IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 6, 2019, accepted May 22, 2019, date of publication May 27, 2019, date of current version June 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919241

Fast and Robust Diffusion Kurtosis Parametric Mapping Using a Three-Dimensional Convolutional Neural Network

ZHIWEI LI[1], TING GONG[2], ZHICHAO LIN[1], HONGJIAN HE[2], (Member, IEEE), QIQI TONG[2], CHEN LI[2], YI SUN[3], FENG YU[1], (Member, IEEE), AND JIANHUI ZHONG[2,4]

[1]Department of Instrument Science & Technology, Zhejiang University, Hangzhou 310027, China
[2]Key Laboratory for Biomedical Engineering, Center for Brain Imaging Science and Technology, College of Biomedical Engineering and Instrumental Science, Ministry of Education, Zhejiang University, Hangzhou 310027, China
[3]MR Collaboration NE Asia, Siemens Healthcare, Shanghai 200432, China
[4]Department of Imaging Sciences, University of Rochester, Rochester, NY 14642, USA

Corresponding authors: Hongjian He (hhezju@zju.edu.cn) and Jianhui Zhong (jzhong@zju.edu.cn)

This work was supported in part by grants from the National Key R&D Program of China under Grant 2017YFC0909200, in part by the National Natural Science Foundation of China under Grant 81871428 and Grant 91632109, in part by the Fundamental Research Funds for the Central Universities under Grant 2019QNA5026, in part by the Major Scientific Project of Zhejiang Lab under Grant 2018DG0ZX01, and in part by the Shanghai Key Laboratory of Psychotic Disorders under Grant 13dz2260500.

ABSTRACT Diffusion kurtosis imaging (DKI) is an advanced diffusion imaging method that captures complex brain microstructural properties; however, it often has a lengthy acquisition time compared to conventional diffusion tensor imaging (DTI). Recently, a deep learning-based method has shown the potential for reducing the number of diffusion-weighted images (DWIs) required to compute the rotationally invariant scalar measures to twelve. In this study, we propose a three-dimensional (3D) convolutional neural network (CNN) to estimate the scalar measures. This network further improves the performance of the deep learning-based method with a largely reduced number of required DWIs. In our approach, all the DTI and DKI measures were estimated using a single network, and a hierarchical structure was introduced to customize the outputs based on their computational complexities and to learn the commonalities of the measures. Moreover, $3 \times 3 \times 3$ convolution kernels were introduced to extract features from the 3D input patches and utilize the spatial context from adjacent neighborhoods, which also strengthened the network's robustness against noise. The proposed method was evaluated with two datasets. The results showed that, compared with the previous method that used an artificial neural network, our proposed hierarchical CNN provided enhanced efficiency for estimating all eight diffusion measures. It also improved the robustness against noise and retained the fine structures with only a few DWIs (as few as eight). This result suggests that it is possible to achieve kurtosis mapping in most clinical scanners within one minute, which could significantly extend the clinical utility of the DKI.

INDEX TERMS 3D convolutional neural networks, deep learning, diffusion kurtosis imaging, diffusion-weighted magnetic resonance imaging, hierarchical structure.

I. INTRODUCTION

Diffusion-weighted magnetic resonance imaging (MRI) is a technique sensitive to the Brownian motion of water molecules and the microenvironment in which diffusion takes place [1]. By acquiring dozens of diffusion-weighted images (DWIs) for different diffusion gradient weightings and directions, and resorting to diffusion models, various microstructural tissue properties can be inferred from the estimated metrics. Diffusion model-derived measures of microstructural properties have been helpful in clinical practice for evaluating structural integrity or damage to it by multiple diseases, including stroke [2], epilepsy [3], brain tumors [4], [5], amyotrophic lateral sclerosis [6], [7], and neurodegenerative disorders [8]. For instance, the diffusion tensor (DT) model based on a Gaussian probability

The associate editor coordinating the review of this manuscript and approving it for publication was Bora Onat.

distribution assumption is one of the most popular models [9]–[11]. However, as the motion of the water molecules in neuronal tissues is generally restricted by cell membranes and compartments of different sizes, the Gaussian approximation is often oversimplified and is unable to accurately characterize the tissue microstructures except in cases in which diffusion in an unrestricted space (such as in cerebrospinal fluid) is considered. In this regard, diffusion kurtosis imaging (DKI) is a successful extension of conventional diffusion tensor imaging (DTI) that accounts for non-Gaussian diffusion by using an additional term of a fourth-order kurtosis tensor (KT) [12], [13]. The KT allows several rotationally invariant kurtosis-related metrics to be calculated [14], [15], some of which are more sensitive to the microstructural pathological changes that occur during stroke [16], glioma [17], and traumatic brain injury [18] than are DTI measures. These DKI measures are believed to reflect the heterogeneity of the intra-voxel diffusion environment; therefore, they are indicators of microstructural complexity in many diseases [19] and significantly extend the DTI-based biomarkers.

In a typical DKI study, a mapping model is designed to map the raw image signals to voxel-wise model-derived measures. The accuracy of these measures can be strongly dependent on the quality of the acquired MR images, including diffusion directions, strengths (b-values) of the diffusion-weighting gradients, signal-to-noise ratio (SNR), etc. It has been proposed that an optimized DKI protocol would require a sample of 140 directions at three different non-zero b-values [20]. More modest protocols use 30 directions at each b-value and two non-zero b-values in total to achieve a good approximation [21]. Theoretically, the fourth-order kurtosis tensor contains 15 independent components that reflect the non-Gaussian property of water diffusion in brain tissue. When the other six independent elements of the ordinary diffusion tensor are added, 21 parameters needed to be estimated from the DWIs. However, routine applications are solely interested in the scalar measures that are further computed from these model parameters. For instance, eight scalar measures often include mean diffusivity (MD), radial diffusivity (RD), axial diffusivity (AD), fractional anisotropy (FA), mean kurtosis (MK), radial kurtosis (RK), axial kurtosis (AK), and kurtosis fractional anisotropy (KFA) [20], [22], [23]. Although the last step in computing scalar measures is simple, the entire model estimation process may take approximately an hour [13]. Therefore, lengthy acquisition and post-processing have become obstacles to the clinical use of DKI.

Several attempts have been made to improve the viability of DKI by applying advanced algorithms. For instance, compressed sensing (CS) theory [24] has been considered to take advantage of the implicit sparsity in MR images. It can recover an under-sampled dataset below the Nyquist rate as far as possible to provide the needed DWIs for the subsequent model-fitting method [25]. However, despite its apparent success, the CS reconstruction method is complicated and can even fail due to convergence problems. Selection and tuning CS hyperparameters are also challenging in practice [26]. An analytical solution for scalar measures has been discussed in recent literature. The MK and MD measures could be directly estimated with a specially designed acquisition of 13 [27] or 19 DWIs [28] and with simple post-processing. Nevertheless, to fully derive all the DKI-derived measures, fast estimation relies on an approximation of axially symmetric DKI and a model-fitting process must still be performed with 19 DWIs [29].

Recently, deep learning has undergone rapid development and attracted substantial attention [30], [31]. It has demonstrated remarkable potential for image analysis and has significantly improved the performance of a variety of medical imaging applications [32]–[37]. Instead of requiring handcrafted features, deep learning-based methods automatically detect and generate features from raw data inputs by adjusting network weights using back-propagation and stochastic gradient descent algorithms [30], [38], [39]. Studies have shown that there is redundant information in q-space and that the most relevant information of diffusion scalar measures can be recovered from only a few DWIs [27], [40], [41]. Based on this observation, Golkov *et al.* [42] proposed an artificial neural network (ANN)-based q-space deep learning (q-DL) framework for scalar measure estimation. According to their results, it is possible to reduce the number of DWIs to as few as twelve using an ANN, and to output scalar measure estimates with limited global error [43], [44]. This result generally far outperformed both the CS method and the analytical solution. It should be noted, however, that limited studies are available regarding evaluating the region-level error, which is particularly important in diffusion MRI because significant inhomogeneity exists among tissues and structures. This provided part of the motivation for the present study.

Meanwhile, human brain structures are strongly related to both brain functions and activities, and the adjacent voxels in DWIs may contain contextual information that is inadvisable to ignore [45], [46]. Recently, many studies have used convolutional neural networks (CNNs) to consider spatial correlations in the surrounding areas and to provide sufficient contextual information for classification tasks [34]–[37]. Such contextual information among the neighborhood voxels could also be beneficial for regression tasks (e.g., the DKI estimation task) and could improve the deep learning-based estimation of DKI measures. In addition, it appears from the investigations mentioned above that deep neural networks are quite powerful; however, most of the relevant studies paid little attention to the complexity of the model-derived measures to be estimated. Several papers have reported that a partially shared network can learn the common features among several related tasks through their shared layers and use these features to learn specific tasks through the remaining layers [47], [48]. In DKI analysis, the non-Gaussian kurtosis-derived measures are higher-order nonlinearity compared to those tensor-model measures. In this

respect, we believed that a single CNN-based network for all scalar measures with consideration of the network structure could further improve the deep learning-based estimation of all DKI measures.

In this study, we propose a hierarchical structured convolutional neural network (H-CNN) to efficiently estimate DKI scalar measures that simultaneously outputs the DKI-derived measures at different depths. Specifically, the eight target measures were separated into two groups and output at two different depths within the hierarchical structure. The partially shared hierarchical structure was introduced to capitalize on the relationship between the target measures while preserving the individualities between different targets. We further extended the training inputs into two-dimensional (2D) and three-dimensional (3D) voxel patches and used convolution kernels to automatically learn the cross-voxel information from the adjacent neighborhoods. The methods were tested on two independent datasets to demonstrate their performances, including an open-access dataset from the Human Connectome Project (HCP) [49].

