Medical Image Analysis 79 (2022) 102454

Contents lists available at ScienceDirect

Medical Image Analysis

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

A transfer learning approach to few-shot segmentation of novel white matter tracts

Qi Lu[a,1], Wan Liu[a,1], Zhizheng Zhuo[b,1], Yuxing Li[a], Yunyun Duan[b], Pinnan Yu[b], Liying Qu[b], Chuyang Ye[a,*], Yaou Liu[b,*]

[a] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China
[b] Department of Radiology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

ARTICLE INFO

Article history:
Received 6 June 2021
Revised 19 March 2022
Accepted 8 April 2022
Available online 12 April 2022

Keywords:
White matter tract segmentation
Novel white matter tract
Few-shot segmentation
Transfer learning

ABSTRACT

Convolutional neural networks (CNNs) have achieved state-of-the-art performance for *white matter* (WM) tract segmentation based on *diffusion magnetic resonance imaging* (dMRI). The training of the CNN-based segmentation model generally requires a large number of manual delineations of WM tracts, which can be expensive and time-consuming. Although it is possible to carefully curate abundant training data for a set of WM tracts of interest, there can also be novel WM tracts—i.e., WM tracts that are not included in the existing annotated WM tracts—that are specific to a new scientific problem, and it is desired that the novel WM tracts can be segmented without repeating the laborious collection of a large number of manual delineations for these tracts. One possible solution to the problem is to transfer the knowledge learned for segmenting existing WM tracts to the segmentation of novel WM tracts with a fine-tuning strategy, where a CNN pretrained for segmenting existing WM tracts is fine-tuned with only a few annotated scans to segment the novel WM tracts. However, in classic fine-tuning, the information in the last task-specific layer for segmenting existing WM tracts is completely discarded. In this work, based on the pretraining and fine-tuning framework, we propose an improved transfer learning approach to the segmentation of novel WM tracts in the few-shot setting, where all knowledge in the pretrained model is incorporated into the fine-tuning procedure. Specifically, from the weights of the pretrained task-specific layer for segmenting existing WM tracts, we derive a better initialization of the last task-specific layer for the target model that segments novel WM tracts. In addition, to allow further improvement of the initialization of the last layer and thus the segmentation performance in the few-shot setting, we develop a simple yet effective data augmentation strategy that generates synthetic annotated images with tract-aware image mixing. To validate the proposed method, we performed experiments on brain dMRI scans from public and private datasets under various experimental settings, and the results indicate that our method improves the performance of few-shot segmentation of novel WM tracts.

1. Introduction

White matter (WM) tract segmentation based on *diffusion magnetic resonance imaging* (dMRI) identifies anatomical WM pathways that connect brain regions of interest (Yeatman et al., 2012; Zhang et al., 2020; Chandio et al., 2020). It provides a useful quantitative tool for the analysis of brain characteristics (O'Donnell and Pasternak, 2015; Mueller et al., 2015; Vanderweyen et al., 2020) and facilitates the studies on brain development, function, and disease. For example, in Jaimes et al. (2020), the development of specific WM tracts in the fetal brain is found to be associated with known cellular processes that occur during pregnancy, and the tissue microstructure of these WM tracts allows the discrimination of normal and abnormal development with high anatomical specificity. In Hula et al. (2020), arcuate and middle longitudinal fasciculi are segmented and analyzed, and the results challenge the well-known dual-stream model of language function. In Toescu et al. (2021), it is observed that damage to the dentato-rubro-thalamo-cortical tract is implicated in cerebellar mutism syndrome, and the novel insights can offer important information for the surgical resection of cerebellar tumors.

WM tracts can be reconstructed from dMRI using the technique of fiber tracking or tractography (Basser et al., 2000; Poulin et al., 2019), where the tracts are represented as 3D fiber stream-

[*] Corresponding authors.
E-mail addresses: chuyang.ye@bit.edu.cn (C. Ye), liuyaou@bjtth.org (Y. Liu).
[1] Co-first authors.

https://doi.org/10.1016/j.media.2022.102454
1361-8415/© 2022 Elsevier B.V. All rights reserved.

lines. Specific WM tracts can be segmented manually by selecting the fiber streamlines according to the knowledge of experts and grouping the selected streamlines into fiber bundles (Stieltjes et al., 2013; Thiebaut de Schotten et al., 2011). However, the manual selection is laborious and subjective, and the reproducibility of manual WM tract segmentation can be poor. Therefore, it is highly desired to develop automated WM tract segmentation approaches for neuroimaging studies.

Previous works have automatically classified the fiber streamlines into anatomically defined WM tracts according to the *regions of interest* (ROIs) through which the streamlines pass (Cook et al., 2005; Wassermann et al., 2016) or reference streamlines defined in streamline atlases (O'Donnell and Westin, 2007; Garyfallidis et al., 2018; Wu et al., 2020). An alternative automated strategy is to directly assign tract labels to the voxels in a dMRI scan based on the anatomical prior knowledge and diffusion features. Since voxels are directly labeled without necessarily performing fiber tracking, we refer to this type of WM tract segmentation approaches as volumetric WM tract segmentation (Lu et al., 2021). For example, a volumetric tract atlas can be registered nonlinearly to test scans using maps of diffusion features to obtain the segmentation results (Oishi et al., 2009). More advanced approaches use machine learning techniques, such as Markov random fields, random forests, or *k*-nearest neighbors, to label the voxels (Bazin et al., 2011; Ye et al., 2015; Ratnarajah and Qiu, 2014).

Recently, *convolutional neural networks* (CNNs) have been successfully applied to WM tract segmentation with remarkably improved performance (Zhang et al., 2020; Wasserthal et al., 2018; Li et al., 2020). Like earlier WM tract segmentation approaches, these CNN-based methods either classify the fiber streamlines into fiber bundles or directly label the voxels according to the WM tracts to which they belong. For example, Zhang et al. (2020) have designed a feature map named FiberMap for each fiber streamline, and a CNN is trained based on FiberMap to perform fiber streamline classification. Wasserthal et al. (2018) use the strategy of volumetric WM tract segmentation, where a CNN named TractSeg is trained to predict the tract labels of each voxel from input fiber orientation maps. A similar volumetric strategy is proposed in Li et al. (2020), and a 3D CNN named Neuro4Neuro is used to label the voxels directly with the input of diffusion tensor images.

As in other image processing tasks, the success of CNNs in WM tract segmentation relies on abundant annotated training scans. However, manually delineating WM tracts on a large number of dMRI scans can be very time-consuming and costly. Although the training data can be carefully curated and accumulated throughout time for a set of WM tracts of interest for a study (Wasserthal et al., 2018), there can still be novel WM tracts—i.e., WM tracts that are not included in the existing annotated WM tracts—that are to be analyzed in a new scientific problem (Toescu et al., 2021; MacNiven et al., 2020; Banihashemi et al., 2021). Repeating the labor-intensive collection of tract annotations on a large number of scans for each new study involving novel WM tracts can be prohibitive, and it is desired that the segmentation of novel WM tracts can be accurately performed with only a few cases of annotations for the novel WM tracts.

One possible solution to the problem described above is to transfer the knowledge learned with the abundant annotated data collected previously for segmenting existing WM tracts to the segmentation of novel WM tracts. In this case, even with only a few manual delineations of novel WM tracts, the CNN can learn adequate knowledge for the segmentation of novel WM tracts. Intuitively, a fine-tuning strategy can be used for the purpose, where a CNN is pretrained to segment the existing WM tracts and it is then fine-tuned to obtain the target model that segments the novel WM tracts. In classic fine-tuning (Tajbakhsh et al., 2016), the weights in the feature extraction layers of the pretrained CNN are copied for initializing the corresponding layers of the target model, whereas the last task-specific layer of the target model is randomly initialized. Then, all weights of the target model are jointly learned with the limited amount of training data that is available for the target task. It is also possible to transfer the knowledge from a source task without a pretrained model (Roy et al., 2020; Feng et al., 2021). However, this type of method requires the access to the training data collected for the source task, which may sometimes be impractical due to privacy concerns in medical imaging (Burton et al., 2015), whereas the access to pretrained models is less of a concern. Therefore, in this work we focus on the use of the pretraining and fine-tuning framework for the few-shot segmentation of novel WM tracts.

Although the classic fine-tuning strategy can be used for the few-shot segmentation of novel WM tracts, it completely discards the information in the last task-specific layer of the pretrained model for segmenting the existing WM tracts. Since different WM tracts can be correlated due to tract crossing or overlapping, the discarded layer may also bear valuable information that is relevant to the novel WM tracts, and thus classic fine-tuning may be suboptimal. Therefore, in this work, we further explore the knowledge transfer from the segmentation of existing WM tracts and propose a transfer learning approach to the segmentation of novel WM tracts in the few-shot setting. As described above, we focus on the scenario where the model pretrained for segmenting existing WM tracts is available, but the access to the training data collected for existing WM tracts is not guaranteed. In addition, we focus on volumetric WM tract segmentation because it does not require performing tractography, the result of which can be sensitive to the choice of tracking algorithms and hyperparameters (Zhang et al., 2021).

Specifically, we assume that the novel WM tracts can be predicted with the logits—the unnormalized predictions before the final activation function—of existing WM tracts. For simplicity, we formulate the prediction as a logistic regression problem, and based on this formulation, we derive a better initialization of the last task-specific layer for segmenting the novel WM tracts using the information in the last task-specific layer pretrained for existing WM tracts. For the feature extraction layers, like in classic fine-tuning the knowledge transfer is still performed by copying the weights pretrained for existing WM tracts for initialization. In this way, all knowledge learned for segmenting existing WM tracts can be transferred to the segmentation of novel WM tracts. Moreover, we show that the derivation of the better initialization of the last layer motivates a more adaptive initialization strategy, which can be simply achieved by inserting a warmup stage before classic fine-tuning. In addition to the improved knowledge transfer strategy, to allow further improvement of the initialization of the last layer and thus the segmentation performance in the few-shot setting, we propose to better exploit the information in the scarce annotations of novel WM tracts and develop a simple yet effective data augmentation strategy TractMix. In TractMix, synthetic annotated images are generated from the real annotated scans with tract-aware image mixing, and the mixing can also vary with different combinations of novel WM tracts to increase the diversity of the synthetic annotated data. These synthetic images are used together with the real annotated images in network training to further improve the initialization of the last layer for segmenting the novel WM tracts.

To validate the proposed method, we performed experiments on the publicly available *Human Connectome Project* (HCP) dataset (Van Essen et al., 2013) and a private dataset comprising both *healthy control* (HC) subjects and patients with *Alzheimer's disease* (AD). The segmentation performance was evaluated under various experimental settings, and the results show that our method improves the quality of the segmentation of novel WM tracts given only a few annotated training images.

A preliminary version of this work has been presented at IPMI 2021 (Lu and Ye, 2021). Compared with the conference version, in the current manuscript we have described the proposed methodology with more details and have further developed TractMix that generates additional synthetic training data to improve the segmentation performance. In addition, we have performed a more comprehensive evaluation of the proposed method, where additional experimental settings and an additional dataset have been considered. In particular, using the additional private dataset, we show that the proposed method can be applied when domain shift (Ganin and Lempitsky, 2015) exists between the data used for segmenting existing and novel WM tracts and it is applicable to both HC subjects and AD patients.

The remaining of the paper is organized as follows. Section 2 describes the proposed approach to the segmentation of novel WM tracts in the few-shot setting. Section 3 presents the results on the public and private datasets under various experimental settings. In Section 4, we discuss the results and future works. Finally, Section 5 summarizes the proposed work.

2. Methods

In this section, we first formulate the problem of segmenting novel WM tracts and introduce classic fine-tuning. Then, we present the proposed transfer learning approach to few-shot segmentation of novel WM tracts, as well as how this approach motivates a better implementation. In addition, we describe how to further benefit the transfer learning by better exploiting the few annotated scans with data augmentation achieved by tract-aware image mixing. Finally, we introduce the backbone CNN for WM tract segmentation and describe the implementation details.

2.1. Problem formulation and classic fine-tuning

Suppose we are given a CNN-based segmentation model pretrained with abundant annotations for segmenting a set of WM tracts, which, for convenience, are referred to as existing WM tracts. We are interested in the segmentation of a novel set of WM tracts that are not included in the training set of the given model.[2] Only a few annotations are available for these novel WM tracts because delineations of WM tracts are generally labor-intensive. Our goal is to achieve decent segmentation accuracy for the novel WM tracts with the scarce annotations. To achieve such a goal, a common practice is to transfer the knowledge learned for segmenting existing WM tracts in the given model to the segmentation of novel WM tracts. Typically, a classic fine-tuning strategy (Tajbakhsh et al., 2016) can be used to perform the knowledge transfer, and its mathematical formulation is given below.

