Artificial Intelligence In Medicine 96 (2019) 12–24

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Reliability-based robust multi-atlas label fusion for brain MRI segmentation☆

Liang Sun[a], Chen Zu[a], Wei Shao[a], Junye Guang[a], Daoqiang Zhang[a,*], Mingxia Liu[b]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, 211106, China

[b] Department of Information Science and Technology, Taishan University, Taian 271000, China

ARTICLE INFO

Keywords:
Label fusion
Label reliability
Spatial reliability
Multi-atlas segmentation
Brain structural MRI

ABSTRACT

Label fusion is one of the key steps in multi-atlas based segmentation of structural magnetic resonance (MR) images. Although a number of label fusion methods have been developed in literature, most of those existing methods fail to address two important problems, *i.e.*, (1) compared with boundary voxels, inner voxels usually have higher probability (or reliability) to be correctly segmented, and (2) voxels with high segmentation reliability (after initial segmentation) can help refine the segmentation of voxels with low segmentation reliability in the target image. To this end, we propose a general reliability-based robust label fusion framework for multi-atlas based MR image segmentation. Specifically, in the *first* step, we perform initial segmentation for MR images using a conventional multi-atlas label fusion method. In the *second* step, for each voxel in the target image, we define two kinds of reliability, including the label reliability and spatial reliability that are estimated based on the soft label and spatial information from the initial segmentation, respectively. *Finally*, we employ voxels with high label-spatial reliability to help refine the label fusion process of those with low reliability in the target image. We incorporate our proposed framework into four well-known label fusion methods, including locally-weighted voting (LWV), non-local mean patch-based method (PBM), joint label fusion (JLF) and sparse patch-based method (SPBM), and obtain four novel label-spatial reliability-based label fusion approaches (called ls-LWV, ls-PBM, ls-JLF, and ls-SPBM). We validate the proposed methods in segmenting ROIs of brain MR images from the NIREP, LONI-LPBA40 and ADNI datasets. The experimental results demonstrate that our label-spatial reliability-based label fusion methods outperform the state-of-the-art methods in multi-atlas image segmentation.

1. Introduction

MR imaging widely used in real-world clinical applications. Accurate segmentation of brain MR images provides quantitative analysis of brain structures, thus facilitating MRI-based pathology detection and brain parcellation. For instance, many clinical applications need the segmentation of MR images to describe how brain structures change during the disease progression. As an example, the *hippocampus* is known to be related with the Alzheimer's disease [1,2], and thus, it is critical to accurately segment hippocampus from the whole brain for computer-aided brain disease diagnosis. On the other hand, we usually segment a brain MR image into multiple regions-of-interest (ROIs) before constructing brain networks for subsequent analysis in brain network analysis [3–7]. However, it is time-consuming and usually error-prone for experts to manually segment those large amounts of MR images [8–10]. Hence, there is a largely unmet need to develop advanced automatic methods for brain ROI segmentation.

Recently, multi-atlas based segmentation methods have shown great successes in segmenting medical images [11–26]. The assumption of multi-atlas segmentation is that a voxel in the target image should have the same label as its corresponding voxel in the atlas image, if their local tissue shapes or appearances are similar. Typically, there are two main steps for multi-atlas segmentation, *i.e.*, (1) image registration [27–32], and (2) label fusion. Specifically, in the image registration step, each atlas image is warped onto the target image. Then, in the label fusion step, labels from different atlases will be propagated to the target image to obtain the final labels. In this work, we focus on the label fusion step under the multi-atlas segmentation framework.

☆ This work was supported in part by the National Natural Science Foundation of China (Nos: 61876082, 61861130366, 61703301, and 61473149), Taishan Scholar Program of Shandong Province in China, Scientific Research Foundation of Taishan University (No. Y-01-2018019) and China Council Scholarship.

* Corresponding author.
E-mail address: dqzhang@nuaa.edu.cn (D. Zhang).

https://doi.org/10.1016/j.artmed.2019.03.004
Received 4 December 2017; Received in revised form 4 March 2019; Accepted 5 March 2019

In literature, numerous label fusion strategies have been proposed for multi-atlas based brain MRI segmentation. Among them, majority voting (MV) is the simplest one, where each atlas image is treated equally when assigning labels to the target image [11]. As a more advanced strategy, locally-weighted voting (LWV) considers patch-wise similarity between the target image and each atlas as the voting weight for label assignment, and it has shown that LWV outperforms MV when segmenting brain MRI [12]. To alleviate the registration errors, the non-local mean patch-based method (PBM) has been proposed to propagate labels from *not only* the same location in the atlases, *but also* the neighboring patches in the atlases. PBM seeks multiple candidates by the pair-wise similarity between the target image patch and atlas image patches within a search region, showing improved accuracy and robustness of the labeling results [13,14]. More recently, the sparse representation based PBM method (SPBM) is proposed for label fusion, where only a small number of image patches (with high similarity to the target image patch) will be selected for the subsequent label fusion by using the l_1-norm based sparsity constraint [15,16]. In addition, several joint label fusion methods for brain MRI segmentation are proposed to measure the joint labeling risk between two patches in atlases, thus reducing the risks of labeling error [18,19]. Several multi-layer dictionary learning methods [24,25] have been proposed for multi-atlas segmentation. Song et al. progressively construct dynamic multi-layer dictionary to reduce the gap between the image domain and the label domain. Zu et al. [25] use a tree-like multi-layer dictionary to represent the hierarchical patch for ROI segmentation. Instead of capturing the complex brain with intensity features, deep learning method is used for learning the representation of the original image patches, and use these learnt features for label fusion [26]. However, most of the existing methods treat each voxel in the target image equally and independently in segmentation, without considering the specific location and reliability of each voxel in the target image.

Previous studies [13,17,33] have shown that most of misclassified voxels locate at the boundary of ROIs, while voxels far from the boundary of ROIs are easier to be segmented correctly. In light of this, we define the reliability for each voxel to measure whether a specific voxel is easy to be correctly segmented. If a voxel has high reliability, we assume that it is easy to correctly segment this voxel; and vice versa. Consequently, those voxels with high reliability in the target image can be used to help refine the label fusion of voxels with low reliability.

With this assumption, in this paper, we present a general reliability-based robust label fusion framework for multi-atlas based MR image segmentation, including three main steps: (1) initial segmentation, (2) estimating voxel reliability, and (3) reliability-based robust label fusion. Specifically, we first perform the initial segmentation using conventional multi-atlas label fusion method (*e.g.*, LWV and PBM, *etc.*), and thus can obtain a normalized voting result (*i.e.*, a soft label with a value between 0 and 1) for each voxel after initial segmentation. In the second step, for each voxel in the target image, we define two kinds of reliability: (1) label reliability, and (2) spatial reliability. The label reliability is estimated based on the soft label, where we assume that a voxel has higher label reliability to be correctly segmented if its soft label has lower entropy. Meanwhile, the spatial reliability is estimated based on the spatial structure of the target image label map from the initial segmentation, where we assume a voxel has high reliability to be correctly segmented if the label map around this voxel is continuous. In the third step, we use voxels with the high label and spatial reliability to help refine the label fusion process of those voxels with low reliability in the target image. Our method is a general framework and can easily be combined with existing state-of-the-art methods. To validate the effectiveness of the proposed framework, we apply the proposed reliability-based strategy to four well-known label fusion approaches, *i.e.*, LWV, PBM, JLF and SPBM, and obtain four novel reliability-based robust label fusion approaches, called label-spatial reliability-based LWV (ls-LWV), label-spatial reliability-based PBM (ls-PBM), label-spatial reliability-based JLF (ls-JLF) and label-spatial reliability-based SPBM (ls-SPBM), respectively. Experimental results on the NIREP, LONI-LPBA40 and ADNI datasets show that our method yields improved performance in ROI segmentation of brain MRI, compared with several state-of-the-art methods.

The major contributions of this work can be summarized as follows. *First*, we estimate the reliability (*i.e.*, label reliability and spatial reliability) of each voxel based on results of initial segmentation, where most of existing conventional multi-atlas label fusion method can be used for initial segmentation. This helps us easily embed conventional label fusion methods into our framework. *Second*, voxels with high reliability are used to help refine the label fusion process of the voxels with low reliability in our framework. To the best of our knowledge, this is among the first attempt to utilize the voxel reliability as the prior knowledge for brain MRI segmentation. *Third*, we apply the proposed reliability-based strategy to several state-of-the-art methods on the NIREP, LONI-LPBA40 and ADNI datasets, with experimental results demonstrating the superior performance of our method over the state-of-the-art approaches in brain MRI segmentation. The code is now publicly available.[1]

The remainder of the paper is organized as follows. In Section 2, we describe the proposed reliability-based robust label fusion framework. We present the materials used in the experiments, experimental settings, and experimental results on the NIREP, LONI-LPBA40 and ADNI datasets in Section 3. In Section 4, we compare the proposed method with the state-of-the-art methods and investigate the influence of parameters. Finally, a conclusion of this paper and the limitations of our method as well as possible future work are presented in Section 5.

2. Method

In this section, we first introduce the notations used in this work, and then present the definitions of the proposed label reliability and spatial reliability. Finally, we elaborate our reliability-based robust label fusion framework.

2.1. Notations

Denote $\boldsymbol{T}$ as the target image to be labeled. Let $\boldsymbol{A} = \{\boldsymbol{A}_s | s = 1, \ldots, N\}$ and $\boldsymbol{L} = \{\boldsymbol{L}_s | s = 1, \ldots, N\}$ represent the N atlases and their corresponding label maps, respectively. We first register each atlas image $\boldsymbol{A}_s$ ($s = 1, \ldots, N$) and its corresponding label map $\boldsymbol{L}_s$ ($s = 1, \ldots, N$) onto the target image space. The process of label fusion aims to determine the label map $\boldsymbol{L}_T$ for the target image. We denote $\boldsymbol{P}_T(y)$ as the patch centered at the voxel y in the target image $\boldsymbol{T}$, and $\boldsymbol{P}_{\boldsymbol{A}_s}(x)$ as the patch centered at the voxel x in the atlas $\boldsymbol{A}_s$. Also, we denote the neighborhood of the voxel y in the target image $\boldsymbol{T}$ and atlases $\boldsymbol{A}$ as $\boldsymbol{N}_T(y)$ and $\boldsymbol{N}_A(y)$, respectively. In the following, we briefly review two widely-used label fusion methods, including (1) locally-weighted voting (LWV) method [12] and (2) non-local mean patch-based method (PBM) [13,14].

