Pretrain Once and Finetune Many Times: How Pretraining Benefits Brain MRI Segmentation

Hao Zhang, Sheng Xu, Wei Ren, Huping Ye, Yi Hong✉
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
yi.hong@sjtu.edu.cn

***Abstract*—Brain MRI segmentation plays an important role in analyzing brain anatomical structures and understanding brain images. In this paper, we consider building a uniform 3D brain MRI segmentation framework using the pre-training and fine-tuning style to fully leverage existing public brain images and segmentation masks. Based on existing Transformer-based 3D image segmentation models, UNETR and Swin UNETR, we study the necessity and benefit of using pre-training, through pre-training on a big collection of over 6,000 brain scans from OASIS, ADNI, and CC359, and fine-tuning with limited segmentation masks to perform three downstream tasks, i.e., skull stripping, 4-structure segmentation, and 33-structure segmentation. Experimental results demonstrate that in most cases the pre-training can help reduce 90% of segmentation masks and half the time. Also, our method outperforms the recent method SynthSeg by a good margin. Our pre-trained model and source code are available online at https://github.com/AllanIverson/medical-segmentation.**

***Index Terms*—Brain MRI segmentation, 3D brain scan, pre-training, fine-tuning, Transformer**

I. Introduction

Brain MRI segmentation, e.g., brain skull stripping, segmentation of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) of a brain scan, is a fundamental task in medical image analysis, which helps understand brain scans and analyze brain anatomical structures of interest. To tackle a specific brain image segmentation task, researchers typically build and train a corresponding segmentation model, as shown in Fig. 1(a). For instance, the current solution often provides one model for brain skull stripping [1], [2], and another one for segmenting GM, WM, and CSF [3], [4], although both work on the same brain images. Also, in practice, annotating medical images at pixel- or voxel-level needs help from experts, which is labor-intensive and time-consuming. Fortunately, many brain image datasets, e.g., OASIS [5], ADNI [6], CC359 [7], are publicly available with a good amount of segmentation masks, although they are collected for different purposes and often used separately in their own tasks [8]. In this paper, we consider building a practical and general brain image segmentation model that fully leverages existing public brain image datasets and then transferring it to handle different brain segmentation tasks with images collected from different domains and limited brain segmentation masks, as shown in Fig. 1(b).

Thanks to the rapid development of self-supervised learning techniques [9], [10], especially the pre-training models, such as MAE [11], I-JEPA [12], DeiT [13], we can pretrain a big model on a large scale of images without annotations and finetune it for the downstream brain segmentation tasks. Since advanced pretraining models are mostly based on two visual backbone networks, i.e., Vision in Transformer (ViT) [14] and Swin-Transformer (Swin-T) [15], for image encoding, we follow these two state-of-the-art (SOTA) methods for handling 3D medical images, i.e., ViT-based UNETR [16] and Swin UNETR [17], and study their potential of handling brain segmentation tasks in one uniform framework. In particular, we aim to answer the following three questions: (1) When do we need pre-training? Is a pre-trained model always helpful? (2) With a pre-trained model, how many image annotations from the target domain are required for fine-tuning to reach the performance of a model trained on a large number of masks? (3) Is the pre-trained model better than a sophisticatedly designed brain segmentation model? By answering these three questions, we know how pre-training benefits brain image segmentation and how to use it in practice.

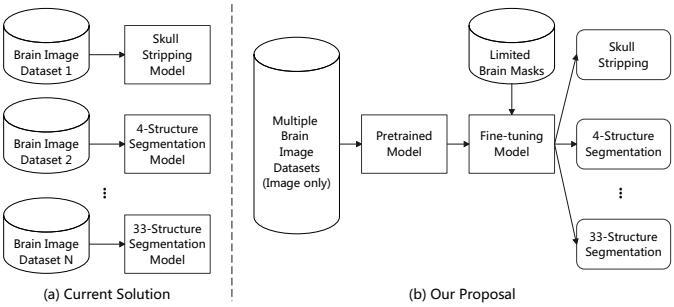


Fig. 1: Comparison between the current solution and our proposal for addressing multiple brain image segmentation tasks in a uniform framework. We pretrain a big self-learning model and finetune it with a low cost for each task.

To answer these questions, we collect over 6,000 3D brain scans from ADNI, OASIS, and CC359, including images with and without a skull, to train a big model based on either 3DViT-B [16] or 3D Swin-T [17]. The downstream tasks of brain segmentation include the simple brain skull stripping (a binary segmentation), a coarse-grained partitioning with four structures (i.e., white matter, subcortical gray matter, cortex, and CSF), and a fine-grained segmentation with 33 structures (e.g., ventricle, brain stem, left thalamus, right thalamus, etc.). To evaluate the contribution from pre-training, we assume that in the source domain, e.g., OASIS, we have many segmentation masks available; however, in the target

domain, e.g., ADNI, we have no or only a few masks available for training, as that we often encounter in practice use. We also compare our models with a recent brain segmentation model SynthSeg [18], to present the advantage of using pre-training.

