Lab Report (WiSe 2023/24)
Visualization and Medical Image Analysis
March 8, 2024

The Impact of Fiber Orientation Features on Direct White Matter Tract Segmentation

Bertan Karacora
University of Bonn, Department of Computer Science
bertan.karacora@uni-bonn.de

Abstract

Tract segmentation is a fundamental step in visualizing and analyzing white matter tracts in the human brain. Traditional methods often rely on tractography, the reconstruction of fiber tract pathways as 3D streamlines in diffusion-weighted Magnetic Resonance Imaging (dMRI). However, a streamline-based segmentation requires either subject-specific manual annotation or complex processing pipelines, in addition to tractography itself. This project is focused on TractSeg, a direct approach that treats the problem as a conventional image segmentation task. TractSeg, as proposed by Wasserthal et al., employs a convolutional neural network (CNN) to achieve a voxel-wise fiber tract identification in fields of fiber orientation distribution function (fODF) peaks. Despite promising results, utilizing fODF peaks as input features to the segmentation model may restrict its performance, since peaks are limited in their expressiveness and susceptible to interference. This project explores the impact of fODF peaks and alternative fiber orientation features on direct tract segmentation methods, using TractSeg as a well-acknowledged representative. Experimental results indicate a slight but notable performance improvement when the model is applied to a more comprehensive representation of the voxel-wise fODF as a fourth-order tensor.

1. Introduction

Accurately identifying and delineating white matter tracts in the human brain is crucial for exploring its structure and functionality. Numerous studies showed that changes in the microstructure of distinct fiber tracts are commonly associated with specific brain diseases and disorders [Fie08, FF16]. Quantitative properties of these tracts can be used to perform statistical analysis within segmented regions or along their pathways, which may help to identify and investigate healthy and abnormal brain morphology [JDML19, BCD*11]. Both pure segmentation and localized analysis are critical in various clinical applications and in neuroscience [AJBTdS19]. Prominent examples include neurosurgery [YYPC21], connectomics [BS09, MHNH*17a], and studies in neurodegenerative diseases [YSA*09].

Traditionally, tract segmentation is achieved by the manual dissection of streamlines [CHPJ02], which in turn are computed via dMRI-based tractography [LBBL*86, JDML19]. As this typically involves segmenting meaningful regions-of-interest (ROIs) by hand, this process is time-consuming, requires deep expert knowledge, and lacks reproducibility [SRP*21]. In recent years, different methods have been established to automate parts of this task [GASZ23]. However, their implementations entail large pipelines of complex processing steps involving atlas construction, image registration, parcellation, and clustering.

TractSeg, as proposed by Wasserthal et al. [WNMH18a], constitutes a direct approach to tract segmentation. Instead of first constructing streamlines and then classifying them, TractSeg uses a convolutional neural network (CNN) to identify voxels belonging to known major fiber tracts. Compared with a dataset of manually validated reference segmentations, the authors demonstrated that TractSeg consistently outperforms six streamline-based segmentation methods both on high-quality and clinical data, with regards to the Dice metric.

For predicting the location, shape, and structure of white matter tracts, direct segmentation methods such as TractSeg may take advantage of representing dMRI measurements in a more condensed and robust manner, particularly in case of high angular resolution. To this end, constrained spherical deconvolution (CSD) [TCC07] has become a popular tool for decomposing the directional diffusion intensities in terms of the orientations and volumes of the underlying fiber tract compartments. The resulting information is contained in the fiber orientation distribution function (fODF), which characterizes the likelihood of arbitrary fiber orientations at a specific voxel. Therefore, the fODF describes properties of the local microstructure rather than the emerging diffusion signal.

In TractSeg, three peaks per voxel, i.e., local maxima of the fODF, serve as input features to the CNN. Intended to indicate the principal fiber directions prevalent in each voxel, simply taking local maxima comes with several limitations. Most significantly, the discrete nature of the peaks, as well as the restriction to only three of them, may oversimplify the fODF, leading to a loss of information, especially in regions with a complex fiber architecture. Moreover, peak vectors represent only one of two equivalent orientations, disregarding the symmetry of the fODF. Finally, peak directions are subject to interference from contributions of crossing or branching fibers [SS08]. As such, peaks do not characterize individual fiber tracts but a mixture of them.

In this project, we address the limitations associated with fODF peaks as input features in TractSeg. Towards this goal, we adapted TractSeg's segmentation model and training pipeline for its application to more comprehensive descriptors of the local fiber orientations. More precisely, our modified implementation takes fields of higher-order fODF tensors as input, estimated by a variation of multi-shell multi-tissue CSD [JTD*14, ALGS16, ALGS17]. At the cost of a marginally increased network complexity, this adds additional information into the model and resolves the inherent directional ambiguity of the peaks.

Supplementary to their results, Wasserthal et al. [WNMH18a] provide the source code of TractSeg. Built upon this foundation, we created a framework for conducting experiments with different input features, including fODF tensors and approximations thereof. Various changes to feature processing, data augmentation, and hyperparameters have been made to account for how these features are encoded and how they transform, as well as any hardware constraints we encountered during this project. Finally, based on experimental results, we discuss the capabilities and limitations of the considered feature inputs. Notably, replacing the peaks with fODF tensors results in a slight but significant improvement of approx. 0.3 Dice points.

From a more general perspective, this work is meant to investigate the effect of fiber orientation features on direct segmentation methods, considering TractSeg as an important archetype. In order to do so, the remainder of this report elaborates on the essential ideas, methods, and experimental results within this project, supplemented with own considerations and interpretations.

2. Related work

2.1. Streamline-based tract segmentation

White matter tracts consist of myelinated nerve fibers that share certain characteristics such as spatial location, connectivity to gray matter regions, and neurological function [MSP18]. As currently the only non-invasive in-vivo modality, dMRI [LBBL*86] allows to study white matter microstructure by measuring local diffusion properties. Complementary, tractography [BPP*00, JDML19] aims to extrapolate the local information and estimate 3D streamlines that represent fiber trajectories. Since tracts are easier recognizable at a global scale, experts are able to virtually dissect the fibers by specifying ROIs that the tracts are known to terminate in or traverse through [CHPJ02]. Notably, streamline-based segmentation is intrinsically constrained by the limitations of tractography [MHNH*17b].

To reduce or even eliminate the need for manual annotation, many automated methods leverage atlases of predefined ROIs in combination with image registration. Hansen at al. [HYL*21] provide an overview of widely-used atlases, while presenting their own, constructed from a large population using state-of-the-art tractography. Despite advancements, atlases may not capture the variability of appropriate ROIs across subjects, age, and pathologies [GASZ23].

As an alternative, streamline clustering techniques may achieve a white matter parcellation based on other properties than ROI-traversal. A review of these and more is given in [ZDH*22]. However, segmenting the clusters into anatomically meaningful tracts again requires atlases or other means of supervision.

2.2. Direct tract segmentation

Given dMRI images of a subject, direct tract segmentation methods perform a voxel-wise binary classification, separately for each tract [WNMH18a]. In contrast to streamline-based segmentation, direct approaches often rely entirely on local features, aiming to avoid the errors, assumptions, and complexity of modeling the white matter structure via tractography. Examples from the past have tried to incorporate anatomical priors in Markov Random Fields [BYB*11], in template matching [ESS*09], or in k-nearest neighbors classification [RQ14]. Nevertheless, these methods often result in lower quality segmentations than streamline-based ones.

Recently, a promising potential is seen with supervised deep learning approaches like TractSeg [WNMH18a], improving both accuracy and efficiency in direct tract segmentation. For a review of deep learning methods, the reader may refer to the work of Ghazi et al. [GASZ23].

A number of studies aimed to improve specific aspects of TractSeg. Tchetchenian et al. [TZZ*23] presented a systematic search for the optimal U-Net [RFB15] architecture and found that the UNet3+ [HLT*20] outperforms TractSeg's mean Dice score by 0.004. Several works explored the efficient fine-tuning of TractSeg for unseen white matter tracts in few- or one-shot settings [LY21, LLZ*22, LZLY23]. Kebiri et al. [KGCK23] randomly sampled six gradient directions from a single shell and utilized a U-Net to segment tracts with a similar accuracy as TractSeg, based only on raw dMRI measurements.

2.3. White matter feature descriptors in deep learning

As one of its greatest strengths, deep learning does not necessitate hand-crafted feature engineering and is capable to learn from raw data. However, in some cases, an initial feature mapping may still be advantageous to reduce dimensionality, alleviate data acquisition effects, and concentrate the network attention on properties more relevant than others.