The remainder of this paper is organized as follows: Sections II and III present the proposed H-CNN network architecture along with the descriptions of the datasets and the experimental setups. The results are presented and discussed in Sections IV and V. Finally, some general conclusions are drawn in Section VI.

II. METHODS AND MATERIALS

The task of estimating the rotationally invariant scalar measures that quantify tissue properties from a few DWIs involves finding a function that maps the DW signals to the corresponding scalar measures. In the DKI model, the DW signal S of voxel v is expressed as a function of b along a given gradient $\mathbf{g}$ in direction $\mathbf{n}$:

$$S_v(b,\mathbf{n}) = S_{v,0}\exp(-b\sum_{i,j=1}^{3} n_i n_j D_{v,ij} + \frac{1}{6}b^2\overline{D}_v^2\sum_{i,j,k,l=1}^{3} n_i n_j n_k n_l W_{v,ijkl}), \quad (1)$$

where S_0 is the signal in the absence of the diffusion encoding gradient $\mathbf{g}$, and n_i represents the element of the direction $\mathbf{n}$. This equation has a diffusion term, that spans a symmetric, positive definite 3×3 DT with six independent elements D_{ij} and mean $\overline{D}$ (i.e., the mean of the eigenvalues of DT), and a kurtosis term, which spans a symmetric $3 \times 3 \times 3 \times 3$ KT with 15 independent elements W_{ijkl} [50]. Given DT and KT, the diffusion coefficient and diffusion kurtosis in an arbitrary direction can be calculated. The eight important rotationally invariant scalar measures are defined in Table 1.

The conventional algorithm usually optimizes the estimation of the signal equation parameters and the calculation of the scalar measures independently, and information is lost at each step [51]. However, in the deep learning-based method, the scalar measures in the DKI model can be directly estimated and jointly optimized by training a network with the raw DWIs as the input and the target measures as the output. Once the network is trained, the input-output relationship is learned and can be applied quickly.

TABLE 1. Equations of all eight DKI measures.

Measures	Equations
Axial diffusivity (AD)	$\mathrm{D}_{\|\|} = \lambda_1$
Radial diffusivity (RD)	$\mathrm{D}_{\perp} = (\lambda_2 + \lambda_3)\,/\,2$
Mean diffusivity (MD)	$\overline{\mathrm{D}} = (\lambda_1 + \lambda_2 + \lambda_3)\,/\,3$
Fractional anisotropy (FA)	$\sqrt{\frac{3}{2}}\frac{\sqrt{(\lambda_1-\overline{\mathrm{D}})^2+(\lambda_2-\overline{\mathrm{D}})^2+(\lambda_3-\overline{\mathrm{D}})^2}}{\sqrt{\lambda_1^2+\lambda_2^2+\lambda_3^2}}$
Axial kurtosis (AK)	$K(\mathbf{v}_1)$
Radial kurtosis (RK)	$\frac{1}{2\pi}\int_0^{2\pi} K(\mathbf{v}_2\cos\varphi + \mathbf{v}_3\sin\varphi)d\varphi$
Mean kurtosis (MK)	$\frac{1}{4\pi}\iint_{\mathbb{S}^2} K(\mathbf{n})d\mathbb{S}^2_{\mathbf{n}}$
Kurtosis fractional anisotropy (KFA)	$\frac{\|\mathbf{W}-\overline{W}\mathbf{I}\|_F}{\|\mathbf{W}\|_F}$

Note: λ_i and $\mathbf{v}_i$ are the eigenvalues and corresponding eigenvectors of DT, with λ_1 being the largest eigenvalue. $\mathbb{S}^2$ is the unit sphere. $K(\mathbf{n})$ is diffusion kurtosis along directon $\mathbf{n}$. $\mathbf{W}$ is the kurtosis tensor, and $\overline{W}$ is equal to $\frac{1}{4\pi}\iint_{\mathbb{S}^2} W(\mathbf{n})d\mathbb{S}^2_{\mathbf{n}}$. $\mathbf{I}$ is a fully symmetric, rank 4 isotropic tensor whose element is defined as $I_{ijjkl} = \frac{1}{3}(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$, where δ_{ij} is the Kronecker delta. $\|\|_F$ denote the Frobenius norm.

To exploit the commonalities among the scalar measures and improve the generalization ability, all eight measures are estimated by a single network. However, considering that the diffusivity-related scalar measures (MD, RD, and AD) are low-order scalar measures and thus comparatively simpler than the other measures, we customized the output structures and designed a hierarchical network structure that simultaneously outputs the three diffusivity-related measures from a shallow layer and the other five measures (FA, MK, RK, AK, and KFA) from a deeper layer. Because adjacent voxels typically share similar microtissue structures, better statistical power is achieved using the intrinsic spatial correlation of neighboring voxels [37], [52]. Thus, small convolution kernels were introduced to take the neighboring voxels into consideration. Moreover, the fully connected (FC) layer in the ANN can be treated as a cross-channel parametric pooling layer that performs weighted linear recombination on the input feature maps, which is equivalent to a convolutional layer with multiple 1×1 convolution kernels [53]. The entire network, therefore, becomes a sequence of convolutional layers capable of being fed with all the 3D brain voxels and directly estimating the scalar measures within a single forward propagation [54].

Taking the above as a whole, we propose a CNN-like network with a partially shared structure that makes use of the correlations of spatial neighborhoods and among target measures, which we call an H-CNN. The structure of the H-CNN is illustrated in Fig. 1 and described in more detail below.

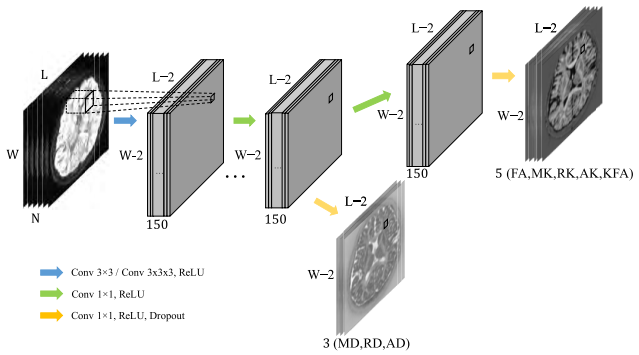


FIGURE 1. The inference process of the proposed network. The shape of the inputs is $L \times W \times H \times N$, where $L \times W \times H$ is the size of the 3D brain volume, and N is the number of down-sampled DWIs (the H dimension is not shown). The diffusivity-related scalar measures, MD, RD, and AD, are outputted at the penultimate hidden layer, and the other five measures are outputted at the last hidden layer.

A. NETWORK ARCHITECTURE

A standard neural network consists of an input layer, an output layer, and several hidden layers. Typically, each hidden layer performs nonlinear transformation:

$$\mathbf{y} = g(f(\mathbf{x}; \mathbf{W})), \tag{2}$$

where the vectors $\mathbf{x}$ and $\mathbf{y}$ are the input and output of the hidden layer, respectively, f and g are the mapping and activation functions that perform linear and nonlinear transformations, respectively, and $\mathbf{W}$ is the weight matrix. To approximate the relationship between the input and the output, the cost function, which is defined as the difference between the desired output $\mathbf{y}_i$ and the network's output $\widehat{\mathbf{y}}_i$, is minimized by solving the following optimization problem:

$$\arg\min_{\mathbf{W}} \sum_i cost(\mathbf{y}_i, \widehat{\mathbf{y}}_i), \tag{3}$$

where the sum is taken over all the training samples i. During the training stage, all the adjustable network weights are updated by a gradient-based optimizer [55]–[57] that operates on the back-propagated gradients [58].

ANNs use dense mapping [$f(\mathbf{x}; \mathbf{W}, \mathbf{b}) = \mathbf{W}\mathbf{x} + \mathbf{b}$] (which is also called the FC layer) in all the hidden layers and have been proven to be universal approximators that can uniformly approximate any continuous function, even with only a single hidden layer [59], [60]. CNNs use sparse mapping by introducing convolutional kernels, which take local patches as input and output a feature map by performing a convolution operation on the input patches [61].

An overview of the proposed H-CNN is presented in Fig. 1. The proposed network includes one input layer, several hidden layers, and two output layers. The first hidden layer is a convolutional layer, which operates as a feature extractor that learns spatial information from the input voxel patches. The shapes of the convolution kernels can be 3×3 or $3 \times 3 \times 3$, and the corresponding network is known as 2D H-CNN or 3D H-CNN, respectively. The other hidden layers and the output layers are all FC layers along the DWI dimension and take the form of convolutional layers with 1×1 kernels for scalability, and that also accelerate the inference stage. The rectified linear unit (ReLU) activation [62]: $g(\mathbf{x}) = \max(\mathbf{x}, 0)$ is adopted in each hidden layer. All the hidden layers except the last one are shared by the two output layers. The shallow output layer is connected to the penultimate hidden layer with three kernels and is responsible for the diffusivity-related measures, whereas the deeper output layer is connected to the last hidden layer with five kernels and is responsible for the other scalar measures with higher complexities. To prevent overfitting, a dropout layer [63] was inserted before each output layer. The number of kernels in each hidden layer and the dropout fraction were chosen as 150 and 0.1, respectively [42].

B. DATASETS

1) DATASET 1

This dataset consisted of data from three healthy subjects, collected using a MAGNETOM Prisma 3T scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 64-channel RF coil. Each subject was scanned three separate times over approximately one week. The local ethics committee approved this human study, and written informed consent was obtained from each participant.