Experimental results demonstrate that with the help of the pre-trained model, for most cases we can reduce 90% of the masks and half of the time to achieve comparable performance with the model trained on all available masks in the target domain. The more difficult the segmentation task is, the more benefits we can gain from using pre-training. Besides, the pre-training is extensible to many different downstream tasks and easy to use once we collect some masks in our target domain. Overall, the contributions of this paper are summarized below:

- We propose a uniform framework for brain MRI segmentation using the pre-training and fine-tuning style, which needs pre-training only once but can be fine-tuned many times for different brain segmentation purposes.
- We pretrain a big model on over 6,000 brain MRI scans and demonstrate its big gain of reducing annotation and training time requirements, via evaluating three brain segmentation tasks with different difficulty levels.
- Our model outperforms the current SOTA method SynthSeg [18], and more importantly, our method can easily use existing images and masks and be straightforwardly extended to a new domain, which is practical and instructive for clinical use.

II. Related Work

Brain MRI segmentation has been an active research topic for decades [19], [20], including traditional methods based on region growing [21], thresholding [22], fuzzy c-means algorithms [23], machine learning based methods like using auto-context [24], and advanced learning methods based on deep neural networks [4], [18], [25]. In recent years, due to the high accuracy and efficiency of deep learning-based methods, they have been the first choice for brain MRI segmentation.

A. CNNs and Brain Image Segmentation

Due to the powerful feature extraction capability of convolutional neural Networks (CNNs), they have been widely used in medical image segmentation, e.g., MR brain image segmentation [26]. Especially, fully convolutional network (FCN) [27] and the family of UNets [28] are two classical CNN designs for image segmentation, which are extended to the 3D version [29], [30] for segmenting medical image volumes like brain scans. SynthSeg [18] is a recent work on brain MRI segmentation based on CNNs, which can handle brain images with any contrast and resolution without retraining. Although CNN-based methods have been successfully used in brain image segmentation, the convolutions used in these models have limited receptive fields, which greatly limits its further improvement on the segmentation accuracy.

B. Transformer and Pretrained Models

In natural language processing (NLP), Transformer-based models allow long-distance modeling by stacking a set of self-attention blocks, which have the flexibility of learning both short- and long-distance information [31]. Also, unlike CNN-based models whose performance rapidly saturates with model capacity, Transformer-based models are more powerful to handle large-scale datasets [32]–[34]. Along with the development of Vision in Transformer (ViT), Transformer-based models have been the new favorite for medical image segmentation [16], [17], [35].

To learn from large-scale datasets with the most unlabelled images, self-supervised learning is a popular and effective choice and allows for pre-training. Models driven by a large number of data have achieved comparable and even better performance, compared to supervised learning methods, as demonstrated in [36]–[38]. In the past decades, more and more datasets become publicly available and easily accessible, which has facilitated the emergence of numerous relevant endeavors on employing self-supervised learning techniques in the domain of medical image analysis. For instance, Tang et al. [39] collect 5050 publicly available CT images to pre-train a Swin-T structure segmentation network to obtain SOTA performance on the Beyond the Cranial Vault (BTCV) Segmentation Challenge with 13 abdominal organs using the Medical Segmentation Decathlon (MSD) dataset. Ghesu et al. [40] collect more than 100 million publicly available medical images in various forms and achieve success in the detection of abnormalities from chest radiography scans and hemorrhage detection on brain CT by using contrast learning. Even Azizi et al. [41] observe that although medical images are apparently different from natural images, the model pre-trained on natural images can still help improve the dermatology and chest X-ray classification performance on medical images.

Therefore, although CNN-based self-learning methods have been explored in medical image analysis [42], we prefer a Transformer-based pretraining strategy due to its success in image analysis like [11]–[13]. Therefore, we collect more than 6,000 brain image volumes to pretrain a transformer-based brain image encoder and perform three downstream brain MRI segmentation tasks with varying difficulties. To our knowledge, our work is the first one to explore pretraining brain MRI segmentation on such a large scale and outperforms the recent method SynthSeg [18] by a good margin.

III. Method

As shown in Fig. 1(b), our brain segmentation model includes two stages, i.e., the pre-training and fine-tuning stages.

A. Pre-training Stage

MAE [11] has been successfully used in pre-training natural images and videos [43]. In medical domains, researchers demonstrate that conducting MAE pre-training on MRIs can substantially enhance the performance of downstream tasks [44]. Hence, we adopt the MAE-like design as in Fig. 2. In particular, our pre-training model adopts the reconstruction-based self-learning technique, by using a heavyweight encoder based on Transformer to extract image features and a lightweight decoder to reconstruct the masked 3D image

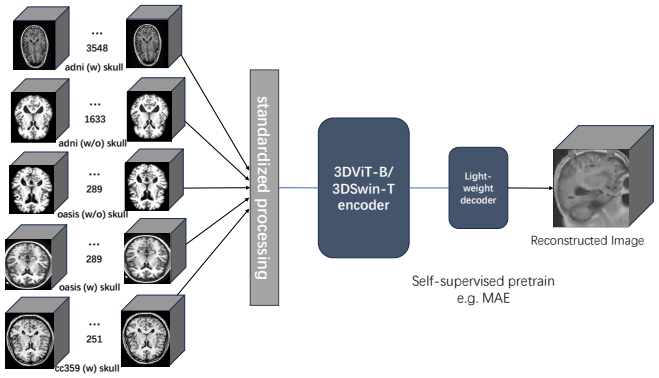


Fig. 2: Pre-training stage: collecting brain images with (w) and without (w/o) brain skulls from multiple datasets, followed by a standardized processing step, a large image encoder but a light-weight decoder, and then reconstructing the input image back by performing self-supervised learning like MAE [11]. No segmentation masks are used in this stage.

