Better Generalization of White Matter Tract Segmentation to Arbitrary Datasets with Scaled Residual Bootstrap

Wan Liu and Chuyang Ye(✉)

School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China
chuyang.ye@bit.edu.cn

Abstract. *White matter* (WM) tract segmentation is a crucial step for brain connectivity studies. It is performed on *diffusion magnetic resonance imaging* (dMRI), and *deep neural networks* (DNNs) have achieved promising segmentation accuracy. Existing DNN-based methods use an annotated dataset for model training. However, the performance of the trained model on a different test dataset may not be optimal due to distribution shift, and it is desirable to design WM tract segmentation approaches that allow better generalization of the segmentation model to arbitrary test datasets. In this work, we propose a WM tract segmentation approach that improves the generalization with scaled residual bootstrap. The difference between dMRI scans in training and test datasets is most noticeably caused by the different numbers of diffusion gradients and noise levels. Since both of them lead to different *signal-to-noise ratios* (SNRs) between the training and test data, we propose to augment the training scans by adjusting the noise magnitude and develop an adapted residual bootstrap strategy for the augmentation. First, with a dictionary-based linear representation of diffusion signals, we compute the signal residuals for the training dMRI scans, which can represent samples drawn from the noise distribution. Then, we adapt the bootstrap procedure by scaling the residuals that are randomly drawn with replacement and adding the scaled residuals to the linear signal representation, where augmented dMRI scans with different SNRs are generated. Finally, the augmented and original images are jointly included in model training. Since it is difficult to know the SNR of the test data *a priori*, we choose to perform the residual scaling with multiple factors. To validate the proposed approach, two dMRI datasets were used, and the experimental results show that our method consistently improved the generalization of WM tract segmentation under various settings.

Keywords: White matter tract segmentation · residual bootstrap · generalization

1 Introduction

White matter (WM) tract segmentation on *diffusion magnetic resonance imaging* (dMRI) provides a valuable quantitative tool for various brain stud-

A. Frangi et al. (Eds.): IPMI 2023, LNCS 13939, pp. 629–640, 2023.
https://doi.org/10.1007/978-3-031-34048-2_48

ies [1,7,21,24]. Manually delineated WM tracts are generally considered the gold standard segmentation, but the annotation process can be time-consuming and requires the expertise of experienced radiologists. Therefore, automated WM tract segmentation approaches are developed, which classify fiber streamlines [4,6] obtained with tractography [2,9] or directly provide voxelwise labeling results [3,19,26]. In particular, methods based on *deep neural networks* (DNNs) have substantially improved the accuracy of WM tract segmentation [15,25,27]. For example, Zhang et al. [27] group fiber streamlines into different WM tracts with a DNN that takes the spatial coordinates of the points along a fiber streamline as input; in [25], fiber orientation maps extracted from dMRI scans are fed into a U-net [20] to directly predict the existence of WM tracts at each voxel.

The DNN-based segmentation model is generally trained on a dataset where both dMRI scans and WM tract annotations are available. However, the performance of the model on an arbitrary test dataset that is different from the training dataset may be degraded due to distribution shift, where the use of different numbers of diffusion gradients and different noise levels are two major contributing factors [18]. Since dMRI scans can be acquired with various protocols, the improvement of the generalization of WM tract segmentation models to arbitrary test data becomes an important research topic. Although domain adaptation techniques [8] may be applied to improve the generalization, they require access to the test data during model training, which is not guaranteed when arbitrary test data is considered, and thus they are out of scope for this work. To account for the different numbers of diffusion gradients between training and test datasets, in [25] additional training scans are obtained by subsampling the diffusion gradients of the training data, and this allows improved segmentation accuracy on test data. However, the segmentation accuracy may still be improved by taking the *signal-to-noise ratio* (SNR) into consideration during model training.

