Supplementary Materials of Domain Aware Multi-Task Pretraining of 3D Swin Transformer for T1-weighted Brain MRI

This Supplementary Materials provide additional details not included in the main paper. In Sec. A, we provide details about the several datasets we employed. Sec. B includes information on training details and network hyperparameters. Sec. C details the pretraining task. Finally, Sec. D details the results of pretraining tasks.

A Datasets

To pretrain and evaluate our proposed methods, we utilized 13,687 samples from several large-scale T1 structural MRI databases, including ADNI, HCP, IXI, ABIDE, DOD ADNI, ICBM, and A4. We further employed four datasets for the model assessment: ADNI, AIBL, OASIS, and PPMI. These datasets are independent of the datasets utilized for pretraining and were considered solely for evaluation.

Alzheimer's Disease Neuroimaging Initiative (ADNI) ADNI [41,61,67] is a research database dedicated to collecting multi-modal neuroimaging (MRI, fMRI, PET, and DTI) and non-imaging data (clinical outcome and genotyping data) related to AD. In our study, we obtained a total of 10,169 T1-weighted MR images from ADNI. These images encompass longitudinal data, different field strengths (1.5T and 3T), and scans from various manufacturers (Philips, Siemens, and GE). Of these images, we employed 8,300 images for pretraining phase and reserved 1,869 images for model evaluation. These data underwent preprocessing [17] and were employed to pretrain the model. For downstream tasks, we utilized a dataset of 1,869 samples, comprising CN: 639, MCI: 886, and AD: 344, for model evaluation.

Human Connectome Project (HCP) HCP [85] is a large-scale initiative aimed at comprehensively mapping the neural connections within the human brain. In our study, a total of 1,104 MR images were acquired. The following parameters were considered to acquire MR scans: manufacturer = Siemens, field strength=3T, TR = 2400 ms, TE = 2.14 ms, Flip angle = 8 degrees, FOV = $224 \times 224 mm^2$, Matrix size = 256×256, and Voxel size = $0.7 \times 0.7 \times 0.7 mm^3$. These data were used to pretrain the model.

Information eXtraction from Images (IXI) IXI [6] contains 581 MR images from healthy participants. These images include various MR scan types such as T1, T2, PD-weighted, MRA, and DWI. The T1-weighted images are

available in two field strengths (1.5T and 3T), and were scanned by different manufacturers (Philips, Siemens, and GE). We employed all these 581 images for pretraining.

Autism Brain Imaging Data Exchange (ABIDE) ABIDE [20, 21] contains 1099 MR images from Autism Spectrum Disorder (ASD) and control. These images include various MR scan types such as T1, resting state fMRI, and DWI. The T1-weighted images are available in two field strengths (1.5T and 3T), and were scanned by different manufacturers (Philips, Siemens, and GE). We used all these 1099 images for pretraining.

Effects of TBI & PTSD on Alzheimer's Disease in Vietnam Vets (DOD ADNI) DOD ADNI [88] focuses on exploring potential connections between traumatic brain injury (TBI), post-traumatic stress disorder (PTSD), MR scans, including longitudinal data. These T1-weighted scans were taken at two field strengths: 1.5T and 3T. The parameters for the 3T scanner were as follows: TR/TE = 2300/2.98ms, TI = 900ms, Flip angle = 9°, with a $1\times1\times1.2mm^3$ voxel size and 256×256 matrix over 170 slices. For the 1.5T scanner, they are: TR/TE = 2400/3.16ms, TI = 1000ms, Flip angle = 8°, with a $1.25\times1.25\times1.2mm^3$ voxel size and 256×256 matrix over 170 slices.

International Consortium for Brain Mapping (ICBM) ICBM [60] consists of 344 MRI images. These images were acquired axially in a 3D type using a body coil. The scans were taken with a SIEMENS TrioTim 3.0 Tesla machine. Key parameters include: Field Strength of 3.0 tesla, Flip Angle of 13.0°, and a GR/IR pulse sequence. The matrix dimensions are $220\times320\times208$ voxels with voxel sizes of $0.8\times0.8\times0.8mm^3$. Other notable parameters were TE = 2.8 ms, TI = 773 ms, and TR = 2200 ms, with a T1 weighting.

Anti-Amyloid Treatment in Asymptomatic Alzheimer's (A4) The A4 [19] provides a unique opportunity to compare MRI findings, such as Amyloid-related imaging abnormalities (ARIA), between cognitively impaired elderly individuals with high or low brain amyloid levels. This dataset includes sequences like T1, T2, GRE, FLAIR, and DWI, captured using a 3T MRI. The specifications for the 3T scanner are: voxel size of $1\times1\times1.2mm^3$ and a 256×256 matrix over 170 slices. For the pretraining of our model, we utilized 1791 T1-weighted images from this dataset.

