MAS-CL: An End-to-End Multi-Atlas Supervised Contrastive Learning Framework for Brain ROI Segmentation

Liang Sun, Yanling Fu, Junyong Zhao, Wei Shao, Qi Zhu, and Daoqiang Zhang, *Senior Member, IEEE*

***Abstract*— Brain region-of-interest (ROI) segmentation with magnetic resonance (MR) images is a basic prerequisite step for brain analysis. The main problem with using deep learning for brain ROI segmentation is the lack of sufficient annotated data. To address this issue, in this paper, we propose a simple multi-atlas supervised contrastive learning framework (MAS-CL) for brain ROI segmentation with MR images in an end-to-end manner. Specifically, our MAS-CL framework mainly consists of two steps, including 1) a multi-atlas supervised contrastive learning method to learn the latent representation using a limited amount of voxel-level labeling brain MR images, and 2) brain ROI segmentation based on the pre-trained backbone using our MSA-CL method. Specifically, different from traditional contrastive learning, in our proposed method, we use multi-atlas supervised information to pre-train the backbone for learning the latent representation of input MR image, *i.e.*, the correlation of each sample pair is defined by using the label maps of input MR image and atlas images. Then, we extend the pre-trained backbone to segment brain ROI with MR images. We perform our proposed MAS-CL framework with five segmentation methods on LONI-LPBA40, IXI, OASIS, ADNI, and CC359 datasets for brain ROI segmentation with MR images. Various experimental results suggested that our proposed MAS-CL framework can significantly improve the segmentation performance on these five datasets.**

***Index Terms*— Contrastive learning, multi-atlas, brain segmentation.**

Manuscript received 11 September 2023; revised 25 March 2024 and 5 June 2024; accepted 16 July 2024. Date of publication 25 July 2024; date of current version 31 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62376123, Grant 62136004, Grant 62006115, Grant 62276130, Grant 62272226, Grant 62076129, Grant 62371234, Grant 82171249, and Grant 82172061; in part by the National Key Research and Development Program of China under Grant 2023YFF1204803; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515011925; in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842; and in part by the Key Research and Development Plan in Jiangsu (Social Development) under Grant BE2022677. The associate editor coordinating the review of this article and approving it for publication was Dr. Jacinto C. Nascimento. *(Corresponding author: Daoqiang Zhang.)*

Liang Sun, Wei Shao, and Daoqiang Zhang are with the College of Artificial Intelligence, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen 518038, China (e-mail: dqzhang@nuaa.edu.cn).

Yanling Fu, Junyong Zhao, and Qi Zhu are with the College of Artificial Intelligence, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

Digital Object Identifier 10.1109/TIP.2024.3431451

I. Introduction

MULTI-ATLAS segmentation methods [1], [2], [3] have shown effectiveness for brain region-of-interest (ROI) segmentation with MR images. The basic idea of multi-atlas segmentation methods is that pair-wise voxels with similar local appearance should have the same labels. The multi-atlas segmentation method propagates the labels of atlas images to the target image based on the calculated local similarity. Hence, the feature representation of brain MR images is an important factor for multi-atlas segmentation methods. However, most traditional multi-atlas segmentation methods calculate the local similarity by using intensity features. The simple intensity feature cannot adequately describe the complex anatomical structure of the brain.

Deep learning methods show great success for feature representation in medical image analysis tasks [4], [5], [6], [7], [8], [9], [10], [11], [12]. The deep learning methods generally need numerous training data to train an effective deep network for feature representation. However, since the voxel-level labeling with such high-dimensional 3D brain MR images is extremely time-consuming, the brain MR images with voxel-level labeling are difficult to acquire in large quantities. Hence, there is insufficient labeling MR images to train a deep learning model for effectively representing the brain MR images. Recently, to solve the problem of limited data, many contrastive learning methods [13], [14], [15], [16], [17], [18], [19], [20], [21] are proposed to learn feature representation with a small number of training data. Contrastive learning methods first employ a pretext task to train a deep neural network, and then use the learned parameters of the trained deep neural network to fine-tune the downstream network. In recent contrastive learning methods, pretext tasks are generally defined in an unsupervised manner. As shown in Fig. 1, the positive and negative pair of samples are generated by using the data augmentation operations, *i.e.*, crop, cutout, and rotate *etc.* However, the human brain has extremely complex anatomical structure, and the intensity MR images are high-resolution to present the details of brain anatomical structure. Hence, most common data augmentation operations are not appropriate for brain ROI segmentation with MR images.

In summary, the multi-atlas segmentation methods can utilize the anatomical prior from atlas images to boost

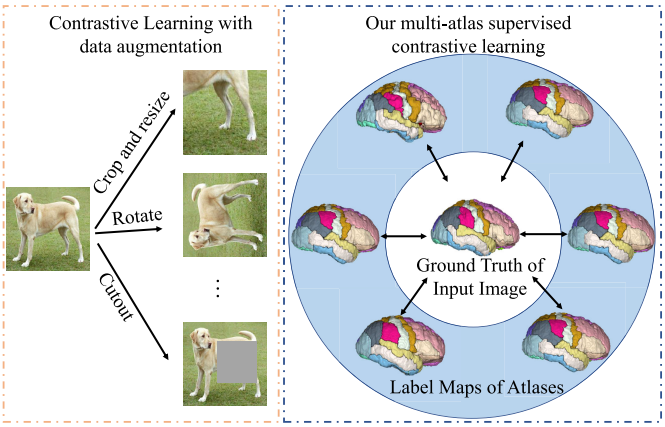

Fig. 1. Our multi-atlas supervised contrastive learning method generates the sample pairs based on the supervised information of multiple atlases.

the segmentation performance. The feature representation is a bottleneck to further improve the performance of multi-atlas segmentation methods for brain ROI segmentation with MR images. Meanwhile, deep neural networks cannot learn an adequate model to represent the MR images with a limited number of voxel-level labeled MR images.

To this end, in this paper, we propose a multi-atlas supervised contrastive learning framework for brain ROI segmentation with MR images, denoted as MAS-CL. Firstly, our MAS-CL employs a multi-atlas supervised contrastive learning module to learn the latent representation of MR images. Specifically, as shown in Fig. 1, to take advantage of anatomical prior used in the multi-atlas segmentation methods, for each target voxel in input MR image, we generate the corresponding voxels in atlas images. Hence, we can define the correlation of each pair of voxels based on the labels of voxels of the target image and atlas images to pre-train the backbone for brain ROI segmentation. Secondly, we introduce a segmentation model to obtain the final label map of the target MR image. Specifically, in this paper, we validate our MAS-CL framework with three kinds of brain ROI segmentation methods (five methods), including a multi-atlas segmentation method (locally-weighed voting), two end-to-end segmentation networks (U-Net and nnUNet), and two deep-learning based multi-atlas segmentation method (label fusion network and anatomical gated U-Net). Various experiments show that our proposed MAS-CL methods achieve superior segmentation results on all datasets when compared to several state-of-the-art methods.

The main contributions of this work are listed in the following three-fold.

- We propose an end-to-end multi-atlas supervised contrastive learning framework for 3D medical image segmentation. Our method can pre-train the backbone network without extra data augmentation. In our MAS-CL framework, we use atlases to define the voxel-level sample pairs to pre-train the backbones, which is more effective for medical image segmentation tasks.
- Our framework is very simple and flexible for 3D medical image segmentation. It can be easily combined with most brain ROI segmentation methods, such as multi-atlas segmentation methods and deep learning-based segmentation methods. Therefore, we implement our proposed MAS-CL framework with three kinds of brain ROI segmentation methods, including the conventional multi-atlas segmentation method, the end-to-end networks, and deep learning-based multi-atlas segmentation methods.
- We validate five methods with our MAS-CL framework on LONI-LPBA40, IXI, OASIS, ADNI, and CC359 datasets. The experimental results on these datasets show that our MAS-CL framework can significantly improve the segmentation results.

II. Related Work

A. Multi-Atlas Segmentation

Multi-atlas segmentation methods achieved great success for brain MR image segmentation [22], [23], [24]. Multi-atlas segmentation methods assume that the voxels should have the same label if they have a similar local appearance pattern. Therefore, the multi-atlas segmentation methods generally consist of two steps, *i.e.*, 1) the image registration step aiming to warp the atlas images to the common space of the to-be-segmented image, and 2) the label fusion step aiming to propagate the labels of the atlas images to the to-be-segmented image. Based on the basic assumption, many multi-atlas segmentation methods are proposed. For example, the weighted voting methods [25], [26] calculate the pair-wise patch similarity of the voxels on the target image and atlas images at the same location. Then, the pair-wise patch similarity is used as a voting weight to determine the final label of the target voxel in a weighted voting manner. Due to the brain MR images exhibiting extremely complex anatomical structures, the registration step is inevitable to produce registration errors. To alleviate the possible registration errors, the idea of non-local is used in multi-atlas segmentation methods. In the non-local based multi-atlas segmentation methods [1], [2], [3], the voxel-wise local appearance similarity is not only calculated with the candidate voxels at the same location in atlas images but also with the candidate voxels within the location-specific region in atlas images. Then, the labels of selected candidate voxels within the location-specific regions are propagated to the target voxels. However, these traditional multi-atlas segmentation methods generally use the intensity features to calculate the similarity. The simple intensity features cannot adequately describe the complex anatomical structure of brain MR images.

