Masked Contrastive Language-Image Modeling For Brain Segmentation

Jianwen Liang[1], Junyan Lyu[1,2⋆], Yixuan Yuan[3], and Xiaoying Tang[1,4] (✉)

[1] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
tangxy@sustech.edu.cn
[2] Queensland Brain Institute, The University of Queensland, Brisbane, Australia
[3] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[4] Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China

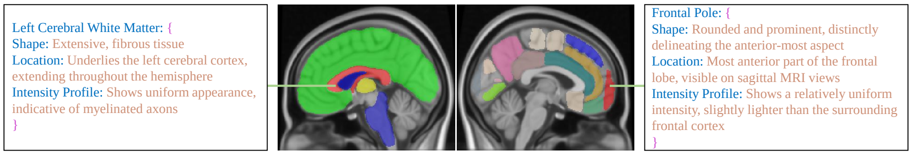
Abstract. Self-supervised learning (SSL) has emerged as a powerful paradigm to mitigate neuroimaging analysis algorithms' reliance on annotated data. However, existing SSL methods for brain MRI often fail to incorporate anatomical priors inherent in brain MRI, limiting their effectiveness. Here, we present Masked Contrastive Language-Image Modeling (MCLIM), a novel SSL framework that integrates knowledge from brain atlases through text-guided representation learning. We first generate structure-specific textual descriptors based on brain atlases, with no need for manually collecting image-text pairs. Then MCLIM employs (1) an image restoration branch that reconstructs randomly masked image patches through an encoder-decoder network, and (2) a cross-modal alignment module that establishes semantic correspondences between image features and atlas-derived text embeddings. These two learning objectives enable the simultaneous capture of fine-grained intensity patterns and whole-brain topological relationships. The proposed method is fine-tuned and evaluated on three brain parcellation datasets with varying granularities and a brain lesion segmentation dataset. Experiment results demonstrate that MCLIM outperforms state-of-the-art SSL methods and reduces annotation effort by at least 40%. Code and pre-trained models will be available at https://github.com/CRazorback/MCLIM.

Keywords: Self-supervised learning · Masked contrastive language-image modeling · Brain segmentation.

1 Introduction

Accurate segmentation of brain structures and lesions from magnetic resonance imaging (MRI) is crucial for quantitative neuroimaging analysis [7,18]. Recent advances in supervised deep learning methods have demonstrated exceptional

⋆ J. Liang and J. Lyu contributed equally to this work.



Left Cerebral White Matter: {
Shape: Extensive, fibrous tissue
Location: Underlies the left cerebral cortex, extending throughout the hemisphere
Intensity Profile: Shows uniform appearance, indicative of myelinated axons
}

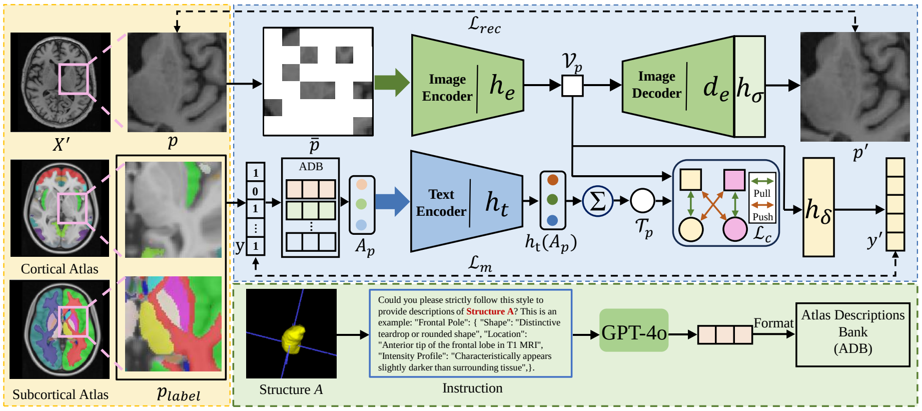
Frontal Pole: {
Shape: Rounded and prominent, distinctly delineating the anterior-most aspect
Location: Most anterior part of the frontal lobe, visible on sagittal MRI views
Intensity Profile: Shows a relatively uniform intensity, slightly lighter than the surrounding frontal cortex
}

Fig. 1. Texture descriptions of a cortical structure and a sub-cortical structure.

performance in brain segmentation [15,22]. Nevertheless, these methods heavily rely on extensively annotated datasets, and the annotation process for brain MRI remains challenging due to the complexity of brain anatomy [12]. In recent years, self-supervised learning (SSL) has emerged as a promising paradigm to address this limitation [4,19,24]. Existing evidence suggests that SSL methods based on image restoration can enhance the downstream segmentation performance for medical images [3,5,25]. However, the intensity profiles across regions in brain MRI exhibit a certain level of similarity [17], which constrains the efficacy of general SSL approaches.