Australian Imaging, Biomarkers and Lifestyle (AIBL) The AIBL [73] aims to provide researchers with new insights into the onset and progression of Alzheimer's disease. The dataset encompasses both AD and control groups. We utilized a total of 525 T1-weighted (T1w) images from this dataset, consisting of 434 CN and 91 AD samples for model validation. The T1 scanner parameters are set as follows: a matrix size of $240\times240\times160$, voxel size of $1\times1\times1.2mm^3$, TE=3.0 ms, TI=900.0 ms and TR=2300.0 ms.

Open Access Series of Imaging Studies (OASIS) We utilized the OASIS [55, 56] dataset, specifically OASIS 3, which includes sequences such as T1w, T2w, FLAIR, ASL, SWI, time of flight, resting-state BOLD, and DTI. Out of these, we used 817 T1-weighted images (comprising 676 CN and 141 AD) for model validation.

Parkinson's Progression Markers Initiative (PPMI) PPMI [57,58] dataset is a collection of a variety of medical data, including demographic and clinical, genetic, and neuroimaging data (i.e., MRI, PET, and SPECT). In our study, we obtained T1-weighted MRI data from a total of 663 images, which were acquired using the following parameters: field strength = 3T, repetition time (TR) = 2300 ms, echo time (TE) = 2.98 ms, and inversion time (TI) = 900 ms. Field of view (FOV) was $256 \times 256 mm^2$, matrix size was 256×256, and voxel size was $1 \times 1 \times 1.2 mm^3$.

Preprocessing The T1-weighted MR images used in our study were collected from various institutions, resulting in different matrix sizes, voxel spacings, and FOV. We employed the standard preprocessing steps [17], including skull stripping, bias field correction, and intensity normalization. Specifically, we skull-stripped MR using FSL-BET [42]. We resampled the voxels to $1.25 \times 1.25 \times 1.25 mm^3$. Then, we normalized the image intensities of all voxels using the zero-mean unit variance method. Brain anatomy was analyzed using the Desikan atlas, which involves dividing the whole brain into 120 regions and 17 subcortical regions, as computed by Freesurfer [18, 26]. Brain morphology measurements, cortical thickness and curvature, are also calculated using Freesurfer on Desikan atlas and 17 subcortical regions, resulting in 274 measurements.

B Implementation Details

Model architecture We employ the Swin transformer as our backbone framework due to its efficiency on 3D data. Table A shows the model configuration. Specifically, the encoder architecture consists of four stages, each containing two transformer blocks except for the third stage, which consists of six transformer blocks, resulting in a total of $L = 24$ layers. Between stages, a patch merging layer is used to reduce the resolution by a factor of 2. In the first stage, the linear embedding layer and transformer blocks maintain the number of tokens at $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$. Additionally, a patch merging layer groups patches with a resolution of $2 \times 2 \times 2$ and concatenates them resulting in a 4C-dimensional feature embedding. A linear layer is then utilized to downsample the resolution by reducing the dimension to 2C. The procedure is repeated in stages 2, 3, and 4, with resolutions of $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$, and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$, respectively. The patch size is set to $2 \times 2 \times 2$, with a feature dimension of $2 \times 2 \times 2 = 8$. The embedding space has a dimension of $C = 48$. The window size for multi-head self-attention is $7 \times 7 \times 7$.

Table A: Our swin Transformer configuration. FLOPs; floating point operations per second

Patch Size	Window size	Feature size	Embedded Dimension
$2 \times 2 \times 2$	$7 \times 7 \times 7$	48	768
Number of Blocks	Number of Heads	Parameters	FLOPs
[2,2,18,2]	[3,6,12,24]	57.16M	82.38G

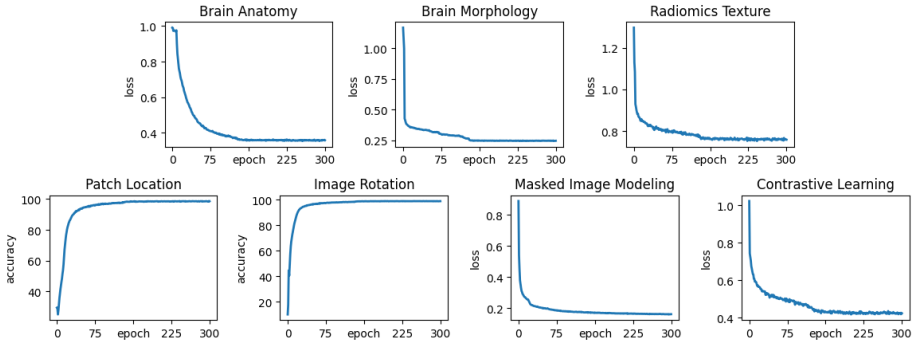

Fig. A: The graphs represent various metrics during pretraining with multi-task learning. The y-axes of the graphs for patch location and image rotation show accuracy, which converges to nearly 100% during training. The y-axes of other tasks show the loss, which converges during 300 epochs.

