Resolution-Conditioned Representation Learning for Unified Brain MRI Modeling

Ke Dong[1], Xihong Hu[2], Shijie Huang[1], Yanqing Xu[3], Zifeng Lian[1], Xiaoye Li[1], Liqin Yang[1], Yimin Fan[1], Dinggang Shen[1,4,5], and Kaicong Sun[1]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China
{dgshen,sunkc}@shanghaitech.edu.cn
[2] Department of Radiology, Children's Hospital of Fudan University, Shanghai 201102, China
[3] Department of Psychiatry and Psychology, Children's Hospital of Fudan University, Shanghai 201102, China
[4] Shanghai Clinical Research and Trial Center, Shanghai 201210, China
[5] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China

Abstract. Clinical magnetic resonance imaging (MRI) scans often exhibit substantial variability in slice thickness owing to heterogeneous imaging protocols and scan-time constraints. However, most deep learning models for brain MRI are developed under fixed-resolution assumptions, limiting their generalization across resolution shifts. To fill the gap, we propose a resolution-conditioned representation learning framework that can effectively handle scans with varying slice thicknesses. Our framework is pretrained via self-supervised reconstruction on 21,156 brain MRI scans from 14 public cohorts, and the learned representations are further refined through multimodal alignment with clinical metadata, including acquisition parameters and demographic characteristics. This alignment enables the model to accommodate diverse acquisition settings while preserving fine-grained anatomical information. The learned representations consistently support multiple clinical applications, such as brain tissue segmentation and diagnosis of autism spectrum disorder and major depressive disorder. Extensive experiments on public benchmarks and in-house clinical cohorts demonstrate that our framework achieves competitive segmentation performance against state-of-the-art models and exhibits superior generalization to external datasets. Notably, our method outperforms task-specific diagnosis models on real-world thick-slice MRI scans, highlighting its robustness under clinically realistic imaging conditions. These findings demonstrate the potential of resolution-conditioned representation learning to enable robust and versatile brain MRI analysis across heterogeneous slice thicknesses.

Keywords: Heterogeneous slice thickness · Multimodal fusion · Tissue segmentation · Disease diagnosis.

1 Introduction

Brain magnetic resonance imaging (MRI) plays a crucial role in precise morphological analysis, tissue segmentation, and neurological disease diagnosis [1, 2]. However,

K. Dong, X. Hu, and S. Huang contributed equally to this work.

acquiring high-resolution volumetric MRI is time-consuming and often impractical in routine clinical workflows. In real-world clinical settings, fast thick-slice MRI protocols are widely adopted, resulting in large volumes of clinical data that remain underutilized by existing AI models due to the limited through-plane structural details [6]. Therefore, developing a unified brain MRI framework that can robustly accommodate heterogeneous slice thicknesses is of substantial clinical importance. Such a framework could improve the reliability and clinical utility of automated neuroimaging analysis pipelines across diverse clinical settings.

Existing deep learning models for brain MRI are predominantly developed on thin-slice datasets [3–5], implicitly assuming consistent slice thicknesses across training and deployment settings. Consequently, their performance often degrades on thick-slice clinical MRI, where diminished through-plane resolution and reduced anatomical detail induce pronounced distributional shifts [7]. This resolution sensitivity remains an underexplored limitation in current neuroimaging models. Moreover, most prior approaches are tailored to specific downstream tasks, such as brain tissue segmentation [8] or disease diagnosis [9, 10]. Although recent unified radiological models such as OmniRad [11] support multiple tasks within a shared architecture, they primarily focus on 2D radiological images and coarse organ-level or pathology-level prediction tasks. As a result, they neither explicitly address slice-thickness heterogeneity in brain MRI nor support fine-grained brain neuroanatomical segmentation and mental disorder classification under varying imaging resolutions. In fact, MRI examinations are accompanied by rich metadata, including imaging parameters (i.e. slice thickness, in-plane spacing) and demographic information. Such metadata provides valuable contextual information for modeling imaging heterogeneity and mitigating domain shifts arising from multi-site protocols and diverse patient populations. Vision–language models such as CLIP [12] and BiomedCLIP [13] have demonstrated the effectiveness of cross-modal alignment in learning robust shared representations. However, metadata-guided representation learning remains largely underexplored in addressing resolution-induced heterogeneity in clinical neuroimaging, particularly when robust performance across diverse downstream tasks and imaging resolutions is required.