In LWV method, the patch-wise similarity between the target and each atlas at the same location is used as the voting weight. Specifically, the voting weight is calculated as follows,

$$w(y_i, x_{s,i}) = \exp \frac{-||\boldsymbol{I}(y_i) - \boldsymbol{I}(x_{s,i})||_2^2}{\delta} \tag{1}$$

with

$$\delta(y_i) = \arg\min_{x_{s,i}} ||\boldsymbol{I}(y_i) - \boldsymbol{I}(x_{s,i})||_2 + \varepsilon$$

where y_i represents the i-th voxel in target image and $x_{s,i}$ denote the i-th voxel in the s-th atlas. $\boldsymbol{I}(y_i)$ and $\boldsymbol{I}(x_{s,j})$ represent the normalized intensity of voxels within the patch $\boldsymbol{P}_T(y_i)$ (extracted from the target image) and

[1] https://github.com/sunmoon91/label-spatial-reliability.

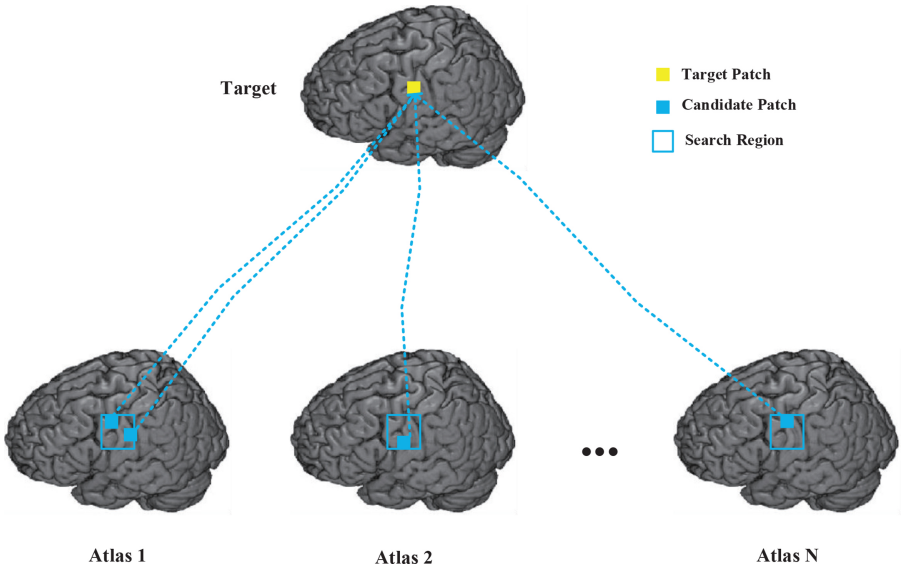

Fig. 1. Overview of patch based multi-atlas segmentation method. The blue rectangles on the atlases represent the search region. The blue patches in the atlases represent the candidate patches and the yellow patch in the target represents the target patch. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the patch $\boldsymbol{P}_{\boldsymbol{A}_s}(y_i)$ (extracted from the atlas $\boldsymbol{A}_s$), respectively. The term $||\cdot||_2$ is the normalized l_2-norm, and ϵ is a small constant to ensure numerical stability. For the voxel y_i of the target image to be segmented, its label is estimated based on a weighted fusion strategy by considering all labeled voxels at the same location within in the atlas images:

$$l(y_i = c) = \frac{\sum_{s=1}^{N} w(y_i, x_{s,i})\{l_{s,i} = c\}}{\sum_{s=1}^{N} w(y_i, x_{s,i})} \tag{2}$$

where $l_{s,i}$ is the label of the voxel $x_{s,i}$ (*i.e.*, the i-th voxel in the s-th atlas $\boldsymbol{A}_s$). Note that $\{l_{s,i} = c\}$ is equal to 1 if $x_{s,i}$ is in the ROI c; and 0, otherwise. Moreover, $w(y_i, x_{s,i})$ is a weight between the voxel y_i in the target image and the voxel $x_{s,i}$ in the atlas, depending on the similarity of these two patches (*i.e.*, $\boldsymbol{P}_{\boldsymbol{T}}(y_i)$ and $\boldsymbol{P}_{\boldsymbol{A}_s}(y_i)$).

Different from LWV, PBM propagates labels from *not only* voxels in the same location, *but also* voxels nearby in the atlas images. It seeks multiple candidates based on the pair-wise similarity between the target image patch and atlas image patches within a certain region, alleviating the registration error and improving the accuracy and robustness of the labeling results [13,14]. As illustrated in Fig. 1, the blue rectangle in each atlas image represents the search region, the blue square in the atlas image represents the candidate patch and the yellow square in the target image represents the target patch. Mathematically, the voting weight in PBM is calculated as follows,

$$w(y_i, x_{s,j}) = \exp \frac{-||\boldsymbol{I}(y_i) - \boldsymbol{I}(x_{s,j})||_2^2}{\delta(y_i)} \tag{3}$$

with

$$\delta(y_i) = \min_{x_{s,j} \in \boldsymbol{N_A}(y_i)} ||\boldsymbol{I}(y_i) - \boldsymbol{I}(x_{s,j})||_2 + \varepsilon$$

where y_i and $x_{s,j}$ represent the i-th voxel in target image and j-th voxel in s-th atlas, respectively. Here, $\boldsymbol{I}(y_i)$ and $\boldsymbol{I}(x_{s,j})$ represent the normalized intensity of the voxels within the patches $\boldsymbol{P}_{\boldsymbol{T}}(y_i)$ and $\boldsymbol{P}_{\boldsymbol{A}_s}(x_{s,j})$, respectively.

For the to-be-segmented voxel y_i, its label is estimated based on a weighted fusion of all labeled voxels inside its neighborhood via

$$l(y_i = c) = \frac{\sum_{s=1}^{N} \sum_{j \in \boldsymbol{N_A}(y_i)} w(y_i, x_{s,j})\{l_{s,j} = c\}}{\sum_{s=1}^{N} \sum_{j \in \boldsymbol{N_A}(y_i)} w(y_i, x_{s,j})} \tag{4}$$

where $l_{s,j}$ is the label of the voxel $x_{s,j}$ (*i.e.*, the j-th voxel in the s-th atlas $\boldsymbol{A}_s$), and the voxel $x_{s,j}$ in the atlas image is within the neighborhood of y_i. The term $w(y_i, x_{s,j})$ is a weight between the voxel y_i and $x_{s,j}$, depending on the similarity between the patch $\boldsymbol{P}_{\boldsymbol{T}}(y_i)$ in the target image and it neighboring patches (*i.e.*, $\boldsymbol{P}_{\boldsymbol{A}_s}(y_i)$ with $j \in \boldsymbol{N_A}(y_i)$) in atlas images. In the next section, we describe the process of label reliability and spatial reliability estimation.

2.2. Proposed label-spatial reliability

For each voxel in the target image, we define two kinds of reliability, including (1) label reliability, and (2) spatial reliability. After initial segmentation by conventional multi-atlas segmentation methods (*e.g.*, LWV and PBM, *etc.*), we get a normalized voting result (within [0,1]) called *soft label* for each voxel in the target image. For each to-be-segmented voxel y_i and the c-th ROI, we assume the soft label $l(y_i = c)$ is the probability of y_i belonging to the region c. Based on the soft label of the voxel y_i, we define its label reliability $lr(y_i)$ using the Shannon entropy, shown as follows,

$$\text{le}(y_i) = -H(y_i) \tag{5}$$

with

$$H(y_i) = -\sum_{c=1}^{C} l(y_i = c) \log l(y_i = c)$$

where $H(y_i)$ is the Shannon entropy of y_i and C is the total number of ROIs in the label map. From Eq. (5), we can observe that the voxel y_i has a high label reliability if its Shannon entropy is low. For example, for a binary segmentation problem, we set the label of y_i to be 1 if the voxel belongs to the foreground; and 0, otherwise. If $l(y_i = 1)$ is close to 1, most of the voters have the same decision, and thus, the voxel will be assigned a high label reliability. On the contrary, when $l(y_i = 1)$ is close to 0.5, half of the voters believe that the voxel belongs to the foreground, and the remaining voters believe the voxel belongs to the background. In this case, it is difficult to determine the real label for this voxel, and hence, the label reliability of y_i defined in Eq. (5) is low. We normalize $le(y_i)$ to the range of [0,1], and get the normalized label reliability $lr(y_i)$.

On the other hand, as reported in previous studies [13,17], voxels along the boundary of ROIs are often easy to be misclassified. Therefore, we also consider the spatial structure information obtained from the initial segmentation. In this study, we assume a voxel has high reliability to be correctly segmented, if its neighborhood voxels have more consistent labels. As an illustration, in Fig. 2, we use the yellow square to represent the voxel that needs to be calculated for spatial

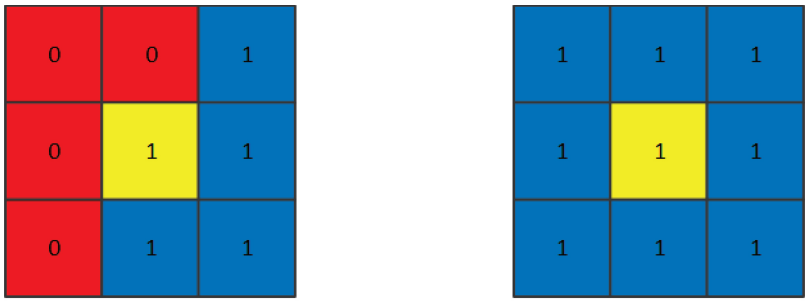

Fig. 2. Two examples in computing spatial reliability. We want to compute the spatial reliability of the yellow squares. The red squares represent voxels belonging to the background, and the blue squares represent voxels belonging to the ROI. The spatial reliability of example (a) is 4/8 = 0.5 and example (b) is 8/8 = 1.0. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reliability, the red squares to represent the voxels with different labels from the yellow square, and the blue squares represent the voxels that have the same label as the yellow square. In Fig. 2(a), the center voxel (*i.e.*, yellow square) and its four neighboring voxels (*i.e.*, red squares) share the same label (*i.e.*, 1). In Fig. 2(b), the center voxel and its eight neighboring voxels have the same label (*i.e.*, 1). Here, we believe that the center voxel in Fig. 2(b) has higher spatial reliability than that in Fig. 2(a). In practice, we calculate the spatial reliability in 3-dimensional image space. With this assumption, we define the spatial reliability of each voxel in the target image in the following. We first compute the *hard label* $L(y_i)$ for each voxel in the target image using conventional label fusion methods as follows

$$L(y_i) = \underset{c=1,\ldots,C}{\arg\max}\ l(y_i = c) \tag{6}$$

Based on the hard labels of voxels in the target image, we define the spatial reliability of each voxel as follows

$$\mathrm{sr}(y_i) = \frac{\#\{L(y_i) = L(y_t)\}}{|y_t|} \tag{7}$$

with

$$y_t \in \boldsymbol{N_T}(y_i)(t = 1, 2, \ldots, n)$$

where $\#\{\Omega\}$ represents the number of true elements in Ω, and $|y_t|$ represents the number of $y_i \in \boldsymbol{N_T}(y_i)$. Accordingly, the spatial reliability of the center voxel in Fig. 2(a) is 4/8 = 0.5, while the spatial reliability of the center voxel in Fig. 2(b) is 8/8 = 1.0.