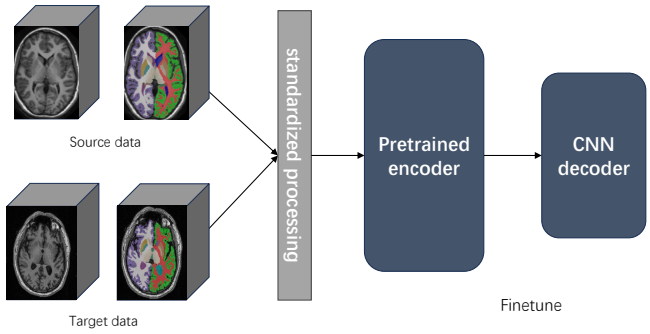


Fig. 3: Fine-tuning stage: using all brain segmentation masks from the source domain and some from the target domain to finetune the pre-trained encoder and a CNN decoder to perform brain image segmentation.

patches. Before inputting brain MRIs into the network, we have a standardized processing step to normalize all images collected from different datasets, as described in Section IV-A.

Encoder. We have two choices for our pre-training encoder, one is 3D ViT-B used in UNETR [16] and the other is 3D Swin-T used in Swin UNETR [17]. Both models are 3D Transformer designs for medical image segmentation. The 3D ViT-B encoder has a 12-layer self-attention-block stack with 12 heads per layer. Self-attention is computed among all 3D patches. Each 3D patch is projected into a 768-dimensional token, which is kept consistent with the vanilla ViT [14].

For the 3D Swin-T encoder, we use an FCN-like down-sampling method [27], which differs from the ViT-based encoder in several ways. Firstly, its self-attention is computed by using a sliding window, and its receptive field is not as wide as ViT. Secondly, it has a mechanism to reduce the resolution and longitudinally stretch the number of channels, which is similar to the convolution operation of CNN. Table I compares the model size of these two encoders in terms of the number of parameters. The 3D ViT-B is twice the size of the 3D Swin-T. Typically, a Transformer model with more parameters is more powerful and can fit a larger dataset, which is also demonstrated in our experiments.

Lightweight Decoder. To perform the reconstruction of masked image patches, we use only four layers of self-attention blocks, compared to the 12 blocks used in the 3D ViT-B encoder. Also, it connects feature maps extracted from the encoder at the same resolution by skipping connections [28], as done in UNETR [16].

Loss Functions. For UNETR with 3D ViT-B encoder, we adopt the mean squared error (MSE), the same loss with MAE [11]. As for Swin UNETR with 3D Swin-T encoder, we follow Tang's work [39], which utilizes masked volume inpainting, rotation angle prediction, and contrastive learning to effectively extract features from image patches.

TABLE I: Model size comparison between two pre-training encoders. The number 16 indicates that the ViT uses a 3D patch of $16\times16\times16$.

Model	Backbone	#Parameter
UNETR	3D ViT-B (16)	89.2M
Swin UNETR	3D Swin-T	42.6M

B. Fine-tuning Stage

In the pre-training stage, we do not use any segmentation masks. To perform brain segmentation, we need to finetune the pre-trained encoder and re-train a lightweight decoder for segmentation. Fig. 3 illustrates the shared network design for different downstream brain segmentation tasks. We assume a scenario where a good amount of segmentation masks are available in the source domain, like a public dataset, and a few masks are provided in the target domain like our private dataset. We use the combination of segmentation masks collected from these two domains for fine-tuning.

After the same standardized processing step as used in the pretraining stage, we finetune the pre-trained encoder and train a CNN decoder to perform each segmentation task. Thanks to the pre-training model, this fine-tuning stage is very efficient and typically takes 10-20 GPU hours to complete its training.

CNN Decoder and Loss Functions. We attempt to use a transformer-based decoder and observe that it cannot effectively integrate features from different resolutions, as done in the CNN-based U-shape network. Therefore, during the fine-tuning stage, we adopt the same CNN decoders used in UNETR [16] and Swin UNETR [17]. For all our downstream tasks, a combination of cross entropy and dice coefficients is used as the loss function.

IV. Experiments

A. Datasets

We collect 3D brain MRI scans and segmentation masks from three datasets, i.e., OASIS [5], ADNI [6], and CC359 [7].

