Artificial Intelligence In Medicine 130 (2022) 102330

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Diffusion tensor estimation with transformer neural networks☆

Davood Karimi *, Ali Gholipour

Department of Radiology at Boston Children's Hospital, and Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
Diffusion MRI
Diffusion tensor imaging
Deep learning
Transformer networks

ABSTRACT

Diffusion tensor imaging (DTI) is a widely used method for studying brain white matter development and degeneration. However, standard DTI estimation methods depend on a large number of high-quality measurements. This would require long scan times and can be particularly difficult to achieve with certain patient populations such as neonates. Here, we propose a method that can accurately estimate the diffusion tensor from only six diffusion-weighted measurements. Our method achieves this by learning to exploit the relationships between the diffusion signals and tensors in neighboring voxels. Our model is based on transformer networks, which represent the state of the art in modeling the relationship between signals in a sequence. In particular, our model consists of two such networks. The first network estimates the diffusion tensor based on the diffusion signals in a neighborhood of voxels. The second network provides more accurate tensor estimations by learning the relationships between the diffusion signals as well as the tensors estimated by the first network in neighboring voxels. Our experiments with three datasets show that our proposed method achieves highly accurate estimations of the diffusion tensor and is significantly superior to three competing methods. Estimations produced by our method with six diffusion-weighted measurements are comparable with those of standard estimation methods with 30–88 diffusion-weighted measurements. Hence, our method promises shorter scan times and more reliable assessment of brain white matter, particularly in non-cooperative patients such as neonates and infants.

1. Introduction

1.1. Diffusion tensor imaging

Diffusion weighted magnetic resonance imaging (DW-MRI) is one of the most common medical imaging modalities. It uses the diffusion of water molecules, and restrictions thereof due to obstacles such as membranes and fibers, to generate contrast. Although its applications are not limited to brain, DW-MRI is presently the best non-invasive tool for studying the brain micro-structure in vivo. Diffusion tensor imaging (DTI) is a specific type of DW-MRI that is widely used in neuroimaging [1,10,41]. DTI is capable of capturing diffusion anisotropy, i.e., the dependence of the diffusion process on the orientation. In DTI, a Gaussian model of diffusion is assumed, whereby the orientation-dependence of diffusion is characterized with a 3×3 symmetric matrix, i.e., a tensor:

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix}. \quad (1)$$

The diffusion tensor formalism can be interpreted as representing the surface of the diffusion front with an ellipsoid. An eigen-decomposition of D gives us the direction of the strongest diffusion as well as parameters such as mean diffusivity (MD) and fractional anisotropy (FA), which are widely used to study brain development and as biomarkers for various diseases [8,9,40,42]. There exist more complex models of tissue micro-structure, e.g. [7,53,66]. Nonetheless, DTI remains the most widely used method in brain micro-structure studies because of easier acquisition and model fitting and wide-spread availability of DTI analysis software.

☆ This study was supported in part by the National Institutes of Health (NIH) grants R01 EB018988, R01 NS106030, R01 EB031849, and R01 EB032366; in part by the Office of the Director of the NIH under award number S10OD0250111; in part by the National Science Foundation under award number 2123061; and in part by a Technological Innovations in Neuroscience Award from the McKnight Foundation. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the McKnight Foundation. D. Karimi and A. Gholipour are with the Computational Radiology Laboratory of the Department of Radiology at Boston Children's Hospital, and Harvard Medical School, Boston, MA, USA (email: davood.karimi@childrens.harvard.edu).

* Corresponding author.
E-mail address: davood.karimi@childrens.harvard.edu (D. Karimi).

https://doi.org/10.1016/j.artmed.2022.102330
Received 5 December 2021; Received in revised form 23 March 2022; Accepted 29 May 2022
Available online 6 June 2022

Given diffusion signals measured with higher and lower diffusion weightings, s_1 and s_0 respectively, one can write:

$$s_1/s_0 = exp\big(-b_{xx}D_{xx} - b_{yy}D_{yy} - b_{zz}D_{zz} - 2b_{xy}D_{xy} - 2b_{xz}D_{xz} - 2b_{yz}D_{yz}\big), \quad (2)$$

where $b_{..}$ are elements of the so-called b-matrix [29,41]. Given a set of m such measurements, after log-transformation, one can express the relation between the diffusion tensor elements and the measurements as a linear system of equations:

$$B\widetilde{D} = S, \quad (3)$$

where B is a design matrix that depends only on the directions and strengths of the applied diffusion-sensitizing gradients, S is the vector of log-transformed diffusion signal measurements, and $\widetilde{D} = [D_{xx}, D_{yy}, D_{zz}, D_{xy}, D_{xz}, D_{yz}]$ is the vector of unknowns.

Many different approaches have been proposed for estimating $\widetilde{D}$. The ordinary least squares solution can be obtained as $\widetilde{D} = (B^TB)^{-1}B^TS$. This solution is based on the assumption of homoskedasticity, i.e., that the variance of the noise is the same for all measurements. Even though this assumption is largely correct for the diffusion signal before the log transformation, it is not correct after the transformation. After log-transformation, the measurement variance is higher for lower signal intensities. Therefore, one can improve upon the ordinary least squares method by introducing weights: $\widetilde{D} = (B^TWB)^{-1}B^TWS$, where W is a diagonal matrix with the diagonal elements proportional to the measurements [36]. Alternatively, one could attempt solving the original non-linear system of equations without log-transforming the measurements. This approach is theoretically more appealing but can suffer from other problems such as sensitivity to the initial solution, convergence to local minimum, and higher computational cost.

There have been attempts to improve upon the least squares-based methods mentioned above. For example, basic least squares-based methods may yield a tensor that has negative eigen-values, which is physically invalid. One approach to enforce positivity of eigenvalues is Cholesky factorization of the diffusion tensor [35,36]. Other notable methods include algorithms that aim at reducing the effect of erroneous measurements due to such factors as high noise, subject head motion, and cardiac pulsation. For example, RESTORE algorithm uses an iterative weighted least squares strategy to detect and remove erroneous measurements [14,15]. Bootstrap methods have been used to improve and quantify the accuracy and uncertainty of DTI parameter estimation [16,35].

The above methods are widely used in practice and constitute the core of the tensor fitting algorithms in common DW-MRI software [21,61]. However, they require a large number of measurements for accurate tensor estimation. Although the diffusion tensor has only six degrees of freedom, practical guidelines recommend acquiring at least 30 measurements with diffusion encoding directions uniformly spread on the sphere [31,58]. It is strongly recommended to increase the number of measurements to much larger than 30, if possible, in order to achieve more accurate and more robust tensor estimation [29]. These requirements highlight the challenging nature of estimating micro-structural parameters of interest from noisy and imperfect measurements. However, they also suggest that standard diffusion tensor estimation methods may be sub-optimal. In particular, these estimation methods are based on biophysical models of diffusion that can only approximate the true underlying signal generation processes. More importantly, the classical estimation methods fit the diffusion signal on a voxel-wise basis. They fail to take into account the correlation between signals in neighboring voxels and to exploit the spatial regularity of diffusion tensor values.

DTI estimation accuracy depends not only on the number of measurements, but also on several other factors. For example, the choice of the diffusion gradient strength (the b-vale) can have a significant effect, as shown by several prior works [2,6,31]. Similarly, the arrangement of the directions of diffusion-sensitizing gradients can also be very important [17,28,58]. Another important factor is the signal to noise ratio (SNR). The effect of noise on diffusion tensor estimation error has been explored in the past and various methods for reducing the impact of measurement noise have been proposed [12,25,30]. Moreover, like any other DW-MRI technique, diffusion tensor imaging suffers from artifacts such as eddy current-induced distortions, magnetic susceptibility effects, subject motion, and cardiac pulsation [47,52,55].

Although many factors contribute to DTI estimation error, as briefly discussed above, in general increasing the number of measurements improves the accuracy and robustness of estimation [28,29]. Acquiring more measurements means longer scan times. Moreover, it may be difficult to achieve with non-cooperative subjects such as infants and young children, where part of the data may have to be discarded due to excessive motion. Therefore, methods that can accurately estimate the diffusion tensor from smaller numbers of measurements are highly desirable. Machine learning methods have a great potential in this regard. Unlike standard estimation methods, they do not need to assume a known mathematical model for the diffusion signal and noise. Instead, they learn the mapping from the diffusion signal to the tensor from training data. Furthermore, they can effectively learn the spatial correlations in the diffusion signal and the parameter(s) of interest. With the increasing availability of very large DW-MRI datasets, the advantage of machine learning methods has grown. It is now possible to train a machine learning model on these large and rich datasets and use the trained model on less perfect in-house datasets.

1.2. Related works

Applications of machine learning and data-driven methods for parameter estimation in DW-MRI have been explored in several prior works. Random forests, support vector regression, and other classical machine learning methods were used in several works [50,51,54,56]. More recently, deep learning has been shown to hold great promise for improving the accuracy and robustness of parameter estimation in DW-MRI [5,24,44,49]. Several recent studies have shown that deep learning can dramatically reduce the number of measurements, and hence the scan times, required for estimating micro-structural parameters of interest. The q-space deep learning (q-DL) was one of the first deep learning methods for diffusion parameter estimation [24]. It showed that a three-layer neural network could accurately estimate diffusion kurtosis as well as neurite orientation dispersion and density measures, while reducing the required number of measurements by a factor of 12. This result was particularly interesting given that a very simple neural network was used and the network performed the prediction on a voxel-wise basis, i.e., without exploiting spatial correlations.

Following the success of q-DL, several studies have used deep learning models to estimate other diffusion parameters on a voxel-wise basis. One study showed that a deep neural network can achieve significantly more accurate estimation of fiber orientations than standard methods such as constrained spherical deconvolution [49]. Examples of other parameters that have been estimated on a voxel-wise basis include the number [38] and orientation [39] of major fibers.

One can expect more accurate and more robust estimation when spatial correlations between the signal in neighboring voxels and the spatial regularity of the parameter(s) of interest are exploited. There are a variety of deep learning models that are capable of learning such spatial patterns. Perhaps the most well-known of these models are convolutional neural networks (CNN). Prior works have applied CNNs on patches of DW images to estimate the fiber orientations [37,45]. One study reported that more accurate estimation of diffusion kurtosis measures could be obtained, compared with the q-DL framework, with a simple CNN [44]. In other studies, CNNs have been used for tract segmentation and tractography analysis [63,65].

