High-order Feature Learning for Multi-atlas based Label Fusion: Application to Brain Segmentation with MRI

Liang Sun, Wei Shao, Mingliang Wang, Daoqiang Zhang*, Mingxia Liu*

***Abstract*—Multi-atlas based segmentation methods have shown their effectiveness in brain regions-of-interesting (ROIs) segmentation, by propagating labels from multiple atlases to a target image based on the similarity between patches in the target image and multiple atlas images. Most of the existing multi-atlas based methods use image intensity features to calculate the similarity between a pair of image patches for label fusion. In particular, using only low-level image intensity features cannot adequately characterize the complex appearance patterns (*e.g.*, the high-order relationship between voxels within a patch) of brain magnetic resonance (MR) images. To address this issue, this paper develops a high-order feature learning framework for multi-atlas based label fusion, where high-order features of image patches are extracted and fused for segmenting ROIs of structural brain MR images. Specifically, an unsupervised feature learning method (*i.e.*, means-covariances restricted Boltzmann machine, mcRBM) is employed to learn high-order features (*i.e.*, mean and covariance features) of patches in brain MR images. Then, a group-fused sparsity dictionary learning method is proposed to jointly calculate the voting weights for label fusion, based on the learned high-order and the original image intensity features. The proposed method is compared with several state-of-the-art label fusion methods on ADNI, NIREP and LONI-LPBA40 datasets. The Dice ratio achieved by our method is 88.30%, 88.83%, 79.54% and 81.02% on left and right *hippocampus* on the ADNI, NIREP and LONI-LPBA40 datasets, respectively, while the best Dice ratio yielded by the other methods are 86.51%, 87.39%, 78.48% and 79.65% on three datasets, respectively.**

***Index Terms*—high-order features, multi-atlas, ROI segmentation**

I. Introduction

ACCURATE segmentation of structural magnetic resonance (MR) images is a key step for clinical diagnosis and pathology detection. As an example, the *hippocampus* plays an important role in memory and cognition. Abnormalities of the *hippocampus* may lead to many diseases such as Alzheimer's disease [1]–[5] and Schizophrenia [6]–[8]. Thus, it is essential to accurately segment the *hippocampus* from the whole brain for clinical analysis and brain disease diagnosis [9], [10]. The most straightforward solution is the manual annotation for regions-of-interest (ROIs) in brain MR images, which is not only time-consuming but also error-prone. It is highly desired to develop automatic methods for accurate segmentation of brain ROIs.

Among various automated segmentation approaches, multi-atlas based methods have shown to be useful in medical image segmentation in recent years [11]–[17]. The basic assumption of multi-atlas segmentation is that the target image voxel should have the same label as the atlas image voxel if they have the similar local appearance. Generally, the multi-atlas based segmentation methods consist of two key components, including 1) image registration [18]–[25] and 2) label fusion [12]–[15]. In the *image registration* step, the atlas images will be registered onto the target image by using affine registration and deformable registration algorithms. In the *label fusion* step, labels of multiple atlases are propagated to the target image, by using a specific metric to measure the similarity between each atlas and the target image. In this work, we focus on the latter step of multi-atlas based segmentation, *i.e.*, label fusion.

Most of the existing label fusion methods in multi-atlas segmentation are based on the weighted voting strategy [11]–[15], where each selected candidate patch in an atlas image will determine the final label of its corresponding patch in the target image according to the voting weight. Specifically, the voting-based label fusion methods propagate labels from atlases to the target image, based on the pairwise similarity between the target patch and the selected candidate patches on in each atlas image. That is, if a candidate patch is very similar to the target patch, it will be assigned with a large weight in label propagation. Therefore, an essential component in label fusion is to accurately compute the voting weights for candidate patches, while these weights are usually defined as the pairwise similarity between target and candidate patches.

Existing voting-based label fusion strategies typically employ the image intensity to calculate the pairwise similarity between patches. However, using only intensity feature cannot adequately describe the complex appearance patterns (*e.g.*, the high-order relationship between voxels) of a patch in brain MR images, since brain MR images usually have large inter-

L. Sun, W. Shao, M. Wang and D. Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China. M. Liu is with the Department of Information Science and Technology, Taishan University, Taian 271000, China, and also with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China.

*Corresponding authors: D. Zhang (dqzhang@nuaa.edu.cn) and M. Liu (mxliu1226@gmail.com).

This work was supported by the National Key Research and Development Program of China (No. 2018YFC2001602) and the National Natural Science Foundation of China (Nos. 61876082, 61861130366, 61703301, 61732006 and 61902183), the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship (No. NAF\R1\180371), the Fundamental Research Funds for the Central Universities (No. NP2018104), Taishan Scholar Program of Shandong Province in China, Shandong Natural Science Foundation for Distinguished Young Scholar in China (No. ZR2019YQ27), and Scientific Research Foundation of Taishan University (No. Y-01-2018019).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2952079, IEEE Transactions on Image Processing

variation (among different subjects) and intra-variation (within the same subject) regarding image intensity. Unfortunately, previous methods seldom consider such higher-order relationship between voxels within a patch. Intuitively, utilizing such high-order relationship could further promote the performance of multi-atlas based label fusion methods.

To this end, in this paper, a **H**igh-order feature learning framework is proposed for multi-atlas based **L**abel **F**usion (HLF). Specifically, the means-covariances Restricted Boltzmann Machine (mcRBM) [26], [27] is introduced to learn high-order patch-level features of brain MR images firstly, which can effectively describe the higher-order relationship between voxels within a patch. Note that mcRBM can not only learn the mean intensity (denoted as *mean feature*) of voxels within a patch, but also acquire the high-order intensity dependency (denoted as *covariance feature*) between each pair of voxels in the patch. Based on the learned mean and covariance features as well as the intensity feature, a group-fused sparsity dictionary learning method is developed to jointly calculate the similarity (corresponding to voting weights) between the target and candidate patches. Finally, the label fusion based on the learned voting weights is performed for multi-atlas based ROI segmentation in brain MRIs.

The preliminary work of this method was reported on ISBI [28]. In this journal paper, new contributions have been offered in the following aspects: 1) developing a new strategy for learning voting weights based on both high-order and original intensity features; 2) evaluating the effectiveness of the proposed method on two additional datasets (*i.e.*, ADNI[1] and LONI-LPBA40 [29]); 3) performing the affine and deformable registration on brain MR images; 4) investigating the influence of primary parameters in our method; 5) comparing our method with state-of-the-art approaches for multi-atlas based segmentation; and 6) performing statistical significance analysis for our method versus the competing methods.

The major contributions of this paper are three-fold. *First*, an unsupervised feature learning method is leveraged to learn the high-order features representation (*i.e.*, mean and covariance features) of patches for multi-atlas based brain ROI segmentation. *Second*, a group-fused sparsity dictionary learning approach is developed to jointly calculate the voting weights for label fusion, based on the learned high-order features and the original intensity feature. *Third*, the proposed method is evaluated in the task of brain ROI segmentation on three datasets, including ADNI, NIREP and LONI-LPBA40. Experimental results have demonstrated that our method achieves superior performance in brain ROI segmentation.

The rest of the paper is organized as follows. Section II reviews the most relevant studies. The proposed method is introduced in Section III. In Section IV, we present materials, experimental settings, competing methods, and experimental results. In Section V, we compare the proposed method with deep learning methods, perform generalization analysis, study the influence of several essential components, and present the limitations of the current work as well as future research directions. Section VI concludes this paper.

II. Related Work

Many label fusion strategies have been proposed for multi-atlas segmentation [12]–[17], [30]–[45]. Given a target image $\boldsymbol{T}$, the goal of multi-atlas segmentation is to automatically determine the label map $\boldsymbol{L_T}$ for the target image based on multiple well-labeled atlases. Let $\boldsymbol{A}$ denote the atlas set that contains a set of atlas images and their corresponding label maps. Denote $\boldsymbol{I}_k$ and $\boldsymbol{L}_k$ as the k-th atlas image and its corresponding label map, respectively. The basic assumption of multi-atlas segmentation is that the voxel in the target image should have the same label as the corresponding voxel in the atlas image if they have the similar local appearance. To generate accurate segmentation, each atlas image and its label map are first registered onto the target image. Then, in the label fusion step, the similarity between the target voxel and its corresponding voxels in atlas images can be calculated. Based on the learned similarity/weight, one can finally produce the label for each voxel in the target image.