By combining the proposed label reliability in Eq. (5) and spatial reliability in Eq. (7) together, we compute the final label-spatial reliability for each voxel in the target image as follows

$$r(y_i) = \mathrm{lr}(y_i) \times \mathrm{sr}(y_i) \tag{8}$$

For clarity, in Fig. 3, we show how to compute the proposed label-spatial reliability for each voxel in the target image. First, we perform the conventional multi-atlas method to calculate the soft label of each voxel in the target image. After initial segmentation, we estimate the label reliability-based on soft labels in Fig. 3(a), and then obtain the hard label via Eq. (6) to assign each voxel to a specific ROI or background in Fig. 3(b). Afterward, in Fig. 3(c), we compute the spatial reliability of each voxel based on the estimated spatial structure. Finally, in Fig. 3(d), we combine the label reliability and spatial reliability to get the final label-spatial reliability for each voxel in the target image.

2.3. Label-spatial reliability-based robust label fusion

Based on the proposed label-spatial reliability for each voxel in the target image, we propose to use voxels with low reliability to guide the remaining voxels in the label fusion process. Here, we denote $\boldsymbol{P}$ and $\boldsymbol{Q}$ as the set of voxels with low reliability and the set of voxels with high reliability, respectively. In Fig. 4, we illustrate the proposed label-spatial reliability-based label fusion process, where the red and blue rectangles are the search regions in the target and atlas images, respectively. The blue patches in the atlases represent the candidate patches. The center voxel of the red/green patch in the target image will be assigned label-spatial reliability. Here, we assume that the center of the red patch is the voxel with high label-spatial reliability, while the center of the green patch denotes the voxel with a low value of label-spatial reliability. Then, we use the red patch (with high label-spatial reliability) to guide the labeling procedure of the green patch in the target image.

To be specific, we use the propagated soft label from atlases and voxels in $\boldsymbol{Q}$ (with high reliability) in the target image to guide the label fusion process for each voxel (*e.g.*, y_i) in $\boldsymbol{P}$. That is, for the voxel $y_i \in \boldsymbol{P}$ with low reliability, the patch-wise similarity between y_i and the voxel y_j with high reliability can be calculated as follows

$$w(y_i, y_j) = \exp\frac{-\|\boldsymbol{I}(y_i) - \boldsymbol{I}(y_j)\|_2^2}{\delta'(y_i)} \tag{9}$$

with

$$\delta'(y_i) = \underset{y_j \in \boldsymbol{N_T}(y_i) \cap \boldsymbol{Q}}{\arg\min}\ \|\boldsymbol{I}(y_i) - \boldsymbol{I}(y_j)\|_2 + \varepsilon$$

where $y_j \in \boldsymbol{N_T}(y_i) \cap \boldsymbol{Q}$ denotes the neighborhood of voxel y_i in $\boldsymbol{P}$.

For the voxel y_i in $\boldsymbol{P}$, its estimated label is defined based on all labeled voxels within its neighborhood (*i.e.*, $\boldsymbol{N_T}(y_i) \cap \boldsymbol{Q}$) via the following

$$l_r(y_i = c) = \frac{\sum_{j \in \boldsymbol{N_T}(y_i) \cap \boldsymbol{Q}} w(y_i, y_j) r(y_j) \{L(y_j) = c\}}{\sum_{j \in \boldsymbol{N_T}(y_i) \cap \boldsymbol{Q}} w(y_i, y_j) r(y_j)} \tag{10}$$

where $r(y_j)$ is defined in Eq. (8) and $L(y_j)$ is defined in Eq. (6). In this way, we can employ voxels with high reliability to help refine the label fusion process of those voxels with low reliability in target image.

Similar to conventional label fusion methods, we also propagate the labels from atlases to the target image via Eq. (2) or Eq. (4). Then, for the voxel y_i in the target image, we can calculate its soft label $l_{new}(y_i = c)$ as follows

$$l_{\mathrm{new}}(y_i = c) = \lambda l(y_i = c) + (1 - \lambda) l_r(y_i = c) \tag{11}$$

where $\lambda \in [0, 1]$ is a tuning-parameter to balance the contributions from the label $l(y_i = c)$ estimated by conventional label propagation strategy and the label $l_r(y_i = c)$ generated by the proposed label-spatial reliability-based label fusion strategy. Accordingly, for each voxel y_i in the target image, its final label $L_T(y_i)$ can be obtained via

$$L_T(y_i) = \underset{c=1,\ldots,C}{\arg\max}\ l_{\mathrm{new}}(y_i = c) \tag{12}$$

From Eqs. (11) and (12), we can observe that one can easily embed conventional label fusion methods into our label-spatial reliability-based label fusion framework, since most of existing conventional multi-atlas methods can be used for initial segmentation. Accordingly, under the proposed framework, we now develop four new label fusion methods, *i.e.*, label-spatial reliability-based LWV (ls-LWV), label-spatial reliability-based PBM (ls-PBM), label-spatial reliability-based JLF (ls-

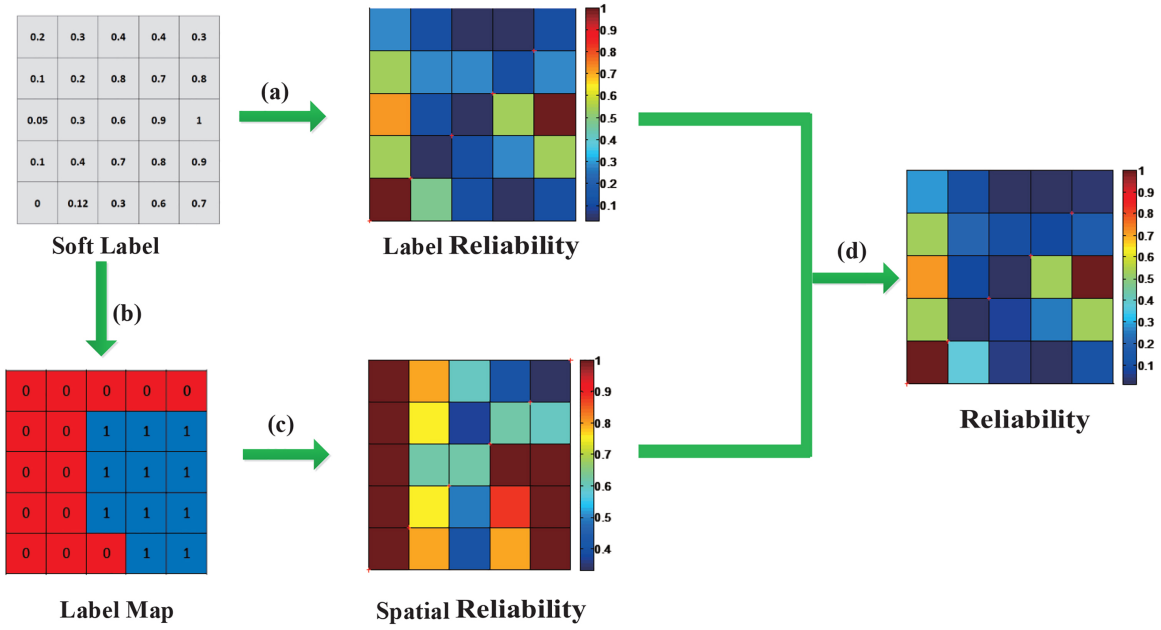

Fig. 3. An illustration of computations of the proposed label-spatial reliability for each voxel in the target image. We first perform a conventional multi-atlas method (*e.g.*, LWV or PBM) to calculate the soft label of each pixel in the target image. After initial segmentation, we estimate the label reliability-based on soft labels in step (a), and then obtain the hard label via Eq. (6) to assign each voxel to a specific ROIs or background in step (b). Afterward, in step (c), we compute the spatial reliability of each voxel based on the estimated spatial structure. Finally, in step (d), we combine the label reliability and spatial reliability together (via Eq. (8)) to get the final label-spatial reliability for each voxel in the target image.

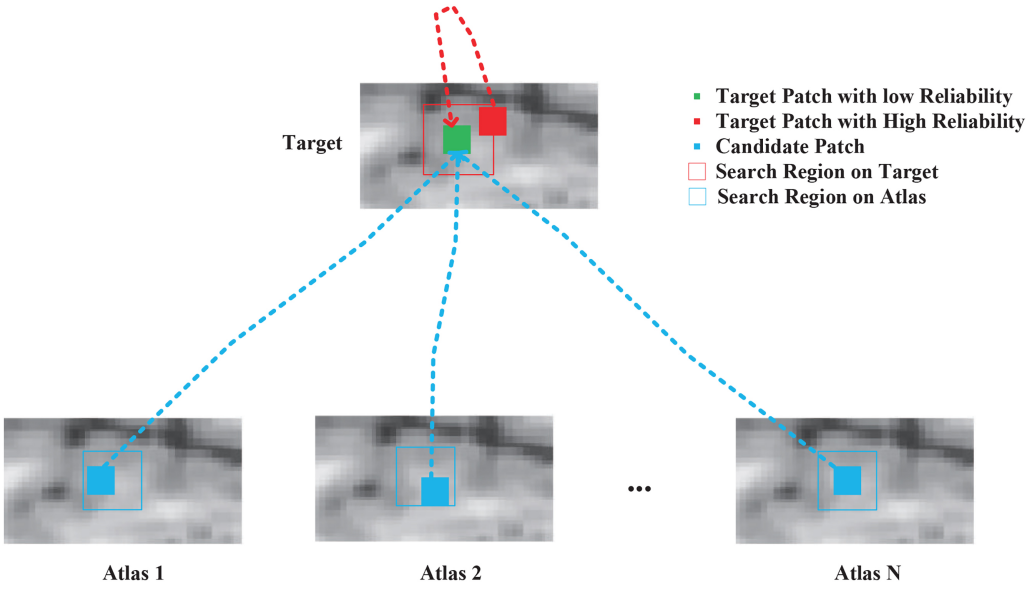

Fig. 4. Overview of proposed label-spatial reliability-based robust label fusion. The red and blue rectangles are the search regions in the target and atlas images, respectively. The blue patches in the atlases represent the candidate patches. The center voxel of the red/green patch in the target image will be assigned a label-spatial reliability, based on both the label reliability of patches in the atlas images and the spatial reliability by their spatial structure in the target image. Here, we assume that the center of the red patch is the voxel with high label-spatial reliability, while the center of the green patch denote the voxel with a low value of label-spatial reliability. Then, we use the red patch (with high label-spatial reliability) to guide the labeling procedure of the green patch. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

JLF) and label-spatial reliability-based SPBM (ls-SPBM).