To address resolution heterogeneity in clinical brain MRI, we propose a resolution-conditioned representation learning framework that enables robust modeling across varying slice thicknesses. The framework is pretrained on 21,156 brain MRI scans from 14 public cohorts via large-scale self-reconstruction to learn anatomy-preserving structural representations. We further incorporate structured clinical metadata, including imaging parameters and demographic attributes, through cross-modal alignment to establish a resolution-aware and acquisition-informed embedding space. The resulting representations generalize across both spatially fine-grained and high-level semantic tasks, supporting brain tissue segmentation and mental disease diagnosis within a unified backbone. The main contributions are: 1) **Unified cross-resolution modeling:** We introduce a shared representation framework for brain MRI that supports both segmentation and disease diagnosis under heterogeneous slice thickness conditions; 2) **Resolution-conditioned representation learning:** We combine large-scale self-reconstruction pretraining with metadata-guided multimodal alignment to learn anatomy-preserving and resolution-aware features; 3) **Extensive multi-cohort valida-**

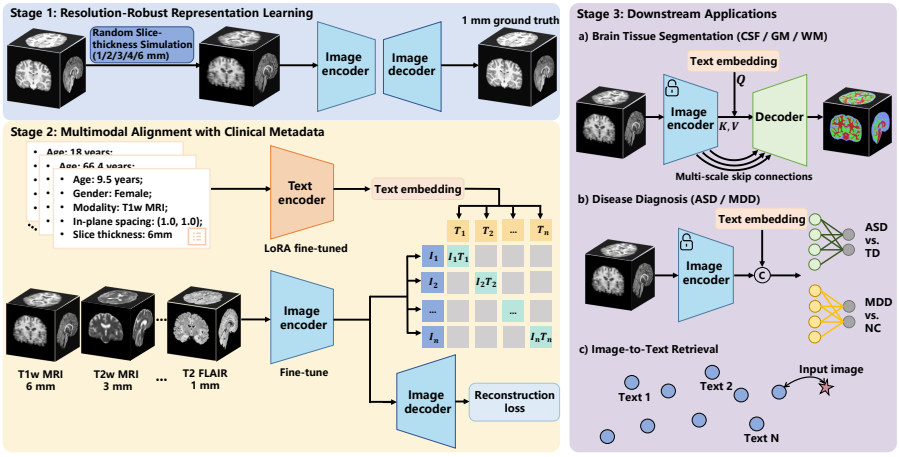

Fig. 1. Schematic illustration of our unified framework for brain MRI analysis supporting heterogeneous slice thickness : 1) Resolution-aware representation learning; 2) Multimodal alignment; and 3) Downstream applications.

tion: Our framework consistently matches or outperforms task-specific models while maintaining strong generalization across CSF/GM/WM segmentation, autism spectrum disorder (ASD) diagnosis, and major depressive disorder (MDD) diagnosis under both simulated and real-world thick-slice MRI settings.

2 Method

The overall framework is illustrated in Fig. 1. It consists of three stages: 1) Resolution-robust representation learning; 2) Multimodal alignment with clinical metadata; and 3) Downstream applications including brain tissue segmentation, multiple brain disease diagnosis, and image-to-text retrieval. The pretraining is performed in two separate stages to reduce the memory burden of jointly optimizing reconstruction and contrastive objectives, enabling more effective contrastive learning while preserving fine anatomical representations. More details are provided below.

2.1 Resolution-Robust Representation Learning

To learn representations robust to heterogeneous slice thicknesses, we adopt a resolution-conditioned self-reconstruction pretraining strategy. Specifically, thin-slice MRI volumes are downsampled along a randomly selected spatial axis using multiple scaling factors (1/2/3/4/6) to simulate clinically realistic slice-thickness variations, and are subsequently resampled to an isotropic resolution of $1\times1\times1$ mm before being fed into the encoder. By exposing the model to diverse degradation scenarios, the network is trained to reconstruct the original high-resolution isotropic volume from incomplete structural observations. The reconstruction network consists of a 3D ResNet-50 encoder [14] and

a lightweight convolutional decoder, optimized using a mean squared error (MSE) reconstruction loss. Encoder–decoder skip connections are deliberately omitted during pretraining to prevent direct copying of low-level spatial details from the input, thereby forcing reconstruction to rely primarily on the latent representation. This design encourages the encoder to capture anatomy-preserving features that are less sensitive to slice-thickness variations and more transferable across downstream classification and segmentation tasks.

