Towards Accurate Microstructure Estimation via 3D Hybrid Graph Transformer

Junqing Yang[1]([⊠]), Haotian Jiang[2]([⊠]), Tewodros Tassew[1], Peng Sun[1], Jiquan Ma[2]([⊠]), Yong Xia[1], Pew-Thian Yap[3,4], and Geng Chen[1]([⊠])

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
`geng.chen@nwpu.edu.cn`

[2] School of Computer Science and Technology, Heilongjiang University, Harbin, China
`majiquan@hlju.edu.cn`

[3] Department of Radiology, University of North Carolina, Chapel Hill, NC, USA

[4] Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC, USA

Abstract. Deep learning has drawn increasing attention in microstructure estimation with undersampled diffusion MRI (dMRI) data. A representative method is the hybrid graph transformer (HGT), which achieves promising performance by integrating q-space graph learning and x-space transformer learning into a unified framework. However, this method overlooks the 3D spatial information as it relies on training with 2D slices. To address this limitation, we propose 3D hybrid graph transformer (3D-HGT), an advanced microstructure estimation model capable of making full use of 3D spatial information and angular information. To tackle the large computation burden associated with 3D x-space learning, we propose an efficient q-space learning model based on simplified graph neural networks. Furthermore, we propose a 3D x-space learning module based on the transformer. Extensive experiments on data from the human connectome project show that our 3D-HGT outperforms state-of-the-art methods, including HGT, in both quantitative and qualitative evaluations.

Keywords: Microstructure Imaging · Graph Neural Network · Transformer · 3D Spatial Domain

1   Introduction

Diffusion microstructure imaging has drawn increasing research attention in recent years. A number of powerful microstructure models have been proposed and shown great success in both clinical and research sides. Typical examples include diffusion kurtosis imaging (DKI) [15], neurite orientation dispersion and density imaging (NODDI) [23], and spherical mean technique (SMT) [16]. However, these models usually rely on diffusion MRI (dMRI) data densely sampled in q-space with a sufficient angular resolution and multiple shells, which are impractical in clinical settings.

To resolve this problem, deep learning has been introduced to learn the mapping between high-quality microstructure indices and q-space undersampled dMRI data. For instance, Golkov et al. [12], for the first time, introduced deep learning to microstructure estimation from undersampled dMRI data with a multilayer perceptron (MLP). Gibbons et al. [11] trained a 2D CNN to generate NODDI and fractional anisotropy parameter maps from undersampled dMRI data. Tian et al. [18] proposed DeepDTI to predict high-fidelity diffusion tensor imaging metrics using a 10-layer 3D CNN. Ye et al. designed MEDN [21] and MEDN+ [22] based on a dictionary-inspired framework to learn the mapping between undersampled dMRI data and high-quality microstructure indices. Zheng et al. [24] proposed a three-stage microstructure estimation model that combines sparse encoding with a transformer. Chen et al. [4] introduced the graph convolutional neural network (GCNN) to learn q-space information for microstructure estimation, which considers the angular information in q-space and achieves promising performance.

Recently, inspired by the fact that dMRI data live in a joint x-q space [3,14], a powerful hybrid model, called HGT [5], was proposed to jointly learn spatial-angular information for accurate microstructure estimation. It achieves superior performance in comparison with various types of existing methods. However, a major limitation of HGT lies in the ignorance of the fact that the spatial domain of dMRI is a 3D space rather than a 2D one. A large number of studies have demonstrated that careful consideration of 3D space is the key to advancing the performance of different medical image analysis tasks [7,9].

To this end, we propose 3D-HGT, an advanced microstructure estimation model capable of making full use of 3D x-space information and q-space information jointly. Specifically, we design an efficient q-space learning module based on simple graph convolution (SGC) [20], which runs with less memory at high speed and is able to alleviate the large computational burden associated with 3D spatial information learning. We further propose a 3D x-space learning module based on a U-shape transformer, which is able to capture the long-range relationships in 3D spatial space for improved performance. Finally, we train two modules end-to-end for predicting microstructure index maps. Our 3D-HGT exploits 3D spatial information and angular information in an efficient manner for more accurate microstructure estimation. We perform extensive experiments on data from the human connectome project (HCP) [19]. The experimental results demon-