OASIS [5]. This dataset contains 416 T1-weighted brain scans of OASIS-1 collected from 416 subjects, who are aged from

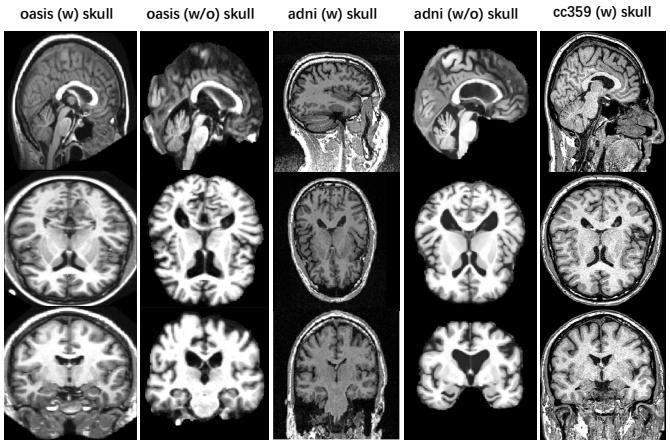


Fig. 4: Visualization of brain MRI scans sampled from our three datasets, i.e., OASIS, ADNI, and CC359.

18 to 96 years old and 100 of whom are clinically diagnosed with mild to moderate Alzheimer's disease. To build our dataset, we use both raw image scans with skulls and their pre-processed images after skull stripping. The images have a size of 256×256×256, with a resolution of 1mm×1mm×1mm. Each image has a binary brain extraction mask (i.e., brain and non-brain regions), a four-structure segmentation mask (i.e., white matter, subcortical gray matter, cortex, CSF), and a 35-structure mask (e.g., brain stem, 3rd ventricle, right amygdala, etc.). Since OASIS has all the masks we need for experiments, we treat it as the source domain and others, e.g., ADNI, CC359, as the target domain, during the fine-tuning stage.

ADNI [6]. This dataset includes over 10,000 1.5T/3T T1-weighted structure MRI scans collected from more than 2,000 subjects, who are healthy or diagnosed with mild cognitive impairment or Alzheimer's Disease. Some subjects have brain scans at multiple time points. The original image size is 256×176×240, with a variety of different voxel spacing. Thanks to [45], we collect brain segmentation masks with 138 structures for over 5,000 raw images. To be consistent with the OASIS segmentation masks, with the help of ChatGPT [46], we group some small structures into a four-structure segmentation mask and a 33-structure segmentation mask, respectively. Since there are two anatomical regions missing in the ADNI segmentation mask, we only use the corresponding 33 structures for both datasets. Figure 5 visualizes the segmentation masks sampled from the OASIS and ADNI datasets. We also have preprocessed images without skulls for ADNI. The pre-processing step includes denoising, bias field correction, skull stripping, and affine registration to the SRI24 atlas. However, we do not have brain segmentation masks for this set.

CC359 [7]. This dataset contains 359 brain MRIs collected from older healthy adults aged from 29 to 80 years, using three scanners, i.e., GE, Siemens, and Philips. This dataset provides brain extraction masks for all image volumes, which are generated using supervised classification. In this data set, MRIs collected from Siemens scanner have a size of 150×288×288 with a resolution of 1mm×0.88mm×0.88mm,

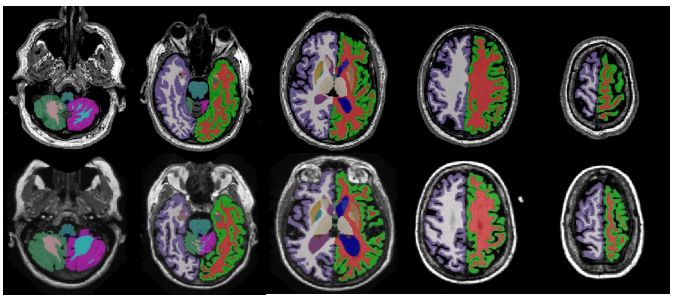


Fig. 5: Visualization of segmentation masks with partial brain structures. Top: OASIS, bottom: ADNI.

scans from Philips scanner have a size of 192×256×256 with a resolution of 1mm×1mm×1mm, and those from GE have a size of 200×256×256 with a resolution of 1mm×1mm×1mm.

We randomly take 70% of images from each dataset to form a large set of 6,010 images for pre-training, including 3,548 images with skull and 1,633 images without skull from ADNI, 289 images with skull and 289 images without skull from OASIS, and 251 images with skull from CC359. For the remaining 30% of the images, we have 858 of them divided for validation and 1,717 for testing. To make sure the test set contains only unseen subjects, we remove the images from those subjects whose images at some time points are used in pre-training. As a result, we have 380 image volumes in total for the test, including 245 brain image volumes from ADNI, 83 from OASIS, and 52 from CC359.

Standardization. Since our images have varying sizes and resolutions, we resample all image volumes, resulting in images with a resolution of 1mm×1mm×1mm. All images are then padded or cropped, resulting in an image size of (192, 192, 192). Limited by our GPU memory, we randomly sample 3D image patches with a size of (160, 160, 160) for UNETR, and (96, 96, 96) for Swin UNETR. Lastly, we normalize the image intensity to [0, 1]. Table II summarizes our three datasets.

TABLE II: Summary of our standardized datasets.