Recent studies have attempted to integrate domain-specific priors to address the aforementioned limitations [10,14]. For instance, MDM [14] leverages brain anatomical priors by predicting atlas-to-subject deformation fields. However, such approaches cannot precisely capture fine-grained anatomical details. Given the rich semantic information of textual data, contrastive language-image pre-training (CLIP) demonstrates superior image understanding capability [16]. Medical CLIP variants [6,20] trained on large-scale medical datasets further enhance anatomical feature representation through multimodal alignment. Integrating image restoration with cross-modal alignment shall make a model learn fine-grained visual patterns and spatial distributions. However, this requires large-scale paired image-text datasets that remain unavailable in neuroimaging research. Notably, as illustrated in Fig. 1, brain atlases contain a wealth of detailed neuroanatomical information where brain structures can be precisely represented through comprehensive textual description. Registration enables atlas-based text generation for image patches by normalizing individual scans to a standard space. Although lacking pixel-level precision, this approach still achieves sufficient anatomical localization fidelity for structural description.

In this work, we propose a novel SSL framework for brain MRI that integrates atlas-derived textual descriptors to enhance the semantic representation capacity of vision models. Our method comprises three core components: (1) a masked image reconstruction branch that learns local intensity patterns through image restoration; (2) a cross-modal alignment module establishing semantic correspondences between masked image patches and neuroanatomical text embeddings; (3) a global structure matching mechanism that enhances the aforementioned alignment. With masked image patches, the image encoder is compelled to comprehensively capture the spatial distribution of brain structures to achieve alignment with textual features. Thus, the vision model can jointly learn the brain's spatial distribution patterns and anatomical semantics.



Could you please strictly follow this style to provide descriptions of **Structure A**? This is an example: "Frontal Pole": { "Shape": "Distinctive teardrop or rounded shape", "Location": "Anterior tip of the frontal lobe in T1 MRI", "Intensity Profile": "Characteristically appears slightly darker than surrounding tissue",}.

Fig. 2. Illustration of the proposed SSL framework.

To our knowledge, this is the first SSL approach that systematically incorporates neuroanatomical text priors without requiring paired image-text training datasets. Our method demonstrates superior performance in segmenting brain structures and lesions, evidenced by its state-of-the-art (SOTA) results across four benchmark datasets, while significantly reducing annotation costs.

2 Methodology

We propose an SSL framework for brain MRI segmentation that leverages textual descriptions of brain structures. We provide an overview of the pipeline in Fig. 2, consisting of the following steps: (1) an input image is affine registered to the MNI152 space, and text is generated for every brain structure; (2) the affine registered image is randomly cropped and masked, followed by joint training of a text encoder and an image encoder-decoder network for cross-modality alignment while reconstructing the image patch of interest.

2.1 Atlas Descriptions Bank

While natural language supervision has proven effective for enhancing visual representation learning in the natural image domain, its application to 3D brain MRI remains unexplored. This gap is primarily due to the absence of large-scale paired MRI-text datasets. Given that brain atlases encapsulate expert-level knowledge, where the delineation of brain structures in atlases is closely tied to the structures' anatomical locations and neurological functions, we leverage brain atlases to generate detailed textual descriptions for brain structures. Specifically, for a given brain structure A in the Harvard-Oxford cortical and subcortical structural atlases [9], we utilize GPT-4o [1] following the instruction shown in Fig. 2 to generate its precise **shape description**, **anatomical**

location, and **intensity profile**. These descriptions are then formatted using a predefined template: "*This is a T1 weighted human brain MRI patch including A, its shape is <* ***shape description*** *>, it <* ***anatomical location*** *>, it <* ***intensity profile*** *>.*" to create structured textual representations of structure A, which are subsequently stored in the Atlas Descriptions Bank (ADB).