We denote the network models for segmenting existing and novel WM tracts by $\mathcal{M}_\mathrm{e}$ and $\mathcal{M}_\mathrm{n}$, respectively. In classic fine-tuning, $\mathcal{M}_\mathrm{e}$ and $\mathcal{M}_\mathrm{n}$ share the same network structure except for the last task-specific layer. We denote the task-specific weights in the last layer L_e of $\mathcal{M}_\mathrm{e}$ and the last layer L_n of $\mathcal{M}_\mathrm{n}$ by $\boldsymbol{\theta}_\mathrm{e}$ and $\boldsymbol{\theta}_\mathrm{n}$, respectively, and the other weights in $\mathcal{M}_\mathrm{e}$ or $\mathcal{M}_\mathrm{n}$ are denoted by $\boldsymbol{\theta}$. Suppose the input image is $\mathbf{X}$; from $\mathbf{X}$ a multi-channel feature map $\mathbf{F}$ is computed with a mapping $f(\mathbf{X};\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$:

$$\mathbf{F} = f(\mathbf{X};\boldsymbol{\theta}), \tag{1}$$

and the segmentation probability map $\mathbf{P}_\mathrm{e}$ or $\mathbf{P}_\mathrm{n}$ for existing or novel WM tracts is computed from $\mathbf{F}$ with L_e or L_n using another mapping $g_\mathrm{e}(\mathbf{F};\boldsymbol{\theta}_\mathrm{e})$ or $g_\mathrm{n}(\mathbf{F};\boldsymbol{\theta}_\mathrm{n})$ parameterized by $\boldsymbol{\theta}_\mathrm{e}$ or $\boldsymbol{\theta}_\mathrm{n}$, respectively:

$$\begin{aligned}\mathbf{P}_\mathrm{e} &= g_\mathrm{e}(\mathbf{F};\boldsymbol{\theta}_\mathrm{e}) = g_\mathrm{e}(f(\mathbf{X};\boldsymbol{\theta});\boldsymbol{\theta}_\mathrm{e}) \text{ and } \mathbf{P}_\mathrm{n} = g_\mathrm{n}(\mathbf{F};\boldsymbol{\theta}_\mathrm{n})\\ &= g_\mathrm{n}(f(\mathbf{X};\boldsymbol{\theta});\boldsymbol{\theta}_\mathrm{n}).\end{aligned} \tag{2}$$

In classic fine-tuning, instead of directly training $\mathcal{M}_\mathrm{n}$ from scratch—i.e., $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\mathrm{n}$ are randomly initialized—using the scarce annotations of novel WM tracts, the information in $\mathcal{M}_\mathrm{e}$ can be exploited. Since $\mathcal{M}_\mathrm{e}$ is trained by minimizing the difference between $\mathbf{P}_\mathrm{e}$ and the abundant annotations of existing WM tracts, the learned values $\tilde{\boldsymbol{\theta}}$ of the weights $\boldsymbol{\theta}$ for segmenting existing WM tracts can provide useful information about feature extraction. Thus, $\tilde{\boldsymbol{\theta}}$ is used to initialize $\boldsymbol{\theta}$ for training $\mathcal{M}_\mathrm{n}$, and only $\boldsymbol{\theta}_\mathrm{n}$ is randomly initialized. In this way, the knowledge learned for segmenting existing WM tracts in the pretrained model $\mathcal{M}_\mathrm{e}$ can be transferred to the segmentation of novel WM tracts, and this classic fine-tuning strategy has been effectively applied to various medical image analysis tasks (Tajbakhsh et al., 2016).

2.2. Improved knowledge transfer for few-shot segmentation of novel WM tracts

Although the classic fine-tuning strategy can be used for the few-shot segmentation of novel WM tracts, it completely discards the information about the weights $\boldsymbol{\theta}_\mathrm{e}$ in the last layer L_e learned for existing WM tracts. Since WM tracts are known to co-occur as crossing or overlapping fiber tracts in a considerable number of voxels (Ginsburger et al., 2019; Maier-Hein et al., 2017), different WM tracts can be correlated. Thus, we hypothesize that the information in the discarded weights that produce the segmentation results for existing WM tracts could also be relevant to the segmentation of novel WM tracts. More specifically, we assume that the novel WM tracts could be predicted from the existing WM tracts, and this assumption allows us to exploit the discarded information in L_e as well for training $\mathcal{M}_\mathrm{n}$, so that the knowledge transfer for the few-shot segmentation of novel WM tracts can be improved. The derivation of the improved knowledge transfer is described below.

Suppose $\mathbf{P}_\mathrm{e}^v$ and $\mathbf{P}_\mathrm{n}^v$ are the vectors of segmentation probabilities at the v-th voxel of $\mathbf{P}_\mathrm{e}$ and $\mathbf{P}_\mathrm{n}$, respectively, where $v \in \{1, 2, \ldots, V\}$ and V is the total number of voxels. Each entry in $\mathbf{P}_\mathrm{e}^v$ or $\mathbf{P}_\mathrm{n}^v$ represents the segmentation result of a tract at voxel v. In existing segmentation networks, L_e and L_n generally use a convolution with a kernel size of one to classify each voxel (e.g., see Wasserthal et al. (2018)), which is equivalent to matrix multiplication (plus a bias vector) at each voxel. Therefore, we rewrite the task-specific weights as $\boldsymbol{\theta}_\mathrm{e} = \{\mathbf{W}_\mathrm{e}, \boldsymbol{b}_\mathrm{e}\}$ and $\boldsymbol{\theta}_\mathrm{n} = \{\mathbf{W}_\mathrm{n}, \boldsymbol{b}_\mathrm{n}\}$, so that the segmentation probabilities can be explicitly expressed as

$$\mathbf{P}_\mathrm{e}^v = \sigma\left(\mathbf{W}_\mathrm{e}\mathbf{F}^v + \boldsymbol{b}_\mathrm{e}\right) \text{ and } \mathbf{P}_\mathrm{n}^v = \sigma\left(\mathbf{W}_\mathrm{n}\mathbf{F}^v + \boldsymbol{b}_\mathrm{n}\right). \tag{3}$$

Here, $\mathbf{F}^v$ represents the feature vector at the v-th voxel of the feature map $\mathbf{F}$, and $\sigma(\cdot)$ is the channel-wise sigmoid activation because there can be multiple WM tracts in a single voxel.

In classic fine-tuning the information about $\mathbf{W}_\mathrm{e}$ and $\boldsymbol{b}_\mathrm{e}$ is completely discarded. However, according to our assumption, it is possible to exploit $\mathbf{W}_\mathrm{e}$ and $\boldsymbol{b}_\mathrm{e}$ to provide a better initialization for $\mathbf{W}_\mathrm{n}$ and $\boldsymbol{b}_\mathrm{n}$. To this end, we investigate the prediction of novel WM tracts with the logits $\mathbf{H}_\mathrm{e}$ of existing WM tracts—i.e., the intermediate output of L_e before the sigmoid activation—given by the trained $\mathcal{M}_\mathrm{e}$. For simplicity, this prediction is achieved with logistic regression. We denote the logit vector at voxel v given by the trained $\mathcal{M}_\mathrm{e}$ by $\mathbf{H}_\mathrm{e}^v = (h_{\mathrm{e},1}^v, \ldots, h_{\mathrm{e},M}^v)^\mathsf{T}$, where M is the number of existing WM tracts. Suppose the total number of novel WM tracts is N; then the prediction $p_{\mathrm{e}\to\mathrm{n},j}^v$ of the j-th ($j \in \{1, \ldots, N\}$) novel WM tract at

[2] Following the terminology that is commonly used in few-shot learning (Li et al., 2019; Lifchitz et al., 2019), we use the word “novel” to represent new classes of WM tracts.

voxel ν from the information of existing WM tracts is given by

$$p^{\nu}_{\mathrm{e}\to\mathrm{n},j} = \frac{1}{1+\exp\left(-(b_j+\sum_{i=1}^{M} w_{ij}h^{\nu}_{\mathrm{e},i})\right)}, \tag{4}$$

where w_{ij} and b_j are the regression parameters to be determined.

Combining the prediction of all novel WM tracts at voxel ν into a vector $\mathbf{P}^{\nu}_{\mathrm{e}\to\mathrm{n}} = (p^{\nu}_{\mathrm{e}\to\mathrm{n},1},\ldots,p^{\nu}_{\mathrm{e}\to\mathrm{n},N})^{\mathrm{T}}$, we simply have

$$\mathbf{P}^{\nu}_{\mathrm{e}\to\mathrm{n}} = \sigma\left(\mathbf{W}\mathbf{H}^{\nu}_{\mathrm{e}} + \boldsymbol{b}\right), \tag{5}$$

where

$$\mathbf{W} = \begin{bmatrix} w_{11} & \cdots & w_{1M} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NM} \end{bmatrix} \text{ and } \boldsymbol{b} = [b_1,\ldots,b_N]^T. \tag{6}$$

Note that $\mathbf{H}^{\nu}_{\mathrm{e}} = \widetilde{\mathbf{W}}_{\mathrm{e}}\widetilde{\mathbf{F}}^{\nu} + \tilde{\boldsymbol{b}}_{\mathrm{e}}$, where $\widetilde{\mathbf{F}}^{\nu}$ corresponds to the ν-th voxel of $\widetilde{\mathbf{F}} = f(\mathbf{X};\tilde{\boldsymbol{\theta}})$ that is computed with the weights $\tilde{\boldsymbol{\theta}}$ learned for segmenting existing WM tracts, and $\widetilde{\mathbf{W}}_{\mathrm{e}}$ and $\tilde{\boldsymbol{b}}_{\mathrm{e}}$ are the values of $\mathbf{W}_{\mathrm{e}}$ and $\boldsymbol{b}_{\mathrm{e}}$ learned for segmenting existing WM tracts, respectively. Then, we have

$$\mathbf{P}^{\nu}_{\mathrm{e}\to\mathrm{n}} = \sigma\left(\mathbf{W}\left(\widetilde{\mathbf{W}}_{\mathrm{e}}\widetilde{\mathbf{F}}^{\nu} + \tilde{\boldsymbol{b}}_{\mathrm{e}}\right) + \boldsymbol{b}\right) = \sigma\left(\mathbf{W}\widetilde{\mathbf{W}}_{\mathrm{e}}\widetilde{\mathbf{F}}^{\nu} + \mathbf{W}\tilde{\boldsymbol{b}}_{\mathrm{e}} + \boldsymbol{b}\right). \tag{7}$$

Comparing $\mathbf{P}^{\nu}_{\mathrm{n}}$ in Eq. (3) and $\mathbf{P}^{\nu}_{\mathrm{e}\to\mathrm{n}}$ in Eq. (7), we notice that instead of being randomly initialized, $\boldsymbol{\theta}_{\mathrm{n}} = \{\mathbf{W}_{\mathrm{n}}, \boldsymbol{b}_{\mathrm{n}}\}$ may be better initialized using the information in $\boldsymbol{\theta}_{\mathrm{e}} = \{\mathbf{W}_{\mathrm{e}}, \boldsymbol{b}_{\mathrm{e}}\}$. Here, $\mathbf{W}$ and $\boldsymbol{b}$ still need to be computed for initializing $\boldsymbol{\theta}_{\mathrm{n}}$, and they can be computed by minimizing the difference between $\mathbf{P}^{\nu}_{\mathrm{e}\to\mathrm{n}}$ and the annotation of novel WM tracts. Note that although there are only a few annotations of novel WM tracts, they are sufficient for the computation of $\mathbf{W}$ and $\boldsymbol{b}$ because the number of unknown parameters is drastically reduced. Then, suppose the estimates of $\mathbf{W}$ and $\boldsymbol{b}$ are $\widetilde{\mathbf{W}}$ and $\tilde{\boldsymbol{b}}$, respectively; $\mathbf{W}_{\mathrm{n}}$ and $\boldsymbol{b}_{\mathrm{n}}$ are initialized as

$$\mathbf{W}_{\mathrm{n}} \leftarrow \widetilde{\mathbf{W}}\widetilde{\mathbf{W}}_{\mathrm{e}} \text{ and } \boldsymbol{b}_{\mathrm{n}} \leftarrow \widetilde{\mathbf{W}}\tilde{\boldsymbol{b}}_{\mathrm{e}} + \tilde{\boldsymbol{b}}. \tag{8}$$

Finally, with $\boldsymbol{\theta}$ initialized by $\tilde{\boldsymbol{\theta}}$ like in classic fine-tuning, all network weights in $\mathcal{M}_{\mathrm{n}}$ are learned jointly using the small number of annotations of novel WM tracts.

