A deep learning-based method for reducing residual motion effects in diffusion parameter estimation

Ting Gong [1,2], Qiqi Tong [1], Zhiwei Li [3], Hongjian He [1], Hui Zhang [2], Jianhui Zhong [1,4]

[1] Centre for Brain Imaging Science and Technology, Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou, China
[2] Department of Computer Science & Centre for Medical Image Computing, University College London, UK
[3] Department of Instrument Science & Technology, Zhejiang University, Hangzhou, China
[4] Department of Imaging Sciences, University of Rochester, Rochester, NY, United States

Correspondence: Hongjian He, hhezju@zju.edu.cn
Address for Correspondence: Zhejiang University, No. 38 Zheda Road, Hangzhou, China, 310027
Hui Zhang and Jianhui Zhong contributed equally to this work.

Abstract

Purpose: Conventional motion correction techniques for diffusion MRI can introduce motion level–dependent bias in derived metrics. To address this challenge, a deep learning-based technique is developed to minimize such residual motion effects.

Methods: The data-rejection approach is adopted in which motion-corrupted data are discarded before model fitting. A deep learning-based parameter estimation algorithm, using hierarchical convolutional neural network (H-CNN), is combined with a procedure of motion assessment and corrupted volume rejection. The method has been designed to overcome limitations of existing methods of this kind that produce parameter estimation whose quality depends strongly on the proportion of the data discarded. Evaluation experiments are conducted for estimation of diffusion kurtosis and diffusion tensor derived measures at both individual and group levels. The performance is compared to robust approach of iteratively reweighted linear least squares (IRLLS) after motion correction with and without outlier replacement.

Results: Compared to IRLLS, the H-CNN-based technique is minimally sensitive to motion effects, as tested at severe motion levels when 70%-90% of the data are rejected, and when random motion is present, showing stable performance independent of the numbers and schemes of data rejection. A further test on dataset from children with attention deficit hyperactivity disorder demonstrates the technique can potentially ameliorate spurious group-level difference due to head motion.

Conclusion: This method shows great potential for reducing residual motion effects in motion-corrupted DWI data, bringing benefits that include reduced bias in derived metrics in individual scans and reduced motion-level dependent bias in population studies employing diffusion MRI.

Keywords: head motion; diffusion kurtosis imaging; diffusion tensor imaging; neural network.

---

1. Introduction

Recent years have seen a growth in diffusion models that can derive unique measures of local microstructural tissue properties from diffusion-weighted MRI (DWI) data. Beyond the most widely used diffusion tensor imaging (DTI) model (1), more advanced models, including diffusion kurtosis imaging (DKI) (2,3), and a variety of microstructural models (4–6), have been developed. These advanced models in general require substantially larger number of DWI volumes and the application of stronger diffusion gradients, with consequently longer acquisition times and more stringent demands on image quality.

However, the imaging principle of DWI makes it vulnerable to various image artefacts (7), especially those due to subject motion, with longer acquisition more prone to such artefacts. Motion introduces broadly two types of artefacts: spatial misalignment and intensity corruption. Spatial misalignment can be caused by subject movements either between acquisition of consecutive DWI volumes, resulting in misalignment between them, or between acquisition of consecutive slices within a single volume, resulting in misalignment within it. Signal dropouts can be caused by motion during diffusion encoding. The number of affected volumes in a scan depends on the level of motion, which if uncorrected, can increase uncertainty in model fitting and introduce bias in derived measures (7–9).

Retrospective methods are commonly used to correct for motion artefacts. While these techniques have proved valuable for mitigating the effect of motion, the correction is never perfect, inevitably retaining some residual artefacts. The extent of such residual motion effects similarly depends on the level of motion in a scan. For example, spatial misalignment is corrected with the common strategy of image registration-based realignment, for which many tools have been developed (10–14). However, the correction procedure requires accurate estimation of motion, which can be more difficult in the presence of large motion, and image interpolation, a step that will blur the corrected volumes. The number of volumes affected, as well as the extent to which the quality of the volumes will be impacted, will depend on the level of motion in a scan. Signal dropouts are even more challenging to correct; the common strategy is to detect them as outliers and to replace them with data-driven predictions (15). The accuracy of the predictions however, will be affected by the number of measurements without outliers. This again can lead to residual motion effects that is motion-level dependent.

These residual effects of head motion can compromise model fitting, introducing bias to model-derived measures. This can negatively impact downstream group-level studies if the level of motion is different between groups. At individual level, residual motion effects have been shown to lead to systematic errors in estimation of DTI-derived measures in white matter (WM) (12,16). At group level, such effects following registration-based motion correction have been demonstrated to introduce spurious group differences in DTI-derived measures (17). Similar effects could also result from correction techniques based on outlier replacement (15), which has been shown to introduce bias to diffusion-derived structural connectivity (18).

The growing recognition of residual motion effects associated with conventional motion correction methods motivates research into alternative motion-mitigating strategies. One such approach is to detect and reject part of the data that is motion corrupted before model fitting for each individual. This has been adopted by a number of techniques, ranging from discarding each affected DWI volume entirely (19,20), to advanced robust estimators based on voxel-wise or slice-wise rejection (12,21–24). However, rejecting data also has detrimental impact on model fitting (25,26). For example, it can result in parameter estimates of greater variations in accuracy and precision if the number of unrejected data points at each voxel varies greatly. In the extreme case, some voxels may not have enough data points to support fitting at all, rendering these voxels unusable in downstream analyses. These effects are especially detrimental to studies of motion-prone populations, such as young children. Moreover, when applied at group level, if more data points are rejected in one group compared to the other, these effects can introduce group-level bias. Hence, the success of such data-rejection approaches requires the development of parameter estimation methods that are extremely robust to the number of input data points.

In this study, we propose to utilize deep learning (DL)-based parameter estimation method to address residual motion effects. This is motivated by the fact that DL-based methods have recently demonstrated excellent performance in estimating diffusion-derived measures from highly under-sampled data (27–31). Here, we demonstrate this new approach by combining a DL-based method utilising a hierarchical convolutional neural network (H-CNN) (29) with a motion assessment and data rejection procedure (Figure 1(C)). The proposed technique is evaluated for recovering DKI and DTI derived measures from motion contaminated data. Its performance is compared to fitting-based methods after motion correction (Figure 1(A), (B)).

2. Methods

This section details the proposed motion-correction pipeline and the evaluation experiments.

2.1 The DL-based technique

Similar to conventional techniques, for each study subject, the proposed technique takes their DWIs, many of which may be motion corrupted, as an input. It aims to estimate the subject's derived diffusion metrics with accuracy and precision comparable to those that would have been derived from the corresponding motion-free DWIs had they been available.

Unlike conventional techniques, the proposed method requires an additional input – a few training subjects. This is because at the heart of the technique is a patch-based H-CNN model (29), trained with supervised

learning, to map any given subset of the full data to the metrics derived from the full data. To train the model, the technique will require minimally motion-contaminated data from 1 to 2 subjects who are able to lie still over the entire diffusion acquisition; these subjects will be referred to as the training subjects henceforth.

The pipeline of the technique, shown in comparison to fitting-based methods in Figure 1, consists of four steps: (1) pre-processing, (2) motion assessment and data rejection, (3) H-CNN model training, and (4) diffusion metric estimation. Briefly, for each subject, its original DWIs are first pre-processed to correct for distortion and motion, with conventional techniques. The motion parameters derived as part of the pre-processing are subsequently used to assess the level of motion of each DWI and to reject the volumes deemed to have moved excessively. Next, the b-values and the diffusion gradient vectors associated with the remaining DWIs are used to select the desired subset from the separate training data to train the H-CNN model for this subject. Finally, the trained model is used to estimate diffusion metrics from the remaining DWIs. The detailed description of each step is given below.

2.1.1 Data Pre-processing

The original DWI data are corrected for geometric distortion, volumetric motion and signal dropouts using TOPUP and EDDY from the FMRIB Software Library. Motion-correcting transformations estimated with EDDY are used at the next step for motion assessment. In more detail, the pre-processing begins with estimating the field map with TOPUP (32), using a pair of b=0 images with reversed phase-encoding directions (See Data Acquisition). Next, the estimated field map is fed into EDDY, allowing an integrated correction of volumetric motion and both eddy-current and B0 field inhomogeneity-induced distortion for the acquired DWIs (13), with the first b=0 volume treated as the reference. When applying EDDY, its outlier detection and replacement feature (15) is used along with its distortion and motion correction functionalities: image slices detected with signal dropouts in each DWI volume are taken as outliers and replaced with predictions made by a Gaussian process using angularly neighbouring measurements. The number of outlier slices is also used for motion assessment. After pre-processing, all DWI volumes are realigned to the first b=0 image and the corresponding gradient vectors are reoriented accordingly (33).The data is only resampled once in the whole procedure.