Geometric descriptors are common in streamline-based segmentation [GASZ23], e.g., curvature, torsion, and Euclidean distances to an atlas of landmarks [LBF*18]. More explicit streamline features may also be considered. For instance, Zhang et al. [ZKH*20] constructed so called FiberMaps from explicit point coordinates of 15 points along each streamline and achieved a labeling with CNNs.

To give some examples of direct segmentation, Li et al. [LDGS*20] utilized diffusion tensor images as input to a CNN. However, they did not compare their results with TractSeg, since they were not able to replicate its performance on their own dataset. They report that fODF peaks led to a lower performance than diffusion tensors and adding T1-weighted MRI images showed no improvements. Dong at al. [DYPW19] also closely follow the methodology of TractSeg but combine fODF peaks with a structural tensor of T1-weighted images. In their reproduction of TractSeg's results, the model achieved a mean Dice of 0.827 over all subjects, whereas the multi-modal approach reached 0.832. Pomiecko et al. [PSF*19] show that a median Dice of approx. 0.7 can be attained using quantitative anisotropy maps alone.

3. Methods

Because of its important role as one of the earliest and most accurate direct segmentation methods, this project is built upon the methodology of TractSeg [WNMH18a]. We adopted TractSeg's encoder-decoder CNN architecture and replace the input fODF peaks with alternative features of local fiber orientation. This section explains the computation of those features, the design of the segmentation model, methods for its training, and mechanisms for its usage to predict 3D tract segmentations during inference. Fig. 1 provides an overview of the training and inference processes.

3.1. Fiber orientation features

With diffusion tensor imaging (DTI) [BML94] being capable to estimate only a single fiber orientation per voxel, high angular resolution diffusion imaging (HARDI) is of special interest in tract segmentation. Consequently, high-quality datasets contain a large set of volumes for each subject, one per applied gradient direction and diffusion weighting. Although this data contains all the available information, it is impractical to input into a CNN in its raw form. Not only would the training process require extensive amounts of memory transfer, but the entire network would need to be re-trained for every change in the data acquisition scheme.

Therefore, it is desirable to extract meaningful features even before the CNN receives the input. If suitable features can be computed in variable dMRI setups (e.g., with different scanners, shells, resolutions, and gradient selections), then an initial feature extraction step might impose some degree of normalization, potentially mitigating acquisition-specific biases and making the model applicable to a wider range of situations. In this project, we focus on input features derived from local fiber orientations.

To evaluate how fiber orientation features affect the tract segmentation performance, we consider three different representations: (a) fODFs as fourth-order tensors, (b) peaks of the fODFs, (c) low-rank approximations of the fODF tensors. All of these leverage multi-shell multi-tissue CSD (MSMT-CSD) [JTD*14] for the fODF estimation in each voxel. Further inputs are tested in section 4.3.

3.1.1. fODF tensors

Arguably, dMRI measurements in the white matter arise from the superimposed contributions of single fiber responses [BWJ*03]. Assuming that each response is identical and only depends on the fiber orientation, this implies that the desired local information is contained within the fODF, an antipodally symmetric and non-negative function of fiber orientation density. Mathematically, the spherical dMRI signal $S(\theta, \phi)$ may be modeled as the convolution of the fODF $F(\theta, \phi)$ with an axially symmetric fiber response kernel $R(\theta)$:

$$S(\theta, \phi) = F(\theta, \phi) * R(\theta), \tag{1}$$

where θ and ϕ are polar and azimuth angles, respectively [SS08]. Accordingly, the fODF can be estimated using a deconvolution algorithm. Despite several disadvantages and strong assumptions in this model, fODFs have proven to be very useful local descriptors of the white matter microstructure.

An fODF can be represented as a higher-order tensor $\mathcal{T}$. In this project, a fourth-order representation $\mathcal{T} \in \mathbb{R}^{3\times3\times3\times3}$ is chosen, describing an fODF as

$$F(\mathbf{v}) = \mathcal{T}(\mathbf{v}) = \sum_{i_1,i_2,i_3,i_4=1}^{3} \mathcal{T}_{i_1 i_2 i_3 i_4} v_{i_1} v_{i_2} v_{i_3} v_{i_4}, \tag{2}$$

where $\mathbf{v} \in \mathbb{S}^2$ is a vector on the unit sphere [SS08].

Typically, fODFs are encoded in a truncated basis of spherical harmonics. This formulation can be though of as a Fourier expansion defined in the fODF domain, i.e., the unit sphere manifold. The basis coefficients hold different degrees of significance with lower-order ones capturing more coarse features and higher-order ones finer details. Spherical harmonics not only simplify the deconvolution but also enable the efficient control over the level of detail. The mapping from the order-l tensor representation to spherical harmonics with a maximum degree of l amounts to only a change of basis, i.e., a nonsingular linear transformation, as shown explicitly in [SFG*14]. Because of that, no meaningful difference from our choice of basis is expected.

To compute the fODF tensors, we employ a variant of MSMT-CSD [ALGS16, ALGS17] and its implementation in Bonndit [AGMS24]. The fODF tensor form is super-symmetric [SS08], i.e., for any permutation σ of the l indices the following holds:

$$\mathcal{T}_{i_1,i_2,\ldots,i_l} = \mathcal{T}_{i_{\sigma(1)},i_{\sigma(2)},\ldots,i_{\sigma(l)}}. \tag{3}$$

Hence, in three spatial dimensions, it is sufficient to operate with $T \in \mathbb{R}^n$ including only $n = \frac{(3+4-1)!}{(3-1)!4!} = 15$ unique components instead of all $3^4 = 81$ ones.

3.1.2. fODF peaks

Peaks are 3D vectors gathered as local maxima of the fODF. Their directions may be interpreted as the principal fiber orientations in a voxel [WNMH18b], and their magnitude as corresponding voxel volume fractions [SPDDG21]. In TractSeg, three different fODF variants are computed, one from MSMT-CSD and two from standard CSD applied on single-shell data subsets at $b = 1000\mathrm{s/mm}^2$ with 30 and 12 gradients, respectively. During training, each sample is randomly chosen from these versions to make the model robust towards various CSD methods. However, within the scope of this project, single-shell settings are not considered. Both the fODF and peak estimation is based on routines from MRtrix3 [TSR*19,TCGC04] with a specified number of three peaks per voxel. If less fiber directions are identified, the remaining feature channels are set to zero.

3.1.3. Low-Rank fODF tensor approximation

Simply taking local maxima does not account for interference in the fODF. Consider a voxel containing two crossing fiber tracts with a small angle between them. If each tract contributes to an fODF

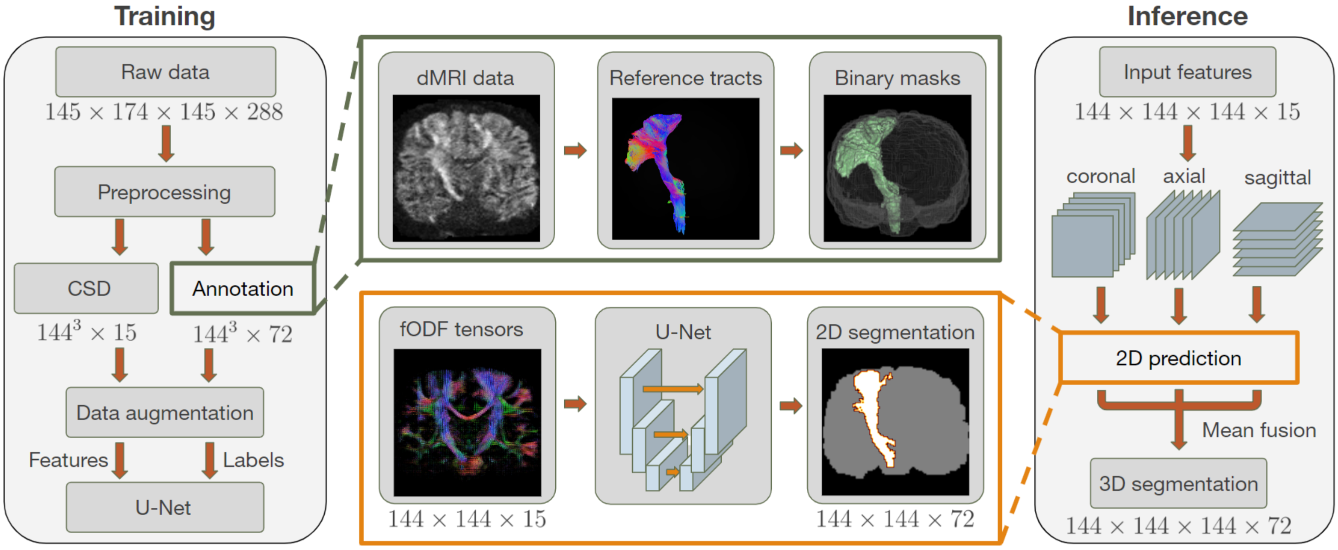

Figure 1: *Method overview. The supervised training requires input features and reference labels from preprocessed data. During inference, the model predicts segmentations in 2D slices from coronal, axial, and sagittal directions, which are fused to 3D by averaging. The numbers below the boxes indicate the data dimensions after each step, exemplary for the case of fODF tensor inputs.*

mode with a certain width, the modes might be shifted closer together or even be merged into a single one. Therefore, the principal fiber orientations may be better understood as the vectors $\mathbf{v}_i \in \mathbb{R}^3$, up to their signs, that characterize the best-fit low-rank approximation of the fODF tensor [SS08, SFG*14]. With λ_i denoting volume fractions, the following approximation can be obtained:

$$\mathcal{T}^{(r)} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i. \tag{4}$$

Bonndit [AGMS24] contains the algorithms used to compute fODF tensor approximations of symmetric rank $r = 3$. Finally, a concatenation of the vectors $\lambda_i^{0.25}\mathbf{v}_i$ is given into the segmentation model.