2.3. A better implementation with warmup

The derivation above suggests a possible way of using all information learned in $\mathcal{M}_{\mathrm{e}}$ for segmenting existing WM tracts to improve the segmentation of novel WM tracts. However, it is possible to have a more convenient implementation. To see that, we let $\mathbf{W}' = \mathbf{W}\widetilde{\mathbf{W}}_{\mathrm{e}}$ and $\boldsymbol{b}' = \mathbf{W}\tilde{\boldsymbol{b}}_{\mathrm{e}} + \boldsymbol{b}$. Then, Eq. (7) becomes

$$\mathbf{P}^{\nu}_{\mathrm{e}\to\mathrm{n}} = \sigma\left(\mathbf{W}'\widetilde{\mathbf{F}}^{\nu} + \boldsymbol{b}'\right). \tag{9}$$

This suggests that we can directly estimate $\mathbf{W}'$ and $\boldsymbol{b}'$ and use the estimated values to initialize $\boldsymbol{\theta}_{\mathrm{n}}$. This is equivalent to inserting a warmup stage before the classic fine-tuning, and the information in $\boldsymbol{\theta}_{\mathrm{e}}$ becomes redundant with such a fine-tuning strategy (but not with classic fine-tuning). Specifically, given the trained model $\mathcal{M}_{\mathrm{e}}$, for $\mathcal{M}_{\mathrm{n}}$ we first initialize $\boldsymbol{\theta}$ as $\tilde{\boldsymbol{\theta}}$. Then, we fix $\boldsymbol{\theta}$ and learn $\boldsymbol{\theta}_{\mathrm{n}}$ (randomly initialized) from the annotations of novel WM tracts. Finally, with the initial value of $\boldsymbol{\theta}_{\mathrm{n}}$ learned in the previous warmup stage, we jointly fine-tune the weights $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\mathrm{n}}$ using the annotations of novel WM tracts. Note that since $\boldsymbol{\theta}$ is already pretrained before $\boldsymbol{\theta}_{\mathrm{n}}$ is learned, in the complete training process (including pretraining), $\boldsymbol{\theta}_{\mathrm{n}}$ does not necessarily go through more iterations than $\boldsymbol{\theta}$. The impacts of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\mathrm{n}}$ on the segmentation results are naturally balanced, because 1) during pretraining and the initialization of the last layer, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\mathrm{n}}$ are updated until training convergence, respectively, and 2) at the final step of fine-tuning, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\mathrm{n}}$ are adjusted simultaneously.

The implementation with warmup not only is more convenient than the derivation in Section 2.2, but also is more adaptive and could achieve better performance for the following reasons. First, the warmup strategy is not restricted to the decomposition in Eq. (8) and allows a more adaptive use of the information in $\boldsymbol{\theta}_{\mathrm{e}}$. It can find the initialization corresponding to the decomposition as well as possibly better initialization that may not be decomposed as Eq. (8). Second, even for the case where the decomposed form allows the best initialization, the separate computation of $\{\widetilde{\mathbf{W}}, \tilde{\boldsymbol{b}}\}$ and $\{\widetilde{\mathbf{W}}_{\mathrm{e}}, \tilde{\boldsymbol{b}}_{\mathrm{e}}\}$ could accumulate the error of each computation and slightly degrade the initialization, whereas directly estimating $\mathbf{W}'$ and $\boldsymbol{b}'$ avoids the problem.

2.4. Data augmentation by tract-aware image mixing

In addition to the improved strategy of knowledge transfer from the pretrained model to the model that segments novel WM tracts, we further seek to better exploit the scarce training data with annotated novel WM tracts during the knowledge transfer. In particular, motivated by the success of mixing-based data augmentation (Yun et al., 2019; Zhang et al., 2018), we propose to generate new training samples with image mixing when there are more than one annotated scans for the novel WM tracts. These synthetic samples can be used to further improve the initialization of the last layer for segmenting novel WM tracts and hence improve the segmentation performance.

Mathematically, suppose we are given a set of images $\mathcal{X} = \{\mathbf{X}_k\}_{k=1}^{K}$ for which the annotations $\mathcal{Y} = \{\mathbf{Y}_k\}_{k=1}^{K}$ of novel WM tracts are available. Here, K is the total number of annotated images, $\mathbf{X}_k$ is the k-th annotated image, and $\mathbf{Y}_k$ is the corresponding annotation. Note that since there are N novel WM tracts of interest, each $\mathbf{Y}_k$ comprises N binary masks: $\mathbf{Y}_k = \{\mathbf{Y}_k^j\}_{j=1}^{N}$, where $\mathbf{Y}_k^j$ represents the annotation mask of $\mathbf{X}_k$ for the j-th novel WM tract.

When more than one images are annotated for the novel WM tracts (i.e., $K > 1$), we propose to mix a pair of annotated images as well as the annotations to generate additional training data. The mixing is achieved by combining different regions of the two annotated images and the corresponding annotations. Specifically, given two annotated images $\mathbf{X}_{k_1}$ and $\mathbf{X}_{k_2}$ randomly drawn from $\mathcal{X}$ and the corresponding annotations $\mathbf{Y}_{k_1}$ and $\mathbf{Y}_{k_2}$ from $\mathcal{Y}$, a synthetic image $\mathbf{X}_{\mathrm{s}}$ and its annotation $\mathbf{Y}_{\mathrm{s}}^j$ ($j \in \{1,\ldots,N\}$) for each novel WM tract can be generated as

$$\mathbf{X}_{\mathrm{s}} = \mathbf{X}_{k_1} \odot \mathbf{M} + \mathbf{X}_{k_2} \odot (1-\mathbf{M}), \tag{10}$$

$$\mathbf{Y}_{\mathrm{s}}^j = \mathbf{Y}_{k_1}^j \odot \mathbf{M} + \mathbf{Y}_{k_2}^j \odot (1-\mathbf{M}). \tag{11}$$

Here, $\odot$ represents voxelwise multiplication, and $\mathbf{M}$ is a binary mask that represents the region of $\mathbf{X}_{k_1}$ used for data mixing. The design of $\mathbf{M}$ is described below.

Considering that for segmentation problems it is beneficial to perform object-aware data mixing (Ghiasi et al., 2021), we design the mixing mask $\mathbf{M}$ such that it is aware of the novel WM tracts, i.e., the mask is determined by these tracts. Also, since the number of annotated scans is small in the few-shot setting, to generate more diverse data, we choose to vary the tract awareness with different combinations of novel WM tracts. Specifically, for each novel WM tract, its probability of contributing to the computation of the mask $\mathbf{M}$ is set to 0.5. We randomly select the novel WM tracts with this probability, and the mask $\mathbf{M}$ is determined as the union of the regions of the selected tracts in the two annotated scans that are to be mixed. Mathematically, it is equivalent to calculate $\mathbf{M}$ as

$$\mathbf{M} = \left\lceil \frac{1}{2N}\sum_{j=1}^{N} a_j(\mathbf{Y}_{k_1}^j + \mathbf{Y}_{k_2}^j) \right\rceil. \tag{12}$$

Here, a_j is sampled from the Bernoulli distribution Bernoulli(0.5) and determines whether the j-th novel WM tract contributes to

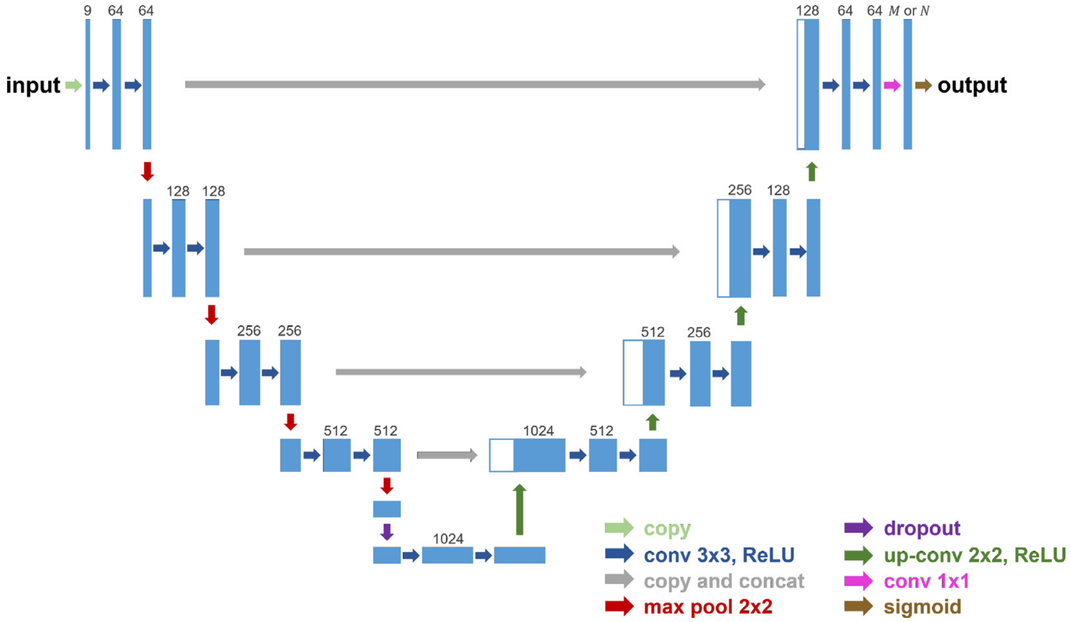

Fig. 1. The detailed network architecture of TractSeg, which is used as the backbone segmentation network in this work. The numbers of channels are indicated for the layers. Note that the number of channels of the last layer is M when the network segments existing WM tracts, and the number is N for segmenting novel WM tracts.

the computation of $\mathbf{M}$; $\lceil\cdot\rceil$ represents the ceiling operation. Since the data mixing in Eqs. (10) and (11) with the mask defined in Eq. (12) mixes the WM tracts in two images, the proposed data augmentation strategy is named TractMix. Note that in general the samples generated by image mixing may not always look realistic, yet they can still benefit the network training (Yun et al., 2019; Zhang et al., 2018).

By repeating the random sampling of the annotated images used for image mixing and WM tracts used for computing $\mathbf{M}$, a number of synthetic annotated images can be generated. With K real images annotated for N novel WM tracts, TractMix can produce $K \times (K-1) \times (2^N - 1)$ different synthetic annotated images at most, and duplicate synthetic samples are not allowed. Suppose the desired number of synthetic annotated images is K_s; in this work, we set $K_s = \min\{100, K \times (K-1) \times (2^N - 1)\}$, so that a large number of unique synthetic images can be produced. The synthetic images and their annotations are used together with the real annotated images $\mathcal{X}$ and real annotations $\mathcal{Y}$ for initializing the last layer of the network that segments novel WM tracts. Note that for the final fine-tuning step where all network weights are jointly updated, only the real images $\mathcal{X}$ and real annotations $\mathcal{Y}$ are used. This is because after the initialization of the last layer, the network weights can be already close to the desired values, and the incorporation of the synthetic samples that may be unrealistic could negatively affect the final fine-tuning.

We choose to perform offline data augmentation with TractMix, where the synthetic samples are generated before network training, so that TractMix can be conveniently integrated with an arbitrary segmentation framework without the need of modifying its code, for example, the code for batch generation. Also, even when only the interface of the segmentation framework is available without the access to its source code, the offline data augmentation can still be applied.

Note that the data augmentation step can be optional. As shown later in the experimental results in Section 3.3 and Appendix A, the proposed data augmentation approach allows substantially improved segmentation accuracy for the more challenging scenario where domain shift exists between the data used for segmenting existing and novel WM tracts. For the less challenging scenario without the domain shift, the segmentation performance of the proposed method without TractMix is already good, and the segmentation accuracy is similar with or without TractMix.

2.5. Backbone network for WM tract segmentation

Our method is generic and agnostic to the structure of the segmentation network. For demonstration, we choose the TractSeg architecture (Wasserthal et al., 2018) as the backbone network, which has achieved state-of-the-art performance and been applied to brain studies (Veraart et al., 2021; Bryant et al., 2021), but other networks can also be used if they are shown superior to the TractSeg architecture.