Dataset	Modality	Resolution	#Subjects	#Images
OASIS	T1	$1mm^3$	416	832
ADNI	T1	$1mm^3$	2334	7401
CC359	T1	$1mm^3$	359	359

B. Downstream Tasks

We have 289 images from OASIS with binary, four-structure, and 33-structure masks as our source domain. With the help of the pre-trained model, we perform the following three downstream tasks in different target domains:

- Skull-stripping. For this task, our target domain is CC359. We have 289 binary brain masks from OASIS for fine-tuning and then explore by adding how many additional masks, we can achieve comparable performance with the fully-supervised segmentation model which uses all 251 masks from the target domain CC359 for training.

Model	Pre-training	Skull Stripping			4-Structure Segmentation			33-Structure Segmentation		
		#Masks (CC359)	Dice Score↑	Fine-tuning (GPUs Hours)	#Masks (ADNI)	Dice Score↑	Fine-tuning (GPU Hours)	#Masks (ADNI))	Dice Score↑	Fine-tuning (GPUs Hours)
UNETR with 3DViT-B	w/o	251	99.2%	13.16	3548	94.7%	79.12	3548	88.8%	109.28
	w	0	97.3%	10.52	0	89.6%	9.96	0	75.8%	13.96
		5	98.5%	11.72 (+1.20)	5	93.1%	15.44 (+5.48)	5	83.0%	17.44 (+3.48)
		10	98.7%	12.04 (+1.52)	10	93.4%	15.80 (+5.84)	10	84.5%	20.52 (+6.56)
		20	98.8%	12.44 (+1.92)	20	93.9%	16.64 (+6.68)	20	85.1%	21.52 (+7.56)
		40	99.0%	13.28 (+2.76)	40	94.3%	18.44 (+8.48)	40	86.0%	23.52 (+9.56)
		60	99.1%	15.92 (+5.40)	100	94.6%	29.60 (+19.64)	100	87.0%	29.24 (+15.28)
		100	99.2%	16.48 (+5.96)	200	94.8%	38.48 (+28.52)	300	88.4%	44.40 (+30.44)
		–	–	–	300	95.0%	47.12 (+37.16)	600	89.0%	52.44 (+38.48)
Swin UNETR with 3DSwin-T	w/o	251	98.7%	10.36	3548	94.9%	95.64	3548	90.9%	165.60
	w	0	91.1%	9.36	0	73.4%	11.68	0	73.4%	11.40
		5	98.3%	10.16 (+0.80)	5	74.0%	19.80 (+8.12)	5	82.3%	18.60 (+7.20)
		10	98.5%	10.40 (+1.04)	10	86.7%	20.20 (+8.52)	10	84.7%	20.60 (+9.20)
		20	98.6%	10.88 (+1.52)	20	91.2%	20.28 (+8.60)	20	85.4%	21.72 (+10.32)
		40	98.8%	11.60 (+2.24)	30	91.4%	21.48 (+9.80)	40	86.4%	25.56 (+14.16)
		100	98.9%	14.32 (+4.96)	100	92.9%	22.42 (+10.74)	300	89.7%	44.08 (+32.68)
		–	–	–	200	94.5%	29.12 (+17.44)	600	90.7%	67.72 (+56.32)
		–	–	–	300	94.9%	41.68 (+30.00)	–	–	–

TABLE III: Comparison between two segmentation models with (w) and without (w/o) pre-training via three brain segmentation tasks. The highlighted rows show results that are produced by using pre-training and are comparable with those using all masks without pre-training. The number after + indicates the additional GPU hours required to finetune the model with masks from the target domain, i.e., CC359 or ADNI, compared to fine-tuning it only using masks from the source domain, i.e., OASIS.

- Four-structure segmentation. For this task, our target domain is ADNI. Similarly, we have 289 brain masks from OASIS with four anatomical regions for fine-tuning. The fully-supervised model for comparison has all 3,548 segmentation masks from ADNI for training.
- 33-class segmentation. This task has the same target domain and the same data for fine-tuning and testing. The only difference is segmenting more anatomical regions.

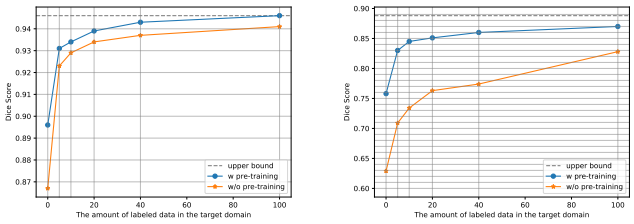
C. Other Settings

We use the AdamW optimizer with an initial learning rate of 1e-4 and a regularized learning rate of 1e-5. Our model is implemented using PyTorch-2.0 and is trained on four NVIDIA GeForce RTX 3090 GPUs, with a batch size of three 3D MRI scans per GPU. We have different training settings for pre-training and fine-tuning. In the pre-training stage, to train the encoder with the 3D ViT-B backbone, we set a maximum of 200 epochs and select the model with the best reconstruction for fine-tuning. For the encoder with the 3D Swin-T backbone, we have 120k iterations during pre-training. In the fine-tuning stage, the maximum number of epochs is 100 and we use the early stopping based on the model's performance on the validation set. Besides, we reduce the learning rate during the fine-tuning phase to avoid excessive modification of the model.

D. Experimental Results

Q1: When do we need pre-training? Is a pre-training model always helpful?

