



A microstructure estimation Transformer inspired by sparse representation for diffusion MRI

Tianshu Zheng^a, Cong Sun^b, Weihao Zheng^c, Wen Shi^a, Haotian Li^a, Yi Sun^d, Yi Zhang^a, Guangbin Wang^b, Chuyang Ye^e, Dan Wu^{a,*}

^aDepartment of Biomedical Engineering, College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, Zhejiang, China

^bDepartment of Radiology, Shandong Medical Imaging Research Institute, Cheeloo College of Medicine, Shandong University, Jinan, China

^cSchool of Information Science and Engineering, Lanzhou University, Lanzhou, China

^dMR Collaboration, Siemens Healthineers Ltd., Shanghai, China

^eSchool of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China

ARTICLE INFO

Article history:

Received 1 May 2013

Received in final form 10 May 2013

Accepted 13 May 2013

Available online 15 May 2013

Communicated by S. Sarkar

Keywords: Diffusion MRI, Microstructural model, Sparse coding, Transformer

ABSTRACT

Diffusion magnetic resonance imaging (dMRI) is an important tool in characterizing tissue microstructure based on biophysical models, which are typically multi-compartmental models with mathematically complex and highly non-linear forms. Resolving microstructures from these models with conventional optimization techniques is prone to estimation errors and requires dense sampling in the q-space with a long scan time. Deep learning based approaches have been proposed to overcome these limitations in dMRI-based microstructure estimation. Motivated by the superior performance of the Transformer in feature extraction than the convolutional structure, in this work, we present a learning-based framework based on Transformer, namely, a *Microstructure Estimation Transformer with Sparse Coding* (METSC) for dMRI-based microstructural parameter estimation with downsampled q-space data. To take advantage of the Transformer while addressing its limitation in large training data requirement, we explicitly introduce an inductive bias—model bias into the Transformer using a sparse coding technique to facilitate the training process. Thus, the METSC is composed with three stages, an embedding stage, a sparse representation stage, and a mapping stage. The embedding stage is a Transformer-based structure that encodes the signal in a high-level space to ensure the core voxel of a patch is represented effectively. In the sparse representation stage, a dictionary is constructed by solving a sparse reconstruction problem that unfolds the *Iterative Hard Thresholding* (IHT) process. The mapping stage is essentially a decoder that computes the microstructural parameters from the output of the second stage, based on the weighted sum of normalized dictionary coefficients where the weights are also learned. We tested our framework on two dMRI models with downsampled q-space data, including the *intra-voxel incoherent motion* (IVIM) model and the *neurite orientation dispersion and density imaging* (NODDI) model. The proposed method achieved up to 11.25 folds of acceleration in scan time while retaining high fitting accuracy, reducing the *mean square error* (MSE) by up to 70% compared with q-space learning. METSC outperformed the other state-of-the-art learning-based methods, including the model-free and model-based methods, and reduced the MSE by most 81%. The network also showed robustness against the noise. The superior performance of METSC indicates its potential to improve dMRI acquisition and model fitting in clinical applications.

© 2022 Elsevier B. V. All rights reserved.

1. Introduction

Diffusion MRI (dMRI) is one of the most important medical imaging tools and the only noninvasively approach that can probe tissue microstructures based on the restricted diffusion of water molecules in biological tissues (Mori and Zhang, 2006). The commonly used diffusion tensor model has shown to be sensitive to pathological changes such as stroke and tumor (Le Bihan *et al.*, 2001), but it is not specific to microstructural properties, such as cell size, axonal diameter, fiber density and orientational dispersion. Advanced dMRI models are developed to characterize specific microstructural features (Novikov *et al.*, 2019), such as *intra-voxel incoherent motion* (IVIM) (Le Bihan *et al.*, 1988), *AxCaliber* (Assaf *et al.*, 2008), *diffusion basis spectrum imaging* (DBSI) (Wang *et al.*, 2011), *neurite orientation dispersion and density imaging* (NODDI) (Zhang *et al.*, 2012), *soma and neurite density imaging* (SANDI) (Palombo *et al.*, 2020), and *imaging microstructural parameters using limited spectrally edited diffusion* (IMPULSED) (Jiang *et al.*, 2016), just to name a few. The majority of the advanced dMRI models consist of multiple compartments with mathematically complex and highly non-linear signal representations. Fitting of these models with conventional nonlinear optimization techniques, such as nonlinear least square fitting (Arun *et al.*, 1987), is prone to estimation errors. Moreover, from the data acquisition perspective, advanced dMRI models require the acquisition of multiple b-values and diffusion directions in the q-space, which is time-consuming and vulnerable to motion artifacts. This is particularly a problem for moving subjects, such as abdominal organs, fetuses, and placentas.

To reduce the estimation error and accelerate the acquisition for advanced dMRI models, many methods have been proposed. Nedjati-Gilani *et al.* (Nedjati-Gilani *et al.*, 2014) and Alexander *et al.* (Alexander *et al.*, 2014, 2017) proposed a random forest method to estimate the microstructural parameters in the Kärger model (Kärger *et al.*, 1988), NODDI model, and *spherical mean technique* (SMT) model (Kaden *et al.*,

2016), respectively. The development of deep learning techniques opens a new avenue for dMRI model fitting. The concept of q-space deep learning (Golkov *et al.*, 2016) is first proposed to directly map the dMRI signals to the DKI parameters using a subset of the q-space data (reduced number of b-value and diffusion directions). The original q-space deep learning (abbreviated as q-DL) only used the three-layer multilayer perceptron (MLP) (Golkov *et al.*, 2016). Gibbons *et al.* used a 2D convolution neural network to estimate the NODDI and generalized fractional anisotropy maps simultaneously (Gibbons *et al.*, 2019). Koppers *et al.* used a residual network to increase the comparability of dMRI signals measured on two different scanners (Koppers *et al.*, 2019). Chen *et al.* used a subset q-space to estimate the NODDI parameters via graph convolutional neural network (Chen *et al.*, 2020). Barbieri *et al.* used three-layer MLP to estimate the IVIM model parameters (Barbieri *et al.*, 2020). Beyond the end-to-end mapping approaches, the model-driven neural networks that introduce domain knowledge into a deep neural network as the prior information have also been proposed to improve network performance and interpretability (Gregor and LeCun, 2010; Yang *et al.*, 2018; Xu and Sun, 2018; Wang and Sun, 2020; Liang *et al.*, 2019). Specifically, the model-driven neural network is designed to unfold the optimization process of a mathematical model through a network (Liang *et al.*, 2019). In contrast to the conventional networks, the model-driven network is not only data-driven, but also incorporates a model prior that makes the network easy to be interpreted (Wang and Sun, 2020), and therefore, gained increasing popularity in the medical image area. Gregor *et al.* first proposed a sparse coding neural network based on the optimization procedure (Gregor and LeCun, 2010). ADMM-Net is one of the commonly used model-driven deep neural networks and was first used in MRI for solving the compressed sensing problem with the learnable model parameters (Yang *et al.*, 2018). Ye *et al.* introduced a model-based neural network for estimating NODDI parameters (Ye, 2017b), and we recently proposed a model-driven sparsity coding deep neural network (SCDNN) to estimate the IVIM parameters in the fetal brain (Zheng *et al.*,

*Corresponding author: email: danwu.bme@zju.edu.cn

2021).