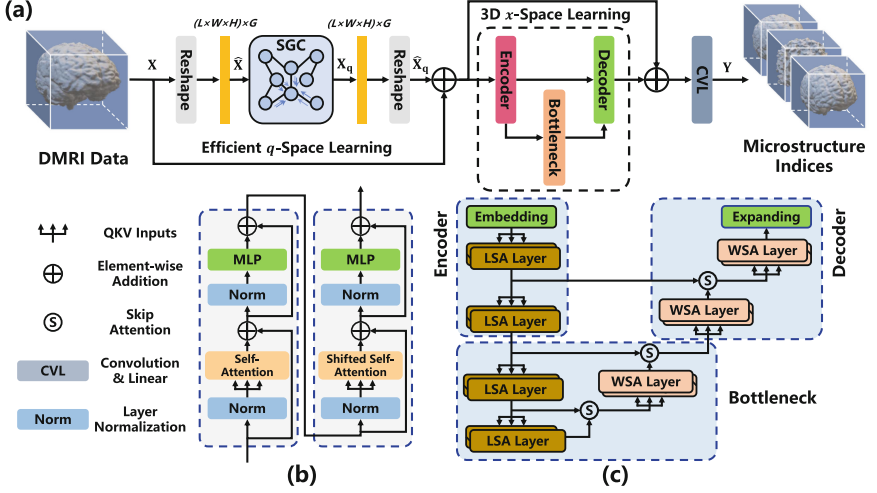strate that our 3D-HGT outperforms cutting-edge methods, including HGT, both quantitatively and qualitatively.

Fig. 1. Overview of 3D-HGT. (a) shows the overall architecture of the model, where the q-space learning module first learns the feature of dMRI data in q-space, and then the x-space learning module learns the 3D spatial information; (b) shows the architecture of the LSA/WSA component, which is a cascaded transformer group with two self-attention blocks; (c) shows the structure of the x-space learning module.

2   3D Hybrid Graph Transformer

2.1   Network Overview

As shown in Fig. 1, our 3D-HGT consists of two key modules: an efficient q-space learning module and a 3D x-space learning module. The q-space learning module is built with SGC, which can extract q-space features from voxel-wise dMRI data more efficiently. Motivated by the nnFormer [25], the x-space learning module mainly consists of a U-shaped network composed of local volume-based multi-head self-attention (LSA) and wide volume-based multi-head self-attention (WSA) components.

The input dMRI data $\mathbf{X} \in \mathbb{R}^{L \times W \times H \times G}$ has four dimensions, where L, W, H, and G denote the length, width, height, and number of gradient directions, respectively. In our model, the data is first reshaped into a two-dimensional tensor $\hat{\mathbf{X}} \in \mathbb{R}^{LWH \times G}$ and fed into the SGC network to learn the q-space features efficiently. The output $\mathbf{X}_{\mathrm{q}}$ is then reshaped to four dimensions $\hat{\mathbf{X}}_{\mathrm{q}} \in \mathbb{R}^{L \times W \times H \times G}$ and enters into the x-space learning module followed by convolutional layers. Finally, our 3D-HGT generates microstructure predictions $\mathbf{Y} \in \mathbb{R}^{L \times W \times H \times M}$, where M denotes the number of microstructure indices.

2.2   Efficient *q*-Space Learning Module

According to [2,6], the geometric structure of the *q*-space of dMRI data can be encoded as a graph $\mathcal{G}$ determined by a binary affinity matrix $\mathbf{A} = \{a_{i,j}\}$. If the angle $\theta_{i,j}$ between two sampling points i and j in *q*-space is less than an angle threshold θ, then $a_{i,j} = 1$; otherwise, $a_{i,j} = 0$. Following such processing, GCNN can be used for *q*-space learning.

Different from HGT, 3D-HGT learns in *q*-space with SGC, which removes the non-linearity between the graph convolutional layers and collapses K graph convolutional layers into one. In this way, the complex computation of a multi-layer network is reduced to a few matrix multiplications. Thus, we can use fast matrix multiplication to speed up computation, reduce network redundancy, and improve computational efficiency. Mathematically, the feature provided by our *q*-space learning module is as follows:

$$\mathbf{X}_{\mathrm{q}} = \hat{\mathbf{A}}^K \hat{\mathbf{X}} \mathbf{\Theta}, \tag{1}$$

where K denotes the number of SGC layers, $\hat{\mathbf{A}}$ is the normalized version of $\mathbf{A}$ with self-loop added, and $\mathbf{\Theta}$ is the product of the weight matrices learned by the *q*-space learning module.