The DWIs were obtained using a simultaneous multi-slice (SMS) diffusion echo-planar imaging sequence. Diffusion weightings of $b = 1000$, 2000, and 3000 s/mm^2 were applied in 30 different directions, with six $b = 0$ images equally temporally separated in the scheme, resulting in a total of 96 DWIs. Uniform coverage across multiple shells and an incremental scheme were ensured using a generalization of electrostatic repulsion [64]. All images with two opposite phase encoding directions (AP and PA) were acquired for distortion correction. Other imaging parameters were as follows: repetition time, 5400 ms; echo time, 71 ms; field-of-view, 220 mm × 220 mm; resolution, 1.5 mm × 1.5 mm × 1.5 mm; number of slices, 93; bandwidth, 1712 Hz/Px; partial Fourier, 6/8; in-plane acceleration factor, 2; and SMS factor, 3. The size of each DWI is 146 × 146 × 92.

2) DATASET 2

To evaluate the generalization capability, we validated the proposed method on a dataset with a larger sample size of 30 randomly selected subjects from the HCP. The HCP datasets were acquired at a 1.25 mm isotropic resolution with diffusion weightings of $b = 1000$, 2000, and 3000 s/mm^2 applied in 90 directions, respectively. Eighteen $b = 0$ s/mm^2 images were equally temporally separated in the scheme, resulting in a total of 288 DWIs [49]. The size of each DWI is 145 × 178 × 145. This dataset also offered a chance to statistically analyze the performance of the 3D H-CNN.

III. EXPERIMENTS

A. DATA PROCESSING

Dataset 1 was first preprocessed for motion and distortion correction [65], followed by an alignment across the three scans for each subject [66]. The preprocessing was

implemented in FSL (FMRIB Software Library, University of Oxford, UK). For both datasets, model fitting was conducted using a constrained weighted linear least square method implemented in DKE (The Center for Biomedical Imaging, Medical University of South Carolina, USA), which ensured physically and/or biologically plausible tensor estimates, thus increasing the model's robustness against noise, motion, and imaging artifacts [15]. The model-fitting results with all DWIs included were defined as the reference standards and were used as training and test labels in all the subsequent experiments. For Dataset 1, averaging the results of multiple repetitions resulted in a high SNR and improved the robustness of the reference standards.

1) PATCH EXTRACTION

The 3×3 patches for 2D H-CNN were extracted by expanding each voxel to include its eight adjacent neighbors in the same slice. The $3 \times 3 \times 3$ patches for 3D H-CNN were extracted by expanding each voxel to include its 26 adjacent neighbors in all directions.

2) DOWN-SAMPLING

Two down-sampling schemes were used to down-sample the training and test data for all the networks. The index of the DWI was defined following the DWI acquisition order in all the DWIs.

- Sequential scheme: This scheme was performed by taking the first N DWIs sequentially from the dataset because the diffusion space acquisition schemes were incremental to ensure that any first N DWIs would result in reasonably uniform coverage of the sampling domains. This scheme ensured effective down-sampling with all three shells included, which is beneficial to the model-fitting process and may also benefit the learning process.
- Selective scheme: This scheme was performed by specifying certain numbers of different b-values and randomly taking the N DWIs that satisfied the specified combination of b-values.

B. EXPERIMENTAL SETTINGS

We constructed a 2D H-CNN and a 3D H-CNN as described in subsection II-A with three hidden layers. For comparison purposes, we also constructed an ANN following the q-DL framework [42] (three hidden layers; 150 hidden units in each layer; 0.1 dropout fraction; ReLU activation) and normal CNNs without hierarchical structures.

1) EXPERIMENTS ON DATASET 1

Voxels or voxel patches of two randomly chosen subjects were used as a training set in all the experiments. The proposed H-CNNs and the q-DL ANN were trained and tested with different numbers of down-sampled DWIs from 96 to 4 (the maximum number of down-sampled DWIs was 30 for 3D H-CNN, owing to the limitation of computer random-access memory). The model-fitting method was also conducted on different down-sampled DWIs. Normal CNNs and ANNs with different numbers of hidden layers were trained and compared. The inputs were down-sampled to eight DWIs for all the networks when performing the depth evaluation. All the previous down-samplings were performed following the sequential scheme.

The effects of the b-values were evaluated by performing the estimation task on the input DWIs down-sampled by the selective scheme.

2) EXPERIMENTS ON DATASET 2

Ten randomly chosen subjects were used to form the training dataset, from which one or several subjects were selected. Then, voxel patches of the selected subjects were used to train the networks. The estimation tasks were performed using the model-fitting, 3D H-CNN, and ANN methods, and the DWIs were down-sampled following the sequential scheme. The estimated scalar measures were first registered to the standard space [66]; then, statistical tests (paired t-test and coefficient of variation (CV)) were performed on the registered results. A one-tailed paired t-test was adopted in these experiments, and the significance level was set to 0.01. To maximally visualize the differences of different methods, all the results were calculated without applying a correction for the false positive rate to show all the positive voxels.

C. NETWORK SETTINGS

1) TRAINING AND INFERENCE PROTOCOL

In the training stage, the extracted $3 \times 3 \times 3$ and 3×3 patches were used as the training inputs for the 3D CNNs and 2D CNNs, respectively. All the trainable weights of the proposed networks were initialized using the Xavier uniform initializer [67]. The gradient-based Adam optimizer [57] was used to train the network (first moment value $= 0.9$, second moment value $= 0.999$, and initial learning rate $= 0.001$). The mini-batch size was set to 256. A simple learning rate scheduler, which decays the learning rate at a decay factor of 0.5 when the reduced value of the training set error in 10 epochs does not exceed a threshold of 0.0001, was used to improve the training performance. Ten percent of the training set was used as a validation set to prevent overfitting. An early termination strategy was also introduced to prevent overfitting when the validation set error stopped decreasing or started increasing within 30 epochs. Both the training and test data were normalized to have zero mean and unit standard deviation to achieve good learning performance. The training parameter settings were based on Bengio's suggestion [68].

In the inference stage, all voxels of the 3D brain volume were fed into the trained network to directly estimate the scalar values.

2) EVALUATION METRIC

The primary evaluation metric was the root-mean-squared error (RMSE) between the estimated scalar measures and

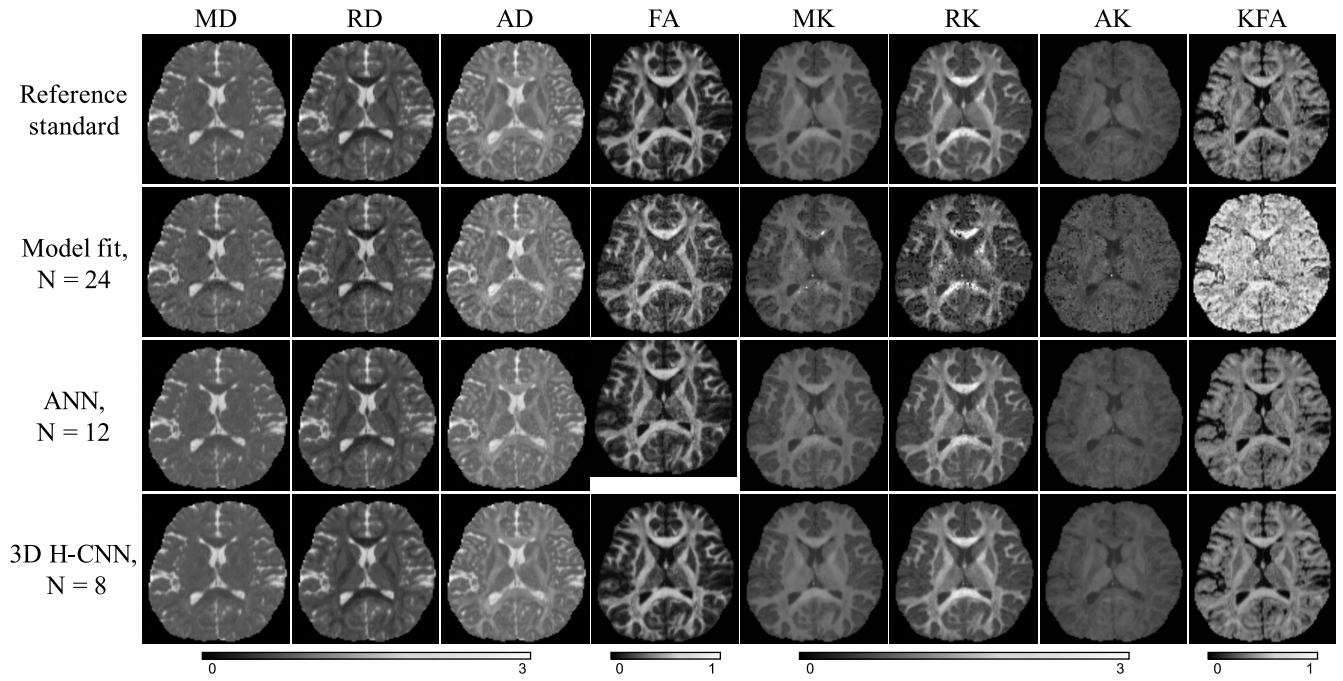


FIGURE 2. Maps of the eight scalar measures with different methods. Slice 49 was visualized for all the measures. From top to bottom: reference standards, results of model fitting with 24 DWIs, results of the ANN with twelve DWIs, and results of the 3D H-CNN with eight DWIs. From left to right: MD, RD, AD, FA, MK, RK, AK, and KFA.

their reference standards over all voxels within the entire brain mask [69]:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{V} (\hat{v}_i - v_{gt_i})^2}{V}}, \tag{4}$$

where V is the total number of voxels, $\hat{v}_i$ is the predicted scalar value, v_{gt_i} is the reference standard, and $\hat{v}_i - v_{gt_i}$ is defined as the error of the estimation result. The RMSEs over all the scalar measures are called the overall RMSEs.

3) IMPLEMENTATION

All the experiments were performed under Linux OS on a desktop computer with 20 × Intel® Core™ i7-6950X @ 3.0 GHz CPU, 2 × NVIDIA® GeForce® GTX 1080 Ti graphics card, and 32 GB DDR3 RAM. All the neural networks were implemented in Python using the Keras [70] framework with a Tensorflow [71] back-end.