As a widely used label fusion method, the local-weighted voting (LWV) [12] strategy first computes the pairwise similarity between the target patch and the patch at the same location in each atlas image. Then, the similarity $w(y, x_{k,y})$ is used as the voting weight via

$$w(y, x_{k,y}) = exp^{\frac{||P(y)-P(x_{k,y})||_2^2}{\sigma}}, \tag{1}$$

where y is the to-be-segmented voxel in the target image, and $x_{k,y}$ is the voxel in the k-th atlas image, which have the same coordinate with y in the space of $\boldsymbol{T}$. Here, $P(y)$ is a cubic patch centered at the voxel y and $P(x_{k,y})$ is a cubic patch centered at the voxel $x_{k,y}$, which encode the local appearance of y and $x_{k,y}$, respectively. $||\cdot||$ is normalized l_2-norm computed between each intensity of the elements of the patches. In addition, the term $\sigma = \min_{x_{k,y}} ||P(y) - P(x_{k,y})||_2 + \varepsilon$, where ε is a small constant to ensure the numerical stability. The disadvantage of LWV is that its segmentation performance is very sensitive to the error in registration from atlas images to the target image.

To alleviate the negative influence of possible registration errors, a non-local mean patch-based method (PBM) has been proposed [13], which propagates the labels not only from the voxel at the same location in atlases, but also from its neighboring locations. Specifically, PBM seeks multiple patches from each atlas for label fusion, which based on the pairwise similarity between the target and candidate patches within a certain region via

$$w(y, x_{k,j}) = exp^{\frac{||P(y)-P(x_{k,j})||_2^2}{\sigma}}, \tag{2}$$

where $x_{k,j}$ is the j-th voxels in y's neighboring regions $N(y)$ in the k-th atlas. Also, the term $\sigma = \min_{x_{k,j}} ||P(y) - P(x_{k,j})||_2 + \varepsilon$. Previous studies have shown that PBM is more robust to registration errors in comparison to LWV, thus generating improved segmentation performance [13]. However, both PBM and LWV methods calculate voting weights independently for each patch, ignoring the fact that different patches may produce similar labeling errors if they have the similar appearance.

[1] http://adni.loni.usc.edu/

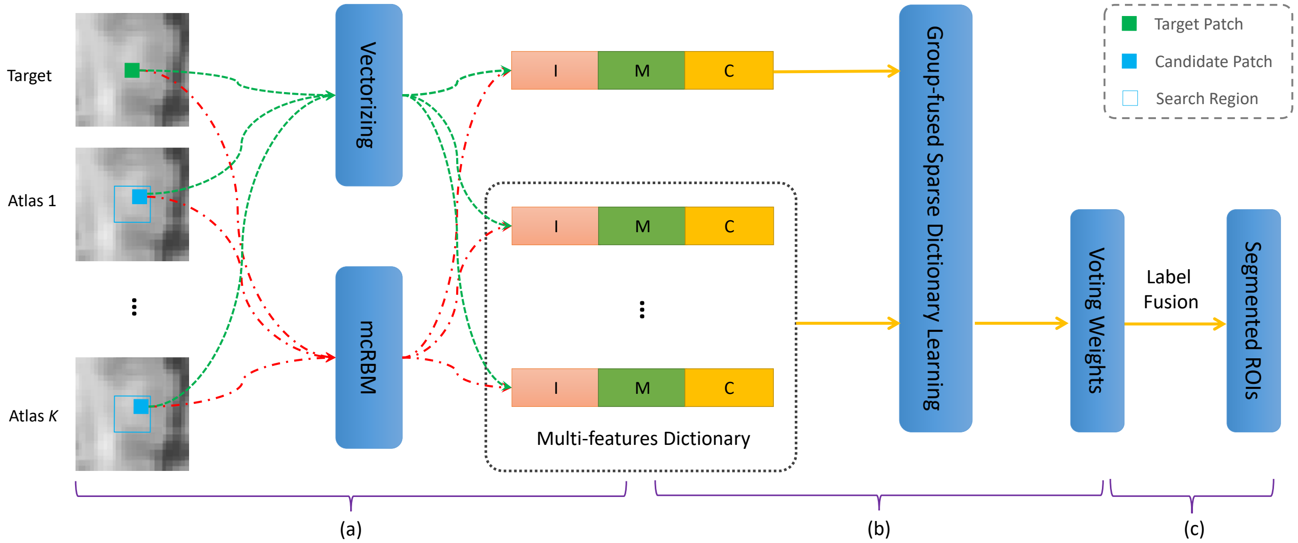


Fig. 1. Overview of proposed high-order feature learning framework for multi-atlas based label fusion. The blue rectangles on the atlases represent the search region. The blue patches in the atlas images represent the selected candidate patches and the green patch in the target image represents the target patch. At stage (a), the mcRBM [26] model is used to extract the high-order patch-level features (*i.e.*, mean feature and covariance feature) of brain MR images. At stage (b), a group-fused sparsity dictionary learning method is developed to jointly calculate the similarity (*i.e.*, voting weight) between the target and candidate patches, based on the learned mean and covariance features as well as the original image intensity features. At stage (c), label fusion using the learned voting weights is performed for multi-atlas based ROI segmentation with brain MR images.

Accordingly, a joint label fusion method [15] is proposed to jointly learn the voting weights of patches. In this method, the weighted voting is formulated to minimize the total expectation of labeling error and the pairwise dependency between atlases is modeled as the joint probability of two atlases making a segmentation error at a voxel. Specifically, the optimal voting weight vector $\mathbf{w}_y^*$ in this method are calculated as follows

$$\begin{aligned}\mathbf{w}_y^* = \arg\min_{\mathbf{w}_y} \mathbf{w}_y^T \mathcal{M}_y \mathbf{w}_y + \lambda ||\mathbf{w}_y||_2, \\ s.t. \sum\nolimits_{k=1}^{K} w(y, x_{k,y}) = 1,\end{aligned} \tag{3}$$

where $\mathbf{w}_y \in R^K$ is the voting weight vector (with the element $w(y, x_{k,y})$) from K atlases for the target voxel y, and $\mathcal{M}_y$ is a pairwise dependency matrix, which determined by the pairwise similarity between two patches in atlases (with each patch extracted from a particular atlas). Moreover, Wang *et al.* [46] also propose an extended joint label fusion (JLF) method, using a non-local search strategy to further alleviate the possible registration errors.

In addition, a sparse patch-based method (SPBM) [32] is proposed for multi-atlas segmentation, where the patch in the target image can be reconstructed by the linear superposition of patches in the atlas images. In SPBM, based on the region-specific dictionary $D(y)$, the target patch $P(y)$ can be linearly reconstructed by a weighting vector $\mathbf{w}_y^*$ using a sparse dictionary learning model with the l_1-norm constraint

$$\mathbf{w}_y^* = \arg\min_{\mathbf{w}_y} \frac{1}{2} ||P(y) - D(y)\mathbf{w}_y||_2^2 + \lambda ||\mathbf{w}_y||_1, \tag{4}$$

where the regularization parameter λ controls the sparsity of the $\mathbf{w}_y$, and hence, only a small number of patches in the atlas images (with high similarity to the target patch) will be selected for the subsequent label fusion process. More recently, several multi-layer dictionary learning methods [16], [17] are also proposed for multi-atlas segmentation. It is worth noting that most of the existing methods [12]–[17], [30]–[43] simply focus on using the image intensity as feature representation of patches. Due to the large inter-variation (among different subjects) and intra-variation (within the same subject) of image intensities, using only intensity feature cannot adequately describe the complex appearance patterns (*e.g.*, the high-order relationship between voxels within a patch) in brain MR images. However, previous multi-atlas based label fusion strategies seldom consider such higher-order relationship between voxels within a patch.

Based on the assumption that similar patches are prone to share the same label, the label of the voxel y in the target image can be inferred by using the weighted voting strategy. In this strategy, the final label of y is calculated as follows:

$$l_y = \arg\max_r \sum\nolimits_{p=1}^{P} w(y, x^p)\delta(l^p, r), r = 1, \cdots, R, \tag{5}$$

where x^p is the p-th $(p = 1, \cdots, P)$ candidate patch extracted from multiple atlases, and l^p denotes the voxel label of the voxel centered at the p-th patch. Besides, $\delta(l^p, r)$ is a Dirac function, which is equal to 1 when $l^p = c$; and 0, otherwise. Also, r $(r = 1, \cdots, R)$ is corresponding to the r-th ROI.

III. METHODOLOGY

In this work, a high-order feature learning framework is proposed for multi-atlas based label fusion, with the pipeline illustrated in Fig. 1. There are three major components in HLF. As shown in Fig. 1 (a), the mcRBM model [26], [27] is employed to learn high-order patch-level features (*i.e.*, mean feature and covariance feature) of brain MR images, where the higher-order relationship between voxels within a patch can be implicitly modeled. Based on the learned mean and covariance features, as well as the original intensity feature, a group-fused sparsity dictionary learning method is developed to jointly calculate the similarity (*i.e.*, voting weights) between the target and candidate patches (see Fig. 1 (b)). As shown in Fig. 1 (c), the label fusion is performed by using the learned voting weights for multi-atlas based ROI segmentation in brain MR images. More details can be found in the following.