2.2 Masked Contrastive Language-Image Modeling

Text Generation For Image Patches. Since brain MRI exhibits strong anatomical consistency across individuals, using full-volume inputs for pre-training would limit text descriptions' diversity. To address this, we employ a patch-wise sampling strategy to ensure that different image patches contain diverse brain structures, significantly enhancing the text descriptions' diversity. First, we align a brain MRI scan X to the MNI152 standard space wherein the atlas resides, obtaining X'. Subsequently, we perform random sampling on X' to get an image patch: $p = crop(X')$, and apply the same sampling to the atlas for identifying the brain structures contained in p, denoted as: $p_{label} = crop(altas)$. Based on p_{label}, we retrieve the corresponding textual descriptions from the pre-constructed ADB.

Text Embedding. We utilize the same text encoder architecture h_t as that in BiomedCLIP [23] to encode the textual descriptions and the pre-trained weights of the text encoder released by BiomedClip are used to initialize h_t. Since these weights are pre-trained on PMC-15M [23], a large-scale medical dataset, using them for initialization makes h_t possess a certain level of understanding for MRI images and the brain. However, directly concatenating the textual descriptions of all brain structures within p and inputting them to h_t lead to collapse, which we attribute to a main reason: the text encoder typically used for training CLIP imposes constraints on the text token's length [16]. The average sequence length of the textual data in PMC-15M is 110 tokens while p containing multiple brain structures may result in excessively long text sequences. The discrepancy in text sequence length between pre-training and downstream training leads to architectural incompatibility, consequently inducing training instability. To overcome this, we adopt a prototype-based approach to represent the text features corresponding to p. Specifically, for each brain structure within p, we retrieve its corresponding description A_p from ADB and compute its feature embedding: $h_t(A_p)$. The final text feature representation for p is then obtained by aggregating these embeddings as follows:

$$\mathcal{T}_p = \sum_{A_p \in p} h_t(A_p). \tag{1}$$

Image Embedding. We incorporate an image restoration task into our SSL framework to learn the MRI's intensity patterns. Specifically, we randomly mask 75% of the regions in p to obtain $\bar{p}$, which is then processed by an image encoder of a U-shape network [8] to extract latent features: $\mathcal{V}_p = h_e(\bar{p})$.

Pre-training for MCLIM. The model is tasked with two primary objectives. The first objective is to reconstruct the original image patch from the latent features. An image decoder d_e and an intensity head h_σ are used to predict the voxel intensity values: $p' = h_\sigma(d_e(\mathcal{V}_p))$. This process encourages the model to learn the intensity distribution of the brain MRI. The model is trained to minimize the difference between the restored patch and the original patch. Specifically, an $l2$-loss is employed on the masked voxels

$$\mathcal{L}_{rec}(p, p') = \frac{1}{\mathit{\Omega}(p_m)} ||p_m - {p'}_m||^2, \tag{2}$$

where p_m is the masked voxels in the original patch, and $\mathit{\Omega}(p_m)$ is the total number of masked voxels. The second objective is to align the masked image patch with the priors of the brain. We align the image features with the texture features by the contrastive objective

$$\mathcal{L}_c^{T\leftarrow I}(\mathcal{T}_p, \mathcal{V}_p) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{exp(\mathcal{T}_p^i \cdot {\mathcal{V}_p^i}^T/\tau)}{\sum_{j=1}^{B} exp(\mathcal{T}_p^i \cdot {\mathcal{V}_p^j}^T/\tau)}, \tag{3}$$

$$\mathcal{L}_c^{I\leftarrow T}(\mathcal{T}_p, \mathcal{V}_p) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{exp(\mathcal{T}_p^i \cdot {\mathcal{V}_p^i}^T/\tau)}{\sum_{j=1}^{B} exp(\mathcal{T}_p^j \cdot {\mathcal{V}_p^i}^T/\tau)}, \tag{4}$$

$$\mathcal{L}_c = \mathcal{L}_c^{T\leftarrow I} + \mathcal{L}_c^{I\leftarrow T}, \tag{5}$$