2.2 Multimodal Alignment with Clinical Metadata

To enhance robustness under heterogeneous imaging conditions, we incorporate structured textual metadata that contains demographic attributes (i.e., age and gender) and imaging parameters, explicitly encoding resolution-related information (i.e. in-plane spacing and slice thickness). Image–text pairs are aligned via contrastive learning to project both modalities into a shared embedding space, encouraging the learning of resolution-aware representations. Each MRI volume and its corresponding textual metadata form a positive pair, while non-matching image–text pairs are treated as negative pairs.

To be specific, the image encoder is initialized with the weights from Stage 1, and during contrastive learning (Stage 2), only the final residual block of the image encoder is fine-tuned while the earlier layers remain frozen. The text encoder is initialized using the pretrained BiomedCLIP [13], and fine-tuned using Low-Rank Adaptation (LoRA). After independently encoding the textual and imaging inputs, their embeddings are projected into a shared 512-dimensional embedding space. Given a mini-batch of paired image and text embeddings $\{(v_i, t_i)\}_{i=1}^{B}$ with B being the batch size, the contrastive alignment loss is defined as:

$$\mathcal{L}_{align} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathrm{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(v_i, t_j)/\tau)}, \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. τ is a learnable parameter that controls the concentration of the similarity distribution and is initialized to $\log(1/0.07)$ following CLIP. To prevent structural drift during multimodal optimization, the self-reconstruction constraint defined in Stage 1 is retained in this stage.

2.3 Application to Downstream Tasks

Tissue Segmentation. Tissue segmentation requires fine-grained spatial information for accurate boundary delineation. During pretraining, encoder–decoder skip connections are intentionally omitted to encourage the encoder to learn compact and semantically meaningful representations. While such representations are beneficial for general-purpose representation learning, they do not explicitly preserve the multi-scale spatial information required for dense prediction tasks. Therefore, for downstream segmentation, we introduce a task-specific decoder with encoder–decoder skip connections to recover fine-grained spatial details and improve boundary localization.

Table 1. Summary of datasets used in this study.

Dataset	Modality	Age (years)	Gender (M/F/U)
Pretraining data [n=21,156]	T1w/T2w/FLAIR	0–100 (37.4 ± 27.9)	9,526/10,374/1,256
IXI [n=174]	T1w/T2w	20–82 (48.1 ± 16.6)	82/92
CBCP [n=275]	T1w	0–7 (3.7 ± 2.1)	–
ABIDE			
ASD [n=266]	T1w	7–58 (15.0 ± 8.1)	228/38
TD [n=262]	T1w	6–45 (15.6 ± 7.1)	234/28
CHFU-ASD			
ASD [n=1,173]	T1w/T2w/FLAIR	0–6 (3.2 ± 1.1)	947/226
TD [n=987]	T1w/T2w/FLAIR	0–7 (4.1 ± 1.4)	522/465
CHFU-MDD			
MDD [n=1,126]	T1w/T2w/FLAIR	9–17 (13.2 ± 1.6)	229/897
NC [n=1,107]	T1w/T2w/FLAIR	8–15 (11.1 ± 2.0)	554/553

Specifically, feature maps from the four encoder stages are concatenated with the corresponding decoder features through skip connections to facilitate multi-scale feature fusion. The decoder progressively reduces the channel dimensionality from 2048 to 1024, 512, and 256 through four stages, each consisting of trilinear upsampling followed by two 3D convolutional blocks. A final $1\times1\times1$ convolution layer is used to produce voxel-wise predictions. Besides, textual metadata embeddings are incorporated into the bottleneck features through a cross-attention module, allowing image representations to be adaptively modulated by acquisition parameters and demographic context. In Stage 3, the image and text encoders are frozen, and only the segmentation decoder is trained. The segmentation decoder is optimized using a composite loss consisting of Dice loss and Focal loss.

Disease Diagnosis. Disease diagnosis primarily relies on high-level structural and semantic patterns. The pretrained encoder therefore provides a suitable representation for diagnosis tasks. Specifically, the extracted image features are compressed via global average pooling, and then concatenated with textual metadata embeddings to reduce confounding effects from age, gender, and acquisition variability. The concatenated features are connected with a lightweight multilayer perceptron, which is optimized by a cross-entropy loss. Note that the image and text encoders are frozen, and only the classification head is trained.