However, convolution-based networks used in the current model-driven frameworks have a fixed receptive field within a single layer (Luo *et al.*, 2016), and repeatedly stacking deeper convolution layers will make the model bloated with sharply increasing computation load (Wang *et al.*, 2018). Thus, the self-attention mechanism that adapts a dynamic receptive field (Vaswani *et al.*, 2017) can be added to the q-space deep learning task to improve its performance, which forms a Transformer-like structure. Because of its superior performance and flexibility, Transformer has gained immense interest in many fields. In image processing area, Vision Transformer (ViT) (Dosovitskiy *et al.*, 2020) has been introduced for classification tasks for computer vision and outperformed convolution networks.

Despite its superior performance, applications of ViT in medical imaging are limited due to its high demand for training data. A typical ViT does not need any inductive bias (Dosovitskiy *et al.*, 2020) but requires a large quantity of data for training ($\sim 300\text{M}$). The inductive bias can be considered as a priori hypothesis (Battaglia *et al.*, 2018) that facilitates the network training process. In standard deep learning networks, convolution has the inductive bias of the locality and invariance of spatial translation. Recurrence has the inductive bias of the sequentiality and invariance of time translation. Graph network has the inductive bias of the arbitrariness with the invariance of the node or edge permutations. Inductive bias is not limited to these forms but can also be incorporated by tailored interventions into a deep neural network architecture (Karniadakis *et al.*, 2021), which can be introduced in the model-driven approach.

In order to address the need for large amounts of data for training Transformer and to enhance the model interpretability, we add a new type of inductive bias—model bias, into the Transformer structure to drive the training process. In addition, sparse-coding is introduced to the model-driven process by converting the nonlinear dMRI models into a linear layout using a dictionary technique. Here, we propose a *Microstructure Estimation Transformer with Sparse Coding* (METSC) for dMRI-based microstructural parameter estimation. This frame-

work can be used to estimate different types of dMRI-based models, by modifying the decoder for a specific model. The major contributions are:

- 1) A new framework with the Transformer structure is proposed for dMRI model estimation, which is the first application of Transformer in a regression task in medical imaging.
- 2) The METSC framework introduces a model bias, via the iterative structure, to allow Transformer to be trained with less data, e.g., only about 300K data.
- 3) This framework enables dMRI model parameter estimation using reduced q-space samples, and achieved a reduction of scan time up to 11.25 folds.
- 4) The proposed network has the superior performance in estimating the microstructure parameters, and reduced up to 81% mean square error (MSE) compared with the former learning based method.
- 5) We performed a thorough investigation of network architecture on two types of dMRI models with different data sizes, and METSC outperformed the other state-of-the-art q-space learning methods for both models.

Specifically, we tested this new framework on the IVIM model, which is a bi-exponential model that separates the microcirculation in the capillaries from water diffusion in the tissues (Le Bihan, 2019) using multiple b-value information, and the NODDI model (Zhang *et al.*, 2012) that estimates the microstructural dendrites and axons using multiple diffusion directions. Particularly, we investigated IVIM of the placenta, which is commonly used to assess the placenta perfusion but is subject to abdominal and fetal motion; for NODDI, we used brain MRI from the *Human Connectome Project* (HCP) (Van Essen *et al.*, 2013).

2. Method

In this section, we will first describe the IVIM model and the NODDI model, and then describe the METSC framework in detail.

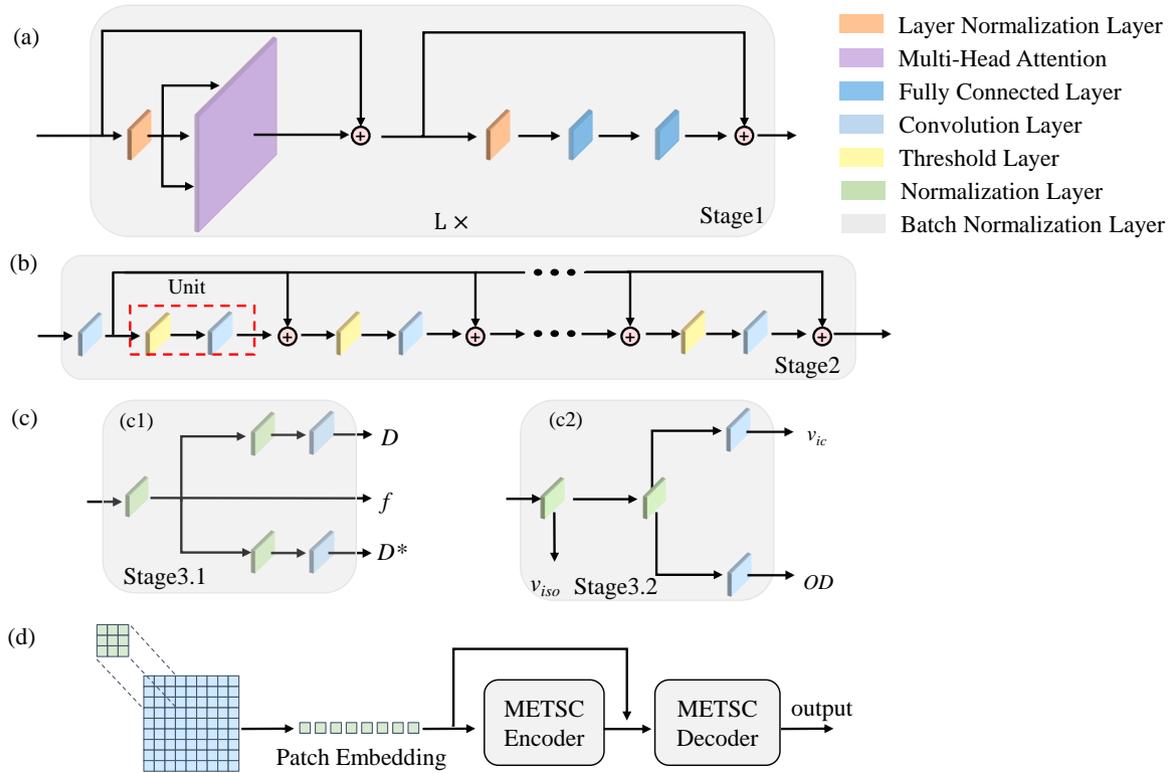


Fig. 1. Overview of the METSC model. (a) Transformer-based encoder architecture, in which images are first split into patches and then fed into a ViT-like Transformer encoder, and additional skip connections are added to patch embedding and METSC encoder. (b) Schematic of the sparsity-based METSC decoder with a cascaded structure that includes model prior in training. A threshold layer plus a convolution layer make up the basic unit in the red dashed box. (c) Stages 3.1 and 3.2 show the mapping stage designed for the IVIM and NODDI model outputs, respectively. (d) The entire METSC framework consists of the Transformer-based encoder and sparsity-based decoder.