Several recent studies have used deep learning models for estimating

the diffusion tensor, FA, and MD. One study used a CNN for direct estimation of the diffusion tensor from DW-MRI measurements [43]. Aliotta et al. used a multi-layer perceptron to estimate MD and FA [4]. Another notable example is the DeepDTI method [59], which proposed a CNN model for estimating the diffusion signal residuals. The input to the CNN is a set of potentially noisy and artifact-full DW scans that are stacked together. The CNN learns to map these low-quality scans to their difference (residual) with respect to high-quality reference DW scans. In order to improve the model accuracy, anatomical (i.e., T1 and T2) images are registered to the DW scans and stacked with the DW scans to enrich the CNN input. In a recent study by our own team, we used CNNs for accurate color-FA estimation for fetuses [32,33]. A combination of a CNN and a multi-layer perceptron was used for FA and MD estimation in a recent work [3]. Another recent study used a multi-scale encoder-decoder CNN for estimating FA and MD [64]. The authors also used a monte-carlo dropout technique to compute the prediction uncertainty of the model predictions.

Despite the importance of the efforts mentioned above, the potential of deep learning for improving the accuracy and robustness of parameter estimation in DW-MRI is highly under-explored. One important shortcoming of most prior studies is their failure to effectively model the relationship between diffusion signal in neighboring voxels. Some methods, such as [24,49], have only used the signal in one voxel. Some other methods, for example [37,45], have used models such as CNNs that have originally been devised for computer vision applications and are not optimal for regression and parameter estimation applications. There are also studies that have used 2D CNNs, which fail to exploit the correlations in all three dimensions [23].

1.3. Contributions of this work

In this work, we propose a novel method for diffusion tensor estimation. Our main idea is to exploit the correlations between the diffusion signal and diffusion tensor parameters in neighboring voxels. Brain tissue micro-structure is spatially regular, in the sense that micro-structural properties such as fiber orientations do not change randomly between adjacent voxels. Rather, there exist strong spatial correlations between neighboring voxels. Moreover, these spatial correlations in the brain tissue micro-structure are largely shared across the brains of different subjects. These correlations in micro-structure, in turn, give rise to correlations between diffusion signals in neighboring voxels. We propose to learn these correlations in order to improve the accuracy and robustness of DTI estimation, especially when the measurements are noisy and few in number.

Our proposed method is based on the attention models, which represent the state of the art in sequence modeling. While prior works have either ignored spatial correlations or have used computer vision models such as CNN to learn spatial correlations, we use attention models that are more flexible and more powerful. The attention mechanism has important advantages over more classical deep learning models such as CNNs and RNNs. It offers higher flexibility since network weights are adapted in an input-dependent fashion. Moreover, each component/location of the output can attend to any component/location of the input, regardless of the distance between the two components in the sequence. In RNNs, for example, it can take up to n steps to propagate the information from one location in the sequence to another location, where n is the sequence length. In attention models, on the other hand, information can be exchanged between any two locations in the sequence in one step. We leverage these advantages to develop a novel method for DTI estimation. We show that our proposed method can accurately estimate the diffusion tensor from only six diffusion-weighted measurements. In particular, we demonstrate the effectiveness of our method on neonatal subjects where tensor estimation is very challenging and standard estimation methods can be highly inaccurate and unreliable.

2. Materials and methods

2.1. Attention models

The concept of attention can be employed in different machine learning models, but here our focus is on neural networks. In a standard neural network, the output of each layer is computed from the output of the preceding layer using a function of the form $x^i = \Psi(W^{iT}x^{i-1})$, where W^i represents the layer weights (for simplicity of presentation we omit the bias term) and Ψ is some non-linear function. In standard neural networks, the weights are optimized on a set of training data and they remain fixed afterwards. Attention models represent an alternative paradigm in which the weights are computed dynamically based on the input. In other words, $x^i = \Psi(f_\theta(x^{i-1})^T x^{i-1})$, where f_θ is a learnable function [48]. In this paradigm, the model weights are not fixed at training; rather, they depend on the input at inference time. The attention mechanism can take different forms. One common form is self-attention, where elements of W^i depend on the pair-wise similarity between elements of the input sequence x^{i-1}. In other words: $W^i{}_{s,t} = \mathbf{score}(x_s^{i-1}, x_t^{i-1})$. The score function, $\mathbf{score}(x_s^{i-1}, x_t^{i-1})$, quantifies the similarity between the two vectors x_s^{i-1} and x_t^{i-1} and can take various forms. One common form is Luong's multiplicative formulation: $\mathbf{score}(a,b) = a^T H b$, for some matrix H [46]. Therefore, with this formulation the network can learn to compute x^i by "paying attention" to the relevant pieces of information in x^{i-1} in a dynamic input-dependent manner.

In this work we follow a self-attention approach similar to that of the transformer network [62]. Let us denote the output of the previous network layer with $x^{i-1} \in \mathrm{IR}^{n,\,d}$, where n is the sequence length and d is the dimension of each element of the sequence. First, a set of query, key, and value sequences are computed via linear projections of x^{i-1}:

$$Q^i = x^{i-1}W_Q^i, \quad K^i = x^{i-1}W_K^i, \text{ and } V^i = x^{i-1}W_V^i \tag{4}$$

The projection matrices W_Q^i, W_K^i, and W_V^i are of size $\mathrm{IR}^{d,\,d_h}$, which means that the query, key, and value sequences will be of size d_h. The self-attention output is then computed as:

$$x^{i*} = \frac{Q^i K^{iT}}{\sqrt{d_h}} V^i, \tag{5}$$

where the scaling factor $1/\sqrt{d_h}$ is introduced for stable computations. In other words, self-attention is formulated based on the similarity between queries and keys, which are both computed from the input x^{i-1}. Note that similarity between query and key vectors in Eq. (5) is computed as a dot product, which is a special case of Luong's attention where H is the identity matrix. Since x^{i*} is in IR^{d_h}, it is passed through another fully-connected layer to generate $x^i \in \mathrm{IR}^d$ as the input for the next stage of the network. A transformer network consists of a succession of such self-attention modules. The output of the last module is projected onto the space of the desired network output using a fully-connected layer.

There are two important variations to the standard transformer model that we also utilize in this work [48,62]. First, in order to improve the expressive power of the learned attention maps, multi-headed attention is used. In this approach, n_h different query, key, and value sequences are computed, each with different projection matrices in Eq. (4). Then, x^{i*} is formed by concatenating the n_h sequences, each computed using Eq. (5). Second, the standard self-attention model lacks a means of knowing the sequence order because it is permutation-invariant. To overcome this limitation, a positional encoding is added to the input sequence [62]. Specifically, the sequence of initial input signals $x_s \in \mathrm{IR}^{n,\,d_s}$ is projected onto IR^d and a sequence of the same shape is added to it: $x^0 = x_s W_s + p$. Here $W_s \in \mathrm{IR}^{d_s,\,d}$ is the signal embedding projection matrix and $p \in \mathrm{IR}^{n,\,d}$ is meant to encode the relative position between elements of the input sequence. Many different forms of positional encoding have been proposed [19,48,57,62]. In this work, because we do not know a priori how diffusion signals and tensors in

neighboring voxels are related, we consider a learnable positional encoding. In other words, in our method p is a free parameter that is learned during training.

2.2. Proposed method

Fig. 1 shows our proposed method, which we refer to as Transformer-DTI because its core is similar to the transformer model [48,62]. It requires only six diffusion weighted measurements. The directions of the diffusion-sensitizing gradients for these six diffusion weighted measurements are clarified in more detail below. To exploit the spatial correlations in the signal and tissue micro-structure, the model uses the signal in a cubic patch of side length equal to L voxels to estimate the diffusion tensor in a cubic patch of the same size. Unless otherwise specified, in the experiments reported in this paper we use $L = 5$.

The proposed model aims to estimate the diffusion tensor in each voxel by exploiting the spatial correlations in the signal and in the diffusion tensor. To this end, our model consists of two sub-models, which are trained separately and sequentially. Model S uses the diffusion signal as the input and estimates the diffusion tensor. Model ST, on the other hand, uses the diffusion signal as well as the diffusion tensor estimated by Model S to provide a more accurate estimation of the diffusion tensor. As we show in the Results section below, Model S on its own can provide accurate tensor estimations. Nonetheless, the final estimations provided by Model ST are significantly more accurate. This is because Model S can only exploit the correlations between diffusion signals in the neighboring voxels. Model ST builds upon the estimation produced by Model S. Although the tensors estimated by Model S are not optimal, they serve as useful additional input for Model ST. Therefore, Model ST which provides our final tensor estimate, can learn to exploit the correlations in the diffusion signal as well as the correlations in the diffusion tensor.

Denoting the number of diffusion measurements in each voxel with n_s, the diffusion signal in a patch will be of shape $\mathrm{IR}^{L,\,L,\,L,\,n_s}$. This is first reshaped into $x_s \in \mathrm{IR}^{n,\,n_s}$, where $n = L^3$, which will serve as the input signal sequence for both Model S and Model ST. In both models, x_s is first embedded into IR^d and a learnable positional encoding sequence is added to form the input to the transformer modules, $x_0 \in \mathrm{IR}^{n,\,d}$. In Model S, this position-encoded sequence is passed through a series of N_{Tr} transformer modules. The output of the last transformer module is projected from IR^d onto IR^6 to form the estimated tensor sequence of size $\mathrm{IR}^{L^3,\,6}$. This output sequence can be reshaped into $\mathrm{IR}^{L,\,L,\,L,\,6}$ to obtain the tensor estimate for the input patch.

Model ST has two branches, each one of them similar to Model S. One of the branches works on the diffusion signal, similar to Model S. The other branch is architecturally identical, but works on the diffusion tensor estimated by Model S. These two branches are meant to learn the spatial correlations between the diffusion signal and the tensor values. The outputs of these two branches are concatenated and passed through a fully-connected layer to estimate the final diffusion tensor estimate for the patch.