2.1.2 Motion assessment and data rejection

Exploiting the estimated rigid transforms and slice outliers identified by EDDY, volume-based motion assessment measures are calculated to monitor both between- and within-volume motion (17,26). By setting stringent thresholds of these measures, only the volumes with minimal motion, defined as having all measures smaller than or equal to the thresholds, are retained for subsequent parameter estimation.

We define two types of motion assessment measures. The first type of measures characterises the aggregate movement of a volume relative to the first b=0 image, the reference. Specifically, its absolute translation and rotation (AT, AR) relative to the reference are computed. Large values of these tend to reduce image quality when the corresponding volumes are realigned to the reference. The second type of measures assess the more transient movement that tends to cause signal dropouts. As this may be reflected by the relative movement between consecutive volumes, the relative translation and rotation (RT, RR) of a volume with respect to the adjacent volume in front are computed. Furthermore, the fraction of slices with signal dropouts (FSD) in a volume, based on the number of slice outliers identified, is also computed to more directly quantify the extent of signal dropouts due to within-volume motion. Slices with too few brain voxels (< 250) to ascertain signal dropouts are excluded. The motion measures for the i-th volume in a scan are calculated as follows:

$$AT_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$$

$$AR_i = |\theta_i| + |\phi_i| + |\psi_i|$$

$$RT_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}$$

$$RR_i = |\theta_i - \theta_{i-1}| + |\phi_i - \phi_{i-1}| + |\psi_i - \psi_{i-1}|$$

$$FSD_i = \frac{NO_i}{N_{total}} \times 100\%$$

where x_i , y_i, z_i are its translation components; θ_i, ϕ_i, ψ_i its rotation components around the X, Y, Z axes; NO_i the number of slices detected with signal dropouts; N_{total} the total number of slices in a volume. Because the first b=0 volume is used as the reference for the remaining ones, all its components ($x_1, y_1, z_1, \theta_1, \phi_1, \psi_1, NO_1$), and the corresponding measures ($RT_1, RR_1, AT_1, AR_1, RSD_1$) are set to zero.

2.1.3 H-CNN model and training

A subject-specific H-CNN model is trained with a separate set of training subjects. The network architecture of H-CNN is fully described in (29), but for completeness it is included here as Supporting Information Figure S1.. Two aspects of the model are relevant here. First, it is a patch-based technique. The training samples are small 3x3x3 patches; as hundreds of thousands of these patches can be drawn from a single subject, training requires only a few subjects. Second, the model is tailored for any desired subset of the full data. For any given subset, its corresponding model is trained to map this subset to the target diffusion metrics derived from the full data with model fitting. In the present context, for any given study subject, the desired subset of the training data is precisely the measurements that have the b-values and the diffusion gradient vectors in common with those associated with the remaining volumes of the study subject after

data rejection. The training time for H-CNN model is 5 to 10 minutes (depending on the number of remaining volumes) with a Tesla k20 graphics card for GPU acceleration, which is comparable to the time of conventional model-fitting methods with 8 CPUs.

2.1.4 Parameter estimation

Once the subject-specific model is trained, the desired motion-free diffusion metrics are estimated by applying the trained model to the remaining volumes of the study subject. This parameter estimation process is nearly instantaneous.

2.2 Evaluation

To assess the proposed DL-based technique in comparison with conventional fitting-based methods, two primary experiments are conducted for (1) quantitative evaluation with ground truth acquired from individuals and (2) quantitative evaluation of robustness to random motion rejection. In addition, a secondary, supporting, group-level demonstration is conducted in experiment (3).

For comparison to conventional fitting-based methods, the weighted linear least squares (WLLS) estimator (34) is chosen to represent method in Figure 1(A) and an advanced approach of iteratively reweighted linear least squares (IRLLS) (23) is chosen to represent method in Figure 1(B). The advanced IRLLS approach fits the model with WLLS after a voxel-wise outlier detection and rejection. The IRLLS approach is chosen because it offers comparable estimation accuracy and precision to more advanced nonlinear estimators, such as RESTORE (21) and iRESTORE (22), but is substantially faster in estimation speed (23).

DKI and DTI derived measures are estimated jointly, including fractional anisotropy (FA), radial kurtosis (RK), mean diffusivity (MD), mean kurtosis (MK), radial diffusivity (RD), axial diffusivity (AD), axial kurtosis (AK), and kurtosis fractional anisotropy (KFA) (35,36). WLLS and IRLLS use the MATLAB codes available at https://github.com/NYU-DiffusionMRI/DESIGNER/tree/master/parameter_estimation. The H-CNN model is implemented in-house using Keras (37) with TensorFlow (38) backend.

2.2.1 Experiment 1: Quantitative evaluation

To allow quantitative evaluation with known ground truth, we collect both motion-corrupted and motion-free dataset from same individuals.

Dataset

Data were collected from 2 healthy subjects (S1, S2) who were scanned twice in the same session. In the first scan they were asked to lie still; these motion-free data serve as the ground truth. In the second scan they were asked to perform deliberate head motion to produce motion-contaminated data. Another two

subjects (S3, S4) were scanned once, lying still, constituting the training dataset for H-CNN. All data were collected on a MAGNETOM Prisma 3T scanner (Siemens Healthineers, Erlangen, Germany) using a 64-channel head-neck coil. The local ethical committee approved this study and written informed consent was obtained from each participant.

The diffusion imaging parameters for each scan were as follows: single-shot echo-planar imaging (EPI) sequence; TR/TE = 7000/67 ms; FOV = 210×210 mm^2; slice number = 50; resolution = $2.5 \times 2.5 \times 2.5$ mm^3; slice acceleration factor = none; phase acceleration factor = 2; phase partial Fourier = none; bandwidth = 2126 Hz/pixel; diffusion weightings of b = 1000, 2000, and 3000 s/mm^2 were applied in 30 distinct directions, respectively, with six $b = 0$ volumes acquired, resulting in a total of 96 volumes. The diffusion weightings and directions were designed using a uniform coverage across multiple shells and an incremental scheme by a generalization of electrostatic repulsion (39), making diffusion vectors from different b-values different and interspersed temporally. A $b = 0$ volume with an opposite phase-encoding direction was also acquired. The total diffusion acquisition time was 12 mins for each scan. T1-weighted images were additionally acquired using an MPRAGE sequence for anatomical reference.

Analysis

Parameter maps from motion-contaminated data are estimated with H-CNN and IRLLS, respectively. They are quantitatively compared to the ground truth, taken as the parameter maps from the data in the still condition estimated with IRLLS. Stringent thresholds of AT < 3 mm, AR < 3°, RT < 2 mm, RR < 2°, and FSD < 5% are applied for DL-based method. The values have been chosen to broadly correspond to the upper limit of these measures seen for the still scans. The target diffusion metrics for the training data (S3 and S4) are estimated from IRLLS.

To assess the effect of motion thresholds on H-CNN, a range of less stringent alternatives are tested. With less stringent thresholds, the number of data points available for parameter estimation is higher, but at the same time, the number of motion-contaminated data points is also higher. This analysis investigates this trade-off for H-CNN. For comparison, WLLS and IRLLS are also tested, including additionally data without outlier replacement. For quantitative evaluation, the root-mean-squared errors (RMSEs) from WM voxels are computed and compared for each measure. WM voxels are determined by a five-tissue segmentation method using T1-weighted images (40).

2.2.2 Experiment 2: Robustness to random motion rejection

To test the robustness of each technique to the number and scheme of the remaining data, random rejection tests are conducted using data from S1 in the still condition from experiment 1.

This experiment controls for data rejection, allowing parameter estimation to be assessed specifically, i.e. comparing WLLS to H-CNN estimation directly. For each tested number of retained volumes N, 100 sub-sampled schemes are drawn randomly from the full scheme (with the first b=0 volume and at least two different b-values always included). Each sub-sampled scheme is used to evaluate both techniques. Additionally, several underdetermined schemes for WLLS are included to assess H-CNN further; they are N = 20, 16, and 12.