2.1. Background

The rationale of selecting the IVIM model and the NODDI model is that IVIM requires multiple b-values densely sampled in the low b-value regime ($b < 800 \text{ s/mm}^2$), while NODDI relies on high-angular resolution (densely sampled diffusion directions) at high b-values ($b > 1000 \text{ s/mm}^2$). They are two representative models to test the generalizability of the proposed framework.

2.1.1. IVIM

The IVIM model separates the water diffusion in the tissue and pseudo-diffusion of microstructural flows in the capillaries, based on the different diffusivity of the two compartments (Le Bihan *et al.*, 1988), with a bi-exponential formulation:

$$S_b = S_0 \left[(1 - f)e^{-bD} + fe^{-bD^*} \right] \quad (1)$$

where S_0 is the non-diffusion-weighted signal and S_b is the diffusion-weighted signals at a b-value of b ; f and D^* are the

fraction and pseudo-diffusivity of microcirculation, and D is the water diffusivity in the tissues.

Traditionally, the IVIM model can be fitted in two ways, including the two-step nonlinear least squares (NLLS) method (Federau *et al.*, 2014) and Bayesian method (Neil and Bretthorst, 1993), and the latter considered to be the best method abdominal imaging (Gustafsson *et al.*, 2018). Therefore, this study used the Bayesian fitting results of the fully sampled IVIM data as the gold standard, and compared the METSC with NLLS, Bayesian, and learning-based methods using downsampled IVIM data.

2.1.2. NODDI

The NODDI model separates the dMRI signal into three parts: intracellular, extracellular, and CSF compartments, and outputs microstructural parameters including the orientation dispersion (OD), and the volume fractions of the intra-cellular compartment (v_{ic}) and the CSF compartment (v_{iso}). The model

can be written as follows:

$$A = (1 - v_{iso})(v_{ic}A_{ic} + (1 - v_{ic})A_{ec}) + v_{iso}A_{iso} \quad (2)$$

where A is the normalized diffusion signal defined as $A = A_b/A_0$, with A_b being the diffusion-weighted signal and A_0 being the non-diffusion-weighted signal. A_{ic} , A_{ec} and A_{iso} represent the signal contributions from the intra-cellular, extra-cellular, and CSF compartments, which can be defined as:

$$A_{ic} = \int_{S^2} M\left(\frac{1}{2}, \frac{3}{2}, \kappa\right)^{-1} e^{\kappa(\mu \cdot \mathbf{n})^2} e^{-b d_{\parallel}(\mathbf{q} \cdot \mathbf{n})^2} d\mathbf{n} \quad (3)$$

A_{ic} is determined by the confluent hypergeometric function M , the diffusion encoding scheme \mathbf{q} according to the gradient direction and b-value, concentration parameter κ , the mean orientation μ , parallel diffusivity d_{\parallel} , and $M\left(\frac{1}{2}, \frac{3}{2}, \kappa\right)^{-1} e^{\kappa(\mu \cdot \mathbf{n})^2} d\mathbf{n}$ gives the probability of finding sticks along orientation \mathbf{n} . Beside \mathbf{q} , M , κ , μ , \mathbf{n} , A_{ec} is also determined by $D(\mathbf{n})$, a cylindrically symmetric tensor with principal orientation \mathbf{n} .

$$A_{ec} = \exp\left(-b\mathbf{q}^T \left(\int_{S^2} M\left(\frac{1}{2}, \frac{3}{2}, \kappa\right)^{-1} e^{\kappa(\mu \cdot \mathbf{n})^2} D(\mathbf{n}) d\mathbf{n} \right) \mathbf{q}\right) \quad (4)$$

The CSF compartment is modeled as a Gaussian diffusion with a diffusivity of d_{iso} :

$$A_{iso} = \exp(-bd_{iso}) \quad (5)$$

The original NODDI fitting with NLLS was relatively accurate but extremely slow. Daducci et al developed the Accelerated Microstructure Imaging via Convex Optimization (AMICO) toolkit (Daducci et al., 2015) that effectively speeded up the process via the sparse representation. Here we used NLLS fitted results of the fully sampled NODDI data as the gold standard, and compared the METSC with AMICO and learning-based methods using downsampled NODDI data.

2.2. METSC

The METSC framework (Fig. 1) can be divided into three parts, a Transformer-based encoder and a sparsity-based decoder that consists of a sparse coding neural network and a model-specific microstructural mapping network.

2.2.1. Transformer based METSC encoder

The Transformer-based METSC encoder is adapted from the ViT structure (Dosovitskiy et al., 2020) and tailored for the model estimation task. It consists of two layer-normalization layers (Ba et al., 2016), a multi-head attention layer, and two fully connected layers (Fig. 1(a)). To accelerate the training of the METSC encoder, a skip connection is added (He et al., 2016) to connect the beginning of one encoder to the end of the encoder. Also to adapt the classification task to the estimation task for dMRI models, the BERT's [class] token is removed. Following the selection of nonlinear activation function in BERT (Devlin et al., 2018), GELU is chosen as our activation function.

First of all, similar to ViT, to accommodate the 2D input, the image $\mathbf{q}_0 \in \mathbb{R}^{H \times W \times C}$ (where H and W are the height and width of the image, C is the number of channels which is the number of b-values in our network) is split into smaller 2D patches $\mathbf{q}_p \in \mathbb{R}^{N \times H_p \times W_p \times C}$ (where H_p and W_p are the height and width of the patch, $N = H/H_p \times W/W_p$ is the number of patches) with a non-overlap design, as the sequence of input to the Transformer. Patch embedding is applied to the sequence of patches, which is learned through training with a linear projection in a fully connected layer.

$$\mathbf{z}_0 = [\mathbf{q}_p^1 \mathbf{E}; \mathbf{q}_p^2 \mathbf{E} \cdots \mathbf{q}_p^N \mathbf{E}] \quad (6)$$

where \mathbf{z}_0 is the sequence of embedded patches and $\mathbf{E} \in \mathbb{R}^{(H_p \times W_p \times C) \times D}$ is the patch embedding projection matrix. Then the data are sent to a Layer Normalization layer followed by the multi-head self-attention (MSA).

$$\mathbf{z}_M = \text{MSA}(\mathbf{z}_I) = [\text{SA}_1(\mathbf{z}_I), \text{SA}_2(\mathbf{z}_I) \cdots \text{SA}_k(\mathbf{z}_I)] \mathbf{U}_{msa} \quad (7)$$

where, \mathbf{z}_I is the normalized \mathbf{z}_0 , $\text{SA}(\cdot)$ is the self-attention layer, $\mathbf{U}_{msa} \in \mathbb{R}^{k \times D_h \times D}$ with k being the number of heads, D being the dimension of the fully connected layer, and D_h being the scalar that is typically set to D/k .

$$\text{SA}(\mathbf{z}_I) = \text{softmax}\left(\frac{qk^T}{\sqrt{D_h}}\right)v \quad (8)$$

where q , k , v correspond to the query, key, and value of the input

sequence z_l and they can be calculated as follows:

$$[q, k, v] = z_l \mathbf{U}_{qkv}, \quad \mathbf{U}_{qkv} \in \mathbb{R}^{D \times 3D_h} \quad (9)$$

The input of embedded patches and the output from the MSA are connected through the skip connection. The output of the skip connection is sent into the Layer Normalization layer and two fully connected layers with the gaussian error linear units (GELU) activation function and dropout.

$$z_{FC} = \text{FFN}(z_{l2}) + z_0 + z_M \quad (10)$$

Where, z_{l2} is the normalized $z_0 + z_M$, and $\text{FFN}(\cdot)$ is a feed-forward network containing two fully connected layers. The flow chart of the Transformer-based METSC encoder is shown in Fig. 2(a). The overall procedure is summarized in the supplementary material Algorithm 1.