2.3. Implementation and training

We selected the model architecture hyper-parameters using preliminary experiments on our training datasets. We set the number of transformer modules, N_{Tr}, in Model S and each of the two branches of Model ST to 4, as shown in Fig. 1. Furthermore, we set $d_h = 512$, and the number of heads in multi-headed self-attention $n_h = 2$. We discuss the effects of some of these hyper-parameters on the performance of the proposed method in the Results section below. We initialized all learnable parameters using He's method [26]. We first trained Model S, by minimizing the square of the difference between estimated ($\hat{D}_{\mathrm{S}}$) and reference (D_{ref}) tensors:

$$\mathscr{L}\left(\hat{D}_{\mathrm{S}}, D_{\mathrm{ref}}\right) = \| \hat{D}_{\mathrm{S}} - D_{\mathrm{ref}} \|_2^2 \tag{6}$$

Once training of Model S was finished, we trained Model ST. Our experiments showed that further fine-tuning Model S during the training of Model ST did not improve the accuracy of Model ST. Therefore, while training Model ST, we kept Model S fixed. Model ST was also trained using a loss function similar to Eq. (6), with $\hat{D}_{\mathrm{ST}}$ in place of $\hat{D}_{\mathrm{S}}$. Both

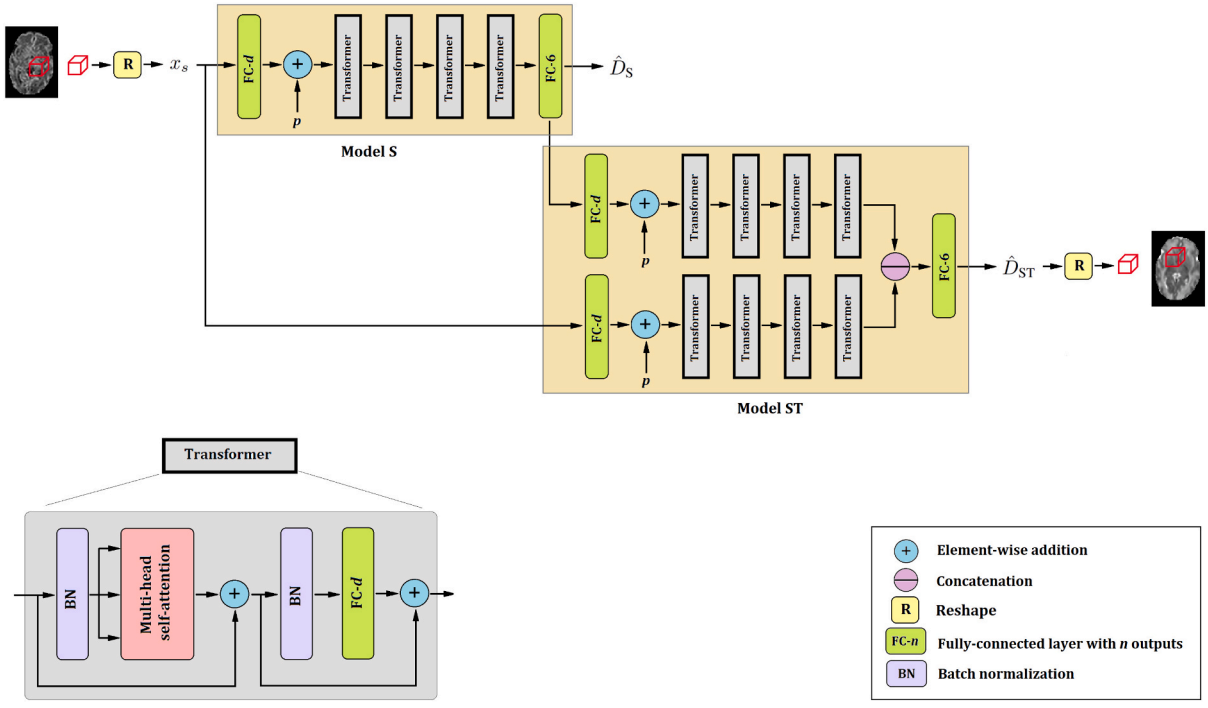

Fig. 1. A schematic of the proposed method. The input image is a volumetric (3D) image with six channels, where each channel represents one of the six normalized diffusion-weighted measurements. The output is a tensor image of the same shape as the input image with six channels, where each of the channels represents one of the six tensor elements.

models were trained with a batch size of 10, and an initial learning rate of 10^{-4} using Adam [34]. We reduced the learning rate by a factor of 0.9 every time the loss on the validation set did not decrease after a training epoch. We stopped the training if the validation loss did not decrease after two consecutive epochs. All models were implemented in TensorFlow. Training and validation were performed on a Linux computer with an NVIDIA GeForce GTX 1080 GPU.

2.4. Data and evaluation approach

Most of the experimental results presented in this paper are with the developing Human Connectome Project (dHCP) dataset [11]. Nonetheless, to demonstrate the applicability of our method to other datasets, we also report experimental results with scans from the Pediatric Imaging, Neurocognition, and Genetics (PING) dataset [27] as well as in-house scans of Vein of Galen Malformation (VOGM) patients at Boston Children's Hospital. This study was approved by the institutional review board. We mostly focus on the dHCP dataset because it is a publicly-available dataset on which the interested reader can train and test our method and because it is a challenging dataset. It consists of DW-MRI images of neonates scanned at 29–45 gestational weeks. The neonatal period represents a critical time in brain development. It is characterized by rapid cortical expansion and formation of connections between distant regions of the brain. Immature myelination of the white matter and patient motion make diffusion tensor imaging of neonates especially challenging.

Each DW-MRI scan in the dHCP dataset includes measurements at three b-values of 400, 1000, and 2600. Following the widely-adopted recommendations [31], we use the $b = 1000$ measurements for DTI estimation. Each scan includes 88 measurements in the $b = 1000$ shell, which are approximately uniformly distributed on the sphere. For each subject, we used all 88 measurements to reconstruct a high-quality "reference" DTI with the constrained weighted linear least squares (CWLLS) method [36]. We refer to this reference reconstruction as D_{ref}. For our proposed method, we selected six of the 88 measurements that were closest to the six optimized directions proposed in [58]. Specifically, the unit vectors indicating these directions are: [0.910, ± 0.416, 0.000], [±0.416,0.000,0.910], and [0.000,0.910, ±0.416]. These six directions have been derived to minimize the condition number of the diffusion tensor transformation matrix [58]. We also selected one of the $b = 0$ measurements to normalize the six $b = 1000$ measurements. We refer to the reconstructions of our proposed method with these six normalized measurements as D_{Tr}. For comparison with existing methods, we apply three methods on the same six diffusion-weighted measurements and one $b = 0$ measurement. These three methods are the following: 1) Constrained Weighted Linear Least Squares (CWLLS) [36], which is the standard estimation method, 2) Constrained Nonlinear Least Squares (CNLS) [36], and 3) The CNN-based method of Lin et al. [45]. This method was originally proposed for estimation of fiber orientation distribution. We simply changed the first and the last layers of the network to match our application. We refer to this method as CNN-DTI.

With the dHCP dataset, we trained our method using scans of 200 subjects aged 40–45 gestational weeks. We used data from 40 of these subjects as validation set during various stages of hyper-parameter selection and training. Once the training of the final model was complete, we tested our method on scans of 40 independent subjects; 20 of these were from the same age range, while the other 20 were younger subjects aged 29–36 gestational weeks. We compare our method with the competing techniques in terms of the norm of the difference between the estimated tensor and the reference tensor, D_{ref}. Specifically, if we denote the predicted tensor with D_{pred}, then we define the tensor estimation error as $\sum_{i=1}^{6} | D^i_{\text{pred}} - D^i_{\text{ref}}|$, where the index i refers to the six tensor elements. Moreover, we compute the error in FA, MD, and the angle of the major eigen-vector of the tensor. To do this, we calculate the eigen-decomposition of the tensor in each voxel and compute the FA and MD from the eigen-values using their standard definitions [29]. For each voxel, the error in FA and MD is defined as the absolute difference between those of the predicted and reference tensors. We compute the angle between the major eigen-vectors of the predicted and reference tensors as the angular error between the two tensors. We also present tractography and connectivity analysis results. To have a fair comparison between different methods, no especial data pre-processing (e.g., smoothing or re-sampling) was performed for our method or any of the compared techniques. Our proposed method does not rely on such data pre-processing steps.

3. Results and discussion

Table 1 shows the error in the estimated tensor and three tensor-derived variables, i.e., FA, MD, and the orientation of the major eigen-vector for the proposed method and the three compared methods. For each of these variables, we computed the average error (across all brain voxels) separately for each test subject, thereby obtaining one error value for each parameter and subject. The numbers in Table 1 and the other tables in this paper show the mean ± standard deviation across subjects. On all 40 test subjects from the dHCP datasets our method achieved lower estimation errors than all other methods. As shown in the table, compared with other methods, our proposed method has reduced the error in MD by factors of approximately 2.6–9.8 and the error in FA and orientation of major eigen-vector by factors of approximately 1.5–2.2. We ran paired t-tests to determine the statistical significance of the differences between our method and the other methods. These tests showed that the errors for the proposed method were significantly lower ($p < 0.001$) than those of the three compared methods.

Fig. 2 shows examples images reconstructed with the proposed method and CWLLS for two test subjects from the dHCP dataset. It shows two of the tensor channels, FA, MD, and color-FA images. Due to space limits, in this figure and the following figures we show the results of selected compared methods. The results shown in Fig. 2 clearly demonstrate a substantial advantage for the proposed method compared with CWLLS. The parameters estimated with the proposed method using six diffusion-weighted measurements are close to the reference images obtained with 88 measurements. On the other hand, CWLLS estimations are very noisy and contain large errors both in the gray matter area as well as in the location of major white matter tracts. Visually, the

Table 1
Estimation error on older (40–45 gestational weeks) and younger (29–36 gestational weeks) test subjects from the dHCP dataset. Bold type indicates statistically smaller errors at $p = 0.001$, computed using paired t-tests to compare our proposed method with every one of the competing methods.