WM RMSEs are calculated for each random rejection case. To allow for a higher-quality reference standard for assessment, two more data repetitions from S1 in the still condition are acquired and combined with the original repetition. Additionally, WM RMSEs calculated with respect to the IRLLS-estimated maps from a single repetition as in experiment 1 are given in Supporting Information Figure S3 for comparison. To statistically test the robustness of both methods over different number of DWIs remained and different rejection schemes, the Levene's test for equal variance (41) on the RMSEs are conducted. Finally, a simulation study is also conducted on the estimated measures to evaluate the influence of these methods on the power of detecting differences. Details can be found in Supporting Information Figure S4.

2.2.3 Experiment 3: Group-level evaluation

To demonstrate the method at group level, data with varying motion levels are employed and divided into a control group with small motion, and a test group with large motion. Voxel-wise statistical analysis is then carried out with tract-based spatial statistics (TBSS) (42) using parameters estimated from IRLLS and proposed technique.

Dataset

Data from 19 children diagnosed with attention deficit hyperactivity disorder (ADHD) (5 females and 14 males; age: 10.45 ± 2.81 yr) are employed from the Healthy Brain Network biobank (43). The diffusion data were collected on a Siemens Prisma 3T scanner with the following parameters: simultaneous multi-slice EPI sequence; resolution = $1.7 \times 1.7 \times 1.7$ mm^3; slice acceleration factor = 3; one b=0 s/mm image, and diffusion weightings of b=1500, and 3000 s/mm^2 applied in the same 64 directions in each shell sequentially. One b=0 image pair in the reversed phase-encoding direction was acquired.

Grouping and estimation

To divide subjects into two groups, a total motion index (TMI) is calculated that summarizes head motion for each subject from all motion measures (17). We divide subjects with TMI<0 into the control group and subjects with TMI>0 into the motion group. The TMI for the i-th subject is calculated as follows:

$$TMI_i = \sum_{j=1}^{5} \frac{x_{ij} - M_j}{Q_j - q_j}$$

where j = 1,…, 5 indexes the five average motion measures across all DWIs ($\overline{AT}$, $\overline{AR}$, $\overline{RT}$, $\overline{RR}$, $\overline{FSD}$); x_{ij} is the value of the j-th average motion measure for the i-th subject; and M_j, Q_j, and q_j are, respectively, the median, upper quartile, and lower quartile of the j-th average motion measure over all subjects included in the group comparison.

To ensure training and testing do not perform on the same subject for the DL-based method, data from 3 of the 19 subjects with small TMI constitute the training dataset, in which data from 2 of the 3 subjects are employed as training data for the other 16 subjects. Meanwhile, for each subject in the 3 training subjects, data from the other 2 subjects are employed for training. During parameter estimation, stringent thresholds of AT < 3 mm, AR < 3°, RT < 1 mm, RR<1°, and FSD < 5% are applied for H-CNN estimation; more stringent thresholds of relative motion measures are employed here to account for the higher image resolution of this dataset.

Analysis

To evaluate whether residual motion or data rejection introduce bias into analysis for the IRLLS and H-CNN methods, two-sample *t*-tests are conducted with the derived FA, MD, and RK measures between the control and motion groups. For each test, 5000 permutations of the data are generated (44). To further investigate whether IRLLS and H-CNN estimations are different from each other, one sample *t*-tests are performed using the difference maps estimated from the two methods for the two groups. Exhaustive sign-flip permutations are run for each test. The false discovery rate is used to correct for multiple comparisons with $P = 0.05$ as the threshold for significance. To further test the effects of method used, motion level and their interaction to estimated diffusion metrics, a two-factor mixed measures statistical test is conducted on the mean of diffusion metrics on the major WM skeleton. Voxel wise interaction is additionally tested by two-sample *t*-tests between two group of subjects on both the arithmetic differences and the absolute differences of IRLLS and H-CNN estimations. Details about the interaction tests can be found in Supporting Information S5 and S6.

3. Results

3.1 Experiment 1

The motion measures of each DWI from S1 and S2 are depicted in Figure 2. Evidently these measures are considerably higher for the data acquired in the moving condition than for those in the still condition. The motion-contaminated scan from S2 has suffered more severe motion than that from S1. The number of DWIs retained for different motion thresholds are listed in Table 1.

Figure 3 demonstrates two representative volumes with large relative motion measures after correction without and with outlier replacement from these subjects. Signal dropouts are evident before outlier replacement, suggesting the relative motion measures are effective for assessing within-volume motion. Outlier replacement appears to improve S1 substantially more than S2 for whom the motion is more severe.

The estimated MD, FA, and RK maps are shown in Figure 4. The H-CNN derived maps are minimally sensitive to residual motion effects, with good image contrast and small difference compared to their still references for these subjects. In contrast, IRLLS performs better for S1 than for S2. The maps for S1 are noisier than its references but they otherwise are almost identical. However, the maps for S2 are severely blurred compared to its ground truth, losing important anatomical details.

Figure 5 and Supporting Information Figure S2 show the quantitative evaluation of the estimation accuracy of diffusion derived measures using WM RMSEs for each data-rejection motion thresholds listed in Table 1. Under the stringent thresholds, H-CNN outperforms both WLLS and IRLLS, despite the number of its retained DWIs being considerably smaller than that of the full data. Moreover, its RMSEs for these subjects are similar, despite the levels of motion and the numbers of retained DWIs are different between them, demonstrating its robustness to different levels of motion. In contrast, IRLLS-derived measures from the full data suffer from evident motion-level dependent residual effects, with higher RMSEs for S2. In addition, the challenge facing fitting-based methods is clearly illustrated with WLLS without outlier replacement: while its RMSEs decrease at first, when the volumes with severe motion are rejected, they later increase when the number of data points decreases. Finally, note that our results replicate the existing findings that IRLLS outperforms WLLS and outlier replacement generally reduces the effects of residual motion with full data.

3.2 Experiment 2

The boxplots of RMSEs from IRLLS and H-CNN to random motion rejections are shown in Figure 6; the mean and standard deviation for each case are given as Supporting Information Table S1; the statistics of equal variance test are given in Table 2. In general, H-CNN method provides more accurate and robust estimates for different rejection number and random rejection schemes than IRLLS. Specifically, for all measures, the median and interquartile range (IQR) of RMSEs for IRLLS increase quickly when the retained volumes decrease from 60 to 30. In contrast, the RMSEs for H-CNN remain comparatively stable

from 60 to 12 volumes for all measures. The statistics from Table 2 further suggest that H-CNN provides parameter estimates more robust to random rejection with much more stable and smaller variances. Finally, the power analysis demonstrates that the effect sizes from IRLLS and H-CNN show improvement over WLLS, with H-CNN producing values closest to the ground truth; see Supporting Information Figure S4 for detail.

3.3 Experiment 3

The motion assessment results of each subject ranked by TMI are depicted in Figure 7. The control group includes 8 subjects (average $\overline{\mathrm{RT}}$ / $\overline{\mathrm{RR}}$ / $\overline{\mathrm{FSD}}$ / number of rejected volume: 0.25 mm / 0.19° / 1.39% / 15) and the motion group includes the other 11 subjects (average $\overline{\mathrm{RT}}$ / $\overline{\mathrm{RR}}$ / $\overline{\mathrm{FSD}}$ / number of rejected volume: 0.53 mm / 0.61° / 2.89% / 63).

The voxel-wise TBSS results in Figure 8 suggest that while the IRLLS alleviates the negative impact of motion on FA measures, some other measures, such as MD and RK, still suffer from the deterioration of model fitting with data of different levels of motion. Specifically, for IRLLS, motion tends to increase MD and decrease RK, demonstrated by over 50% and 80% of the WM skeleton showing significantly lower MD and significantly higher RK respectively in the control group than the motion group. The one-sample *t*-tests of difference between estimations of the two methods further demonstrate the ability of H-CNN to reduce the residual motion effects: there is no significant difference between IRLLS and H-CNN for MD, FA and RK in the control group. In contrast, in the motion group, Estimated MD and RK are significantly different between H-CNN and IRLLS for over 60% of the WM skeleton. Statistics from the two-factor mixed measures design provided in the Supporting Information Figure S5 and Supporting Information Table S2 further show that (1) for FA measure, there is no significant impact from the method used, the motion level, and their interaction; (2) but for RK and MD measure there is significant impact from both factors and their interaction, with the presence of large motion and the use of IRLLS inducing the strongest deviation in these metrics from the control group. This finding agrees with the voxel-wise TBSS results in Figure 8 and Supporting Information Figure S6.