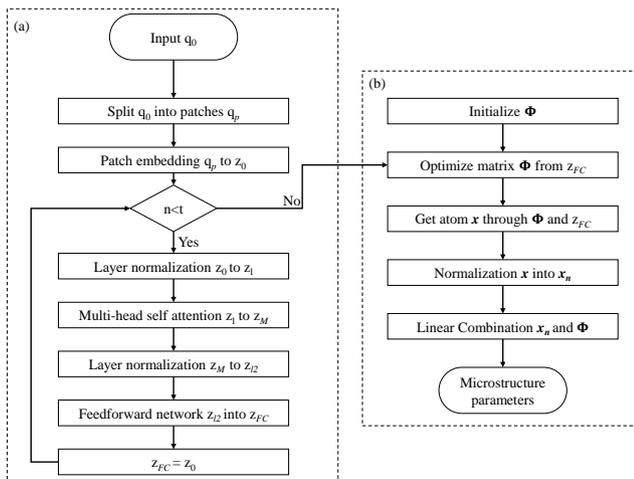


Fig. 2. The algorithmic flow of the METSC. (a) Transformer-based encoder flow chart. (b) Sparsity-based decoder flow chart. Here t is the number of Transformer-based encoders.

2.2.2. Sparsity-based METSC decoder

The sparsity-based METSC decoder is a model-driven deep neural network that provides an inductive bias in METSC. As mentioned above, the inductive bias can be seen as a type of prior information (Battaglia *et al.*, 2018; Karimi *et al.*, 2021), which takes the form of the IVIM model or NODDI model in the current study. In this section, the METSC decoder configuration will be shown for IVIM model and NODDI model, separately. The method of construction sparsity-based METSC decoder can be briefly summarized as dictionary construction and linear combination.

2.2.2.1. Decoder for IVIM Model.

Sparse coding. The main challenge of a model-driven network is the choice of the optimization algorithm. The IVIM model follows a nonlinear bi-exponential function (Eq. [1]), in which the D and D^* are the exponential terms, and f is coupled with exponents. Thus, the model cannot directly be translated into the network through the optimization procedure. Inspired by AMICO (Daducci *et al.*, 2015), the nonlinear models can be represented via dictionary learning. In this work, the IVIM model is linearized through a sparse-coding based dictionary learning framework as below:

$$z_{FC} = \Phi \mathbf{x} + \eta \quad (11)$$

where $z_{FC} = (z_1, \dots, z_n)^T$ is a vector comprised of the encoded dMRI signals from the Transformer-based encoder that is acquired at n different b-values; Φ is a dictionary vector ($\Phi \in \mathbb{R}^{1 \times 2j}$, where j corresponds to the length of the discretized D and D^*); \mathbf{x} is a vector of the dictionary coefficients ($\mathbf{x} \in \mathbb{R}^{2j \times 1}$), and η is a noise term. The dictionary can be established through the discretized D and D^* , and the x corresponds to the fraction of f :

$$\Phi = [\Phi_D, \Phi_{D^*}] \quad (12)$$

$$\mathbf{x} = [x_{1-f}, x_f]^T \quad (13)$$

The signals are normalized to the b_0 signal and fall into the interval of $[0, 1]$, and thus, the three parameters can be reformulated as below:

$$\mathbf{x} = \frac{\mathbf{x} + \tau}{\|\mathbf{x} + \tau\|_1} \quad (14)$$

$$x_{1-f} = \frac{x_{1-f} + \tau}{\|x_{1-f} + \tau\|_1} \quad (15)$$

$$x_f = \frac{x_f + \tau}{\|x_f + \tau\|_1} \quad (16)$$

$$f = \mathbf{I}_1 \mathbf{x} \quad (17)$$

$$D = \frac{\Phi \mathbf{I}_2 x_{1-f}}{\mathbf{I}_2 x_{1-f}} \quad (18)$$

$$D^* = \frac{\Phi \mathbf{I}_3 x_f}{\mathbf{I}_3 x_f} \quad (19)$$

where, $\tau = 1e^{-10}$ is set to avoid 0 in the denominator. $\mathbf{I}_1 \in \mathbb{R}^{1 \times 2j}$, $\mathbf{I}_2, \mathbf{I}_3 \in \mathbb{R}^{2j \times j}$, and $\mathbf{I}_4 \in \mathbb{R}^{1 \times 2j}$ are defined as:

$$\mathbf{I}_4 = [0 \dots 0, 1 \dots 1]^T \quad (20)$$

$$\mathbf{I}_2 = \begin{pmatrix} \mathbf{I}_5 \\ \mathbf{0} \end{pmatrix} \quad (21)$$

$$\mathbf{I}_3 = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_5 \end{pmatrix} \quad (22)$$

$$\mathbf{I}_4 = [1 \cdots 1, 0 \cdots 0]^T \quad (23)$$

where, $\mathbf{I}_5 \in \mathbb{R}^{j \times j}$ is an identity matrix.

Network construction. According to Eq.[11], the bi-exponential IVIM model can be converted into a linear model. The next step is to establish a dictionary that optimally represents the signals, with the objective function as below:

$$\min_{\mathbf{x}} \|\mathbf{y} - \Phi \mathbf{x}\|_2^2 + \beta \|\mathbf{x}\|_0 \quad (24)$$

where β controls the sparsity of matrix \mathbf{x} . Here, the *Iterative Hard Thresholding* (IHT) (Blumensath and Davies, 2009) method is used for optimization, which is formulated:

$$\mathbf{x}^{n+1} = H_M(\mathbf{W}z_{FC} + \mathbf{S}\mathbf{x}^n) \quad (25)$$

$$= H_M[\mathbf{x}^n + \Phi^H(z_{FC} - \Phi)\mathbf{x}^n] \quad (26)$$

$$= H_M[\Phi^H z + (\mathbf{I} - \Phi^H \Phi)\mathbf{x}^n] \quad (27)$$

where, $\mathbf{W} = \Phi^H$, $\mathbf{S} = \mathbf{I} - \Phi^H \Phi$, and H_M denotes a nonlinear operator:

$$H_M(\mathbf{x}) \begin{cases} 0 & \text{if } |\mathbf{x}| < \lambda \\ \mathbf{x} & \text{if } |\mathbf{x}| \geq \lambda \end{cases} \quad (28)$$

where λ is a positive threshold. In the IVIM model, the model parameters are nonnegative, and thus, the nonlinear operator can be simplified as Eq.[29]:

$$H_M(\mathbf{x}) \begin{cases} 0 & \text{if } \mathbf{x} < \lambda \\ \mathbf{x} & \text{if } \mathbf{x} \geq \lambda \end{cases} \quad (29)$$

Thus, the network can be designed by unfolding the iterative process using the sparsity-based METSC decoder. \mathbf{W} and \mathbf{S} are the shared weights among the layers including the dictionary layer, which do not need to be pre-trained. The decoder design algorithm for IVIM can be summarized in the supplementary material Algorithm 2.1. After the dictionary is trained in Fig. 1(b), the parameters can be estimated through Eq. [17], Eq. [18], and Eq. [19], as shown in Fig. 1(c1).