In this work, we seek to further improve the generalization of WM tract segmentation from the perspective of SNR.[1] We focus on volumetric WM tract segmentation that directly obtains volumes of WM tract labels without requiring the tractography step. We assume that by producing diverse SNRs for training data, the training data can better represent the test data, and the trained model can better generalize to the test data. Therefore, we propose a scaled residual bootstrap strategy that augments the training scans with adjusted noise magnitude. First, we estimate a linear dictionary-based representation of diffusion signals and compute the residuals of the representation. These residuals are considered samples drawn from the noise distribution [10]. Then, for each diffusion gradient, the residual is drawn with replacement, and we adapt the standard residual bootstrap by scaling the residual. The scaled residuals are added to the linear representation of diffusion signals to generate augmented dMRI scans with different SNRs. Finally, the augmented images are used together with the original images for model training. Since it is difficult to know the SNR of the

[1] Note that the use of different numbers of diffusion gradients implicitly leads to different SNRs of measures derived from dMRI as well.

test data *a priori*, we choose to perform the residual scaling with multiple factors. The proposed approach was evaluated on two brain dMRI datasets, where various experimental settings of training and test scans were considered. The results show that our method consistently improved the generalization of WM tract segmentation under these various settings.

2 Methods

2.1 Problem Formulation

Suppose we are given a set of dMRI scans from a training dataset and the set of their annotations of WM tracts. We seek to train a WM tract segmentation model with good generalization, i.e., it performs well on an arbitrary test dataset. Like existing volumetric WM tract segmentation approaches [14,25], the model input is fiber orientation maps computed from dMRI. Two major factors that cause the difference between the training and test dMRI data are the use of different numbers of diffusion gradients and different noise levels. Since increasing/decreasing the number of diffusion gradients also leads to increased/decreased SNRs in the fiber orientation maps, respectively, we assume that adjusting the SNR of the dMRI scans for model training can effectively improve the generalization of the trained model to other datasets. Although existing approaches have considered SNR manipulation in the data augmentation operations of model training [25], it is applied to the network input of fiber orientation maps. As fiber orientations are orientations with unit lengths, adding realistic noise that is consistent with imaging physics to them is nontrivial. Therefore, we seek to further explore data augmentation with SNR adjustment in model training to improve the generalization of WM tract segmentation models.

2.2 Model Training with Scaled Residual Bootstrap

To produce training data with diverse SNRs and realistic noise distributions, we propose a scaled residual bootstrap strategy for model training. For convenience, we denote the diffusion weighted signals at each voxel of a training dMRI scan by a vector $\boldsymbol{y}$, where $\boldsymbol{y} \in \mathbb{R}^{N_\mathrm{d}}$ and N_d is the number of diffusion gradients. It has been shown that diffusion weighted signals can be linearly represented with a properly designed dictionary [16,17]:

$$\boldsymbol{y} = \mathbf{D}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{N_\mathrm{d} \times N_\mathrm{a}}$ is the dictionary with N_a dictionary atoms, $\boldsymbol{x} \in \mathbb{R}^{N_\mathrm{a}}$ is the vector of dictionary coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^{N_\mathrm{d}}$ represents the noise.

If the distribution of $\boldsymbol{\epsilon}$ is known, different levels of realistic noise can be added to the noise-free linear representation to provide training data with different SNRs. This motivates us to adopt a residual bootstrap strategy, which provides a feasible way of approximating the noise distribution. Then, by modifying the noise distribution, we achieve the goal of augmenting the SNR levels of training data. There are two major steps in the proposed method, which are 1) residual computation and 2) data generation with scaled residuals.

Residual Computation. Like the standard residual bootstrap, we first estimate $\boldsymbol{x}$ with the pseudoinverse of $\mathbf{D}$:

$$\hat{\boldsymbol{x}} = (\mathbf{D}^{\mathsf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\boldsymbol{y}, \tag{2}$$

where $\hat{\boldsymbol{x}}$ is the estimated coefficient vector. Then, the linear representation of the diffusion weighted signals can be estimated as

$$\hat{\boldsymbol{y}} = \mathbf{D}\hat{\boldsymbol{x}} = \mathbf{D}(\mathbf{D}^{\mathsf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\boldsymbol{y}. \tag{3}$$

The residuals $\hat{\boldsymbol{\epsilon}}$ of the signal representation can be simply computed by subtracting $\hat{\boldsymbol{y}}$ from $\boldsymbol{y}$

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{D}(\mathbf{D}^{\mathsf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}})\boldsymbol{y}. \tag{4}$$