3.2. Segmentation model

The CNN architecture of TractSeg is based on the U-Net [RFB15], a widely used encoder-decoder network for image segmentation tasks. The U-Net comprises two stages. In the encoding part, explicit input data is compressed into a hierarchy of increasingly abstract feature maps The decoder, on the other hand, is designed to restore spatial information by upsampling the abstracted features and combining them with more detailed ones from earlier hidden layers. Built upon the U-Net, TractSeg adds deep supervision [LXG*15] to facilitate the back-propagation to earlier layers. Moreover, zero-padding is applied in each convolution step to ensure that the final output segmentation maintains the same resolution as the input image.

The segmentation model of TractSeg takes a 2D image of shape 144×144 as input, with 9 channels representing the coordinates of three 3D peak vectors. To accommodate alternative fiber orientation features as input, the channel width of the first convolutional layer is reconfigured with respect to the dimension of the respective feature domain. Subsequent layers have been kept identical. Fig. 2 depicts the network architecture for the tensor representation case.

As output, the model generates an image of the same resolution with 72 channels, each each corresponding to the probability map for a specific fiber tract. Notably, the final activation function employed is a sigmoid function. Unlike functions requiring voxel-wise probabilities to sum to one, such as softmax, the sigmoid function allows for independent channel probabilities. This enables multi-label segmentation, which is crucial to account for regions of crossing fiber tracts.

3.3. Inference

During inference, the model segments slices along the coronal, axial, and sagittal planes. Subsequently, a full 3D segmentation is constructed from fusing the three outputs per voxel and tract, either by taking the mean or using a second-level CNN. While a slight performance increase was observed for the latter approach, Wasserthal et al. recommended mean fusion, which is less complex and much faster. We chose to disregard the second CNN, because its training requires inputs containing $3 \cdot 72 = 216$ channels. Within this project, this would have led to an excessive increase of hardware requirements and training duration. Finally, the binary segmentation per tract is achieved by thresholding the voxel-wise mean of the three outputs at 0.5.

3.4. Training

Wasserthal et al. continued to improve TractSeg after its initial publication [WNMH18a] in subsequent works [WNMH18b, WNHMH19]. As part of their work, they publicly shared a Pytorch [PGM*19] implementation.

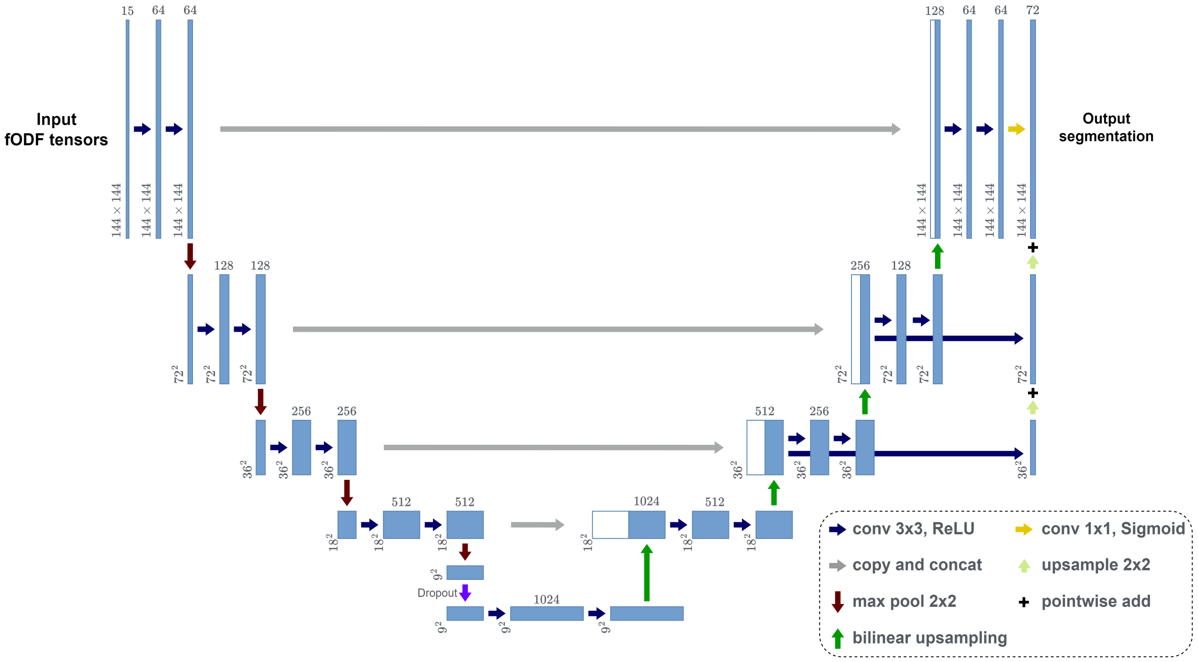

Figure 2: *CNN architecture. Blue and white boxes represent generated and copied feature maps, respectively. Numbers denote their dimensions. Network operations are illustrated by the colored arrows and symbols shown at the bottom right. To learn from different fiber orientation such as fODF tensors, the channel width of the first convolutional layer is adapted to fit the feature dimension.*

In this project, the training pipeline builds upon the most recent version of TractSeg. Our implementation extends TractSeg by a framework that provides scripts for feature extraction and label preparation, as well as experiment configurations and various utilities, e.g., visualizations and monitoring with Tensorflow's Tensorboard [AAB*16]. Moreover, automatic mixed precision [MNA*17] has been added to speed up the training process.

Like Wasserthal et al., we conduct our experiments on a dataset of 105 selected subjects from the Human Connectome Project (HCP) [VESB*13]. The HCP data has 1.25mm isotropic resolution and 288 channels including 270 gradient directions across three shells at $b \in \{1000, 2000, 3000\}[\mathrm{s/mm}^2]$, as well as 18 $b = 0\mathrm{s/mm}^2$ volumes. The minimal preprocessing pipeline described in [GSW*13] has already been applied to the data provided.

Wasserthal et al. utilized 5-fold cross-validation, dividing the population into folds of 21 subjects each. However, because of the long training time, we decided to use a fixed train-validation-test split instead, avoiding to train the model multiple times in each experiment. This allowed us to concentrate our resources on experiments with additional input features like the ones presented in section 4.3. In practice, we selected three of the folds for training, one for validation, and one for testing. After training has converged, the performance on the validation subset determines the best-performing epoch, from which the network weights are selected for the final evaluation on the test subset. For explorative purposes such as optimizing hyperparameters and data augmentation configurations, we evaluated the model performance on the validation subset.

For training supervision, Wasserthal et al. provided a carefully curated dataset containing high-quality tract dissections of 72 major white matter tracts for all 105 HCP subjects. A detailed description of how the reference tractographies have been created is given in the original publication of TractSeg [WNMH18a]. Within our extended framework, we mask the reference streamlines into binary volumes and concatenate them into a 72-channel 3D image To minimize memory transfer, we cropped both the labels and features to the subject's brain bounding box and padded them to 144 × 144 during training.

The CNN is trained using the binary cross-entropy loss

$$\mathrm{loss}(\hat{y}, y) = -\frac{1}{n}\sum_{i=1}^{n}(y[i], \log(\hat{y}[i]) + (1 - y[i])\log(1 - \hat{y}[i])), \quad (5)$$

where $n = 72$ is the number of tract labels, $y \in \{0, 1\}^n$ are the reference labels per subject per voxel and $\hat{y} \in [0, 1]^n$ is the model prediction.