2.2.2.2. Decoder for NODDI Model.

Sparse coding. In the NODDI model (Eq.[2]), the signal can also be linearized following Eq. [11]. Similarly in NODDI, $z_{FC} = (z_1, \cdots, z_n)^T$ is a vector comprised of the encoded dMRI signals from the Transformer-based encoder that is acquired at n different diffusion gradients; Φ is a dictionary vector ($\Phi \in \mathbb{R}^{1 \times 2j+i}$, $\Phi_t \in \mathbb{R}^{1 \times 2j}$, $\Phi_i \in \mathbb{R}^{1 \times i}$ here j corresponds to the length of the discretized v_{ic} and κ , and the length of v_{iso} is i), and \mathbf{x} is a vector of the dictionary coefficients ($\mathbf{x} \in \mathbb{R}^{1 \times 2j+i}$, $\mathbf{x}_t \in \mathbb{R}^{2j \times 1}$, $\mathbf{x}_i \in \mathbb{R}^{i \times 1}$, \mathbf{x}_t is the coefficient of anisotropic signals including v_{ic} , κ , and \mathbf{x}_i is the coefficient of the isotropic v_{iso}). The dictionary can be established through the discretized v_{ic} , κ , v_{iso} , and Φ , \mathbf{x} can be defined below:

$$\Phi = [\Phi_t, \Phi_i] \quad (30)$$

$$\mathbf{x} = [\mathbf{x}_t, \mathbf{x}_i]^T \quad (31)$$

The components in \mathbf{x}_t need to be normalized into the interval of [0,1]. Then, the three parameters v_{ic} , κ and v_{iso} can be obtained as below:

$$v_{iso} = \mathbf{I}_6 \mathbf{x}_i \quad (32)$$

$$\mathbf{x}_t = \frac{\mathbf{x}_t + \tau}{\|\mathbf{x}_t + \tau\|_1} \quad (33)$$

$$v_{ic} = \frac{\Phi_t \mathbf{I}_7 \mathbf{x}_t}{\mathbf{I}_9 \mathbf{x}_t} \quad (34)$$

$$\kappa = \frac{\Phi_t \mathbf{I}_8 \mathbf{x}_t}{\mathbf{I}_9 \mathbf{x}_t} \quad (35)$$

where $\mathbf{I}_5 \in \mathbb{R}^{j \times j}$ is an identity matrix, and $\mathbf{I}_6 \in \mathbb{R}^{1 \times i}$, $\mathbf{I}_7 \in \mathbb{R}^{2j \times 2j}$, $\mathbf{I}_8 \in \mathbb{R}^{2j \times 2j}$, $\mathbf{I}_9 \in \mathbb{R}^{1 \times 2j}$ are defined as:

$$\mathbf{I}_6 = [1 \cdots 1]^T \quad (36)$$

$$\mathbf{I}_7 = \begin{pmatrix} \mathbf{I}_5 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad (37)$$

$$\mathbf{I}_8 = [1 \cdots 1]^T \quad (38)$$

$$\mathbf{I}_9 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_5 \end{pmatrix} \quad (39)$$

Finally, OD can be calculated through:

$$\text{OD} = \frac{2}{\pi} \text{atctan} \left(\frac{1}{\kappa} \right) \quad (40)$$

Network construction. The network can be designed in the same way as in Section 2.2.2.1. After the dictionary is

trained in Fig. 1(b), the parameters can be estimated through Eq.[32], Eq.[34], and Eq.[40] as shown in Fig. 1(c2). The decoder design algorithm for NODDI can be summarized in the supplementary material Algorithm 2.2. Overall, the flow chart of the METSC decoder can be shown in Fig. 2(b). Finally, the METSC encoder and decoder can be combined with a skip connection to connect the input and the framework, as illustrated in Fig. 1(d). The code will be provided at <https://github.com/Tianshu996/METSC> upon publication.

3. Results

In this section, we tested the METSC framework on both the IVIM and NODDI models to find the optimal setup. We also compared our method with the state-of-the-art networks for dMRI-based microstructural estimation.

3.1. IVIM Model

In subsection, we used placental IVIM data to train the model and determined the optimal hyper-parameter with a set of ablation experiments. We then compared our approach to existing optimization-based and learning-based methods. We also evaluate the generalizability of our framework on an independent dataset obtained from a different center.

3.1.1. Dataset

The placental IVIM data were acquired on a 1.5T GE scanner (SIGNA HDXT) from 24 normal pregnant women (gestational age 13-37 weeks) under Institutional Research Board approval at the local hospital. Diffusion gradients were applied in three orthogonal directions at 10 b-values of 0, 10, 20, 50, 80, 100, 150, 200, 300, 500 s/mm^2 with the following acquisition parameters: repetition time/echo time = 3000/76 ms , in-plane resolution = $1.25 \times 1.25 mm^2$, field of view = $320 \times 320 mm^2$, 15 slices with a slice thickness of 4 mm . The data was preprocessed through bias correction and registration for motion correction, and signals within the placenta mask were used in the following experiment in a voxelwise manner. To obtain the gold standard (f , D , and D^*), we performed Bayesian estimation of these parameters using the full

dataset (10 b-values) using the Matlab fitting toolbox (Gustafsson et al., 2018).

The dataset was then divided into the training, validation, and testing datasets, resulting in 324016 voxels from 15 subjects for training with 10% of the training samples used for validation (Golkov et al., 2016), and 248059 voxels from 9 subjects for testing. All datasets were split into overlapping patches with a step size of 1 in zero-padded images.

We also generated simulation data (S) using the gold standard model parameters according to Eq.[41] and added noise to generate data at different signal-to-noise ratio (SNR) levels according to (Daducci et al., 2014):

$$S_{\text{simulated}} = \sqrt{(S + \xi_1)^2 + (\xi_2)^2} \quad (41)$$

where, $\xi_1, \xi_2 \sim N(0, \sigma^2)$, and $\sigma = S_0/\text{SNR}$. Similar to (Daducci et al., 2014), we assumed $S_0 = 1$ and SNR varied from 10 to 70.

3.1.2. Training

All the models were trained using Adam as the optimizer with the total epochs of 2000 and batch size of 512. We used the cosine warm-up method in the first 200 epochs and a reducing learning rate with an initial learning rate of 1×10^{-4} . The experiments were performed on a Linux machine with eight NVIDIA GeForce RTX 3090 GPUs.