Settings of 3D ViT We set up the 3D patch embedding of size $16 \times 16 \times 16$ and a projection dimension of 2048. For 3D Swin transformer, we set the patch size to $2 \times 2 \times 2$, with feature dimensions of 8. The dimensions of the embedding space are $C = 48$. For multi-head self-attention, the window size was set to $7 \times 7 \times 7$.

Data augmentation Two strategies were employed for data augmentation. First, we used multi-view (i.e., global and local views) augmentation inspired by DINO [9] for 3D input images. A global view was obtained by cropping and resizing the full image to remove the background to $128 \times 128 \times 128$, which included the entire brain. The local view, on the other hand, is a randomly cropped patch of size $56 \times 56 \times 56$ to focus on specific brain structures and that is further resized to $64 \times 64 \times 64$. Three local and one global views were considered for each sample. Second, we used a series of operations such as rotation and shifted intensity to augment the data. For contrast training, each view was augmented twice, yielding two enhanced views from the same sample. Furthermore, only one of these augmented views is masked, allowing for the simultaneous execution of contrastive learning and masked image modeling. All pretext tasks were applied

to both global and local views, except for the patch location prediction task, owing to the nature of the task.

Hyperparameters We conducted training on four NVIDIA A100 GPUs, each with a batch size of 2. The pretraining phase involved an initial learning rate of 0.0001 for 300 epochs with a cosine annealing scheduler and linear warm-up. We utilized AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Settings of Other SSL Frameworks We tried to keep the original settings of SSL frameworks (i.e., MoCo v2 [15, 37], BYOL [33], and DINO [9]) in the comparative experiments as much as possible. However, our dataset consists of single channel 3D images and has a relatively small number of samples compared to previous studies. Therefore, we made some modifications to several hyperparameters. For the common augmentation method between ours and other SSL, we followed their implementations but replaced the color jitter with intensity scaling and shifting due to the single-channel nature of our medical images. The image size was cropped to $128 \times 128 \times 128$, and a pretrain batch size of 2 per GPU was used for 300 epochs.

MoCov2 We modified the default queue size to 12,288, because the total number of subjects in our dataset is 13,687.

DINO We leveraged a global view of size $128 \times 128 \times 128$ and local views of size $56 \times 56 \times 56$. We used two global views and eight local views for the training process.

C Pretraining Task Details

Brain Anatomy Prediction This task involved predicting the brain parcellation of a given patch. Only the regions belonging to the patch are considered during training, and the other regions are masked out during the loss calculation. For example, we are likely to consider only a few anatomically neighboring regions in a given patch. The segmentation task was performed by adding a simple CNN decoder to form a UNet-like structure, which is based on a previous study that employed a Swin Transformer as an encoder [34]. A total of 120 regions are predicted.

Brain Morphology Prediction This task involved predicting the morphological features of each brain region. We predict the average thickness and curvature in each of the 137 brain regions. Similar to the brain anatomy prediction, only the regions within the patch are considered during training and other regions are masked out during the loss calculation. The morphology values were predicted using a morphology head composed of a simple multilayer perception (MLP) consisting of two FC layers.

Radiomics Texture Prediction This task aimed to predict the radiomics texture features of the gray matter, white matter, and CSF regions. For each region, 20 GLCM features and four GLSZM features are extracted, resulting in 72 features (3 regions with 24 features each). These features were extracted using Pyradiomics v3.0.1 [86]. To execute this task, representation z from the swin transformer was passed through a two-layer perceptron for regression prediction.

Patch Location For the patch location task, an eight-way classification was conducted to estimate the location of $2 \times 2 \times 2$ sub-patches within the 3D images. This task is performed only locally. For patch location, the representation was trained with a single FC layer to perform an 8-way classification.

Image Rotation In our 3D rotation prediction task, we randomly rotate 3D input patches by a degree chosen from a set of 12 possible degrees (i.e., 0, 90, 180, 270 degrees along each axis), then train the model to predict rotation degree in a classification manner. Since the zero-degree rotation of the x, y, and z axes were the same, only 10 possible rotation degrees were available for our classification task. In our study, we added a single FC layer as the image rotation head for 10-way classification.