Test subjects	Method	Tensor (×1000), mm^2s^{-1}	MD (×1000), mm^2s^{-1}	FA	Angle (degrees)
Older neonates ($n = 20$)	CWLLS	0.450 ± 0.040	0.413 ± 0.042	0.155 ± 0.017	20.2 ± 2.45
	CNLS	0.438 ± 0.044	0.410 ± 0.041	0.149 ± 0.019	20.3 ± 2.48
	CNN-DTI	0.393 ± 0.037	0.345 ± 0.037	0.124 ± 0.011	17.2 ± 2.20
	Proposed	**0.118 ± 0.019**	**0.042 ± 0.029**	**0.071 ± 0.003**	**11.6 ± 2.07**
Younger neonates ($n = 20$)	CWLLS	0.430 ± 0.058	0.376 ± 0.050	0.141 ± 0.019	18.5 ± 2.85
	CNLS	0.420 ± 0.060	0.370 ± 0.050	0.144 ± 0.022	18.7 ± 2.87
	CNN-DTI	0.401 ± 0.049	0.325 ± 0.041	0.137 ± 0.017	16.4 ± 2.34
	Proposed	**0.122 ± 0.018**	**0.123 ± 0.022**	**0.073 ± 0.012**	**10.3 ± 2.75**

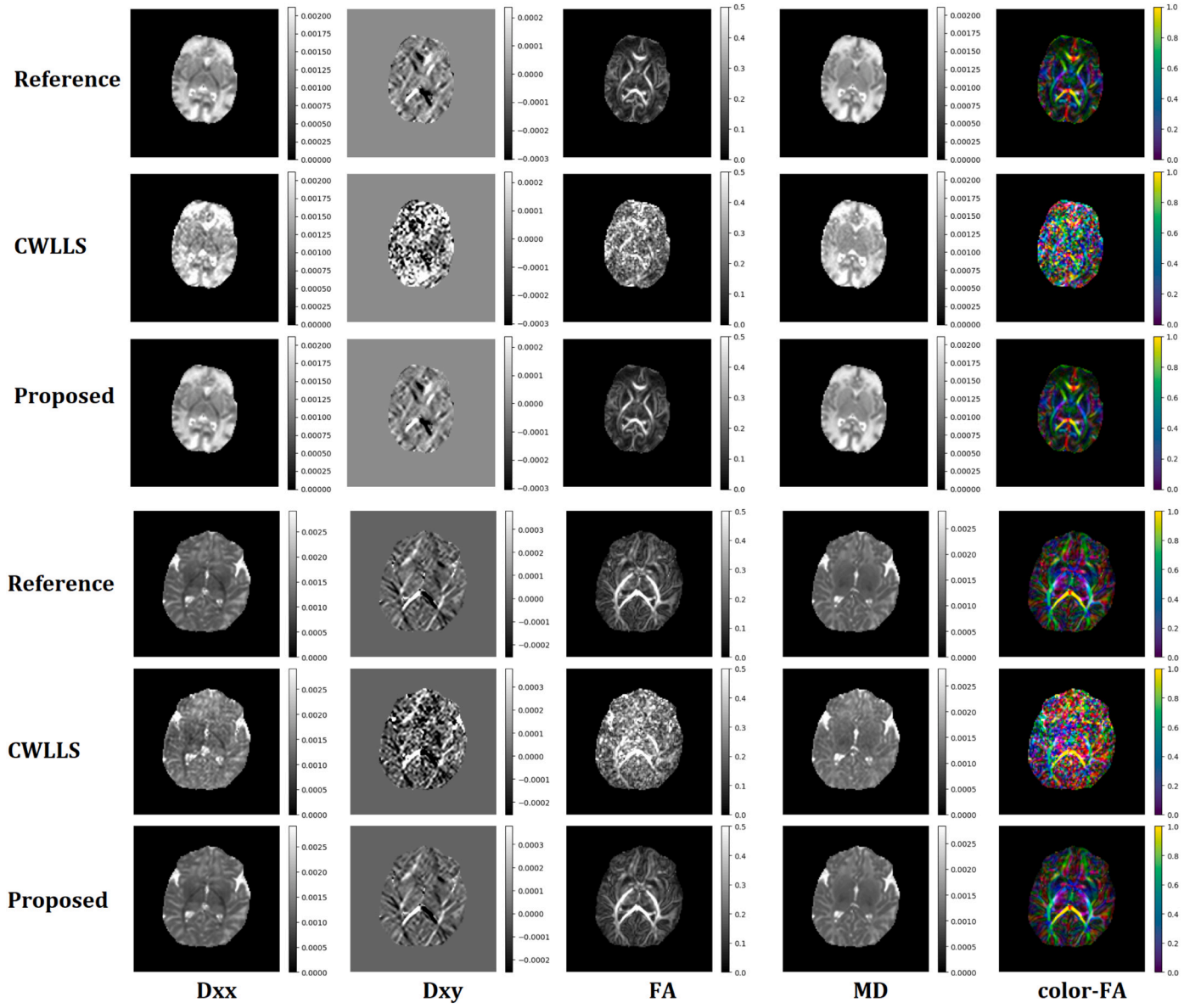

Fig. 2. Example tensor images estimated with CWLLS and the proposed method and their corresponding tensor-derived parameters. These images were reconstructed from scans of two neonatal subjects from the dHCP dataset. The top subject has a gestational age of 41 weeks, which is within the age range of subjects that have been used for training (i.e., 40–45 gestational weeks). The bottom subject has a gestational age of 31 weeks, which is much younger than the age range of the training subjects. For each subject, we have shown two of the six tensor channels (i.e., D_{xx} and D_{xy}), FA, MD, and color-FA.

superiority of our method compared with CWLLS can be best seen in the color-FA images. These are standard color-coded FA images that display the tensor anisotropy and the orientation of the major eigenvector in a single image. These images show that our method can accurately estimate the diffusion tensor throughout the brain.

From the examples shown in Fig. 2, the images reconstructed with our method seem to be less noisy than the reference image. Although the reference image is based on 88 measurements, it is computed by fitting the diffusion tensor on a voxel-wise basis, i.e., by considering the diffusion signal in each voxel, one at a time. Our proposed method, on the other hand, uses the correlations between the diffusion signal and tensor values among L^3 neighboring voxels. For further visual assessment of the reconstruction results, in Fig. 3 we have shown example one-dimensional profiles from FA and MD images reconstructed with our method and CWLLS compared with the reference. Furthermore, in Fig. 4 we have shown glyph visualization of the tensors. The profiles in Fig. 3 show that, compared with the reference, the reconstructions produced by our method show no significance loss of edge sharpness and are close to the reference image in almost all locations. The reconstructions produced with CWLLS are noisy and very different from the reference. The tensor visualizations presented in Fig. 4 further show that our method is close to the reference for most brain regions. CWLLS reconstructions, on the other hand, show large estimation errors in this figure.

In order to investigate the effect of measurement noise on the performance of our proposed method, we conducted a simulation experiment where we added simulated noise to the scans from the dHCP dataset. Specifically, independent and identically-distributed Rician noise with signal-to-noise ratio (SNR) in the range [15,50] dB was added. An SNR of 15 dB is close to the lowest SNR reported or used for simulation in prior works [18,60]. Table 2 shows the results of this experiment in terms of error in FA and the orientation of the major tensor eigen-vector for the proposed method and CWLLS. The noise had a much smaller effect on our proposed method than on CWLLS. For our method, the error in FA and the orientation of the major eigenvector

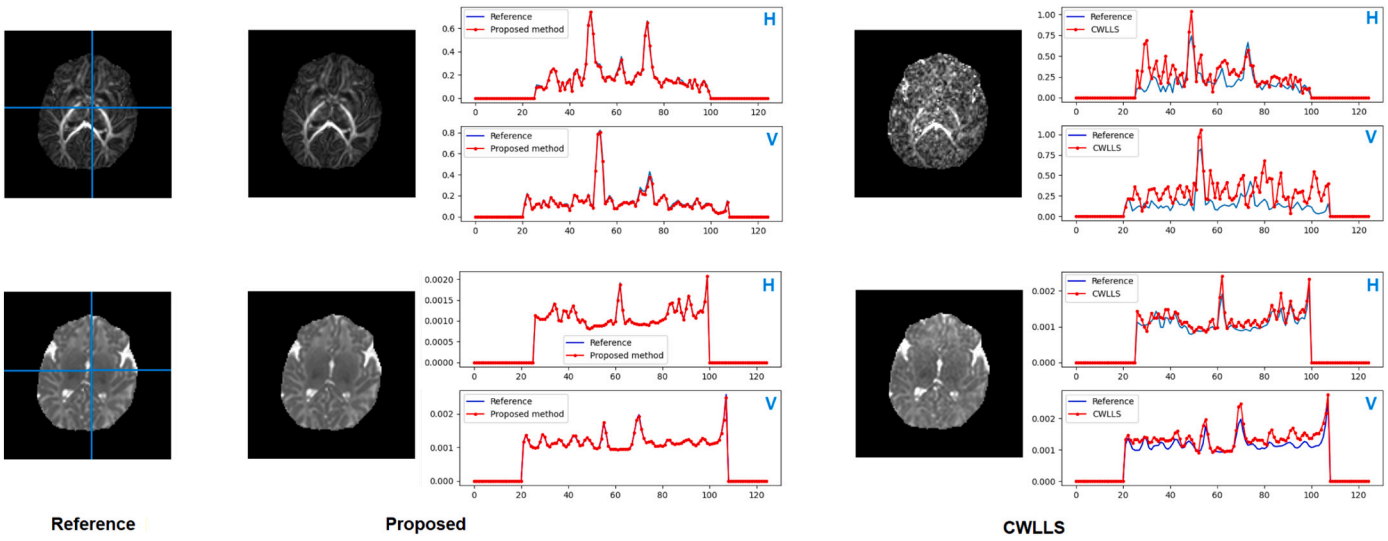

Fig. 3. Example one-dimensional profiles of FA (top) and MD (bottom) images reconstructed by the proposed method and CWLLS. For each image, we have shown one horizontal (H) and one vertical (V) profile. The locations of these profiles have been marked with the blue lines on the reference images (left-most column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

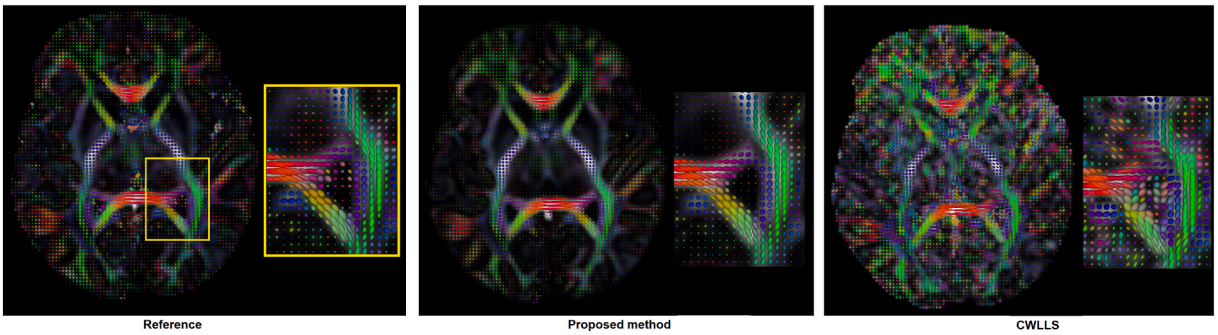

Fig. 4. Example glyph visualizations of the diffusion tensor images reconstructed with the proposed method and CWLLS alongside the reference tensors.

Table 2
Reconstruction error in terms of the angle of the major tensor eigen-vector and FA for CWLLS and the proposed method for different amounts of simulated noise added to the signal. The test subjects in this experiment were the 20 older neonates that were used in the experiment reported in Table 1. The first column shows the case where no noise was added to the signal. Bold type indicates statistically smaller errors at $p = 0.001$, computed using paired t-tests to compare our proposed method with CWLLS.