where B is the batch size and τ is the temperature parameter. Before computing the contrastive loss, both $\mathcal{T}_p$ and $\mathcal{V}_p$ are normalized using the $l2$ norm: $\frac{\mathcal{T}_p}{||\mathcal{T}_p||_2}, \frac{\mathcal{V}_p}{||\mathcal{V}_p||_2}$. The masked image patch enhances the model's learning capacity, as the image encoder must develop a comprehensive understanding of the brain's spatial architecture to achieve effective cross-modal alignment. Furthermore, we enhance the masked image modeling process with the guidance of language, which enables the visual model to capture robust anatomical priors regarding the brain's relatively fixed structural relationships. This ability is particularly crucial for segmenting cortical structures that exhibit similar intensity profiles but of different positions. To enhance the correlation between the image representations and the brain anatomy's characteristics, a matching head is employed to match $\mathcal{V}_p$ with the structure label y by the binary cross entropy loss

$$\mathcal{L}_m(y, y') = -\frac{1}{N}\sum_{A\in p} y_A log{y'}_A, \tag{6}$$

where N is the number of structures in the atlases, $y_A \in y$ indicates whether structure A is in p, and $\{y'_A = h_\delta(\mathcal{V}_p) | y'_A \in y'\}$ is the corresponding prediction probability. The framework is pre-trained by optimizing the overall loss

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_m + \mathcal{L}_c. \tag{7}$$

For downstream segmentation tasks, we transfer the pre-trained parameters of h_e and d_e to the segmentation model and replace the intensity head h_σ with a segmentation head h_ω.

Table 1. Quantitative DSC comparisons with SOTA SSL methods on brain segmentation. Datasets A, B, C, and D are respective Mindboggle-101, CADNI, JHU, and ATLAS2. Top 1 results are highlighted in **bold**.

Methods	Structure				Lesion	Average
	A	B	C	Average	D	
Scratch	76.71	86.86	80.24	81.27	52.13	73.98
Model Genesis [25]	76.78	86.58	80.25	81.20	52.44	74.01
ASA [10]	79.01	86.96	80.30	82.09	53.50	74.94
PCRLv2 [24]	79.63	86.95	80.91	82.49	56.36	75.96
Tang et al. [19]	79.41	87.10	80.95	82.48	57.16	76.15
MDM [14]	79.76	**87.20**	81.39	82.78	57.56	76.47
MCLIM	**80.54**	87.17	**81.97**	**83.22**	**58.53**	**77.04**

3 Experiments and Results

3.1 Datasets and Implementation

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset is used for pre-training in this work. Specifically, we select 5,714 T1-weighted MRI scans from 818 subjects, adhering to the criteria of isotropic resolution and a slice thickness of less than 1.2 mm. All images are affine registered to the MNI152 template using ANTs [2]. We evaluate MCLIM on three brain parcellation datasets: Mindboggle-101 [12], CANDI [11], and JHU [21]. These datasets exhibit variability in segmentation complexity, with the number of brain structures ranging from 30 to 289, and sample sizes varying between 37 and 101. We also evaluate MCLIM on ATLAS2 [13], a brain lesion segmentation dataset with a sample size of 1,271. Following MDM [14], we randomly split each dataset into training and testing sets at a ratio of 60%/40%. To ensure consistent intensity normalization, the intensity values of all images are scaled to the range $[0, 1]$ based on the 1st and 99th percentiles, followed by zero-mean and unit-variance normalization.

Foreground image patches are randomly cropped to $96 \times 96 \times 96$. Random rotation, gamma correction, and random scaling are applied to enlarge the training set. All experiments are conducted using PyTorch 1.13.1 with NVIDIA RTX A6000 GPUs. The temperature coefficient τ is set as 0.05. The AdamW optimizer with a cosine learning rate scheduler is employed to optimize the pre-training and fine-tuning objectives. For pre-training, we train the model for 200 epochs with 10 warm-up epochs. For fine-tuning, due to varying numbers of training samples across datasets, we train the model for 25,000 iterations with 400 warm-up iterations for all downstream datasets. The batch size is set to 48 and 2 and the initial learning rate is 1e-4 and 4e-4 for pre-training and fine-tuning.