Then, to ensure that the variances of the residuals $\hat{\boldsymbol{\epsilon}}$ are consistent with those of the noise $\boldsymbol{\epsilon}$, the residuals are corrected with the following normalization [5,10]:

$$\hat{\epsilon}'_i = \frac{\hat{\epsilon}_i}{\sqrt{1 - h_{ii}}}. \tag{5}$$

Here, $\hat{\epsilon}_i$ is the i-th entry of $\hat{\boldsymbol{\epsilon}}$, $\hat{\epsilon}'_i$ is the corresponding corrected residual, and h_{ii} is the i-th diagonal entry of $\mathbf{H} = \mathbf{D}(\mathbf{D}^{\mathsf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}$. The set $\mathcal{E} = \{\hat{\epsilon}'_i\}_{i=1}^{N_\mathrm{d}}$ of corrected residuals is then used in the bootstrap procedure that provides training data with diverse SNRs, and the procedure is described next.

Data Generation with Scaled Residuals. The corrected residuals $\mathcal{E}$ can be viewed as samples drawn from the noise distribution [5], and in the standard residual bootstrap, they are randomly drawn with replacement and added to the linear representation $\hat{\boldsymbol{y}}$. For our purpose of better generalization, we seek to generate samples with diverse SNRs. Therefore, the standard bootstrap procedure is modified with a scaling operation. Specifically, for the i-th diffusion gradient, we sample from $\mathcal{E}$ with replacement, and the sampled residual is denoted by $\tilde{\epsilon}_i$. The vector comprising the sampled residuals for all diffusion gradients is represented as $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{N_\mathrm{d}})$. Then, a bootstrap signal $\tilde{\boldsymbol{y}}$ is generated as

$$\tilde{\boldsymbol{y}} = \hat{\boldsymbol{y}} + r\tilde{\boldsymbol{\epsilon}}, \tag{6}$$

where r is the scaling factor that controls the magnitude of noise. r is selected from a predefined candidate set $\mathcal{R}$. By repeating the scaled residual bootstrap in Eq. (6) for each voxel, bootstrap diffusion weighted images can be generated.

Note that in dMRI acquisition, the $b0$ image without diffusion weighting is also acquired, and when more than one $b0$ images are available, their SNR can be adjusted as well. We denote the j-th $b0$ signal at each voxel by y_j^0, and the number of $b0$ images is denoted by N_0. Then, the residual $\hat{\epsilon}_j^0$ for the j-th $b0$ signal is calculated by

$$\hat{\epsilon}_j^0 = y_j^0 - \bar{y}^0, \tag{7}$$

where $\bar{y}^0 = \frac{1}{N_0}\sum_{j=1}^{N_0} y_j^0$ is the mean value of all $b0$ signals. These residuals form a set $\mathcal{E}^0$. For each j, a sample is drawn from $\mathcal{E}^0$ with replacement, which is denoted by $\tilde{\epsilon}_j^0$, and the bootstrap $b0$ signal is generated as

$$\tilde{y}_j^0 = \bar{y}^0 + r\tilde{\epsilon}_j^0. \tag{8}$$

Here, r has the same value as in Eq. (6). Equation (8) is repeated for each voxel to obtain bootstrap $b0$ images.

After bootstrap $b0$ images and diffusion weighted images are generated, they are combined to obtain new dMRI scans with different SNRs. These bootstrap dMRI scans are used to train the segmentation model together with the original dMRI scans based on the WM tract annotations.

2.3 Implementation Details

Our method is agnostic to the architecture of the segmentation model. For demonstration, the state-of-the-art TractSeg architecture [25] is used as the backbone network, but other network structures [13,14] may also be applied. As in [25], we extract fiber orientation maps from dMRI scans with *constrained spherical deconvolution* (CSD) [22] (for single-shell dMRI data) or *multi-shell multi-tissue CSD* (MSMT-CSD) [11] (for multi-shell dMRI data), and use these maps as network input. At most three fiber orientations are allowed, and all WM tracts are jointly segmented [25].

We use the SHORE basis[2] [17] for the linear representation of diffusion signals, which is a common choice. To generate bootstrap training data with diverse SNRs, the set of candidate scaling factors is $\mathcal{R} = \{2, 3, 4\}$. Since it is difficult to predetermine the SNR of arbitrary test data, all values in $\mathcal{R}$ are used for bootstrap, and each value is used once for each training scan.