A. High-order Feature Learning

As mentioned above, using only the intensity feature of image patches cannot adequately describe the complex appearance patterns (*e.g.*, the high-order relationship between voxels within a patch) in brain MR images. In this work, the mcRBM model [26], [27] is employed to learn the high-order representation of image patches in brain MR images, with the network architecture shown in Fig. 2. The mcRBM [26], [27] is a generative model having two sets of hidden units, including the mean hidden units (the green rectangle in Fig. 2) and covariance hidden units (the yellow rectangle in Fig. 2). Here, the mean hidden units represent the mean intensity information of the input patch, while the covariance hidden units denote the covariances (*i.e.*, pair-wise dependencies) information among voxels in the patch. Typically, the model only has the full connection between the input patch and hidden units, without any connection between the mean hidden units and covariance hidden units. In particular, the covariance part has two layers, including 1) the factor layer that connects twice to the input with the same filters, and 2) covariance layer that connects once to the factor layer. Also, the mean part has one layer where the mean layer connects once to the input. The probabilistic mcRBM model can be defined in two terms of energy functions, *i.e.*, 1) the mean intensity of the input energy function E^m, and 2) the covariance of the input energy function E^c. Specifically, the mean of the visible units energy E^m is defined as

$$E^m(v,h^m) = -\sum_{j=1}^{N_m}\sum_{i=1}^{N_v} M_{ij}h_j^m v_i - \sum_{j=1}^{N_m} b_j^m h_j^m, \tag{6}$$

where h_j^m represents the j-th mean hidden units ($j = 1,\cdots,N_m$), v_i is the i-th visible units, b_j^m is the j-th bias and M_{ij} is the connection weight between the mean hidden units h_j^m and the visible units v_i. As once direct connection, the model can describe the mean structure of patch. To capture the high-order dependency between each pair of voxels in the patch, we calculate the weighted sum of products between pairs of voxels in the patch. The covariance of the visible units energy function E^c is defined as:

$$E^c(v,h^c) = -\frac{1}{2}\sum_{s=1}^{N_f}\sum_{n=1}^{N_c} Q_{sn}h_n^c(\sum_{i=1}^{N_v} C_{is}v_i)^2 - \sum_{n=1}^{N_c} b_n^c h_n^c, \tag{7}$$

where h_n^c represents the n-th ($n = 1,\cdots,N_c$) covariance hidden unit, C is the connection weight matrix between visible units and factors, Q is the connection weight matrix between factors and covariance hidden units. Since, the factors are connected twice to the voxels in the patch, and once to the hidden unit. E^c can be seen as the energy of the Restricted Boltzmann Machine [47] that models the covariance structure of patch. Hence, the overall energy function is as follows:

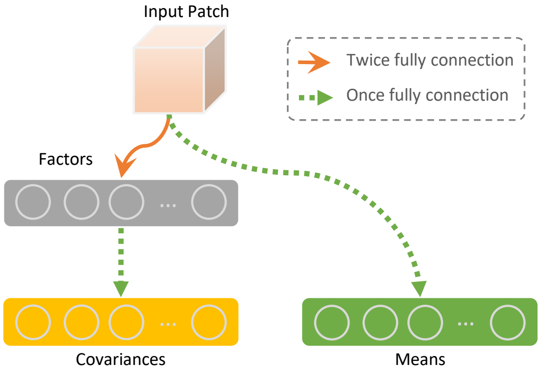


Fig. 2. Architecture of mcRBM [26] used in this study. The orange line and green line represent twice and once connection, respectively. The mcRBM model only has the full connection between the input patch and hidden units, without any connection between the means hidden units and covariance hidden units. The covariance part has two layers, including the factor layer that connects twice to the input with the same filters, and the covariance layer that connects once to the factor layer. Also, the means part has one layer where the mean hidden units connect once to the input patch.

$$E(v,h^c,h^m) = E^c(v,h^c) + E^m(v,h^m). \tag{8}$$

The probability density function can be defined as:

$$p(v,h^c,h^m) = \frac{1}{Z}exp^{E(v,h^c,h^m)}, \tag{9}$$

where $Z = \sum_{h^c,h^m} E(v,h^c,h^m)$ is a normalized factor. The marginal distribution over the visible units is $p(v) = \sum_{h^c,h^m} p(v,h^c,h^m)$. The parameters of the mcRBM model can be learned by a stochastic gradient ascent algorithm with the log likelihood in terms of joint energy $E(v,h^c,h^m)$. The optimization problem can be solved by using the Contrastive Divergence [48] and Hybrid Monte Carlo methods [49]. Once the mcRBM model is trained, it can learn two kinds of feature representations for each input patch, including the mean feature and covariance feature. The mean feature describes the mean intensity of the patch, and the covariance feature captures the higher-order relationship between voxels within a patch.

B. Voting Weight Learning for Multiple Features

For patches in both target and atlas images, two types of features (*i.e.*, mean feature and covariance feature) can be learned via the trained mcRBM model. Besides, as shown in Fig. 1 (a), the image intensity of patches is also used as feature representation in HLF. To effectively fuse these three types of features for multi-atlas based segmentation, a group-fused sparse dictionary learning method is developed to compute the similarity between the target patch and each candidate patch.

Denote F_1^y, F_2^y, F_3^y as the original intensity feature, mean feature and covariance feature of the patch centered at the voxel y, respectively. By concatenating these features, a multi-feature matrix $F^y = [F_1^y, F_2^y, F_3^y]$ can be constructed to describe each patch. Based on these three kinds of features, three feature-specific dictionaries are constructed, including 1) $D_1(y)$ that is the original intensity feature dictionary), 2)

$D_2(y)$ that is the mean feature dictionary, and 3) $D_3(y)$ that is the covariance feature dictionary. Then, F^y can be linearly reconstructed via a weight matrix $W = [w_1, w_2, w_3]$. The group-fused sparse dictionary learning model [50], [51] can be defined as follows:

$$W^* = \min_W \frac{1}{2} \sum\nolimits_{f=1}^{3} ||F_f^y - D_f(y) w_f||_2^2 + \lambda_1 ||W||_{2,1} \\ + \lambda_2 \sum\nolimits_{f=1}^{2} ||w_{f+1} - w_f||_1, \tag{10}$$

where the second term is a $l_{2,1}$-norm based sparsity constraint [52]. The regularization term $||W||_{2,1}$ is used to ensure that W is sparse in rows (with each row corresponds to a patch). Hence, only a small number of patches (with non-zero coefficient in W) are selected to reconstruct the target patch based on intensity, mean and covariance features. The third term is the fused smoothness regularization term [53], encouraging the reconstruction coefficients of different features for the same patch to be similar. Here, λ_1 controls the group sparsity of the linear model, and λ_2 controls the smoothness of the linear model. Thus, the reconstruction coefficients in the learned W can be regarded as the similarity metric between the candidate and target patches with different features, and use such similarity to obtain the voting weights of candidate patches for label fusion.

C. Weighted Label Fusion

For each voxel y in the target image, its label is inferred by using the weighted voting strategy to fuse all labeled voxels in atlases (within the neighborhood of the voxel y). Specifically, the probability of y belonging to the r-th ROI can be calculated as follows

$$p(l_y = r) = \frac{1}{Z} \sum\nolimits_{k=1}^{N} \sum\nolimits_{j \in N(y)} \sum\nolimits_{f=1}^{3} w_f(y, x_{k,j}) \delta(l_{k,j}, r), \tag{11}$$

where $Z = \sum_{k=1}^{N} \sum_{j \in N(y)} \sum_{f=1}^{3} w_f(y, x_{k,j})$ is a normalization factor, $N(y)$ denotes the neighbors of the voxel y, and $l_{k,j}$ is the label of voxel $x_{k,j}$ at the j-th voxel in k-th atlas. And $w_f(y, x_{k,j})$ is a weight between voxel y and $x_{k,j}$ based on the f-th feature. Note that $\delta(l_{k,j}, r)$ is equal to 1 if $x_{k,j}$ belongs to the r-th ROI; and 0, otherwise. Then, the final label is obtained by using the maximum a posteriori (MAP) criterion, *i.e.*, $\arg\max_r p(l_y = r)$. The details of the proposed algorithm are summarized in Algorithm 1.