Deep neural networks show great feature representation ability. Therefore, to take advantage of deep neural networks for representation learning, deep neural networks are used to learn the representation of MR images for multi-atlas segmentation. For instance, a deep neural network is employed [27] to learn the discriminative features of image patches. Then, they used the learned deep features

to calculate the voting weights for label fusion based on the conventional multi-atlas segmentation methods. The unsupervised deep learning method [28] is employed to learn the latent representation of brain MR images to calculate the similarity of patches. However, both conventional multi-atlas methods and these deep feature representation-based multi-atlas methods are performed in a voxel-by-voxel manner, these methods are very time-consuming for brain ROI segmentation with MR images.

B. Convolutional Neural Networks

Convolutional neural networks are widely used in image segmentation. Especially, the encoder-decoder architecture [29], [30], [31], [32], [33] can map the target image from its image space to label map directly. Hence, the encoder-decoder networks are much faster for image segmentation. The U-Net and its variants [30], [31], [34] show great performance for medical image segmentation. U-Net can reuse the high-resolution spatial feature maps as complementary local details to improve performance. Recently, many convolutional neural networks have been proposed for brain ROI segmentation with MR images. For example, DARTS [35] employs a dense U-Net for brain ROI segmentation. E2D [36] uses three independently modified U-Net to segment the hippocampus in sagittal, coronal, and axial views and then fuses the segmented results in different views to obtain the final label map. In addition, some deep learning methods with prior are proposed for brain ROI segmentation. DeepNAT [37] uses coordinate information to train the network for brain ROI segmentation. Furthermore, AG-UNet [38] learns the anatomical prior from multi-atlas by the convolutional neural network to boost the brain ROI segmentation performance. However, convolutional neural networks need a large amount of labeling data to train the deep model. The voxel-level labeling brain MR images is very rare. Hence, it is a remaining challenge to train an effective model for brain MR segmentation.

C. Contrastive Learning

Contrastive learning methods [16], [39], [40], [41], [42], [43], [44] have been widely used to learn a latent representation from the limited training data. The classical contrastive learning methods generally first define the positive pairs and negative pairs by data augmentation methods. Then, the defined pairs are used to train the backbone. Hence, the pre-trained network can be trained with the target dataset. However, most of these methods train the network for image-level tasks. For the voxel-level brain MR image segmentation tasks, it is not very appropriate to define the pair-wise relationship to calculate the voxel-level contrastive loss for high-dimensional brain MR images. Inspired by the idea of multi-atlas segmentation methods, in this paper, we use the anatomical prior from multiple atlases to define the pair-wise relationship to pre-train the network.

III. Methodology

In this section, we first introduce the multi-atlas supervised contrastive learning method. We then present the implementation details of our multi-atlas supervised contrastive learning method for brain ROI segmentation.

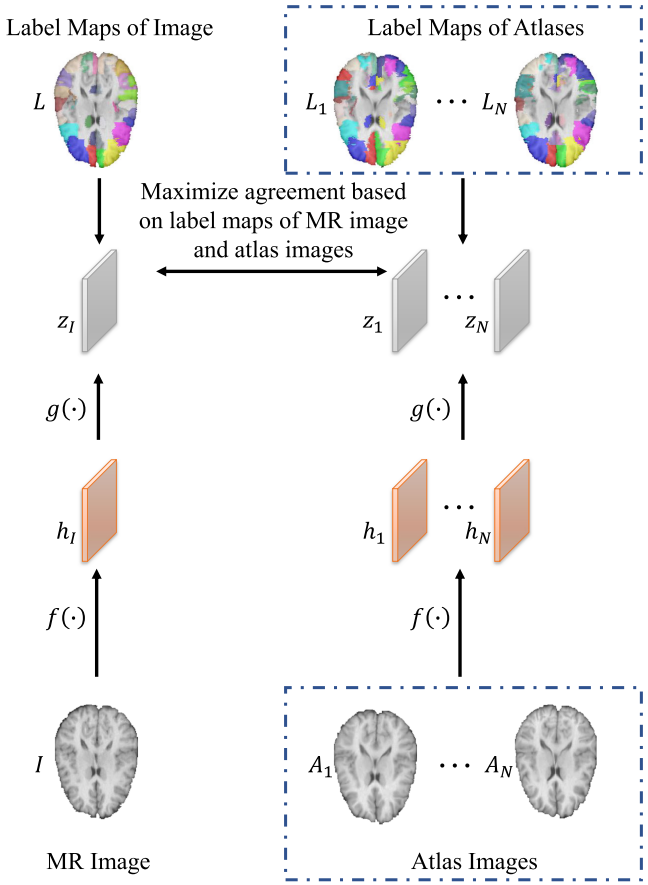

Fig. 2. Illustration of multi-atlas supervised contrastive learning framework. The input MR image and multiple atlas images are fed into an encoder-decoder network. Then, a voxel-level projection head is used to map the learned features into the latent space. Finally, we maximize agreement based on label maps of MR image and atlas images in voxel-level.

A. Multi-Atlas Supervised Contrastive Learning

Fig. 2 shows our proposed multi-atlas supervised contrastive learning framework. Different from the conventional contrastive learning methods, we do not use any data augmentation operation to generate sample pairs. We obtain the sample pairs from the multiple atlas images. Specifically, we define the voxel v_i in input MR image I and $v_{n,j}$ in atlas image A_n are positive pair if v_i and $v_{n,j}$ have same labels (*i.e.*, $l(v_i) = l(v_{n,j})$), otherwise are negative pair.

According to the definition, the input MR image I and atlas image set $A = \{A_n | n = 1, \ldots, N\}$ are fed into a backbone encoder-decoder network $f(\cdot)$ to learn the feature maps $h_I = f(I)$ and $h_A = \{h_n = f(A_n) | n = 1, \ldots, N\}$, respectively. Where N is the number of atlas images.

Inspired by the projection head widely used in contrastive learning, we then employ a voxel-level projection head based on the $1 \times 1 \times 1$ convolutional layer to map these latent features onto a new space to calculate the contrastive loss. As shown in Fig. 3, the projection head used in our work is $g(\cdot) = \text{Conv}^{(2)}(\text{ReLU}(\text{Conv}^{(1)}(\cdot)))$, where $\text{Conv}(\cdot)$ is the convolutional layer with $1 \times 1 \times 1$ kernels, and $\text{ReLU}(\cdot)$ is

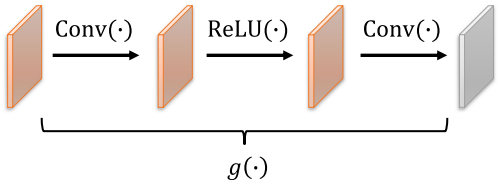

Fig. 3. Illustration of the projection head $g(\cdot)$ used in our proposed multi-atlas contrastive learning module. Conv(·) is the convolutional layer with $1 \times 1 \times 1$ kernels, and ReLU(·) is ReLU non-linearity.

ReLU non-linearity function. Hence, $z_I = g(h_I)$ and $z_A = g(h_A) = \{g(h_n)|n = 1, \ldots, N\}$.

Finally, we maximize agreement based on the label maps of MR image and atlas images. Herein, inspired by the non-local strategy used in multi-atlas segmentation methods, we calculate the contrastive loss for our multi-atlas contrastive learning module in a search region, which can generate more potential sample pairs. Specifically, the contrastive loss is not only calculated voxels at the same location in the MR image and the atlas images, but also between the voxels in neighboring region $R(v_i)$ of v_i in atlas images. More specifically, we can implement the non-local strategy by shifting atlas images with the step sizes. The loss function is defined as,

$$loss_i = -\log \frac{\sum_{n,j} \delta_{[l(v_i)=l(v_{n,j})]} \exp(sim(z_i, z_{n,j})/\tau)}{\sum_{n,j} \exp(sim(z_i, z_{n,j})/\tau)}, \quad (1)$$

where $v_{n,j} \in R(v_i)$. $\delta_{[l(v_i)=l(v_{n,j})]}$ is an indicator function, which is 1 if $l(v_i) = l(v_{n,j})$, and 0 otherwise. τ is a temperature parameter. $sim(z_i, z_{n,j})$ is the voxel-wise similarity between voxels z_i and $z_{n,j}$. In this paper, we define the voxel-wise similarity as follows,

$$sim(z_i, z_{n,j}) = \frac{z_i^\top z_{n,j}}{||z_i||||z_{n,j}||}, \quad (2)$$

where $\top$ and $||a||$ are the transpose operation and l_2-norm of vector a, respectively. The overall loss is computed on the whole voxel of images in a mini-batch. Algorithm 1 summarizes our proposed multi-atlas supervised contrastive learning method.

B. Implementation for Brain MR Image Segmentation

Our proposed multi-atlas supervised contrastive learning method is a flexible framework, which can easily combine with most brain segmentation methods. In this paper, we implement our proposed multi-atlas supervised contrastive learning framework with three backbones, including U-Net, nnUNet, and anatomical gated U-Net (AG-UNet), to learn the feature representation of the input MR images. The last layer in these backbones is replaced with our proposed projection heads. Then, these networks are trained with Eq. 1.