3 Experiments and Results

3.1 Datasets and Data Preprocessing

Our framework is developed and evaluated on a large-scale multi-cohort brain MRI collection. Detailed data statistics are summarized in Table 1.

Pretraining datasets. We collected 21,156 thin-slice brain MRI scans from 14 public cohorts covering the full lifespan, from neonates to older adults. Neonatal scans were primarily obtained from dHCP [16] (n=1,575). Infant scans were obtained from BCP [15] (n=1,121) and NDAR [19] (n=915). Adolescent scans were obtained from ABCD [17] (n=2,208), ABIDE [26] (n=528), HCP [20] (n=498), and ADHD [18] (n=707). Adult and older-adult scans were obtained from HCP [20] (n=1,831), ADNI [22] (n=353), AIBL [21] (n=1,310), OASIS-3 [23] (n=2,637), and UK Biobank [24] (n=4,517). Each dataset was randomly split into training, validation, and test subsets using a 7:1:2 ratio. All images were processed using a standardized pipeline comprising bias-field correction, skull stripping, and rigid registration to the MNI template to ensure spatial consistency [28]. To simulate clinical thick-slice acquisitions, high-resolution volumes were downsampled along a randomly selected spatial axis using factors corresponding to typical clinical slice thicknesses, followed by resampling to an isotropic resolution of $1\times1\times1$ mm. To standardize the input size, all the data were cropped or padded to $192 \times 192 \times 192$.

Segmentation datasets. Segmentation evaluation was conducted on both internal and external datasets. The internal test set consisted of 20% of the pretraining dataset and spanned the full lifespan. Two public datasets, IXI [29] and CBCP [25], were used exclusively for external segmentation evaluation.

Disease diagnosis datasets. For disease diagnosis, we used the public ABIDE dataset (1-mm resolution) for ASD vs. typically developing (TD) classification, as well as two in-house clinical cohorts acquired at Children's Hospital of Fudan University (CHFU) using routine thick-slice MRI protocols (6–6.5 mm), including an ASD cohort (CHFU-ASD; ASD vs. TD) and an MDD cohort (CHFU-MDD; MDD vs. normal controls, NC). All clinical scans were processed using the same preprocessing pipeline without simulated downsampling.

3.2 Implementation Details

All models were implemented in PyTorch and trained on NVIDIA A100 GPUs with 80 GB RAM. The reconstruction network in Stage 1 was trained on a single GPU using Adam optimizer with an initial learning rate of 1×10^{-4} for 50 epochs. The mini-batch size was set to 1. The multimodal alignment in Stage 2 was performed on eight GPUs with a mini-batch size of 16, and the model was trained for 100 epochs using a cosine annealing learning-rate schedule.

3.3 Brain Tissue Segmentation

Internal evaluation on test dataset. For brain tissue segmentation, we compared our method with SwinUNETR [8], MedNeXt [30], and SynthSeg [7] under the same experimental settings. To assess robustness to slice-thickness variations, volumes were downsampled to simulate slice thicknesses of 1, 3, and 6 mm, and all methods were trained on the resulting mixed-resolution dataset. As shown in Table 2, although slightly inferior to

Table 2. Quantitative comparison of tissue segmentation under different slice-thickness settings (1/3/6 mm) on the internal test dataset.

Method	1 mm		3 mm		6 mm	
	Dice (%)	HD95 (mm)	Dice (%)	HD95 (mm)	Dice (%)	HD95 (mm)
MedNeXt	**93.3 ± 4.7**	**1.07 ± 0.38**	91.3 ± 4.2	1.12 ± 0.40	89.1 ± 4.0	1.18 ± 0.43
SwinUNETR	92.7 ± 5.0	1.11 ± 0.53	90.5 ± 4.5	1.12 ± 0.54	87.3 ± 4.1	1.35 ± 0.56
SynthSeg	66.2 ± 30.7	15.26 ± 30.18	64.8 ± 30.3	15.75 ± 30.92	60.3 ± 27.7	14.31 ± 27.34
Ours	92.4 ± 5.0	1.10 ± 0.43	**91.5 ± 4.9**	**1.11 ± 0.45**	**89.5 ± 3.5**	**1.14 ± 0.35**