Masked Image Modeling In global and local perspectives, 75% of the 3D volume within the patch was masked out. We employed a patch size of 16 and randomly generated the cut-out regions. A single-layer projection with pixel shuffle served as the MIM head. During pretraining, the L1 loss was calculated between the original and reconstructed patches.

Contrastive Learning To perform contrastive Learning to randomly augmente the patches to generate positive and negative pairs. Specifically, because we set the batch size to two, one positive pair and two negative pairs were available for i-th augmented patch. Then, we computed the latent representation z of each augmented patch using linear projection, where the dimension of the latent representation was 512. Finally, the contrastive coding loss is computed using eq.6. In our study, we applied a contrastive learning task to both global and local views to learn multiscale representations.

D Results of Pretraining Tasks

Learning progress of each task To demonstrate the effectiveness of our multitasking approach, we evaluated the performance of each task during the pretraining phase. Fig. A showcases the metrics for each task during the training phase. The y-axes of the graphs for patch location and image rotation represent accuracy, whereas masked image modeling and brain morphology use L1 loss, brain anatomy uses the Dice coefficient, and contrastive learning uses information noise

and contrastive estimation loss. Each task demonstrated that the learning metrics converged during training. The model shows pretraining on various aspects of the brain's structural features across seven tasks.

Effectiveness of Pretraining We compared the convergence speed of training with the scratch model and our pretrained model. Fig. B depicts the convergence graphs of the training losses for the two swin transformer models. Our pretrained model not only converges faster in the early epoch, but also has a lower loss than the scratch model at all epochs. The results demonstrate the effectiveness of our pretraining method using multi-task learning.

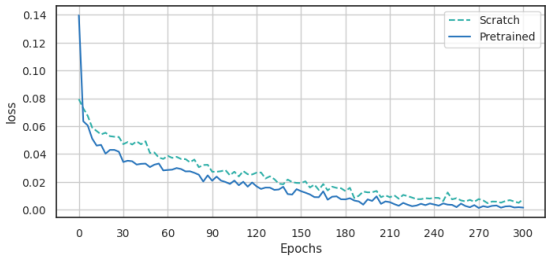

Fig. B: The train loss graphs of the scratch and pretrained Swin Transformer

Results of MIM Fig. C illustrates the reconstruction process for MIM. To pretrain the encoder, we attached a single projection layer to reconstruct the masked 3D volume. Despite performing reconstruction through a simple single projection layer, it is evident that the masked areas are effectively encoded, enabling the identification and restoration of the corresponding structure.

Results of Brain Anatomy Prediction To assess task performance, we visualized ground truth and prediction parcellation. Fig. D illustrates the process of predicting brain anatomy. Fig. E shows a 3D rendering comparing the ground truth with the predicted brain parcellation. Our objective was to train the encoder. Therefore, we utilized a lightweight CNN decoder and observed its ability to reasonably predict the locations of rough parcellations.