Figure 6 shows the multi-class segmentation results on ADNI with and without pre-training. For both cases, the segmentation accuracy increases as we add more and more segmentation masks from the ADNI training set. For both



(a) 4-structure segmentation (b) 33-structure segmentation

Fig. 6: ADNI image segmentation with (w) and without (w/o) pre-training using UNETR with 3D ViT-B as the backbone encoder. The upper bound is calculated by using all 3,548 brain masks in the ADNI training set.

4-structure and 33-structure segmentation tasks, using pre-training achieves higher segmentation accuracy and gradually approaches the upper bound computed by using all masks in the ADNI training set. Also, for difficult tasks like the 33-structure segmentation, pre-training gains more improvement on the segmentation accuracy with limited annotations. That is, pre-training is quite helpful when handling the segmentation of many anatomical regions with limited brain masks.

Table III reports more quantitative results on three downstream tasks with different difficulty levels. For simple brain segmentation like the skull-stripping task, if we have more than 250 brain masks, segmentation models like UNETR or Swin-UNETR can achieve over 98% of segmentation accuracy. The only benefit provided by pre-training is the number reduction of required brain segmentation masks. However, as the difficulty level of a segmentation task increases, we need more masks to achieve good segmentation accuracy. For

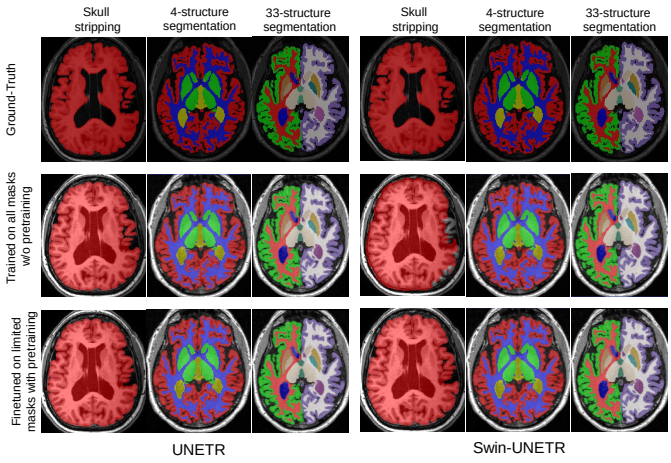


Fig. 7: Visualization of segmentation results on three downstream tasks. Top to bottom: the ground-truth mask, masks predicted by the model trained on all ADNI masks without pre-training, and masks predicted by the model after pre-training and fine-tuning and having comparable results, corresponding to the highlighted ones in Table III.

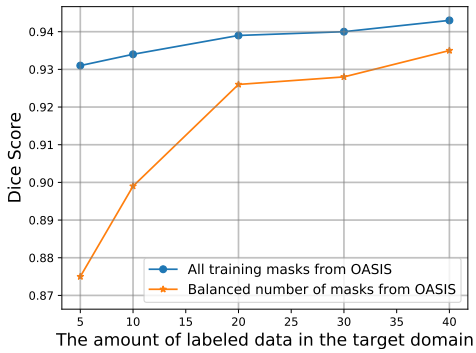


Fig. 8: Comparison between using the same number of masks from the source (OASIS) and target (ADNI) domains and using all segmentation masks from the source domain.

instance, we need more than 3,500 brain masks to achieve a 95% Dice score for 4-structure segmentation and 89%-90% for 33-structure segmentation. In such cases, pre-training can greatly reduce the number of required brain masks for training.

In summary, we need pre-training when we have limited brain segmentation masks in our target domain, especially for the difficult segmentation tasks with many anatomical regions of interest. For some simple segmentation tasks like binary segmentation, pre-training may not be very helpful, especially when we can easily obtain hundreds of segmentation masks for images in our target domain. Even like this, using pre-training with zero masks in the target domain can still provide a good starting point for segmentation.

Q2: With a pre-trained model, how many image annotations from the target domain are required for fine-tuning to reach the performance of a model trained on a large number of masks?

In Table III, using UNETR with 3DViT-B encoder, we

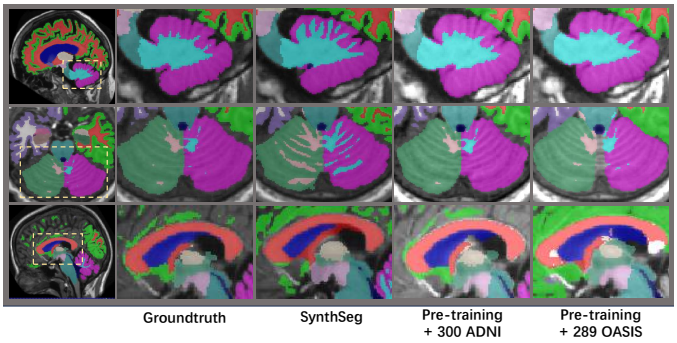


Fig. 9: OASIS result comparison between SynthSeg and ours.