In Algorithm 1, we show an overview of the proposed label-spatial reliability-based robust label fusion framework. Specifically, the input of the proposed method includes a set of MR images as atlases and the target image to be segmented, and these images have been aligned to a common space. We first compute the soft label for each voxel in the target image via a conventional label fusion method (*e.g.*, LWV or PBM). Then, we estimate the label reliability-based on the soft label and the spatial reliability-based on the spatial structure, followed by combing the label reliability and spatial reliability to get the final label-spatial reliability. Finally, voxels with high reliability to be correctly segmented are employed to guide the label fusion process of voxels with low reliability, and then we can obtain the final label map for the target image.

Algorithm 1. Reliability-based robust label fusion for multi-atlas segmentation

Input: Atlas $\boldsymbol{A} = \{\boldsymbol{A}_s | s = 1, \ldots, N\}$, the label maps $\boldsymbol{L} = \{\boldsymbol{L}_s | s = 1, \ldots, N\}$ for atlas images, and the target image $\boldsymbol{T}$.
Output: Label map $\boldsymbol{L}_T$ for the target image $\boldsymbol{T}$.
1 Compute the soft label by a conventional label fusion method.
2 Estimate the label reliability of each voxel in $\boldsymbol{T}$ using Eq. (5).
3 Estimate the spatial reliability of each voxel in $\boldsymbol{T}$ via Eq. (7).
4 Combine the proposed label reliability and spatial reliability of each voxel in $\boldsymbol{T}$ via Eq. (8).
5 Propagate the labels of voxels with high reliability to voxels with low reliability using Eq. (10).
6 Reliability-based robust label fusion via Eq. (11).

7 Assign voxels in $\boldsymbol{T}$ to a specific ROI or background based on their estimated labels using Eq. (12).

3. Experiments

3.1. Data and image pre-processing

In the experiments, we validate our proposed methods and those competing methods on the segmentation of regions-of-interest (ROIs) in brain MR images.

(1) **NIREP dataset** [34]: This dataset consists of 16 subjects with T1-weighted MR images, including 8 normal male adults and 8 female adults. The MR images were obtained in a General Electric Signa scanner operating at 1.5 T, using the following protocol: SPGR/50, TR 24, TE 7, NEX 1 matrix 256 × 192, FOV 24 cm. 124 contiguous coronal slices were obtained, with 1.5 or 1.6 mm thick, and with an interpixel distance of 0.94 mm. The images are resized from voxel dimensions 0.7 mm × 0.7 mm × 1.5 mm to 0.7 mm × 0.7 mm × 0.7 mm, and the image size changed from 256 × 256 × 124 to 256 × 300 × 256. These MR images have been manually segmented into 32 ROIs. For each of the ROIs, a Leave-One-Out (LOO) cross-validation is performed to test the segmentation performance on each LOO fold. That is, each of 16 subjects are alternatively used as the target subject, and MR images of the remaining 15 subjects are treated as the atlases (with images aligned onto the target image).

(2) **LONI-LPBA40 dataset** [35]: This dataset is provided by the Laboratory of Neuro Imaging (LONI) at UCLA, containing 40 brain MR images and corresponding label maps that were created manually to annotate the brain structures. High-resolution 3D Spoiled Gradient Echo (SPGR) MRI volumes were acquired on a GE 1.5 Tesla system as 124 contiguous 1.5 mm coronal brain slices (TR range 10.00–12.50 ms; TE range 4.22–4.50 ms; FOV = 220 mm or 200 mm) with in-plane voxel resolution of 0.86 mm or 0.7 mm. Besides, these MRI volumes are rigidly aligned to the MNI305 template. Specifically, each MR image has 54 manually labeled ROIs. We randomly select 20 subjects as the atlas and remained 20 subjects as testing.

(3) **Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset**[2]: Similar to [17,24], we randomly select 60 subjects from the ADNI dataset for *hippocampus* segmentation, which include 20 Alzheimer's disease (AD) subjects, 20 mild cognitive impairment (MCI) subjects and 20 normal control (NC) subjects. These images were acquired sagittally, with the in-plane resolution of 1 mm × 1 mm and the slice thickness of 1.2 mm. We perform skull removal [36], N4-based bias field correction [37] and intensity standardization to normalize the intensity range [38] for pre-processing. A LOO cross-validation is performed to test the segmentation performance on each LOO fold.

For each MR image, we first perform affine registration by FLIRT in the FSL toolbox [27], which using the normalized mutual information similarity metric, 12 degrees of freedom and the search range ± 20 in all directions. Then after the affine registration, a deformable registration is performed using the Diffeomorphic Demons method [39] with smooth sigma 2.0 and iterations in low middle and high resolutions as 20 × 10 × 5.

3.2. Competing methods

As mentioned in the Section 2.3, our proposed four label-spatial reliability-based robust label fusion methods (called ls-LWV, ls-PBM, ls-JLF and ls-SPBM) are built on LWV [12], PBM [13], JLF [40] and SPBM [16], respectively. To evaluate the effectiveness of the proposed label-spatial reliability-based robust label fusion framework, we compare the ls-LWV, ls-PBM, ls-JLF and ls-SPBM methods with their conventional counterparts, *i.e.*, LWV, PBM, JLF and SPBM, on the NIREP, LONI-LPBA40 and ADNI datasets.

3.3. Experimental settings

For segmentation results achieved by a specific algorithm, we use the Dice ratio to measure the overlap between the region A and the region B. Dice ratio is defined as follows

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{13}$$

where A and B denote the estimated ROI achieved by a particular algorithm and the manual segmented ROI (*i.e.*, ground truth), respectively. The term ∩ denotes the overlap between automatic segmentation and ground truth, and $|\cdot|$ denotes the number of voxels in the ROI. When calculating the Dice ratio for each of multiple ROIs, we perform an independent binary assessment for each ROI. That is, we set the label of a voxel as 1 if it belongs to the specific ROI; and 0, otherwise.

For the fair comparison, in all methods, we use the size of 7 × 7 ×7 neighborhood search region in the atlas images, and 7 × 7 ×7 neighborhood search region in the target image. We also use the spatial size of 7 × 7 ×7 for estimating the spatial reliability of each voxel in the target image. In addition, in our experiments, we empirically fix the patch size as 7 × 7 ×7 for all competing methods. We perform patch pre-selection [41] to reduce the computational time according to the similarity between any pair of patches. In the step of label-spatial reliability-based label fusion, we divided the voxels into 20 subsets according to their reliability with 0.05 as the interval (*i.e.*, [0.95, 0.90, …, 0.05, 0]). We gradually use the subsets with high reliability to help refine the subset with low reliability.

3.4. Results on NIREP

We first compare the segmentation results of different methods on NIREP dataset. Table 1 reports the segmentation results in terms of average Dice ratio of 32 ROIs in brain MR images. From Table 1, one can observe that the proposed ls-LWV, ls-PBM, ls-JLF and ls-SPBM achieve the improvement of 2.57%, 1.14%, 1.40% and 1.42% over their conventional counterparts (*i.e.*, LWV, PBM, JLF and SPBM) in the terms of Dice ratio, respectively. These results demonstrate that using our proposed label-spatial reliability as guidance information provide a practical solution in promoting the segmentation performance of conventional label fusion algorithms.

We also report the Dice ratio achieved by eight different methods for each of 32 ROIs in Fig. 5. It can be seen from Fig. 5 that the proposed ls-LWV, ls-PBM, ls-JLF and ls-SPBM methods consistently outperform their conventional counterparts (*i.e.*, LWV, PBM, JLF and SPBM) in segmenting all ROIs on the NIREP dataset.

In Fig. 6, we visually plot the segmentation results for the region of *L insula gyrus*, achieved by eight methods. For comparison, we also show the original image in Fig. 6(a) and the ground truth in Fig. 6(b). We can see from Fig. 6 that our proposed ls-LWV, ls-PBM, ls-JLF and ls-SPBM achieve the better visual quality of segmentation results, compared with LWV, PBM, JLF, SPBM, respectively. This further validates the effectiveness of the proposed label-spatial reliability-based label fusion framework in ROI segmentation of brain MRI.

3.5. Results on LONI-LPBA40

In this section, we compare the segmentation results of different methods on the LONI-LPBA40 dataset. Table 2 shows the results of Dice ratio for 54 ROIs. Our proposed ls-LWV, ls-PBM, ls-JLF and ls-SPBM methods yield the improvement of 2.21%, 1.04%, 1.44% and 1.26%

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

Table 1
Segmentation results of Dice ratio achieved by eight different methods (*i.e.*, LWV, PBM, JLF, SPBM and our proposed ls-LWV, ls-PBM, ls-JLF, ls-SPBM methods) on NIREP dataset. The second and third rows are mean and standard deviation (std) of Dice ratio for different ROIs, respectively.

Method	LWV	ls-LWV	PBM	ls-PBM	JLF	ls-JLF	SPBM	ls-SPBM
Mean (%)	75.02	**77.59**	76.74	**77.88**	78.25	**79.65**	78.48	**79.90**
Std (%)	4.86	4.60	4.57	4.50	4.05	4.23	4.14	4.27

Table 2
Segmentation results of Dice ratio achieved by eight different methods (*i.e.*, LWV, PBM, JLF, SPBM and our proposed ls-LWV, ls-PBM, ls-JLF, ls-SPBM methods) on LONI-LPBA40 dataset. The second and third rows are mean and standard deviation (std) of Dice ratio for different ROIs, respectively.