The detailed network architecture of TractSeg is shown in Fig. 1. TractSeg uses an encoder-decoder CNN based on the 2D U-net architecture (Ronneberger et al., 2015) to segment WM tracts. The inputs to the CNN are fiber orientation maps, so that the network can be applied to data acquired with various protocols. The fiber orientations are computed with *multi-shell multi-tissue constrained spherical deconvolution* (MSMT-CSD) (Jeurissen et al., 2014; Tournier et al., 2019) for multi-shell dMRI data and *constrained spherical deconvolution* (CSD) (Tournier et al., 2007) for single-shell dMRI data. For each voxel, the maximum number of fiber orientations is set to three, and thus there are nine input channels.[3] Given a 3D image of fiber orientations, the network performs 2D segmentation for each image view—the coronal, axial, or sagittal view—separately, and then these results are merged for the final segmentation. Note that the same network structure is used to segment existing or novel WM tracts, except that the number of channels of the output layer is M for existing WM tracts and N for novel WM tracts.

2.6. Implementation details

We have implemented the proposed method based on the open-source code of TractSeg at https://github.com/MIC-DKFZ/TractSeg/ using PyTorch (Paszke et al., 2019). To pretrain the network for segmenting existing WM tracts, we follow Wasserthal et al. (2018) and minimize the cross-entropy loss,

[3] If there are fewer than three fiber orientations, the intensities are set to zero in the corresponding channels.

Table 1
A list of the 12 novel WM tracts and their abbreviations.

	WM tract name	abbreviation		WM tract name	abbreviation
1	Corticospinal tract left	CST_left	7	Optic radiation left	OR_left
2	Corticospinal tract right	CST_right	8	Optic radiation right	OR_right
3	Fronto-pontine tract left	FPT_left	9	Inferior longitudinal fascicle left	ILF_left
4	Fronto-pontine tract right	FPT_right	10	Inferior longitudinal fascicle right	ILF_right
5	Parieto-occipital pontine left	POPT_left	11	Uncinate fascicle left	UF_left
6	Parieto-occipital pontine right	POPT_right	12	Uncinate fascicle right	UF_right

where Adamax (Kingma and Ba, 2015) is used as the optimizer with a learning rate of 0.001 and a batch size of 47 (Wasserthal et al., 2019). In addition, dropout with a probability of 0.4 is used like in Wasserthal et al. (2018). Network training is performed with 300 epochs to ensure convergence, and the model corresponding to the epoch with the highest Dice score on a validation set is selected. Similarly for the novel WM tracts, the training specification described above is also used at each step of parameter learning, including the initialization of the network weights of the last layer and the final fine-tuning. Since in TractSeg traditional data augmentation (elastic deformation, scaling, intensity perturbation, etc.) is applied to training images online by default, these operations are also performed online during pretraining and at each training step of the proposed transfer learning approach (for the synthetic samples generated offline by TractMix as well).

3. Results

In this section, we present the validation of our method on the publicly available HCP dataset (Van Essen et al., 2013) and a private dataset comprising both HC subjects and AD patients. In the experiments, the proposed method was evaluated under various experimental settings. We first introduce the datasets and experimental settings, and then we describe the experimental results on the two datasets.

3.1. Data description and experimental settings

3.1.1. The HCP dataset

We first selected the dMRI scans from the HCP dataset (Van Essen et al., 2013) for evaluation. The dMRI scans were acquired with 270 diffusion gradients ($b = 1000,\ 2000$, and $3000\ \mathrm{s/mm^2}$) and an isotropic spatial resolution of 1.25 mm (Sotiropoulos et al., 2013), and they have been processed by the minimal preprocessing pipeline (Glasser et al., 2013). For these dMRI scans, 72 major WM tracts were annotated, and the annotations are provided by Wasserthal et al. (2018). For the complete list of the annotated WM tracts, we refer the readers to Wasserthal et al. (2018).

We split the 72 WM tracts into a set of existing WM tracts and a set of novel WM tracts, which comprised 60 and 12 tracts, respectively, i.e., $M = 60$ and $N = 12$. The 12 novel WM tracts were randomly selected from the bilateral WM tracts, and the names and abbreviations of the novel WM tracts are listed in Table 1. The existing WM tracts correspond to the remaining WM tracts.

We considered the experimental setting where abundant annotations were available for existing WM tracts and only a few annotated scans were available for novel WM tracts. Specifically, 65 dMRI scans were used to pretrain the network that segments existing WM tracts, together with the corresponding annotations of the existing tracts. During pretraining, these 65 dMRI scans were split into a training set of 52 dMRI scans and a validation set of 13 dMRI scans. Then, for segmenting the novel WM tracts, we selected four other dMRI scans to fine-tune the network with their annotations of the novel tracts. Three of them were used as the training set, and the other one was used as the validation set. Another set of 30 annotated dMRI scans that were different from all the training and validation scans described above was selected as the test set to evaluate the segmentation performance of the proposed method for the novel WM tracts. The results under this experimental setting will be presented in Section 3.2.1.

To further investigate the impact of the number of annotated training scans for the novel WM tracts on the segmentation performance, we considered two additional experimental settings. In the first setting, only one annotated scan was kept in the training set for fine-tuning, and only itself was used as the validation scan. In the second setting, additional dMRI scans with annotations of novel WM tracts were included in the training and validation sets, so that the total numbers of training and validation scans for fine-tuning were five and two, respectively. In both settings, we did not alter the test set, and the same pretrained model for segmenting existing WM tracts was used. The results under these experimental settings will be shown in Section 3.2.2.

To show that the proposed method is applicable to different numbers of novel WM tracts, we also varied the selection of the novel WM tracts, where only a subset of the 12 novel WM tracts was included in fine-tuning and evaluation (still with three annotated training scans and one annotated validation scan). Specifically, two selections were considered. In the first selection (referred to as Selection One), CST_left, CST_right, POPT_left, POPT_right, OR_left, and OR_right were included as the novel WM tracts, and in the second selection (referred to as Selection Two), CST_left, CST_right, OR_left, and OR_right were included as the novel WM tracts. The other training and evaluation settings were not changed. The results under these experimental settings will be presented in Section 3.2.3.

In addition, like Wasserthal et al. (2018) and Lu et al. (2021) we investigated the impact of data quality on the segmentation performance. Specifically, we generated clinical quality dMRI scans by downsampling the original HCP dMRI scans to the spatial resolution of 2.5 mm isotropic and then selecting only 34 diffusion gradients at $b = 1000\ \mathrm{s/mm^2}$. The annotations were also downsampled accordingly. The sets of the training, validation, and test subjects were the same as the sets specified for the experiments with the original HCP dMRI scans, but only the generated clinical quality data was used at each step (including both pretraining and fine-tuning). Note that the same three cases of the numbers of training and validation scans for fine-tuning were considered here for the 12 novel WM tracts, where the numbers were 1/0, 3/1, and 5/2 for the training/validation scans, respectively. The results under these experimental settings will be shown in Section 3.2.4.

Note that for all experimental results presented in Section 3.2 on the HCP dataset, where the scans used for segmenting existing and novel WM tracts are from the same dataset with the same imaging settings, TractMix was not applied, as the proposed method already achieved good segmentation quality without TractMix and further incorporation of TractMix would lead to similar segmentation performance. A detailed description of the comparison between the results of the proposed method achieved with and without TractMix will be given in Appendix A.

3.1.2. The private dataset

To show that the proposed method is not just applicable to the HCP dataset, we also used a private dataset (Qin et al., 2021) to evaluate the segmentation performance. This dataset contained both HC subjects and AD patients. The dMRI scans in the private dataset were acquired on a GE Premier scanner with an isotropic spatial resolution of 1.7 mm and 270 diffusion gradients ($b = 1000,\ 2000$, and $3000\ \mathrm{s/mm^2}$). These dMRI scans were preprocessed with the FSL topup (Andersson et al., 2003) and eddy (Andersson and Sotiropoulos, 2016) tools for distortion and motion correction.

For this dataset, 10 WM tracts were annotated according to the annotation protocol described in Wasserthal et al. (2018), and they included the tracts listed in Table 1 except ILF_left and ILF_right. The ten annotated tracts were used as the set of novel WM tracts. Note that since other WM tracts were not annotated for this dataset, we selected the network pretrained with the original HCP dMRI scans for segmenting existing WM tracts as the pretrained model, and this pretrained model was then fine-tuned with the annotated dMRI scans in the private dataset for segmenting the novel WM tracts. Specifically, during fine-tuning, one annotated dMRI scan of an HC subject and one annotated dMRI scan of an AD patient were used as the training scans, and they were also used for model selection. The annotated dMRI scans of five other HC subjects and five other AD patients were used as the test set to evaluate the segmentation quality. In addition, like the experiments on the HCP dataset, we varied the selection of the novel WM tracts and evaluated the segmentation performance. Here, the same two subsets of novel WM tracts selected for the HCP dataset—i.e., Selection One and Selection Two—were used. The other training and evaluation settings were kept unchanged.

Note that the proposed method was applied both with and without TractMix for the private dataset, where the segmentation was more challenging due to the domain shift between the data used for pretraining and fine-tuning. The results for the private dataset will be reported in Section 3.3.

3.2. Results on the HCP dataset

3.2.1. Evaluation of segmentation quality

We first evaluated the performance of the proposed method with 12 novel WM tracts using the original HCP dataset, where four annotated scans (three for training and one for validation) were used for fine-tuning as described in Section 3.1.1. The fine-tuning of the pretrained model was performed with either the initialization strategy proposed in Section 2.2 or the more convenient implementation in Section 2.3. For convenience, hereinafter we refer to the approaches proposed in Sections 2.2 and 2.3 as Ours1 and Ours2, respectively. Ours1 and Ours2 were compared with three competing methods. The first one is the baseline TractSeg network that was trained from scratch with the annotations of novel WM tracts, where the pretrained model was not used. The second one is a representative conventional registration-based segmentation method Atlas FSL developed in Wasserthal et al. (2019), where a volumetric tract atlas was created with the few dMRI scans annotated for the novel WM tracts and then registered to the test scans. The third competing method is the classic fine-tuning method based on the pretrained model for segmenting existing WM tracts and the annotations of novel WM tracts, and this method is referred to as FT.

The qualitative evaluation results are given in Fig. 2 for the proposed and competing methods, where the 3D renderings of the segmentation results and their cross-sectional views overlaid on the *fractional anisotropy* (FA) map are shown for representative test subjects and WM tracts. Here, the manual delineations—i.e., annotations—of the tracts are also shown for reference. In addition, zoomed views of the highlighted regions in the cross-sectional views are shown in Fig. 2. Compared with the competing methods, the geometry and spatial coverage of the tracts given by Ours1 and Ours2 are more similar to those of the manual delineations.

The proposed method was also evaluated quantitatively, where the Dice coefficient was computed for each WM tract on each test scan. These Dice coefficients are displayed in the boxplots in Fig. 3 for the proposed and competing methods. For reference, the *upper bound* (UB) Dice coefficient was also computed and is shown in Fig. 3, which represents the segmentation performance achieved with abundant annotations of novel WM tracts. Specifically, to compute the UB Dice coefficient, the 65 dMRI scans originally used for pretraining the network that segments existing WM tracts were directly used to train the network that segments novel WM tracts (using their annotations of novel WM tracts), together with the four dMRI scans originally used for network fine-tuning, and the network was trained from scratch. Note that the validation subjects originally used for pretraining and fine-tuning were combined as the validation set for computing the UB performance.

From Fig. 3, we can see that our method (either Ours1 or Ours2) achieved higher Dice coefficients than the baseline method, Atlas FSL, and FT, and these Dice coefficients are much closer to the UB performance. The results of Ours1 and Ours2 are similar. Note that the variation of the Dice coefficient can be greater for some WM tracts than the others. This may be caused by the different variability of WM tracts. Some WM tracts may have greater shape variability across different subjects than the other tracts, and thus the difficulty of segmenting them varies more across subjects.

Besides, Ours1 or Ours2 was compared with the competing methods for each novel WM tract using paired Student's t-tests[4], and the effect sizes (Cohen's d) were measured for the comparison. These results are shown in Table 2. It can be seen that for each tract both Ours1 and Ours2 highly significantly ($p < 0.001$ after Bonferroni correction for multiple comparisons) outperform the baseline method, Atlas FSL, and FT with large effect sizes ($d > 0.8$).