D. Optimization

For the group-fused sparse dictionary learning model, an iterative projected gradient descent method [54] is used to solve Eq. 10. Specifically, the objective function in Eq. 10 can be divided into two parts, including 1) a smooth part and 2) a non-smooth part. The smooth part $\phi(W)$ is as follows

$$\phi(W) = \frac{1}{2} \sum\nolimits_{f=1}^{3} ||F_f^y - D_f(y) w_f||_2^2. \tag{12}$$

And the non-smooth part $\psi(W)$ consists a group Lasso regularizer and a fuse Lasso regularizer, listed below.

$$\psi(W) = \lambda_1 ||W||_{2,1} + \lambda_2 \sum\nolimits_{f=1}^{2} ||w_{f+1} - w_f||_1. \tag{13}$$

Algorithm 1: High-order feature learning framework for multi-atlas based label fusion

```
Input: Atlas $A = \{A_s | s = 1, \cdots, N\}$; label map $L = \{L_s | s = 1, \cdots, N\}$; and target image $T$
Output: Label map $L_T$ of the target image
Training:
    1: Random sampling patches from both atlas and target image;
    2: Training the mcRBM model to extract both mean and covariance features.
Testing:
    1: Extracting the high-order features by applying the trained mcRBM;
    2: Learning the voting weights using Eq. 10;
    3: Label fusion based on learned voting weights via Eq. 11 and MAP criteria.
```

At each iteration t, an intermediate variant V^t is computed firstly via the following

$$V^t = W^t - r^t g(\phi(W^t)), \tag{14}$$

where $g(\phi(W^t))$ is the gradient of $\phi(W)$ at the t-th iteration, and r^t is the step size (that can be determined by line search). Then, W is updated via

$$W^{t+1} = \arg\min_W \frac{1}{2} ||W - V^t||_2^2 + \psi(W), \tag{15}$$

which is the proximal operator associated with the non-smooth term $\psi(W)$. Eq. 15 can be solved by alternatively optimized the group Lasso and fused Lasso regularizers.

Following [54], an accelerated gradient descent method is further used to speed up the optimization of Eq. 15. Specifically, the search point $\mathcal{S}^t$ is computed via

$$\mathcal{S}^t = W^t + \alpha^t (W^t - W^{t-1}), \tag{16}$$

where α^t is a constant. By replacing W^t in Eq. 14 with $\mathcal{S}^t$, Eq. 14 can be rewritten as follows

$$V^t = \mathcal{S}^t - r^t g(\phi(\mathcal{S}^t)). \tag{17}$$

Finally, the approximate solution of W in Eq. 15 is calculated by using Eq. 16 and Eq. 17.

IV. Experiment

In this section, we first introduce materials, experimental settings and competing methods used in the experiments, and then present the segmentation results of different methods.

A. Materials

The proposed HLF method is evaluated on three public datasets, including 1) Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [55], 2) Non-rigid Image Registration Evaluation Project (NIREP) dataset [56] and 3) LONI-LPBA40 dataset [29]. More details can be found as follows.

1) **ADNI [55]:** Following [16], [33], 60 T1-weighted brain MR images are randomly selected from ADNI for *hippocampus* segmentation. A Leave-One-Out (LOO) cross-validation strategy is used to perform experiments. Specifically, each brain MRI is treated as a to-be-segmented target image, while the remaining 59 images

are regarded as atlas images. Several pre-processing steps are performed for MR images in this dataset, including skull removal [57], N4-based bias field correction [58], and intensity standardization [59].

2) **NIREP [56]:** This dataset consists of 16 T1-weighted brain MR images, acquired from 8 normal male adults and 8 female adults. These MR images were obtained in a General Electric Signa scanner (1.5 Tesla), using the following protocol: SPGR/50, TR 24, TE 7, NEX 1 matrix 256×192, FOV $24\,cm$. In this dataset, each MR image is manually segmented into 32 ROIs. The LOO cross-validation is performed on this dataset. That is, 15 of 16 subjects are used as atlases and aligned onto the remaining image that used as the target image.

3) **LONI-LPBA40 [29]:** The LONI-LPBA40 dataset is provided by the Laboratory of Neuro Imaging (LONI) at UCLA. This dataset contains 40 brain MR images, and each image has 54 manually labeled ROIs. Specifically, these labeled ROIs were created manually to annotate brain structures of images in the LONI-LPBA40 dataset. And all MRI volumes were rigidly aligned to the MNI305 template [60]. For this dataset, 20 subjects are randomly selected as atlases and the remaining 20 subjects are used as target images.

B. Experimental Settings

For all pre-processed brain MR images, affine registration is applied via FLIRT in the FSL [61] toolbox, using the normalized mutual information as similarity metric, 12 degrees of freedom, and the search range ± 20 in all directions. After the affine registration, a deformable registration is performed using the Diffeomorphic Demons method [20] with default parameters (*i.e.*, smoothing kernel size of 2.0, and iterations in low, middle and high resolutions as $20 \times 10 \times 5$). For training the mcRBM model, we first randomly sample $500,000$ patches with the patch size of $7 \times 7 \times 7$. And the mcRBM network is trained with 343 factor units, 343 covariance hidden units and 343 mean hidden units. For the proposed HLF method, the patch size is fixed as $7 \times 7 \times 7$ and the search region size is set as $7 \times 7 \times 7$. The parameters λ_1 and λ_2 in Eq. 10 are empirically set as 0.1 and 0.01, respectively.

Two evaluation metrics are used in the experiments. Specifically, the first metric is the Dice ratio (*DR*), defined as

$$DR = \frac{2|R_1 \cap R_2|}{|R_1| + |R_2|}, \tag{18}$$

where $\cap$ denotes the overlap between automatic segmented regions R_1 and ground truth of region R_2, and $|\cdot|$ denotes the number of voxels belonging to each ROI. The second metric is the average surface distance (*ASD*), defined as

$$\begin{aligned} ASD = &\frac{1}{2}(\frac{1}{n_1} \sum_{r_1 \in S(R_1)} d(r_1, S(R_2)) \\ &+ \frac{1}{n_2} \sum_{r_2 \in S(R_2)} d(r_2, S(R_1))), \end{aligned} \tag{19}$$

where $d(\cdot,\cdot)$ measures the Euclidean distance, and n_1 and n_2 are the numbers of vertices in surface $S(R_1)$ and surface $S(R_2)$, respectively.

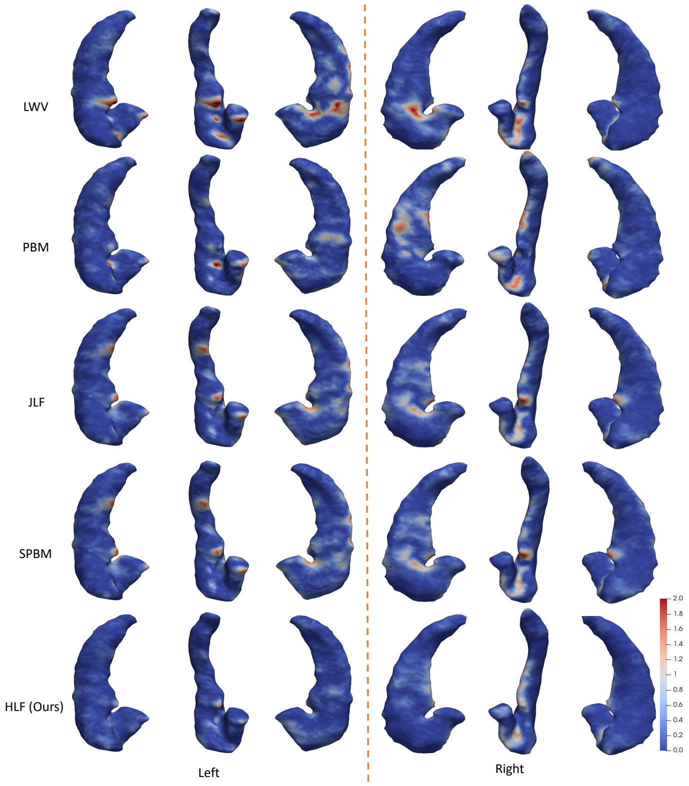


Fig. 3. Visual illustration of surface distance between the segmentation results of different methods and ground truth on the left and right *hippocampus*.

C. Competing Methods

The proposed HLF method is compared with four well-known multi-atlas based segmentation methods, including LWV [12], PBM [13], JLF [15], [46] and SPBM [32]. In our experiments, two competing methods (*i.e.*, JLF[2] and SPBM) are implemented by using the publicly available code, while LWV [12] and PBM [13] are implemented by ourselves using default parameters. For the fair comparison, the same experimental settings are used for HLF and four competing methods, including the patch size, search region size, atlas number, and registration methods. In this way, we can ensure that the comparison between HLF and four competing methods in the experiments is fair. Note that the proposed HLF method can employ both mean and covariance features (via mcRBM) and the original image intensity features, while four competing methods (*i.e.*, LWV, PBM, JLF, and SPBM) rely on only the original image intensity features. As multi-atlas based methods, three methods (*i.e.*, JLF, SPBM, and HLF) are learning-based methods that can optimize the voting weights for label fusion, while the remaining two methods (*i.e.*, LWV and PBM) are non-learning-based approaches.