need 100 (40% of 251) masks for skull stripping, 100 (2.8% of 3548) for 4-structure segmentation, and 300 (8.5% of 3548) for 33-structure segmentation, to achieve the UNETR's performance trained on all masks. On the other hand, Swin-UNETR needs fewer masks for skull stripping, only 40 (16% of 251), but more masks for 4-structure and 33-structure segmentation. The qualitative results are visualized in Fig. 7. If we are satisfied with just over 90% of segmentation accuracy, for the skull-stripping tasks, we do not need any masks from our own target domain. Only using the OASIS mask for fine-tuning, we can obtain good results that are ready to use, especially using the UNETR network with pre-training. For the 4-structure segmentation, we need a few masks from our target domain to obtain over 90% accuracy. With 5 ADNI masks and UNETR, we can obtain a 93.1% Dice score. The 33-structure segmentation is more difficult, but with 20 ADNI masks, we can also obtain over 85% Dice score for both UNETR and Swin-UNETR. More importantly, we need less than half the time for fine-tuning to obtain a good multi-class segmentation model, compared to training a model on all masks. However, this is not the case for the skull-stripping task. Overall, by using pre-training, we can reduce the required amount of segmentation masks and for most cases, we can reduce 90%. Further, pre-training can further reduce the training time when working on multi-class segmentation, typically half the time.

We also explore the possibility of reducing the masks in the source domain of OASIS, like using the same amount of masks in both source and target domains. As shown in Fig. 8, using all available segmentation masks in the source domain is a better choice. That is, we probably would like to use all masks we can collect from the public datasets to help the segmentation in our private dataset. Therefore, in all other experiments, we fully leverage the brain segmentation masks provided in our OASIS training set.

Q3: Is the pre-trained model better than a sophisticatedly designed brain segmentation model?

Table IV and Figure 9&10 show the comparison between SynthSeg [18] and ours using pre-training and different numbers of brain masks for fine-tuning. SynthSeg was trained on more than 1,000 brain scans, including 500 masks from ADNI. It obtains an 83.2% Dice score on our OASIS test set. We first randomly choose 300 segmentation masks from the ADNI

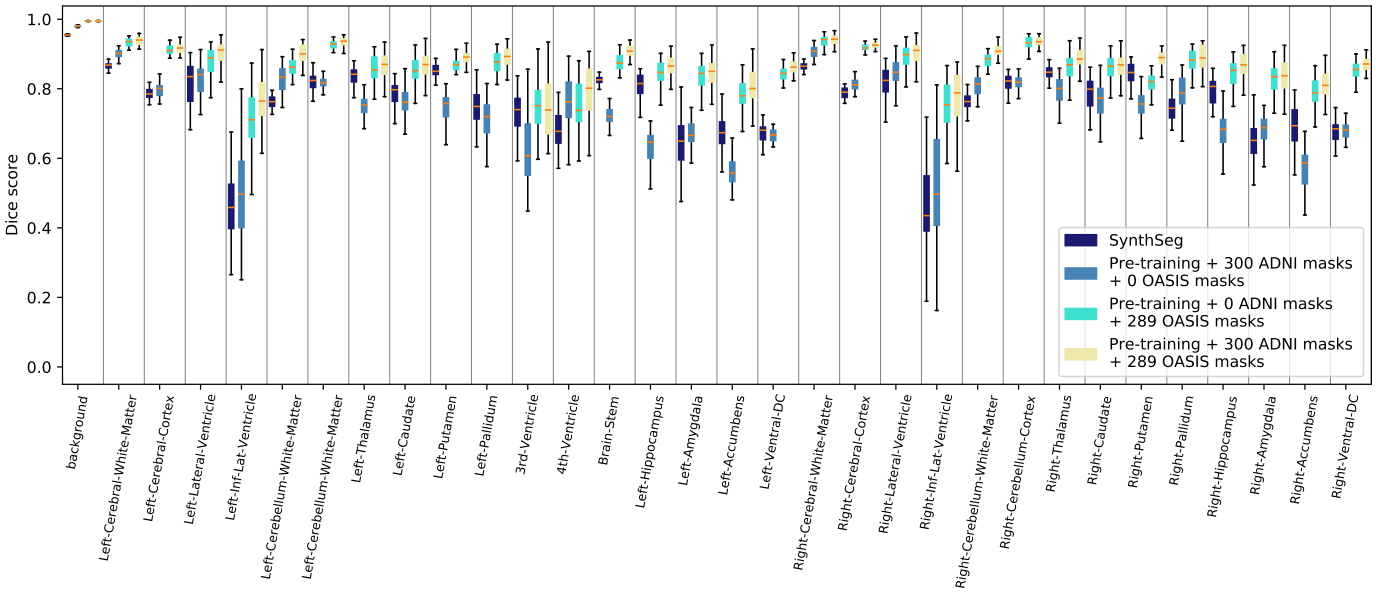


Fig. 10: Comparison between our models with a recent brain segmentation model SynthSeg [18], tested on the OASIS dataset.

TABLE IV: Comparison between our pre-training models and SynthSeg [18] on OASIS (the target domain). Our source domain only contains 300 ADNI masks and SynthSeg has 500 ADNI masks, 500 HCP [47] masks, and 20 T1-39 [48] masks.