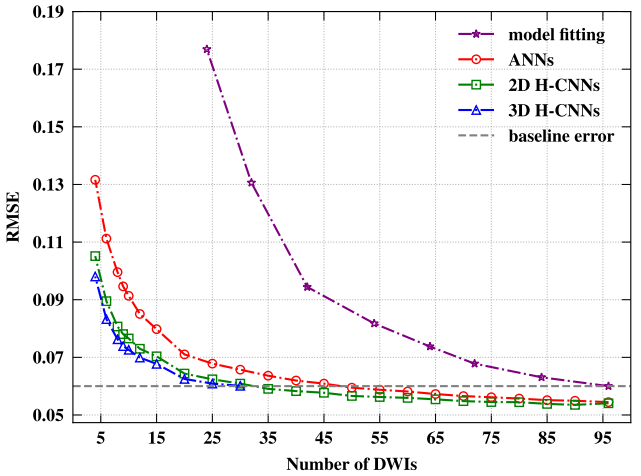


FIGURE 3. Overall RMSEs of different methods with different numbers of DWIs down-sampled following the sequential scheme. All the networks contain three hidden layers. The baseline error was defined as the RMSE between the results of the model fitting with 96 DWIs and their reference standards.

IV. RESULTS

A. RESULTS ON DATASET 1

Fig. 2 displays the typical results for the maps of the reference standards and the estimation results of the model-fitting method, the ANN, and the proposed H-CNNs. Notably, the 3D H-CNN demonstrated the best results; the 3D H-CNN with only eight DWIs showed similar fine structure visualizations compared to the ANN results with twelve DWIs. The overall RMSEs of the scalar measures estimated by these methods were also compared (Fig. 3), and our proposed 3D H-CNN achieved the lowest RMSE at an equivalent number of DWIs. The difference between the overall RMSEs of the 2D H-CNN and the ANN was rather small (0.0003) when fully sampled DWIs were used. As the number of DWIs decreased, this difference continued to increase; it finally reached 0.027 when four DWIs were used. With more neighboring voxels introduced, the 3D H-CNN slightly outperformed the 2D H-CNN; its overall RMSE was 0.0765 when only eight DWIs were used, which was smaller than the RMSE of the ANN with twelve DWIs (0.0852). Both the graphical and numerical results demonstrated that the

TABLE 2. RMSEs of each scalar measure estimated by 3D CNNs trained with eight DWIs down-sampled by the sequential scheme. Best results are highlighted in bold.

Method	RMSEs								
	MD	RD	AD	FA	MK	RK	AK	KFA	overall
3D CNN with 2 hidden layers	**0.07075**	**0.07618**	0.09472	0.05182	0.06455	0.11442	0.06608	0.06446	0.07765
3D CNN with 3 hidden layers	0.07205	0.07736	0.09473	0.05109	0.06383	0.11376	0.06520	0.06412	0.07757
3D CNN with hierarchical structure	0.07169	0.07713	**0.09382**	**0.05013**	**0.06230**	**0.11162**	**0.06385**	**0.06363**	**0.07654**

TABLE 3. RMSEs of each scalar measure estimated by the 3D H-CNN trained with eight DWIs down-sampled by the selective scheme. Best results are highlighted in bold.

Combination of b-values	RMSEs								
	MD	RD	AD	FA	MK	RK	AK	KFA	overall
1	0.07615	0.08020	0.09654	**0.04771**	0.07286	0.11964	0.07110	**0.06240**	0.08093
2	0.07383	0.07889	0.09426	0.04934	0.06987	0.11675	0.06835	0.06610	0.07949
3	0.08927	0.09235	0.11274	0.05189	0.06167	0.11096	0.06402	0.06964	0.08438
4	**0.07087**	**0.07680**	**0.09316**	0.05040	0.06742	0.11785	0.06738	0.06313	0.07834
5	0.07718	0.08194	0.09909	0.05091	**0.06107**	**0.11057**	**0.06305**	0.06634	0.07859
6	0.07169	0.07713	0.09382	0.05013	0.06230	0.11162	0.06385	0.06363	**0.07654**

The chosen combinations of b-value in the eight DWIs: 1) seven $b = 1000$ s/mm^2 with single $b = 0$ s/mm^2; 2) seven $b = 2000$ s/mm^2 with single $b = 0$ s/mm^2; 3) seven $b = 3000$ s/mm^2 with single $b = 0$ s/mm^2; 4) four $b = 1000$ s/mm^2, and three $b = 2000$ s/mm^2, with single $b = 0$ s/mm^2; 5) four $b = 2000$ s/mm^2, and three $b = 3000$ s/mm^2, with single $b = 0$ s/mm^2; 6) two $b = 1000$ s/mm^2, three $b = 2000$ s/mm^2, and two $b = 3000$ s/mm^2, with single $b = 0$ s/mm^2;

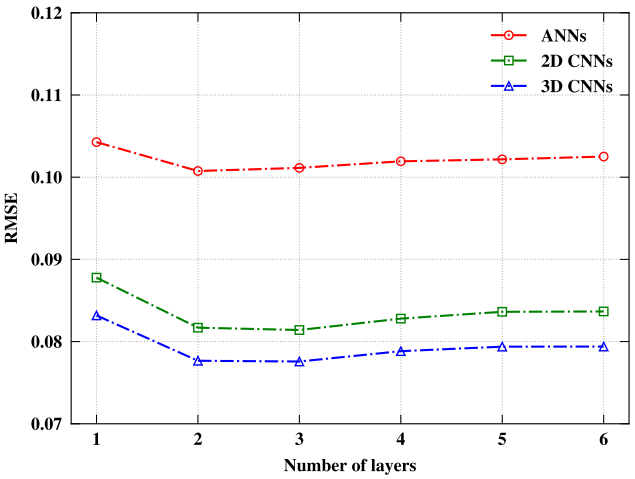


FIGURE 4. Overall RMSEs of networks with different numbers of hidden layers. The number of DWIs were down-sampled to eight by the sequential scheme for all the networks.

3D H-CNN maintained comparable performance even under largely reduced DWIs.

As shown in Fig. 4, with the 3D convolution kernels, the best overall RMSE of the 3D CNN was over 20% lower than that of the ANN when eight DWIs were used. Thus, the introduced convolution kernels improved the robustness against noise by considering the spatial information of the adjacent voxels, which reduced the overall RMSE. Meanwhile, all the networks achieved their best performances when the number of hidden layers was two or three; their performances deteriorated as the number of hidden layers continued to increase. This result indicates that networks with more than three hidden layers have higher capacities than required for the estimation task; such networks may become highly sensitive to subtle differences and unable to capture the general similarity between voxels, thus leading to overfitting [61]. We conducted a detailed comparison between the convolutional networks with two and three hidden layers. Table 2 shows the RMSEs of each single scalar measure estimated by the 3D CNNs. The diffusivity-related measures estimated by the network with two hidden layers had smaller RMSEs than those estimated by the network with three hidden layers; the opposite results occurred for the other measures. With the partly shared hierarchical structure, the proposed network outperformed the CNNs with two or three hidden layers on the estimation of almost all the scalar measures.

Table 3 shows the RMSEs of the DKI measures estimated by the 3D H-CNN trained with DWIs down-sampled following the selective scheme. The overall RMSEs were slightly higher when trained with DWIs from a single b-value. For individual measures, a single low b-value of 1000 s/mm^2 can provide a good estimation of DT-based measures. However, for some KT-based measures, a higher b-value is needed. In general, the lowest overall RMSE and more balanced performance were achieved when all the b-values were introduced in combination 6. The results, therefore, offer a trade-off between performance and the choice of b-values.

B. RESULTS ON DATASET 2

Fig. 5 shows the RMSEs of all the subjects as a heat map. When performing estimation tasks on the test subjects, the 3D H-CNNs trained with data from a single subject had similar (or even lower) RMSEs compared to the model-fitting

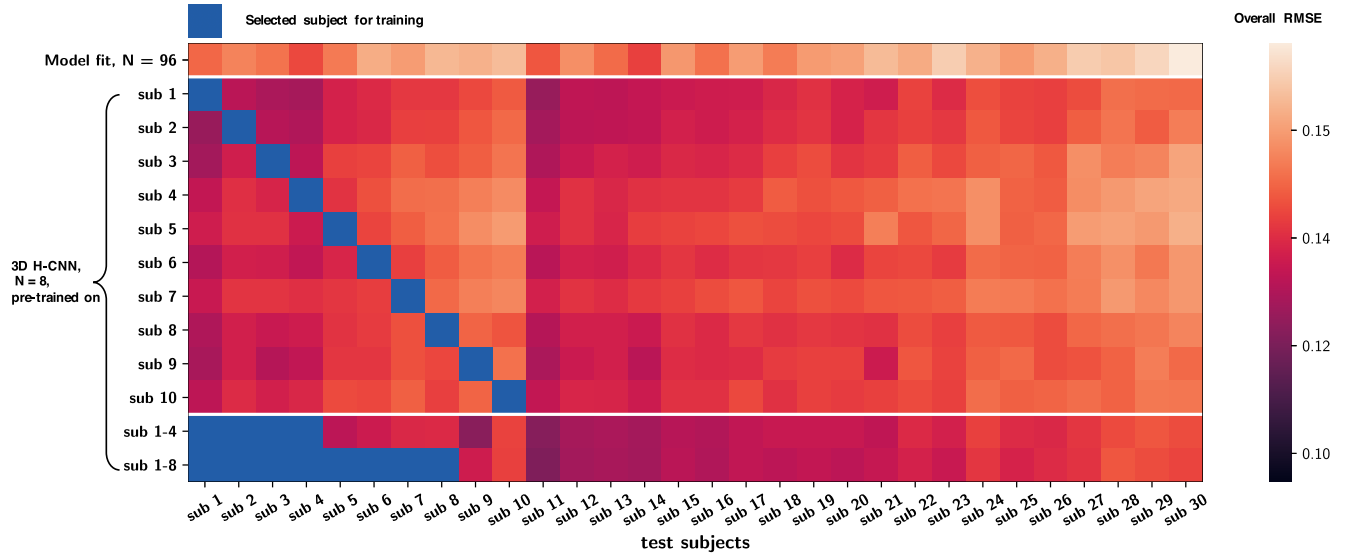


FIGURE 5. Heat map of the overall RMSEs of DKI measures evaluated with different methods on all 30 subjects. From top to bottom: Model fitting with 96 DWIs, ten versions of 3D H-CNNs pre-trained on different individuals, 3D H-CNN pre-trained on four subjects, and 3D H-CNN pre-trained on eight subjects. The subjects selected for training are masked in blue. The number of DWIs for each training subject was down-sampled to eight.

results with 96 DWIs. Moreover, selecting different training subjects had little effect on the RMSEs of the test results, which implies that the 3D H-CNNs could learn the underlying mapping of the estimation tasks even with a single training subject. When trained with four subjects, the RMSEs of the test subjects decreased distinctly, which demonstrated that using data from more subjects can improve the training results. However, when the number of subjects was further increased to eight, no substantial improvement occurred in the RMSE means compared to the results with four subjects.