To evaluate the feature representation ability of our MAS-CL framework, we first implement our MAS-CL with three kinds of brain segmentation methods based on the U-Net backbone, including locally-weighted voting (LWV) [25], U-Net, and label fusion network. The implementation details are as follows,

Algorithm 1 Multi-Atlas Supervised Contrastive Learning Algorithm

Input: Input MR image I, label map of input MR image L_I, Atlas images $A = \{A_n|n = 1, \cdots, N\}$, label maps of atlas images $L_A = \{L_n|n = 1, \cdots, N\}$, temperature parameter τ, networks $f(\cdot)$ and $g(\cdot)$
for sampled MR image I
  for all voxel v_i in I **do**
    for $A_n \in A$ **do**
      for voxel $v_{n,j}$ in A_n and $v_{n,j} \in R(v_i)$ **do**
        $sim(z_i, z_{n,j}) = z_i^\top z_{n,j}/(||z_i||||z_{n,j}||)$
      end for
    end for
    Calculate $loss_i$ using Eq. 1
  end for
  $\mathcal{L} = Average(\sum_i loss_i)$
  update networks $f(\cdot)$ and $g(\cdot)$ to minimize $\mathcal{L}$
end for
return network $f(\cdot)$

1) LWV-CL: LWV is a multi-atlas-based segmentation method. LWV calculates the voting weights between the to-be-segmented voxel and candidate voxels in atlases at the same location for label fusion. Herein, we use the feature representation learned by our multi-atlas supervised contrastive learning framework based on U-Net to calculate the similarity.

2) U-Net-CL: U-Net is a classical end-to-end network for image segmentation. Specifically, a $1 \times 1 \times 1$ convolutional layer with a Softmax unit following the output of the U-Net backbone is used to obtain the probability map of the target MR image. We use the pre-trained U-Net to initial U-Net-CL.

3) LF-CL: We design an end-to-end label fusion network for brain ROI segmentation, denoted as LF. The architecture of LF is shown in Fig. 4. Specifically, the proposed network first uses the backbone $f(\cdot)$ to learn the deep representation of MR images and atlas images, respectively. Then, the learned feature maps $h_I = f(I)$ and $h_A = \{h_n = f(A_n)|n = 1, \ldots, N\}$ are fed into the convolutional layers $q(\cdot)$ and $k(\cdot)$ with $1 \times 1 \times 1$ kernels to obtain the feature maps $z_I = q(h_I)$ and $z_A = \{z_n = k(h_n)|n = 1, \ldots, N\}$, respectively. We further measure the voxel-wise similarity between voxels z_i and $z_{n,j}$ by Eq. 2. Then, we obtain the probability of voxel v_i belongs to c-th ROI by weighted voting,

$$p(v_{i,c}) = \frac{\sum_{n,j} \delta_{[l(v_i)=l(v_{n,j})]} * \exp(sim(z_i, z_{n,j}))}{\sum_{n,j} \exp(sim(z_i, z_{n,j}))} \quad (3)$$

Hence, in the training stage, we can obtain the probability map by Eq 3. Then, the probability map can be used to calculate the loss and optimize the network. Specifically, in this paper, we employ a cross-entropy loss to train the network as follows

$$loss_{ce} = -\frac{1}{N \times w \times h \times d} \sum_{j=1}^{N} \sum_{i=1}^{w \times h \times d} \sum_{c=1}^{C} \delta_{[l(v_i)=c]} \log p(v_{i,c}), \quad (4)$$

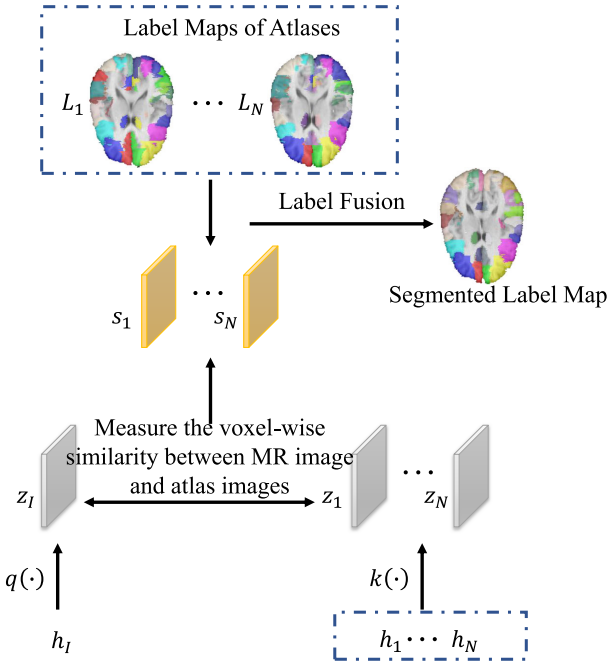

Fig. 4. Illustration of label fusion network. The learned feature maps of MR images and atlas images are fed into the convolutional layers, respectively. Then, a voxel-wise similarity between the MR image and atlas images is calculated. Finally, a label fusion based on the calculated voxel-wise similarity is performed to obtain the final segmented label map.

where N is batch size. w, h, and d are the dimensions of the input MR image. In the testing stage, we further use the maximum a posteriori (MAP) criterion to obtain the final labels for each voxel v_i in the target brain MR image,

$$l(v_i) = \arg\max_c \{p(v_{i,c})\}_{c=1}^{C}. \tag{5}$$

Then, we implement our MAS-CL with two state-of-the-art methods, *i.e.*, nnUNet and AG-UNet. The implementation details are as follows,

4) nnUNet-CL: We adopt nnU-Net as the backbone. Similarly, a $1 \times 1 \times 1$ convolutional layer with a Softmax unit following the nnUNet backbone is used to obtain the label probability map of the target MR image. We use the pre-trained parameters to initial nnUNet-CL.

5) AG-UNet-CL: We adopt the AG-UNet as the backbone. A $1 \times 1 \times 1$ convolutional layer with a Softmax unit following the AG-UNet backbone is used to obtain the label probability map of the target MR image. We use the pre-trained parameters to initial AG-UNet-CL.

IV. Experiment

In this section, we first present datasets and experimental settings used in our study. Then, we show the experimental results for ROI segmentation with brain MR images.

A. Materials

We perform our methods on two public datasets for brain ROI segmentation with MR images *i.e.*, LONI-LPBA40 [45], IXI [46], [47], OASIS [48], ADNI [49] and CC359 [50] datasets. More details of these five datasets are listed as follows,

1) **LONI-LPBA40 [45]:** The LONI-LPBA40 dataset is provided by the Laboratory of Neuro Imaging (LONI) for whole brain ROI segmentation. This dataset consists of 40 brain MR images and manually annotated label maps. The MR images were acquired on a GE 1.5 Tesla system with 124 contiguous 1.5 mm coronal brain slices. More specifically, TR is 10.00-12.50 ms, TE is 4.22-4.50 ms, FOV is 220 mm or 200 mm, in-plane voxel resolution is of 0.86 mm (38 subjects) or 0.78 mm (2 subjects).All MR images are also resampled to the $1 \times 1 \times 1\, mm^3$ resolution by using trilinear interpolation methods. Furthermore, these MR images have already been rigidly registered to the MNI305 template [45]. For the LONI-LPBA40 dataset, we also first randomly select 20 MR images as the atlases, and the remaining 20 MR images are randomly divided into 2 subsets for 2-fold cross-validation for whole brain segmentation on LONI-LPBA40 dataset.
2) **IXI [46], [47]:** The IXI dataset included 30 adult brain atlases with 95 ROIs. MRI scans were obtained on the 1.5 Tesla GE Signa Echospeed scanner with voxel sizes of $0.9375 \times 0.9375 \times 1.5$ mm. For each MR image, we also perform the skull removal algorithm [51], N4-based bias field correction algorithm [52], and intensity standardization algorithm [53]. For the IXI dataset, we also first randomly select 20 MR images as the atlases, and the remaining 10 MR images are randomly divided into 2 subsets for 2-fold cross-validation for whole brain segmentation on the IXI dataset.
3) **OASIS [48]:** The OASIS dataset consists of a cross-sectional collection of 416 subjects aged 18 to 96. The MR images were acquired with the in-plane resolution of $1\,mm \times 1\,mm$ and the slice thickness of $1.25\,mm$. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD). We randomly select 20 MR images as the atlases, 100 MR images as the training images, and the remaining MR images as the testing images.
4) **ADNI [49]:** We employ 60 subjects from ADNI for hippocampus segmentation. These brain MR images were acquired in the sagittal view, with the in-plane resolution of $1\,mm \times 1\,mm$ and the slice thickness of $1.2\,mm$. All images are resampled to have the resolution of $1 \times 1 \times 1\, mm^3$ with trilinear interpolation. The ground-truth label maps were created manually to annotate the right and left hippocampus regions in the brain. We perform pre-processing for all MR images via three procedures, including skull removal [51], N4-based bias field correction [52], and intensity standardization [53]. We randomly select 20 subjects as atlases, and the remaining images are randomly split into 2 subsets for 2-fold cross-validation on ADNI.
5) **CC359 [50]:** The CC359 dataset consists of 359 subjects with an age range from 29 to 80 years. The MR images were acquired on scanners from three vendors

(Siemens, Philips, and General Electric) at both 1.5 T and 3 T. Post-hoc testing with Bonferroni correction demonstrated that only the Philips 3 T and Siemens 3 T group age. We also perform the hippocampus segmentation task on the CC359 dataset. We randomly select 20 subjects as atlases, and the remaining images are randomly split into 2 subsets for 2-fold cross-validation.

B. Experimental Settings

We compared our LWV-CL, U-Net-CL, and LF-CL with their counterparts without using our multi-atlas supervised contrastive learning framework, denoted as LWV, U-Net, and LF, respectively. We use the cross-entropy loss to train the deep learning models. The learning rate and epoch are set to 0.001 and 100, respectively. Similar to literature [16], we also froze the backbone network. Meanwhile, we do not use the non-local strategy in our experiments. Instead, we only generate the sample pairs at the same location of the target image and atlas images. Hence, the segmentation performance is used as a proxy for representation quality.