For model training, following [25], we use the binary cross entropy loss function, which is minimized by Adamax [12] with a batch size of 56 and 300 training epochs; the initial learning rate is set to 0.001. We select the model that has the best segmentation accuracy on a validation dataset. Traditional data augmentation implemented online in TractSeg, such as intensity perturbation and spatial transformation, is also applied online in the proposed method.

3 Results

3.1 Datasets and Experimental Settings

We used two dMRI datasets to evaluate our method. The first one is the publicly available *Human Connectome Project* (HCP) dataset [23], and the second one is an in-house dMRI dataset. A detailed description of the two datasets and their experimental settings is given below.

[2] The default setting given in https://dipy.org/documentation/1.4.1./reference/dipy.reconst/#dipy.reconst.shore.ShoreModel is used.

The HCP Dataset. The dMRI scans in the HCP dataset were acquired with 270 diffusion gradients ($b = 1000$, 2000, and 3000 s/mm^2) and an isotropic image resolution of 1.25 mm. 18 $b0$ images were also acquired for each dMRI scan. 72 WM tracts were manually delineated for the HCP dataset[3]. We used 100 scans in our experiments, where 55 and 15 scans were used as the training set and validation set, respectively, and the remaining 30 scans were used for testing. To improve the generalization of the segmentation model to different imaging protocols, in TractSeg [25], subsampling of diffusion gradients was performed on the original training dMRI scans, where dMRI scans with 12 and 90 diffusion gradients associated with $b = 1000$ s/mm^2 were generated for model training together with the original dMRI scans.[4] Here, we followed [25] and performed the subsampling as well for the original and bootstrap training data for model training. For convenience, the original HCP dataset is referred to as HCP_1.25mm_270, and the subsampled datasets with 12 and 90 diffusion gradients are referred to as HCP_1.25mm_12 and HCP_1.25mm_90, respectively.

To evaluate the performance of the proposed method on test scans that were acquired with different protocols, we generated additional test sets from the 30 original test scans. First, like the training data in HCP_1.25mm_12 and HCP_1.25mm_90, the 12 and 90 diffusion gradients associated with $b = 1000$ s/mm^2 were selected from the 30 test scans, respectively. Second, 34 diffusion gradients associated with $b = 1000$ s/mm^2 were selected for the test scans, so that their imaging protocol was different from the original and subsampled training data, and the images associated with this subsampling are referred to as HCP_1.25mm_34. Only three $b0$ images were kept for HCP_1.25mm_34. Finally, another test set HCP_1.25mm_36 was generated from the test scans by selecting 18 diffusion gradients associated with $b = 1000$ s/mm^2 and 18 diffusion gradients associated with $b = 2000$ s/mm^2, which also produced dMRI scans that used a different imaging protocol than the training data. Only one $b0$ image was kept for HCP_1.25mm_36. A summary of these different datasets is listed in Table 1.

In addition, to investigate the impact of the amount of training data on the segmentation, three other experimental settings were considered, where 10, 20, or 30 training subjects were used and the other settings were not changed.

The In-House Dataset. The segmentation models trained on the HCP dataset were also applied to an in-house dataset for further evaluation. The dMRI scans in the in-house dataset were acquired with 270 diffusion gradients ($b = 1000$, 2000, and 3000 s/mm^2) and one $b0$ image. The spatial resolution is 1.7 mm isotropic. These scans were acquired on a scanner that is different from that of the HCP dataset. Due to the annotation cost, only ten of the 72 annotated WM tracts of the HCP dataset were manually delineated, and the delineation was performed on 17 in-house dMRI scans. These annotations were used only to evaluate the segmentation accuracy. This in-house dataset is referred to

[3] The annotations can be downloaded at https://doi.org/10.5281/zenodo.1088277.

[4] All $b0$ images were kept for these two cases.