3.2 Evaluation Results

Comparisons with SOTA. For a fair comparison purpose, all evaluated methods are pre-trained from scratch according to their published code with the same

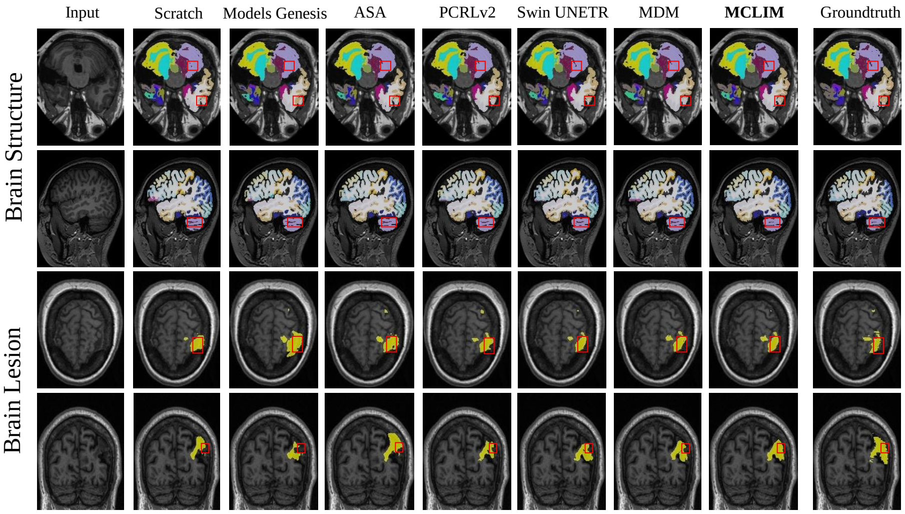


Fig. 3. Qualitative results of MCLIM and comparative methods on representative images. Regions with visual improvements are highlighted with bounding boxes.

network architectures and datasets. We quantitatively measure the segmentation performance using the Dice similarity coefficient (DSC) in all experiments. We denote model training from scratch as 'scratch' in subsequent sections. As summarized in Table 1, our MCLIM obtains the highest average DSC of 83.22% on brain parcellation, with an improvement of 0.44% over the previous best method, i.e., MDM. Across the three brain parcellation datasets, MCLIM demonstrates significant superiority over SOTA methods on two datasets, except that MDM achieves a marginal advantage over the proposed approach on the CADNI dataset. For brain lesion segmentation, MCLIM also obtains the best performance with the mean DSC of 58.53%, achieving an improvement of 0.97% over the previous best method. Collectively, MCLIM obtains the best average DSC across all four datasets, demonstrating its robustness. Qualitative comparisons in Fig. 3 show that our method demonstrates superior discrimination capability in segmenting the temporal lobe and the boundary of the dentate nucleus over comparative approaches while achieving notably lower false positive rates in lesion segmentation. These experimental results demonstrate that the proposed method has effectively learned the medical prior of the human brain.

Ablation Studies. We evaluate the effectiveness of each component in MCLIM. As shown in Table 2, learning the medical priors of the brain improves the downstream segmentation performance of brain structures and lesions over scratch by 1.95% and 6.40% in DSC (the 1^{st} line *vs.* the *last* line). Meanwhile, integrating medical knowledge can improve the mask image modeling (MIM) baseline (the 2^{nd} line *vs.* the *last* line). We also observe that the performance improve-

Table 2. Ablation analysis for the loss terms. Datasets A, B, C, and D are respective Mindboggle-101, CADNI, JHU, and ATLAS2. Top 1 results are highlighted in **bold**.

L_{rec}	L_m	L_c	Structure				Lesion	Average
			A	B	C	Average	D	
Scratch			76.71	86.86	80.24	81.27	52.13	73.98
✓			77.31	86.93	81.33	81.85	54.78	75.08
	✓	✓	77.27	87.14	81.06	81.82	56.32	75.44
✓	✓		77.55	87.09	81.42	82.02	56.27	75.58
✓		✓	79.52	87.11	81.65	82.76	58.09	76.59
✓	✓	✓	**80.54**	**87.17**	**81.97**	**83.22**	**58.53**	**77.04**

ment over MIM mainly comes from L_c rather than L_m (the 4^{th} line *vs.* the 5^{th} line), indicating that the supervision signals in the form of language provide the image encoder with a deeper understanding of brain anatomy. We also evaluate MCLIM's data efficiency power. As shown in Fig. 4, compared to scratch, MCLIM can effectively reduce the annotation effort by at least 40%.