We set the number of atlas images to 20. We set the mini-batch size as 1. Hence, each mini-batch has 20 sample pairs. Besides, we perform the affine registration [54] and deformable registration [55] algorithms on the atlas images to map the atlas images onto the same space as the target image.

Two main evaluation metrics are leveraged to evaluate the segmentation performance of our proposed methods and their competing methods. We first employ the Dice coefficient (*DC*) to assess the segmentation performance, which is defined as,

$$DC = \frac{2|R_1 \cap R_2|}{|R_1| + |R_2|}, \tag{6}$$

where R_1 and R_2 denote the segmented region and the ground truth, respectively. The term $\cap$ is used to reflect the overlap between R_1 and R_2, *i.e.*, the number of voxels of the intersecting regions. $|a|$ is the number of voxels of region a. Meanwhile, we use the sensitivity (*SEN*) to evaluate the performance of different segmentation methods for brain ROI segmentation, which is defined as,

$$SEN = = \frac{TP}{TP + FN}, \tag{7}$$

where TP and FN are True Positive and False Negative, respectively.

C. Results on LONI-LPBA40

We first validate our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL methods on the LONI-LPBA40 dataset for ROI segmentation with brain MR images. The average Dice coefficient (*DC*) and sensitivity (*SEN*) are reported in Table I.

As can be seen from Table I, our AG-UNet-CL achieves the best segmentation results in the Dice coefficient. AG-UNet-CL gains with 0.0364, 0.0194, 0.0354, 0.0198, 0.0215, 0.0040, 0.0238, 0.0102, and 0.0163 increments in Dice coefficient over LWV, LWV-CL, U-Net, U-Net-CL, nnUNet, nnUNet-CL, LF, and LF-CL, respectively. Meanwhile, AG-UNet-CL also achieves the best performance in terms of sensitivity. In addition, using our proposed MAS-CL framework, LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet achieve better segmentation performance over LWV, U-Net, nnUNet, LF, and AG-UNet in both Dice coefficient and sensitivity, respectively. We perform the Wilcoxon signed rank test in Dice coefficient results achieved by our methods with their counterparts, respectively. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL suggest significant improvement ($p < 0.05$) over LWV ($p = 8.8575e-05$), U-Net ($p = 1.0335e-04$), nnUNet($p = 8.8575e-05$), LF ($p = 8.8575e-05$), and AG-UNet ($p = 8.8575e-05$) for brain ROI segmentation task, respectively. These results imply that our proposed multi-atlas supervised contrastive learning framework can learn better latent representation for brain ROI segmentation on LONI-LPBA40 datasets.

TABLE I

SEGMENTATION RESULTS ACHIEVED BY DIFFERENT METHODS ON THE LONI-LPBA40 DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY. THE SYMBOL '*' INDICATES THAT OUR PROPOSED METHOD CAN SIGNIFICANTLY IMPROVE ITS CONVENTIONAL COUNTERPART BASED ON THE WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	*DC*	*SEN*
LWV	0.7822 ± 0.0088	0.7724 ± 0.0061
*LWV-CL	0.7992 ± 0.0105	0.8029 ± 0.0124
U-Net	0.7832 ± 0.0118	0.7759 ± 0.0134
*U-Net-CL	0.7988 ± 0.0116	0.7959 ± 0.0134
nnUNet	0.7971 ± 0.0118	0.7955 ± 0.0137
*nnUNet-CL	0.8146 ± 0.0100	0.8183 ± 0.0114
LF	0.7948 ± 0.0105	0.7987 ± 0.0127
*LF-CL	0.8084 ± 0.0103	0.8136 ± 0.0105
AG-UNet	0.8023 ± 0.0106	0.8032 ± 0.0100
*AG-UNet-CL	$\mathbf{0.8186 \pm 0.0113}$	$\mathbf{0.8246 \pm 0.0113}$

Fig. 5 plots the violin visualization of Dice coefficient values achieved by U-Net-CL, nnUNet-CL, LF-CL, AG-UNet-CL, and their counterparts on 54 ROIs. As shown in Fig. 5, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL with our multi-atlas supervised contrastive learning framework outperforms U-Net, nnUNet, LF, and AG-UNet in most ROIs, respectively.

In Fig. 6, we further show the surface distance between the segmentation results of different methods and ground truth on three different sizes of ROIs, including the superior frontal gyrus with more than 100,000 voxels, the precuneus with about 20,000 voxels, and the hippocampus with about 6,000 voxels. As shown in Fig. 6, our proposed methods achieved better quality segmentation results when compared with their conventional counterparts, respectively.

D. Results on IXI

We then compared our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL with their counterparts on the IXI dataset for brain ROI segmentation. The segmentation results achieved by different methods are reported in Table II.

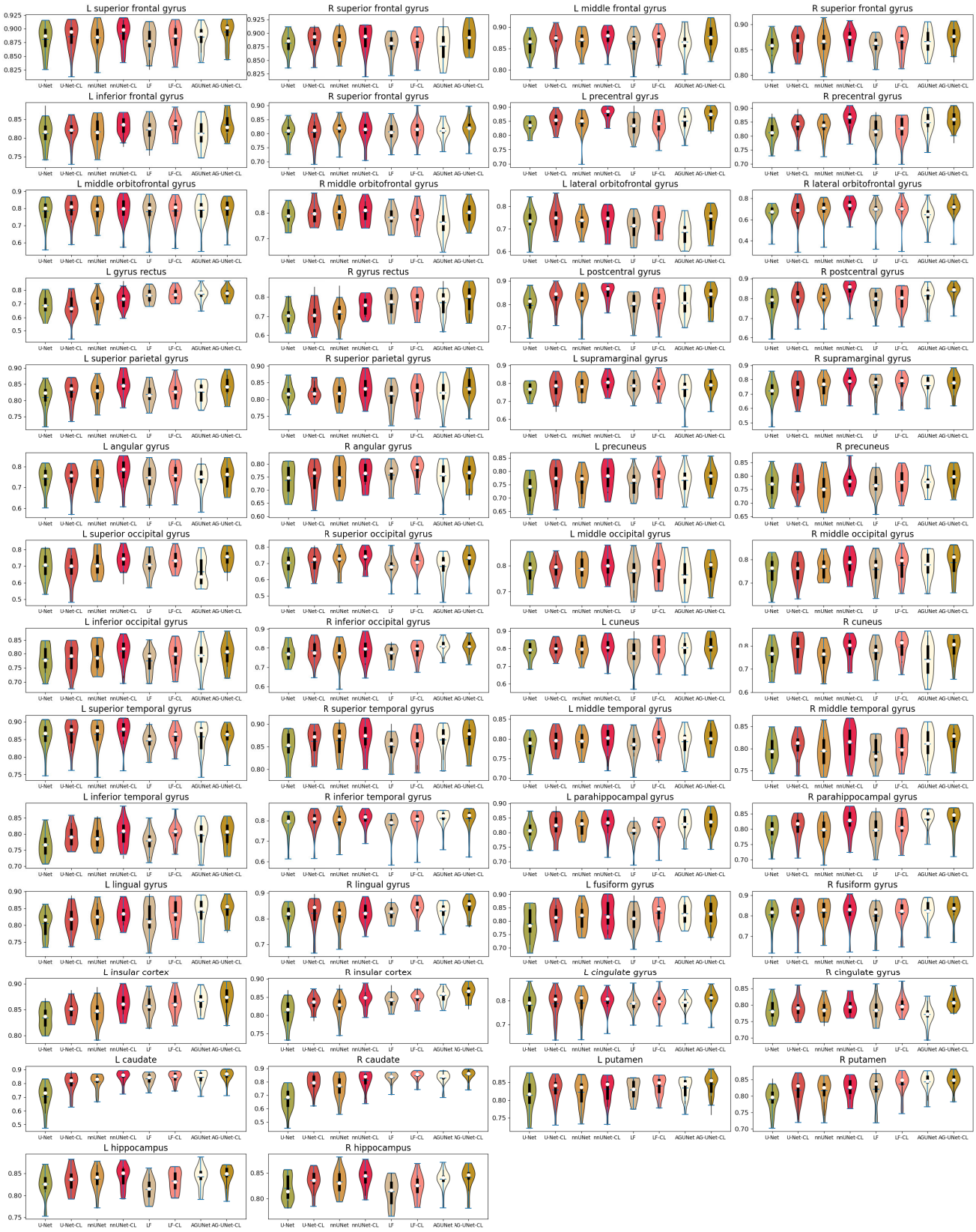

Fig. 5. Segmentation result of 54 ROIs on the LONI-LPBA40 dataset achieved by U-Net, U-Net-CL, nnUNet, nnUNet-CL, LF, LF-CL, AG-UNet, and AG-UNet-CL in terms of Dice coefficient values.