We observed a satisfactory memory usage in all experiments using a batch size of 40, which amounts to 226 batches or 9040 slices per epoch. Generally, the model is trained until convergence, which typically took 180 to 230 epochs. Adamax [KB14] has been employed as optimizer. The initial learning rate is set to 0.001 and is scheduled to be reduced by one order of magnitude each time the validation loss does not decrease for at least 20 epochs. Originally in TractSeg, the axial slicing direction is chosen with a probability of 0.5, whereas coronal and sagittal slices are sampled with a probability of 0.25 each. Since no explanation for this was found, we replaced this by uniform sampling.

3.5. Data augmentation

TractSeg uses heavy data augmentation to increase the model robustness and prevent overfitting. We maintained most transformations and hyperparameters as in [WNHMH19].

Prior to any augmentation step, every slice is normalized (jointly over all channels) to zero mean and unit variance. Besides this, no further normalization in-between CNN layers is performed. Subsequently, the following transformations are applied to each slice in every batch, based on implementations in the Batchgenerators library [IJW*20]:

- Rotation by angle $\varphi \sim \mathcal{U}[-\frac{\pi}{4}, \frac{\pi}{4}]$. Peaks and tensors are rotated accordingly.
- Elastic deformation with $(\alpha, \sigma) \sim (\mathcal{U}[90, 120], \mathcal{U}[9, 11])$. A field of displacement vectors $d_{xy} \sim \mathcal{U}[-1, 1]^2$ is sampled, smoothed by a Gaussian filter with standard deviation σ and scaled by α.
- Displacement by $(\Delta x, \Delta y) \sim \mathcal{U}[-10, 10]$.
- Zooming by a scale $\lambda \sim \begin{cases} \mathcal{U}[0.9, 1] & \text{if Bernoulli(0.5) = 0} \\ \mathcal{U}[1, 1.5] & \text{else.} \end{cases}$
- Gaussian blurring with standard deviation $\sigma \sim \mathcal{U}[0, 1]$.
- Gaussian noise with mean and standard deviation $(\mu, \sigma) \sim (0, \mathcal{U}[0.05])$.

This augmentation strategy remained unchanged when the model was trained on either fODF peaks or rank-3 approximations of fODF tensors. However, several important adjustments have been introduced for the training on fODF tensors, as described in the following.

As mentioned before, a fourth-order fODF tensor $\mathcal{T}$ may be written in a vectorized form $T \in \mathbb{R}^n$ with only $n = 15$ components due to symmetry. Consider the symmetry-induced equivalence classes of index tuples of $\mathcal{T}$. Let m be the mapping from the indices of T to the corresponding equivalence class such that $\{T_i\} = \{\mathcal{T}_{i_1 i_2 i_3 i_4} | (i_1, i_2, i_3, i_4) \in m(i)\}$ for all indices i.

First off, to simulate the effect of Gaussian noise with $\mu = 0$ and a sampled but fixed standard deviation σ on the fODF tensor, we proceed as follows. For restoring symmetry, one could add the noise to all components of $\mathcal{T}$ and average the elements corresponding to each equivalence class. To simplify the operation and avoid the expansion into the full tensor form, we directly apply a scaled noise to T instead, which accounts for the multiplicities $|m(i)|$ of the components in $\mathcal{T}$. In other words, we add Gaussian noise with a differently scaled standard deviation $\frac{\sigma}{|m(i)|}$ separately to each component of T.

For keeping the local information consistent with their spatial arrangement, it is necessary to transform the fiber orientation features according to the coordinate transformation applied to their domain, i.e., the image slices [SFG*14]. In case of fODF peaks and rank-3 approximations, elastic deformation violates this principle. However, we chose to keep the elastic deformation augmentation for its empirically observed potential to prevent overfitting and improve the models accuracy in case of peaks and approximations as input. Conversely, we observed slightly worse segmentation performance if we apply elastic deformation to fields of fODF tensors and restrained from this augmentation step in that case.

Equally important, the fiber orientations need to be rotated according to the image slice rotation. For 3D vectors, this operation is straight-forward and already implemented in TractSeg. However, significant effort was dedicated to the efficient implementation of tensor rotation that prevents a drastic increase in memory requirements. To achieve this goal, it is crucial to exploit the tensor symmetry and avoid expanding the vectorized representation T to the full tensor $\mathcal{T}$.

Let $R \in \mathbb{R}^{3\times 3}$ represent the rotation matrix applied to the image slice. This matrix defines a change of basis in image space (MNI space in case of HCP data) and may be constructed from Euler angles with respect to the slicing direction. It is worth noting that the fODFs may be defined in a different coordinate system and R needs to be transformed accordingly. As shown in [SFG*14], the change of basis of a tensor may be expressed as the result of a multilinear matrix product

$$\mathcal{T}'_{i_1 i_2 i_3 i_4} = \sum_{j_1, j_2, j_3, j_4 = 1}^{3} R_{i_1 j_1} R_{i_2 j_2} R_{i_3 j_3} R_{i_4 j_4} \mathcal{T}_{j_1 j_2 j_3 j_4}. \tag{6}$$

In the reduced form, we computed the rotated tensor T' via a single matrix-vector multiplication as

$$T'_i = \sum_{j=1}^{n} A_{ij} \mathcal{T}_{[m(j)]} = \sum_{j=1}^{n} A_{ij} T_j, \tag{7}$$

where $[m(i)]$ denotes an arbitrary representative of the equivalence class $m(i)$. We construct matrix A according to equation 6 as

$$A_{ij} = \sum_{(j_1, j_2, j_3, j_4) \in m(j)} \prod_{k=1}^{l} R_{[m(i)]_k j_k}. \tag{8}$$

The construction of this matrix is required only once per image slice. The rotation of the fODF tensors is efficiently implemented in a vectorized fashion using Numpy [HMVDW*20].

4. Experiments and results

For the experimental evaluation, we established a fixed train-validation-test split of the 105 HCP subjects. In our setup, 63 subjects formed the training set, with an additional 21 subjects designated for model validation in each epoch. Therefore, 21 test subjects remained for the final evaluation. As a consequence, our results may not be directly comparable to those of Wasserthal et al. in [WNMH18a] and in [WNHMH19]. In their work, they employed a 5-fold cross-validation approach and reported the mean Dice score for the entire subject population. Although our evaluation, conducted on a subset of this population, may still provide indications regarding the consistency between TractSeg and our adaptations, we aimed to replicate the results of TractSeg, while bearing in mind the aforementioned changes to the training pipeline.

To evaluate the impact of fiber orientation features on the segmentation performance, we consider three main experiments. Naturally, our replication of TractSeg using fODF peaks as input features serves as the baseline. In the two other setups, we investigate the impact of replacing the peaks with forth-order tensor representations of the fODFs, and with rank-3 approximations of the fODF tensors, respectively. Moreover, we investigate the effect of concatenating the features with additional inputs, specifically, a T1-weighted MRI image and a partial volume map of the white matter.

An NVIDIA GeForce GTX 1080 GPU and an Intel Core i7-4930K CPU with 6 physical cores have been used for training. Convergence of the training loss is typically observed after approx. 28 to 42 hours, depending on the input feature dimension.

We adopted the Dice score [TH15] as the evaluation metric, a widely used performance measure for image segmentation that quantifies the overlap between prediction and reference. To calculate this score, we computed the Dice coefficient individually for each of the 72 reference tracts for every test subject. The final Dice score per test subject is obtained by averaging these tract-specific scores.

4.1. Quantitative evaluation

Figure 3 shows the segmentation accuracy for the three primary experiments. Our replication of TractSeg achieves a mean Dice score of approx. 0.8375 on the test subjects, which is consistent with the results reported by Wasserthal et al. [WNMH18a] for the entire subject population. The model trained and tested on fourth-order representations of the fODFs outperforms the baseline by approximately 0.3 Dice points on average. On the other hand, the rank-3 tensor approximations exhibit a slightly lower mean Dice score than the peaks. Generally, the relative distributions of Dice scores within the test subset are very similar to each other.

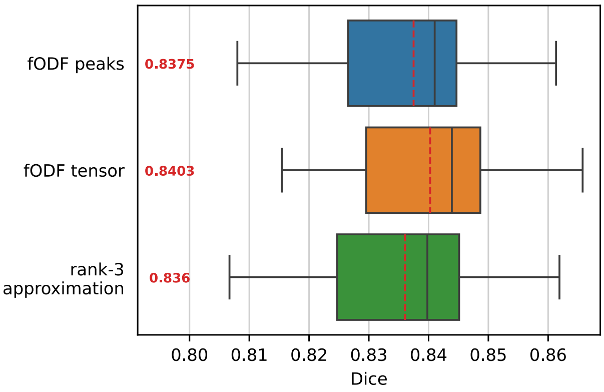

Figure 3: *Comparison of segmentation accuracy using different fiber orientation features. The box plot shows the distributions of Dice scores (mean over all 72 tracts per subject). The mean across all test subjects scores is indicated by the dashed red line and the red number on the left.*

A statistical analysis of the results is conducted using a significance level of $\alpha = 0.05$, considering p-values below this threshold as significant. Initially, the Friedman test [Fri37,Fri40] is employed to assess if there are significant differences across the segmentation performances with different inputs. This non-parametric test evaluates the joint distribution of rank sums for the models concerning their Dice scores for each of the 21 test subjects. We obtained the test statistic $\chi^2 \approx 29.81$, yielding a p-value of $p << 0.001$. Thus, we reject the null hypothesis of equally accurate models.