4. Discussion

In this study, a DL-based technique is proposed to reduce the effects of residual motion in diffusion parameter estimation. Such a technique is needed because motion-level dependent residual effects are increasingly recognised as being commonly present following standard motion correction. While the existing data rejection approaches provide an improvement, their reliance on conventional model fitting renders their performance dependent on the number of remaining data points. Our approach takes advantage

of recent advances in diffusion parameter estimation with DL models. Results suggest that the proposed technique provides robust estimations of DKI- and DTI-derived measures with minimum effects of residual motion at both individual and group levels. Overall, the technique provides great potential to make full use of motion-corrupted data.

Compared to fitting-based methods for reducing residual motion effects, the advantage of the DL-based technique is its robustness to large rejection number and different rejection schemes. For a fitting-based method such as IRLLS, data redundancy is required. It has been shown that the minimum number of distinct gradient directions necessary for robust estimation of FA values is approximately 30 (45). For higher-order DKI measures such as RK and MK, the estimation quality is closely related to the number of DWIs as well as multiple b-values. Hence, a different rejection number in different voxels or subjects could introduce a bias to the estimation performance. Moreover, as suggested by the random rejection experiment, the accuracy of fitting varies across the random rejection scheme even with the same rejection number. Some studies have also pointed out that the angular distribution of estimation precision is inhomogeneous (46). The robustness of DL-based technique enables the use of data that would otherwise be abandoned.

The robustness of the DL-based method is gained from its supervised learning process with large-scale training samples. Voxel-wise model fitting may be compromised by an inadequate number of measurements or an orientation-unbalanced sampling scheme. The DL-based method, on the other hand, benefits from the joint optimization of large number of training voxels containing ample and varying tissue properties and orientation information from the whole brain. Combining large and rich training samples with a strong inference ability, the DL-based method could reduce the number of needed DWIs with very steady estimation performance. This lays the foundation for rejecting outlier volumes without deteriorating the estimation performance.

Utilization of motion-contaminated data can be maximised with suitable acquisition design. First, it is advisable to acquire dataset using an incremental scheme for b-value arrangement like the one in experiment 1. Interspersing diffusion vectors of different b-values temporally maximises the probability that the remaining data contain multiple b-values, even in the event of an early-terminated scan. This is especially important for DKI and other microstructural models where data of multiple b-values are crucial. Second, it is recommended to sample different diffusion vectors for different b-values. With this strategy, any subset data would be more likely to result in a denser coverage of the angular space. These considerations are important for conventional model fitting and could similarly benefit the proposed approach, by maximising the richness of information available to train DL-based models.

The current study has several potential limitations. First, the proposed method requires at least one high-quality DWI training dataset with the same imaging parameters. This, however, should not be a major

burden, since typical studies include enough subjects from which a few with minimal motion can be identified as the training dataset, as in experiment 3. As long as one training dataset is available, the network can be trained and applied to any possible motion-affected patterns. This limitation could additionally be addressed by performing diffusion data simulations (47), which is an ongoing area of our current research.

The second potential limitation is the possibility of not rejecting the volumes with intra-volume motion but no signal dropouts. This can be remedied using more computationally expensive slice-to-volume registration tools (48) to identify and reject such volumes, which could improve the proposed technique further.

Another limitation is that experiment 3 for group level demonstration lacks ground truth to draw firm conclusions, as one could not exclude the existence of true differences between the control and motion groups. Nevertheless, this proof-of-concept demonstration has shown that the proposed method is able to reduce false positives due to residual motion compared to conventional techniques.

One possible improvement for our method is to extend our work to more diffusion models. The current study has evaluated the DKI- and DTI-derived measures, which, thus far, are the most widely used diffusion metrics in diffusion MRI. Other models benefitting from the combination of multi-shell protocols with high angular resolution (4,49) will likely face the same challenge. Future studies taking these models into considerations will further test the utility of the method. Another important avenue for improvement is to include uncertainty quantification (50,51) , which could be beneficial for quantifying reliability.

5. Conclusion

With quantitative and statistical benefits demonstrated in this study, the proposed DL-based technique could be a powerful new tool for reducing residual motion effects in motion-contaminated data, providing increased utilization of diffusion data for quantitative studies.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (81871428); the Shanghai Key Laboratory of Psychotic Disorders (13dz2260500); the Major Scientific Project of Zhejiang Lab (No.2018DG0ZX01); the Fundamental Research Funds for the Central Universities (2019QNA5026, 2019XZZX001-01-08); and the Zhejiang University Education Foundation Global Partnership Fund.

Data Availability Statement: The codes for the DL-based pipeline will be made openly available as a toolbox at https://github.com/Tinggong/DLmotion after the necessary documentation has been completed.

References

1. Basser PJ, Mattiello J, Lebihan D. Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo. J. Magn. Reson. Ser. B 1994;103:247–254 doi: 10.1006/jmrb.1994.1037.

2. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: The quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. Magn. Reson. Med. 2005;53:1432–1440 doi: 10.1002/mrm.20508.

3. Jensen JH, Helpern JA. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR Biomed. 2010;23:698–710 doi: 10.1002/nbm.1518.

4. 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;61:1000–1016 doi: 10.1016/j.neuroimage.2012.03.072.

5. Assaf Y, Basser PJ. Composite hindered and restricted model of diffusion (CHARMED) MR imaging of the human brain. Neuroimage 2005 doi: 10.1016/j.neuroimage.2005.03.042.

6. Assaf Y, Blumenfeld-Katzir T, Yovel Y, Basser PJ. AxCaliber: A method for measuring axon diameter distribution from diffusion MRI. Magn. Reson. Med. 2008;59:1347–1354 doi: 10.1002/mrm.21577.

7. Le Bihan D, Poupon C, Amadon A, Lethimonnier F. Artifacts and pitfalls in diffusion MRI. J. Magn. Reson. Imaging 2006 doi: 10.1002/jmri.20683.

8. Aksoy M, Liu C, Moseley ME, Bammer R. Single-step nonlinear diffusion tensor estimation in the presence of microscopic and macroscopic motion. Magn. Reson. Med. 2008 doi: 10.1002/mrm.21558.

9. Jones DK, Basser PJ. “Squashing peanuts and smashing pumpkins”: How noise distorts diffusion-weighted MR data. Magn. Reson. Med. 2004 doi: 10.1002/mrm.20283.

10. Rohde GK, Barnett AS, Basser PJ, Marenco S, Pierpaoli C. Comprehensive Approach for Correction of Motion and Distortion in Diffusion-Weighted MRI. Magn. Reson. Med. 2004 doi: 10.1002/mrm.10677.

11. Pierpaoli C, Walker L. TORTOISE: an integrated software package for processing of diffusion MRI data. … Process. Diffus. … 2010.

12. Oguz I, Farzinfar M, Matsui J, et al. DTIPrep: quality control of diffusion-weighted images. Front. Neuroinform. 2014 doi: 10.3389/fninf.2014.00004.

13. Andersson JLR, Sotiropoulos SN. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 2016;125:1063–1078 doi: 10.1016/j.neuroimage.2015.10.019.

14. Andersson JLR, Graham MS, Drobnjak I, Zhang H, Filippini N, Bastiani M. Towards a comprehensive framework for movement and distortion correction of diffusion MR images: Within volume movement. Neuroimage 2017;152:450–466 doi: 10.1016/j.neuroimage.2017.02.085.

15. Andersson JLR, Graham MS, Zsoldos E, Sotiropoulos SN. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. Neuroimage 2016;141:556–572 doi: 10.1016/j.neuroimage.2016.06.058.

16. Liu B, Zhu T, Zhong J. Comparison of quality control software tools for diffusion tensor imaging. Magn. Reson. Imaging 2015 doi: 10.1016/j.mri.2014.10.011.

17. Yendiki A, Koldewyn K, Kakunoori S, Kanwisher N, Fischl B. Spurious group differences due to head motion in a diffusion MRI study. Neuroimage 2014;88:79–90 doi: 10.1016/j.neuroimage.2013.11.027.