From Table II, we can see that the Dice coefficient on the IXI dataset are 0.7796, 0.7714, 0.7900, 0.7862, and 0.8007 achieved by our LWV-CL, U-Net-CL, nnUNet, LF-CL, and AG-UNet-CL, which are better than their counterpart (*i.e.*, LWV, U-Net, nnUNet, LF, and AG-UNet), respectively. Meanwhile, our LWV-CL, U-Net-CL, nnUNet,

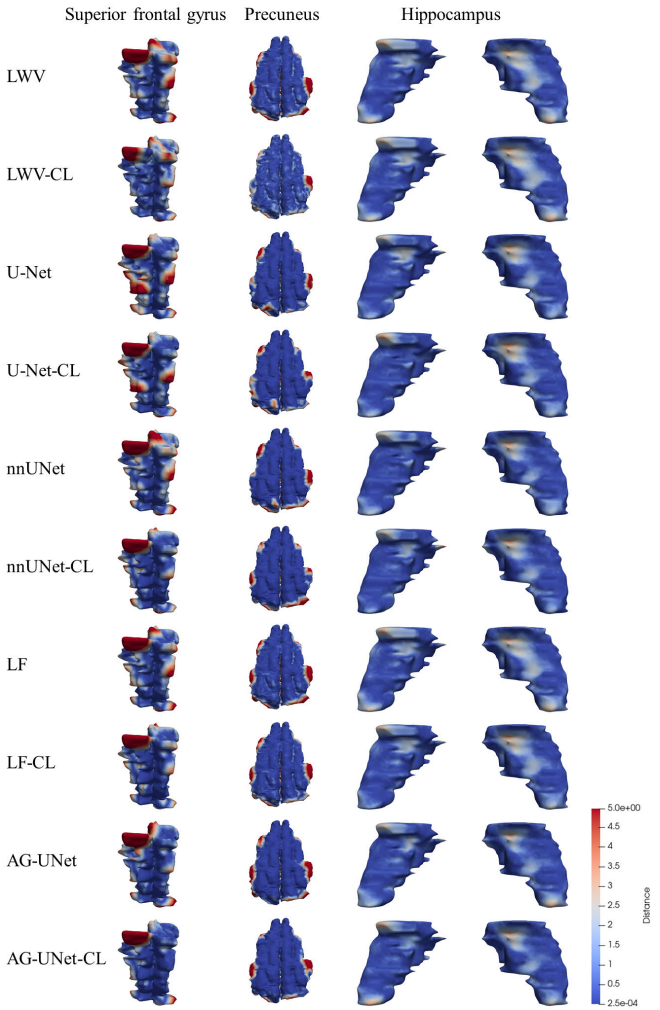

Fig. 6. Visual illustration of the surface distance between the segmentation results of different methods and ground truth on the superior frontal gyrus, precuneus, and hippocampus.

TABLE II

SEGMENTATION RESULTS ACHIEVED BY DIFFERENT METHODS ON THE IXI DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY. THE SYMBOL '*' INDICATES THAT OUR PROPOSED METHOD CAN SIGNIFICANTLY IMPROVE ITS CONVENTIONAL COUNTERPART BASED ON THE WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	*DC*	*SEN*
LWV	0.7292 ± 0.0078	0.7214 ± 0.0060
*LWV-CL	0.7796 ± 0.0105	0.7706 ± 0.0771
U-Net	0.7626 ± 0.0139	0.7502 ± 0.0174
*U-Net-CL	0.7714 ± 0.0095	0.7623 ± 0.0091
nnUNet	0.7729 ± 0.0160	0.7635 ± 0.0150
*nnUNet-CL	0.7900 ± 0.0142	0.7866 ± 0.0142
LF	0.7752 ± 0.0095	0.7662 ± 0.0070
*LF-CL	0.7862 ± 0.0110	0.7825 ± 0.0091
AG-UNet	0.7888 ± 0.0105	0.7826 ± 0.0102
*AG-UNet-CL	$\mathbf{0.8007 \pm 0.0140}$	$\mathbf{0.8026 \pm 0.0141}$

LF-CL, and AG-UNet-CL methods also achieve better results in terms of sensitivity. We also perform the Wilcoxon signed rank test in Dice coefficient results achieved by different methods. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL still achieve a significant improvement ($p < 0.05$) over LWV ($p = 0.0020$), U-Net ($p = 0.0098$), nnUNet ($p = 1.9531e-3$), LF ($p = 0.0020$), and AG-UNet ($p = 1.9531e-3$) for brain ROI segmentation task, respectively. These results further show that using our proposed multi-atlas supervised contrastive learning framework can boost the brain ROI segmentation results. Fig. 7 shows that our U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL outperforms U-Net, nnUNet, LF, and AG-UNet in most ROIs, respectively.

TABLE III

SEGMENTATION RESULTS ACHIEVED BY DIFFERENT METHODS ON THE OASIS DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY. THE SYMBOL '*' INDICATES THAT OUR PROPOSED METHOD CAN SIGNIFICANTLY IMPROVE ITS CONVENTIONAL COUNTERPART BASED ON THE WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	*DC*	*SEN*
LWV	0.7730 ± 0.0292	0.7627 ± 0.292
*LWV-CL	0.7939 ± 0.0392	0.7968 ± 0.0272
U-Net	0.8666 ± 0.0140	0.8662 ± 0.0193
*U-Net-CL	0.8877 ± 0.0120	0.8875 ± 0.0151
nnUNet	0.8734 ± 0.0130	0.8694 ± 0.0162
*nnUNet-CL	0.8918 ± 0.0124	0.8879 ± 0.0136
LF	0.8403 ± 0.0244	0.8374 ± 0.0229
*LF-CL	0.8519 ± 0.0235	0.8495 ± 0.0220
AG-UNet	0.8837 ± 0.0125	0.8835 ± 0.0151
*AG-UNet-CL	$\mathbf{0.8982 \pm 0.0117}$	$\mathbf{0.8933 \pm 0.0137}$

E. Results on OASIS

In the third group of experiments, we compare our methods with their counterparts on the OASIS for the brain ROI segmentation. The segmentation results on the OASIS dataset are reported in Table III.

As shown in Table III, the methods with our MAS-CL framework archive better segmentation results than their counterparts in terms of Dice coefficient and sensitivity. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL archive 0.0209, 0.0229, 0.0248, 0.0363, and 0.0141 improvement in Dice coefficient over LWV, U-Net, nnUNet, LF and AG-UNet, respectively. Fig. 8 plots the violin visual of Dice coefficient values achieved by U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL, and their counterparts on 32 ROIs. As shown in Fig. 8, our U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL also outperforms U-Net, nnUNet, LF, and AG-UNet in most ROIs.

We also perform the Wilcoxon signed rank test in Dice coefficient results achieved by different methods. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL still achieve a significant improvement ($p < 0.05$) over LWV ($p = 1.9369e-34$), U-Net ($p = 2.3502e-37$), nnUNet ($p = 3.4273e-37$), LF ($p = 8.0048e-38$), and AG-UNet ($p = 2.3502e-37$) for brain ROI segmentation task, respectively.

F. Results on ADNI

In the fourth group of experiments, we perform our methods and their counterparts on the ADNI dataset for hippocampus

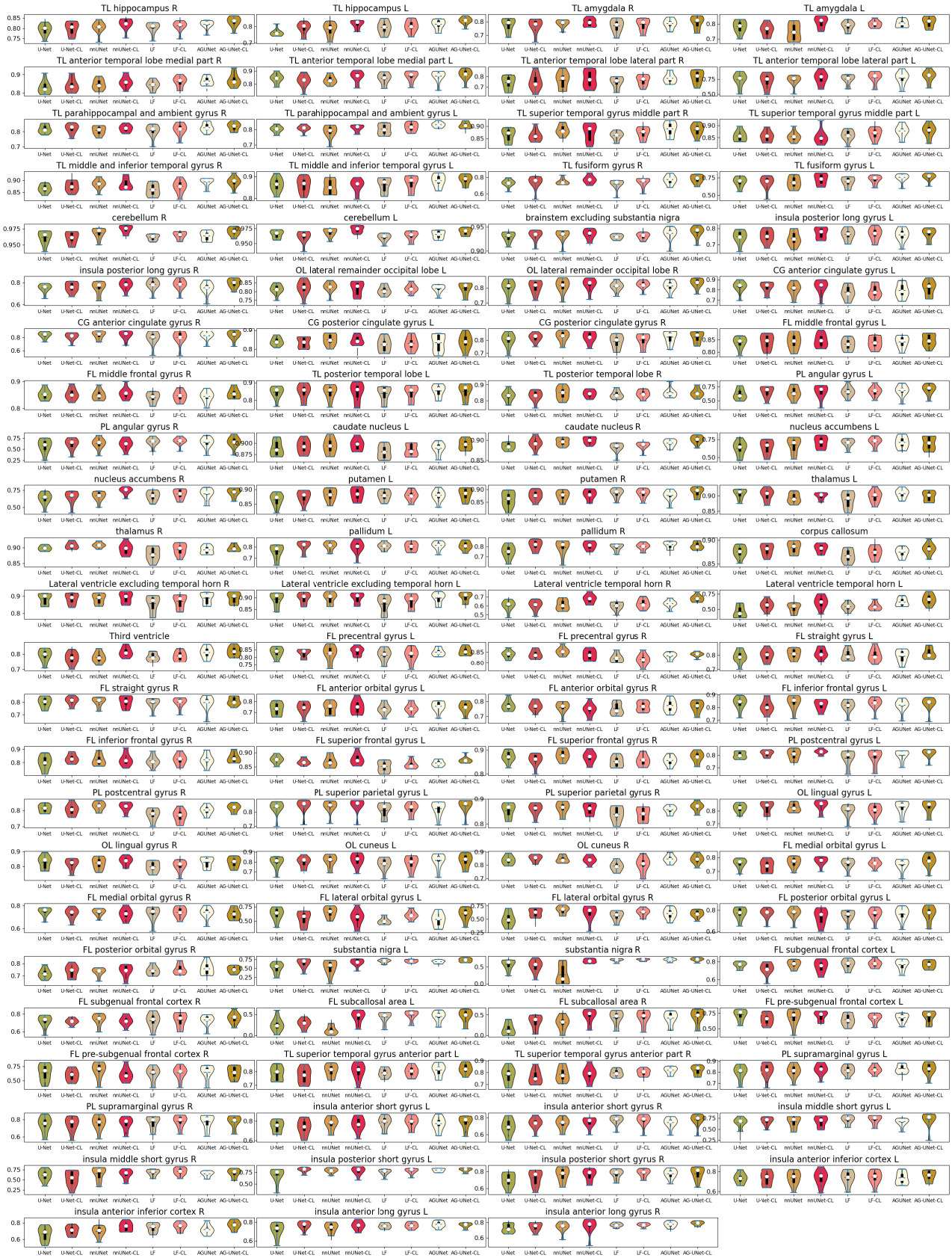

Fig. 7. Segmentation result of 95 ROIs on the IXI dataset achieved by U-Net, U-Net-CL, nnUNet, nnUNet-CL, LF, LF-CL, AG-UNet, and AG-UNet-CL in terms of Dice coefficient values.

segmentation. Table IV reports the results achieved by different methods on the ADNI dataset.