D. Experimental Results on ADNI

In the first group of experiments, we perform *hippocampus* segmentation on the ADNI dataset. Table I shows the *DR* and *ASD* values achieved by different methods in segmenting both the left and right *hippocampus*. As can be seen

[2] https://www.nitrc.org/projects/picsl_malf

TABLE I
SEGMENTATION RESULTS OF LWV, PBM, JLF, SPBM AND THE PROPOSED HLF METHOD ON LEFT AND RIGHT *hippocampus* SEGMENTATION ON ADNI DATASET. THE SYMBOL "*" INDICATES SIGNIFICANT IMPROVEMENT (p<0.05) BY THE PROPOSED METHOD. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY.

Method	Left Hippocampus		Right Hippocampus	
	DR (%)	*ASD* (*mm*)	*DR* (%)	*ASD* (*mm*)
LWV*	85.15 ± 2.31	0.492 ± 0.142	85.75 ± 2.02	0.451 ± 0.058
PBM*	85.47 ± 5.11	0.521 ± 0.232	87.08 ± 3.57	0.440 ± 0.142
JLF*	86.51 ± 2.31	0.452 ± 0.173	87.35 ± 3.15	0.402 ± 0.083
SPBM*	86.32 ± 3.58	0.460 ± 0.189	87.39 ± 2.77	0.399 ± 0.115
HLF (Ours)	$\mathbf{88.30 \pm 3.34}$	$\mathbf{0.418 \pm 0.152}$	$\mathbf{88.83 \pm 2.77}$	$\mathbf{0.381 \pm 0.085}$

TABLE II
SEGMENTATION RESULTS OF LWV, PBM, JLF, SPBM AND OUR PROPOSED METHOD ON THE NIREP DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY.

Method	*DR* (%)	*ASD* (*mm*)
LWV	75.02 ± 1.39	1.116 ± 0.048
PBM	76.74 ± 1.43	1.069 ± 0.056
JLF	78.25 ± 1.70	1.043 ± 0.064
SPBM	78.48 ± 1.90	1.132 ± 0.075
HLF (Ours)	$\mathbf{79.54 \pm 1.78}$	$\mathbf{1.035 \pm 0.065}$

TABLE III
SEGMENTATION RESULTS OF LWV, PBM, JLF, SPBM AND THE PROPOSED HLF METHOD ON THE LONI-LPBA40 DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY.

Method	*DR* (%)	*ASD* (*mm*)
LWV	78.22 ± 0.88	1.234 ± 0.049
PBM	78.81 ± 0.91	1.170 ± 0.056
JLF	79.26 ± 1.07	1.181 ± 0.061
SPBM	79.65 ± 1.00	1.196 ± 0.048
HLF (Ours)	$\mathbf{81.02 \pm 1.06}$	$\mathbf{1.136 \pm 0.059}$

from Table I, the proposed HLF method achieves the best performance in segmenting the left and right *hippocampus* regarding both the *DR* and *ASD* metrics. For instance, HLF yields the improvement of 3.15%, 2.83%, 1.79% and 1.98% over LWV, PBM, JLF, and SPBM, respectively, in terms of *DR* on the left *hippocampus*. For the right *hippocampus*, HLF increases the *DR* value by 3.08%, 1.75%, 1.48%, and 1.44%, respectively, compared to LWV, PBM, JLF, and SPBM. The possible reason for the improvement is that our HLF method uses the high-order features that reflect the high-order relationship between voxels within a patch to learn the similarity. Since learning based methods (*i.e.*, JLF, SPBM, and HLF) can automatically learn the coefficients, the learning methods can better measure the pairwise similarity between the target patch and candidate patch. Hence, compared with the non-learning based methods (*i.e.*, LWV and PBM), the learning based methods usually achieves better segmentation performances. Moreover, compared with SPBM, the proposed HLF can jointly learn the similarity based on the high-order information and the original intensity feature of image patches.

The paired t-test algorithm is performed between HLF and each of four competing methods based on the *DR* values, with results shown in Table I. Note that the symbol "$*$" in Table I indicates that the proposed HLF method achieves a significant improvement over a specific competing method. Table I suggests that HLF shows the significant improvement ($p < 0.05$) over LWV, PBM, JLF, and SPBM on both left and right *hippocampus* segmentation. In Fig. 3, we visually show the *ASD* values between the automatic segmentation images and ground truth on both left and right *hippocampus*, achieved by LWV, PBM, JLF, SPBM and our HLF method, respectively. As can be seen from Fig. 3, HLF generates segmentation with better visual quality, compared with four competing methods. This further validates the effectiveness of the proposed method in *hippocampus* segmentation.

E. Experimental Results on NIREP

In the second group of experiments, we perform segmentation of 32 ROIs on the NIREP dataset. Table II shows the average *DR* and *ASD* values achieved by five different methods. The *DR* values achieved by LWV, PBM, JLF, SPBM and our proposed HLF method on each of 32 ROIs are further reported in Fig. 4. Here, the symbols "\$" and "*" denote that HLF is significantly better than JLF and SPBM based on paired t-test ($p < 0.05$), respectively.

Table II shows that the average *DR* and *ASD* values for segmenting 32 ROIs are 79.54% and $1.035\,mm$ achieved by our HLF method, which are superior to the second best results (*i.e.*, *DR*=78.48% achieved by SPBM and *ASD*=$1.043\,mm$ yielded by JLF). It can be seen from Fig. 4 that, in terms of *DR*, HLF generally significantly outperforms the competing methods in segmenting 32 ROIs on the NIREP dataset. For instance, HLF yields significant improvement on each of all 32 ROIs, compared with LWV.

F. Experimental Results on LONI-LPBA40

In the third group of experiments, we perform segmentation of 54 ROIs on the LONI-LPBA40 dataset, with results shown in Table III and Fig. 5. From Table III, one can see that HLF increases the average *DR* value by 2.80%, 2.21%, 1.76% and 1.37%, respectively, compared to LWV, PBM, JLF, and SPBM. Also, the average *ASD* value generated by HLF on 54 ROIs is $1.136\,mm$, which is better than that of each competing method (*i.e.*, LWV, PBM, JLF, and SPBM). Fig. 5 suggests that HLF consistently outperforms those four methods in term of *DR*