3.1.3. Ablation Experiments on Network Architecture

Eight pairs of ablation experiments were performed to compare model-free decoder / METSC decoder, convolution encoder / METSC encoder, non-patched / patched image inputs, different sizes of the training data, different combinations of b-value, the different number of b-values, the different dictionary size, and the different patch size.

(1) **Model-free decoder versus Sparsity decoder** were tested in combination with METSC encoder with patched inputs at five b-values (20, 50, 150, 300, 500 s/mm^2). The model-free decoder was designed following Golkov that was composed of three fully connected layers and the nonlinear activation ReLU (Golkov et al., 2016). The results in Fig. 3(a)

Table 1. MSE of estimated f , D , and D^* using METSC with different encoders, decoders, and input forms, on 9 testing data (about 248059 voxels). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ by paired t-test between different methods on the nine patients.

		$f \times 10^{-6}$	$D (\times 10^{-4} \mu\text{m}^2/\text{ms})$	$D^* (\times 10^{-2} \mu\text{m}^2/\text{ms})$
Decoder (with METSC encoder / patched input)	Model-free	46.5(**)	4.59 (*)	47(**)
	METSC	6.3	2.15	1.4
Encoder (with METSC decoder / patched input)	Convolutional	8.45 (**)	2.78	8.1 (**)
	METSC	6.33	2.15	1.4
Input (with METSC encoder / METSC decoder)	Non-patch	3.60×10^3 (***)	2.44×10^6 (***)	1.99×10^3 (***)
	Patch	6.33	2.151	1.4

Table 2. The MSE of estimated f , D , and D^* using the different number of training data with METSC and ViT.

	$f (\times 10^{-4})$		$D (\times 10^{-4} \mu\text{m}^2/\text{ms})$		$D^* (\times 10^{-2} \mu\text{m}^2/\text{ms})$	
	ViT	METSC	ViT	METSC	ViT	METSC
10K	7.2	0.83	2.02×10^6	2.02×10^3	630	39
40K	2.1	0.29	400	3.9	160	9.5
200K	2.3	0.11	4.34	2.2	171	2.6
300K	2.8	0.061	3.93	1.8	241	1.3

showed the model parameters estimated from the METSC decoder achieved higher correlations with the gold standard compared to the model-free decoder. It was also clear from the error maps (gold standard – estimated parameters, Fig. 3(a)) that the METSC decoder resulted in lower estimation errors than the model-free decoder, especially for the Δf and the ΔD^* maps.

(2) **Convolution encoder versus Transformer encoder** were tested in combination with METSC decoder with patched inputs at five b-values (20, 50, 150, 300, 500 s/mm^2). The convolution encoder consisted of 2D convolution layers, Batch Normalization layers, and the ReLU activation. The results in Fig. 2(b) showed that the Transformer encoder provided a more accurate estimation of the IVIM model parameters according to the correlation plots and the Δf , ΔD , and ΔD^* maps.

(3) **Patch versus non-patch-based inputs** were tested with the METSC encoder and decoder at five b-values (20, 50, 150, 300, 500 s/mm^2). The results in Table 1 (third row) showed that METSC with patch-based inputs outperformed the non-patch inputs.

(4) **Effect of training data size.** Based on the optimal

METSC setup (Transformer encoder + Sparsity decoder with patched inputs) obtained above, we tested the network performance using varying training data sizes of 10K, 40K, 200K, 300K. We found that with insufficient training data (10K), METSC performed worse than the SCDNN, which is a model-driven neural network without the Transformer encoder (Zheng *et al.*, 2021). As the training data increased to 200K, METSC reached a comparable accuracy to the SCDNN. The performance of METSC further increased and outperformed SCDNN as the number of training data increased to 300K (Fig. 4). We also tested the ViT structure without sparsity decoder, and showed that METSC outperformed the ViT structure for all training data sizes between 10K-300K (Table 2).

Table 3. The MSE of estimated IVIM model parameters using different combinations of b-values and different number of b-values. The combinations with top performance were highlighted in bold.

	f ($\times 10^{-4}$)	D ($\times 10^{-4} \mu\text{m}^2/\text{ms}$)	D^* ($\times 10^{-2} \mu\text{m}^2/\text{ms}$)
3 b-values			
(20, 150, 500) s/mm^2	0.41	2.84	8.4
5 b-values Comb1 (ours)	0.063	2.2	1.4
(20, 50, 150, 300, 500) s/mm^2			
5 b-values Comb2	0.4	2.1	1.8
(20, 50, 150, 200, 500) s/mm^2			
5 b-values Comb3	0.063	5.6	1.8
(20, 50, 200, 300, 500) s/mm^2			
5 b-values Comb4	0.097	1.7	1.7
(20, 100, 150, 300, 500) s/mm^2			
5 b-values Comb5	0.094	1.7	2.0
(20, 80, 150, 300, 500) s/mm^2			
7 b-values			
(20, 50, 100, 150, 200, 300, 500) s/mm^2	0.057	1.1	1.9

(5) **Effect of the choice of b-values.** In Experiment 1-4, we used a selected subset of 5 (out of 10) b-values at 20, 50, 150, 300, 500 s/mm^2 as the diffusion-weighted signals at