Table 1. A summary of the datasets used in the experiments

Dataset	Resolution	Diffusion gradients	Usage
HCP_1.25mm_270	1.25 mm	$90 \times b = 1000\,s/mm^2$ $90 \times b = 2000\,s/mm^2$ $90 \times b = 3000\,s/mm^2$ $18 \times b = 0\,s/mm^2$	Training & Test
HCP_1.25mm_12	1.25 mm	$12 \times b = 1000\,s/mm^2$ $18 \times b = 0\,s/mm^2$	Training & Test
HCP_1.25mm_90	1.25 mm	$90 \times b = 1000\,s/mm^2$ $18 \times b = 0\,s/mm^2$	Training & Test
HCP_1.25mm_34	1.25 mm	$34 \times b = 1000\,s/mm^2$ $3 \times b = 0\,s/mm^2$	Test
HCP_1.25mm_36	1.25 mm	$18 \times b = 1000\,s/mm^2$ $18 \times b = 2000\,s/mm^2$ $1 \times b = 0\,s/mm^2$	Test
IH_1.7mm_270	1.7 mm	$90 \times b = 1000\,s/mm^2$ $90 \times b = 2000\,s/mm^2$ $90 \times b = 3000\,s/mm^2$ $1 \times b = 0\,s/mm^2$	Test
IH_1.7mm_36	1.7 mm	$18 \times b = 1000\,s/mm^2$ $18 \times b = 2000\,s/mm^2$ $1 \times b = 0\,s/mm^2$	Test

as IH_1.7mm_270. We also synthesized another dataset IH_1.7mm_36 from IH_1.7mm_270 for evaluation, where 18 diffusion gradients of $b = 1000\,s/mm^2$ and 18 diffusion gradients of $b = 2000\,s/mm^2$ were selected from the original scans. These two datasets are also summarized in Table 1.

3.2 Evaluation of Segmentation Results on the HCP Dataset

We first present the evaluation of the segmentation results on the HCP dataset. Our method was compared with TractSeg without using bootstrap (but with the subsampling of diffusion gradients), which is referred to as the baseline method.

Examples of the segmentation results are shown in Fig. 1. For demonstration, here we show the results of representative WM tracts on HCP_1.25mm_90, HCP_1.25mm_36, and HCP_1.25mm_34 when 55 training subjects were used. For reference, the gold standard (manual delineation) is also displayed. In Fig. 1, cross-sectional views of the WM tracts are given, and regions are highlighted with zoomed views for better comparison. It can be seen that the segmented tracts of the proposed method have more similar spatial coverage to the gold standard than the baseline method.