We also computed the average Dice coefficient for each tract and each method, and the results are summarized in the boxplots in Fig. 4 together with the UB performance. The mean value of these average Dice coefficients is indicated for each method and UB in Fig. 4 as well. Consistent with Fig. 3, both Ours1 and Ours2 outperform the competing methods, and their results are close to the UB performance. In addition, Ours2 has a slightly higher mean Dice coefficient than Ours1. Then, we compared the average Dice coefficients of the tracts between Ours2 and the other methods (including Ours1) using paired Student's t-tests[5] and measured the effect sizes (Cohen's d). These results are also shown in Fig. 4. The performance of Ours2 is highly significantly ($p < 0.001$) better than those of the competing methods (the baseline method, Atlas FSL, and FT) with large effect sizes ($d > 0.8$). Although the difference between Ours1 and Ours2 is significant ($p < 0.05$), the effect size is very small. This indicates that Ours2 is better than Ours1 for most tracts, but the difference is very small.

To confirm the benefit of each step in the proposed transfer learning approach (both Ours1 and Ours2), the mean Dice coefficient achieved with only the initialization of the last layer without the final fine-tuning step is shown in Table 3, and the final results of Ours1 and Ours2 are listed again for reference. For both Ours1 and Ours2, after the initialization of the last layer only (Step One),

[4] The paired Student's t-test was selected because the difference between the two methods being compared is independent across the scans belonging to different subjects and its distribution resembles the Gaussian distribution.

[5] Since each WM tract has a distinct definition, it is reasonable to assume that the difference between the two methods being compared is independent across the tracts. Also, since the distribution of the difference resembles the Gaussian distribution, the paired Student's t-test was selected here again.

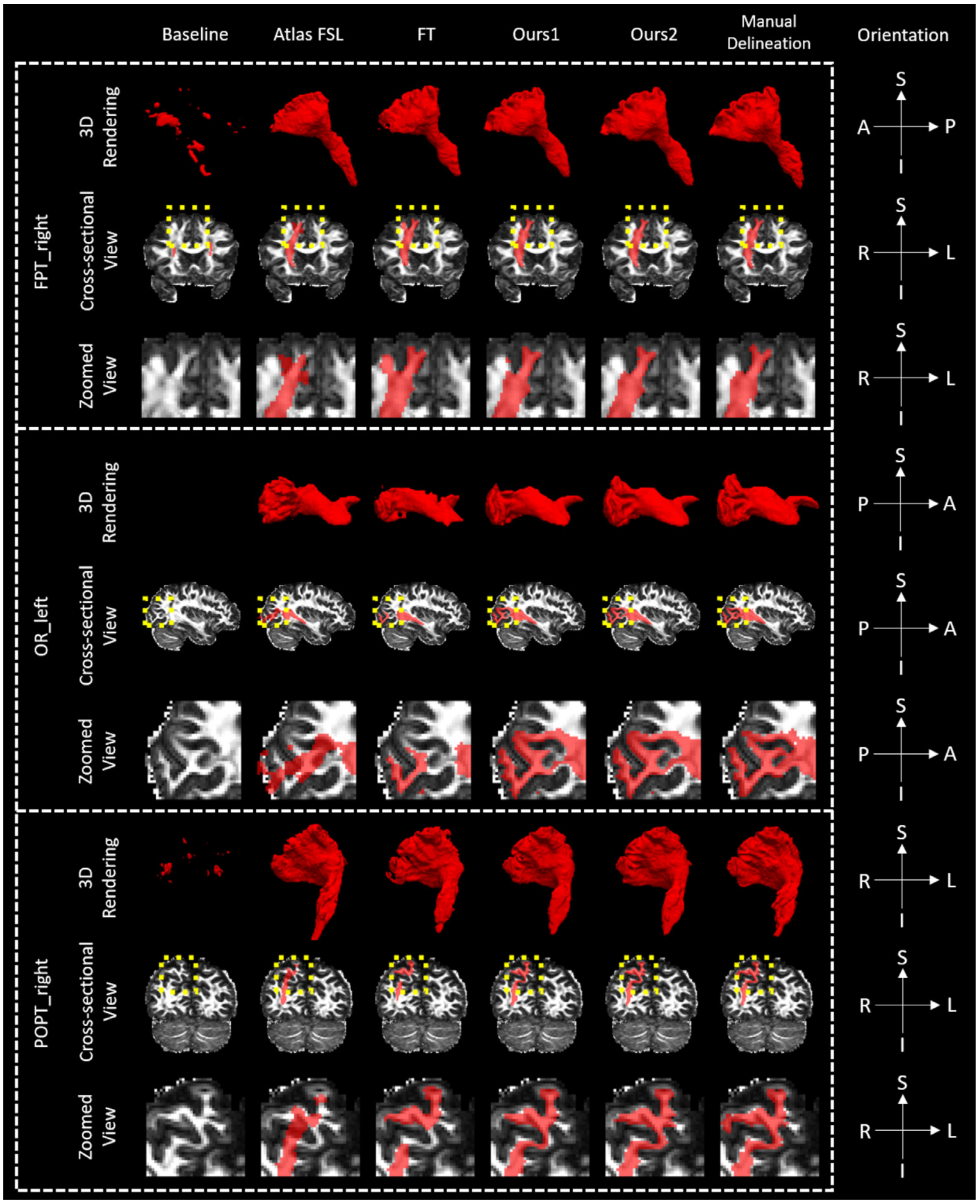

Fig. 2. 3D renderings and cross-sectional views of the segmentation results (red) for representative test subjects and WM tracts. The cross-sectional views are overlaid on the FA map. The zoomed views of the highlighted regions in the cross-sectional views are also shown. The results of the proposed and competing methods are shown together with the manual delineations. The image orientations are indicated in the rightmost column. Note the 3D renderings for comparing the tract geometry and the highlighted regions for comparing the spatial coverage of tracts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the segmentation network can provide a moderate mean Dice coefficient, and after the final fine-tuning step (Step Two), the segmentation accuracy is substantially improved.

Like Lu et al. (2021), we considered an additional metric for evaluation, which is the *relative volume difference* (RVD) (Yeghiazaryan and Voiculescu, 2018) between the segmented WM tract and the corresponding manual delineation. When similar segmentation accuracy is achieved in terms of the Dice coefficient, a smaller RVD value is desired to reduce the bias of the quantification of tract volumes. We computed the average RVD for each tract, and the results of each method and UB are shown in Fig. 5. The mean of the average RVD values is also indicated for each method and UB in Fig. 5. Again, like in Fig. 4 the average RVD values of Ours2 were compared with those of the other methods (including Ours1) using paired Student's t-tests, and the effect sizes were computed. We can see that both Ours1 and Ours2 have better (smaller) RVD values than the competing methods, and their results are close to the UB performance. Ours2 is significantly ($p < 0.05$ or $p < 0.001$) better than the competing methods with large ($d > 0.8$) or medium (d close to 0.5) effect sizes. Although the mean RVD of Ours2 is slightly smaller than that of Ours1, the performance of Ours1 and Ours2 is comparable, as indicated by the very small effect size and non-significant difference between them.

3.2.2. Impact of the number of training scans annotated for novel WM tracts

We further investigated the impact of the number of training scans that were annotated for novel WM tracts. As described in Section 3.1.1, we considered two additional experimental settings, where the numbers of annotated scans in the training/validation set for network fine-tuning were 1/0 and 5/2, respectively.

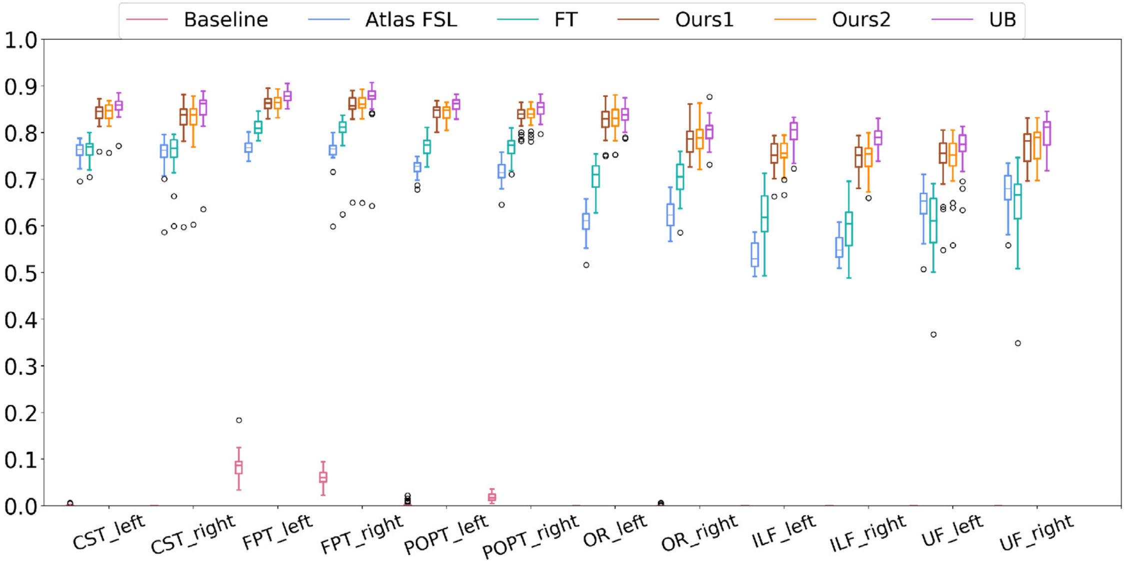

Fig. 3. Boxplots of the Dice coefficients on the test scans for all 12 novel WM tracts. Our method (either Ours1 or Ours2) achieved higher Dice coefficients than the competing methods, and these Dice coefficients are also much closer to the UB performance.

Table 2
The effect sizes (Cohen's d) for the comparison of Dice coefficients between the proposed method (Ours1 or Ours2) and the competing methods for each novel WM tract. Asterisks (***) indicate that the difference between the proposed and competing methods is highly significant ($p < .001$) using a paired Student's t-test after Bonferroni correction for multiple comparisons.

Ours1 v.s.		Baseline	Atlas FSL	FT	Ours2 v.s.		Baseline	Atlas FSL	FT
CST_left	d	53.86	3.85	3.59	CST_left	d	52.32	3.86	3.61
	p	***	***	***		p	***	***	***
CST_right	d	24.23	1.71	1.61	CST_right	d	24.57	1.71	1.61
	p	***	***	***		p	***	***	***
FPT_left	d	35.60	5.88	3.08	FPT_left	d	36.16	6.16	3.29
	p	***	***	***		p	***	***	***
FPT_right	d	25.45	2.59	1.26	FPT_right	d	25.09	2.59	1.28
	p	***	***	***		p	***	***	***
POPT_left	d	70.62	7.34	3.81	POPT_left	d	69.52	7.25	3.75
	p	***	***	***		p	***	***	***
POPT_right	d	50.88	5.42	2.88	POPT_right	d	51.12	5.50	2.95
	p	***	***	***		p	***	***	***
OR_left	d	41.42	7.30	3.90	OR_left	d	40.79	7.34	3.97
	p	***	***	***		p	***	***	***
OR_right	d	33.23	5.26	2.30	OR_right	d	32.93	5.34	2.40
	p	***	***	***		p	***	***	***
ILF_left	d	35.44	7.24	2.96	ILF_left	d	36.78	7.49	3.07
	p	***	***	***		p	***	***	***
ILF_right	d	34.06	6.55	3.32	ILF_right	d	31.48	6.24	3.24
	p	***	***	***		p	***	***	***
UF_left	d	19.35	2.07	2.31	UF_left	d	20.31	2.09	2.32
	p	***	***	***		p	***	***	***
UF_right	d	27.62	2.35	2.09	UF_right	d	29.09	2.53	2.20
	p	***	***	***		p	***	***	***

Table 3
The mean value of the average Dice coefficients of the novel WM tracts after each step of the proposed transfer learning approach (both Ours1 and Ours2). For convenience, we refer to the initialization of the last layer only without the final fine-tuning step as Step One, and the final fine-tuning step is referred to as Step Two.

Ours1		Ours2	
Step One	Step Two	Step One	Step Two
0.361	0.807	0.413	0.808

For each additional experimental setting, we computed the average Dice coefficient and the average RVD for each tract. The means of the average Dice coefficients and the average RVD values are reported for each method in Tables 4 and 5, respectively. The UB performance was also computed and listed for reference. We can see that under these two settings, either Ours1 or Ours2 is better than the competing methods, as indicated by the higher Dice coefficients and lower RVD values, and their results are closer to the UB performance. In particular, the performance of either Ours1 or Ours2 with only one annotated training scan for the novel WM tracts is better than the performance of classic fine-tuning with five annotated training scans.