Parameter	Method	No added noise	SNR = 50 dB	SNR = 30 dB	SNR = 20 dB	SNR = 15 dB
angle (degrees)	CWLLS	20.2 ± 2.45	20.5 ± 2.41	21.2 ± 2.49	22.8 ± 2.55	24.6 ± 2.56
	Proposed	**11.6 ± 2.07**	**11.7 ± 2.09**	**11.9 ± 2.05**	**12.0 ± 2.11**	**12.2 ± 2.15**
FA	CWLLS	0.155 ± 0.017	0.157 ± 0.017	0.165 ± 0.020	0.172 ± 0.025	0.183 ± 0.026
	Proposed	**0.071 ± 0.003**	**0.071 ± 0.003**	**0.071 ± 0.005**	0.073 ± 0.006	**0.074 ± 0.006**

increased by 4.2 % and 5.2 %, respectively, as we reduced the SNR from 50 dB to 15 dB. Comparatively, the errors for CWLLS increased by 18.1 % and 20.0 %, respectively. Fig. 5 shows example slices and profiles of FA and MD images reconstructed with our proposed method at SNR = 15. They show that, even at this low SNR, reconstructions of our proposed method are close to the reference image.

Fig. 6 shows three example whole-brain connectomes generated from the diffusion tensors estimated with the proposed method and CNN-DTI. Tractography provides an indirect assessment of diffusion tensor estimation accuracy because the tractography results depend on the settings of the fiber tracing algorithm. For a fair comparison, we used the same seed locations and the same fiber tracing algorithm for different methods. Specifically, we used the white matter mask provided as part of the dHCP dataset for seeding. Moreover, we used the EuDX tractography algorithm [20], which is a fiber tracing algorithm that relies heavily on voxel-wise fiber directions instead of imposing global fiber priors. We used a step size of 0.5 mm. As shown in these example figures, the connectome produced with our proposed method from only 6 measurements is similar to the one produced based on D_{ref} reconstructed from 88 measurements. The connectome produced with CNN-DTI, on the other hand, lacks much of the major white matter tracts. Fig. 7 shows example connectivity maps between 87 brain regions provided for each dHCP scan, computed from the whole-brain connectomes. As shown in this example, the connectivity map for the proposed method from only 6 measurements is similar to that of the reference connectome computed from 88 measurements. The connectivity maps for CNN-DTI and CWLLS, on the other hand, are very different and lack many of the connections that are present in the

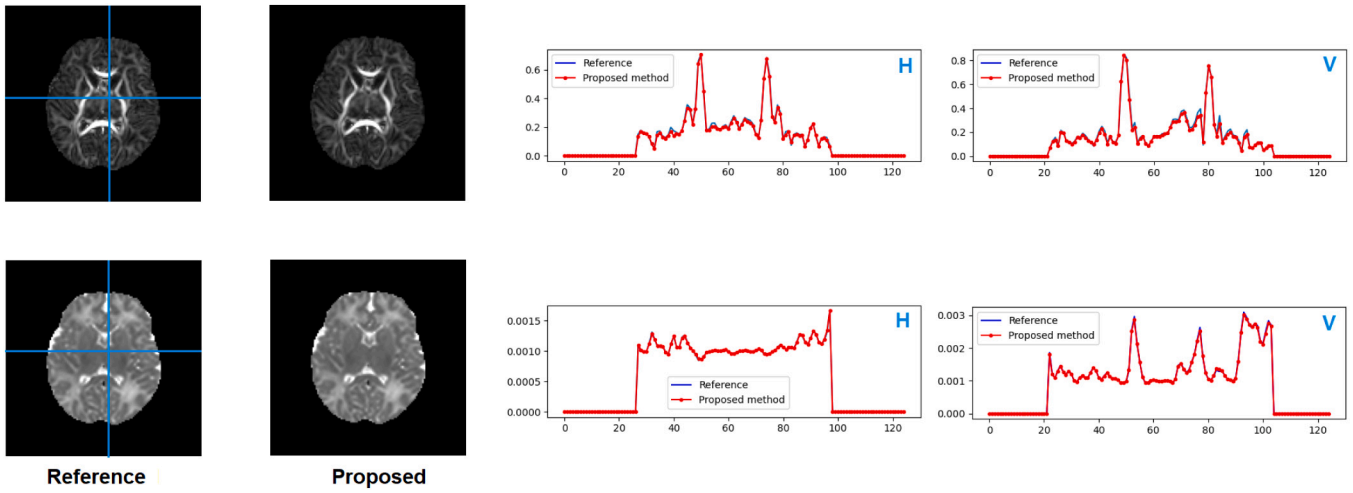

Fig. 5. Example slices (left) and one-dimensional profiles (right) from the FA and MD images reconstructed with the proposed method with added simulated Rician noise with an SNR of 15 dB. For each image, we have shown one horizontal (H) and one vertical (V) profile. The locations of these profiles have been marked with the blue lines on the corresponding reference image slice (left-most column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

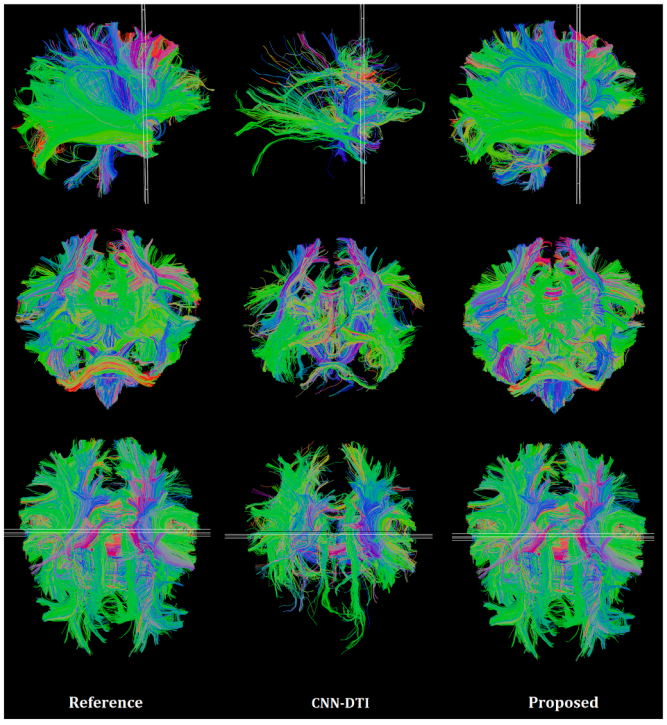

Fig. 6. Example whole-brain connectomes produced with the EuDX tractography algorithm based on the reference diffusion tensor estimated from 88 measurements and the diffusion tensors estimated with our proposed method and CNN-DTI from only 6 measurements.

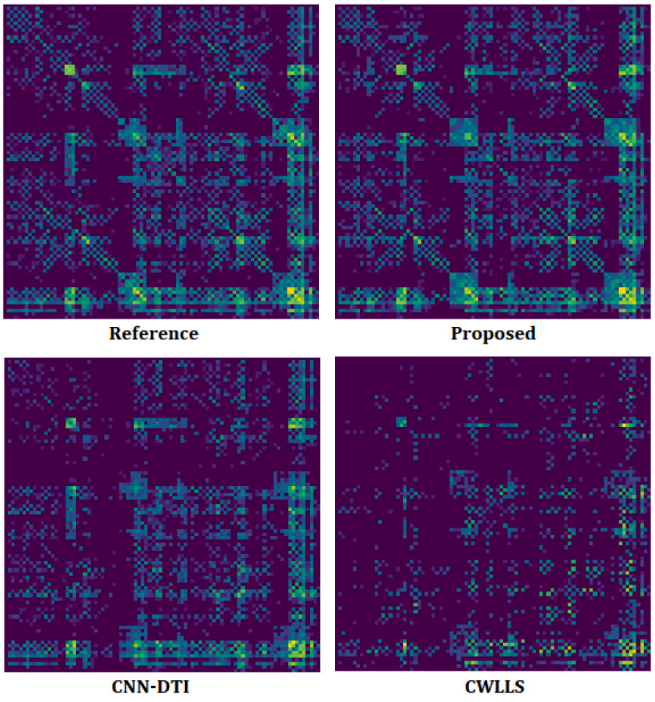

Fig. 7. The connectivity map between 87 brain regions for a test subject in the dHCP dataset. The connectivity maps were computed from the whole-brain connectomes derived based on the reference diffusion tensor (estimated from 88 measurements) and the diffusion tensors estimated with the proposed method, CWLLS, and CNN-DTI (estimated from 6 measurements).

reference connectivity map. For a quantitative comparison of the connectivity matrices, we considered 100 off-diagonal elements of the reference connectivity matrix with the largest values, i.e., the 100 pairs of brain regions with the strongest connections. We compared the values of those matrix elements between the reference connectivity matrix and the connectivity matrix computed with our method using a paired t-test with a significance threshold of $p = 0.001$. The test showed that the connectivity matrix for our method was not different from the Reference ($p = 0.32$). On the other hand, the comparison of the Reference connectivity matrix and the connectivity matrix estimated with CWLLS using the same statistical significance test showed a significant difference ($p < 0.001$). Similarly, for CNN-DTI the difference with the Reference connectivity matrix was significant ($p < 0.001$).

With regard to architectural hyper-parameters, increasing the patch size (L) to 7 did not improve the estimation accuracy of the proposed method. Note that the number of signals in the sequence grows cubically with L. Increasing L from 5 to 7 would increase the sequence length from 125 to 343. Our experiments show that the transformer modules become difficult to train with $L \geq 6$. On the other hand, reducing L to 3 significantly increased the method's estimation error, obviously because of the

reduced spatial context to utilize in the estimation. We also found that increasing the number of transformer modules in Models S and ST beyond 4 did not improve the accuracy of our method. Furthermore, increasing the projection dimension and the number of self-attention heads to values larger than our default values (i.e., $d_h = 512$ and $n_h = 2$) did not improve the estimation accuracy. On the other hand, our two-stage estimation approach proved to be effective.

Fig. 8 shows example tensor and FA images reconstructed with Model S and Model ST and the corresponding error maps, i.e., the difference between the estimated values and the reference. Model ST consistently improved the estimation accuracy of Model S. For example, the tensor estimation error for Model S on older neonates and younger neonates from the dHCP dataset were, respectively, 0.154 ± 0.028 and 0.168 ± 0.031. Paired t-tests showed that these errors were significantly higher than the errors for Model ST reported in Table 1, although they were significantly lower than the errors for CWLLS, CNLS, and CNN-DTI ($p < 0.001$). We performed extensive experiments to investigate whether the accuracy of Model S could be improved to match Model ST by changing hyper-parameters (L, d_h, n_h, number of transformer modules, and training settings). However, Model ST was consistently more accurate than Model S, regardless of hyper-parameter settings. These observations support and justify our proposed two-stage approach that exploits the spatial correlations in the diffusion signal as well as in the tensors.