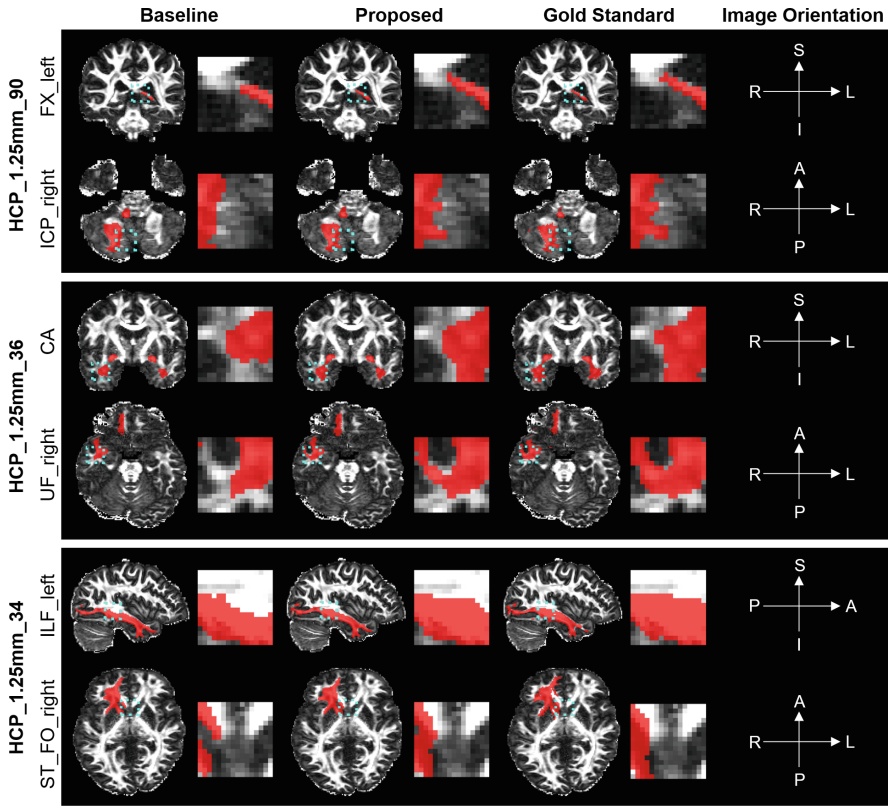

Fig. 1. Representative segmentation results (red) for the HCP dataset, together with the gold standard (manual annotation) for reference. The cross-sectional views of the segmented tracts are shown, and they are overlaid on fractional anisotropy maps. Zoomed views of the highlighted regions are also displayed for better comparison. The image orientation is shown in the rightmost column. For the meaning of the tract abbreviations, we refer readers to [25]. (Color figure online)

We then quantitatively evaluated the proposed method by computing the Dice coefficient between the segmentation results and the gold standard. The mean Dice coefficient of all 72 WM tracts for each test dataset and each number of training subjects is shown in Table 2. As some WM tracts can be more challenging to segment [14] and the improvement of the segmentation of these tracts is important, in Table 2 we also show the individual average Dice coefficients of the three most challenging WM tracts, which are the anterior commissure (CA), left fornix (FX_left), and right fornix (FX_right) [14,25]. Compared with the baseline method, the proposed method can consistently improve the Dice coefficients across the different cases, and the improvement is more prominent for the three most challenging WM tracts. In addition, the Dice coefficients of the proposed method were compared with those of the baseline method using paired Student's t-tests, and the p-values are listed in Table 2. It can be seen that the improvement of the proposed method is statistically significant in all cases.

Table 2. The mean Dice coefficient (%) of all 72 WM tracts and the individual average Dice coefficients (%) of the three most challenging tracts for the HCP dataset across different settings. The proposed method was compared with the baseline method using paired Student's t-tests, and asterisks indicate that the difference between the two methods is statistically significant (***: $p < 0.001$).

Dataset	Tract	Method	Number of training subjects							
			10		20		30		55	
HCP_1.25mm_270	All	Baseline	80.0	***	81.8	***	82.3	***	83.4	***
		Proposed	80.9		83.1		83.5		84.0	
	CA	Baseline	52.4	***	61.6	***	63.7	***	67.3	***
		Proposed	56.6		65.7		68.0		69.4	
	FX_left	Baseline	55.8	***	67.6	***	68.1	***	70.7	***
		Proposed	65.0		73.6		73.9		73.7	
	FX_right	Baseline	51.1	***	59.5	***	61.4	***	64.9	***
		Proposed	55.3		68.0		68.3		69.5	
HCP_1.25mm_90	All	Baseline	79.0	***	80.9	***	81.4	***	82.9	***
		Proposed	80.2		82.8		83.2		83.7	
	CA	Baseline	51.0	***	61.1	***	63.6	***	66.3	***
		Proposed	56.9		65.1		67.4		68.7	
	FX_left	Baseline	55.6	***	63.5	***	65.9	***	68.9	***
		Proposed	62.8		72.2		72.6		72.6	
	FX_right	Baseline	47.8	***	57.9	***	57.1	***	63.6	***
		Proposed	56.4		67.5		67.1		68.1	
HCP_1.25mm_12	All	Baseline	77.9	***	80.2	***	80.8	***	82.2	***
		Proposed	79.7		82.4		82.8		83.4	
	CA	Baseline	47.7	***	59.2	***	61.9	***	64.8	***
		Proposed	56.2		64.1		66.8		68.0	
	FX_left	Baseline	50.6	***	61.8	***	62.9	***	65.5	***
		Proposed	58.5		71.6		71.9		72.3	
	FX_right	Baseline	40.6	***	54.2	***	53.3	***	59.6	***
		Proposed	51.4		66.1		65.6		66.5	
HCP_1.25mm_36	All	Baseline	79.2	***	80.9	***	81.5	***	82.7	***
		Proposed	80.6		82.9		83.3		83.9	
	CA	Baseline	48.6	***	60.0	***	61.6	***	65.4	***
		Proposed	55.2		65.2		67.4		68.7	
	FX_left	Baseline	53.9	***	66.3	***	65.9	***	68.4	***
		Proposed	63.5		72.9		73.2		73.5	
	FX_right	Baseline	46.3	***	56.7	***	58.8	***	62.6	***
		Proposed	52.3		66.9		67.3		68.5	
HCP_1.25mm_34	All	Baseline	78.4	***	80.6	***	81.1	***	82.6	***
		Proposed	80.1		82.7		83.1		83.6	
	CA	Baseline	49.3	***	60.3	***	63.3	***	65.9	***
		Proposed	56.6		65.3		67.3		68.8	
	FX_left	Baseline	51.6	***	61.5	***	64.3	***	67.2	***
		Proposed	61.1		72.1		72.0		72.2	
	FX_right	Baseline	43.9	***	55.3	***	54.6	***	62.0	***
		Proposed	54.0		66.5		66.2		67.5	