Method	LWV	ls-LWV	PBM	ls-PBM	JLF	ls-JLF	SPBM	ls-SPBM
Mean (%)	78.22	**80.43**	78.81	**79.85**	79.26	**80.70**	79.65	**80.91**
Std (%)	4.73	4.38	5.50	4.85	4.46	4.39	4.59	4.69

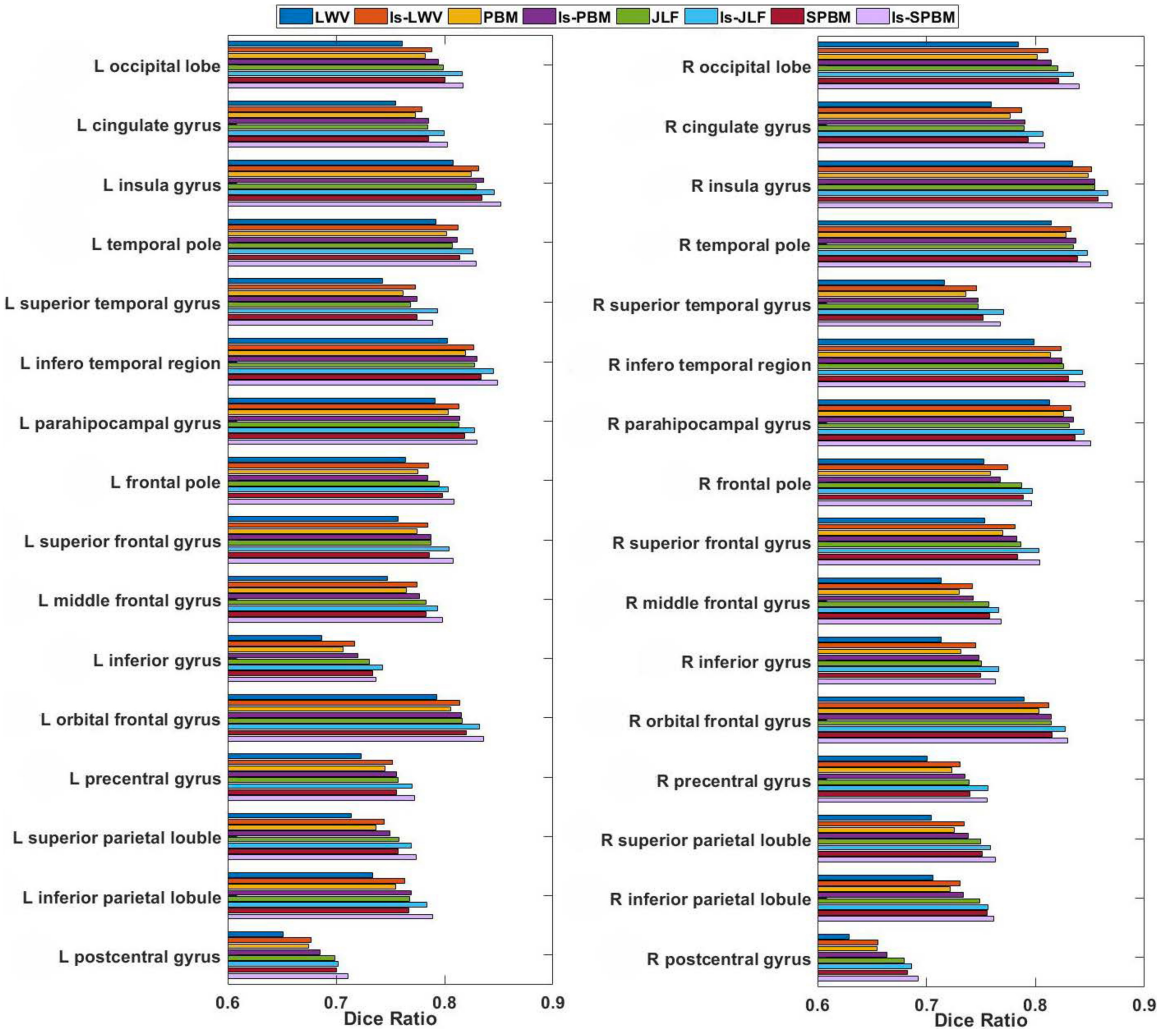

Fig. 5. Segmentation results of 32 ROIs achieved by LWV, ls-LWV, PBM, ls-PBM, JLF, ls-JLF, SPBM and ls-SPBM on the NIREP dataset.

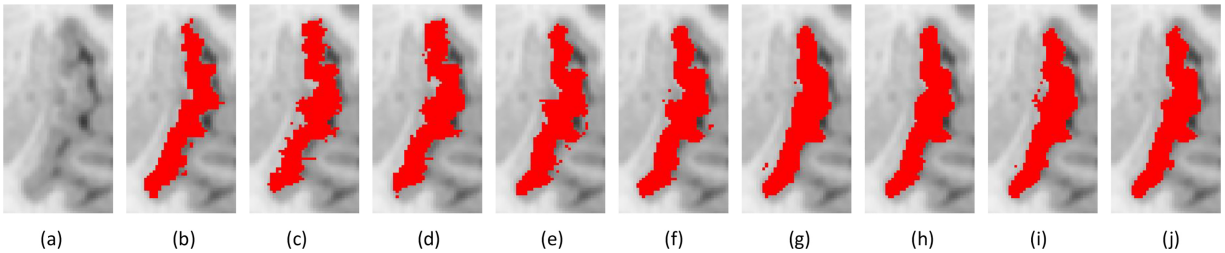

Fig. 6. Visual views on original image (a), ground truth (b) and segmentation results of LWV (c), ls-LWV (d), PBM (e), ls-PBM (f), JLF (g), ls-JLF (h), SPBM (i) and ls-SPBM (j) for the region of *L insula gyrus* on the NIREP dataset.

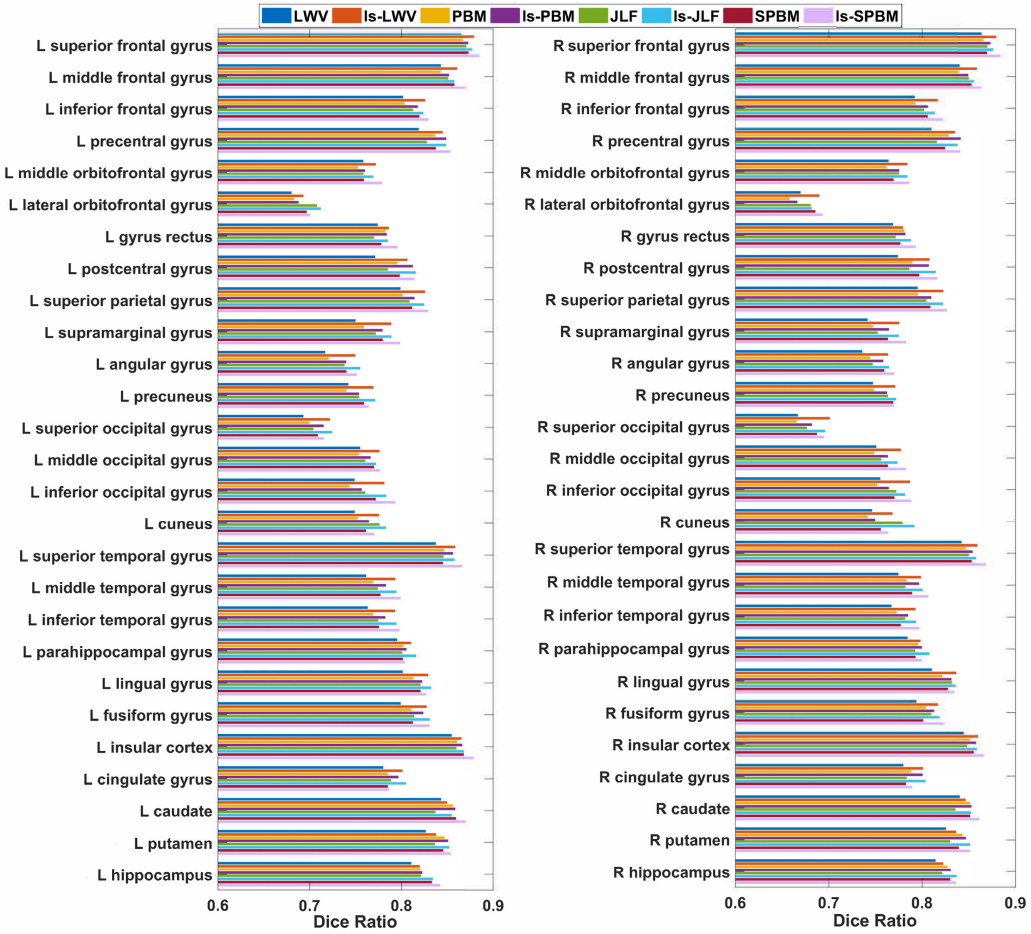

Fig. 7. Segmentation results of 54 ROIs achieved by LWV, ls-LWV, PBM, ls-PBM, JLF, ls-JLF, SPBM and ls-SPBM on the LONI-LPBA40 dataset.

Table 3
The Dice ratio achieved by eight different methods (*i.e.*, LWV, PBM, JLF, SPBM and our proposed ls-LWV, ls-PBM, ls-JLF, ls-SPBM methods) on the ADNI dataset for *hippocampus* segmentation. The second and third rows are mean and standard deviation (std) of Dice ratio for different subjects, respectively.

Method	LWV	*ls-LWV	PBM	*ls-PBM	JLF	*ls-JLF	SPBM	*ls-SPBM
Mean (%)	85.48	**87.21**	86.31	**87.65**	87.08	**88.32**	87.19	**88.43**
Std (%)	1.95	1.71	3.87	2.61	3.05	2.76	3.10	2.93

over LWV, PBM, JLF and SPBM, respectively, by introducing the proposed label-spatial reliability. Fig. 7 shows the Dice ratio achieved by 8 methods on each of 54 ROI segmentation on the LONI-LPBA40 dataset. It can be observed from Fig. 7 that the proposed methods obtain overall better performance, compared with their conventional counterparts.

3.6. Results on ADNI

We perform experiments on the ADNI dataset for *hippocampus* segmentation. We report the segmentation results achieved by our proposed methods (*i.e.*, ls-LWV, ls-PBM, ls-JLF and ls-SPBM) and the conventional state-of-the-art counterparts (*i.e.*, LWV, PBM, JLF, and SPBM) in Table 3. As can be seen from Table 3, the average Dice ratio achieved by ls-LWV, ls-PBM, ls-JLF and ls-SPBM methods are 87.21 ± 1.71%, 87.65 ± 2.61%, 88.32 ± 2.76% and 88.43 ± 2.93%, respectively, which are superior to LWV, PBM, JLF and SPBM, respectively. These results suggest that our proposed methods outperform these state-of-the-art methods, thus further demonstrating the effectiveness of our label-spatial reliability-based label fusion strategy.

4. Discussion

In this section, we first compare our proposed framework with a corrective learning method [40]. We then analyze the statistical significance of the difference between our methods and each of competing methods, and analyze the robustness of the proposed method. We also investigate the influence of the proposed label reliability and spatial reliability, as well as the effect of the search region in the target image and the parameter λ in Eq. (11).

4.1. Comparison with corrective learning method

To evaluate the effectiveness of our proposed label-spatial reliability-based label fusion framework, we also compare our proposed ls-JLF with the joint label fusion with corrective learning (JLF-CL) [40]. The JLF-CL method first segments the images using JLF. Then, JLF-CL segment each atlas image using the remaining atlases, and using these segmented atlas images to train the corrective classifiers. As can be seen in Table 4, our proposed ls-JLF method achieves the competitive results with JLF-CL on NIREP, LONI-LPBA40 and ADNI datasets for ROI segmentation. It is worth noting that, our proposed label refine method is a

Table 4
Segmentation results of JLF-CL and our proposed ls-JFL on NRIEP, LONI-LPBA40 and ADNI dataset. The terms a and b in "$a \pm b$" denote the mean and standard deviation, respectively.