18. Baum GL, Roalf DR, Cook PA, et al. The impact of in-scanner head motion on structural connectivity derived from diffusion MRI. Neuroimage 2018;173:275–286 doi: 10.1016/j.neuroimage.2018.02.041.

19. Jiang S, Xue H, Counsell S, et al. Diffusion Tensor Imaging (DTI) of the brain in moving subjects: Application to in-utero fetal and ex-utero studies. Magn. Reson. Med. 2009 doi: 10.1002/mrm.22032.

20. Tymofiyeva O, Hess CP, Ziv E, et al. A DTI-Based Template-Free Cortical Connectome Study of Brain Maturation. PLoS One 2013 doi: 10.1371/journal.pone.0063310.

21. Chang LC, Jones DK, Pierpaoli C. RESTORE: Robust estimation of tensors by outlier rejection. Magn. Reson. Med. 2005 doi: 10.1002/mrm.20426.

22. 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 doi: 10.1002/mrm.24173.

23. Collier Q, Veraart J, Jeurissen B, Den Dekker AJ, Sijbers J. Iterative reweighted linear least squares for accurate, fast, and robust estimation of diffusion magnetic resonance parameters. Magn. Reson. Med. 2015 doi: 10.1002/mrm.25351.

24. Sairanen V, Leemans A, Tax CMW. Fast and accurate Slicewise OutLIer Detection (SOLID) with informed model estimation for diffusion MRI data. Neuroimage 2018 doi: 10.1016/j.neuroimage.2018.07.003.

25. Chen Y, Tymofiyeva O, Hess CP, Xu D. Effects of rejecting diffusion directions on tensor-derived parameters. Neuroimage 2015;109:160–170 doi: 10.1016/j.neuroimage.2015.01.010.

26. Ling J, Merideth F, Caprihan A, Pena A, Teshiba T, Mayer AR. Head injury or head motion? Assessment and quantification of motion artifacts in diffusion tensor imaging studies. Hum. Brain Mapp. 2012;33:50–62 doi:

10.1002/hbm.21192.

27. Golkov V, Dosovitskiy A, Sperl JI, et al. q-Space Deep Learning: Twelve-Fold Shorter and Model-Free Diffusion MRI Scans. IEEE Trans. Med. Imaging 2016 doi: 10.1109/TMI.2016.2551324.

28. Lin Z, Gong T, Wang K, et al. Fast Learning of Fiber Orientation Distribution Function for MR Tractography Using Convolutional Neural Network. Med. Phys. 2019:mp.13555 doi: 10.1002/mp.13555.

29. Li Z, Gong T, Lin Z, et al. Fast and Robust Diffusion Kurtosis Parametric Mapping Using a Three-dimensional Convolutional Neural Network. IEEE Access 2019;PP:1–1 doi: 10.1109/ACCESS.2019.2919241.

30. Gibbons EK, Hodgson KK, Chaudhari AS, et al. Simultaneous NODDI and GFA parameter map generation from subsampled q-space imaging using deep learning. Magn. Reson. Med. 2019 doi: 10.1002/mrm.27568.

31. Ye C, Li X, Chen J. A deep network for tissue microstructure estimation using modified LSTM units. Med. Image Anal. 2019 doi: 10.1016/j.media.2019.04.006.

32. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. Neuroimage 2003;20:870–888 doi: 10.1016/S1053-8119(03)00336-7.

33. Leemans A, Jones DK. The B-matrix must be rotated when correcting for subject motion in DTI data. Magn. Reson. Med. 2009 doi: 10.1002/mrm.21890.

34. Veraart J, Sijbers J, Sunaert S, Leemans A, Jeurissen B. Weighted linear least squares estimation of diffusion MRI parameters: Strengths, limitations, and pitfalls. Neuroimage 2013 doi: 10.1016/j.neuroimage.2013.05.028.

35. Veraart J, Poot DHJ, Van Hecke W, et al. More accurate estimation of diffusion tensor parameters using diffusion kurtosis imaging. Magn. Reson. Med. 2011 doi: 10.1002/mrm.22603.

36. Glenn GR, Helpern JA, Tabesh A, Jensen JH. Quantitative assessment of diffusional kurtosis anisotropy. NMR Biomed. 2015 doi: 10.1002/nbm.3271.

37. Chollet F, others. Keras. 2015.

38. GoogleResearch. TensorFlow: Large-scale machine learning on heterogeneous systems. Google Res. 2015 doi: 10.1207/s15326985ep4001.

39. Caruyer E, Lenglet C, Sapiro G, Deriche R. Design of multishell sampling schemes with uniform coverage in diffusion MRI. Magn. Reson. Med. 2013;69:1534–1540 doi: 10.1002/mrm.24736.

40. Smith RE, Tournier JD, Calamante F, Connelly A. Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. Neuroimage 2012 doi: 10.1016/j.neuroimage.2012.06.005.

41. Levene H. Robust tests for equality of variances. Contrib. to Probab. Stat. Essays … 1960.

42. Smith SM, Jenkinson M, Johansen-Berg H, et al. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. Neuroimage 2006 doi: 10.1016/j.neuroimage.2006.02.024.

43. Alexander LM, Escalera J, Ai L, et al. Data Descriptor: An open resource for transdiagnostic research in pediatric mental health and learning disorders. Sci. Data 2017 doi: 10.1038/sdata.2017.181.

44. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. Neuroimage 2014 doi: 10.1016/j.neuroimage.2014.01.060.

45. Jones DK. The Effect of Gradient Sampling Schemes on Measures Derived from Diffusion Tensor MRI: A Monte Carlo Study. Magn. Reson. Med. 2004 doi: 10.1002/mrm.20033.

46. Sprenger T, Sperl JI, Fernandez B, et al. Bias and precision analysis of diffusional kurtosis imaging for different acquisition schemes. Magn. Reson. Med. 2016 doi: 10.1002/mrm.26008.

47. Graham MS, Drobnjak I, Zhang H. Realistic simulation of artefacts in diffusion MRI for validating post-processing correction techniques. Neuroimage 2016;125:1079–1094 doi: 10.1016/j.neuroimage.2015.11.006.

48. Andersson JLR, Graham MS, Drobnjak I, Zhang H, Filippini N, Bastiani M. Towards a comprehensive framework for movement and distortion correction of diffusion MR images: Within volume movement. Neuroimage 2017 doi: 10.1016/j.neuroimage.2017.02.085.

49. Jeurissen B, Tournier JD, Dhollander T, Connelly A, Sijbers J. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. Neuroimage 2014;103:411–426 doi: 10.1016/j.neuroimage.2014.07.061.

50. Tanno R, Worrall DE, Ghosh A, et al. Bayesian image quality transfer with CNNs: Exploring uncertainty in dMRI super-resolution. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). ; 2017. doi: 10.1007/978-3-319-66182-7_70.

51. Ye C, Li Y, Zeng X. An improved deep network for tissue microstructure estimation with uncertainty quantification. Med. Image Anal. 2020;61 doi: 10.1016/j.media.2020.101650.

Table 1. The number of volumes retained (N) with different motion measure thresholds for S1 and S2 in motion condition. The first b = 0 volume is always retained. There are multiple b-values included in all motion thresholding conditions.

Thresholds					N	
AT/mm	AR/°	RT/mm	RR/°	RSD	S1	S2
3	3	2	2	5%	29	10
5	5	2.5	2.5	8%	46	12
6	6	3	3	10%	64	22
8	8	4	4	15%	87	50
10	10	5	5	20%	93	64

Table 2. Statistics (F statistics and p value) for the Levene's test of equal variance of RMSEs among different remaining volumes N for IRLLS and H-CNN.