One can observe from Table IV, our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL still achieve better

Fig. 8. Segmentation result of 32 ROIs on the OASIS dataset achieved by U-Net, U-Net-CL, nnUNet, nnUNet-CL, LF, LF-CL, AG-UNet, and AG-UNet-CL in terms of Dice coefficient values.

TABLE IV

SEGMENTATION RESULTS ACHIEVED BY DIFFERENT METHODS ON THE ADNI DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY. THE SYMBOL '*' INDICATES THAT OUR PROPOSED METHOD CAN SIGNIFICANTLY IMPROVE ITS CONVENTIONAL COUNTERPART BASED ON THE WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	*DC*	*SEN*
LWV	0.8415 ± 0.0270	0.8660 ± 0.269
*LWV-CL	0.8571 ± 0.0479	0.8849 ± 0.0408
U-Net	0.8878 ± 0.0550	0.8776 ± 0.0550
*U-Net-CL	0.9107 ± 0.0493	0.9187 ± 0.0431
nnUNet	0.8860 ± 0.0849	0.8843 ± 0.0826
*nnUNet-CL	0.9108 ± 0.0279	0.9188 ± 0.0279
LF	0.8403 ± 0.0244	0.8374 ± 0.0229
*LF-CL	0.8776 ± 0.0399	0.9033 ± 0.2782
AG-UNet	0.8980 ± 0.0191	0.8894 ± 0.0287
*AG-UNet-CL	$\mathbf{0.9121 \pm 0.0158}$	$\mathbf{0.9191 \pm 0.0206}$

results on one brain ROI segmentation task. For example, our AG-UNet-CL gains 0.0141 and 0.0297 improvement over AG-UNet in terms of Dice coefficient and sensitivity, respectively. We perform the Wilcoxon signed rank test in Dice coefficient results achieved by different methods. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL still achieve a significant improvement ($p < 0.05$) over LWV ($p = 5.2463e-4$), U-Net ($p = 3.5694e-08$), nnUNet ($p = 3.7583e-07$), LF ($p = 3.5694e-08$), and AG-UNet ($p = 1.4758e-07$) for hippocampus segmentation task, respectively. These results suggest that our multi-atlas supervised contrastive learning framework also can improve the segmentation performance on one brain ROI segmentation task.

Fig. 9 plots the surface distance between the hippocampus segmentation results of different methods and ground truth. As can be seen from Fig. 9, the methods with our MAS-CL framework generate better visual quality, compared with their counterparts, respectively. These results further demonstrate the effectiveness of our proposed multi-atlas supervised contrastive learning framework in hippocampus segmentation tasks on the ADNI dataset.

G. Results on CC359

In the fifth group of experiments, we validate our methods on the CC359 dataset for hippocampus segmentation, with the results reported in Table V.

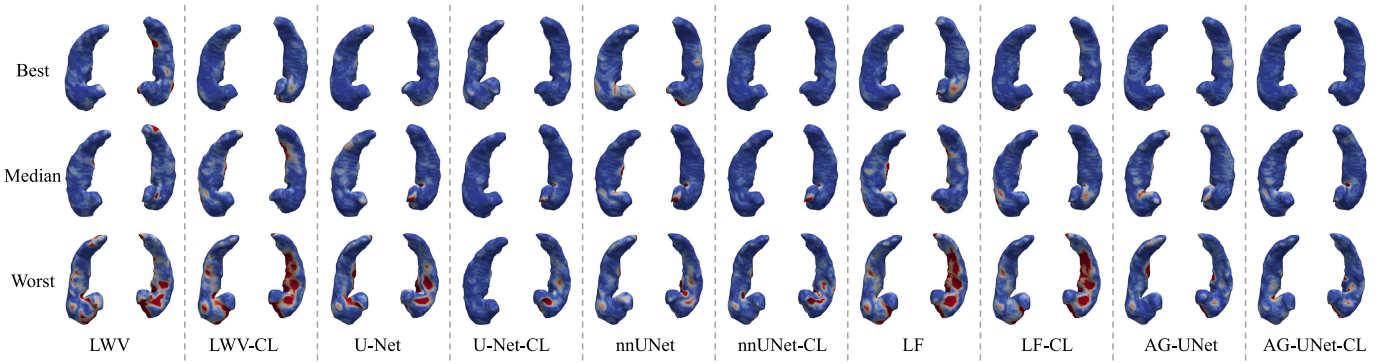

Fig. 9. Visual illustration of the surface distance between the hippocampus segmentation results of different methods and ground truth on the ADNI dataset.

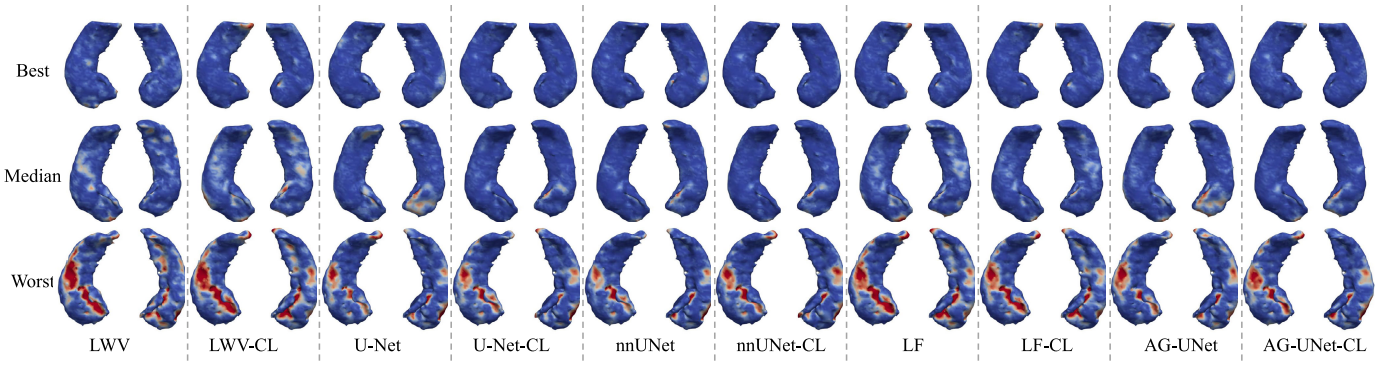

Fig. 10. Visual illustration of the surface distance between the hippocampus segmentation results of different methods and ground truth on the CC359 dataset.

TABLE V

SEGMENTATION RESULTS ACHIEVED BY DIFFERENT METHODS ON THE CC359 DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY. THE SYMBOL '*' INDICATES THAT OUR PROPOSED METHOD CAN SIGNIFICANTLY IMPROVE ITS CONVENTIONAL COUNTERPART BASED ON WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	*DC*	*SEN*
LWV	0.8710 ± 0.0312	0.8704 ± 0.0190
*LWV-CL	0.8943 ± 0.0396	0.8781 ± 0.0274
U-Net	0.9218 ± 0.0422	0.9150 ± 0.0361
*U-Net-CL	$\mathbf{0.9385 \pm 0.0352}$	0.9356 ± 0.0352
nnUNet	0.9278 ± 0.0511	0.9281 ± 0.0425
*nnUNet-CL	0.9384 ± 0.0352	$\mathbf{0.9512 \pm 0.0206}$
LF	0.9009 ± 0.0384	0.8889 ± 0.0242
*LF-CL	0.9152 ± 0.0380	0.9032 ± 0.0221
AG-UNet	0.9193 ± 0.0389	0.9215 ± 0.0219
*AG-UNet-CL	0.9337 ± 0.0378	0.9386 ± 0.0206

As shown in Table V, our methods also achieve better results than their counterparts in both Dice coefficient and sensitivity. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL achieve 0.0233, 0.0167, 0.0106, 0.0143, and 0.0144 improvement over LWV, U-Net, nnUNet, LF, and AG-UNet, respectively. We also perform the Wilcoxon signed rank test in Dice coefficient results achieved by different methods. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL still achieve a significant improvement ($p < 0.05$) over LWV ($p = 2.8556e-51$), U-Net ($p = 3.5872e-56$), nnUNet ($p = 9.8809e-51$), LF ($p = 4.1842e-56$), and AG-UNet ($p = 3.9272e-56$) for brain ROI segmentation task, respectively.