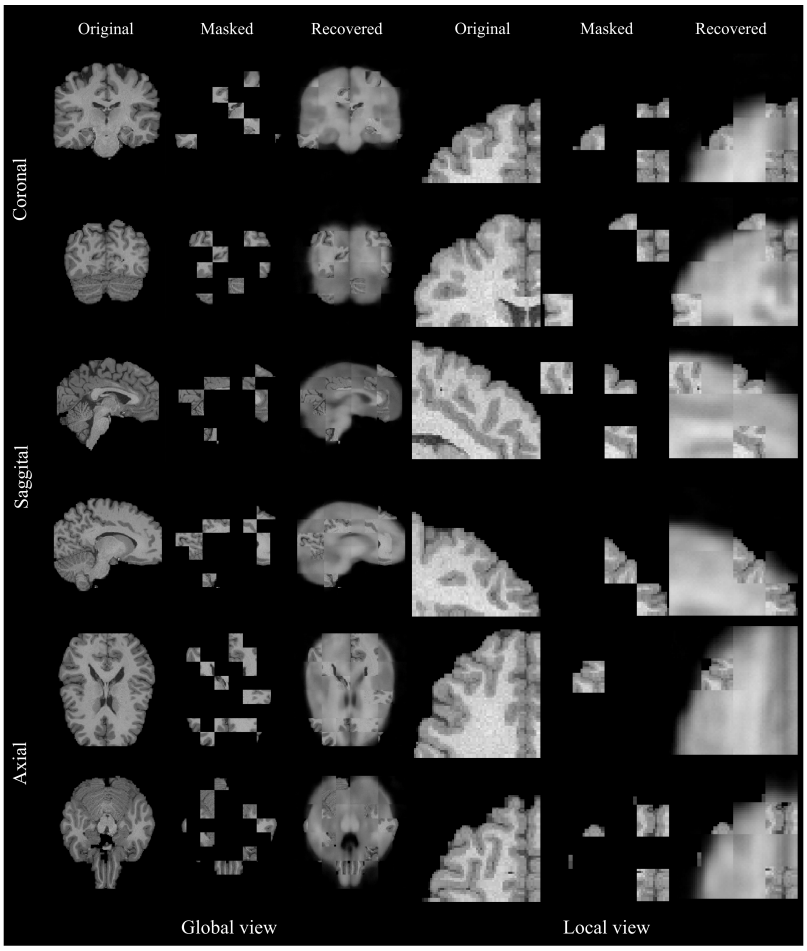

Fig. C: Illustration of the training process for the masked image modeling task. Original: source image. Masked: Image from the original with 75% masked out. Recovered: Image restored after passing the masked image through a single projection head. The model is trained using the L1 Loss between the original and recovered. Given that the input is a 3D volume, we present three planes: coronal, sagittal, and axial. Each plane displays two distinct views. Top: coronal, Middle: saggital, Bottom: axial. Left: global view, Right: local view.

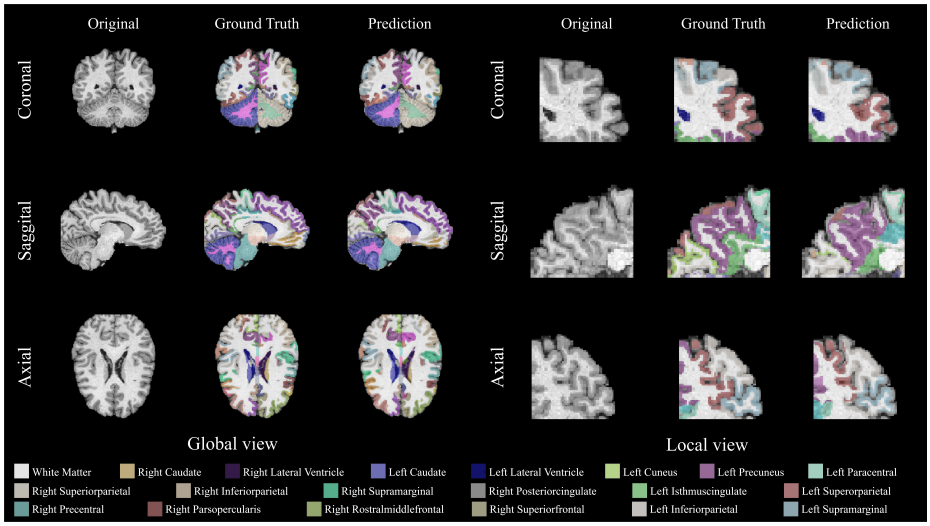

Fig. D: Illustration of the training process for the brain anatomy prediction task. Original: Source image. Ground Truth: Image overlayed with the ground truth brain parcellation on the original. Predicted: Image overlayed with the predicted brain parcellation on the original. It shows that the pretrained encoder with our multi-task effectively predicts brain parcellation. Left: global view, Right: local view

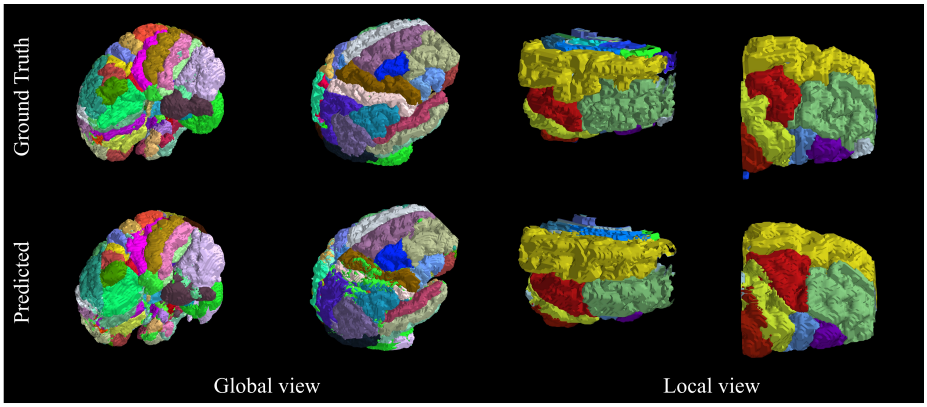

Fig. E: 3D rendering of both the ground truth and predicted brain parcellation. Two different viewing angles are presented. Left: global view, Right: local view