	IRLLS		**H-CNN**			
	Levene's statistic (N=60~30)		Levene's statistic (N=60~30)		Levene's statistic (N=60~12)	
	F	p	F	p	F	p
AD	7.699	<0.001	3.204	0.041	4.528	<0.001
MD	24.62	<0.001	1.969	0.141	1.459	0.201
RD	13.232	<0.001	6.57	0.001	3.875	0.001
FA	12.639	<0.001	0.32	0.725	1.534	0.177
AK	15.342	<0.001	1.93	0.146	1.713	0.129
MK	19.193	<0.001	1.557	0.212	3.449	0.004
RK	24.032	<0.001	5.198	0.006	4.594	<0.001
KFA	39.607	<0.001	0.691	0.501	1.12	0.348

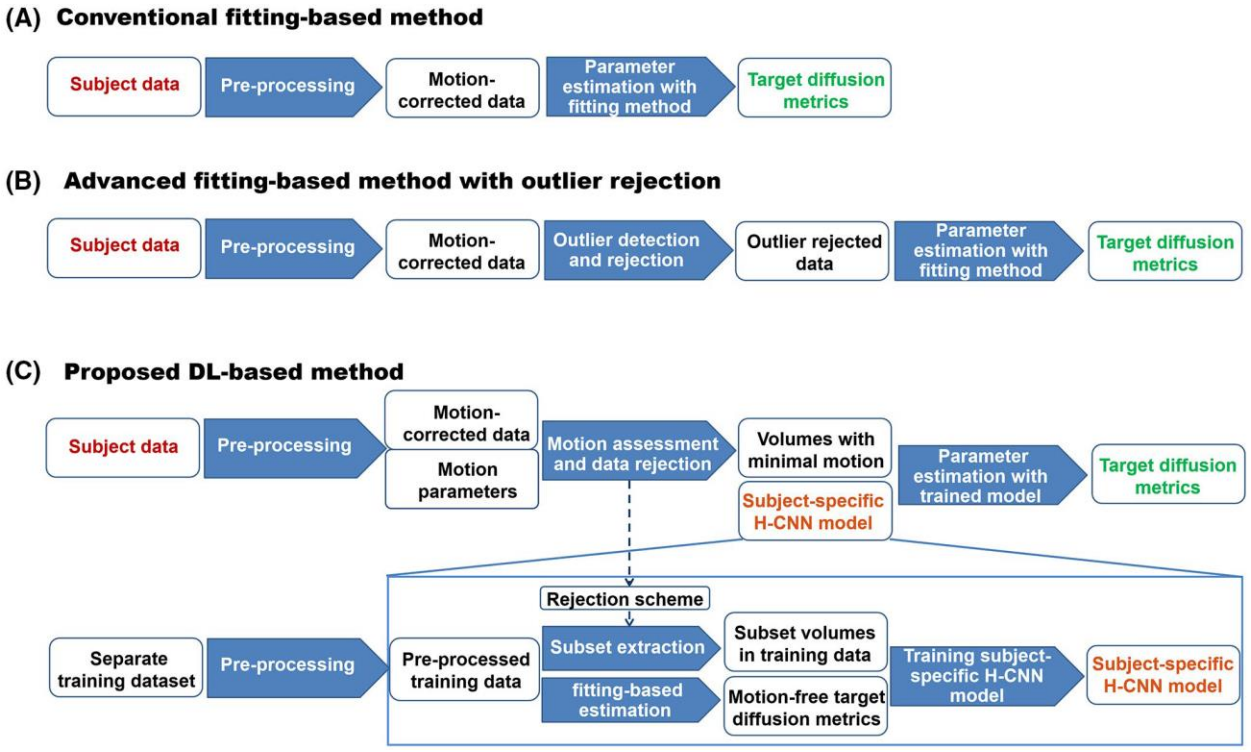

Figure 1. The pipeline of the proposed DL-based technique (C) in comparison to fitting-based methods (A-B). The DL-based method first corrects distortion and motion as fitting-based methods. The motion parameters estimated during pre-processing are used to calculate volume-based motion assessment measures. Stringent thresholds of motion measures are applied to reject the volumes deemed to have moved excessively. The b-values and the diffusion gradient vectors associated with the rejection scheme are used to extract desired subset from the separate training dataset, and a subject-specific network model is then trained using the selected subset data and derived metrics from the full data with model fitting. Finally, the trained model is applied to the remaining data of the study subject to compute its diffusion metrics.

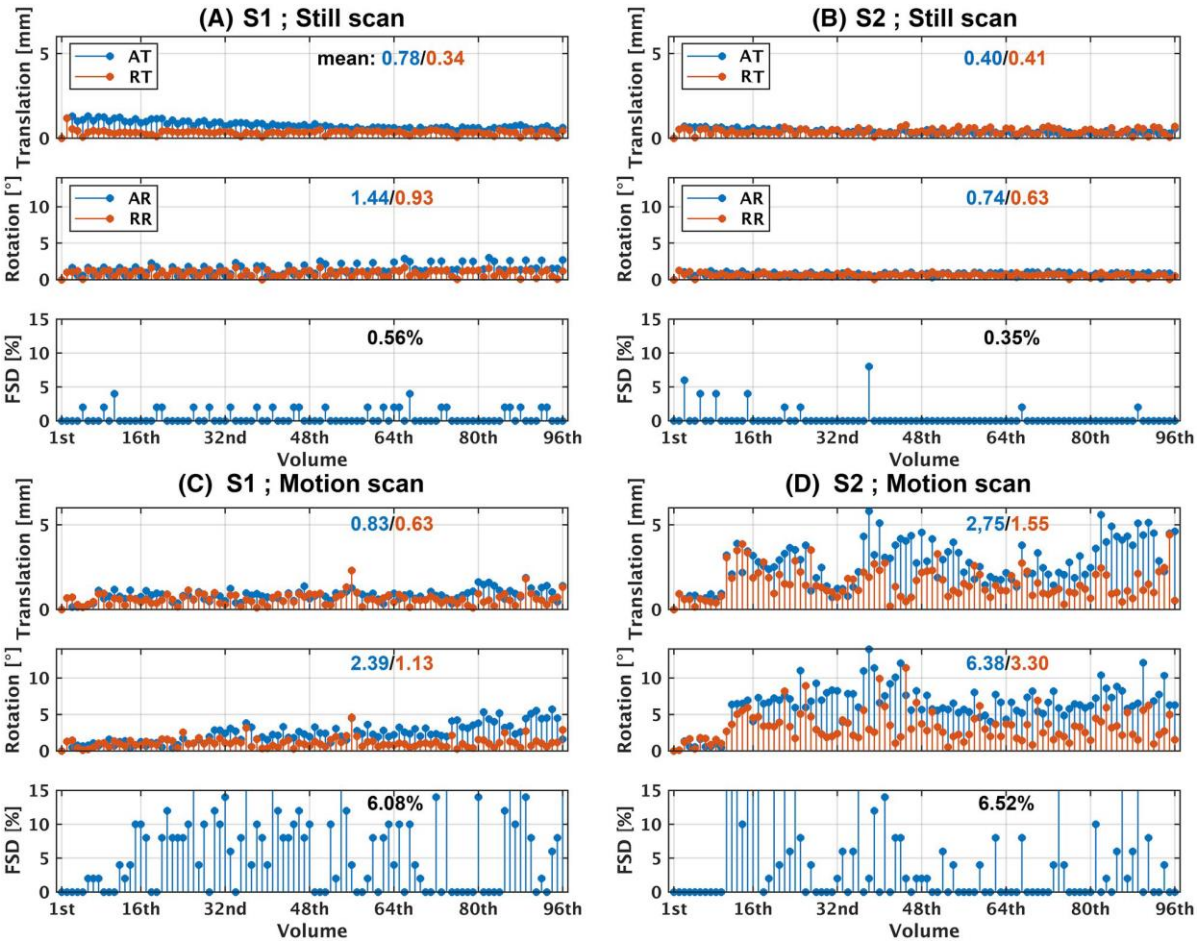

Figure 2. Demonstration of the motion measures (AT, RT, AR, RR and FSD) of the 96 volumes from two test subjects in still (A)(B) and motion scans (C) (D), respectively. The mean measures across all volumes are also depicted in the figure.