Fig. 6 (a) visualizes the paired *t*-test results of the DKI measures. Similar to the visualization results in Fig. 2, the proposed 3D H-CNN demonstrated the best results. For the model-fitting method, a large percentage of the voxels were significantly higher (33% of MD, 15% of RD, 46% of AD, 14% of FA, and 64% of KFA) or lower (31% of MK, 23% of RK, and 24% of AK) than the reference standards when 32 DWIs were used. When the number of DWIs increased to 96, 22% and 47% voxels in AK and KFA, respectively, were still significantly different from the reference standards. However, in the results of the 3D H-CNN trained with twelve DWIs, only approximately 7% of the voxels were significantly different from the reference standards for all measures; moreover, the proportion remained below 10% when the number of DWIs was reduced to eight. In the visualizations, slight group-level differences were observed in the sub-cortical regions and the complex crossing-fiber structures (e.g., caudate body, putamen, and sub-gyral white matter in the temporal and frontal lobe from the Talairach atlas [72]) in the proposed method. These differences could be due to the complex structures, relatively smaller number of voxel samples available for training, and the large variability among subjects. We further evaluated those regions that showed many significant differences with the deep learning-based methods by performing a region-wise analysis. The mean absolute error (MAE) of all 30 subjects in those regions was calculated and visualized. As shown in Fig. 7, although significant deviations exist, our proposed 3D H-CNN with eight DWIs had fewer MAEs in most regions than did the ANN method with twelve DWIs. The paired *t*-test results on the absolute difference from the reference standard are visualized in Fig. 6 (b). In the results of the 3D H-CNN trained with twelve DWIs, only 2% of the voxels had significantly higher absolute error than did the model fitting with 96 DWIs, and 15% of the voxels were significantly lower in the KFA measure. When the DWI number was reduced to eight, the high-order measures still maintained an acceptable performance (below 5%)—especially KFA, where 12% of the voxels had significantly lower absolute differences from the reference standard.

To validate the ability of H-CNN to capture inter-subject variability, the CV (defined as the ratio of the standard deviation to the mean) maps of the DKI measures on the 30 subjects in Dataset 2 were calculated. The CV maps of the RK measure are displayed in Fig. 8. As shown, the CV maps of the 3D H-CNN with eight DWIs were similar to the CV map of the model fitting with 288 DWIs (the reference standard). When compared in detail (e–g), the performance of the 3D H-CNN with eight DWIs was comparable to that of the model fitting with 96 DWIs. In addition, the number of subjects in the training set had little effect on the CV values. This result indicated that the inter-subject variability would not be significantly reduced in the proposed 3D H-CNN method.

In the experiments, the proposed method took approximately 30–60 s per epoch (depending on the number of DWIs) at the training stage and less than 10 s (with coverage of the entire brain volume) at the test stage.

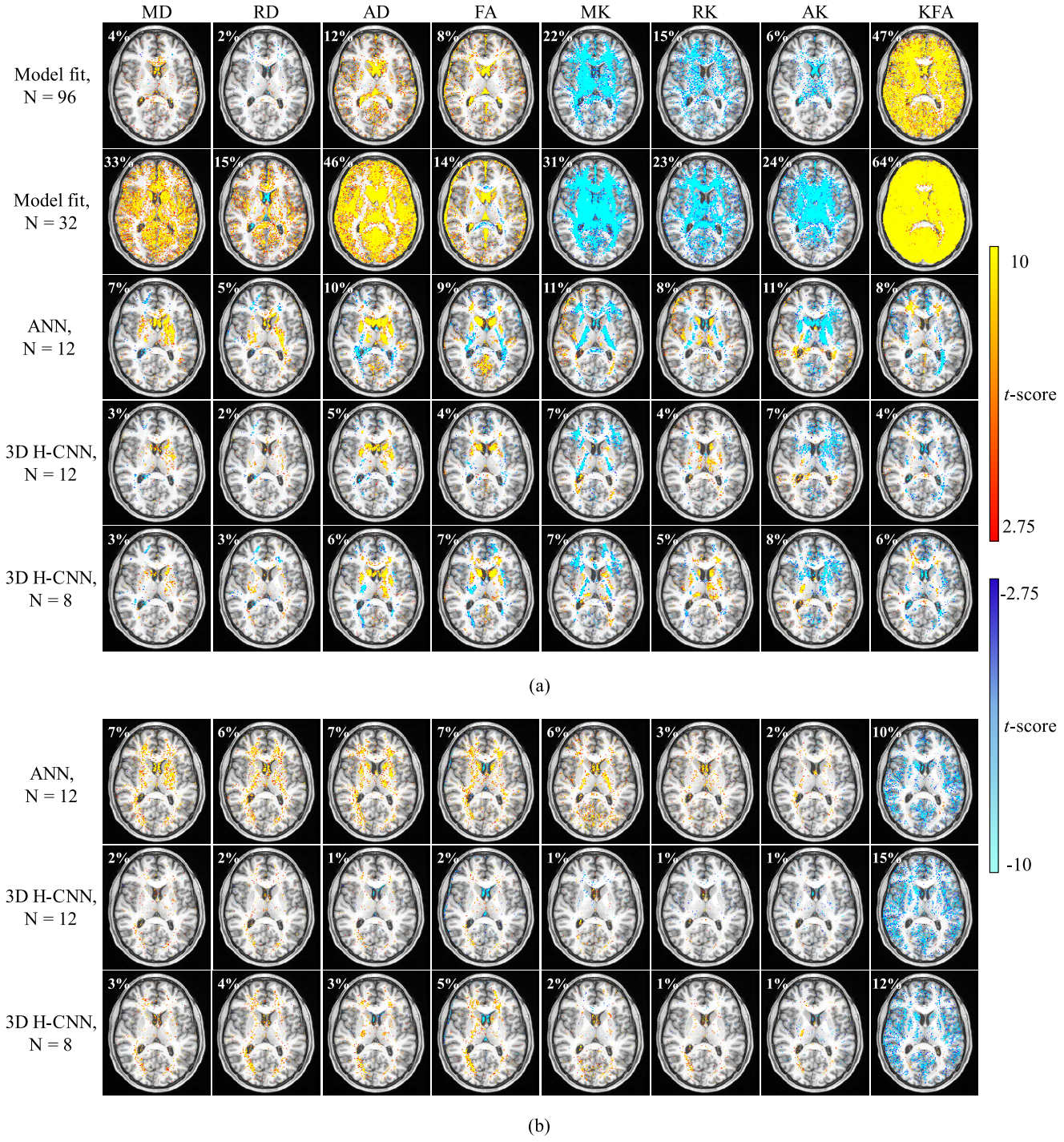


FIGURE 6. Pairwise *t*-test statistics at group level. Voxels with |*t*-score| > 2.75 ($P < 0.01$) were colored, and slice 65 was selected to be visualized for all the measures. The percentages of colored voxels were calculated and displayed. (a) Comparison of DKI scalar measures by the evaluated methods (annotated on the left side) with those by the reference method (model fit, N = 288); Regions with over-estimated measures are in hot color while underestimated regions are shown in blue; (b) Comparison of the absolute deviation from the reference standard of the evaluated methods (annotated on the left side) with that of the model fitting (N = 96); Regions with higher deviation are shown in hot color while those with lower deviation are shown in blue. From left to right: MD, RD, AD, FA, MK, RK, AK, and KFA.