	NIREP		LONI-LPBA40		ADNI	
	JLF-CL	ls-JLF	JLF-CL	ls-JLF	JLF-CL	ls-JLF
Dice ratio(%)	79.26 ± 4.26	79.65 ± 4.23	80.48 ± 4.24	80.70 ± 4.39	88.47 ± 2.42	88.32 ± 2.76

Table 5
The p-values based on paired t-test in terms of Dice ratio on NIREP dataset for the proposed ls-LWV, ls-PBM, ls-JLF, ls-SPBM with their conventional counterparts (*i.e.*, LWV, PBM, JLF, SPBM), respectively.

	ls-LWV		ls-PBM		ls-JLF		ls-SPBM	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*
Occipital lobe	2.03e−08	8.06e−10	2.36e−06	7.28e−08	7.12e−04	3.12e−05	0.0018	5.81e−06
Cingulate gyrus	1.44e−08	1.53e−12	3.61e−07	1.88e−09	1.72e−04	1.12e−06	4.18e−05	2.27e−04
Insula gyrus	7.16e−11	8.42e−11	1.43e−08	9.41e−06	2.07e−07	1.26e−05	4.63e−07	3.94e−05
Temporal pole	5.06e−08	2.64e−07	6.50e−05	5.32e−07	2.18e−04	1.76e−04	0.0052	0.0028
Superior temporal gyrus	6.30e−10	1.01e−08	6.01e−08	2.19e−05	1.54e−05	9.18e−04	0.0062	0.0140
Infero temporal region	1.27e−11	2.68e−10	1.52e−09	4.79e−07	2.14e−06	5.86e−06	2.20e−06	7.34e−04
Parahippocampal gyrus	2.08e−12	2.44e−13	1.26e−10	5.33e−09	4.87e−06	2.41e−05	0.0038	8.07e−05
Frontal pole	4.03e−06	1.47e−05	5.69e−04	9.77e−04	0.3070	0.1601	0.1346	0.2287
Superior frontal gyrus	1.49e−11	1.42e−09	2.96e−09	6.06e−07	6.38e−04	0.0018	2.68e−05	1.22e−04
Middle frontal gyrus	5.96e−09	9.38e−09	5.36e−06	3.59e−06	0.0242	0.0122	0.0031	9.11e−04
Inferior gyrus	8.16e−08	3.67e−10	1.78e−05	3.14e−08	0.0816	0.0181	0.6981	0.0092
Orbital frontal gyrus	1.54e−09	2.27e−08	1.40e−06	1.81e−06	4.84e−05	0.0016	2.92e−06	1.84e−05
Precentral gyrus	9.79e−10	4.29e−10	2.55e−06	4.33e−07	0.0088	3.76e−04	0.0014	0.0012
Superior parietal lobule	4.83e−10	4.08e−09	1.20e−07	2.97e−07	0.0369	0.1211	0.0153	0.0442
Inferior parietal lobule	3.45e−10	1.77e−08	1.56e−08	1.71e−06	0.0015	0.1966	6.26e−04	0.1701
Postcentral gyrus	5.60e−05	1.63e−05	0.0103	0.0057	0.6083	0.1917	0.1425	0.0881

Table 6
The p-values based on paired t-test in terms of Dice ratio on LONI-LPBA40 dataset for the proposed ls-LWV, ls-PBM, ls-JLF, ls-SPBM with their conventional counterparts (*i.e.*, LWV, PBM, JLF, SPBM), respectively.

	ls-LWV		ls-PBM		ls-JLF		ls-SPBM	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*
Superior frontal gyrus	3.73e−16	5.29e−17	3.18e−14	5.99e−12	7.02e−06	6.82e−04	2.61e−10	1.48e−11
Middle frontal gyru	8.31e−13	5.38e−14	4.06e−12	9.42e−14	7.80e−04	5.07e−04	1.09e−08	1.73e−09
Inferior frontal gyrus	3.32e−13	1.65e−11	4.29e−11	6.82e−12	3.70e−04	5.70e−04	0.0026	4.82e−06
Precentral gyrus	2.32e−16	6.13e−15	3.91e−15	1.44e−12	5.92e−08	6.68e−09	4.04e−07	2.30e−09
Middle orbitofrontal gyrus	8.54e−09	4.12e−09	3.52e−06	4.67e−08	0.0073	0.1916	5.57e−06	0.0020
Lateral orbitofrontal gyrus	1.47e−04	9.86e−05	0.0485	0.0151	0.5634	0.8714	0.5043	0.1570
Gyrus rectus	2.60e−06	2.25e−05	0.4867	0.5357	0.0010	2.21e−04	1.38e−05	1.25e−04
Postcentral gyrus	1.11e−17	4.96e−17	1.49e−13	2.20e−12	1.97e−09	2.40e−07	2.45e−06	2.21e−07
Superior parietal gyrus	2.90e−13	1.73e−13	2.04e−11	9.31e−12	1.62e−04	1.03e−05	3.49e−05	7.64e−06
Supramarginal gyrus	4.90e−12	5.06e−10	8.98e−10	8.37e−07	0.0132	1.64e−04	0.0018	3.58e−04
Angular gyrus	1.17e−12	1.08e−11	5.22e−12	5.70e−09	1.81e−04	0.0026	0.0356	0.0597
Precuneus	5.09e−12	1.47e−12	3.56e−10	3.33e−10	7.15e−06	0.0059	0.2715	0.8848
Superior occipital gyrus	6.75e−08	8.12e−10	1.19e−06	1.61e−07	0.0184	0.0079	0.3672	0.2427
Middle occipital gyrus	2.21e−08	8.19e−12	3.62e−09	3.46e−11	0.0028	2.91e−05	0.0453	9.89e−05
Inferior occipital gyrus	3.13e−08	5.77e−13	1.70e−07	2.23e−08	0.0018	0.0213	2.05e−05	1.08e−05
Cuneus	2.82e−09	3.84e−06	3.56e−06	0.0011	0.2723	0.0656	0.3920	0.3801
Superior temporal gyrus	5.45e−15	4.07e−12	1.31e−11	4.37e−09	3.38e−06	0.0067	2.62e−08	6.22e−06
Middle temporal gyrus	3.00e−12	2.79e−12	7.57e−12	5.46e−14	4.35e−06	2.10e−07	4.40e−06	3.03e−04
Inferior temporal gyrus	4.10e−15	2.39e−13	2.93e−11	1.71e−11	3.29e−06	3.24e−04	3.16e−05	1.05e−05
Parahippocampal gyrus	7.46−08	2.45e−07	0.0183	0.0103	9.06e−04	0.0108	0.5623	0.0766
Lingual gyrus	3.36e−16	4.40e−12	1.33e−06	1.62e−07	2.85e−04	0.2567	0.1639	0.0555
Fusiform gyrus	9.16e−12	4.29e−12	2.96e−10	7.08e−07	6.97e−04	0.0245	8.017e−06	6.06e−08
Insular cortex	3.41e−09	5.34e−13	1.95e−06	1.66e−08	8.95e−06	1.00e−04	2.26e−06	6.99e−05
Cingulate gyrus	1.88e−10	9.73e−10	1.72e−09	3.33e−08	2.43e−04	4.85e−07	0.8125	0.2688
Caudate	0.0014	0.0049	0.0676	0.3311	2.36e−05	7.97e−06	9.54e−08	4.57e−07
Putamen	2.99e−11	6.78e−12	2.36e−04	2.56e−04	3.18e−05	6.87e−05	2.51e−04	0.0022
Hippocampus	9.94e−07	2.42e−07	0.2005	0.0064	2.59e−04	0.0018	0.0014	0.0012

lazy learning method that does not need to learn models to refine labels. In contrast, after segmentation, JLF-CL further performs the segmentation on the atlas images and learning a model to refine segmentation results. Hence, our proposed framework is much faster than JLF-CL. For example, the time consumption of the refining process in our proposed ls-JLF is about 12 s for each *hippocampus*, compared with JLF-CL need 45 min to refine the segmentation result.

Table 7
The p-values based on paired t-test in terms of Dice ratio on ADNI dataset for the proposed ls-LWV, ls-PBM, ls-JLF, ls-SPBM with their conventional counterparts (*i.e.*, LWV, PBM, JLF, SPBM), respectively.

Method	ls-LWV	ls-PBM	ls-JLF	ls-SPBM
p-value	6.17e−29	**0.0091**	2.68e−04	**7.30e−04**

Table 8
The mean and trimmed mean segmentation results of Dice ratio achieved by our proposed ls-LWV and ls-PBM methods on the NIREP and LONI-LPBA40 datasets.

	NIREP		LONI-LPBA40	
	ls-LWV	ls-PBM	ls-LWV	ls-PBM
Mean (%)	77.59	77.88	80.43	79.85
Trimmed mean (%)	77.90	78.14	80.67	80.15

4.2. Significance analysis

To validate whether our proposed methods are statistically significantly better than their conventional counterparts, we perform paired t-test in terms of Dice ratio on each ROI for our proposed methods and their conventional counterparts. Here, we report the p-values in each ROI for the NIREP, LONI-LPBA40 and ADNI datasets, with results are reported in Table 5, Table 6, and Table 7, respectively. It can be seen from the Tables 5 and 6, our proposed ls-LWV, ls-PBM, ls-JLF and ls-SPBM show significant improvement in the most of ROIs over LWV, PBM, JLF and SPBM on NIREP and LONI-LPBA40 datasets, respectively. Also, one can observe from Table 7 that all of our proposed methods achieved significant improvement over the conventional methods in terms of Dice ratio on ADNI for *hippocampus* segmentation.

4.3. Robustness analysis

To validate the robustness of our proposed methods, we calculate the 10% trimmed mean of the experimental results, and report these results in Table 8. As can be seen from Table 8, compared with the average Dice ratio of 32 ROIs on NIREP (*i.e.*, 77.59% by ls-LWV and 77.88% by ls-PBM), the trimmed Dice ratio mean are 77.90% by ls-LWV and 78.14% by ls-PBM, respectively. Besides, on the LONI-LPBA dataset, the trimmed Dice ratio mean of 54 ROIs are 80.67% by ls-LWV and 80.15% by ls-PBM, while the average Dice ratio for these ROIs is 80.43% of ls-LWV and 79.85% of ls-PBM. These results imply that the segmentation results achieved by our ls-LWV and ls-PBM methods are less prone to be affected by outliers, thus suggesting the robustness of our proposed methods.

Table 9
Segmentation results of Dice ratio achieved by PBM, l-PBM, s-PBM and ls-PBM on the NIREP dataset. The terms a and b in "$a \pm b$" denote the mean Dice ratio and standard deviation for different ROIs, respectively.

Method	PBM	l-PBM	s-PBM	ls-PBM
Dice ratio (%)	76.74 ± 4.57	77.46 ± 4.50	77.21 ± 4.70	77.88 ± 4.5

4.4. Effect of label reliability and spatial reliability

We now investigate the effect of the proposed label reliability in Eq. (5), spatial reliability in Eq. (7), and label-spatial reliability in Eq. (8). Using PBM for initial segmentation, we denote l-PBM, s-PBM, and ls-PBM as reliability-based label fusion methods using only label reliability, only spatial reliability and label-spatial reliability, respectively. Here, we employ the error rate (*ER*) to measure the ratio of misclassified voxels with certain reliability. Specifically, the error rate is defined as: $\mathrm{ER}(\theta) = \frac{m}{M}$, where m is the number of misclassified voxels with reliability larger than θ, M is the total number of voxels with reliability larger than θ.