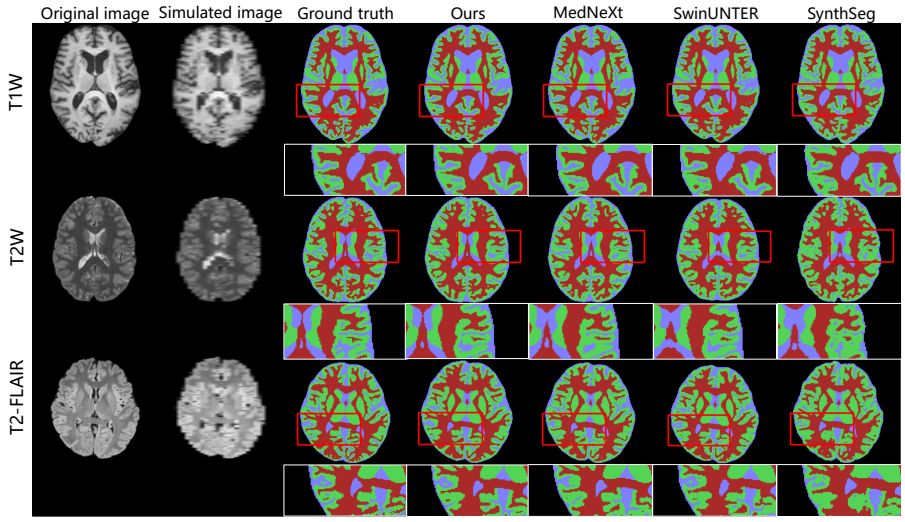

Fig. 2. Comparison of tissue segmentation results on MRI data with a simulated slice thickness of 6 mm.

MedNeXt at 1 mm, our method achieves competitive performance at 3 mm and obtains the highest Dice and the lowest HD95 under the clinically common 6 mm setting, with statistically significant improvements over other methods ($p < 0.05$). Moreover, the performance degradation of our model with increased slice thickness is smaller than the compared methods, indicating its robustness to resolution variation. Besides, we observe that SynthSeg exhibits several near-complete segmentation failures on neonatal and early developmental T2-weighted MRI scans, resulting in a lower average Dice score and higher standard deviation. This may be related to the substantial anatomical and contrast discrepancies between these scans and adult brain MRI.

External Evaluation on IXI and CBCP. We further evaluated the model generalizability on two unseen cohorts, IXI (adults) and CBCP (early childhood). We show that without fine-tuning, our model achieves the strongest zero-shot performance in Table 3, outperforming the SOTA methods significantly ($p < 0.05$).

Table 3. External validation on IXI and CBCP datasets with simulated slice thickness of 6 mm.

Method	IXI		CBCP	
	Dice (%)	HD95 (mm)	Dice (%)	HD95 (mm)
MedNeXt	87.9 ± 0.5	1.30 ± 0.05	85.0 ± 6.7	1.63 ± 0.58
SwinUNETR	86.1 ± 0.8	1.39 ± 0.10	84.3 ± 5.5	1.70 ± 0.56
SynthSeg	82.4 ± 0.7	1.50 ± 0.10	19.6 ± 2.7	36.86 ± 1.17
Ours	$\mathbf{88.4 \pm 0.4}$	$\mathbf{1.29 \pm 0.05}$	$\mathbf{85.9 \pm 5.8}$	$\mathbf{1.61 \pm 0.69}$

Table 4. Comparison of disease diagnosis performance on the ABIDE dataset (1 mm) and two in-house clinical datasets acquired using routine thick-slice MRI protocols (6–6.5 mm).

Dataset	Method	ACC	F1	AUC
ABIDE	3D ViT	0.679 ± 0.126	0.636 ± 0.140	0.740 ± 0.171
	Ours	$\mathbf{0.739 \pm 0.110}$	$\mathbf{0.724 \pm 0.143}$	$\mathbf{0.821 \pm 0.117}$
CHFU-ASD	3D ViT	0.861 ± 0.098	0.872 ± 0.095	0.916 ± 0.127
	Ours	$\mathbf{0.907 \pm 0.087}$	$\mathbf{0.912 \pm 0.089}$	$\mathbf{0.974 \pm 0.056}$
CHFU-MDD	3D ViT	0.792 ± 0.096	0.770 ± 0.188	0.849 ± 0.141
	Ours	$\mathbf{0.824 \pm 0.129}$	$\mathbf{0.834 \pm 0.123}$	$\mathbf{0.917 \pm 0.106}$