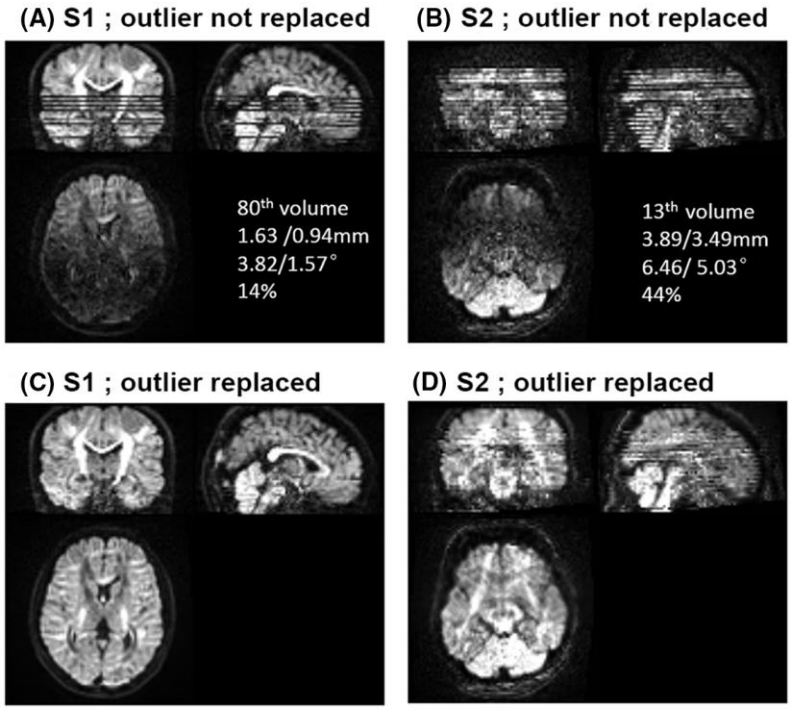

Figure 3. Typical volumes with large relative motion measures after pre-processing without (top) and with (bottom) outlier replacement. (A)(C) the 80th volume from S1, and (B)(D) the 13th volume from S2 are shown. The motion measures are shown in (A) and (B) in the order of AT/RT in mm, AR/RR in degrees, and FSD in percentage.

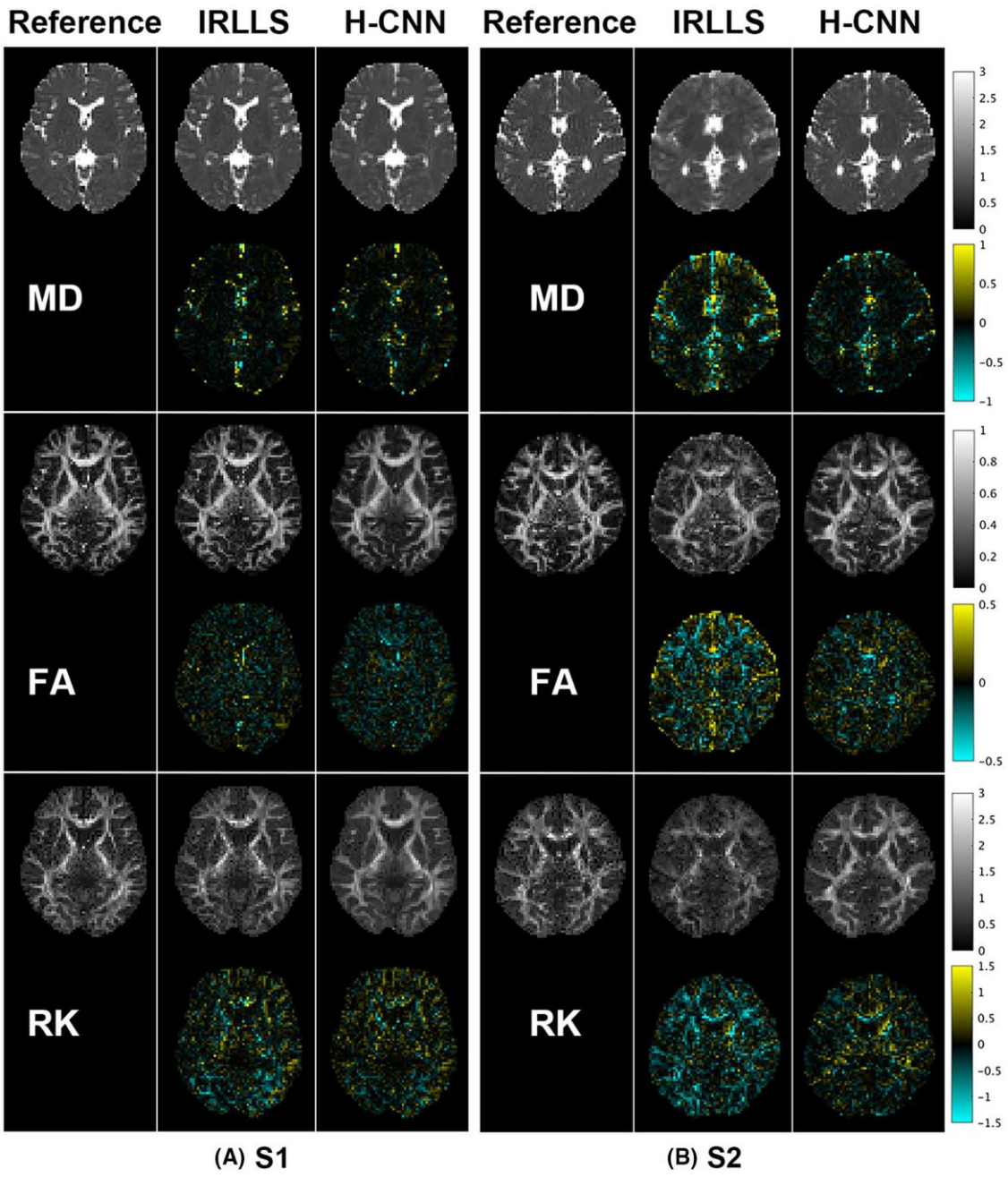

Figure 4. Estimated MD, FA, and RK maps from S1 (A) and S2 (B). In (A) and (B), three planes are shown for the IRLLS estimated reference maps from data in the still condition (first column), IRLLS (second column) and H-CNN (third column) estimated maps from data in the motion condition; their differences to reference maps are shown below the maps. For the proposed DL-based pipeline, there are only 29 volumes and 10 volumes left from a total of 96 volumes respectively for S1 and S2 for parameter estimation after motion rejection.

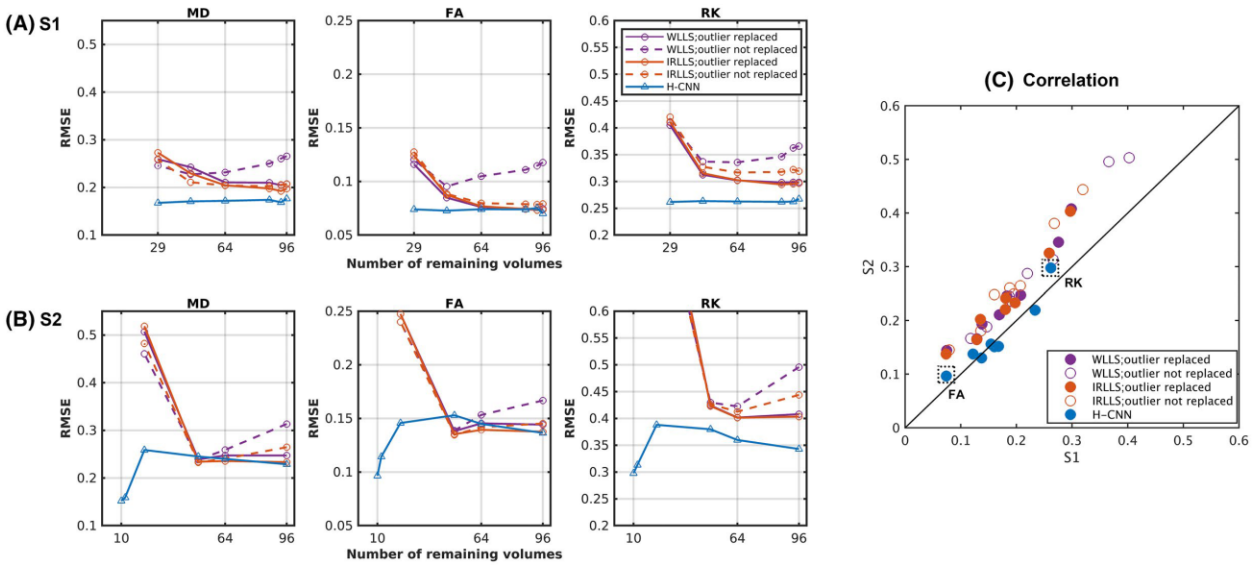

Figure 5. The WM RMSEs as a function of the remaining number of volumes selected by different motion thresholds as listed in Table 1 from S1 (A) and S2 (B) (measures of MD, FA and RK are shown). As demonstrated by the correlation of all eight measures from the two subjects (C), RMSEs from H-CNN for S1 (N=29) and S2 (N=10) were similar with different level of motion for most measures, while RMSEs from IRLLS and WLLS with full data (N=96) show motion-level dependency (S2>S1).