TABLE VI

SEGMENTATION RESULTS ACHIEVED BY U-NET ON THE LONI-LPBA40 DATASET WITH DIFFERENT DATA AUGMENTATION METHODS. THE SYMBOL '*' INDICATES THAT OUR U-NET-CL METHOD ACHIEVES SIGNIFICANT IMPROVEMENT OVER THE COMPETING METHOD BASED ON THE WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	*DC*	*SEN*
*U-Net	0.7832 ± 0.0118	0.7759 ± 0.0134
*U-Net-RO	0.7908 ± 0.0156	0.7900 ± 0.0180
*U-Net-RE	0.7907 ± 0.0111	0.7827 ± 0.0133
U-Net-CL	0.7988 ± 0.0116	0.7959 ± 0.0134

Fig. 10 plots the surface distance between the hippocampus segmentation results of different methods and ground truth. Our LWV-CL, U-Net-CL, nnUNet-CL, LF-CL, and AG-UNet-CL also generate better visual quality, compared with their counterparts.

H. Comparison With Data Augmentation Methods

Data augmentation methods can help train deep networks. To compare our proposed MAS-CL framework with the data augmentation method, we compared our methods with two commonly used data augmentation methods in deep learning-based medical image segmentation, *i.e.*, rotation, and registration. Specifically, we train the U-Net with the rotation

TABLE VII

THE DICE COEFFICIENT VALUES ACHIEVED BY DIFFERENT METHODS ON THE LONI-LPBA40 DATASETS. SYMBOL # IS THE TRAINING NUMBER. THE SYMBOL '*' INDICATES THAT OUR AG-UNET-CL METHOD ACHIEVES SIGNIFICANT IMPROVEMENT OVER THE STATE-OF-THE-ART METHOD BASED ON THE WILCOXON SIGNED RANK TEST IN TERMS OF *DC*

Method	#	Dice	SEN
*DeepNAT	16	0.7789 ± 0.0271	0.7743 ± 0.0231
*SLANT	16	0.7873 ± 0.0490	0.7883 ± 0.0516
*DARTS	16	0.7828 ± 0.0126	0.7929 ± 0.0156
*AG-UNet	16	0.8067 ± 0.0383	0.8074 ± 0.0381
*LF	16	0.8021 ± 0.0105	0.8078 ± 0.0107
LF-CL	10	0.8084 ± 0.0103	0.8136 ± 0.0105
AG-UNet-CL	10	$\mathbf{0.8186 \pm 0.0113}$	$\mathbf{0.8246 \pm 0.0113}$

and registration data augmentation methods, denoted as U-Net-RO and U-Net-RE, respectively. More specifically, we rotate the 3D MR image across sagittal, coronal, and axial sections. Meanwhile, we use the FLIRT method in the FSL [54] toolbox to register each pair of training data. Hence, we have 40 and 100 training data to train the deep learning model for brain ROI segmentation on LONI-LPBA40 datasets. The experimental results are reported in Table VI.

One can observe from Table VI, all our U-Net-CL and the data augmentation methods can improve the brain ROI segmentation on LONI-LPBA40 datasets. Furthermore, our U-Net-CL achieves a better brain ROI segmentation performance, when compared with the U-Net trained with data augmentation methods. We also perform the Wilcoxon signed rank test in Dice coefficient results achieved by our UNet-CL with the data augmentation methods. Our UNet-CL method suggests significant improvement ($p < 0.05$) over U-Net-RO ($p = 8.8575e-05$) and U-Net-RE ($p = 8.8575e-05$) for brain ROI segmentation task, respectively. These results suggest that our proposed multi-atlas supervised contrastive learning framework can learn a better latent representation of MR images than data augmentation by rotation and registration operations.

I. Comparison With Deep Learning Methods

In this section, we compared our LF-CL and AG-UNet-CL methods with four deep learning methods for 54 brain ROI segmentation on LONI-LPBA40 dataset, including DeepNAT [37], SLANT [56], DARTS [35], and AG-UNet [38]. To better demonstrate our proposed MAS-CL framework can train an effective model with limited data, we train the competing state-of-the-art methods with more training data. Herein, we perform 5-fold cross-validation for the competing methods. Hence, the competing methods have 16 subjects to train the deep models on the LONI-LPBA40 dataset. The state-of-the-art methods use the default settings provided in the literature. In addition, we also reported the LF method trained with 16 subjects.

As shown in Table VII, our proposed LF-CL and AG-UNet-CL achieve the second-best and best Dice coefficient and sensitivity values on the LONI-LPBA40 dataset. For example, our AG-UNet-CL achieves 0.0397, 0.0313, 0.0358, 0.0119, and 0.0165 improvement over DeepNAT, SLANT, DARTS, AG-UNet, and LF on LONI-LPBA40 dataset, respectively. It is worth noting that our LF-CL and AG-UNet-CL only use 10 MR images to train our LF-CL and AG-UNet-CL model on the LONI-LPBA40 dataset, while the competing methods use 16 MR images to train their corresponding models. In addition, the segmentation results of the LF and AG-UNet methods using 16 MR images are better than those using 10 MR images. It implies that the training number is an important factor for deep learning methods. However, using our MAS-CL framework, the LF-CL and AG-UNet-CL still achieve better segmentation performance with fewer training data. These results further demonstrate using our multi-atlas supervised contrastive learning framework can learn better feature representation with a limited amount of labeling images for brain ROI segmentation tasks.

We perform the Wilcoxon signed rank test in Dice coefficient results achieved by our AG-UNet-CL with state-of-the-art deep learning methods. Our AG-UNet-CL method suggests significant improvement ($p < 0.05$) over DeepNAT ($p = 8.8575e-05$), SLANT ($p = 1.2042e-05$), DARTS ($p = 8.8575e-05$), AG-UNet ($p = 1.4013e-04$) and LF ($p = 8.8575e-05$) for brain ROI segmentation task, respectively.

V. DISCUSSION

In this section, we first study the influence of search region size for our proposed multi-atlas supervised contrastive learning framework and LF-CL method. Then, we study the influence of the number of atlases for our proposed method. Finally, we present the limitations of this work as well as possible future research directions.

A. Influence of Search Region Size for Multi-Atlas Supervised Contrastive Learning

The size of search region $R(v_i)$ is an important factor in multi-atlas segmentation methods. To investigate the influence of search region size for our proposed multi-atlas supervised contrastive learning step, we generate sample pairs in larger search region sizes to pre-train the networks for brain ROI segmentation with MR images. Herein, we set the search region size as $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $5 \times 5 \times 5$, respectively. For a fair comparison, we randomly generate 20 sample pairs in each mini-batch. The results achieved by LF-CL with different search region sizes and training epochs in multi-atlas supervised contrastive learning step on the LONI-LPBA40 dataset are shown in Fig. 11.

We perform the Wilcoxon signed rank test in Dice coefficient results achieved by the LF-CL with different search region sizes and training epochs in the multi-atlas supervised contrastive learning step. The larger search region size and training epochs do not show significant differences with smaller search region size and training epochs. These results demonstrate that our MAS-CL method is not very sensitive to the search region size when compared with the conventional multi-atlas methods. The possible reason is that the deep learning model can learn high-level contextual features from

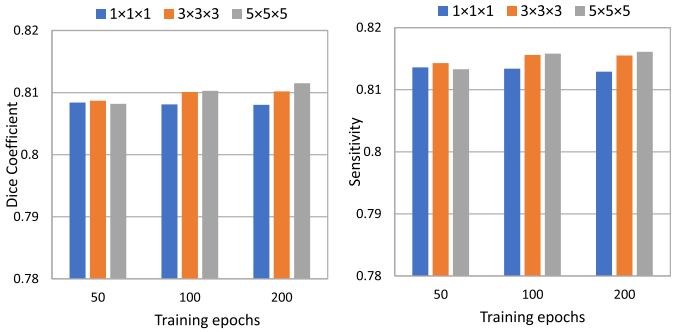

Fig. 11. Segmentation results achieved by LF-CL with different search region size and training epochs in multi-atlas supervised contrastive learning step on LONI-LPBA40 dataset.

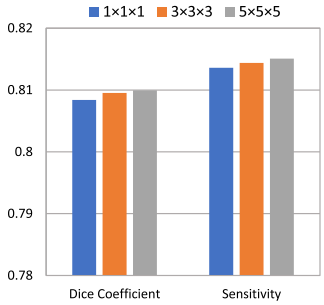

Fig. 12. Segmentation results achieved by LF-CL with different search region sizes in label fusion step on LONI-LPBA40 dataset.

multiple local receptive fields, which can also provide non-local information about brain MR images. Meanwhile, the experimental results also demonstrate our proposed MAS-CL framework has good convergence. However, as shown in Fig. 11, we find that a larger search region size in a multi-atlas supervised contrastive learning step can improve the performance of our LF-CL method with a larger training epoch number.

B. Influence of Search Region Size for LF-CL Method

The non-local strategy can improve segmentation performance in the label fusion step. Thus, we also study the influence of search region size for our LF-CL method. We set the search region size at the label fusion stage as $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $5 \times 5 \times 5$, respectively. To control variables, we use the pre-trained backbone with $1 \times 1 \times 1$ search region size to initialize model parameters. The results achieved by LF-CL with different search region sizes in the label fusion step on the LONI-LPBA40 dataset are shown in Fig. 12.

As shown in Fig. 12, we find that larger search region size in our LF-CL method also can improve the performance of brain ROI segmentation on LONI-LPBA40 datasets. However, we perform the Wilcoxon signed rank test in Dice coefficient results achieved by the LF-CL with different search region sizes in the label fusion step. The larger search region size also does not show significant differences with the smaller search region. These results further indicate that deep learning-based methods can learn better local features of brain MR images, which are not very sensitive to the search region size.