3.4 Disease Diagnosis

We further evaluated the learned representations for disease diagnosis using three independent cohorts: the public ABIDE dataset and in-house CHFU-ASD dataset for ASD vs. TD classification, and in-house CHFU-MDD dataset for MDD vs. NC classification. Performance was assessed using ACC, F1, and AUC (mean ± std over five-fold cross-validation). As shown in Table 4, our method consistently outperforms 3D ViT [31] across all the datasets. Notably, superior performance on the private thick-slice cohorts highlights enhanced robustness under real-world clinical imaging conditions.

3.5 Image-to-Text Retrieval

To evaluate the effectiveness of multimodal alignment, we conducted an image-to-text retrieval experiment in the shared embedding space. We randomly sampled 100 scans from the test dataset and generated five slice-thickness variants (1/2/3/4/6 mm) along the axial axis per scan, resulting in 500 image–text pairs. Retrieval is performed by ranking all candidate texts according to cosine similarity with the image embedding. Our proposed framework achieves a Top-1 retrieval accuracy of 37% and a Top-5 retrieval accuracy of 58%, indicating effective alignment across heterogeneous slice-thickness settings.

3.6 Ablation Study

We conducted ablation studies on simulated 6-mm slice-thickness data derived from the internal test set by removing skip connections (No Skip), replacing text with “unknown” tokens (Text-Unknown), and removing multimodal alignment (Image-Only). Results in

Table 5. Ablation study of brain tissue segmentation on the internal test dataset.

Metric	No Skip	Text-Unknown	Image-Only	Ours
Dice (%)	84.9 ± 3.7	87.9 ± 4.0	88.3 ± 3.9	$\mathbf{89.5 \pm 3.5}$
HD95 (mm)	1.39 ± 0.54	1.19 ± 0.43	1.19 ± 0.57	$\mathbf{1.14 \pm 0.35}$

Table 5 show a noticeable performance drop for all the variants. Our model achieves the highest Dice and lowest HD95, confirming the effectiveness of the skip connections, multimodal alignment, and textual metadata in improving segmentation performance.

4 Conclusions

In this work, we presented a unified multimodal framework for brain MRI analysis that is robust to heterogeneous slice thicknesses, particularly in real-world clinical settings. By learning resolution-aware representations and integrating acquisition-specific metadata, our framework supports a wide range of downstream tasks including brain tissue segmentation, brain disease diagnosis, and image-to-text retrieval across heterogeneous slice thicknesses and diverse demographic characteristics. Experimental results demonstrate that our framework achieves promising performance on external datasets without finetuning, demonstrating strong generalization ability. Moreover, our model shows superior diagnosis performance over the other methods on real-world clinical data, revealing the effectiveness of the learned representations. Overall, our framework provides a foundation for robust brain MRI analysis across heterogeneous slice thicknesses in real-world clinical workflows.

Acknowledgments. This work was supported in part by National Key Research and Development Program of China (grant number 2024YFC2707604), National Natural Science Foundation of China (grant numbers 62301318, U23A20295, 62131015, 82441023), the China Ministry of Science and Technology (STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), The Key Area Research and Development Program of Guangdong Province (grant number 2023B0303040001), and HPC Platform of ShanghaiTech University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI, pp. 424–432 (2016)
2. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
3. Sun, K., Zhang, Y., Liu, J., Yu, L., Zhou, Y., Xie, F., Guo, Q., Zhang, H., Wang, Q., Shen, D.: Achieving multi-modal brain disease diagnosis performance using only single-modal images through generative AI. Communications Engineering **3**, 96 (2024). doi:10.1038/s44172-024-00245-w