Subsequently, we conduct a Nemenyi post-hoc test [Nem63] for multiple comparisons. This test allows for pairwise evaluation of average ranks between models (ranks range from 1 to 3, with 3 being the best), while correcting the p-values with regard to the family-wise error. The fourth-order tensor representation resulted in a significant improvement of the Dice metric, if compared against the peaks ($p \approx 0.002$), and if compared against the tensor approximations ($p \approx 0.001$). The comparison between peaks and tensor approximations yields no validation of a significant difference ($p \approx 0.112$). A critical distance diagram [Dem06] in Fig. 4 serves as visualization.

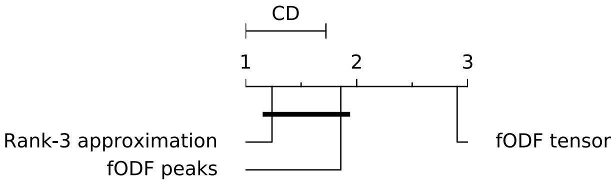

Figure 4: *Critical distance diagram for the Nemenyi post-hoc test. A difference in the performance ranks smaller than the critical distance (CD) indicates statistical significance (with $p < 0.05$).*

Fig. 5 illustrates the Dice scores for each tract individually. Overall, the model trained on fODF tensors consistently outperforms the others for almost all tracts. Variations in the segmentation performances among the trained models are most evident in inferior scores. The Fornix (FX) and the Commissure Anterior (CA) tracts exhibit the lowest Dice scores, which is consistent with the observations of Wasserthal et al. [WNMH18a,WNHMH19] In their own work, they suggest that this might be due to the thin tract shapes, in combination with incomplete reference segmentations of these tracts.

4.2. Qualitative evaluation

To assess the segmentation performance qualitatively, we chose a representative test subject (987983) whose subject-specific Dice scores are closest to the mean Dice score of all three experiments. Following Maier-Hein et al.'s categorization of tracts by empirical reconstruction difficulty [MHNH*17b], we focus on three tracts: the inferior occipito-frontal fascicle (IFO), the corticospinal tract (CST), and the anterior commissure (CA), in ascending order of difficulty level.

Fig. 6 depicts exemplary segmentation results corresponding to the considered fiber orientation features. While not many differences can be observed in 3D renderings, the blue and yellow colored voxels in the slices, indicating false positives and false negatives, respectively, may provide insights into each of the models segmentation capability. In addition, a T1-weighted MRI scan of the subject is shown for context. It is worth noting that mapping the segmentation back to an co-registered but uncropped image (like the T1-weighted scan) requires information of the bounding box shape and position that was previously cropped out.

Each model variant achieves accurate and spatially coherent 3D segmentations that closely resemble the reference. However,

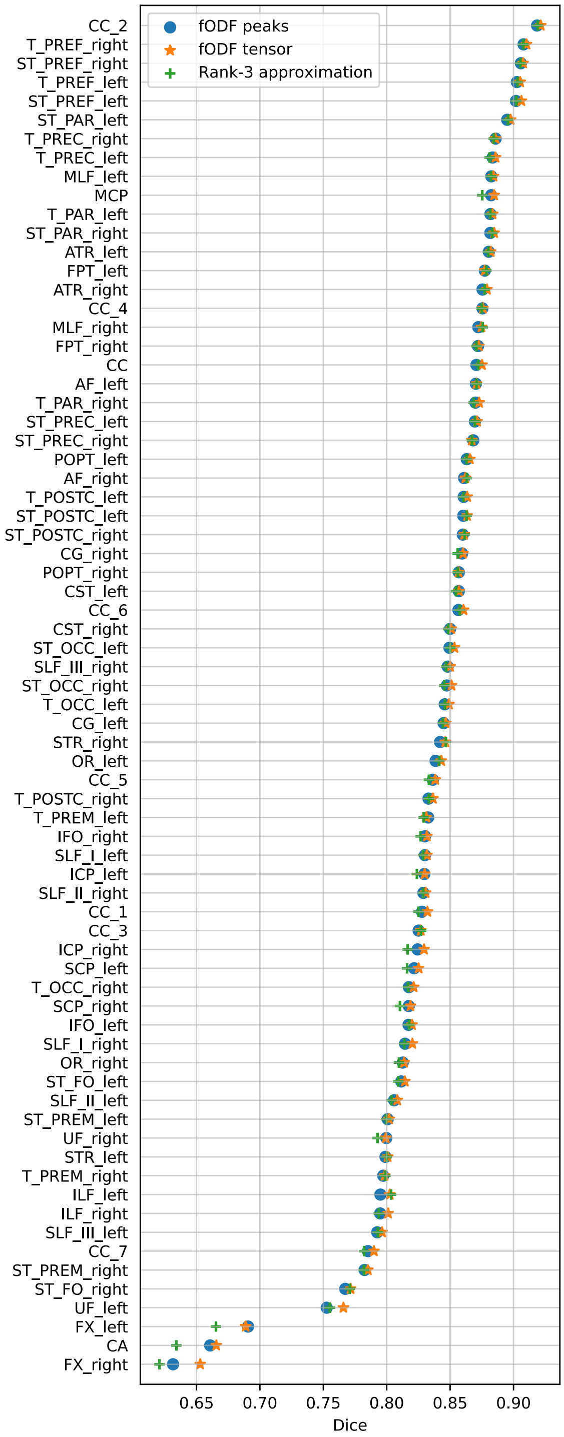

Figure 5: *Mean Dice scores over all test subjects of each tract per model input. A list of all 72 tract names is given in appendix A.*

the inspection of 2D slices reveals considerable over- and under-segmentations at the tract borders. In particular, all models appear more susceptible to under-segment, especially in ending regions, but show convincing capabilities to delineate the fiber trajectories in-between, even in thin tract sections.

Since the between-model deviations are very small, a meaningful qualitative comparison is challenging. Therefore, the following observations and interpretations should be considered with great caution.

In some instances, both the results of fODF peaks and tensor approximations show a higher incidence of false positives than the fourth-order tensors. This might suggest that fODF tensors lead to a more fine-grained but consequently less homogeneous segmentation. Although the tensor approximation vectors are similar to peaks, their segmentation results shows notable discrepancies compared to both peaks and tensors, e.g., in the depicted slice of the CA. This might be a consequence of multiple vectors of the tensor approximation representation corresponding to the same tract in regions of spreading fibers. As the vectors are concatenated in order of their magnitude, spreading vectors of similar lengths might get confused in order within neighboring voxels. In contrast, both peaks and tensors might be more spatially consistent within each input channel. Lastly, the segmentations of fODF tensors, compared with those of the peaks, may be more complete, e.g., in the posterior area of the IFO. However, the improved performance measured in the quantitative evaluation is not evident in single slices.

4.3. Concatenation with structural features

Furthermore, we explore the influence of adding structural features on TractSeg's performance. To this end, we concatenate the fODF tensor coefficients with two additional channels containing a T1-weighted MRI image and a partial volume map of the white matter. Co-registered T1 images are available from the preprocessed HCP data. For computing the partial volume map, we employed the automated segmentation tool [ZBS01] from the FMRIB Software Library (FSL) [JBB*12]. The underlying method models the T1 image intensities as a mixture of Gaussians representing different tissue types, spatially regularizing the tissue labels with a hidden Markov random field, and optimizing via expectation-maximization.

As a result of this experiment, a mean Dice score of approx. 0.8290 is measured, suggesting that the simple concatenation either introduced conflicting information or obstructed the model's interpretation of the fODF tensors. Likely, since T1 images contain values of a much greater numerical range than the fODF tensor coefficients, there arise considerable issues from the joint normalization over all input channels. However, considering the scope of this project, we leave the detailed investigation of concatenation and fusion strategies for future work.

5. Discussion

While expressive fiber orientation descriptors show a great potential as input features for direct segmentation methods, several limitations remain. This section is meant to discuss their strengths and weaknesses based on the presented experimental results, along with future prospectives and suggestions on further investigations.