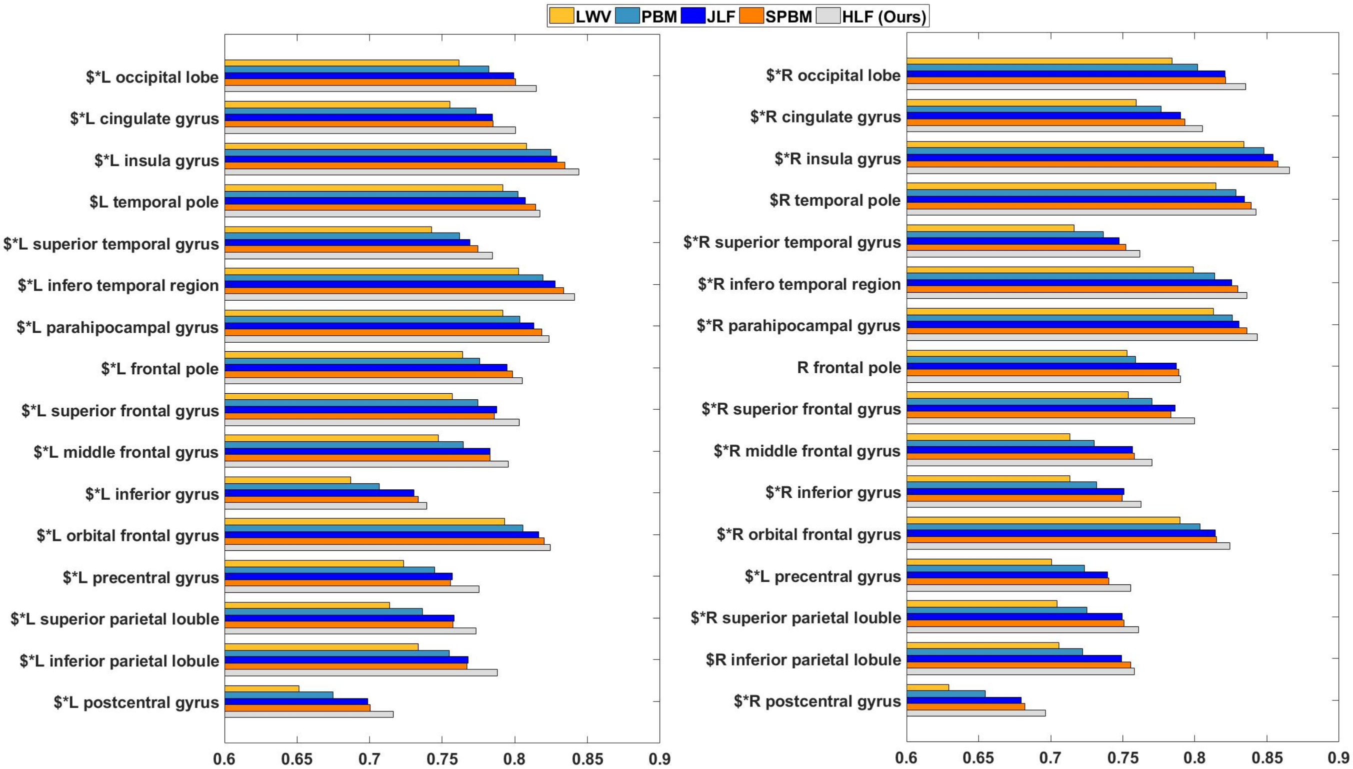


Fig. 4. Segmentation results of 32 ROIs on the NIREP dataset achieved by LWV (yellow), PBM (cyan), JLF (blue), SPBM (orange) and our proposed HLF method (gray) in terms of *DR* values. Here, "$" and "*" denote that our method is significantly better than JLF and SPBM based on paired t-test ($p < 0.05$), respectively. Meanwhile, our method is significantly better than the LWV and PBM in segmenting all ROIs.

TABLE IV
THE *DR* (%) VALUES YIELDED BY HLF AND TWO DEEP LEARNING METHODS (*i.e.*, U-NET AND IVOXELDCNA) ON THE ADNI AND LONI-LPBA40 DATASETS, RESPECTIVELY.

Method	ADNI	LONI-LPBA40
U-Net	82.20	77.42
iVoxelDCNa	83.34	79.13
HLF (Ours)	**88.59**	**81.02**

TABLE V
RESULTS OF THE PROPOSED HLF METHOD FOR *hippocampus* SEGMENTATION ON THE ADNI DATASET. THE TERMS "HLF-N", "HLF-L" AND "HLF" DENOTE THAT THE HLF MODELS ARE TRAINED ON THE NIREP, LONI-LPBA40 AND ADNI DATASETS, RESPECTIVELY.

Method	*DR* (%)	*ASD* (mm)
HLF-N	88.42 ± 4.48	0.395 ± 0.112
HLF-L	88.34 ± 4.47	0.397 ± 0.112
HLF	88.59 ± 2.73	0.399 ± 0.098

on 54 ROIs, and the improvement of HLF over the competing methods is significant in most ROIs.

V. DISCUSSION

A. Comparison with Deep Learning Methods

Besides the comparison with two learning-based methods (*i.e.*, JLF and SPBM) that use image intensity features, the proposed HLF method is further compared with deep learning methods, *i.e.*, U-Net [62] and iVoxelDCNa [63]. The experimental results on the ADNI and LONI-LPBA40 datasets are reported in Table IV, where the results of U-Net and iVoxelDCNa are obtained from [63]. One can observe from Table IV that the proposed HLF method shows better performance in comparison to two deep learning methods. The underlying reason could be that HLF can utilize the prior anatomical structure of the brain provided by multiple atlases to partly alleviate the challenge caused the extremely low image intensity around the ROI boundary, thus resulting in better performance in brain ROI segmentation.

B. Generalization Analysis

To validate the generalization capability of the proposed HLF method, we perform experiments for *hippocampus* segmentation by measuring the inter-dataset performance. Specifically, we first treat the NIREP and LONI-LPBA40 datasets as the training sets for models training, respectively. Then, we use the ADNI dataset as the *independent testing set* for performance evaluation. The experimental results are reported in Table V, where the term "HLF-N" and "HLF-L" denote that HLF models are trained on the NIREP and LONI-LPBA40 datasets, respectively. Also, the term "HLF" denotes that HLF model is trained on the ADNI dataset. From Table V, one can observe that our HLF method yields comparable results in terms of intra-dataset performance and inter-dataset performance in the task of *hippocampus* segmentation, suggesting the good generalization capability of HLF.

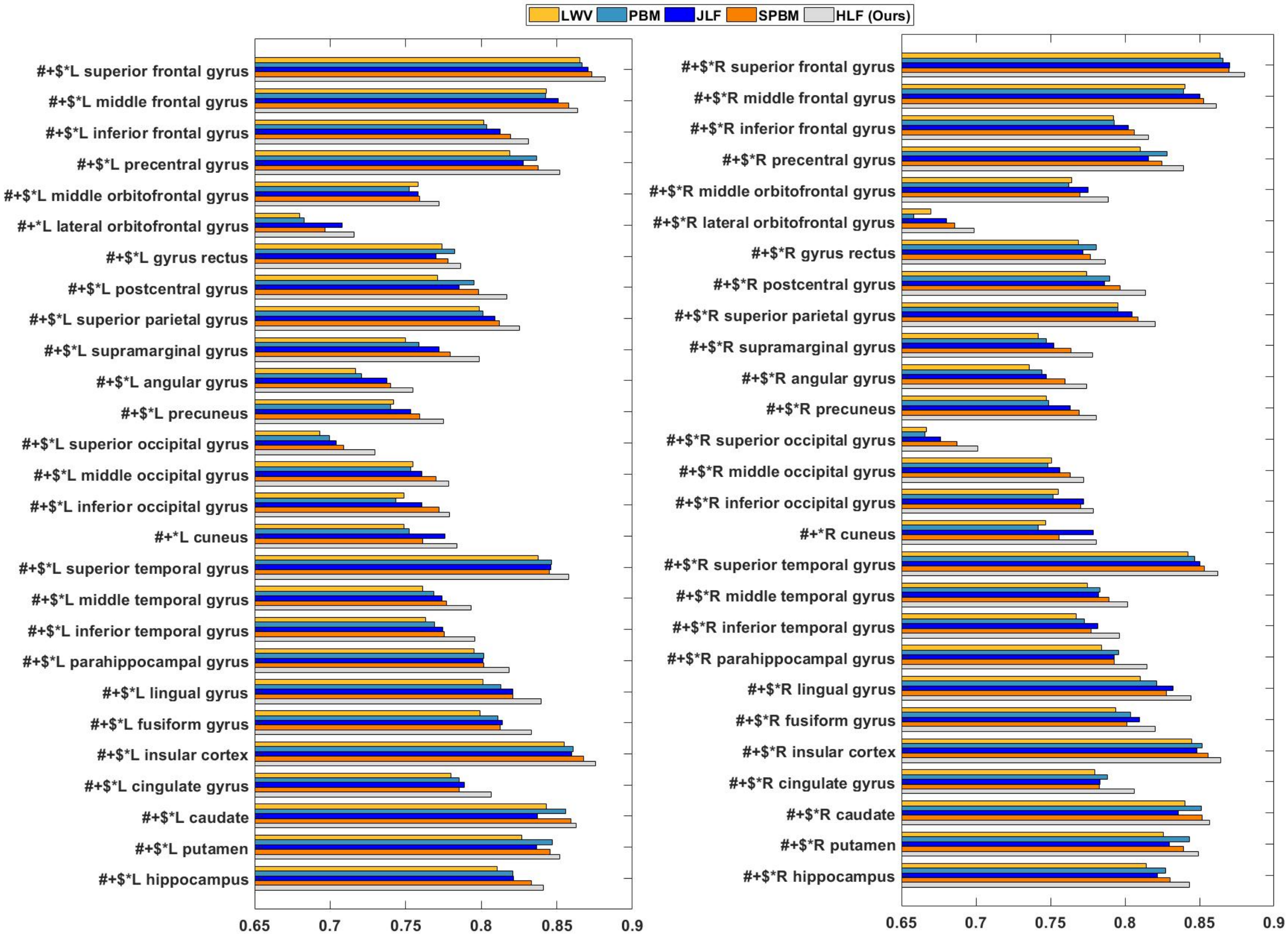


Fig. 5. Segmentation result of 54 ROIs on the LONI-LPBA40 dataset achieved by LWV (yellow), PBM (cyan), JLF (blue), SPBM (orange) and our proposed HLF method (gray) in terms of *DR* values. "#" , "+" , "$" and "*" denote that our method is significantly better than LWV, PBM, JLF and SPBM based on paired t-test ($p < 0.05$), respectively.