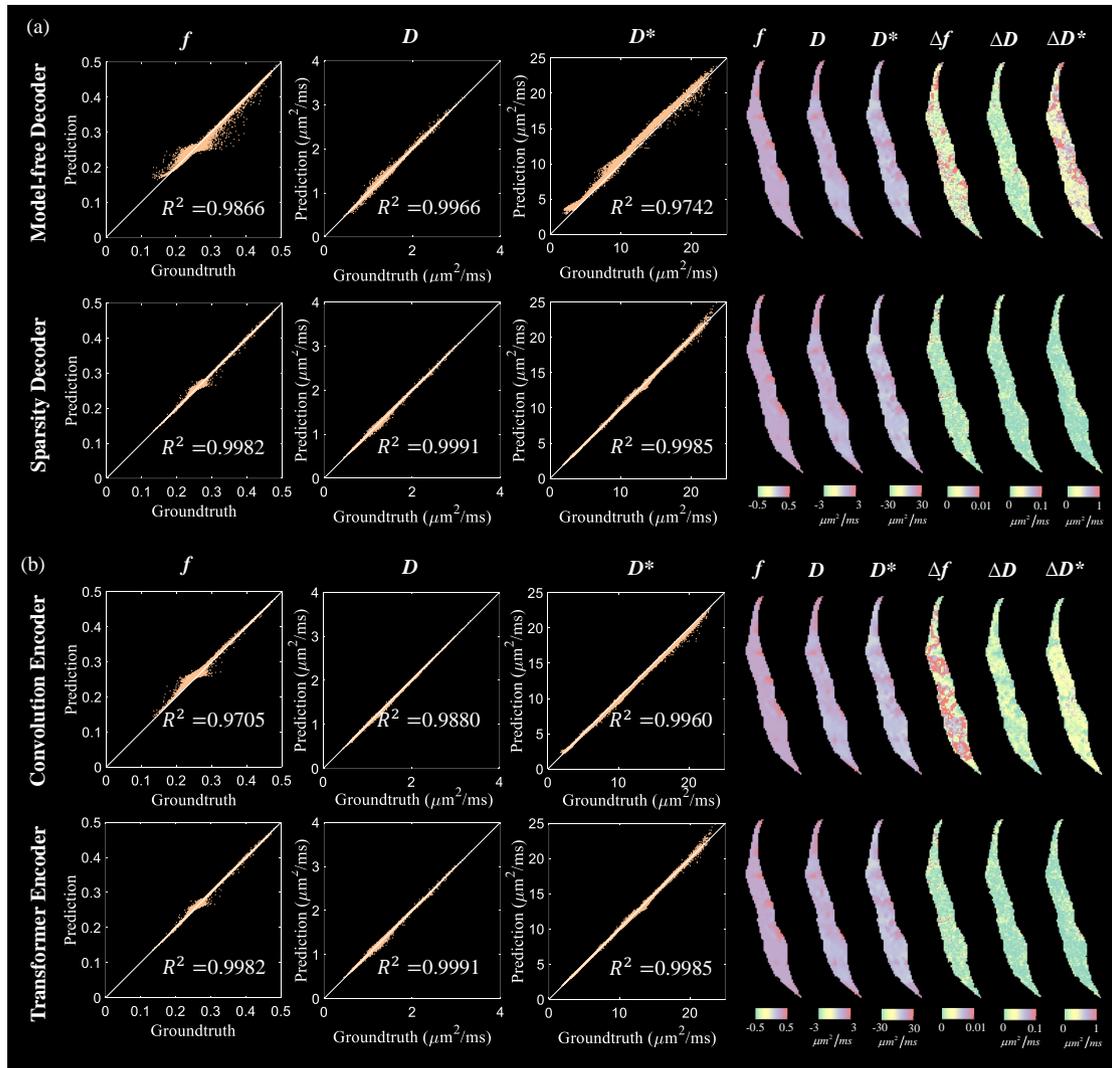


Fig. 3. Ablation experiments on the performance of decoder and encoder of the METSC framework. (a) Estimated model parameters using the model-free versus sparsity-based METSC decoders, based on the voxelwise correlation between the estimated values and ground truth, the estimated parameter maps (f , D , and D^*), and the error maps (Δf , ΔD , and ΔD^*). (b) Estimated model parameters using the convolution and Transformer-based METSC encoders.

these b-values best characterize the signal decay curve. Here we tested another four combinations of five b-values as listed in Table 3, and the combinations with top performance were highlighted in bold for each model parameter. Overall, the b-value combination of 20, 50, 150, 300, 500 s/mm^2 achieved the optimal balance of estimation accuracy for all the three IVIM parameters.

(6) **Effect of the number of b-values.** As expected, the estimation accuracy increased with the number of b-values. Table 3 showed that the MSE considerably reduced as the number of b-values increased from 3 to 5, but the further increase of b-values from 5 to 7 had a limited improvement (15% reduction in the sum of validation loss) at the expense of 1.4 longer scan

time.

(7) **Effect of the dictionary size** was investigated based on the validation loss. The validation loss here was defined as the sum of loss of parameters f , D , and D^* on the validation set. Results in Fig. 5 indicated that the validation loss decreased as the dictionary size changed from 200 to 300, but increased as dictionary size changed from 300 to 400. We went one step further to test dictionary sizes greater than 400 and found that validation loss further decreased and stayed nearly stable after 600. Because the training time increased dramatically as the dictionary size exceeds 600, e.g., the training time for an 800 dictionary was about twice that of a 600 dictionary, we determined the optimal dictionary size to be 600.

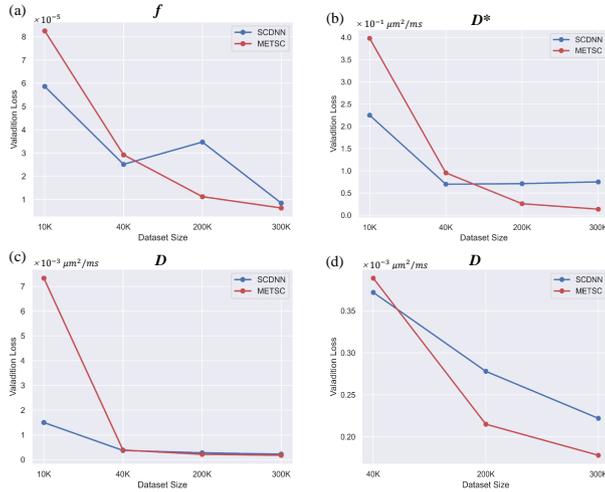


Fig. 4. (a-c) The MSE of estimated f , D , and D^* using the different number of training data with METSC and SCDNN, which is a model-driven learning method without the Transformer encoder (Zheng et al., 2021). (d) Zoom-in view of the MSE of D in the range of training data size 40K-300K.

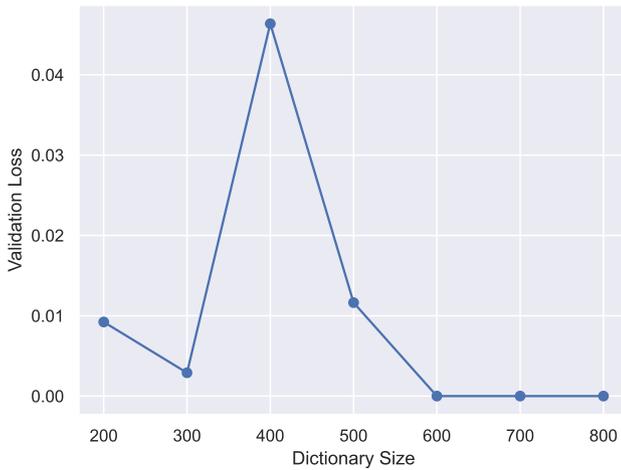


Fig. 5. The results of overall MSE ($f + D + D^*$) on the validation set with different dictionary size. When the dictionary size is larger than 600, the loss on the validation set tends to be stable.

(8) **Effect of the patch size** was also tested according to the validation loss. Fig. 6 showed a minimum loss at patch size of 3, compared to patch sizes of 5 and 7. Moreover, with the increase in patch size, training time increased sharply from 11h to 61h.

From ablation experiments 1-8, we can conclude the optimal METSC structure consists of the Transformer encoder and the sparsity decoder with patched inputs at patch size of 3, using 5 b-values of 20, 50, 150, 300, 500s/mm².

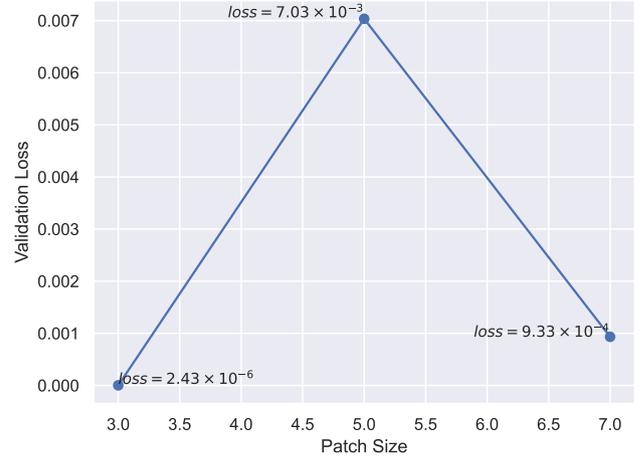


Fig. 6. The impact of different patch sizes on the validation loss. The loss increases first and then decreases with the increase of patch size, and the minimal loss is achieved at patch = 3.