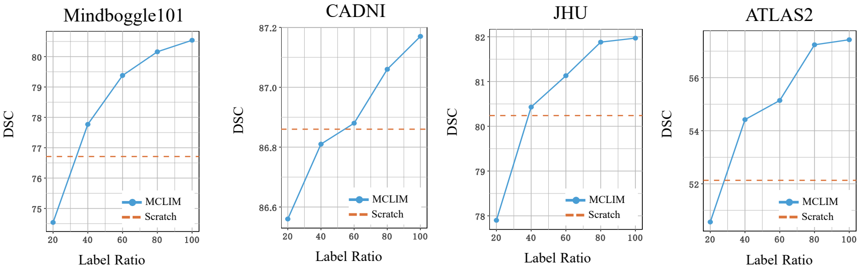


Fig. 4. Segmentation results of MCLIM with different annotation rates. The dotted line represents the performance of scratch with 100% training data.

4 Conclusion

In this paper, we propose MCLIM, a novel SSL framework for brain segmentation. Guided by brain atlases, MCLIM generates text descriptions for brain structures without requiring paired image-text training data. A contrastive language-image loss and a brain structure matching loss are integrated into the context restoration SSL framework to enhance the learning capability of the image encoder. We successfully demonstrate that MCLIM outperforms SOTA methods in brain segmentation while significantly reducing annotation costs. Furthermore, we validate that the success of our method primarily stems from text-based supervision signals. Future work shall include enhancing MCLIM by incorporating real-world data and extending it to brain classification tasks.

Acknowledgment. This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (T2422012, 62071210); the Guangdong Basic and Applied Basic Research (2024B1515020088); the Shenzhen Science and Technology Program (RCYX20210609103056042); the High Level of Special Funds (G030230001, G03034K003).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ants). Insight j **2**(365), 1–35 (2009)
3. Cai, Z., Lin, L., He, H., Cheng, P., Tang, X.: Uni4eye++: A general masked image modeling multi-modal pre-training framework for ophthalmic image classification and segmentation. IEEE Transactions on Medical Imaging (2024)
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
5. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3d medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1970–1980 (2023)
6. Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., Tang, X.: Prior: Prototype representation joint learning from medical images and reports. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21361–21371 (2023)
7. Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O.: Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. Hippocampus **19**(6), 579–587 (2009)
8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
9. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al.: An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage **31**(3), 968–980 (2006)
10. Huang, J., Li, H., Li, G., Wan, X.: Attentive symmetric autoencoder for brain mri segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 203–213. Springer (2022)
11. Kennedy, D.N., Haselgrove, C., Hodge, S.M., Rane, P.S., Makris, N., Frazier, J.A.: Candishare: a resource for pediatric neuroimaging data. Neuroinformatics **10**, 319–322 (2012)

12. Klein, A., Hirsch, J.: Mindboggle: a scatterbrained approach to automate brain labeling. NeuroImage **24**(2), 261–280 (2005)
13. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
14. Lyu, J., Bartlett, P.F., Nasrallah, F.A., Tang, X.: Masked deformation modeling for volumetric brain mri self-supervised pre-training. IEEE Transactions on Medical Imaging (2024)
15. Lyu, J., Xu, P., Nasrallah, F., Tang, X.: Learning ontology-based hierarchical structural relationship for whole brain segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 385–394. Springer (2023)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
17. Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T.: Evaluating intensity normalization on mris of human brain with multiple sclerosis. Medical image analysis **15**(2), 267–282 (2011)
18. Tang, X., Qin, Y., Wu, J., Zhang, M., Zhu, W., Miller, M.I.: Shape and diffusion tensor imaging based integrative analysis of the hippocampus and the amygdala in alzheimer's disease. Magnetic resonance imaging **34**(8), 1087–1099 (2016)
19. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
20. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
21. Wu, D., Ma, T., Ceritoglu, C., Li, Y., Chotiyanonta, J., Hou, Z., Hsu, J., Xu, X., Brown, T., Miller, M.I., et al.: Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on t1-weighted mri. Neuroimage **125**, 120–130 (2016)
22. Wu, J., Tang, X.: Brain segmentation based on multi-atlas and diffeomorphism guided 3d fully convolutional network ensembles. Pattern Recognition **115**, 107904 (2021)
23. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
24. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: Pcrlv2: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. arXiv preprint arXiv:2301.00772 (2023)
25. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)