2.3   3D *x*-Space Learning Module

The overall design of the *x*-space learning module is a U-shaped network composed of three parts: an encoder, a bottleneck, and a decoder. The three components are made up of cascaded self-attention transformer blocks.

The encoder and decoder are similar in structure, with two self-attention layers. However, there are two key differences between them. Firstly, the former is composed of two local volume-based layers, while the latter has two wide volume-based layers. Secondly, before entering the network, data from the encoder passes through an embedding layer, while data from decoders pass through an extension layer for feature integration before outputting results. The bottleneck layer comprises two LSAs and one WSA, which allows skip attention to integrate shallow and deep attention between the encoder and decoder layers. Notably, we omit the pooling or un-pooling operations in the encoder-decoder architecture since pooling can lead to the loss of features, especially when dealing with small-sized 3D patches.

The Embedding Layer. The embedding layer divides the input $\hat{\mathbf{X}}_{\mathrm{q}}$ into high-dimensional patches, which are then sent into the transformer block. Unlike the nnFormer, our embedding layer consists of two convolutional layers, a GELU [13] layer and a normalization layer [1], where both the convolutional kernel and step size are set to one.

Local Volume-Based Multi-head Self-attention (LSA). Based on Swin Transformer [17], LSA is an offset window self-retaining module that defines the

window as a local 3D volume block. The multi-head self-attention is computed in this local volume, which reduces the computational cost significantly.

Let $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{S_{\mathrm{L}} \times S_{\mathrm{W}} \times S_{\mathrm{H}} \times D}$ for LSA be the query, key, and value matrices, and $\mathbf{B} \in \mathbb{R}^{S_{\mathrm{L}} \times S_{\mathrm{W}} \times S_{\mathrm{H}}}$ be the relative position matrix, where $\{S_{\mathrm{L}}, S_{\mathrm{W}}, S_{\mathrm{H}}\}$ denotes the size of the local volume and D is the dimension of query/key. The self-attention is then computed in each 3D local volume as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}} + \mathbf{B})\mathbf{V}. \tag{2}$$

The LSA computational complexity can be expressed as follows:

$$\Omega(\text{ LSA }) = 4N_{\mathrm{P}}C^2 + 2S_{\mathrm{L}}S_{\mathrm{W}}S_{\mathrm{H}}N_{\mathrm{P}}C, \tag{3}$$

where N_{P} is the size of a patch, and C is the length of the embedding sequence.

Wide Volume-Based Multi-head Self-attention (WSA). Although LSA is efficient, its receptive field is limited. Thus, by increasing the size of the window to $\{m \times S_{\mathrm{L}}, m \times S_{\mathrm{W}}, m \times S_{\mathrm{H}}\}$ with m denoting a constant (two by default), we have WSA, which improves the global context awareness ability of LSA. The computational complexity of WSA is as follows:

$$\Omega(\text{ WSA }) = 4N_{\mathrm{P}}C^2 + 2m^3 S_{\mathrm{L}}S_{\mathrm{W}}S_{\mathrm{H}}N_{\mathrm{P}}C. \tag{4}$$

We extend the self-attention field of view in the bottleneck using three wide-field transformer blocks, where six WSA layers are used.

Skip Attention. This component effectively combines shallow and deep attention. A single-layer neural network decomposes the output $\mathbf{X}_l$ at layer l of the encoder into a key matrix $\mathbf{K}_{l'}$ and a value matrix $\mathbf{V}_{l'}$, while the output of $\mathbf{X}_{l'}$ at layer l' of the decoder is used as a query matrix $\mathbf{Q}_{l'}$. Mathematically, the self-attention is defined as:

$$\text{Attention}(\mathbf{Q}_{l'}, \mathbf{K}_{l'}, \mathbf{V}_{l'}) = \text{Softmax}(\frac{\mathbf{Q}_{l'}\mathbf{K}_{l'}^{\mathrm{T}}}{\sqrt{D_{l'}}} + \mathbf{B}_{l'})\mathbf{V}_{l'}, \tag{5}$$

where $\mathbf{B}_{l'}$ is the relative position encoding matrix, and $D_{l'}$ is the dimension of query/key.