Table 3. The mean Dice coefficient (%) of all ten annotated WM tracts and the individual average Dice coefficients (%) of two challenging tracts for the in-house dataset across different settings. The proposed method was compared with the baseline method using paired Student's t-tests, and asterisks indicate that the difference between the two methods is statistically significant (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, n.s.: $p \geq 0.05$).

Dataset	Tract	Method	Number of training subjects							
			10		20		30		55	
IH_1.7mm_270	All	Baseline	58.7	n.s.	60.4	**	61.2	**	61.7	n.s.
		Proposed	59.0		62.0		61.9		61.9	
	UF_left	Baseline	48.4	***	50.7	***	53.1	**	55.1	n.s.
		Proposed	51.9		59.3		56.7		57.1	
	UF_right	Baseline	52.9	n.s.	56.5	***	57.3	***	59.1	***
		Proposed	53.4		61.3		60.1		61.4	
IH_1.7mm_36	All	Baseline	57.2	**	58.7	***	59.3	***	60.4	**
		Proposed	57.8		61.5		61.5		61.3	
	UF_left	Baseline	46.2	n.s.	46.3	***	47.3	***	51.6	**
		Proposed	47.9		58.0		55.7		55.3	
	UF_right	Baseline	48.9	***	54.2	***	53.6	***	56.3	***
		Proposed	51.2		59.1		58.5		60.5	

By comparing the results achieved with different numbers of training subjects, we observe that the overall improvement of the proposed method tends to be greater when the number is moderate (20 and 30) than when the number is small (10) or large (55). Moreover, the Dice coefficients of the proposed method obtained with 20 training subjects are comparable to or higher than the baseline performance achieved with 55 training subjects. Also, when the number of training subjects increases from 20 to 30 or 55, the Dice coefficients of the proposed method are relatively stable, whereas the Dice coefficients of the baseline method can still increase. This is possibly because the proposed method augments the training data and thus reduces the requirement for manual annotation.

3.3 Evaluation of Segmentation Results on the In-House Dataset

The proposed method was next applied to the in-house test datasets IH_1.7mm_270 and IH_1.7mm_36, and the mean Dice coefficients of all ten annotated WM tracts are summarized in Table 3. In addition, the individual average Dice coefficients of two challenging tracts, the left uncinate fasciculus (UF_left) and right uncinate fasciculus (UF_right) [14], are also shown in Table 3.[5] In each case, the proposed method achieves a higher Dice coefficient than the baseline method, and the improvement is more prominent for the two

[5] CA, FX_left, and FX_right were not annotated for the in-house dataset.

challenging tracts and for IH_1.7mm_36 that has a smaller number of diffusion gradients. We also performed paired Student's t-tests to compare the two methods in Table 3, and the difference between the proposed and competing methods is statistically significant in most cases.

Like the results on the HCP dataset, the improvement of the proposed method over the baseline method is greater when the number of training subjects is 20 or 30 than 10 or 55, and its performance becomes stable after the number of training subjects reaches 20. Also, the Dice coefficients of the proposed method obtained with 20 training subjects are already better than the baseline performance achieved with 55 training subjects.

4 Conclusion

We have proposed a WM tract segmentation approach that better generalizes to arbitrary test datasets. In the proposed method a scaled residual bootstrap strategy is developed, where the SNR levels of the training data are adjusted based on the residuals of a linear signal representation. This reduces the discrepancy between training and test data and thus improves the generalization of the trained segmentation model. Our method was validated on public and in-house datasets under various data settings, and the results show that it consistently improved the segmentation accuracy in the different cases.