Method	#Masks (Source Domain)	#Masks (Target Domain)	Dice Score↑
SynthSeg	1020	0	83.2%
Fine-tuning 1	300	0	84.3%
Fine-tuning 2	0	289	93.2%
Fine-tuning 3	300	289	**93.8%**

dataset to finetune our pre-trained model, it achieves a 1% Dice improvement over SynthSeg. By only using OASIS masks (fine-tuning 2) and having both ADNI and OASIS masks (fine-tuning 3), our pre-trained model can further improve the Dice score by 8.9% and 9.5%, respectively. This result demonstrates the potential of a pre-trained model, which can fully leverage masks from different domains and benefit a lot by using masks from the target domain. On the contrary, fine-tuning SynthSeg is non-trivial. That is, our pre-trained model is general and extensible to downstream brain segmentation takes.

V. Conclusion and Future Work

In this paper, we propose a uniform framework for brain MRI segmentation, which is pre-trained once on over 6,000 brain MRI scans and fine-tuned many times for different downstream brain segmentation tasks. The pre-training technique can greatly reduce the demand for brain segmentation masks and the training time for segmentation. One limitation of our work is the lack of zero-shot testing, i.e., evaluating another unseen brain image dataset. Also, we are interested in applying our model in practice, using a private dataset collected from hospitals to evaluate the performance of our pre-trained model. We will explore these two directions in future work.

Acknowledgments

This work was supported by NSFC 62203303 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

References

[1] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in mri," *Neuroimage*, vol. 22, no. 3, pp. 1060–1075, 2004.

[2] R. Dey and Y. Hong, "Compnet: Complementary segmentation network for brain mri extraction," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer, 2018, pp. 628–636.

[3] N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*. Springer, 2015, vol. 9351.

[4] R. Gandhi and Y. Hong, "Mda-net: Multi-dimensional attention-based neural network for 3d image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging*. IEEE, 2021, pp. 822–826.

[5] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 09 2007.

[6] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[7] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, and R. Lotufo, "An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482–494, 2018, segmenting the Brain.

[8] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A. Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "Clip-driven universal model for organ segmentation and tumor detection," *arXiv:2301.00785*, 2023.

[9] J. Xu, "A review of self-supervised learning methods in the field of medical image analysis," *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol. 13, no. 4, pp. 33–46, 2021.

[10] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *PeerJ Computer Science*, vol. 8, p. e1045, 2022.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[12] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[16] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[17] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[18] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, J. E. Iglesias *et al.*, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining," *Medical image analysis*, vol. 86, p. 102789, 2023.

[19] I. Despotović, B. Goossens, W. Philips *et al.*, "Mri segmentation of the human brain: challenges, methods, and applications," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[20] M. K. Singh and K. K. Singh, "A review of publicly available automatic brain segmentation methodologies, machine learning models, recent advancements, and their comparison," *Annals of Neurosciences*, vol. 28, no. 1-2, pp. 82–93, 2021.

[21] H. Tang, E. Wu, Q. Ma, D. Gallagher, G. Perera, and T. Zhuang, "Mri brain image segmentation by multi-resolution edge detection and region selection," *Computerized Medical Imaging and Graphics*, vol. 24, no. 6, pp. 349–357, 2000.

[22] G. E. Sujji, Y. Lakshmi, and G. W. Jiji, "Mri brain image segmentation based on thresholding," *International Journal of Advanced Computer Research*, vol. 3, no. 1, p. 97, 2013.

[23] L. Szilagyi, Z. Benyo, S. M. Szilágyi, and H. Adam, "Mr brain image segmentation using an enhanced fuzzy c-means algorithm," in *Proceedings of the 25th annual international conference of the IEEE engineering in medicine and biology society (IEEE Cat. No. 03CH37439)*, vol. 1. IEEE, 2003, pp. 724–726.

[24] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 10, pp. 1744–1757, 2009.

[25] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[26] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of mr brain images with a convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[29] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "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*. Springer, 2016, pp. 424–432.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[35] Y. Jiang, Y. Zhang, X. Lin, J. Dong, T. Cheng, and J. Liang, "Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer," *Brain sciences*, vol. 12, no. 6, p. 797, 2022.

[36] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[38] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, 2019, pp. 384–393.

[39] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "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*, 2022, pp. 20 730–20 740.

[40] F. C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, D. Neumann, P. Patel, R. Vishwanath, J. M. Balter, Y. Cao, S. Grbic *et al.*, "Self-supervised learning from 100 million medical images," *arXiv preprint arXiv:2201.01283*, 2022.

[41] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3478–3488.

[42] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3d self-supervised methods for medical imaging," *Advances in neural information processing systems*, vol. 33, pp. 18 158–18 172, 2020.

[43] C. Feichtenhofer, Y. Li, K. He *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, vol. 35, pp. 35 946–35 958, 2022.

[44] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image classification and segmentation," *arXiv preprint arXiv:2203.05573*, 2022.

[45] C. Ledig, A. Schuh, R. Guerrero, R. Heckemann, and D. Rueckert, "Structural brain imaging in alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database," *Scientific Reports*, 2018.

[46] OpenAI. (2021) ChatGPT: Conversational ai language model.

[47] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss *et al.*, "The human connectome project: a data acquisition perspective," *Neuroimage*, vol. 62, no. 4, pp. 2222–2231, 2012.

[48] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness *et al.*, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.