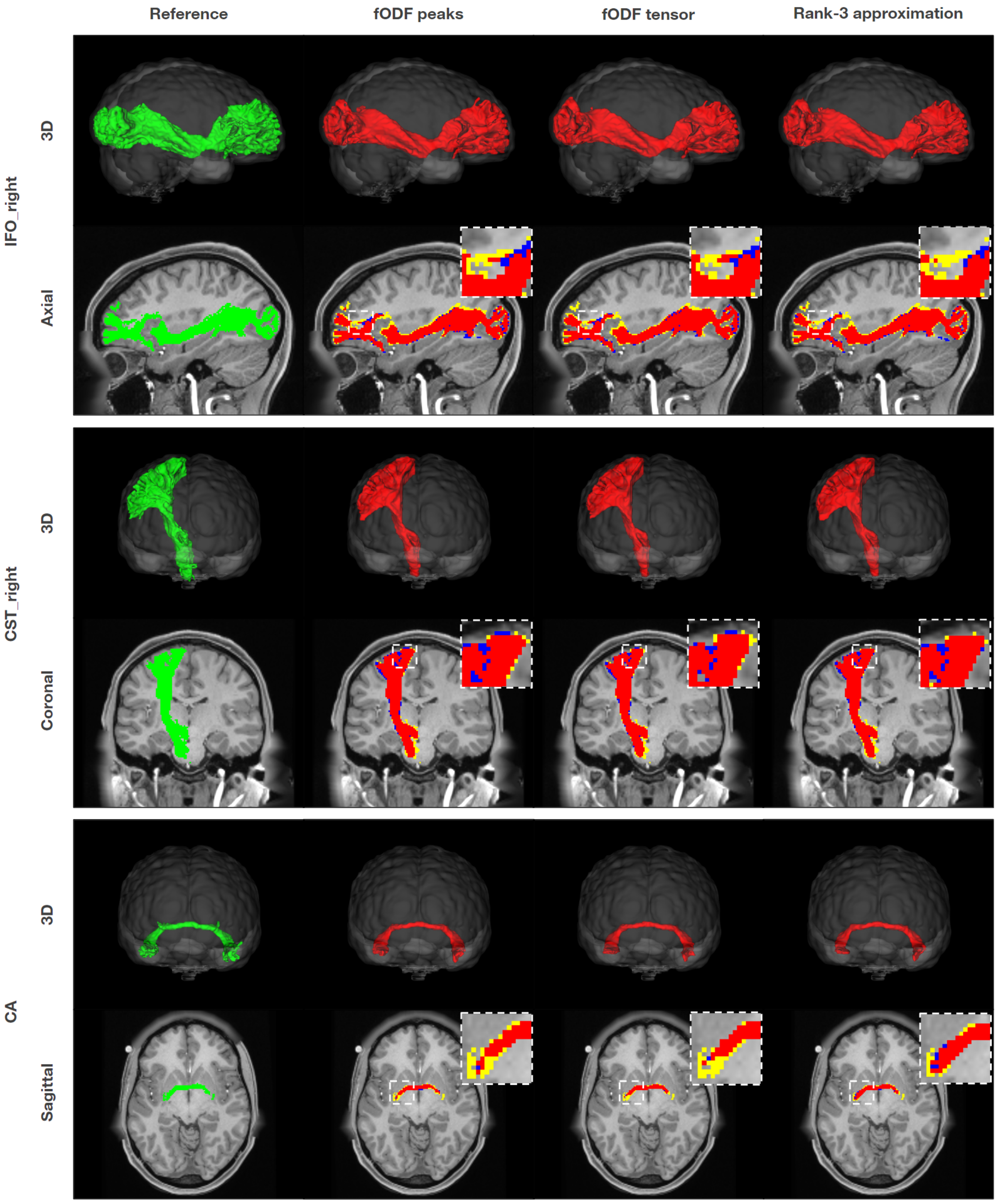

Figure 6: *Qualitative results using different fiber orientation features. Reconstruction of right inferior occipito-frontal fascicle (IFO_right), right corticospinal tract (CST_right), and commissure anterior (CA) of HCP subject 987983. Reference segmentation and model prediction are shown in green and red, respectively. Blue and yellow indicate false positives and false negatives, respectively. Figure created by virtue of MRtrix3 [TSR* 19] and Slicer3D [FBKC* 12].*

To begin with, identifying white matter tracts is an ill-posed problem in the sense that there exist different anatomical definitions of the same tracts [RPC*20] and no possibility to validate segmentations in an unambiguous way [ZDH*22]. In addition, any segmentation in dMRI data depends on scanner effects, vendor effects, shells, subjects, resolution, diffusion sampling scheme, and diffusion sensitization [STR*21]. Representations as peaks, rank-k approximations, or other features of the fODF mitigate the severity of these effects. They are also the basis for modern tractography and thereby virtual dissection, which helps to specify tract definitions.

Models for direct segmentation aim for a voxel-wise labeling and consider only voxel-wise input features. Since fiber orientation features are very local with respect to the overall geometry of fibers, they might lead to difficulties in the segmentation of images with high anatomical variability, for example, images with brain tumors or lesions [GASZ23] CSD is a complex operation that naturally entails estimation errors. Even in theory, CSD makes strong assumptions about the measured dMRI signal and the estimation of a fiber response kernel. Furthermore, the accurate computation of fiber orientation features requires dMRI data with high angular and spatial resolution as well as multiple shells, which is costly to attain and not available in many clinical settings. For these reasons, multiple studies have proposed to perform direct tract segmentation within raw dMRI measurements [KGCK23, GASZ23].

In this project, we focused on a supervised deep approach, which is the currently the most relevant in direct segmentation [GASZ23]. When using CNNs, using a slightly higher-dimensional input such as the fourth-order fODF tensors instead of peaks increases the network complexity only by a negligible amount. On the other hand, the spatial arrangement of voxel-based features needs to be accounted for by the CNN. Since 2D CNNs are much more efficient than their 3D counterparts [LCQ*20], the locality of fiber orientation features necessitates inconvenient solutions like mean fusion to construct 3D segmentations. Additionally, a border effect may arise from repeated zero padding in an architecture like the U-Net. These are some of the considerations that appear for local descriptors like fiber orientation features rather than the commonly used geometric features in streamline-based segmentation.

For supervision, accurate reference segmentations are required. Besides their challenging creation, this step also determines the number and selection of tracts the trained model will be able to identify. Despite recent propositions for efficient finetuning methods of a pretrained model for unseen tracts [LY21, LLZ*22, LZLY23], the quantity and quality of the annotated data arguably is the most decisive aspect of TractSeg and similar supervised learning methods, which is why challenges in generalization of the method to unseen data persist.

Our experimental results demonstrate a small but significant increase of the Dice score if more comprehensive features like fourth-order fODF tensors are utilized instead of peaks. Notably, we did not perform any cross-validation and only tested on a subset of the HCP subjects considered in TractSeg. However, our findings are consistent with the original and show indications that increasing the expressiveness and resolving the directional ambiguity of the inputs lead to an improved segmentation quality.

Among various state-of-the-art segmentation methods investigated by Schilling et al. [STR*21], as the only method using machine learning on local fiber orientation features, TractSeg shows the highest degree of reproducibility. Primarily, this affects the consistency of the resulting binary segmentations. Thereupon, the resulting tract masks may be used to guide tractography [WNMH18b] and for seeding or filtering streamlines belonging to specific tracts [WNHMH19], mitigating the common problem of false positive streamline reconstructions. Finally, this affects tract-specific quantitative analysis and tractometry [BCD*11]. Our results suggest that the more expressive features help delineating the tracts that are most difficult to segment, which might represent a prospective for segmentation improvement in general restricted settings, such as clinical applications.

In consideration of these applications, the relevance of expressive fiber orientation features in direct segmentation seem apparent. In this project, this has been explored in a series of specific experiments. In future work, it might be of interest to evaluate the generalization of our findings to datasets of different data quality, and varying levels of detail in the model inputs. Lastly, fusing local fiber orientation features with suitable representation of structural features in a multi-modal segmentation model could be usefully investigated in further research.

6. Conclusion

In conclusion, this report investigated the impact of fiber orientation descriptors as input features to direct tract segmentation methods. Focusing on TractSeg, a CNN-based approach, we described methods for computing these features and incorporating efficient representations of them into TractSeg's segmentation model, adapting its preprocessing, training, and data augmentation procedures. In a series of experiments, we compared fODF peaks, fODFs as fourth-order tensors, and rank-3 tensor approximations both quantitatively and qualitatively. We demonstrated a slight but notable improvement of 0.3 Dice points when replacing the fODF peaks originally used in TractSeg with the more comprehensive and unambiguous fODF tensors. Finally, we discussed the importance, strengths, and limitations of expressive fiber orientation features for the accurate identification of white matter tracts, especially in difficult regions, as well as future prospectives and challenges. The code of this project is openly available at `https://github.com/bertan-karacora/TractSeg`.