4. Basaia, S., Sarasso, E., Sciancalepore, F., Balestrino, R., Musicco, S., Pisano, S., Stankovic, I., Tomic, A., De Micco, R., Tessitore, A., Salvi, M., Meiburger, K.M., Kostic, V., Molinari, F., Agosta, F., Filippi, M.: Multi-center 3D CNN for Parkinson's disease diagnosis and prognosis using clinical and T1-weighted MRI data. NeuroImage: Clinical **48**, 103859 (2025). https://doi.org/10.1016/j.nicl.2025.103859
5. Praveen, M.A., Evuri, N., Pakala, S.R., Samantula, S., Chebrolu, S.: 3D Swin-Res-SegNet: A hybrid transformer and CNN model for brain tumor segmentation using MRI scans. J. Inst. Eng. (India): Series B **106**, 1489–1499 (2025). doi:10.1007/s40031-024-01166-0
6. Hu, J., Qin, Y., Wang, H., Han, J.: MS2CAM: Multi-scale self-cross-attention mechanism-based MRI super-resolution. Displays **88**, 103033 (2025). https://doi.org/10.1016/j.displa.2025.103033
7. Billot, B., Greve, D.N., Puonti, O., et al.: SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. Medical Image Analysis **86**, 102789 (2023)
8. Hatamizadeh, A., Tang, Y., Nath, V., et al.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: CVPR, pp. 574–584 (2022)
9. Pradhan, S., Chakraborty, B., Mukherjee, P.: A data-driven multimodal deep learning approach for autism spectrum disorder detection using brain MRI images. In: 2025 IEEE 17th International Conference on Computational Intelligence and Communication Networks (CICN), Goa, India, pp. 252–259. IEEE (2025). https://doi.org/10.1109/CICN67655.2025.11368281
10. Wang, L., Shu, C., Wu, Y., Yang, J.: 3D Res-Spnet: 3D ResNet with heterogeneous branches and spatial attention for MDD diagnosis via sMRI. In: Proceedings of the 2nd International Conference on Smart Healthcare and Wearable Intelligent Devices (SHWID), pp. 193–198. ACM (2025). https://doi.org/10.1145/3788112.3788142
11. Zedda, L., Loddo, A., Di Ruberto, C.: OmniRad: A Radiological Foundation Model for Multi-Task Medical Image Analysis. In: arXiv preprint, arXiv:2602.04547 (2026). https://doi.org/10.48550/arXiv.2602.04547
12. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)
13. Zhang, Y., Jiang, Z., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from large-scale image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Howell, B.R., Styner, M.A., Gao, W., et al.: The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development. NeuroImage **185**, 891–905 (2019)
16. Makropoulos, A., Robinson, E.C., Schuh, A., et al.: The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. NeuroImage **173**, 88–112 (2018)
17. Casey, B.J., Cannonier, T., Conley, M.I., et al.: The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. Developmental Cognitive Neuroscience **32**, 43–54 (2018)
18. Bellec, P., Chu, C., Chouinard-Decorte, F., et al.: The neuro bureau ADHD-200 preprocessed repository. NeuroImage **144**, 275–286 (2017)
19. Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., et al.: Brain volume findings in 6-month-old infants at high familial risk for autism. American Journal of Psychiatry **169**(6), 601–608 (2012)
20. Van Essen, D.C., Smith, S.M., Barch, D.M., et al.: The WU-Minn Human Connectome Project: An overview. NeuroImage **80**, 62–79 (2013)

21. Ellis, K.A., Bush, A.I., Darby, D., et al.: The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International Psychogeriatrics **21**(4), 672–687 (2009)
22. Petersen, R.C., Aisen, P.S., Beckett, L.A., et al.: Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. Neurology **74**(3), 201–209 (2010)
23. LaMontagne, P.J., Benzinger, T.L.S., Morris, J.C., et al.: OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. MedRxiv (2019)
24. Sudlow, C., Gallacher, J., Allen, N., et al.: UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine **12**(3), e1001779 (2015)
25. Xu, X., et al.: Behavioral observation and assessment protocol for language and social-emotional development study in children aged 0–6: The Chinese Baby 700 connectome project. BMC Psychology **12**, 533 (2024)
26. Di Martino, A., et al.: Enhancing studies of the connectome in autism using the Autism Brain Imaging Data Exchange II. Scientific Data **4**, 170010 (2017)
27. National Alzheimer's Coordinating Center (NACC): NACC database. https://naccdata.org/ (accessed 2024)
28. Smith, S.M., et al.: Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage **23**(Suppl. 1), S208–S219 (2004)
29. IXI Dataset, https://brain-development.org/ixi-dataset/, accessed 2025/02/20
30. Roy, S., Koehler, G., Ulrich, C., et al.: MedNeXt: Transformer-driven scaling of convnets for medical image segmentation. In: MICCAI 2023. LNCS, vol. 14223, pp. 405–415. Springer (2023)
31. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. In: ICLR (2021)