In the first group of experiments, we vary the value of reliability in the range of [0.0, 0.1, …, 1.0], and record the *ER* achieved by three different methods for segmenting the region of *L insula gyrus* on the NIREP dataset in Fig. 8. Note that, in Fig. 8, the term "reliability" for three methods has different meanings: (1) the label reliability for l-PBM, (2) the spatial reliability for s-PBM, and (3) the label-spatial reliability for ls-PBM. It can be seen from Fig. 8 that our ls-PBM method with label-spatial reliability consistently outperforms l-PBM (using only the label reliability) and s-PBM (with only the spatial reliability). Particularly, when the reliability is equal to 1, we can obtain the best

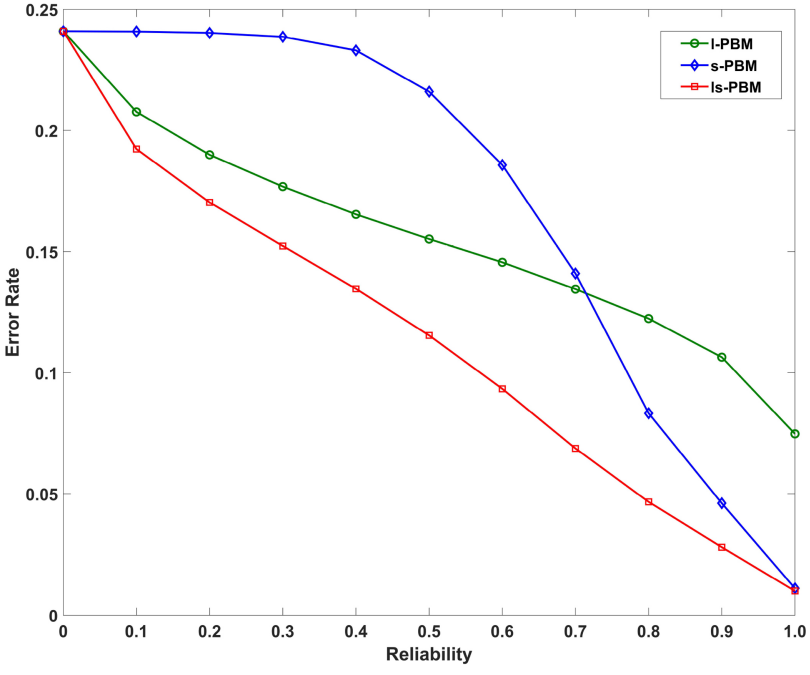

Fig. 8. The trends of error rate along with reliability. The horizontal axis represents reliability. The vertical axis represents segmentation error rate achieved by different methods with different values of reliability. Note that the term "reliability" for three methods has different meanings: (1) the label reliability for l-PBM, (2) the spatial reliability for s-PBM, and (3) the label-spatial reliability for ls-PBM.

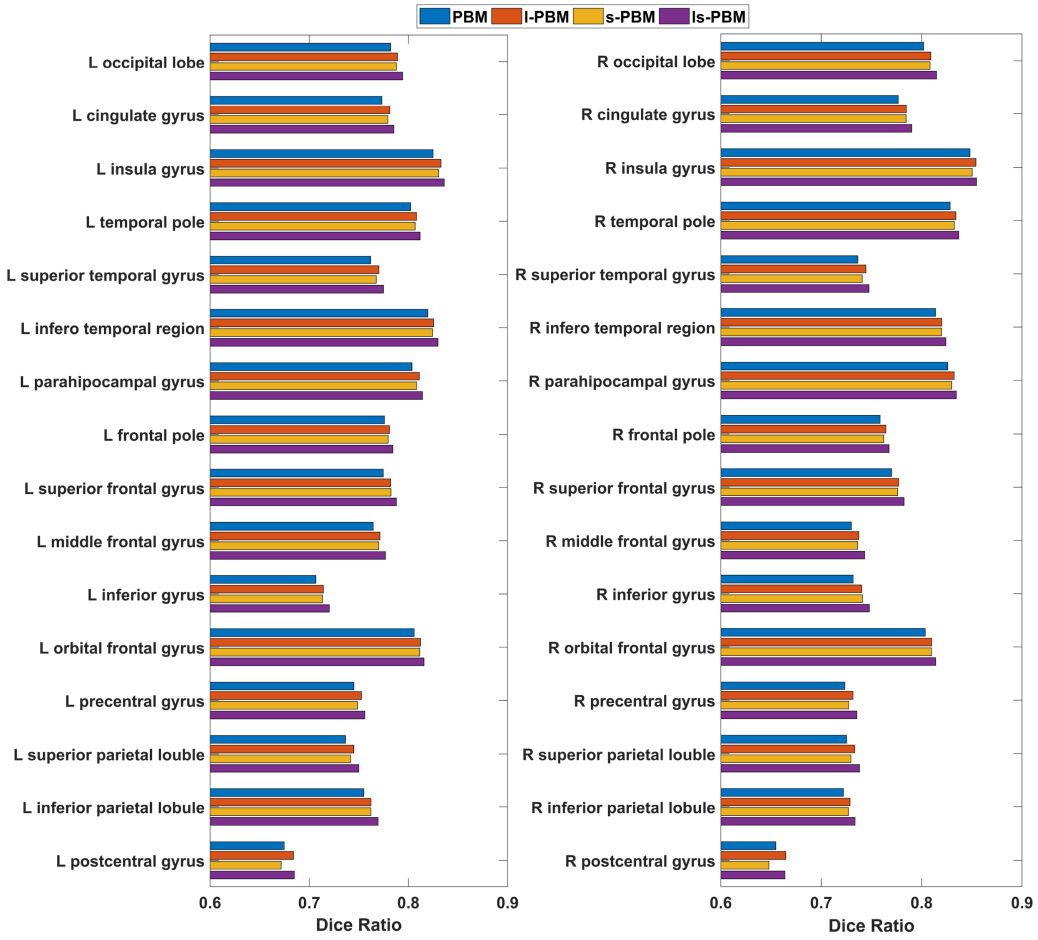

Fig. 9. Segmentation results of 32 ROIs achieved by PBM, l-PBM, s-PBM and ls-PBM on the NIREP dataset.

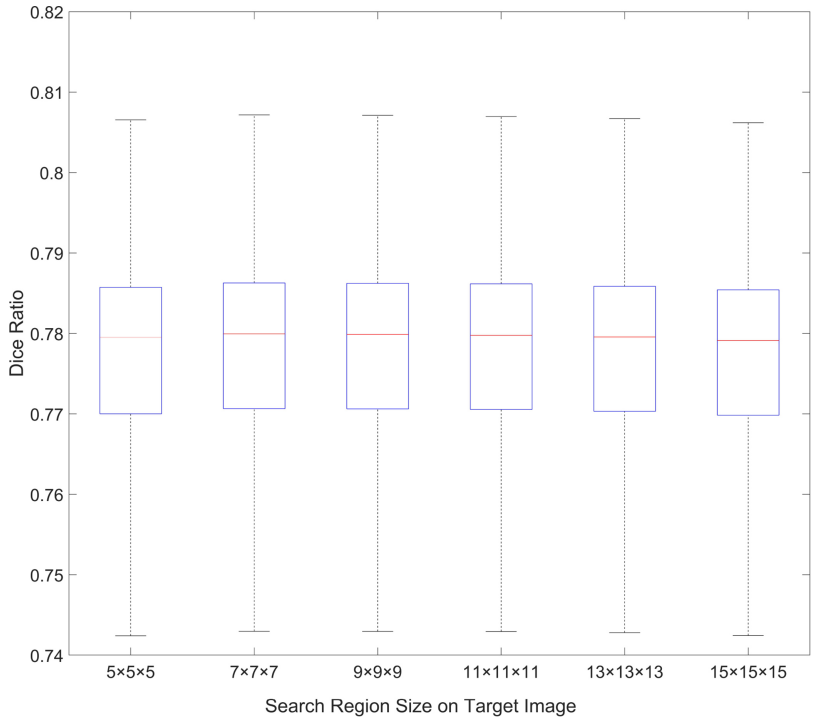

Fig. 10. Performance of the proposed ls-PBM method in segmenting the 32 ROIs on the NIREP dataset, using different search regions in the target image.

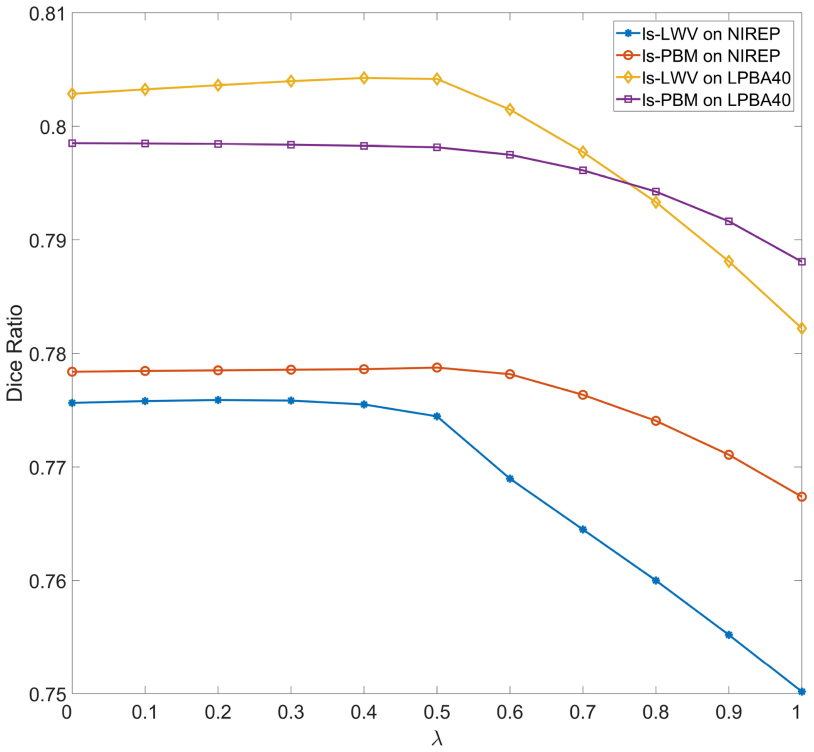

Fig. 11. Dice ratio achieved by the proposed ls-LWV and ls-PBM methods, using different values of λ on the NIREP and LONI-LPBA40 datasets.

results, *i.e.*, the segmentation error rates of 7.48%, 1.11% and 1.00% by l-PBM, s-PBM and ls-PBM, respectively. This implies that using our proposed high reliability would generate low segmentation error rate.