In addition, in Tables 4 and 5, we compared the average Dice coefficients and average RVD values of Ours2 with those of the other methods using paired Student's t-tests and measured the effect sizes. For both Dice coefficients and RVD values, Ours2 significantly ($p < 0.05$ or $p < 0.001$) outperforms the baseline method, Atlas FSL, and FT, mostly with large effect sizes ($d > 0.8$). Although the Dice coefficient of Ours2 is slightly higher than that of Ours1,

Table 4
The means of the average Dice coefficients of the novel WM tracts achieved with different numbers of annotated training scans. Our results are highlighted in bold. The effect sizes (Cohen's d) for comparing the average Dice coefficients between Ours2 and the other methods are also listed. Asterisks indicate that the difference between Ours2 and the other method is significant using a paired Student's t-test. ($^{**}p < .01$, $^{***}p < .001$, n.s. $p \geq 0.05$).

Annotated training scans		Baseline	Atlas FSL	FT	Ours1	Ours2	UB
1	Dice	0.000	0.645	0.590	**0.777**	**0.784**	0.828
	d	20.444	1.919	1.944	0.131	-	-
	p	***	***	***	**	-	-
5	Dice	0.052	0.683	0.757	**0.811**	**0.812**	0.830
	d	10.362	2.004	1.000	0.021	-	-
	p	***	***	***	n.s.	-	-

Table 5
The means of the average RVD values of the novel WM tracts achieved with different numbers of annotated training scans. Our results are highlighted in bold. The effect sizes (Cohen's d) for comparing the average RVD values between Ours2 and the other methods are also listed. Asterisks indicate that the difference between Ours2 and the other method is significant using a paired Student's t-test. ($^{*}p < .05$, $^{***}p < .001$, n.s. $p \geq 0.05$).

Annotated training scans		Baseline	Atlas FSL	FT	Ours1	Ours2	UB
1	RVD	1.000	0.182	0.392	**0.156**	**0.151**	0.105
	d	17.458	0.403	1.854	0.067	-	-
	p	***	*	***	n.s.	-	-
5	RVD	0.955	0.199	0.158	**0.129**	**0.131**	0.105
	d	11.234	0.815	0.372	0.036	-	-
	p	***	*	*	n.s.	-	-

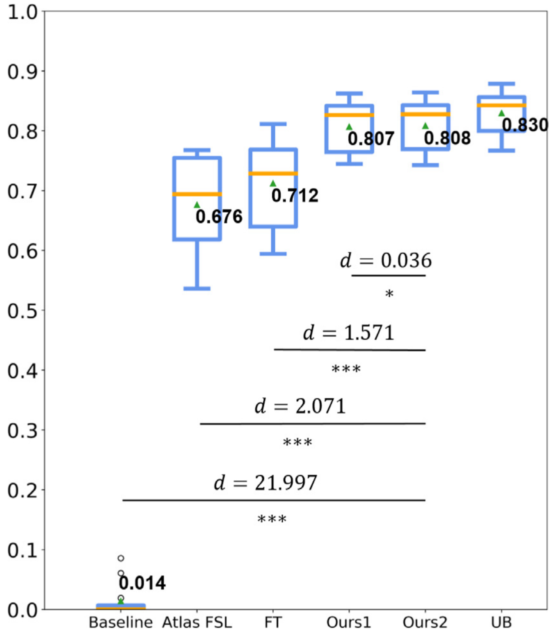

Fig. 4. Boxplots of the average Dice coefficient for each tract. The means of the average Dice coefficients are indicated. The effect sizes (Cohen's d) for comparing the average Dice coefficients between Ours2 and the other methods are also listed. Asterisks indicate that the difference between Ours2 and the other method is significant ($^{*}p < .05$ and $^{***}p < .001$) using a paired Student's t-test.

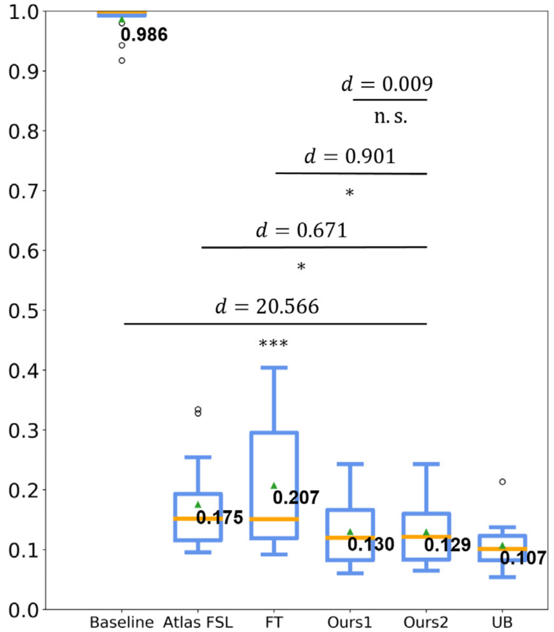

Fig. 5. Boxplots of the average RVD for each tract. The means of the average RVD values are indicated. The effect sizes (Cohen's d) for comparing the average RVD values between Ours2 and the other methods are also listed. Asterisks indicate that the difference between Ours2 and the other method is significant ($^{*}p < .05$ and $^{***}p < .001$) using a paired Student's t-test. Note that n.s. represents non-significant ($p \geq 0.05$) difference.

their performance is still comparable, as indicated by the small effect sizes ($d < 0.2$) of the Dice coefficient and RVD.

3.2.3. Segmentation performance with different selections of novel WM tracts

We then varied the selection of the novel WM tracts as described in Section 3.1.1 and evaluated the segmentation performance for the different selections (with three annotated training scans). For each selection, we computed the average Dice coefficient for each novel WM tract, and the results are summarized in Table 6 for each method and the UB. Like the results achieved with 12 novel WM tracts in Section 3.2.1, for these different selections of novel WM tracts, the proposed method also outperforms the competing methods, and its performance is close to the UB.

3.2.4. Impact of data quality

We also investigated the impact of data quality on the segmentation performance as described in Section 3.1.1. In particular, we used 52/13 clinical quality dMRI scans as the training/validation

Table 6
The average Dice coefficient for each novel WM tract when the selection of the novel WM tracts varied (with three annotated training scans). Our results are highlighted in bold.

Selection One	Baseline	Atlas FSL	FT	Ours1	Ours2	UB
CST_left	0.000	0.761	0.750	**0.835**	**0.842**	0.856
CST_right	0.002	0.754	0.740	**0.825**	**0.830**	0.848
POPT_left	0.000	0.723	0.760	**0.841**	**0.846**	0.859
POPT_right	0.000	0.714	0.763	**0.834**	**0.835**	0.850
OR_left	0.000	0.603	0.697	**0.823**	**0.823**	0.836
OR_right	0.000	0.623	0.707	**0.783**	**0.780**	0.801
Selection Two	Baseline	Atlas FSL	FT	Ours1	Ours2	UB
CST_left	0.000	0.761	0.802	**0.841**	**0.841**	0.855
CST_right	0.000	0.754	0.802	**0.831**	**0.829**	0.849
OR_left	0.000	0.603	0.775	**0.828**	**0.827**	0.836
OR_right	0.000	0.623	0.754	**0.786**	**0.788**	0.803

set to pretrain the model for segmenting existing WM tracts, respectively. Then, 1/0, 3/1, and 5/2 clinical quality dMRI scans were used as the training/validation set to fine-tune the segmentation network for the 12 novel WM tracts, respectively.

The average Dice coefficient and average RVD were computed for each tract. Their mean values are summarized for each method in Tables 7 and 8, together with the UB performance achieved with the clinical quality data. Also, the average Dice coefficients and average RVD values of Ours2 were compared with those of the other methods using paired Student's t-tests, and the effect sizes were computed. These results are given in Tables 7 and 8 too.

Both Ours1 and Ours2 outperform the competing methods with higher Dice coefficients, and the improvement of Ours2 is highly significant ($p < 0.001$) with large effect sizes ($d > 0.8$) compared with the competing methods. For the RVD value, Ours1 is better than the competing methods in most cases; in all cases, Ours2 is better than the competing methods, and the difference is significant with large effect sizes in most cases. The performance of Ours2 is either comparable to that of Ours1 with small effect sizes ($d < 0.2$) or better than that of Ours1 with medium effect sizes (d close to 0.5).

3.3. Results on the private dataset

In addition to the results on the HCP dataset, we evaluated our approach on a private dataset comprising both HC subjects and AD patients as described in Section 3.1.2. Specifically, the segmentation network that was pretrained on the original HCP dataset for existing WM tracts was fine-tuned with one HC dMRI scan and one AD dMRI scan from the private dataset for segmenting novel WM tracts.

We first evaluated the segmentation performance with all ten novel WM tracts. The average Dice coefficient and average RVD value were computed for each tract using all the test scans, and their means are summarized in Table 9 for each method. Both Ours1 and Ours2 have better Dice coefficients than the competing methods. When Ours1 and Ours2 are combined with TractMix (referred to as Ours1 + TractMix and Ours2 + TractMix, respectively), the Dice coefficients are further improved. For the additional RVD metric, Ours1 and Ours2 have better results than the baseline method and FT, which are CNN-based competing methods, and Ours1 + TractMix and Ours2 + TractMix further reduce the RVD; however, Atlas FSL has a lower RVD value than the proposed strategies. For the comparison among the results of the proposed method, the performance of Ours2 or Ours2 + TractMix is better than that of Ours1 or Ours1 + TractMix, respectively, and Ours2 + TractMix has the best performance.

In addition, we compared the average Dice coefficients and average RVD values of Ours2 + TractMix with the results of the other methods using paired Student's t-tests and measured the effect sizes. Regarding the Dice coefficients, Ours2 + TractMix outperforms the competing methods highly significantly ($p < .001$) with large effect sizes ($d > 0.8$), and it also outperforms Ours1, Ours2, and Ours1 + TractMix with statistical significance ($p < 0.001$ or $p < 0.01$). Except for Atlas FSL, the RVD result of Ours2 + TractMix is highly significantly better than those of the competing methods with large effect sizes, and it is also better than the RVD results of Ours1, Ours2, and Ours1 + TractMix with statistical significance ($p < 0.001$ or $p < 0.01$).

We also investigated the segmentation performance for the HC subjects and the AD patients separately, where the means of the average Dice coefficients and average RVD values were computed for these two individual groups. The results are listed in Table 9 as well, and they are consistent with the results computed with all test subjects. For both of the HC and AD groups, our method (Ours1, Ours2, Ours1 + TractMix, or Ours2 + TractMix) has better performance than the competing methods, except for the RVD results of Atlas FSL, and the incorporation of TractMix improves the segmentation quality for both Ours1 and Ours2.

Next, we varied the selection of the novel WM tracts as described in Section 3.1.2 and evaluated the segmentation performance for these cases. For each selection, the means of the average Dice coefficients and average RVD values of the novel WM tracts are summarized in Table 10 for each method. Like the results achieved with ten novel WM tracts, our method achieves better performance than the competing methods, except for the RVD values of Atlas FSL, and the incorporation of TractMix benefits both Ours1 and Ours2.

4. Discussion

Although a large number of annotated scans can be carefully curated for training networks that segment WM tracts, it is possible that certain brain studies focus on novel WM tracts that are not included in the existing annotated WM tracts. Because annotating WM tracts is laborious, it is desired that the knowledge learned for segmenting existing WM tracts can be transferred to the segmentation of novel WM tracts, so that with only a few scans that are annotated for the novel WM tracts, the segmentation can be performed accurately. To this end, we propose a fine-tuning strategy that allows all the knowledge learned for segmenting existing WM tracts to be exploited for segmenting the novel WM tracts, and this strategy is further improved with a more convenient and adaptive implementation with warmup. Our method was evaluated on different datasets under various settings, as well as on both healthy subjects and patients. The results show the benefit of the proposed fine-tuning method compared with other segmentation approaches, including classic fine-tuning. In addition, the results indicate that the more convenient and adaptive implementation with warmup has better or comparable performance compared with the original implementation. This is consistent with the argument in Section 2.3.

In addition to the improved fine-tuning strategy, we have proposed a simple yet effective data augmentation approach TractMix, which can further benefit the segmentation of novel WM tracts given a small number of training scans annotated for these tracts. In TractMix, tract-aware mixing of pairs of annotated images is performed, and to generate diverse synthetic training data, the annotated images are mixed with different combinations of WM tracts. The experimental results demonstrate that TractMix is beneficial for the few-shot segmentation of novel WM tracts in the more challenging scenario where domain shift exists between the data used for segmenting existing and novel WM tracts. For the less challenging scenario without the domain shift, the segmentation accuracy achieved with or without TractMix is similar

Table 7
The means of the average Dice coefficients of the novel WM tracts for the clinical quality data. Our results are highlighted in bold. The effect sizes (Cohen's d) for comparing the average Dice coefficients between Ours2 and the other methods are also listed. Asterisks indicate that the difference between Ours2 and the other method is significant using a paired Student's t-test. ($^{*}p < .05$, $^{***}p < .001$).