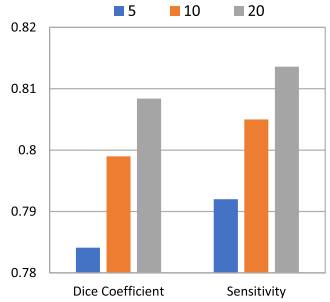

Fig. 13. Segmentation results achieved by LF-CL with different atlas numbers on LONI-LPBA40 dataset.

C. Influence of Atlas Number

We further study the influence of atlas number for our proposed multi-atlas supervised contrastive learning framework. We set the atlas numbers as 5, 10 and 20. The segmentation results achieved by LF-CL with different atlas numbers on the LONI-LPBA40 dataset are shown in Fig. 13.

One can observe that the larger atlas number can improve the brain ROI segmentation results on LONI-LPBA40 datasets. We further perform the Wilcoxon signed rank test in Dice coefficient results achieved by the LF-CL with different atlas numbers. LF-CL with 20 atlases shows a significant improvement ($p < 0.05$) over LF-CL with 5 ($p = 8.8575e - 05$) and 10 ($p = 8.8575e - 05$) atlases. There are two main reasons. First, more atlas images provide more sample pair combinations to improve the feature representation ability in the multi-atlas supervised contrastive learning step. Second, more atlas images can provide more sufficient anatomical prior of brain structure to improve the segmentation performance in the label fusion step.

D. Discussion of Results on All Datasets

We perform our MAS-CL framework on five datasets (*i.e.*, LONI-LPBA40, IXI, OASIS, ADNI, and CC359). LONI-LPBA40, IXI, and OASIS datasets are used for multiple ROI segmentation tasks, and the remaining ADNI and CC359 datasets are used for the hippocampus segmentation. The results on all datasets suggest that using the pre-trained parameters to initial the backbones can significantly improve the segmentation performance. We still have two observations as follows.

Firstly, training models with larger datasets yield better segmentation results. For example, the DC values of AG-UNet-CL are 0.8007, 0.8186, and 0.8982 for multiple ROI segmentation on IXI, LONI-LPBA40, and OASIS datasets with 5, 10, and 100 training data, respectively. Similarly, the DC values of AG-UNet-CL are 0.9121 and 0.9337 for hippocampus segmentation on ADNI and CC395 datasets with 20 and 166/167 training data, respectively.

Secondly, the multiple ROI segmentation tasks benefit more from the anatomical structure prior to guide the segmentation process. As shown in Tables I, II, and III, the AG-UNet-CL achieves the best segmentation performance for multiple ROI segmentation tasks. Tables IV and V, the U-Net-CL, nnUNet-CL, and AG-UNet-CL achieve similar

segmentation performance for hippocampus segmentation. U-Net-CL and nnUNet-CL even achieve better segmentation performance on the CC359 dataset. This discrepancy likely arises because multiple ROI segmentation tasks are inherently more complex and require detailed anatomical structure priors to accurately capture the brain's complex anatomy. In contrast, hippocampus segmentation is relatively simpler, requiring only the differentiation between the hippocampus and surrounding regions. Especially, the U-Net-CL and nnUNet-CL models are able to effectively learn features with the increased training data available on the CC359 dataset.

E. Limitations and Future Work

There are still several limitations in the current work. First, multi-atlas methods are widely used in the field of medical image segmentation. In our work, we only apply our MAS-CL framework to the brain ROI segmentation task. In future work, we will evaluate our proposed MAS-CL in other medical image segmentation tasks. Second, our proposed MAS-CL method is a general framework for end-to-end networks, we only evaluate it with three classical methods for brain ROI segmentation. In the future, more commonly used networks, *e.g.*, TransUNet [57] and Swin-UNet [58], *etc.*, can be trained by our proposed MAS-CL framework for medical image segmentation tasks.

VI. Conclusion

In this paper, we propose an end-to-end multi-atlas supervised contrastive learning framework for brain ROI segmentation with MR images. In our MAS-CL framework, we use supervised anatomical structure information of brain MR images to pre-train the network. By using our proposed framework, we can easily generate the contrastive sample pairs at voxel-level to train the end-to-end networks. Extensive experimental results on LONI-LPBA40, IXI, OASIS, ADNI, and CC359 datasets demonstrate our proposed contrastive learning framework can learn more useful feature representation for brain ROI segmentation.

References

[1] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, Jan. 2011.

[2] D. Zhang, Q. Guo, G. Wu, and D. Shen, "Sparse patch-based label fusion for multi-atlas segmentation," in *Proc. Int. Workshop Multimodal Brain Image Anal.* Cham, Switzerland: Springer, 2012, pp. 94–102.

[3] H. Wang and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion and corrective learning—An open source implementation," *Frontiers Neuroinform.*, vol. 7, p. 27, 2013.

[4] S. Pang et al., "SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 262–273, Jan. 2021.

[5] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos, "Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2196–2210, Oct. 2018.

[6] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, "Fast ScanNet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1948–1958, Aug. 2019.

[7] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2020.

[8] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, Feb. 2016.

[9] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Med. Image Anal.*, vol. 43, pp. 157–168, Jan. 2018.

[10] S. Ali et al., "A deep learning framework for quality assessment and restoration in video endoscopy," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101900.

[11] Z. Zhu et al., "3D pyramid pooling network for abdominal MRI series classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1688–1698, Apr. 2022.

[12] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 574–584.

[13] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[14] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[15] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 776–794.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[17] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[18] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[19] I. Dave, R. Gupta, M. N. Rizve, and M. Shah, "TCLR: Temporal contrastive learning for video representation," *Comput. Vis. Image Understand.*, vol. 219, Jun. 2022, Art. no. 103406.

[20] Y. Tian, Y. Yan, G. Zhai, L. Chen, and Z. Gao, "CLSA: A contrastive learning framework with selective aggregation for video rescaling," *IEEE Trans. Image Process.*, vol. 32, pp. 1300–1314, 2023.

[21] Z. Lin, W. Pei, F. Chen, D. Zhang, and G. Lu, "Pedestrian detection by exemplar-guided contrastive learning," *IEEE Trans. Image Process.*, vol. 32, pp. 2003–2016, 2023.

[22] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, Aug. 2015.

[23] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen, "A generative probability model of joint label fusion for multi-atlas based brain segmentation," *Med. Image Anal.*, vol. 18, no. 6, pp. 881–890, Aug. 2014.

[24] L. Wang et al., "Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation," *NeuroImage*, vol. 89, pp. 152–164, Apr. 2014.

[25] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, Oct. 2006.

[26] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.

[27] G. Sanroma et al., "Learning non-linear patch embeddings with neural networks for label fusion," *Med. Image Anal.*, vol. 44, pp. 143–155, Feb. 2018.

[28] L. Sun, W. Shao, M. Wang, D. Zhang, and M. Liu, "High-order feature learning for multi-atlas based label fusion: Application to brain segmentation with MRI," *IEEE Trans. Image Process.*, vol. 29, pp. 2702–2713, 2020.

[29] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[31] Ö. **c**Ciccek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432.

[32] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[33] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1290–1299.

[34] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[35] A. Kaku et al., "DARTS: DenseUnet-based automatic rapid tool for brain segmentation," 2019, *arXiv:1911.05567*.

[36] D. Carmo, B. Silva, C. Yasuda, L. Rittner, and R. Lotufo, "Hippocampus segmentation on epilepsy and Alzheimer's disease studies with multiple convolutional neural networks," *Heliyon*, vol. 7, no. 2, Feb. 2021, Art. no. e06226.

[37] C. Wachinger, M. Reuter, and T. Klein, "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, vol. 170, pp. 434–445, Apr. 2018.

[38] L. Sun, W. Shao, D. Zhang, and M. Liu, "Anatomical attention guided deep networks for ROI segmentation of brain MR images," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2000–2012, Jun. 2020.

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[40] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. NIPS*, Dec. 2020, pp. 9912–9924.

[41] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[42] X. Zhao et al., "Contrastive learning for label efficient semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10623–10633.

[43] X. He, L. Fang, M. Tan, and X. Chen, "Intra- and inter-slice contrastive learning for point supervised OCT fluid segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 1870–1881, 2022.

[44] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, May 2023.

[45] D. W. Shattuck et al., "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.

[46] A. Hammers et al., "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Human Brain Map.*, vol. 19, no. 4, pp. 224–247, Aug. 2003.

[47] I. Faillenot, R. A. Heckemann, M. Frot, and A. Hammers, "Macroanatomy and 3D probabilistic atlas of the human insula," *NeuroImage*, vol. 150, pp. 88–98, Apr. 2017.

[48] 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," *J. Cognit. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[49] C. R. Jack et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.

[50] R. Souza et al., "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, Apr. 2018.

[51] F. Shi, L. Wang, Y. Dai, J. H. Gilmore, W. Lin, and D. Shen, "LABEL: Pediatric brain extraction using learning-based meta-algorithm," *NeuroImage*, vol. 62, no. 3, pp. 1975–1986, Sep. 2012.

[52] N. J. Tustison et al., "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010.

[53] A. Madabhushi and J. K. Udupa, "New methods of MR image intensity standardization via generalized scale," *Med. Phys.*, vol. 33, no. 9, pp. 3426–3434, Aug. 2006.

[54] S. M. Smith et al., "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, Jan. 2004.

[55] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, Mar. 2009.

[56] Y. Huo et al., "3D whole brain segmentation using spatially localized atlas network tiles," *NeuroImage*, vol. 194, pp. 105–119, Jul. 2019.

[57] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[58] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2023, pp. 205–218.