In the second group of experiments, we compare PBM and our proposed l-PBM, s-PBM, and ls-PBM methods in segmenting 32 ROIs on the NIREP dataset. Table 9 gives the results of the average Dice ratio (mean ± standard deviation). One can see that, compared with PBM, l-PBM and s-PBM only utilizing label reliability and spatial reliability achieve the improvement of 0.72% and 0.47% improvement, respectively, while ls-PBM (with label-spatial reliability) yields an improvement of 1.14%. These results further suggest that the method using the proposed label-spatial reliability could further boost the segmentation results, compared with those using only the label reliability or only the spatial reliability. Fig. 9 shows the Dice ratio for these 32 ROIs by PBM, l-PBM, s-PBM and ls-PBM. It can be seen that ls-PBM consistently outperforms PBM, l-PBM and s-PBM in segmenting all 32 ROIs. In addition, the proposed l-PBM and s-PBM methods do not show significant improvement when comparing to PBM. These results further suggest that methods using the proposed label-spatial reliability could further boost the segmentation results, compared with those using only the label reliability or just the spatial reliability.

4.5. Effect of search region

We also study the effect of the search region in the target image. In this group of experiments, PBM is used for initial segmentation, and we report the performance of our ls-PBM method in segmenting 32 ROIs on the NIREP dataset. Specifically, for computing the spatial reliability in ls-PBM, we vary the size of search region within the range of $[5 \times 5 \times 5, 7 \times 7 \times 7, \ldots, 15 \times 15 \times 15]$, and record the segmentation results in Fig. 10. From Fig. 10, we can see that the overall performance of our ls-PBM method is stable by using different size of search regions, while the best performance (*i.e.*, 77.88%) is achieved by using the search region of $7 \times 7 \times 7$. These results demonstrate the size of the search region has little effect on the performance of our proposed ls-PBM method.

4.6. Effect of parameter λ

The parameter λ in Eq. (11) is used as a trade-off for the contributions of results estimated by conventional label propagation strategy and the proposed label-spatial reliability-based strategy. Here, we investigate the effect of the parameter λ, by varying its value in the range of [0.0, 0.1, …, 1.0]. In Fig. 11, we show the segmentation results achieved by our ls-LWV and ls-PBM methods with different values of λ on NIREP and LONI-LPBA40 dataset.

We can see from Fig. 11 that, on the NIREP and LONI-LPBA40 datasets, our proposed ls-LWV and ls-PBM methods achieve stable, promising results with $\lambda \in [0, 0.5]$. On the NIREP dataset, the best result of ls-LWV is achieved by using $\lambda = 0.2$, while the best result of ls-PBM is obtained with $\lambda = 0.5$. On the LONI-LPBA40 dataset, the best result of ls-LWV is achieved by using $\lambda = 0.4$, while the best result of ls-PBM is obtained with $\lambda = 0.1$. On the other hand, when $\lambda > 0.5$ (*i.e.*, with less contribution of the proposed label-spatial reliability-based label fusion), the performance of these two methods slightly drop with the increase of λ. These results imply that the proposed label-spatial reliability-based label fusion strategy plays an essential role in the final performance of both ls-LWV and ls-PBM.

5. Conclusion

In this paper, we proposed a novel label-spatial reliability-based robust label fusion framework for multi-atlas MRI segmentation. Specifically, we first perform initial segmentation using conventional label fusion methods for the target image. Then, for each voxel in the target image, we define the label reliability and spatial reliability-based on the soft label and the spatial structure from the initial segmentation, respectively. We then estimate the label-spatial reliability for each voxel in the target image. Finally, we employ voxels with high reliability to help refine the label fusion process of those voxels with low reliability. We validate the proposed framework in segmenting ROIs from brain MR images on the NIREP, LONI-LPBA40 and ADNI datasets, with results demonstrating that our label-spatial reliability-based label

fusion methods outperform several state-of-the-art methods in multi-atlas based brain MRI segmentation.

Although the proposed label-spatial reliability-based label fusion framework works well in segmenting ROIs for brain MRI in the NIREP, LONI-LPBA40 and ADNI datasets, there are still several limitations in the current work. *First*, compared with conventional multi-atlas segmentation methods (such as LWV and PBM, *etc.*), the proposed methods require two additional steps, including (1) estimating label-spatial reliability of each voxel in the target image, and (2) label-spatial reliability-based label fusion. These two additional procedures will increase the computational burden for label fusion. For instance, in the task of segmenting the region of *L insula gyrus* on the NIREP dataset, the proposed ls-PBM method requires about 247 s, while the conventional PBM method requires only 229 s. It is interesting to implement the proposed framework in a parallel manner, which will be our future work. *Besides*, we simply combine the label reliability and spatial reliability via Eq. (8), without considering the different contributions of these two types of reliability. As another future work, we will explicitly consider the contributions of the label and spatial reliability for further performance improvement.

References

[1] Liu M, Zhang D. Feature selection with effective distance. Neurocomputing 2016;215:100–9.
[2] Liu M, Zhang D. Pairwise constraint-guided sparse learning for feature selection. IEEE Trans Cybern 2016;46:298–310.
[3] Liu M, Zhang D, Chen S, Xue H. Joint binary classifier learning for ECOC-based multi-class classification. IEEE Trans Pattern Anal Mach Intell 2016;38:2335–41.
[4] Zu C, Jie B, Liu M, Chen S, Shen D, Zhang D. Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. Brain Imaging Behav 2016;10:1148–59.
[5] Jie B, Liu M, Zhang D, Shen D. Sub-network kernels for measuring similarity of brain connectivity networks in disease diagnosis. IEEE Trans Image Process 2018;27:2340–53.
[6] Jie B, Zhang D, Wee C, Shen D. Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification. Hum Brain Mapp 2014;35:2876–97.
[7] Zhang D, Huang J, Jie B, Du J, Tu L, Liu M. Ordinal pattern: a new descriptor for brain connectivity networks. IEEE Trans Med Imaging 2018;37:1711–22.
[8] Lian C, Ruan S, Denœux T, Li H, Vera P. Spatial evidential clustering with adaptive distance metric for tumor segmentation in FDG-PET images. IEEE Trans Biomed Eng 2018;65:21–30.
[9] Lian C, Ruan S, Denœux T, Li H, Vera P. Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions. IEEE Trans Image Process 2019;28:755–66.
[10] Liu M, Zhang D. Sparsity score: a novel graph-preserving feature selection method. Int J Pattern Recognit Artif Intell 2014;28:1450009.
[11] Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 2006;33:115–26.
[12] Artaechevarria X, Munozbarrutia A, Ortizdesolorzano C. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans Med Imaging 2009;28:1266.
[13] Coupe P, Manjon JV, Fonov V, Pruessner JC, Robles M, Collins DL. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 2011;54:940–54.
[14] Rousseau F, Habas PA, Studholme C. A supervised patch-based approach for human brain labeling. IEEE Trans Med Imaging 2011;30:1852–62.
[15] Tong T, Wolz R, Coupe P, Hajnal JV, Rueckert D. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage 2013;76.
[16] Zhang D, Guo Q, Wu G, Shen D. Sparse patch-based label fusion for multi-atlas segmentation. International workshop on multimodal brain image analysis. Springer; 2012. p. 94–102.
[17] Wu G, Kim M, Sanroma G, Wang Q, Munsell BC, Shen D. Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition. NeuroImage 2015;106:34.
[18] Wang H, Suh JW, Das SR, Pluta J, Craige C, Yushkevich PA. Multi-atlas segmentation with joint label fusion. IEEE Trans Pattern Anal Mach Intell 2013;35:611–23.
[19] Wu G, Wang Q, Zhang D, Nie F, Huang H, Shen D. A generative probability model of joint label fusion for multi-atlas based brain segmentation. Med Image Anal 2014;18:881.
[20] Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 2009;46:726–38.
[21] Bai W, Shi W, Oregan D, Tong T, Wang H, Jamilcopley S, et al. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. IEEE Trans Med Imaging 2013;32:1302–15.
[22] Langerak TR, Der Heide UAV, Kotte ANTJ, Viergever MA, Van Vulpen M, Pluim JPW. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans Med Imaging 2010;29:2000–8.
[23] Lotjonen J, Wolz R, Koikkalainen J, Thurfjell L, Waldemar G, Soininen H, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 2010;49:2352–65.
[24] Song Y, Wu G, Bahrami K, Sun Q, Shen D. Progressive multi-atlas label fusion by dictionary evolution. Med Image Anal 2017;36:162–71.
[25] Zu C, Wang Z, Zhang D, Liang P, Shi Y, Shen D, et al. Robust multi-atlas label propagation by deep sparse representation. Pattern Recognit 2017;63:511–7.
[26] Sanroma G, Benkarim OM, Piella G, Camara O, Wu G, Shen D, et al. Learning non-linear patch embeddings with neural networks for label fusion. Med Image Anal 2018;44:143–55.
[27] Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansenberg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 2004;23:208–19.
[28] Kim M, Wu G, Wang Q, Lee S, Shen D. Improved image registration by sparse patch-based deformation estimation. NeuroImage 2015;105:257–68.
[29] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. IEEE Trans Med Imaging 2013;32:1153–90.
[30] Li Z, Mahapatra D, Tielbeek JAW, Stoker J, Van Vliet LJ, Vos FM. Image registration based on autocorrelation of local structure. IEEE Trans Med Imaging 2016;35:63–75.
[31] Schlachter M, Fechter T, Jurisic M, Schimek-Jasch T, Oehlke O, Adebahr S, et al. Visualization of deformable image registration quality using local image dissimilarity. IEEE Trans Med Imaging 2016;35:2319–28.
[32] Yousefi S, Kehtarnavaz N, Gholipour A. Improved labeling of subcortical brain structures in atlas-based segmentation of magnetic resonance images. IEEE Trans Biomed Eng 2012;59:1808–17.
[33] Lian C, Ruan S, Denœux T. An evidential classifier based on feature selection and two-step classification strategy. Pattern Recognit 2015;48:2318–27.
[34] Christensen GE, Geng X, Kuhl JG, Bruss J, Grabowski TJ, Pirwani IA, et al. Introduction to the non-rigid image registration evaluation project (NIREP). IEEE Trans Magn 2006;30:2972–5.
[35] Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, et al. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 2008;39:1064.
[36] Shi F, Wang L, Dai Y, Gilmore JH, Lin W, Shen D. LABEL: pediatric brain extraction using learning-based meta-algorithm. NeuroImage 2012;62:1975–86.
[37] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20.
[38] Madabhushi A, Udupa JK. New methods of MR image intensity standardization via generalized scale. Med Phys 2006;33:3426–34.
[39] Vercauteren T, Pennec X, Perchant A, Ayache N. Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 2009;45:61–72.
[40] Wang H, Yushkevich P. Multi-atlas segmentation with joint label fusion and corrective learning – an open source implementation. Front Neuroinform 2013;7:27.
[41] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600–12.