V. DISCUSSION

As shown in Fig. 3, the model-fitting method was easily affected by noise when a reduced number of DWIs was used. It became underdetermined when fewer than 24 DWIs (including images with 0 *b*-value) were used, whereas the deep learning-based methods still achieved good estimation

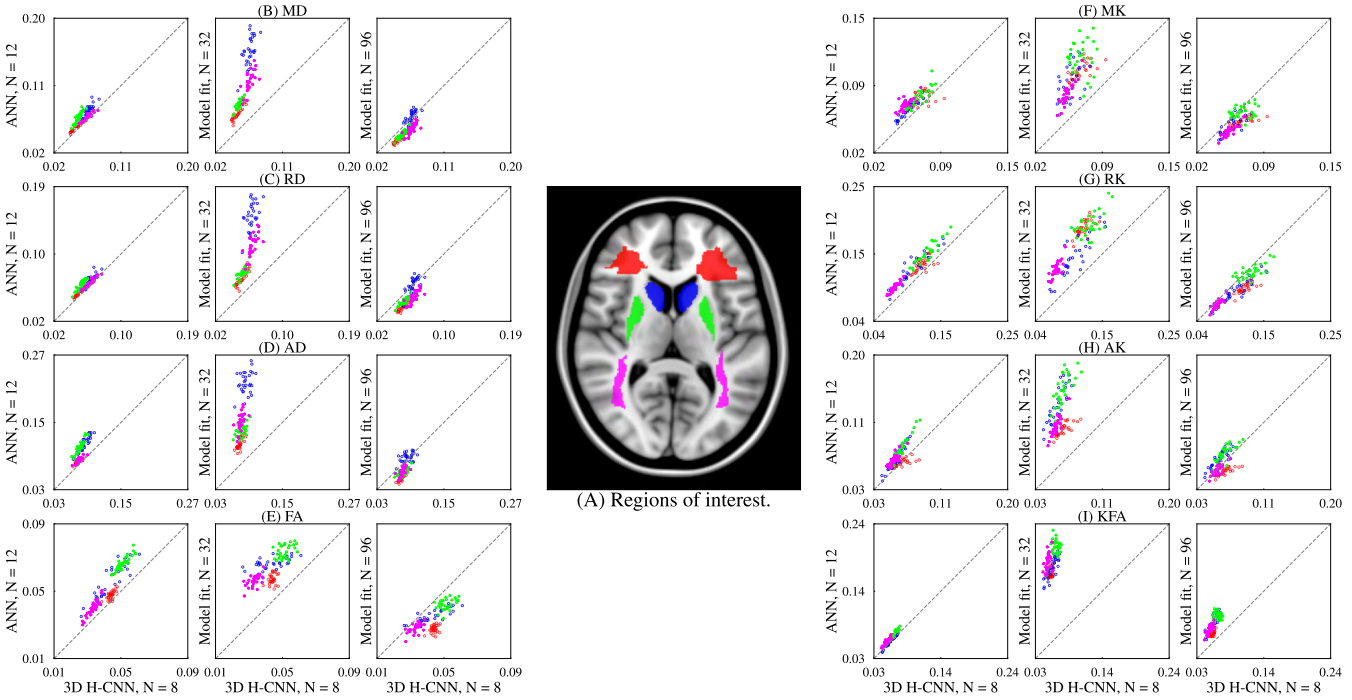


FIGURE 7. Region-wise mean absolute errors. (A) The regions of interest projected on standard space, including sub-gyral white matter in the frontal lobe (red), caudate body (blue), putamen (green), and sub-gyral white matter in the temporal lobe (magenta). (B–I) Scatter plots of the mean absolute errors in each region from each subject for the eight measures of MD, RD, AD, FA, MK, RK, AK, and KFA. Each dot in the plots represents a single subject. In each plot, the left column shows the comparison between the 3D H-CNN with 8 DWIs vs. the ANN with 12 DWIs, the middle column compares the 3D H-CNN with 8 DWIs vs. the model fitting with 32 DWIs, and the right column compares the 3D H-CNN with 8 DWIs vs. the model fitting with 96 DWIs.

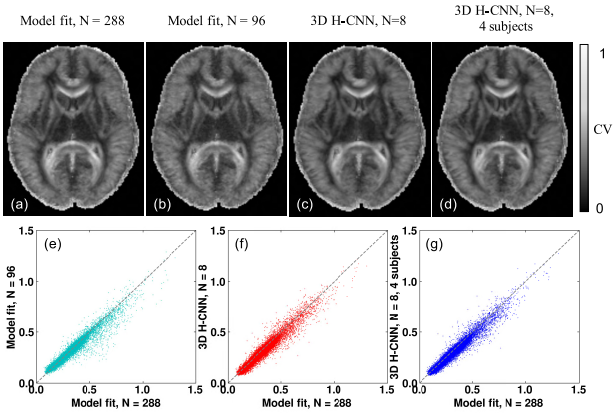


FIGURE 8. Maps of the CV results of RK. Slice 65 was selected for visualization. (a–d) Model fitting results with 288 and 96 DWIs, 3D H-CNN with eight DWIs trained with only one subject and four subjects; (e–g) distributions of CV map (b–d) against the reference standard (a).

performances (the fourth row). As described in Table 1, the eight target measures are highly correlated (all the measures are calculated based on the 21 independent elements of DT and KT, and the AD, MD, and RD measures are linearly dependent on each other), which indicates that redundant information exists among these measures. With deep learning-based methods, these relationships could be used to help learn the underlying mapping functions. By defining the baseline error as the overall RMSE between the model-fitting results with 96 DWIs and the reference standards, we found that the smallest number of DWIs required to maintain the same RMSE as the baseline error was 25 for the 3D H-CNN in Dataset 1, whereas the 3D H-CNN with eight DWIs had smaller RMSEs than the baseline error in Dataset 2 (Fig. 5). A reasonable explanation for this result is that the reference standards were of different qualities. The reference standard for Dataset 2 was estimated with 288 DWIs; thus, it had better quality than the reference standard for Dataset 1, which was the average of three repetitions with 96 DWIs. Finally, compared to the model-fitting method, better results were achieved for the network trained on Dataset 2. Thus, it is reasonable to conclude that better training dataset leads to better training results.

It is notable that there are small ''holes'' in the maps of the reference standards in Fig. 2, particularly in the maps of the MK, RK, and AK measures, which were ''recovered'' in the deep learning-based method results. Because information regarding only a single voxel was used in the model-fitting method, signal noise may have had a strong influence on the estimation results and caused the zero values in the measures. For the deep learning-based methods, the weights learned during the training stage may provide extra information and strengthen their robustness against noise; thus, signal noise had a smaller effect on the estimation results. Moreover, the proposed H-CNNs further improved the robustness by considering the neighbors, which further reduced the ''holes'' in the visualized measure maps.

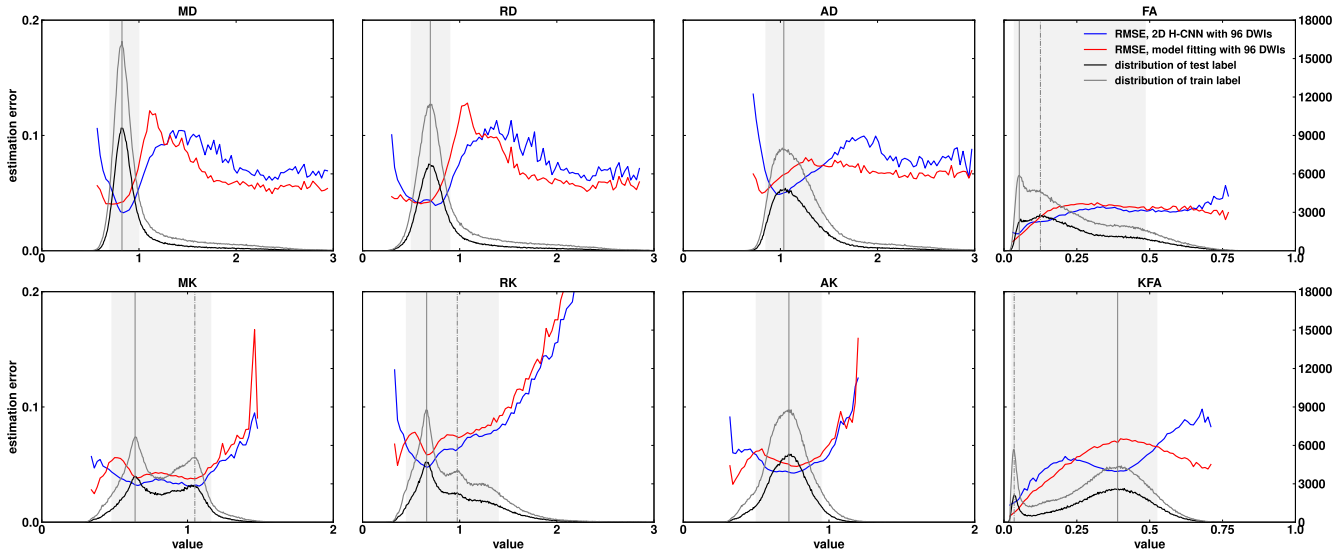


FIGURE 9. Distributions of the reference standard scalar measures from training sets and test sets. The RMSE in each bin of the distribution (except for those bins that had less than 100 voxels) is also plotted.

In the current study, the networks took only raw DWIs as input; they were unaware of the *b*-value or *b*-direction of each DWI. However, the same indexes of the DWIs were used during the training and test stages in the experiments described above, which provided the same *b*-values and similar *b*-directions. Therefore, training a network that understands the input scheme by using *b*-values and *b*-directions as auxiliary information may further improve the generalizability of the method, which can be a subject of future research.

The deep learning-based methods require a large amount of labeled data to feasibly train the network. In our experiments, approximately 0.3 million samples were gathered from each subject (Dataset 1), which seemed to be sufficient, and the results achieved a good performance. However, the distributions of training set were unbalanced. Fig. 9 provides a visualization of the distributions of Dataset 1 and shows that the measures' values of most voxels lay within a small range, particularly for MR, RD, AD, and AK. In the RMSE curves of the 2D H-CNN, almost all the local minima occurred in the bins containing the highest number of voxels in the training set; these were also lower than the corresponding RMSEs of the model-fitting results. For those bins that contained few voxels in the training set, the RMSEs of the test voxels were higher, and the measures' values were biased to the values of most voxels, which smoothed over the measure maps. A similar phenomenon occurs in the classification task called the "class imbalance problem," in which the classifiers tend to be overwhelmed by the majority class and ignore the minority class [73]. Because of the similarity among human brains, all the subjects' data shared a similar distribution; thus, the biased results can still achieve lower RMSEs on the test subjects' data. Introducing data from multiple subjects eliminated the distribution differences between different subjects, and a relatively diverse training set could be provided. However, because of the similarity, the dataset could not be further improved (balanced) by adding more training subjects (Fig. 5). Resampling the data is the most direct solution for this "class imbalance problem."