3   Experiments

3.1   Implementation Details

Experimental Settings. We use the mean square error as the loss function, Adam as the optimizer, and an initial learning rate of 9e-4. We set the number of epochs to 100, batch size to 2, and iterations to 10. The angular threshold θ for constructing graph is set at $45°$, and the number of graph convolutional layers (i.e., K) is set to 2. In the x-space learning module, the embedding dimension

is set to 192, and the window sizes of the LSA and WSA layers are set to 4 and 8, respectively. The model was implemented using PyTorch 1.11 and PyTorch-Geometric 2.1.0, and trained on a server equipped with an RTX 3090 GPU.

Comparison Methods. We compare our 3D-HGT with various methods, including AMICO [8], MLP [12], GCNN [4], MEDN [21], MEDN+ [22], U-Net [10], U-Net++ [26], and HGT [5].

Table 1. Performance evaluation of microstructre estimation using single-shell undersampled data with b=1000 s/mm^2, 30 gradient directions in total. All: A combination of ICVF, ISOVF, and ODI.

Method	PSNR ↑				SSIM ↑			
	ICVF	ISOVF	ODI	All	ICVF	ISOVF	ODI	All
AMICO [8]	8.09	14.90	12.03	10.80	0.020	0.273	0.380	0.210
MLP [12]	19.15	24.07	20.31	20.72	0.670	0.706	0.667	0.681
MEDN [21]	19.18	24.43	20.54	20.87	0.674	0.737	0.684	0.698
GCNN [4]	19.27	24.48	20.55	20.92	0.676	0.742	0.681	0.700
MEDN+ [22]	19.14	24.67	21.30	21.14	0.684	0.754	0.717	0.719
U-Net [10]	19.95	25.11	20.41	21.28	0.727	0.787	0.651	0.722
U-Net++ [26]	19.94	25.49	20.87	21.51	0.750	0.798	0.688	0.745
HGT [5]	21.96	27.13	22.12	23.16	0.814	**0.850**	0.757	0.807
3D-HGT	**23.60**	**27.98**	**22.43**	**24.10**	**0.837**	0.845	**0.765**	**0.816**

3.2   Dataset and Evaluation Metrics

Dataset. Following [5], we randomly select 21 subjects from the HCP to construct our dataset. For each dMRI scan, q-space undersampling is performed on a single shell with b=1000 s/mm^2 to extract 30 gradients uniformly. The undersampled data is then normalized by dividing by the average of b_0 images. Finally, we extract patches from the normalized data with a dimension of $32 \times 32 \times 32 \times 30$. The validation and test sets are also processed in the same way. In our experiments, the ratio between training, validation, and test sets is 10:1:10. We train our model to predict NODDI-derived indices, including intracellular volume fraction (ICVF), isotropic volume fraction (ISOVF), and orientation dispersion index (ODI). The gold standard microstructural indices are computed using the complete HCP data with 270 gradients using AMICO [8].

Evaluation Metrics. We evaluate the quality of predicted NODDI-derived indices with the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM).
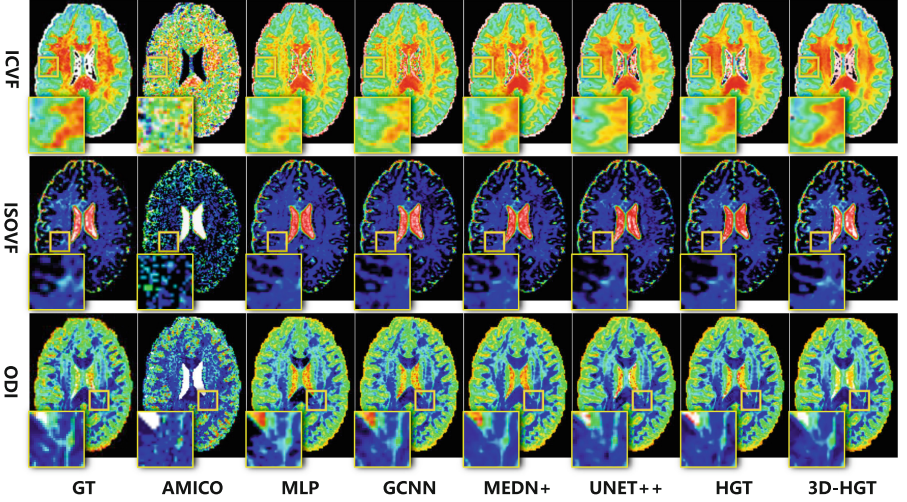
Fig. 2. Visual comparison of prediction results.