Acknowledgements. This work is supported by the Fundamental Research Funds for the Central Universities.

References

1. Banihashemi, L., et al.: Opposing relationships of childhood threat and deprivation with stria terminalis white matter. Hum. Brain Mapp. **42**(8), 2445–2460 (2021)
2. Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A.: In vivo fiber tractography using DT-MRI data. Magn. Reson. Med. **44**(4), 625–632 (2000)
3. Bazin, P.L., et al.: Direct segmentation of the major white matter tracts in diffusion tensor images. Neuroimage **58**(2), 458–468 (2011)
4. Cook, P.A., et al.: An automated approach to connectivity-based partitioning of brain structures. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 164–171. Springer, Heidelberg (2005). https://doi.org/10.1007/11566465_21
5. Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. No. 1, Cambridge University Press (1997)
6. Garyfallidis, E., et al.: Recognition of white matter bundles using local and global streamline-based registration and clustering. Neuroimage **170**, 283–295 (2018)
7. Girard, G., et al.: On the cortical connectivity in the macaque brain: a comparison of diffusion tractography and histological tracing data. Neuroimage **221**, 117201 (2020)
8. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
9. Jeurissen, B., Descoteaux, M., Mori, S., Leemans, A.: Diffusion MRI fiber tractography of the brain. NMR Biomed. **32**(4), e3785 (2019)

10. Jeurissen, B., Leemans, A., Jones, D.K., Tournier, J.D., Sijbers, J.: Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. Hum. Brain Mapp. **32**(3), 461–479 (2011)
11. Jeurissen, B., Tournier, J.D., Dhollander, T., Connelly, A., Sijbers, J.: Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. Neuroimage **103**, 411–426 (2014)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Li, B., et al.: Neuro4Neuro: A neural network approach for neural tract segmentation using large-scale population-based diffusion imaging. Neuroimage **218**, 116993 (2020)
14. Liu, W., et al.: Volumetric segmentation of white matter tracts with label embedding. Neuroimage **250**, 118934 (2022)
15. Lu, Q., Li, Y., Ye, C.: Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. Med. Image Anal. **72**, 102094 (2021)
16. Merlet, S., Caruyer, E., Deriche, R.: Parametric dictionary learning for modeling EAP and ODF in diffusion MRI. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 10–17. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33454-2_2
17. Merlet, S.L., Deriche, R.: Continuous diffusion signal, EAP and ODF estimation via compressive sensing in diffusion MRI. Med. Image Anal. **17**(5), 556–572 (2013)
18. Ning, L., et al.: Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: algorithms and results. Neuroimage **221**, 117128 (2020)
19. Ratnarajah, N., Qiu, A.: Multi-label segmentation of white matter structures: application to neonatal brains. Neuroimage **102**, 913–922 (2014)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Toescu, S.M., Hales, P.W., Kaden, E., Lacerda, L.M., Aquilina, K., Clark, C.A.: Tractographic and microstructural analysis of the dentato-rubro-thalamo-cortical tracts in children using diffusion MRI. Cereb. Cortex **31**(5), 2595–2609 (2021)
22. Tournier, J.D., Calamante, F., Connelly, A.: Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. Neuroimage **35**(4), 1459–1472 (2007)
23. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K.: Wu-Minn HCP consortium: the WU-Minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
24. Veraart, J., Raven, E.P., Edwards, L.J., Weiskopf, N., Jones, D.K.: The variability of MR axon radii estimates in the human white matter. Hum. Brain Mapp. **42**(7), 2201–2213 (2021)
25. Wasserthal, J., Neher, P., Maier-Hein, K.H.: TractSeg - fast and accurate white matter tract segmentation. Neuroimage **183**, 239–253 (2018)
26. Ye, C., Yang, Z., Ying, S.H., Prince, J.L.: Segmentation of the cerebellar peduncles using a random forest classifier and a multi-object geometric deformable model: Application to spinocerebellar ataxia type 6. Neuroinformatics **13**(3), 367–381 (2015)
27. Zhang, F., Karayumak, S.C., Hoffmann, N., Rathi, Y., Golby, A.J., O'Donnell, L.J.: Deep white matter analysis (DeepWMA): fast and consistent tractography segmentation. Med. Image Anal. **65**, 101761 (2020)