C. Influence of Voting Weight Learning

In our previous work [28], we simply concatenate the learned high-order features and intensity feature together to calculate the voting weights. In the paper, the group and fused Lasso regularization terms are employed to jointly learn the voting weights based on three kinds of features (*i.e.*, intensity feature, mean feature, and covariance feature). To evaluate the effectiveness of such joint learning strategy, we compare HLF with SPBM [32] that uses a combination strategy. That is, SPBM is used to learn the voting weights based the combined features of image intensity features and the mean and covariance features (learned by HLF), and this method is denoted as *hv*-SPBM in this work. The comparison between HLF and *hv*-SPBM is performed on the ADNI dataset for *hippocampus* segmentation, with results reported in Table VI. It can be seen from Table VI that, the *DR* and *ASD* values of HLF are 88.59% and $0.399\,mm$, respectively, which are better than those of *hv*-SPBM (*DR*=87.62% and *ASD*=$0.423\,mm$). These results suggest that the proposed joint weight learning strategy (used in HLF) is superior to the combination strategy (used in *hv*-SPBM) that simply combines the intensity feature and the learned features (*i.e.*, mean and covariance features).

TABLE VI
SEGMENTATION RESULTS OF *hv*-SPBM AND OUR PROPOSED METHOD ON *hippocampus* SEGMENTATION ON THE ADNI DATASET. THE TERMS a AND b IN "$a \pm b$" DENOTE THE MEAN AND STANDARD DEVIATION FOR DIFFERENT SUBJECTS, RESPECTIVELY.

Method	*DR* (%)	*ASD* (mm)
hv-SPBM	87.62 ± 3.44	0.423 ± 0.110
HLF (Ours)	$\mathbf{88.59 \pm 2.73}$	$\mathbf{0.399 \pm 0.098}$

D. Influence of Number of Atlases

The number of atlases is an important parameter for multi-atlas based segmentation. To validate the impact of the number of atlases on the performance of the proposed HLF, we randomly selected 20 subjects in the ADNI dataset as the target images and all the remaining subjects are treated as the candidate atlas images. As the number of atlases varies within the range of $[10, 20, 30, 40]$, Fig. 6 records the *DR* and *ASD* values achieved by HLF using different numbers of atlases on ADNI for *hippocampus* segmentation. One can observe from Fig. 6 that the best performance is achieved by HLF when the number of atlases is within $[30, 40]$. These results indicate that incorporating more atlases in HLF yields better segmentation

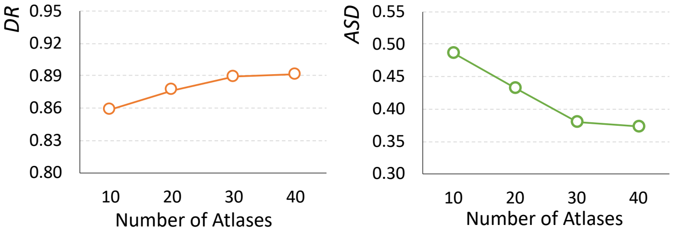


Fig. 6. *DR* and *ASD* (mm) values achieved by the proposed HLF method using different numbers of atlases for *hippocampus* segmentation on the ADNI dataset.

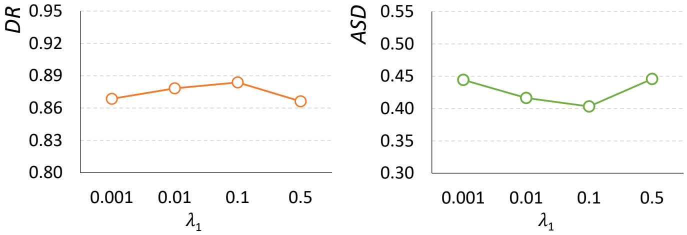


Fig. 7. *DR* and *ASD* (mm) values achieved by the proposed HLF method, using different values of λ_1 and and $\lambda_2 = 0.01$ for *hippocampus* segmentation on the ADNI dataset.

performance by using more anatomical prior information of brain MR images, but using more atlases will increase the computational time.

E. Parameter Analysis

In Eq. 10, λ_1 controls the group sparsity and λ_2 controls the smoothness of the linear model. To study their influences, we perform *hippocampus* segmentation on the ADNI dataset, by varying the values of λ_1 and λ_2 within the range of $[0.001, 0.01, 0.1, 0.5]$. The segmentation results achieved by the proposed HLF method with different values of λ_1 are reported in Fig. 7, while results using different values of λ_2 are shown in Fig. 8. As can be seen from Fig. 7 and Fig. 8 that HLF yields relatively stable results using different parameter values, and the best results are achieved when $\lambda_1 = [0.01, 0.5]$ and $\lambda_2 = [0.001, 0.1]$ on the ADNI dataset. For instance, as shown in Fig. 7, the best *DR* and *ASD* values are 88.38% and $0.403\,mm$ using $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$, respectively. These results imply that the proposed HLF method is not very sensitive to these two parameters (*i.e.*, λ_1 and λ_2).

F. Limitations and Future Work

There are still several limitations in the current work. *First*, in our experiments, the mean and covariance features have the same size as image patches, while the size of mean and covariance features can be arbitrary. Since the number of candidate patches is fixed, the proposed group fusion sparse dictionary algorithm can be applied to problems with any feature dimensions. *Second*, the high dimension of features and the two additional regularization terms (*i.e.*, group Lasso and fused Lasso) will increase the computational burden for learning the voting weights. It is interesting to implement the proposed framework in a parallel manner, which will be our future work. *Besides*, the current method only considers three types of features (*i.e.*, mean feature, covariance feature, and image intensity feature). Actually, the proposed method is a general framework, where any type of pre-defined or automatically learned features can be utilized (*e.g.*, LBP [64], SIFT [65] and HOG [66], *etc.*). As another future work, we plan to apply more features for further performance improvement. *Finally*, the segmentation of each ROI is performed independently, without considering their spatial relationship. It is interesting to design a multi-task learning framework to jointly segment multiple ROIs in brain MR images.

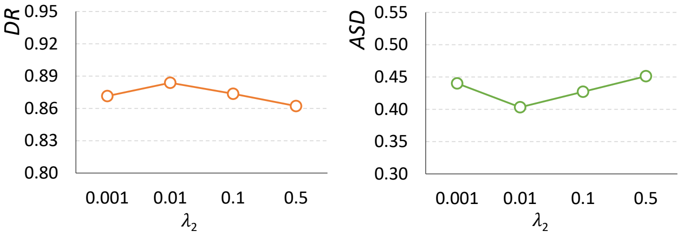


Fig. 8. *DR* and *ASD* (mm) values achieved by the proposed HLF method using $\lambda_1 = 0.1$ and different values of λ_2 for *hippocampus* segmentation on the ADNI dataset.

VI. Conclusion

In this paper, we propose a high-order feature learning framework for multi-atlas based label fusion. Specifically, the mcRBM [26], [27] is used to learn high-order patch-level features of brain MR images, which can effectively describe the higher-order relationship between voxels within a patch (via mean and covariance features). Based on the learned mean and covariance features as well as the original image intensity feature, a group-fused sparsity dictionary learning method is then developed to jointly calculate the similarity (*i.e.*, voting weights) between the target and candidate patches. Finally, weighted label fusion is performed using the learned voting weights for multi-atlas based ROI segmentation with brain MR images. Experiments on ADNI, NIREP and LONI-LPBA40 datasets suggest that the proposed HLF method can achieve superior results on ROI segmentation of brain MR images, compared with several state-of-the-art methods.

References

[1] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[2] D. P. Devanand, G. Pradhaban, X. Liu, A. Khandji, S. De Santi, S. Segal, H. Rusinek, G. H. Pelton, L. S. Honig, R. Mayeux *et al.*, "Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease." *Neurology*, vol. 68, no. 11, pp. 828–836, 2007.

[3] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Transactions on Medical Imaging*, vol. 35, no. 6, pp. 1463–1474, 2016.

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

[5] C. Lian, J. Zhang, M. Liu, X. Zong, S.-C. Hung, W. Lin, and D. Shen, "Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in 7T MR images," *Medical Image Analysis*, vol. 46, pp. 106–117, 2018.

[6] S. Heckers, S. L. Rauch, D. C. Goff, C. R. Savage, D. L. Schacter, A. J. Fischman, and N. M. Alpert, "Impaired recruitment of the hippocampus during conscious recollection in schizophrenia," *Nature Neuroscience*, vol. 1, no. 4, pp. 318–323, 1998.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2952079, IEEE Transactions on Image Processing

[7] Y. Zhou, N. Shu, Y. Liu, M. H. Song, Y. Hao, H. Liu, C. Yu, Z. Liu, and T. Jiang, "Altered resting-state functional connectivity and anatomical connectivity of hippocampus in schizophrenia," *Schizophrenia Research*, vol. 100, no. 1, pp. 120–132, 2008.