References

[AAB*16] ABADI M., AGARWAL A., BARHAM P., BREVDO E., CHEN Z., CITRO C., CORRADO G. S., DAVIS A., DEAN J., DEVIN M., ET AL.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016). 5

[AGMS24] ANKELE M., GRÜN J., MORELLE O., SCHULTZ T.: Bonndit: Diffusion imaging tools developed in Bonn. https://github.com/MedVisBonn/bonndit, 2024. 3, 4

[AJBTdS19] ASSAF Y., JOHANSEN-BERG H., THIEBAUT DE SCHOTTEN M.: The role of diffusion MRI in neuroscience. *NMR in Biomedicine 32*, 4 (2019), e3762. 1

[ALGS16] ANKELE M., LIM L.-H., GROESCHEL S., SCHULTZ T.: Fast and accurate multi-tissue deconvolution using SHORE and H-psd tensors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016), Springer, pp. 502–510. 2, 3

[ALGS17] ANKELE M., LIM L.-H., GROESCHEL S., SCHULTZ T.: Versatile, robust, and efficient tractography with constrained higher-order tensor fODFs. *International journal of computer assisted radiology and surgery 12* (2017), 1257–1270. 2, 3

[BCD*11] BELLS S., CERCIGNANI M., DEONI S., ASSAF Y., PASTERNAK O., EVANS C., LEEMANS A., JONES D.: Tractometry–comprehensive multi-modal quantitative assessment of white matter along specific tracts. In *Proc. ISMRM* (2011), vol. 678. 1, 10

[BML94] BASSER P. J., MATTIELLO J., LEBIHAN D.: Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B 103*, 3 (1994), 247–254. 3

[BPP*00] BASSER P. J., PAJEVIC S., PIERPAOLI C., DUDA J., ALDROUBI A.: In vivo fiber tractography using DT-MRI data. *Magnetic resonance in medicine 44*, 4 (2000), 625–632. 2

[BS09] BULLMORE E., SPORNS O.: Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience 10*, 3 (2009), 186–198. 1

[BWJ*03] BEHRENS T. E., WOOLRICH M. W., JENKINSON M., JOHANSEN-BERG H., NUNES R. G., CLARE S., MATTHEWS P. M., BRADY J. M., SMITH S. M.: Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 50*, 5 (2003), 1077–1088. 3

[BYB*11] BAZIN P.-L., YE C., BOGOVIC J. A., SHIEE N., REICH D. S., PRINCE J. L., PHAM D. L.: Direct segmentation of the major white matter tracts in diffusion tensor images. *Neuroimage 58*, 2 (2011), 458–468. 2

[CHPJ02] CATANI M., HOWARD R. J., PAJEVIC S., JONES D. K.: Virtual in vivo interactive dissection of white matter fasciculi in the human brain. *Neuroimage 17*, 1 (2002), 77–94. 1, 2

[Dem06] DEMŠAR J.: Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research 7* (2006), 1–30. 7

[DYPW19] DONG X., YANG Z., PENG J., WU X.: Multimodality white matter tract segmentation using CNN. In *Proceedings of the ACM Turing Celebration Conference-China* (2019), pp. 1–8. 2

[ESS*09] ECKSTEIN I., SHATTUCK D. W., STEIN J. L., MCMAHON K. L., DE ZUBICARAY G., WRIGHT M. J., THOMPSON P. M., TOGA A. W.: Active fibers: Matching deformable tract templates to diffusion tensor images. *Neuroimage 47* (2009), T82–T89. 2

[FBKC*12] FEDOROV A., BEICHEL R., KALPATHY-CRAMER J., FINET J., FILLION-ROBIN J.-C., PUJOL S., BAUER C., JENNINGS D., FENNESSY F., SONKA M., ET AL.: 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic resonance imaging 30*, 9 (2012), 1323–1341. 9

[FF16] FILLEY C. M., FIELDS R. D.: White matter and cognition: making the connection. *Journal of neurophysiology 116*, 5 (2016), 2093–2104. 1

[Fie08] FIELDS R. D.: White matter in learning, cognition and psychiatric disorders. *Trends in neurosciences 31*, 7 (2008), 361–370. 1

[Fri37] FRIEDMAN M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association 32*, 200 (1937), 675–701. 7

[Fri40] FRIEDMAN M.: A comparison of alternative tests of significance for the problem of m rankings. *The annals of mathematical statistics 11*, 1 (1940), 86–92. 7

[GASZ23] GHAZI N., AARABI M. H., SOLTANIAN-ZADEH H.: Deep Learning Methods for Identification of White Matter Fiber Tracts: Review of State-of-the-Art and Future Prospective. *Neuroinformatics* (2023), 1–32. 1, 2, 10

[GSW*13] 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., ET AL.: The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage 80* (2013), 105–124. 5

[HLT*20] HUANG H., LIN L., TONG R., HU H., ZHANG Q., IWAMOTO Y., HAN X., CHEN Y.-W., WU J.: Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2020), IEEE, pp. 1055–1059. 2

[HMVDW*20] HARRIS C. R., MILLMAN K. J., VAN DER WALT S. J., GOMMERS R., VIRTANEN P., COURNAPEAU D., WIESER E., TAYLOR J., BERG S., SMITH N. J., ET AL.: Array programming with NumPy. *Nature 585*, 7825 (2020), 357–362. 6

[HYL*21] HANSEN C. B., YANG Q., LYU I., RHEAULT F., KERLEY C., CHANDIO B. Q., FADNAVIS S., WILLIAMS O., SHAFER A. T., RESNICK S. M., ET AL.: Pandora: 4-D white matter bundle population-based atlases derived from diffusion MRI fiber tractography. *Neuroinformatics 19* (2021), 447–460. 2

[IJW*20] ISENSEE F., JÄGER P., WASSERTHAL J., ZIMMERER D., PETERSEN J., KOHL S., SCHOCK J., KLEIN A., ROSS T., WIRKERT S., NEHER P., DINKELACKER S., KÖHLER G., MAIER-HEIN K.: batchgenerators - a python framework for data augmentation, Jan. 2020. 6

[JBB*12] JENKINSON M., BECKMANN C. F., BEHRENS T. E., WOOLRICH M. W., SMITH S. M.: Fsl. *Neuroimage 62*, 2 (2012), 782–790. 8

[JDML19] JEURISSEN B., DESCOTEAUX M., MORI S., LEEMANS A.: Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine 32*, 4 (2019), e3785. 1, 2

[JTD*14] JEURISSEN B., TOURNIER J.-D., DHOLLANDER T., CONNELLY A., SIJBERS J.: Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. *NeuroImage 103* (2014), 411–426. 2, 3

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 5

[KGCK23] KEBIRI H., GHOLIPOUR A., CUADRA M. B., KARIMI D.: Direct segmentation of brain white matter tracts in diffusion MRI. *ArXiv* (2023). 2, 10

[LBBL*86] LE BIHAN D., BRETON E., LALLEMAND D., GRENIER P., CABANIS E., LAVAL-JEANTET M.: MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology 161*, 2 (1986), 401–407. 1, 2

[LBF*18] LAM P. D. N., BELHOMME G., FERRALL J., PATTERSON B., STYNER M., PRIETO J. C.: TRAFIC: fiber tract classification using deep learning. In *Medical Imaging 2018: Image Processing* (2018), vol. 10574, SPIE, pp. 257–265. 2

[LCQ*20] LIU L., CHENG J., QUAN Q., WU F.-X., WANG Y.-P., WANG J.: A survey on U-shaped networks in medical image segmentations. *Neurocomputing 409* (2020), 244–258. 10

[LDGS*20] 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.: Neuro4Neuro: A neural network approach for neural tract segmentation using large-scale population-based diffusion imaging. *Neuroimage 218* (2020), 116993. 2

[LLZ*22] LIU W., LU Q., ZHUO Z., LIU Y., YE C.: One-Shot Segmentation of Novel White Matter Tracts via Extensive Data Augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), Springer, pp. 133–142. 2, 10

[LXG*15] LEE C.-Y., XIE S., GALLAGHER P., ZHANG Z., TU Z.: Deeply-supervised nets. In *Artificial intelligence and statistics* (2015), Pmlr, pp. 562–570. 4

[LY21] LU Q., YE C.: Knowledge transfer for few-shot segmentation of novel white matter tracts. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27* (2021), Springer, pp. 216–227. 2, 10

[LZLY23] LIU W., ZHUO Z., LIU Y., YE C.: One-shot segmentation of novel white matter tracts via extensive data augmentation and adaptive knowledge transfer. *Medical Image Analysis 90* (2023), 102968. 2, 10