In the current study, we employed data from healthy subjects to demonstrate our proposed method; however, because pathological alterations could change the tissue properties, the values of some scalar measures may appear in the bins with only a small number of voxels in the distribution. Thus, the trained network may not be directly applicable to diseased cases. However, our study explored the ability of neural networks to learn the features of structure-associated tissue properties in DWIs. Therefore, the network could also be trained to reflect abnormal changes of tissue structure—providing sufficient voxel samples can be gathered to learn the possible features of the diseased tissues. This is similar to the problem of finding a more balanced dataset as discussed above, and it is an important aspect of our future work.

Beyond the eight scalar measures discussed here, there have been other DKI-related measures in the literature but not included in the present study. For example, the mean of the kurtosis tensor (MKT) is an efficient analytical solution introduced by fast kurtosis technique. MKT provides a good approximation to the MK measure and shares similar contrast information with MK [27]. As a fast kurtosis technique, its estimation needs only acquisition of 13 DWIs, and the computation could be done in seconds without any burden in data-training as neural network method does. It would be more complete to include all these measurements in the neural network in further studies. Additionally, the current network was designed for the DKI model. In theory, many other informative diffusion models, such as CHARMED [74] and AxCaliber [75], could also be treated as mapping tasks from the input DWIs to some outputs that reflect different tissue properties. However, considering that those models may have different complexity or model assumptions,

different network structures should be designed and tuned for those models. Nevertheless, it is definitely worth trying to apply neural networks to other clinically important diffusion models using this method in future work.

VI. CONCLUSIONS

In this study, we proposed a hierarchical CNN to estimate DKI-derived scalar measures. The proposed method introduced small convolution kernels to learn the spatial information among the neighboring voxels and improve the robustness against noise. The experimental results demonstrated that by introducing the convolution kernels and the partially shared structure, the proposed network outperformed previous methods from the literature and achieved efficient estimation. This highly accelerated DKI estimation method may provide a clinically feasible acquisition scheme that could promote clinical applications of DKI. The hierarchical structure could also be applied to other complex models with multiple related targets, thus further extending its potential applications.

ACKNOWLEDGMENT

(Zhiwei Li and Ting Gong contributed equally to this work.)

REFERENCES

[1] E. O. Stejskal and J. E. Tanner, "Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient," *J. Chem. Phys.*, vol. 42, no. 1, pp. 288–292, 1965.

[2] G. Thomalla, V. Glauche, M. A. Koch, C. Beaulieu, C. Weiller, and J. Röther, "Diffusion tensor imaging detects early Wallerian degeneration of the pyramidal tract after ischemic stroke," *NeuroImage*, vol. 22, no. 4, pp. 1767–1774, Aug. 2004.

[3] D. W. Gross, "Diffusion tensor imaging in temporal lobe epilepsy," *Epilepsia*, vol. 52, pp. 32–34, Jul. 2011.

[4] T. Inoue, K. Ogasawara, T. Beppu, A. Ogawa, and H. Kabasawa, "Diffusion tensor imaging for preoperative evaluation of tumor grade in gliomas," *Clin. Neurol. Neurosurg.*, vol. 107, no. 3, pp. 174–180, Apr. 2005.

[5] S. Lu, D. Ahn, G. Johnson, M. Law, D. Zagzag, and R. I. Grossman, "Diffusion-tensor MR imaging of intracranial neoplasia and associated peritumoral edema: Introduction of the tumor infiltration index," *Radiology*, vol. 232, no. 1, pp. 221–228, Jul. 2004.

[6] M. Sach, G. Winkler, V. Glauche, J. Liepert, B. Heimbach, M. A. Koch, C. Büchel, and C. Weiller, "Diffusion tensor MRI of early upper motor neuron involvement in amyotrophic lateral sclerosis," *Brain*, vol. 127, no. 2, pp. 340–350, Feb. 2004.

[7] M. M. van der Graaff, C. A. Sage, M. W. A. Caan, E. M. Akkerman, C. Lavini, C. B. Majoie, A. J. Nederveen, A. H. Zwinderman, F. Vos, and F. Brugman, "Upper and extra-motoneuron involvement in early motoneuron disease: A diffusion tensor imaging study," *Brain*, vol. 134, no. 4, pp. 1211–1228, Apr. 2011.

[8] J. Goveas, L. O'Dwyer, M. Mascalchi, M. Cosottini, S. Diciotti L. Passamonti, S. De Santis, C. Tessa, N. Toschi, and M. Giannelli, "Diffusion-MRI in neurodegenerative disorders," *Magn. Reson. Imag.*, vol. 33, no. 7, pp. 853–876, Sep. 2015.

[9] P. J. Basser, J. Mattiello, and D. Lebihan, "Estimation of the effective self-diffusion tensor from the NMR spin echo," *J. Magn. Res. B*, vol. 103, no. 3, pp. 247–254, Mar. 1994.

[10] C. Pierpaoli and P. J. Basser, "Toward a quantitative assessment of diffusion anisotropy," *Magn. Reson. Med.*, vol. 36, no. 6, pp. 893–906, 1996.

[11] P. J. Basser, "Relationships between diffusion tensor and *q*-space MRI," *Magn. Reson. Med.*, vol. 47, no. 2, pp. 392–397, Feb. 2002.

[12] J. H. Jensen, J. A. Helpern, A. Ramani, H. Lu, and K. Kaczynski, "Diffusional kurtosis imaging: The quantification of non-Gaussian water diffusion by means of magnetic resonance imaging," *Mag. Reson. Med.*, vol. 53, no. 6, pp. 1432–1440, 2005.

[13] H. Lu, J. H. Jensen, A. Ramani, and J. A. Helpern, "Three-dimensional characterization of non-Gaussian water diffusion in humans using diffusion kurtosis imaging," *NMR Biomed.*, vol. 19, no. 2, pp. 236–247, Apr. 2006.

[14] J. H. Jensen and J. A. Helpern, "MRI quantification of non-Gaussian water diffusion by kurtosis analysis," *NMR Biomed.*, vol. 23, no. 7, pp. 698–710, 2010.

[15] A. Tabesh, J. H. Jensen, B. A. Ardekani, and J. A. Helpern, "Estimation of tensors and tensor-derived measures in diffusional kurtosis imaging," *Magn. Reson. Med.*, vol. 65, no. 3, pp. 823–836, Mar. 2011.

[16] R. A. Weber, E. S. Hui, J. H. Jensen, X. Nie, M. F. Falangola, J. A. Helpern, and D. L. Adkins, "Diffusional kurtosis and diffusion tensor imaging reveal different time-sensitive stroke-induced microstructural changes," *Stroke*, vol. 46, no. 2, pp. 545–550, Feb. 2015.

[17] S. Van Cauter, J. Veraart, J. Sijbers, R. R. Peeters, U. Himmelreich, F. De Keyzer, S. W. Van Gool, F. Van Calenbergh, S. De Vleeschouwer, W. Van Hecke, and S. Sunaert, "Gliomas: Diffusion kurtosis MR imaging in grading," *Radiology*, vol. 263, no. 2, pp. 492–501, May 2012.

[18] J. Zhuo, S. Xu, J. L. Proctor, R. J. Mullins, J. Z. Simon, G. Fiskum, and R. P. Gullapalli, "Diffusion kurtosis as an *in vivo* imaging marker for reactive astrogliosis in traumatic brain injury," *NeuroImage*, vol. 59, no. 1, pp. 467–477, Jan. 2012.

[19] A. J. Steven, J. Zhuo, and E. R. Melhem, "Diffusion kurtosis imaging: An emerging technique for evaluating the microstructural environment of the brain," *Amer. J. Roentgenol.*, vol. 202, no. 1, pp. W26–W33, Jan. 2014.

[20] D. H. Poot, A. J. den Dekker, E. Achten, M. Verhoye, and J. Sijbers, "Optimal experimental design for diffusion kurtosis imaging," *IEEE Trans. Med. Imag.*, vol. 29, no. 3, pp. 819–829, Mar. 2010.

[21] E. Fieremans, J. H. Jensen, and J. A. Helpern, "White matter characterization with diffusional kurtosis imaging," *NeuroImage*, vol. 58, no. 1, pp. 177–188, Sep. 2011.

[22] G. R. Glenn, J. A. Helpern, A. Tabesh, and J. H. Jensen, "Quantitative assessment of diffusional kurtosis anisotropy," *NMR Biomed.*, vol. 28, no. 4, pp. 448–459, Feb. 2015.

[23] B. Hansen and S. N. Jespersen, "Kurtosis fractional anisotropy, its contrast and estimation by proxy," *Sci. Rep.*, vol. 6, Apr. 2016, Art. no. 23999.

[24] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[25] B. Bilgic, K. Setsompop, J. Cohen-Adad, A. Yendiki, L. L. Wald, and E. Adalsteinsson, "Accelerated diffusion spectrum imaging with compressed sensing using adaptive dictionaries," *Magn. Reson. Med.*, vol. 68, no. 6, pp. 1747–1754, Dec. 2012.

[26] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, 2018.

[27] B. Hansen, T. E. Lund, R. Sangill, and S. N. Jespersen, "Experimentally and computationally fast method for estimation of a mean kurtosis," *Magn. Reson. Med.*, vol. 69, no. 6, pp. 1754–1760, Jun. 2013.

[28] B. Hansen, T. E. Lund, R. Sangill, E. Stubbe, J. Finsterbusch, and S. N. Jespersen, "Experimental considerations for fast kurtosis imaging," *Magn. Reson. Med.*, vol. 76, no. 5, pp. 1455–1468, Nov. 2016.

[29] B. Hansen, N. Shemesh, and S. N. Jespersen, "Fast imaging of mean, axial and radial diffusion kurtosis," *NeuroImage*, vol. 142, pp. 381–393, Nov. 2016.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[31] G. Wang, "A perspective on deep imaging," *IEEE Access*, vol. 4, pp. 8914–8924, 2016.