3.3   Experimental Results

Table 1 compares our 3D-HGT model with cutting-edge models in control experiments under the same conditions using various evaluation metrics. By taking into account the 3D x-space data from the dMRI data, our model is able to learn rich features across a broader range of spatial domains. 3D-HGT outperforms HGT in terms of PSNR and SSIM. The most notable improvements among all indices can be observed in the PSNR of the ICVF, which has increased from 21.96 to 23.60 dB, and the SSIM of the ICVF, which has increased from 0.814 to 0.837. Furthermore, compared with the deep learning models (i.e., MLP to HGT) in Table 1, 3D-HGT improves the PSNR of ALL by 16.3%, 15.5%, 15.2%, 14.0%, 13.3%, 12.0%, and 4.1%, respectively. Although HGT takes into account x-space and q-space, it only makes use of the 2D spatial features of the data. In contrast, 3D-HGT explicitly considers the 3D spatial domain and improves the PSNR and SSIM of ALL by 4.1% and 1.1%, respectively.

Figure 2 depicts the visual comparison of 3D-HGT and other models' microstructure predictions. In particular, the close-up views, shown in the bottom row of Fig. 2, indicate that our model provides the best result, which is much closer to the ground truth.

3.4   Ablation Study

To investigate the effectiveness of the proposed modules, we perform ablation experiments with different ablated versions. The ablation results are shown in Table 2. It should be noted that the ablated versions "(A)" and "(E)" correspond to HGT and the full version of 3D-HGT, respectively.

Table 2. Ablation study for 3D-HGT. "2D Trans." denotes the 2D x-space learning module of HGT. "3D Trans." denotes the 3D x-space learning module of 3D-HGT, as shown in Fig. 1(c). The metric "Time" indicates the time cost (in seconds) for training the corresponding ablation version in one epoch.

Model	q-Space Module		x-Space Module		PSNR ↑				Time ↓
	TAGCN	SGC	2D Trans.	3D Trans.	ICVF	ISOVF	ODI	All	
(A)	✓		✓		21.96	27.13	22.12	23.16	121
(B)		✓	✓		21.97	27.21	22.33	23.28	**109**
(C)				✓	23.38	27.78	22.40	23.97	122
(D)	✓			✓	23.44	27.80	22.35	23.98	226
(E)		✓		✓	**23.60**	**27.98**	**22.43**	**24.10**	197

Effectiveness of Efficient q-Space Learning Module. As shown in Table 2, "(C)" is the ablated version with only the 3D x-space learning module. It can be observed that "(E)" outperforms "(C)", demonstrating that our q-space learning module can effectively improve performance. Moreover, we perform additional experiments using the undersampled dMRI data with a higher angular resolution, i.e., 45 gradients. The results show that our q-space learning module provides a larger improvement when the dMRI data is with a higher angular resolution.

The q-space learning with SGC also shows advantages over TAGCN. Both "(B)" and "(E)" improve the performance and computational efficiency in comparison with their corresponding ablated versions equipped with TAGCN, i.e., "(A)" and "(D)". In particular, compared with "(D)", the training time cost of "(E)" is reduced by nearly 30 s for one epoch, verifying the high efficiency of our q-space learning module.

Effectiveness of 3D x-Space Learning Module. As shown in Table 2, two sets of comparisons, "(A)" vs. "(D)" and "(B)" vs. "(E)", indicate that modifying the x-space learning module to a 3D transformer improves performance. More specifically, compared with "(A)", "(D)" improves the PSNR of ICVF by 1.48 dB, ISOVF by 0.67 dB, ODI by 0.23 dB and ALL by 0.82 dB. Compared with "(B)", "(E)" improves the PSNR of ICVF by 1.63 dB, ISOVF by 0.77 dB, ODI by 0.10 dB and ALL by 0.82 dB. The improvement owes to the 3D x-space learning module, which is equipped with LSAs and WSAs, allowing the model to capture long-term dependencies with a large 3D receptive field.