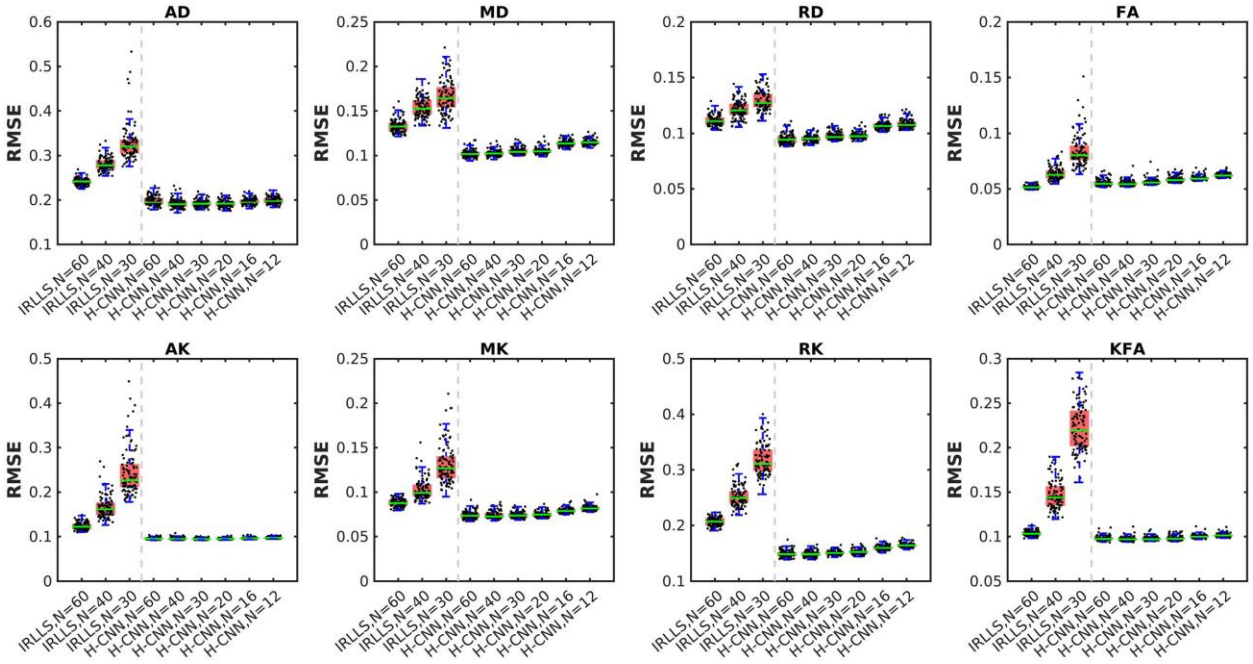

Figure 6. Boxplots of WM RMSEs from random rejection tests. Each red box represents the RMSE of the interquartile range (IQR) from 100 random sub-sampled schemes for each retained volume number N, and

the whiskers indicate the highest and lowest values within 1.5 IQR of the nearer quartile. The detailed RMSEs are plotted as black dots with their median values shown by green lines.

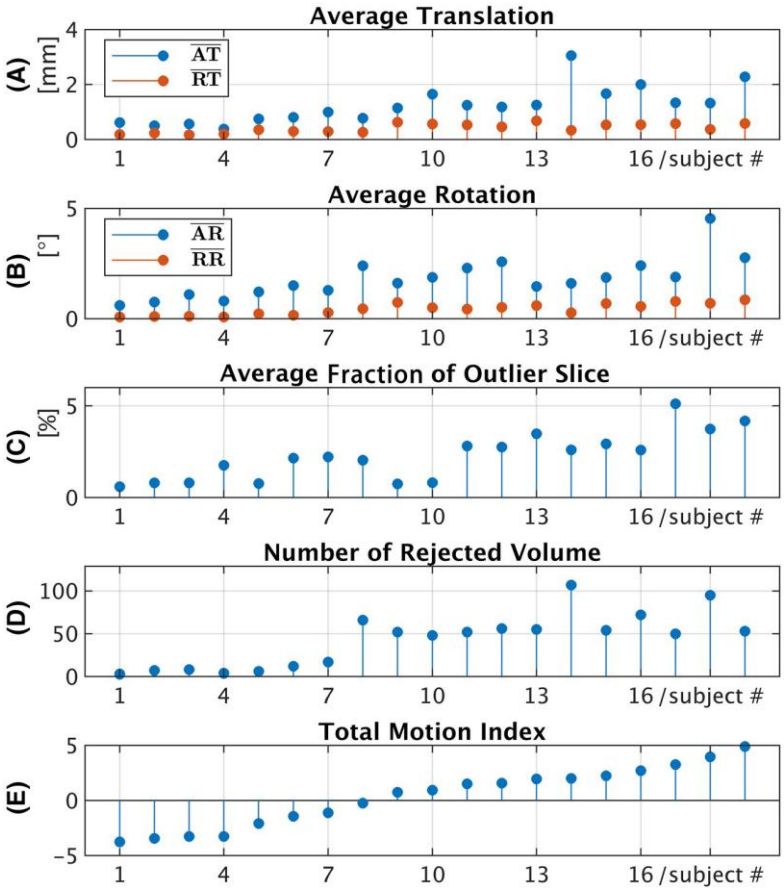

Figure 7. Motion assessment results for each subject. (A)–(C) The averaged five motion measures across all volumes. (D) Number of rejected volumes for H-CNN estimation (E) The TMI calculated taking the five motion measures into consideration. The 8 subjects with TMI < 0 are divided into a small motion control group and the other 11 subjects are divided into a large motion group.

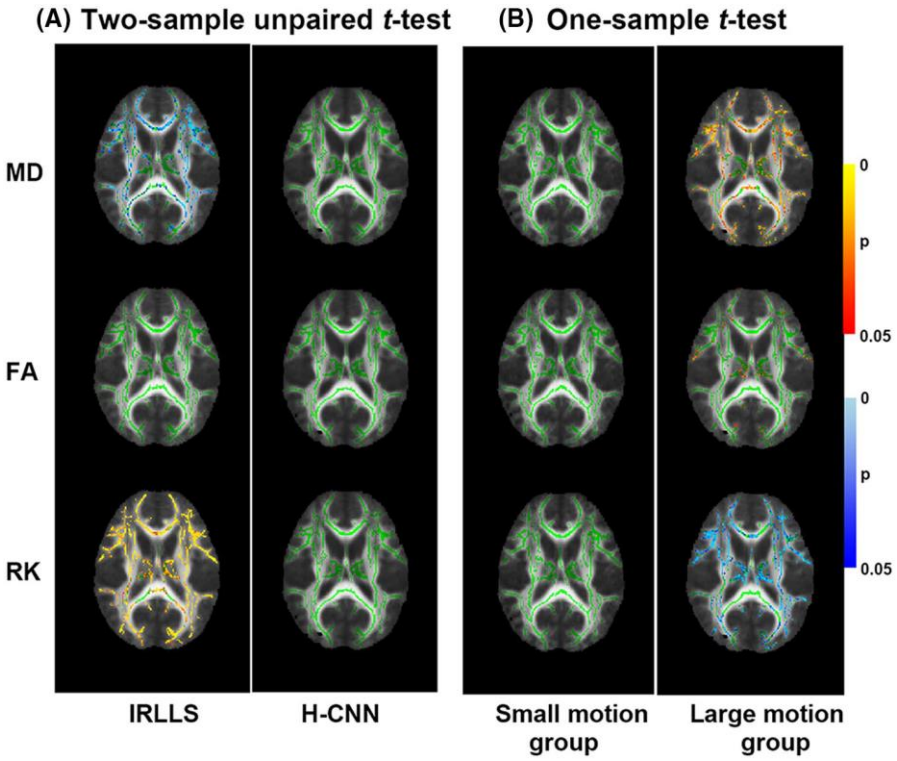

Figure 8. Voxel-wise group statistics of MD, FA, and RK measures on the FA skeleton. (Green: FA skeleton; Red-Yellow: significantly higher; Blue-Light Blue: significantly lower) (A) Two-sample *t*-tests between subjects from the small motion control group and large motion group using IRLLS estimation (left) and H-CNN estimation (right). (B) One-sample *t*-tests of the difference maps between IRLLS and H-CNN estimations from the small motion control group (left) and large motion group (right).