The error maps shown in Fig. 8 display some spatial structure for both Model S and Model ST. To examine these spatial structures, we computed FA and MD estimation errors separately for white matter (WM), gray matter (GM) and CSF voxels. Furthermore, we computed these errors for four different structures in WM. The results of this analysis are presented in Table 4. The first observation from this table is that, for both FA and MD, the reconstruction errors in WM are less than those in GM and CSF. However, for the four different WM sub-structures the errors are very similar. We performed paired t-tests to compare the FA and MD reconstruction errors for these four structures, separately for Model S and Model ST. None of these tests showed statistical significance at $p = 0.001$. On the other hand, on all three tissue types (WM, GM, CSF) and all four WM structures shown in this table, the FA and MD estimation errors for Model ST were significantly smaller ($p < 0.001$) than those for Model S.

Fig. 9 shows example Bland-Altman plots for the FA and MD errors for Model S and Model ST. We have shown plots for Model T versus the reference (left column), Model ST versus the reference (middle column), and Model T versus Model ST (right column). In the plots for Model T versus Model ST, because none of them can be considered the reference, we have used the average of the two models as the horizontal axis as suggested in [13,22]. These plots show that for both FA and MD, both models tend to have larger errors for larger values of FA and MD. However, for both FA and MD, Model ST shows fewer large errors. For both FA and MD, the bias (i.e., the mean error compared with the reference) of Model ST is smaller than the bias of Model S. The plots in the right column in this figure also show that the "corrections" made by Model ST over Model S are over all values of FA and MD.

Table 3 and Fig. 10 show the results of further evaluations of our method on 20 test subjects from the PING dataset and 7 VOGM patients. Each of the scans in these two datasets included 30 measurements in the $b = 1000$ shell. For each scan in these datasets, we used all 30 measurements to reconstruct the reference image. We then selected six of the $b = 1000$ measurements (and one $b = 0$ measurement for normalization) for reconstruction with the proposed method and the three competing methods in the same way as described above for the experiments with the dHCP dataset. The VOGM patients' age range was 0–3 years and the PING subjects' age range was 3–20 years. Because of higher myelination

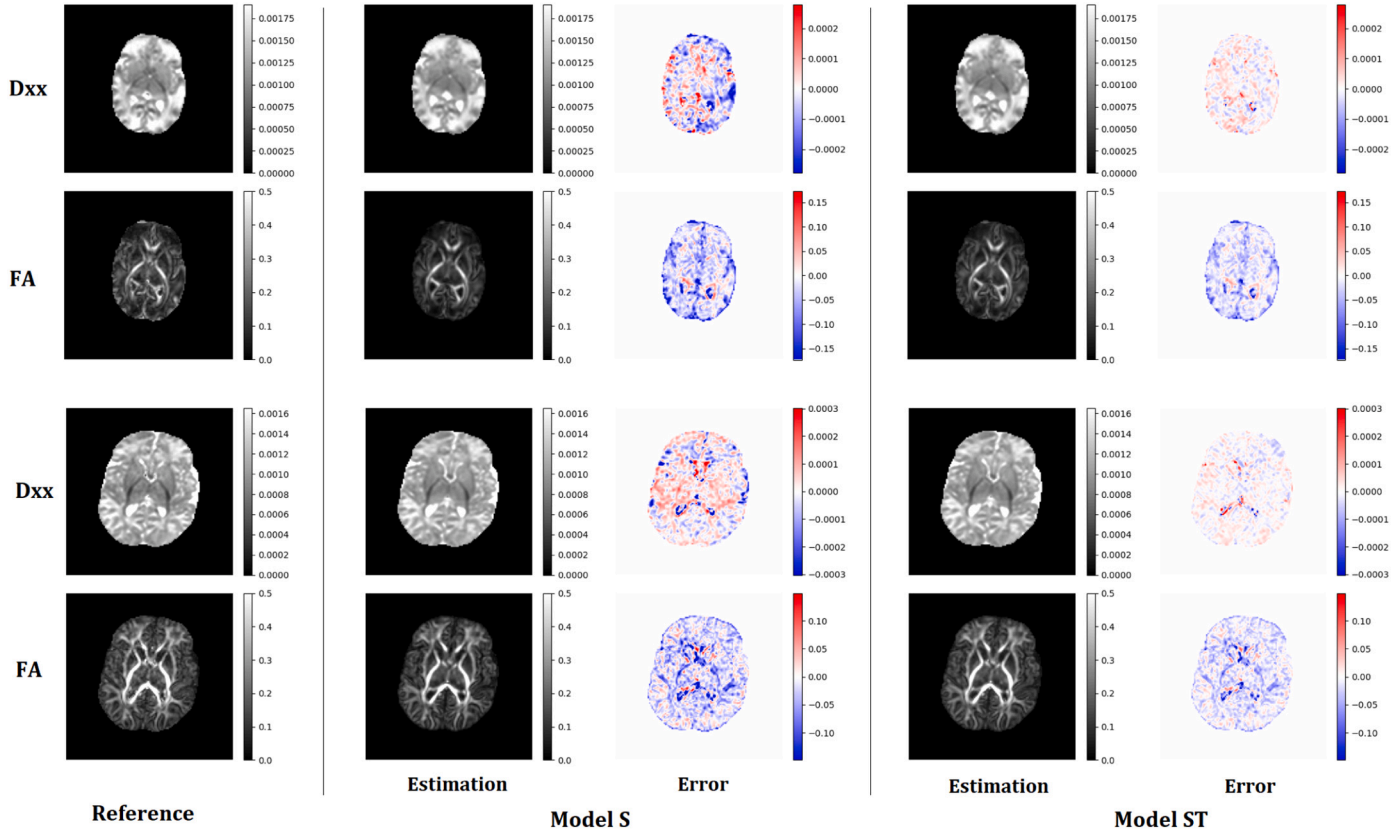

Fig. 8. Comparison of estimation errors for Model S and Model ST in our proposed method. We have shown the results on two subjects from the dHCP dataset; an older subject (top) and a younger subject (bottom). As shown in Fig. 1, Model S uses the diffusion signals to predict the tensor values. Model ST, on the other hand, builds upon the estimations of Model S by learning to incorporate the spatial correlations in the diffusion signal and tensor values.

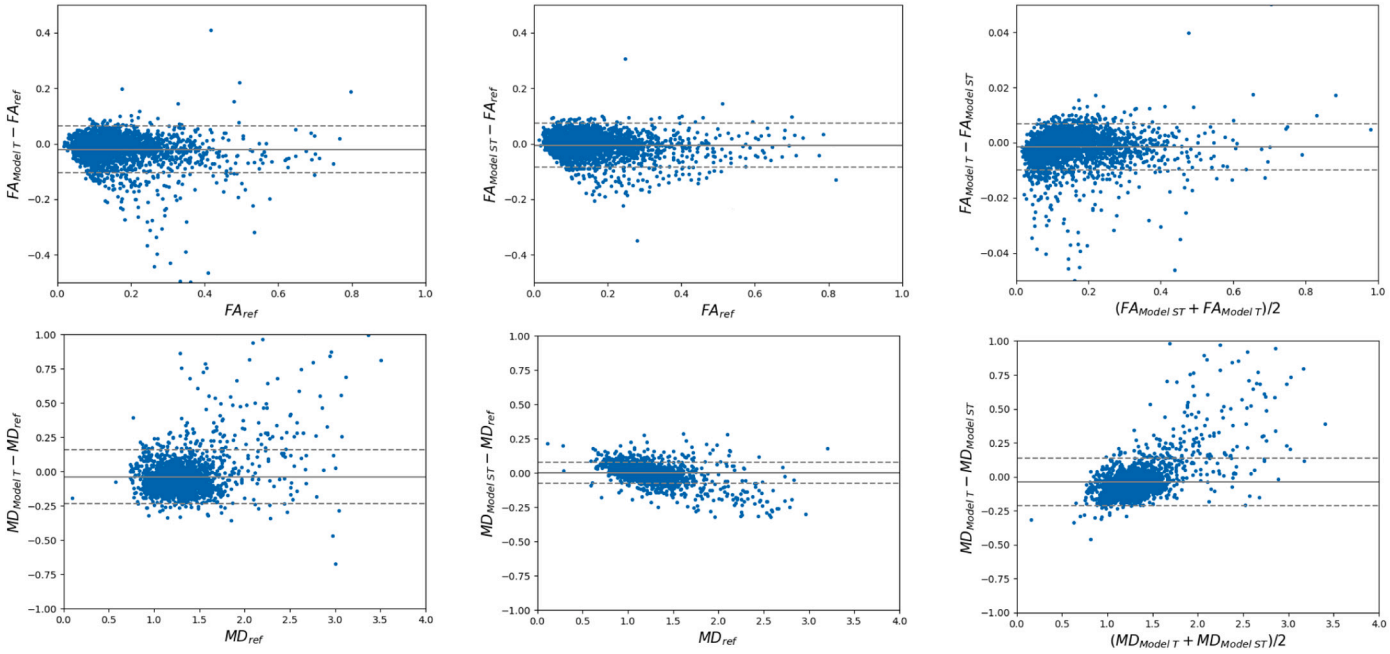

Fig. 9. Bland Altman plots for FA (top) and MD (bottom). Plots have been shown for Model T versus the reference (left column), Model ST versus the reference (middle column), and Model T versus Model ST (right column). To simplify the presentation we have multiplied values of MD by 1000 and they are in units of mm^2s^{-1}.

Table 3
Average and standard deviation of the estimation error for CWLLS and the proposed method on scans of 20 subjects from the PING dataset and 7 VOGM patients. Bold type indicates statistically lower errors at $p = 0.001$ computed using paired t-tests to compare our proposed method with every one of the competing methods.