4   Conclusion

In this paper, we proposed 3D-HGT, an improved microstructure estimation model that makes full use of 3D x-space information and q-space information. Our x-space learning is achieved with a 3D transformer architecture, allowing the model to thoroughly learn the long-term dependencies of features in the 3D

spatial domain. To alleviate the large computational burden associated with 3D x-space learning, we further propose an efficient q-space learning module, which is built with a simplified graph learning architecture. Extensive experiments on the HCP demonstrate that, compared with HGT, 3D-HGT effectively improves the quality of microstructure estimation.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Chen, G., Dong, B., Zhang, Y., Lin, W., Shen, D., Yap, P.T.: Denoising of diffusion MRI data via graph framelet matching in x-q space. IEEE Trans. Med. Imaging **38**(12), 2838–2848 (2019)
3. Chen, G., Dong, B., Zhang, Y., Lin, W., Shen, D., Yap, P.T.: XQ-SR: joint x-q space super-resolution with application to infant diffusion MRI. Med. Image Anal. **57**, 44–55 (2019)
4. Chen, G., et al.: Estimating tissue microstructure with undersampled diffusion data via graph convolutional neural networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12267, pp. 280–290. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_28
5. Chen, G., et al.: Hybrid graph transformer for tissue microstructure estimation with undersampled diffusion MRI data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS, vol. 13431, pp. 113–122. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_11
6. Chen, G., Wu, Y., Shen, D., Yap, P.T.: Noise reduction in diffusion MRI using non-local self-similar information in joint x-q space. Med. Image Anal. **53**, 79–94 (2019)
7. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage **170**, 446–455 (2018)
8. Daducci, A., et al.: Accelerated microstructure imaging via convex optimization (AMICO) from diffusion MRI data. NeuroImage **105**, 32–44 (2015)
9. Dou, Q., et al.: Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Trans. Med. Imaging **35**(5), 1182–1195 (2016)
10. Falk, T., et al.: U-net: deep learning for cell counting, detection, and morphometry. Nat. Methods **16**(1), 67–70 (2019)
11. Gibbons, E.K., et al.: Simultaneous NODDI and GFA parameter map generation from subsampled q-space imaging using deep learning. Magnet. Resonan. Med. **81**(4), 2399–2411 (2019)
12. Golkov, V., et al.: q-Space deep learning: twelve-fold shorter and model-free diffusion MRI scans. IEEE Trans. Med. Imaging **35**(5), 1344–1351 (2016)
13. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)
14. Hong, Y., Chen, G., Yap, P.-T., Shen, D.: Multifold acceleration of diffusion MRI via deep learning reconstruction from slice-undersampled data. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 530–541. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_41

15. Jensen, J.H., Helpern, J.A., Ramani, A., Lu, H., Kaczynski, K.: Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magnet. Resonan. Med. **53**(6), 1432–1440 (2005)

16. Kaden, E., Kruggel, F., Alexander, D.C.: Quantitative mapping of the per-axon diffusion coefficients in brain white matter. Magnet. Resonan. Med. **75**(4), 1752–1763 (2016)

17. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

18. Tian, Q., et al.: DeepDTI: high-fidelity six-direction diffusion tensor imaging using deep learning. NeuroImage **219**, 117017 (2020)

19. Van Essen, D.C., et al.: The WU-Minn human connectome project: an overview. NeuroImage **80**, 62–79 (2013)

20. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., Weinberger, K.: Simplifying graph convolutional networks. In: International Conference on Machine Learning, pp. 6861–6871. PMLR (2019)

21. Ye, C.: Estimation of tissue microstructure using a deep network inspired by a sparse reconstruction framework. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 466–477. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_37

22. Ye, C.: Tissue microstructure estimation using a deep network inspired by a dictionary-based framework. Med. Image Anal. **42**, 288–299 (2017)

23. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.: NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. NeuroImage **61**(4), 1000–1016 (2012)

24. Zheng, T., et al.: A microstructure estimation transformer inspired by sparse representation for diffusion MRI. Med. Image Anal. **86**, 102788 (2023)

25. Zhou, H.Y., et al.: nnFormer: volumetric medical image segmentation via a 3D transformer. IEEE Trans. Image Process. (2023)

26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1