[MHNH*17a] 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., ET AL.: The challenge of mapping the human connectome based on diffusion tractography. *Nature communications 8*, 1 (2017), 1349. 1

[MHNH*17b] 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., ET AL.: The challenge of mapping the human connectome based on diffusion tractography. *Nature communications 8*, 1 (2017), 1349. 2, 7

[MNA*17] MICIKEVICIUS P., NARANG S., ALBEN J., DIAMOS G., ELSEN E., GARCIA D., GINSBURG B., HOUSTON M., KUCHAIEV O., VENKATESH G., ET AL.: Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017). 5

[MSP18] MANDONNET E., SARUBBO S., PETIT L.: The nomenclature of human white matter association pathways: proposal for a systematic taxonomic anatomical classification. *Frontiers in neuroanatomy 12* (2018), 94. 2

[Nem63] NEMENYI P. B.: *Distribution-free multiple comparisons.* Princeton University, 1963. 7

[PGM*19] 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.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (2019), Wallach H., Larochelle H., Beygelzimer A., d'Alché Buc F., Fox E., Garnett R., (Eds.), Curran Associates, Inc., pp. 8024–8035. 4

[PSF*19] POMIECKO K., SESTILI C., FISSELL K., PATHAK S., OKONKWO D., SCHNEIDER W.: 3D convolutional neural network segmentation of white matter tract masks from MR diffusion anisotropy maps. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019), IEEE, pp. 1–5. 3

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (2015), Springer, pp. 234–241. 2, 4

[RPC*20] RHEAULT F., POULIN P., CARON A. V., ST-ONGE E., DESCOTEAUX M.: Common misconceptions, hidden biases and modern challenges of dMRI tractography. *Journal of neural engineering 17*, 1 (2020), 011001. 9

[RQ14] RATNARAJAH N., QIU A.: Multi-label segmentation of white matter structures: application to neonatal brains. *NeuroImage 102* (2014), 913–922. 2

[SFG*14] SCHULTZ T., FUSTER A., GHOSH A., DERICHE R., FLORACK L., LIM L.-H.: Higher-order tensors in diffusion imaging. In *Visualization and Processing of Tensors and Higher Order Descriptors for Multi-Valued Data* (2014), Springer, pp. 129–161. 3, 4, 6

[SPDDG21] SEDLAR S., PAPADOPOULO T., DERICHE R., DESLAURIERS-GAUTHIER S.: Diffusion MRI fiber orientation distribution function estimation using voxel-wise spherical U-net. In *Computational Diffusion MRI: International MICCAI Workshop, Lima, Peru, October 2020* (2021), Springer, pp. 95–106. 3

[SRP*21] SCHILLING K. G., RHEAULT F., PETIT L., HANSEN C. B., NATH V., YEH F.-C., GIRARD G., BARAKOVIC M., RAFAEL-PATINO J., YU T., ET AL.: Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset? *Neuroimage 243* (2021), 118502. 1

[SS08] SCHULTZ T., SEIDEL H.-P.: Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1635–1642. 1, 3, 4

[STR*21] SCHILLING K. G., TAX C. M., RHEAULT F., HANSEN C., YANG Q., YEH F.-C., CAI L., ANDERSON A. W., LANDMAN B. A.: Fiber tractography bundle segmentation depends on scanner effects, vendor effects, acquisition resolution, diffusion sampling scheme, diffusion sensitization, and bundle segmentation workflow. *Neuroimage 242* (2021), 118451. 9, 10

[TCC07] TOURNIER J.-D., CALAMANTE F., CONNELLY A.: Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage 35*, 4 (2007), 1459–1472. 1

[TCGC04] TOURNIER J.-D., CALAMANTE F., GADIAN D. G., CONNELLY A.: Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *Neuroimage 23*, 3 (2004), 1176–1185. 3

[TH15] TAHA A. A., HANBURY A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging 15*, 1 (2015), 1–28. 7

[TSR*19] TOURNIER J.-D., SMITH R., RAFFELT D., TABBARA R., DHOLLANDER T., PIETSCH M., CHRISTIAENS D., JEURISSEN B., YEH C.-H., CONNELLY A.: MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage 202* (2019), 116137. 3, 9

[TZZ*23] TCHETCHENIAN A., ZHU Y., ZHANG F., O'DONNELL L. J., SONG Y., MEIJERING E.: A comparison of manual and automated neural architecture search for white matter tract segmentation. *Scientific Reports 13*, 1 (2023), 1617. 2

[VESB*13] VAN ESSEN D. C., SMITH S. M., BARCH D. M., BEHRENS T. E., YACOUB E., UGURBIL K., CONSORTIUM W.-M. H., ET AL.: The WU-Minn human connectome project: an overview. *Neuroimage 80* (2013), 62–79. 5

[WNHMH19] WASSERTHAL J., NEHER P. F., HIRJAK D., MAIER-HEIN K. H.: Combined tract segmentation and orientation mapping for bundle-specific tractography. *Medical image analysis 58* (2019), 101559. 4, 6, 7, 10

[WNMH18a] WASSERTHAL J., NEHER P., MAIER-HEIN K. H.: TractSeg-Fast and accurate white matter tract segmentation. *NeuroImage 183* (2018), 239–253. 1, 2, 3, 4, 5, 6, 7, 14

[WNMH18b] WASSERTHAL J., NEHER P. F., MAIER-HEIN K. H.: Tract orientation mapping for bundle-specific tractography. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11* (2018), Springer, pp. 36–44. 3, 4, 10

[YSA*09] YAMADA K., SAKAI K., AKAZAWA K., YUEN S., NISHIMURA T.: MR tractography: a review of its clinical applications. *Magnetic resonance in medical sciences 8*, 4 (2009), 165–174. 1

[YYPC21] YANG J. Y.-M., YEH C.-H., POUPON C., CALAMANTE F.: Diffusion MRI tractography for neurosurgery: the basics, current state, technical reliability and challenges. *Physics in Medicine & Biology 66*, 15 (2021), 15TR01. 1

[ZBS01] Zhang Y., Brady M., Smith S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging 20*, 1 (2001), 45–57. 8

[ZDH*22] Zhang F., Daducci A., He Y., Schiavi S., Seguin C., Smith R. E., Yeh C.-H., Zhao T., O'Donnell L. J.: Quantitative mapping of the brain's structural connectivity using diffusion MRI tractography: A review. *Neuroimage 249* (2022), 118870. 2, 9

[ZKH*20] Zhang F., Karayumak S. C., Hoffmann N., Rathi Y., Golby A. J., O'Donnell L. J.: Deep white matter analysis (DeepWMA): fast and consistent tractography segmentation. *Medical Image Analysis 65* (2020), 101761. 2

Appendix A: Tract names

AF	Arcuate fascicle (left/right)
ATR	Anterior Thalamic Radiation (left/right)
CA	Commissure Anterior
CC_1	Rostrum
CC_2	Genu
CC_3	Rostral body (Premotor)
CC_4	Anterior midbody (Primary Motor)
CC_5	Posterior midbody (Primary Somatosensory)
CC_6	Isthmus
CC_7	Splenium
CG	Cingulum (left/right)
CST	Corticospinal tract (left/right)
MLF	Middle longitudinal fascicle (left/right)
FPT	Fronto-pontine tract (left/right)
FX	Fornix (left/right)
ICP	Inferior cerebellar peduncle (left/right)
IFO	Inferior occipito-frontal fascicle (left/right)
ILF	Inferior longitudinal fascicle (left/right)
MCP	Middle cerebellar peduncle
OR	Optic radiation (left/right)
POPT	Parieto-occipital pontine (left/right)
SCP	Superior cerebellar peduncle (left/right)
SLF_I	Superior longitudinal fascicle I (left/right)
SLF_II	Superior longitudinal fascicle II (left/right)
SLF_III	Superior longitudinal fascicle III (left/right)
STR	Superior Thalamic Radiation (left/right)
UF	Uncinate fascicle (left/right)
CC	Corpus Callosum
T_PREF	Thalamo-prefrontal (left/right)
T_PREM	Thalamo-premotor (left/right)
T_PREC	Thalamo-precentral (left/right)
T_POSTC	Thalamo-postcentral (left/right)
T_PAR	Thalamo-parietal (left/right)
T_OCC	Thalamo-occipital (left/right)
ST_FO	Striato-fronto-orbital (left/right)
ST_PREF	Striato-prefrontal (left/right)
ST_PREM	Striato-premotor (left/right)
ST_PREC	Striato-precentral (left/right)
ST_POSTC	Striato-postcentral (left/right)
ST_PAR	Striato-parietal (left/right)
ST_OCC	Striato-occipital (left/right)

Table 1: *Abbreviations and names of all 72 tracts that have been annotated in the reference dataset [WNMH18a].*