Annotated training scans		Baseline	Atlas FSL	FT	Ours1	Ours2	UB
1	Dice	0.000	0.634	0.043	**0.694**	**0.724**	0.788
	d	15.097	1.196	9.934	0.445	-	-
	p	***	***	***	***	-	-
3	Dice	0.000	0.659	0.473	**0.758**	**0.764**	0.790
	d	18.147	1.476	2.075	0.102	-	-
	p	***	***	***	***	-	-
5	Dice	0.000	0.662	0.635	**0.757**	**0.761**	0.784
	d	15.879	1.334	1.352	0.057	-	-
	p	***	***	***	*	-	-

Table 8
The means of the average RVD values of the novel WM tracts for the clinical quality data. Our results are highlighted in bold. The effect sizes (Cohen's d) for comparing the average RVD values between Ours2 and the other methods are also listed. Asterisks indicate that the difference between Ours2 and the other method is significant using a paired Student's t-test. ($^{*}p < .05$, $^{***}p < .001$, n.s. $p \geq 0.05$).

Annotated training scans		Baseline	Atlas FSL	FT	Ours1	Ours2	UB
1	RVD	1.000	0.224	41.447	**0.258**	**0.213**	0.171
	d	14.318	0.131	1.314	0.583	-	-
	p	***	n.s.	*	***	-	-
3	RVD	1.000	0.238	0.562	**0.159**	**0.157**	0.160
	d	16.683	0.975	2.841	0.032	-	-
	p	***	***	***	n.s.	-	-
5	RVD	1.000	0.271	0.362	**0.183**	**0.176**	0.178
	d	12.906	0.995	1.594	0.075	-	-
	p	***	***	***	*	-	-

Table 9
The means of the average Dice coefficients and average RVD values of the novel WM tracts for the private dataset. Our results are highlighted in bold. The effect sizes (Cohen's d) for comparing the average Dice coefficients or average RVD values between Ours2 + TractMix and the other methods are also listed. Asterisks indicate that the difference between Ours2 + TractMix and the other method is significant ($^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$) using a paired Student's t-test. The means computed with the HC subjects and the AD patients separately are shown as well.

		Baseline	Atlas FSL	FT	Ours1	Ours2	Ours1 + TractMix	Ours2 + TractMix
All	Dice	0.008	0.587	0.452	**0.645**	**0.694**	**0.715**	**0.728**
	d	19.116	2.881	2.121	1.529	0.676	0.241	-
	p	***	***	***	***	***	**	-
	RVD	0.991	0.242	0.603	**0.398**	**0.326**	**0.305**	**0.277**
	d	10.168	0.333	2.118	1.285	0.504	0.287	-
	p	***	*	***	***	***	**	-
HC	Dice	0.009	0.610	0.463	**0.667**	**0.712**	**0.732**	**0.743**
	RVD	0.990	0.241	0.583	**0.358**	**0.294**	**0.275**	**0.252**
AD	Dice	0.008	0.565	0.441	**0.622**	**0.675**	**0.698**	**0.712**
	RVD	0.992	0.243	0.623	**0.439**	**0.358**	**0.335**	**0.301**

Table 10
The means of the average Dice coefficients and average RVD values of the novel WM tracts for the private dataset when the selection of the novel WM tracts varied. Our results are highlighted in bold.

Selection		Baseline	Atlas FSL	FT	Ours1	Ours2	Ours1 + TractMix	Ours2 + TractMix
One	Dice	0.008	0.602	0.520	**0.687**	**0.713**	**0.720**	**0.717**
	RVD	0.989	0.267	0.545	**0.344**	**0.286**	**0.284**	**0.280**
Two	Dice	0.000	0.589	0.372	**0.540**	**0.662**	**0.670**	**0.682**
	RVD	0.985	0.303	0.729	**0.538**	**0.366**	**0.359**	**0.346**

(see Appendix A), possibly because the segmentation accuracy is already good and close to the upper bound even without TractMix. Note that as brains are generally located at the center of dMRI scans, we do not perform image registration before image mixing, which would require nontrivial interpolation of multiple fiber orientations. Although there can be misalignment between the annotated images that leads to less realistic synthetic data, the generated training samples can still be beneficial to network training. This is in agreement with previous observations that there is a tradeoff between the authenticity and diversity of the samples produced by data augmentation (Gontijo-Lopes et al., 2021).

Our method was evaluated using the original HCP dataset, the clinical quality data generated from the HCP dataset, and the private dataset. Compared with the results on the original HCP dataset, the segmentation quality is decreased for the clinical quality data, as indicated by the lower Dice coefficients and higher RVD values (e.g., compare Tables 7 and 8 with Figs. 4 and 5). This is possibly due to the reduced image quality, which increases the segmentation difficulty. This observation is also consistent with previous works (Lu et al., 2021; Wasserthal et al., 2018), and more training data is desired for datasets with lower image quality (Lu et al., 2021). In addition, although the image quality of the private dataset is better than the generated clinical quality data in terms of spatial resolution and the number of diffusion gradients, the segmentation performance (the Dice coefficient and RVD) on the private dataset is not better than the performance on the clinical quality data. This is likely due to the domain shift caused by the difference between the HCP dataset that was used for pretraining and the target private dataset, where the feature extraction for the HCP dataset may not be fully suitable for the private dataset. This domain shift may also explain the observation that for the private dataset CNN-based approaches did not preserve the volume of the tracts as well as the registration-based method Atlas FSL, where the smoothness constraint during registration could regularize the variations of tract volumes. However, this regularization does not guarantee the correctness of the spatial coverage of the tracts, and the Dice coefficient of the proposed method exceeds that of Atlas FSL by a large margin for the private dataset, indicating that the proposed method is still much more accurate than Atlas FSL. Note that the data quality of the private dataset is close to that of the original high-quality HCP dataset. If both domain shift exists and data quality is reduced, we expect that the segmentation problem becomes even more challenging. The domain shift problem can be further explored in future work to improve the segmentation performance. For example, dMRI harmonization (Mirzaalian et al., 2016) can be taken into consideration.

In this work, we use the TractSeg architecture as the backbone network, which performs volumetric WM tract segmentation and has achieved state-of-the-art segmentation performance. Our method may also be integrated with more advanced segmentation networks. For example, the integration of transformer with U-net has been shown to improve the performance for several medical image segmentation tasks (Chen et al., 2021). Similar improvement can be made to the TractSeg structure, and our method can be integrated with the improved network. The proposed method may also be applicable to CNN-based WM tract segmentation methods that classify fiber streamlines, such as Zhang et al. (2020), where the last layer can be better initialized with the knowledge learned for existing WM tracts.

In a related work (Lu et al., 2021), a self-supervised learning approach is developed to segment WM tracts with limited annotations by exploiting a large amount of unlabeled data. This method and the proposed approach address the problem of scarce annotations under different settings. The proposed approach performs the segmentation of novel WM tracts given a model that segments existing WM tracts, whereas the method in Lu et al. (2021) segments WM tracts without needing the information about segmenting existing WM tracts but with a sufficient number of unannotated scans. Depending on the resources that are accessible, one of these two approaches can be selected to segment WM tracts. In addition, it is possible to integrate the proposed method with Lu et al. (2021) when both a model for segmenting existing WM tracts and a large amount of unannotated data are available, and this integration can be investigated in future work.

In transfer learning, the pretrained model can have knowledge that is irrelevant to or even misleads the target task. During transfer learning, such knowledge may also be transferred to the target task, which is referred to as negative transfer and degrades the performance of the target task (Wang et al., 2019). Thus, it is desirable to regularize the transfer learning process to reduce the effect of negative transfer. However, in the proposed approach no explicit suppression of negative transfer is incorporated yet. Future work could explore the integration of negative transfer suppression (Chen et al., 2019) into the few-shot segmentation of novel WM tracts.

Because TractMix combines two annotated images for generating new training samples, one limitation of TractMix is that it requires at least two annotated scans and cannot be applied to the one-shot setting. Also, the number of possible synthetic samples decreases when fewer novel WM tracts are of interest. However, we have shown in the experiments that even without TractMix, the proposed transfer learning strategy already leads to improved segmentation accuracy, and when more than one annotated scans are available with multiple tracts of interest, the proposed data augmentation approach can introduce additional benefits. Future work could further explore data augmentation strategies that can be performed even for one-shot segmentation of a single novel WM tract.

In the proposed method, we assume that the logits of existing WM tracts can inform the prediction of novel WM tracts based on the observation that WM tracts can co-occur as crossing or overlapping tracts in a large number of voxels, and this assumption leads to the proposed improved fine-tuning strategy. Our fine-tuning strategy may be applied to other segmentation tasks with similar characteristics, where existing classes of anatomical structures correlate with the novel classes. Hierarchical brain parcellation can be an example, where a pretrained model performs coarse parcellation of the brain into the cortical gray matter, white matter, subcortical structures, and cerebrospinal fluid, and the target task is to parcellate the brain into fine-grained regions (such as different gyri and sulci) with only a few annotations based on the pretrained model.

5. Conclusion

We have proposed a transfer learning approach to few-shot segmentation of novel WM tracts with the knowledge learned from the segmentation of existing WM tracts. Unlike classic fine-tuning, we not only exploit the information in the pretrained feature extraction layers, but also take advantage of the learned knowledge in the task-specific layer for segmenting existing WM tracts. The incorporation of this knowledge allows better initialization for the network that segments novel WM tracts. In addition, a simple yet effective data augmentation strategy TractMix is developed to better exploit the information in the few annotated scans during the knowledge transfer, where synthetic training images are generated with tract-aware image mixing. The proposed method was evaluated on brain dMRI scans from public and private datasets under various experimental settings, and the results indicate that our method improves the performance of segmenting novel WM tracts in the few-shot setting.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Qi Lu: Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Wan Liu:** Methodology, Software, Validation, Writing – original draft. **Zhizheng Zhuo:** Methodology, Validation, Investigation, Data curation. **Yuxing Li:** Investigation, Data curation. **Yunyun Duan:** Investigation, Data curation, Supervision. **Pinnan Yu:** Investigation, Data curation. **Liying Qu:** Investigation, Data curation. **Chuyang Ye:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Formal analysis, Visualization, Funding acquisition. **Yaou Liu:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Formal analysis, Funding acquisition.

Acknowledgments

Chuyang Ye is supported by the Beijing Municipal Natural Science Foundation (L192058). Yaou Liu is supported by the National Natural Science Foundation of China (81870958 & 81571631), the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (JQ20035), and the Special Fund of the Pediatric Medical Coordinated Development Center of Beijing Hospitals Authority (XTYB201831). The HCP dataset was provided by the Human Connectome Project, WU-Minn Consortium and the McDonnell Center for Systems Neuroscience at Washington University.

Appendix A. Segmentation accuracy of the proposed method achieved with and without TractMix for the HCP dataset

In this appendix, we present and compare the results of the proposed method on the HCP dataset achieved with and without TractMix. In particular, the mean value of the average Dice coefficients of the novel WM tracts is summarized in Table A:1 for each experimental setting in Section 3.2. Here, HQ and CQ represent the experiments on the original high-quality and the generated clinical quality scans, respectively. Note that TractMix was only applicable with more than one annotated scans. The results show that for the experiments on the HCP dataset, where there was no domain shift between the scans used for segmenting existing and novel WM tracts, the performance of the proposed method achieved with or without TractMix is similar.

Table A:1
The means of the average Dice coefficients of the novel WM tracts of the proposed method achieved with and without TractMix for the HCP dataset. For convenience, we refer to the experiments on the original high-quality and the generated clinical quality data as HQ and CQ, respectively. Selection All refers to the use of all 12 novel WM tracts.

Data	Annotated training scans	Selection	Ours1	Ours2	Ours1 + TractMix	Ours2 + TractMix
HQ	5	All	0.811	0.812	0.812	0.812
HQ	3	All	0.807	0.808	0.809	0.809
HQ	3	One	0.824	0.826	0.828	0.827
HQ	3	Two	0.821	0.821	0.821	0.821
CQ	5	All	0.757	0.761	0.761	0.763
CQ	3	All	0.758	0.764	0.766	0.769

References

Andersson, J.L., Skare, S., Ashburner, J., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20 (2), 870–888.