Test subjects	Method	tensor (×1000), mm^2s^{-1}	MD (×1000), mm^2s^{-1}	FA	Angle (degrees)
PING ($n = 20$)	CWLLS	0.320 ± 0.035	0.317 ± 0.040	0.107 ± 0.022	18.5 ± 3.30
	CNLS	0.322 ± 0.037	0.319 ± 0.040	0.110 ± 0.025	18.9 ± 3.30
	CNN-DTI	0.299 ± 0.030	0.277 ± 0.032	0.103 ± 0.020	17.8 ± 3.07
	Proposed	**0.174 ± 0.021**	**0.101 ± 0.033**	**0.073 ± 0.011**	**12.4 ± 2.28**
VOGM ($n = 7$)	CWLLS	0.407 ± 0.054	0.350 ± 0.049	0.125 ± 0.016	18.0 ± 2.53
	CNLS	0.401 ± 0.055	0.364 ± 0.050	0.127 ± 0.017	18.2 ± 2.77
	CNN-DTI	0.346 ± 0.034	0.329 ± 0.042	0.114 ± 0.012	15.9 ± 2.22
	Proposed	**0.147 ± 0.031**	**0.176 ± 0.028**	**0.082 ± 0.014**	**11.5 ± 2.69**

and better data quality, competing methods achieved more accurate estimations on these two datasets than on the dHCP dataset. Moreover, compared with the experiments on the dHCP dataset above, in these experiments D_{ref} is more biased towards D_{CWLLS} and, to some extent, D_{CNLS}. This is because D_{CWLLS} and D_{CNLS} are reconstructed using 6 out of the 30 measurements used to reconstruct D_{ref} and they are all reconstructed using least-square principles. Nonetheless, our proposed method achieved significantly lower errors on both of these additional datasets, as indicated in Table 3. Compared with the competing methods, our proposed method reduced the estimation error in MD by factors of $1.9 - 3.2$, error in FA by factors of $1.4 - 1.55$, and error in the orientation of the major eigen-vector by factors of $1.38 - 1.58$. The superiority of the proposed method can be visually observed in the examples shown in Fig. 10. Due to space limits, we have shown the results of our proposed method and CNN-DTI, which was slightly more accurate than CWLLS and CNLS. The differences between our method and CNN-DTI is especially more clear in the color-FA images that encode both the degree of anisotropy and the direction of the major eigen-vector in the same image.

Our study has some limitations that need to be mentioned. First, we only considered the diffusion tensor reconstruction from six measurements. The goal of this paper was to demonstrate the potential gains of exploiting the spatial correlations and the capability of attention-based neural networks to learn these correlations. Therefore, to enhance the focus of the study we used a fixed measurement scheme as input to our network. Although we anticipate that the use of these neural network models should be useful for other DW-MRI models and other measurement schemes, those could be investigated in future works. Moreover, although we have evaluated our method on three different and independent datasets, an application-specific evaluation may be warranted if our method is to be used for a specific clinical application. For example, it is not clear how our method (using six measurements) would compare with the reference image (using much larger measurement sets) if the reconstructions are to be used for detecting subtle changes in the brain micro-structure due to lesions or other pathologies. Therefore, further evaluation of our proposed method for specific clinical patient populations may be useful.

4. Conclusions

Diffusion tensor imaging is increasingly being used to study brain development and degeneration. However, the same least squares-based estimation methods [36] have been commonly used in the past two decades. Standard estimation methods, which are based on linear or non-linear fitting of measurements on a per-voxel basis, can be highly sub-optimal. The main intuition of our work is that state of the art deep neural networks for modeling signal sequences can be adapted to develop accurate DTI estimation methods. In particular, the transformer models that have been used in this work are capable of learning very complicated correlations between signals in a sequence. Using this

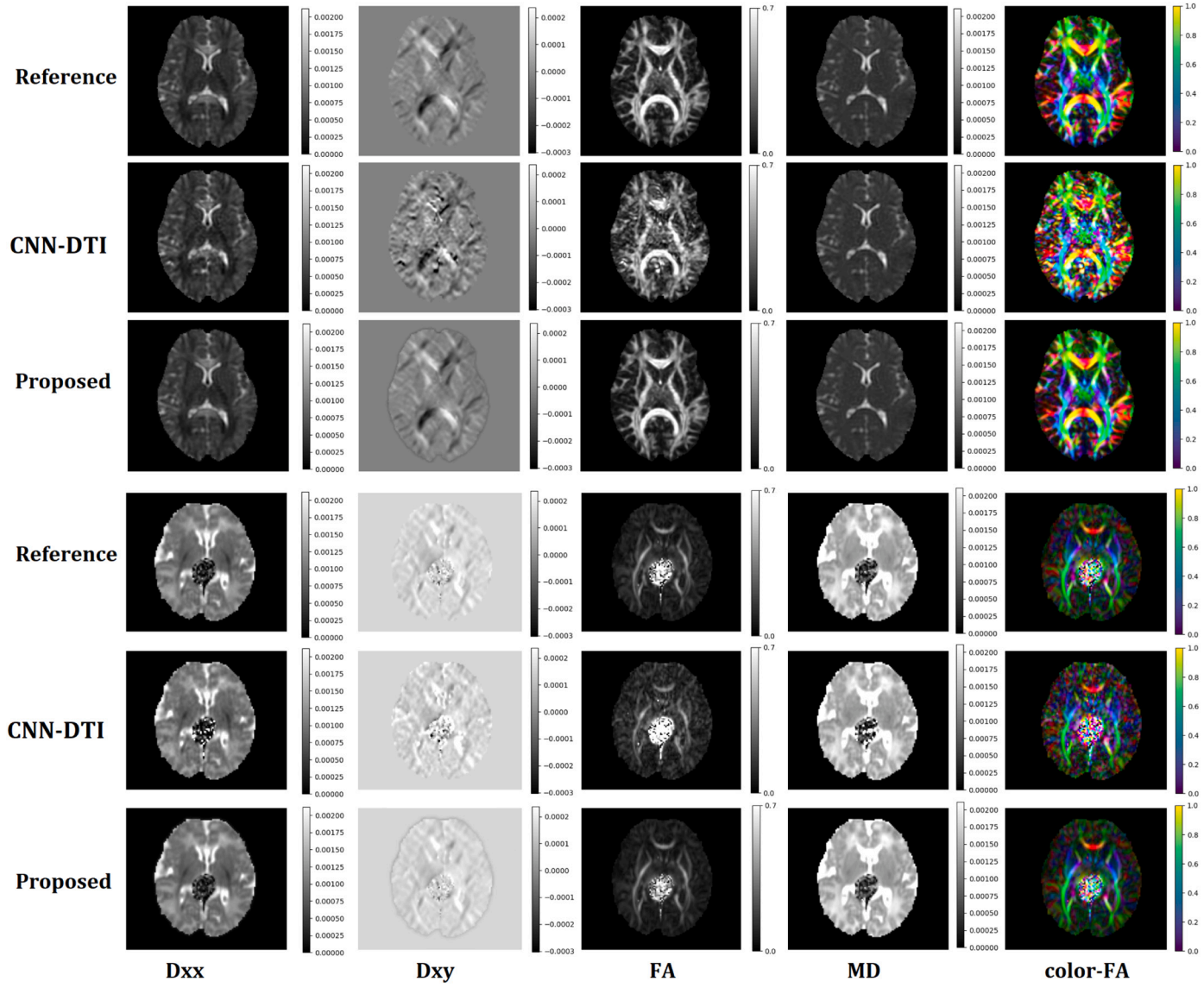

Fig. 10. Example tensor images and tensor-derived parameters obtained with CNN-DTI and the proposed method on a subject from the PING dataset (top) and a VOGM patient (bottom). The PING subject was 14 years old at scan time. The VOGM patient was 2 months old at scan time. The location of malformation in the brain of the VOGM patient is clearly visible in the reference image. For each subject, we have shown two tensor channels (D_{xx} and D_{xy}), FA, MD, and color-FA.

Table 4
FA and MD reconstruction errors for Model S and Model ST for selected tissue types: white matter (WM), gray matter (GM), and Cerebrospinal Fluid (CSF). Additionally, errors have been shown for four white matter structures: corpus callosum (CC), temporal lobe white matter (TLWM), occipital lobe white matter (OLWM), and frontal lobe white matter (FLWM). The test subjects in this experiment were 20 older neonates from the dHCP dataset. Bold type indicates statistically smaller errors at $p = 0.001$, computed using paired t-tests. As in the other tables, to simplify the presentation we have multiplied the MD errors by 1000 and they are in units of mm^2s^{-1}.

	Method	WM	GM	CSF	CC	TLWM	OLWM	FLWM
FA	Model S	0.058 ± 0.004	0.066 ± 0.006	0.092 ± 0.007	0.060 ± 0.004	0.058 ± 0.005	0.057 ± 0.004	0.057 ± 0.006
	Model ST	**0.054 ± 0.004**	**0.061 ± 0.005**	**0.082 ± 0.007**	**0.053 ± 0.006**	**0.055 ± 0.004**	**0.054 ± 0.005**	**0.052 ± 0.005**
MD	Model S	0.038 ± 0.020	0.045 ± 0.018	0.058 ± 0.015	0.037 ± 0.022	0.037 ± 0.024	0.040 ± 0.020	0.040 ± 0.023
	Model ST	**0.035 ± 0.018**	**0.041 ± 0.017**	**0.050 ± 0.016**	**0.034 ± 0.020**	**0.034 ± 0.021**	**0.036 ± 0.017**	**0.035 ± 0.015**

model, we developed a method that was able to learn spatial correlations between diffusion signals and tensor values in neighboring voxels for accurate tensor estimation. On the challenging neonatal DW-MRI scans from the dHCP dataset, our method reduced the estimation error, compared with three competing methods, by factors of 1.5–9.8. The estimations of the proposed method were close to the reference estimations that were obtained using 88 measurements. Our method also led to superior tractography and connectivity analysis. Furthermore, our method showed highly accurate and robust estimation on two additional datasets. These observations demonstrate the significant potential of the proposed method for improving the accuracy of DTI estimation, especially for challenging cohorts such as neonates and infants. Therefore, our method can facilitate DTI studies to detect subtle changes in brain, while also reducing the scan time.

Acknowledgments

The dHCP Data were provided by the developing Human Connectome Project, KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. [319456]. We are grateful to the families who generously supported this trial. Collection and sharing of the PING dataset has been funded by the Pediatric Imaging, Neurocognition and Genetics Study (PING) (National Institutes of Health Grant RC2DA029475). PING is funded by the National Institute on Drug Abuse and the Eunice Kennedy Shriver National Institute of Child Health \& Human Development. PING data are disseminated by the PING Coordinating Center at the Center for Human Development, University of California, San Diego.

Declaration of competing interest

The authors do not have any conflicts of interest (financial or otherwise) to dosclose.