[32] S. Wang, Y. Jiang, X. Hou, H. Cheng, and S. Du, "Cerebral micro-bleed detection based on the convolution neural network with rank based average pooling," *IEEE Access*, vol. 5, pp. 16576–16583, 2017.

[33] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[34] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[35] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.

[36] Z. Liu, C. Cao, S. Ding, Z. Liu, T. Han, and S. Liu, "Towards clinical diagnosis: Automated stroke lesion segmentation on multi-spectral MR image using convolutional neural network," *IEEE Access*, vol. 6, pp. 57006–57016, 2018.

[37] L. Yuan, X. Wei, H. Shen, L.-L. Zeng, and D. Hu, "Multi-center brain imaging classification using a novel 3D CNN approach," *IEEE Access*, vol. 6, pp. 49925–49934, 2018.
[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.
[39] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[40] G. Nedjati-Gilani, M. G. Hall, C. A. M. Wheeler-Kingshott, and D. C. Alexander, "Learning microstructure parameters from diffusion-weighted MRI using random forests," in *Proc. Joint Annu. Meet. ISMRM-ESMRMB*, 2014, p. 2626.
[41] C. Liao, Y. Chen, X. Cao, S. Chen, H. He, M. Mani, M. Jacob, V. Magnotta, and J. Zhong, "Efficient parallel reconstruction for high resolution multi-shot spiral diffusion data with low rank constraint," *Magn. Reson. Med.*, vol. 77, no. 3, pp. 1359–1366, Mar. 2017.
[42] V. Golkov, A. Dosovitskiy, J. I. Sperl, M. I. Menzel, M. Czisch, P. Sämann, T. Brox, and D. Cremers, "*q*-space deep learning: Twelve-fold shorter and model-free diffusion MRI scans," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1344–1351, May 2016.
[43] J. Veraart, J. Sijbers, S. Sunaert, A. Leemans, and B. Jeurissen, "Weighted linear least squares estimation of diffusion MRI parameters: Strengths, limitations, and pitfalls," *NeuroImage*, vol. 81, pp. 335–346, Nov. 2013.
[44] H. Zhang, T. Schneider, C. A. Wheeler-Kingshott, and D. C. Alexander, "NODDI: Practical *in vivo* neurite orientation dispersion and density imaging of the human brain," *NeuroImage*, vol. 61, no. 4, pp. 1000–1016, Jul. 2012.
[45] S. Shen, W. Sandham, M. Granat, and A. Sterr, "MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 3, pp. 459–467, Sep. 2005.
[46] Z. Lu, Q. Zheng, W. Yang, Q. Feng, and W. Chen, "Adaptive image segmentation based on local neighborhood information and Gaussian weighted Chi-square distance," in *Proc. ISBI*, May 2012, pp. 1240–1243.
[47] R. Caruana, "Learning many related tasks at the same time with backpropagation," in *Proc. NIPS*, 1994, pp. 657–664.
[48] J. Ghosn and Y. Bengio, "Multi-task learning for stock selection," in *Proc. NIPS*, Dec. 1996, pp. 946–952.
[49] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, K. Ugurbil, and WU-Minn HCP Consortium, "The WU-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, Oct. 2013.
[50] M. I. Menzel, E. T. Tan, K. Khare, J. I. Sperl, K. F. King, X. Tao, C. J. Hardy, and L. Marinelli, "Accelerated diffusion spectrum imaging in the human brain using compressed sensing," *Magn. Reson. Med.*, vol. 66, no. 5, pp. 1226–1233, Nov. 2011.
[51] A. Chuhutin, B. Hansen, and S. N. Jespersen, "Precision and accuracy of diffusion kurtosis estimation and the influence of b-value selection," *NMR Biomed.*, vol. 30, no. 11, pp. e3777-1–e3777-25, Aug. 2017.
[52] T. Zhu, R. Hu, W. Tian, S. Ekholm, G. Schifitto, X. Qiu, and J. Zhong, "Spatial regression analysis of diffusion tensor imaging (SPREAD) for longitudinal progression of neurodegenerative disease in individual subjects," *Magn. Reson. Imag.*, vol. 31, no. 10, pp. 1657–1667, Dec. 2013.
[53] M. Lin, Q. Chen, and S. Yan, "Network in network," 2014, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400
[54] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. ICLR*, Apr. 2014, pp. 1–16.
[55] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 791–803, 1964.
[56] Y. E. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Sov. Math. Dokl.*, vol. 27, no. 2, pp. 372–376, 1983. [Online]. Available: https://mpawankumar.info/teaching/cdt-big-data/nesterov83.pdf
[57] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[58] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[59] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, 1991.
[60] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.
[61] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[62] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Apr. 2011, pp. 315–323.
[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[64] E. Caruyer, C. Lenglet, G. Sapiro, and R. Deriche, "Design of multishell sampling schemes with uniform coverage in diffusion MRI," *Magn. Reson. Med.*, vol. 69, no. 6, pp. 1534–1540, Jun. 2013.
[65] J. L. R. Andersson and S. N. Sotiropoulos, "An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging," *NeuroImage*, vol. 125, pp. 1063–1078, Jan. 2016.
[66] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, Oct. 2002.
[67] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, May 2010, pp. 249–256.
[68] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*, 2nd ed. G. Montavon, G. B. Orr, and K.-R. Müller, Eds. New York, NY, USA: Springer, 2012, pp. 437–478.
[69] S. M. Smith, "Fast robust automated brain extraction," *Hum. Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.
[70] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras
[71] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. [Online]. Available: http://tensorflow.org/
[72] J. Talairach and P. Tournoux, *Co-Planar Stereotaxic Atlas of the Human Brain: 3-D Proportional System: An Approach to Cerebral Imaging*. New York, NY, USA: Thieme, 1988.
[73] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Proc. ICNC*, 2008, pp. 192–201.
[74] Y. Assaf, R. Z. Freidlin, G. K. Rohde, and P. J. Basser, "New modeling and experimental framework to characterize hindered and restricted water diffusion in brain white matter," *Magn. Reson. Med.*, vol. 52, no. 5, pp. 965–978, Nov. 2004.
[75] Y. Assaf, T. Blumenfeld-Katzir, Y. Yovel, and P. J. Basser, "Axcaliber: A method for measuring axon diameter distribution from diffusion MRI," *Magn. Reson. Med.*, vol. 59, no. 6, pp. 1347–1354, Jun. 2008.

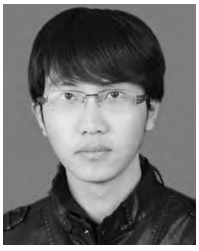
ZHIWEI LI received the B.Eng. degree from the Department of Instrument Science & Technology, Zhejiang University (ZJU), Hangzhou, China, in 2013, where he is currently pursuing the Ph.D. degree. His research interests include network communication, image processing, and artificial intelligence.

TING GONG received the B.S. degree in communication engineering from the China University of Geosciences, Wuhan, China, in 2015. She is currently pursuing the Ph.D. degree with biomedical engineering with Zhejiang University. Her research interests include image acquisition, model reconstruction, and fiber tractography in diffusion MRI.

ZHICHAO LIN received the B.Eng. degree from the Department of Instrument Science & Technology, Zhejiang University (ZJU), Hangzhou, China, in 2013, where he is currently pursuing the Ph.D. degree. His research interests include image processing, digital signal processing, and artificial intelligence.

HONGJIAN HE with received the Ph.D. degree in physics from Zhejiang University, China, in 2011, where he is currently an Associate Professor. His research interests include the advanced methodology development and application of magnetic resonance imaging. He has authored or coauthored more than 20 scientific papers. He is an Associate Member of the International Society of Magnetic Resonance in Medicine (ISMRM).

QIQI TONG received the B.S. degree from the Department of Biomedical Engineering, Zhejiang University (ZJU), Hangzhou, China, in 2013, where she is currently pursuing the Ph.D. degree. Her research interests include image acquisition, model reconstruction, and fiber tractography in diffusion MRI.



CHEN LI received the B.S. degree in biomedical engineering from Yan Shan University (YSU), Qinhuangdao, China, in 2016. He is currently pursuing the master's degree with biomedical engineering with Zhejiang University. His research interests include MRI image processing, MRI quality control, and artificial neural networks.

YI SUN received the Ph.D. degree in medical science from Siemens Healthcare, China, in 2014, where he is currently a Lead Research Scientist in MR collaboration. He has authored or coauthored more than 20 scientific papers. His research interests include advanced methodology development and the application of magnetic resonance imaging.

FENG YU (M'03) received the B.S. degree in semiconductor from Nankai University, in 1988, and the M.S. and Ph.D. degrees in instrumentation engineering from Zhejiang University (ZJU), in 1991 and 2007, respectively, where he is currently a Professor with the College of Biomedical Engineering and Instrument Science. His current research interests include heterogeneous computing architecture, computational biology, and emerging technologies for MRI image processing.

JIANHUI ZHONG is currently a Professor with the University of Rochester and Zhejiang University. He has more than 30 years experience in magnetic resonance imaging (MRI) physics, engineering and biomedical/clinical/neuroscience applications, with more than 190 publications. He is one of the leading researchers for complex relaxation and diffusion effects and interactions of magnetic susceptibility and diffusion in tissues, the application of diffusion in brain electrical activities and functions, and is one of the major researchers for intermolecular multiple quantum coherence brain imaging. He is a grantee of many national and private funds, awardees in research excellence by the NIH, Yale and the University of Rochester, and he holds five U.S. patents. His research interests include brain diffusion tensor imaging, the quantitative analysis of brain functional imaging and network analysis, fast MR imaging techniques and their applications in neuroscience, translational medicine, and individual imaging analysis.