Andersson, J.L., Sotiropoulos, S.N., 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 125, 1063–1078.

Banihashemi, L., Peng, C.W., Verstynen, T., Wallace, M.L., Lamont, D.N., Alkhars, H.M., Yeh, F.-C., Beeney, J.E., Aizenstein, H.J., Germain, A., 2021. Opposing relationships of childhood threat and deprivation with stria terminalis white matter. Hum Brain Mapp 42 (8), 2445–2460.

Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. In vivo fiber tractography using DT-MRI data. Magn Reson Med 44 (4), 625–632.

Bazin, P.-L., Ye, C., Bogovic, J.A., Shiee, N., Reich, D.S., Prince, J.L., Pham, D.L., 2011. Direct segmentation of the major white matter tracts in diffusion tensor images. Neuroimage 58 (2), 458–468.

Bryant, L., McKinnon, E.T., Taylor, J.A., Jensen, J.H., Bonilha, L., de Bezenac, C., Kreilkamp, B.A., Adan, G., Wieshmann, U.C., Biswas, S., Marson, A.G., Keller, S.S., 2021. Fiber ball white matter modeling in focal epilepsy. Hum Brain Mapp 42, 2490–2507.

Burton, P.R., Murtagh, M.J., Boyd, A., Williams, J.B., Dove, E.S., Wallace, S.E., Tasse, A.-M., Little, J., Chisholm, R.L., Gaye, A., Hveem, K., Brookes, A.J., Goodwin, P., Fistein, J., Bobrow, M., Knoppers, B.M., 2015. Data safe havens in health research and healthcare. Bioinformatics 31 (20), 3241–3248.

Chandio, B.Q., Risacher, S.L., Pestilli, F., Bullock, D., Yeh, F.-C., Koudoro, S., Rokem, A., Harezlak, J., Garyfallidis, E., 2020. Bundle analytics, a computational framework for investigating the shapes and profiles of brain pathways across populations. Sci Rep 10 (1), 1–18.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, X., Wang, S., Fu, B., Long, M., Wang, J., 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In: Advances in Neural Information Processing Systems, pp. 1908–1918.

Cook, P.A., Zhang, H., Avants, B.B., Yushkevich, P., Alexander, D.C., Gee, J.C., Ciccarelli, O., Thompson, A.J., 2005. An automated approach to connectivity-based partitioning of brain structures. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 164–171.

Feng, R., Zheng, X., Gao, T., Chen, J., Wang, W., Chen, D.Z., Wu, J., 2021. Interactive few-shot learning: limited supervision, better medical image segmentation. IEEE Trans Med Imaging 40 (10), 2575–2588.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189.

Garyfallidis, E., Côté, M.-A., Rheault, F., Sidhu, J., Hau, J., Petit, L., Fortin, D., Cunanne, S., Descoteaux, M., 2018. Recognition of white matter bundles using local and global streamline-based registration and clustering. Neuroimage 170, 283–295.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., Zoph, B., 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2918–2928.

Ginsburger, K., Matuschke, F., Poupon, F., Mangin, J.-F., Axer, M., Poupon, C., 2019. MEDUSA: a GPU-based tool to create realistic phantoms of the brain microstructure using tiny spheres. Neuroimage 193, 10–24.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124.

Gontijo-Lopes, R., Smullin, S., Cubuk, E.D., Dyer, E., 2021. Tradeoffs in data augmentation: An empirical study. In: International Conference on Learning Representations.

Hula, W.D., Panesar, S., Gravier, M.L., Yeh, F.-C., Dresang, H.C., Dickey, M.W., Fernandez-Miranda, J.C., 2020. Structural white matter connectometry of word production in aphasia: an observational study. Brain 143 (8), 2532–2544.

Jaimes, C., Machado-Rivas, F., Afacan, O., Khan, S., Marami, B., Ortinau, C.M., Rollins, C.K., Velasco-Annis, C., Warfield, S.K., Gholipour, A., 2020. In vivo characterization of emerging white matter microstructure in the fetal brain in the third trimester. Hum Brain Mapp 41, 3177–3185.

Jeurissen, B., Tournier, J.-D., Dhollander, T., Connelly, A., Sijbers, J., 2014. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. Neuroimage 103, 411–426.

Kingma, D., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Li, A., Luo, T., Xiang, T., Huang, W., Wang, L., 2019. Few-shot learning with global class representations. In: International Conference on Computer Vision, pp. 9715–9724.

Li, B., de Groot, M., Steketee, R.M., Meijboom, R., Smits, M., Vernooij, M.W., Ikram, M.A., Liu, J., Niessen, W.J., Bron, E.E., 2020. Neuro4Neuro: a neural network approach for neural tract segmentation using large-scale population-based diffusion imaging. Neuroimage 218, 116993.

Lifchitz, Y., Avrithis, Y., Picard, S., Bursuc, A., 2019. Dense classification and implanting for few-shot learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9258–9267.

Lu, Q., Li, Y., Ye, C., 2021. Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. Med Image Anal 72, 102094.

Lu, Q., Ye, C., 2021. Knowledge transfer for few-shot segmentation of novel white matter tracts. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 216–227.

MacNiven, K.H., Leong, J.K., Knutson, B., 2020. Medial forebrain bundle structure is linked to human impulsivity. Sci Adv 6 (38), eaba4788.

Maier-Hein, K.H., Neher, P.F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W.E., Glass, J.O., Chen, D.Q., Feng, Y., Gao, C., Wu, Y., Ma, J., Renjie, H., Li, Q., Westin, C.-F., Deslauriers-Gauthier, S., Ocegueda González, J.O., Paquette, M., St-Jean, S., Girard, G., Rheault, F., Sidhu, J., Tax, C.M.W., Guo, F., Mesri, H.Y., Dávid, S., Froeling, M., Heemskerk, A.M., Leemans, A., Boré, A., Pinsard, B., Bedetti, C., Desrosiers, M., Brambati, S., Doyon, J., Sarica, A., Vasta, R., Cerasa, A., Quattrone, A., Yeatman, J., Khan, A.R., Hodges, W., Alexander, S., Romascano, D., Barakovic, M., Auria, A., Esteban, O., Lemkaddem, A., Thiran, J.-P., Cetingul, H.E., Odry, B.L., Mailhe, B., Nadar, M.S., Pizzagalli, F., Prasad, G., Villalon-Reina, J.E., Galvis, J., Thompson, P.M., Requejo, F.D.S., Laguna, P.L., Lacerda, L.M., Barrett, R., Dell'Acqua, F., Catani, M., Petit, L., Caruyer, E., Daducci, A., Dyrby, T.B., Holland-Letz, T., Hilgetag, C.C., Stieltjes, B., Descoteaux, M., 2017. The challenge of mapping the human connectome based on diffusion tractography. Nat Commun 8 (1), 1–13.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C.E., Morey, R.A., Flashman, L.A., George, M.S., McAllister, T.W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R.D., Coleman, M.J., Kubicki, M., Westin, C.F., Stein, M.B., Shenton, M.E., Rathi, Y., 2016. Inter-site and inter-scanner diffusion MRI data harmonization. Neuroimage 135, 311–323.

Mueller, B.A., Lim, K.O., Hemmy, L., Camchong, J., 2015. Diffusion MRI and its role in neuropsychology. Neuropsychol Rev 25 (3), 250–271.

O'Donnell, L.J., Pasternak, O., 2015. Does diffusion MRI tell us anything about the white matter? an overview of methods and pitfalls. Schizophr. Res. 161 (1), 133–141.

O'Donnell, L.J., Westin, C.-F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans Med Imaging 26 (11), 1562–1575.

Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J.T., Miller, M.I., van Zijl, P.C., Albert, M., Lyketsos, C.G., Woods, R., Toga, A.W., Pike, G.B., Neto, P.R., Evans, A., Mazziotta, J., Mori, S., 2009. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. Neuroimage 46 (2), 486–499.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035.

Poulin, P., Jörgens, D., Jodoin, P.-M., Descoteaux, M., 2019. Tractography and machine learning: current state and open challenges. Magn Reson Imaging 64, 37–48.

Qin, Y., Li, Y., Zhuo, Z., Liu, Z., Liu, Y., Ye, C., 2021. Multimodal super-resolved *q*-space deep learning. Med Image Anal 71, 102085.

Ratnarajah, N., Qiu, A., 2014. Multi-label segmentation of white matter structures: application to neonatal brains. Neuroimage 102, 913–922.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C., 2020. 'Squeeze & excite' guided few-shot segmentation of volumetric images. Med Image Anal 59, 101587.

Sotiropoulos, S.N., Jbabdi, S., Xu, J., Andersson, J.L., Moeller, S., Auerbach, E.J., Glasser, M.F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D.A., Yacoub, E., Lenglet, C., Van Essen, D.C., Ugurbil, K., Behrens, T.E.J., 2013. Advances in diffusion MRI acquisition and processing in the human connectome project. Neuroimage 80, 125–143.

Stieltjes, B., Brunner, R.M., Fritzsche, K., Laun, F., 2013. Diffusion Tensor Imaging: Introduction and Atlas. Springer Science & Business Media.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35 (5), 1299–1312.

Thiebaut de Schotten, M., ffytche, D.H., Bizzi, A., Dell'Acqua, F., Allin, M., Walshe, M., Murray, R., Williams, S.C., Murphy, D.G., Catani, M., 2011. Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography. Neuroimage 54 (1), 49–59.

Toescu, S.M., Hales, P.W., Kaden, E., Lacerda, L.M., Aquilina, K., Clark, C.A., 2021. Tractographic and microstructural analysis of the dentato-rubro-thalamo-cortical tracts in children using diffusion MRI. Cerebral Cortex 31, 2595–2609.

Tournier, J.-D., Calamante, F., Connelly, A., 2007. Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. Neuroimage 35 (4), 1459–1472.

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.-H., Connelly, A., 2019. MRTrix3: a fast, flexible and open software framework for medical image processing and visualisation. Neuroimage 202, 116137.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Wu-Minn HCP Consortium, 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79.

Vanderweyen, D.C., Theaud, G., Sidhu, J., Rheault, F., Sarubbo, S., Descoteaux, M., Fortin, D., 2020. The role of diffusion tractography in refining glial tumor resection. Brain Structure and Function 225, 1413–1436.

Veraart, J., Raven, E.P., Edwards, L.J., Weiskopf, N., Jones, D.K., 2021. The variability of MR axon radii estimates in the human white matter. Hum Brain Mapp 42 (7), 2201–2213.

Wang, Z., Dai, Z., Póczos, B., Carbonell, J., 2019. Characterizing and avoiding negative transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11293–11302.

Wassermann, D., Makris, N., Rathi, Y., Shenton, M., Kikinis, R., Kubicki, M., Westin, C.-F., 2016. The white matter query language: a novel approach for describing human white matter anatomy. Brain Structure and Function 221 (9), 4705–4721.

Wasserthal, J., Neher, P.F., Hirjak, D., Maier-Hein, K.H., 2019. Combined tract segmentation and orientation mapping for bundle-specific tractography. Med Image Anal 58, 101559.

Wasserthal, J., Neher, P.F., Maier-Hein, K.H., 2018. TractSeg - Fast and accurate white matter tract segmentation. Neuroimage 183, 239–253.

Wu, Y., Hong, Y., Ahmad, S., Lin, W., Shen, D., Yap, P.-T., 2020. Tract dictionary learning for fast and robust recognition of fiber bundles. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 251–259.

Ye, C., Yang, Z., Ying, S.H., Prince, J.L., 2015. 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.

Yeatman, J.D., Dougherty, R.F., Myall, N.J., Wandell, B.A., Feldman, H.M., 2012. Tract profiles of white matter properties: automating fiber-tract quantification. PLoS ONE 7 (11), e49790.

Yeghiazaryan, V., Voiculescu, I.D., 2018. Family of boundary overlap metrics for the evaluation of medical image segmentation. J. Med. Imaging 5 (1), 1–19.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032.

Zhang, F., Daducci, A., He, Y., Schiavi, S., Seguin, C., Smith, R., Yeh, C.-H., Zhao, T., O'Donnell, L.J., 2021. Quantitative mapping of the brain's structural connectivity using diffusion MRI tractography: a review. arXiv preprint arXiv:2104.11644.

Zhang, F., Karayumak, S.C., Hoffmann, N., Rathi, Y., Golby, A.J., O'Donnell, L.J., 2020. Deep white matter analysis (deepWMA): fast and consistent tractography segmentation. Med Image Anal 65, 101761.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations.