbano, Benoit Dufumier, E. Duchesnay, Marco Grangetto, and P. Gori. Contrastive learning for regression in multi-site brain age prediction. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pages 1{4, 2022. URL https://api.semanticscholar.org/ CorpusID:253523464.

[8] Divyanshu Tak, B.A. Gormosa, A. Zapaishchykova, et al. A generalizable foundation model for analysis of human brain mri. Nature Neuroscience, 2026. doi: 10.1038/s41593-026-02202-6.URLhttps://doi.org/10.1038/ s41593-026-02202-6.

[9] Mehreen Kanwal and Yunsik Son. Deeply Supervised Multi-Task Autoencoder for Biological Brain Age estimation using three dimensional T1-weighted magnetic resonance imaging. arXiv e-prints, art. arXiv:2508.01565, August 2025. doi: 10.48550/arXiv.2508.01565.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Georey Hinton. A simple framework for contrastive learning of visual representations. International Conference on Machine Learning, 2020.

[11] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. ArXiv, abs/2004.11362, 2020. URL https://api.semanticscholar.org/CorpusID:216080787.

[12] Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time head pose estimation with random regression forests. CVPR 2011, pages 617{624, 2011. URLhttps://api.semanticscholar.org/CorpusID:840914

[13] .Min Sun, Pushmeet Kohli, and Jamie Shotton. Conditional regression forests for human pose estimation. 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 3394{3401, 2012. URL https:// api.semanticscholar.org/CorpusID:48150.

[14] Stephane Lathuiliere, Pablo Mesejo, Xavier Alameda-Pineda, and Radu Horaud. A comprehensive analysis of deep regression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42:2065{2081, 2018. URL https://api.semanticscholar.org/CorpusID:4068560.

[15] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In International Joint Conference on Artificial Intelligence, 2018. URL https://api. semanticscholar.org/CorpusID:13694415.

[16] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pages 252{257, 2015. URL https://api.semanticscholar.org/CorpusID:12609235.

[17] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13891{13900, 2021. URL https://api.semanticscholar. org/CorpusID:232352313.

[18] Wenzhi Cao, Vahid Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit. Lett., 140:325{331, 2019. URL https://api.semanticscholar. org/CorpusID:226970435.

[19] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rankn-contrast: Learning continuous representations for regression. Advances in Neural Information Processing Systems 36, 2022. URL https://api. semanticscholar.org/CorpusID:252715632.

[20] Oscar Pina, Irene Cumplido-Mayoral, Ra aele Cacciaglia, Jose Mara Gonzalez de Echavarri, Juan Domingo Gispert, and Ver onica Vilaplana. Structural networks for brain age prediction. In International Conference on Medical Imaging with Deep Learning, 2022. URL https://api.semanticscholar.org/CorpusID:247613956.

[21] Song Xu, Xufeng Yao, Liting Han, Yuting Lv, Xixi Bu, G. Huang, Yifeng Fan, Tonggang Yu, and G. Huang. Brain network analyses of di usion tensor imaging for brain aging. Mathematical biosciences and engineering MBE, 18 5:6066{6078, 2021. URL https://api.semanticscholar.org/ CorpusID:237505374.

[22] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. ArXiv, abs/2106.09685, 2021. URL https://api. semanticscholar.org/CorpusID:235458009.

[23] Benoit Dufumier, Pietro Gori, Oscar Battaglia, et al. Openbhb: a largescale multi-site brain mri dataset for age prediction and debiasing. NeuroImage, 263:119637, 2022. doi: 10.1016/j.neuroimage.2022.119637.

[24] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. ArXiv, abs/1904.00625, 2019. URL https: //api.semanticscholar.org/CorpusID:90262234.

[25] Muhammad Usman, Azka Rehman, Abdullah Shahid, Abd Ur Rehman, Sung-Min Gho, Aleum Lee, Tariq Mahmood Khan, and Imran Razzak. Advancing metaverse-based healthcare with multimodal neuroimaging fusion via multi-task adversarial variational autoencoder for brain age estimation. IEEE Journal of Biomedical and Health Informatics, 30:1867{1875, 2024. URL https://api.semanticscholar.org/CorpusID:274282104.