[8] A. A. Grace, "Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression," *Nature Reviews Neuroscience*, vol. 17, no. 8, pp. 524–532, 2016.

[9] M. Liu, J. Zhang, D. Nie, P.-T. Yap, and D. Shen, "Anatomical landmark based deep feature representation for MR images in brain disease diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1476–1485, 2018.

[10] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2019.

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

[12] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[13] P. Coupé, J. V. Manjón, V. Fonov, J. C. 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, 2011.

[14] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, and D. Rueckert, "Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling," *NeuroImage*, vol. 76, no. 1, pp. 11–23, 2013.

[15] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 611–623, 2013.

[16] Y. Song, G. Wu, K. Bahrami, Q. Sun, and D. Shen, "Progressive multi-atlas label fusion by dictionary evolution," *Medical Image Analysis*, vol. 36, pp. 162–171, 2017.

[17] C. Zu, Z. Wang, D. Zhang, P. Liang, Y. Shi, D. Shen, and G. Wu, "Robust multi-atlas label propagation by deep sparse representation," *Pattern Recognition*, vol. 63, pp. 511–517, 2017.

[18] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[19] M. Jenkinson and S. M. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

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

[21] M. Schlachter, T. Fechter, M. Jurisic, T. Schimek-Jasch, O. Oehlke, S. Adebahr, W. Birkfellner, U. Nestle, and K. Bühler, "Visualization of deformable image registration quality using local image dissimilarity," *IEEE Transactions on Medical Imaging*, vol. 35, no. 10, pp. 2319–2328, 2016.

[22] S. Yousefi, N. Kehtarnavaz, and A. Gholipour, "Improved labeling of subcortical brain structures in atlas-based segmentation of magnetic resonance images," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 7, pp. 1808–1817, 2012.

[23] Z. Li, D. Mahapatra, J. A. W. Tielbeek, J. Stoker, L. J. Van Vliet, and F. M. Vos, "Image registration based on autocorrelation of local structure," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 63–75, 2016.

[24] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[25] M. Kim, G. Wu, Q. Wang, S. W. Lee, and D. Shen, "Improved image registration by sparse patch-based deformation estimation," *NeuroImage*, vol. 105, pp. 257–268, 2015.

[26] M. Ranzato and G. E. Hinton, "Modeling pixel means and covariances using factorized third-order boltzmann machines," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2551–2558.

[27] J. Kivinen, C. Williams, and N. Heess, "Visual boundary prediction: A deep neural prediction network and quality dissection," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, vol. 33, 2014, pp. 512–521.

[28] L. Sun, W. Shao, and D. Zhang, "High-order Boltzmann machine-based unsupervised feature learning for multi-atlas segmentation," in *International Symposium on Biomedical Imaging*. IEEE, 2017, pp. 507–510.

[29] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[30] 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, 2006.

[31] F. Rousseau, P. Habas, and C. Studholme, "A supervised patch-based approach for human brain labeling," *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1852–1862, 2011.

[32] D. Zhang, Q. Guo, G. Wu, and D. Shen, "Sparse patch-based label fusion for multi-atlas segmentation," in *International Workshop on Multimodal Brain Image Analysis*. Springer, 2012, pp. 94–102.

[33] G. Wu, M. Kim, G. Sanroma, Q. Wang, B. C. Munsell, and D. Shen, "Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition," *NeuroImage*, vol. 106, pp. 34–46, 2015.

[34] 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," *Medical Image Analysis*, vol. 18, no. 6, pp. 881–890, 2014.

[35] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.

[36] W. Bai, W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert, "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1302–1315, 2013.

[37] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Transactions on Medical Imaging*, vol. 29, no. 12, pp. 2000–2008, 2010.

[38] J. M. P. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.

[39] S. H. Park, Y. Gao, and D. Shen, "Multi-atlas based segmentation editing with interaction-guided constraints," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 198–206.

[40] D. Zhang, G. Wu, H. Jia, and D. Shen, "Confidence-guided sequential label fusion for multi-atlas based segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 14. Springer, 2011, pp. 643–650.

[41] G. Sanroma, O. M. Benkarim, G. Piella, O. Camara, G. Wu, D. Shen, J. D. Gispert, J. L. Molinuevo, and M. A. G. Ballester, "Learning non-linear patch embeddings with neural networks for label fusion," *Medical Image Analysis*, vol. 44, pp. 143–155, 2018.

[42] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, no. 6, pp. 671–684, 2013.

[43] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[44] L. Sun, C. Zu, W. Shao, J. Guang, D. Zhang, and M. Liu, "Reliability-based robust multi-atlas label fusion for brain MRI segmentation," *Artificial Intelligence in Medicine*, vol. 96, pp. 12 – 24, 2019.

[45] Y. Zhu, L. Wang, M. Liu, C. Qian, A. Yousuf, A. Oto, and D. Shen, "MRI-based prostate cancer detection with high-level representation and hierarchical classification," *Medical Physics*, vol. 44, no. 3, pp. 1028–1039, 2017.

[46] H. Wang and P. Yushkevich, "Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation," *Frontiers in Neuroinformatics*, vol. 7, pp. 27–27, 2013.

[47] Y. Freund and D. Haussler, "Unsupervised learning of distributions of binary vectors using two layer networks," *Advances in Neural Information Processing Systems*, vol. 4, pp. 912–919, 1992.

[48] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[49] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 1996, vol. 118.

[50] D. Zhang, J. Liu, and D. Shen, "Temporally-constrained group sparse learning for longitudinal data analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 264–271.

[51] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 1095–1103.

[52] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[53] J. Liu, L. Yuan, and J. Ye, "An efficient algorithm for a class of fused lasso problems," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, pp. 323–332.

[54] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[55] 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.

[56] G. E. Christensen, X. Geng, J. G. Kuhl, J. Bruss, T. J. Grabowski, I. A. Pirwani, M. W. Vannier, J. S. Allen, and H. Damasio, "Introduction to the non-rigid image registration evaluation project (NIREP)," in *Proceeding WBIR'06 Proceedings of the Third International Conference on Biomedical Image Registration*. Springer, 2006, pp. 128–135.

[57] 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, 2012.

[58] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

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

[60] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *Nuclear Science Symposium and Medical Imaging Conference*. IEEE, 1993, pp. 1813–1817.

[61] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. 208–219, 2004.

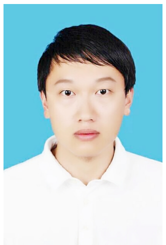
[62] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[63] Y. Chen, H. Gao, L. Cai, M. Shi, D. Shen, and S. Ji, "Voxel deconvolutional networks for 3D brain image labeling," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1226–1234.

[64] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[65] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1150–1157.

[66] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 886–893.

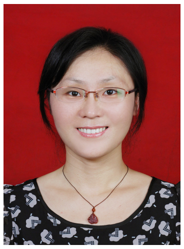


Liang Sun received the B.S degree from Shandong University of Science and Technology, China, in 2014. He is currently a Ph.D. candidate in Computer Science from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include machine learning and medical image segmentation.

Wei Shao received the B.S. and M.S. degrees from Nanjng University of Technology, Jiangsu, China, in 2009 and 2012, respectively, and Ph.D. degree in Software Engineering from Nanjing University of Aeronautics and Astronautics (NUAA), China, in 2018. His current research interests include machine learning and bioinformatics.

Mingliang Wang received the B.S. degree in Nanjing University of Information Science and Technology in 2012, and the M.S. degree from Shanghai Institute of Computing Technology in 2015. He is currently working toward the Ph.D. degree in Software Engineering from Nanjing University of Aeronautics and Astronautics (NUAA), China. His current research interests include machine learning and medical image analysis.

Daoqiang Zhang received the B.S. degree and Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999 and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a Lecturer in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published over 100 scientific articles in refereed international journals such as IEEE Trans. on Pattern Analysis and Machine Intelligence, IEEE Trans. on Medical Imaging, IEEE Trans. on Image Processing, NeuroImage, Medical Image Analysis, Pattern Recognition, Artificial Intelligence in Medicine, IEEE Trans. on Neural Networks; and conference proceedings such as NIPS, IJCAI, AAAI, SDM, ICDM. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation (CCF).



Mingxia Liu received the B.S. and M.S. degrees from Shandong Normal University, Shandong, China, in 2003 and 2006, respectively, and the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2015. She is a Senior Member of IEEE. Her current research interests include machine learning, pattern recognition, and medical image analysis.