References

[1] Alexander AL, Lee JE, Lazar M, Field AS. Diffusion tensor imaging of the brain. Neurotherapeutics 2007;4:316–29.
[2] Alexander DC, Barker GJ. Optimal imaging parameters for fiber-orientation estimation in diffusion mri. Neuroimage 2005;27:357–67.
[3] Aliotta E, Nourzadeh H, Patel SH. Extracting diffusion tensor fractional anisotropy and mean diffusivity from 3-direction dwi scans using deep learning. Magn Reson Med 2021;85:845–54.
[4] Aliotta E, Nourzadeh H, Sanders J, Muller D, Ennis DB. Highly accelerated, model-free diffusion tensor mri reconstruction using neural networks. Med Phys 2019;46:1581–91.
[5] de Almeida Martins JP, Nilsson M, Lampinen B, Palombo M, While PT, Westin CF, Szczepankiewicz F. Neural networks for parameter estimation in microstructural mri: application to a diffusion-relaxation model of white matter. Neuroimage 2021;244:118601.
[6] Armitage P, Bastin M. Utilizing the diffusion-to-noise ratio to optimize magnetic resonance diffusion tensor acquisition strategies for improving measurements of diffusion anisotropy. Magn Reson Med 2001;45:1056–65.
[7] Assaf Y, Basser PJ. Composite hindered and restricted model of diffusion (charmed) mr imaging of the human brain. Neuroimage 2005;27:48–58.
[8] Barnea-Goraly N, Kwon H, Menon V, Eliez S, Lotspeich L, Reiss AL. White matter structure in autism: preliminary evidence from diffusion tensor imaging. Biol Psychiatry 2004;55:323–6.
[9] Barnea-Goraly N, et al. White matter development during childhood and adolescence: a cross-sectional diffusion tensor imaging study. Cereb Cortex 2005;15:1848–54.
[10] Basser PJ, Pierpaoli C. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri. J Magn Reson 2011;213:560–70.
[11] Bastiani M, et al. Automated processing pipeline for neonatal diffusion mri in the developing human connectome project. Neuroimage 2019;185:750–63.
[12] Basu S, Fletcher T, Whitaker R. Rician noise removal in diffusion tensor mri. In: International conference on medical image computing and computer-assisted intervention. Springer; 2006. p. 117–25.
[13] Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;327:307–10.
[14] Chang LC, Jones DK, Pierpaoli C. Restore: robust estimation of tensors by outlier rejection. Magn Reson Med 2005;53:1088–95.
[15] Chang LC, Walker L, Pierpaoli C. Informed restore: a method for robust estimation of diffusion tensor from low redundancy datasets in the presence of physiological noise artifacts. Magn Reson Med 2012;68:1654–63.
[16] Chung S, Lu Y, Henry RG. Comparison of bootstrap approaches for estimation of uncertainties of dti parameters. Neuroimage 2006;33:531–41.
[17] Cook PA, Symms M, Boulby PA, Alexander DC. Optimal acquisition orders of diffusion-weighted mri measurements. J. Magn. Reson. Imaging 2007;25:1051–8.
[18] Dell'Acqua F, Scifo P, Rizzo G, Catani M, Simmons A, Scotti G, Fazio F. A modified damped richardson–lucy algorithm to reduce isotropic background effects in spherical deconvolution. Neuroimage 2010;49:1446–58.
[19] Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint; 2020. arXiv:2010.11929.
[20] Garyfallidis E. Towards an accurate brain tractography. Ph.D. thesis. University of Cambridge; 2013.
[21] Garyfallidis E, et al. Dipy, a library for the analysis of diffusion mri data. Front Neuroinform 2014;8:8.
[22] Giavarina D. Understanding bland altman analysis. Biochem Med 2015;25:141–51.
[23] Gibbons EK, et al. Simultaneous noddi and gfa parameter map generation from subsampled q-space imaging using deep learning. Magn Reson Med 2019;81:2399–411.
[24] Golkov V, et al. Q-space deep learning: twelve-fold shorter and model-free diffusion mri scans. IEEE Trans Med Imaging 2016;35:1344–51.
[25] Goodlett C, Fletcher PT, Lin W, Gerig G. Quantification of measurement error in dti: theoretical predictions and validation. In: Ayache N, Ourselin S, Maeder A, editors. Medical image computing and computer-assisted intervention – MICCAI 2007. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 10–7.
[26] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: IEEE international conference on computer vision (ICCV) 2015; 2015.
[27] Jernigan TL, et al. The pediatric imaging, neurocognition, and genetics (ping) data repository. Neuroimage 2016;124:1149–54.
[28] Jones DK. The effect of gradient sampling schemes on measures derived from diffusion tensor mri: a Monte Carlo study. Magn Reson Med 2004;51:807–15.
[29] Jones DK. Gaussian modeling of the diffusion signal. In: Diffusion MRI. Elsevier; 2009. p. 37–54.
[30] Jones DK, Basser PJ. "squashing peanuts and smashing pumpkins": how noise distorts diffusion-weighted mr data. Magn Reson Med 2004;52:979–93.
[31] Jones DK, Horsfield MA, Simmons A. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. Magn Reson Med 1999;42:515–25.
[32] Karimi D, Jaimes C, Machado-Rivas F, Vasung L, Khan S, Warfield SK, Gholipour A. Deep learning-based parameter estimation in fetal diffusion-weighted mri. Neuroimage 2021;243:118482.
[33] Karimi D, Vasung L, Machado-Rivas F, Jaimes C, Khan S, Gholipour A. Accurate parameter estimation in fetal diffusion-weighted mri-learning from fetal and newborn data. In: International conference on medical image computing and computer-assisted intervention. Springer; 2021. p. 487–96.
[34] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations (ICLR); 2014.
[35] Kingsley PB. Introduction to diffusion tensor imaging mathematics: part iii. tensor calculation, noise, simulations, and optimization. Concepts Magn. Reson. A 2006;28:155–79.
[36] Koay CG, Chang LC, Carew JD, Pierpaoli C, Basser PJ. A unifying theoretical and algorithmic framework for least squares methods of estimation in diffusion tensor imaging. J Magn Reson 2006;182:115–25.
[37] Koppers S, Friedrichs M, Merhof D. Reconstruction of diffusion anisotropies using 3d deep convolutional neural networks in diffusion imaging. In: Modeling, analysis, and visualization of anisotropy. Springer; 2017. p. 393–404.
[38] Koppers S, Haarburger C, Edgar JC, Merhof D. Reliable estimation of the number of compartments in diffusion mri. In: Bildverarbeitung für die Medizin 2017. Springer; 2017. p. 203–8.
[39] Koppers S, Merhof D. Direct estimation of fiber orientations using deep learning in diffusion imaging. In: International workshop on machine learning in medical imaging. Springer; 2016. p. 53–60.
[40] Kubicki M, et al. A review of diffusion tensor imaging studies in schizophrenia. J Psychiatr Res 2007;41:15–30.
[41] Le Bihan D, et al. Diffusion tensor imaging: concepts and applications. J Magn Reson Imaging 2001;13:534–46.
[42] Lebel C, Walker L, Leemans A, Phillips L, Beaulieu C. Microstructural maturation of the human brain from childhood to adulthood. Neuroimage 2008;40:1044–55.
[43] Li H. Superdti: Ultrafast dti and fiber tractography with deep learning. URL: Magn Reson Med 2021;0:1–14. https://doi.org/10.1002/mrm.28937. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.28937, https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.28937.
[44] Li Z, et al. Fast and robust diffusion kurtosis parametric mapping using a three-dimensional convolutional neural network. IEEE Access 2019;7:71398–411.
[45] Lin Z, et al. Fast learning of fiber orientation distribution function for mr tractography using convolutional neural network. Med Phys 2019;46:3101–16.
[46] Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv preprint; 2015.
[47] Morgan PS, Bowtell RW, McIntyre DJ, Worthington BS. Correction of spatial distortion in epi due to inhomogeneous static magnetic fields using the reversed gradient method. J Magn Reson Imaging 2004;19:499–507.
[48] Murphy KP. Machine learning: a probabilistic perspective. 2012.
[49] Nat V, et al. Deep learning reveals untapped information for local white-matter fiber reconstruction in diffusion-weighted mri. Magn Reson Imaging 2019;62:220–7.
[50] Nedjati-Gilani GL, et al. Machine learning based compartment models with permeability for white matter microstructure imaging. Neuroimage 2017;150:119–35.
[51] Neher PF, Cote MA, Houde JC, Descoteaux M, Maier-Hein KH. Fiber tractography using machine learning. Neuroimage 2017;158:417–29.
[52] Nunes RG, Jezzard P, Clare S. Investigations on the efficiency of cardiac-gated methods for the acquisition of diffusion-weighted images. J Magn Reson 2005;177:102–10.
[53] Palombo M, Ianus A, Guerreri M, Nunes D, Alexander DC, Shemesh N, Zhang H. Sandi: a compartment-based model for non-invasive apparent soma and neurite imaging by diffusion mri. Neuroimage 2020;215:116835.
[54] Reisert M, Kellner E, Dhital B, Hennig J, Kiselev VG. Disentangling micro from mesostructure by diffusion mri: a bayesian approach. Neuroimage 2017;147:964–75.
[55] Rohde GK, Barnett A, Basser P, Marenco S, Pierpaoli C. Comprehensive approach for correction of motion and distortion in diffusion-weighted mri. Magn Reson Med 2004;51:103–14.

[56] Schultz T. Learning a reliable estimate of the number of fiber directions in diffusion mri. In: International conference on medical image computing and computer-assisted intervention. Springer; 2012. p. 493–500.
[57] Shiv V, Quirk C. Novel positional encodings to enable tree-based transformers. Adv Neural Inf Proces Syst 2019;32:12081–91.
[58] Skare S, Hedehus M, Moseley ME, Li TQ. Condition number as a measure of noise performance of diffusion tensor data acquisition schemes with mri. J Magn Reson 2000;147:340–52.
[59] Tian Q, et al. Deepdti: high-fidelity six-direction diffusion tensor imaging using deep learning. Neuroimage 2020;219:117017.
[60] Tournier JD, Calamante F, Connelly A. Robust determination of the fibre orientation distribution in diffusion mri: non-negativity constrained super-resolved spherical deconvolution. Neuroimage 2007;35:1459–72.
[61] Tournier JD, et al. Mrtrix3: a fast, flexible and open software framework for medical image processing and visualisation. Neuroimage 2019;202:116137.
[62] Vaswani A, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017: 5998–6008.
[63] Wasserthal J, Neher PF, Hirjak D, Maier-Hein KH. Combined tract segmentation and orientation mapping for bundle-specific tractography. Med Image Anal 2019; 58:101559.
[64] Yigit Avci M, Li Z, Fan Q, Huang S, Bilgic B, Tian Q. Quantifying the uncertainty of neural networks using monte carlo dropout for deep learning based quantitative mri. arXiv e-prints; 2021. arXiv–2112.
[65] Zhang F, Karayumak SC, Hoffmann N, Rathi Y, Golby AJ, O'Donnell LJ. Deep white matter analysis (deepwma): fast and consistent tractography segmentation. Med Image Anal 2020;65:101761.
[66] Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC. Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 2012;61:1000–16.