3.2. Performance Evaluation

We evaluated the performance of the optimized METSC network in terms of its robustness against noise (SNR), estimation accuracy compared to the other state-of-the-art algorithms, and its generalizability on multicenter data.

(1) **Effects of SNR.** Different levels of noise were added into the signal according to Eq. [41] resulting in SNR levels from 10 to 70, and we evaluated the relative error (percentage of the gold standard at the different SNR levels.) Fig. 7(a) showed that the relative error of f decreased gradually as SNR increased and stabilized at SNR above 40. In contrast, the estimation of D was relatively insensitive to SNR (Fig. 7(b)). The relative error of D^* changed slightly with SNR and stabilized at SNR above 40. The results indicated the estimated parameters were relatively robust against noise and an SNR above 40 could ensure optimal accuracy.

(2) **Comparison with other algorithms.** Four different algorithms were compared, including two optimization methods—the NLLS and Bayesian method (Gustafsson et al., 2018; Jalnefjord et al., 2018), and three learning-based methods—qDL (Golkov et al., 2016), IVIM-NET (Barbieri et al., 2020) and SCDNN (Zheng et al., 2021). Compared with NLLS and Bayesian methods, the learning-based methods provided significantly higher estimation accuracy. Among the learning-based methods, the model-driven methods (SCDNN and METSC)

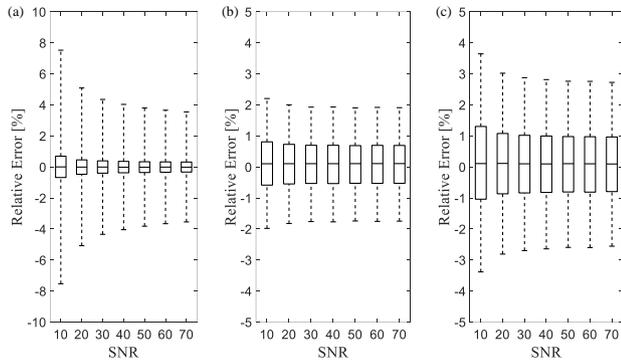


Fig. 7. Relative errors (percentage of the gold standard at the different SNR levels) of the estimated IVIM parameters at SNR levels from 10 to 70.

outperformed the prior-information free q-DL and IVIM-Net, and the METSC demonstrated the best performance among the six methods (Table 4).

Table 4. Comparison of six methods in estimating IVIM model parameters using a reduced number of b-values (5 b-values at 20, 50, 150, 300, 500 s/mm^2). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ by paired t-test between each of the algorithms with METSC.

	NLLS	Bayesian	q-DL	IVIM-NET	SCDNN	METSC
f ($\times 10^{-4}$)	0.54(***)	12(***)	0.7(***)	25(***)	0.072(*)	0.063
D ($\times 10^{-4} \mu m^2/ms$)	2.3×10^3 (***)	22(**)	5.1(***)	2.8×10^3 (***)	3.1	2.2
D^* ($\times 10^{-2} \mu m^2/ms$)	19.8×10^3 (***)	42(***)	38(***)	1.3×10^3 (***)	1.8(*)	1.4

(3) Comparison with other algorithms in b-value choices.

To decouple the effects of the q-space sampling scheme and network performance, the three learning-based methods (q-DL, SCDNN, and METSC) were compared against the different b-value setups in the ablation experiment 6-7 in Section 3.1.3. The results in Table A1 demonstrated that METSC achieved the best performance compared to other methods for all b-value combinations, and the optimal b-value choice was consistent with the ablation experiments (Table 3).

(4) **Multicenter validation.** The previous tests were performed using training, validation, and testing data acquired on a 1.5T GE SIGNA HDXT scanner, and here we tested the network on data acquired on a 3.0T GE 750W scanner at another hospital with the same acquisition protocol. The new testing data included 2 patients (37194 voxels). The results in Table 5 showed that the METSC achieved the least estimation error using the reduced number of b-values (5 b-values at 20, 50, 150, 300, 500 s/mm^2) compared to the other algorithms.

The estimation accuracy was slightly reduced on the multicenter data compared to that on the single center, but still sufficient for parameter estimation, with R^2 between predicted values and ground truth over 0.996 (Fig. 8).

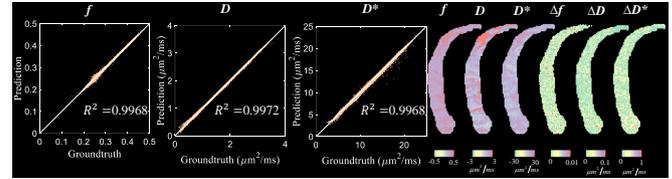


Fig. 8. Estimated IVIM parameters in comparison with ground truth on the multicenter test data by METSC.

Table 5. Evaluation of estimation errors on IVIM data acquired from another center with different scanner and field strength, using the five algorithms with five b-values.

	NLLS	Bayesian	q-DL	SCDNN	METSC
f ($\times 10^{-4}$)	320	30	29	4.4	0076
D ($\times 10^{-4} \mu m^2/ms$)	97	21	12	0.72	0.027
D^* ($\times 10^{-1} \mu m^2/ms$)	973	45	34	8.4	0.48

3.3. NODDI Model

In this part, we first describe how we selected the public dataset and we compared our proposed framework with all other published methods for NODDI parameter estimation in terms of accuracy and precision. Also, we tested our framework with different q-space downsample schemes.

3.4. Dataset and Training

The multi-shell dMRI from HCP data (Van Essen *et al.*, 2013) were acquired on a 3T MR scanner with 3 b-values ($b = 1000, 2000, 3000 s/mm^2$) and 90 diffusion directions per b-value. We randomly selected 26 subjects and used 5 of them for training (with 10% of the training samples as validation (Golkov *et al.*, 2016)), and 21 subjects for testing. To evaluate the proposed METSC, the dataset was downsampled to 30 gradient directions per b-shell at $b = 1000$ and $b = 2000 s/mm^2$ for comparison with other studies (Ye, 2017b). The gold standard microstructural parameters were computed by the NODDI Matlab Tool Box (Zhang *et al.*, 2012) using all 270 q-space data. Similar to section 3.1, all datasets were split into overlapping patches with a step size of 1 in zero-padded images.

3.5. Performance Evaluation

The performance of the proposed network on NODDI was compared with six algorithms including the conventional dictionary learning based method AMICO (Daducci *et al.*, 2015), a traditional q-space learning method q-DL (Golkov *et al.*, 2016), two model-driven learning-based methods (Ye, 2017b; Ye *et al.*, 2020). As MEDN (Ye, 2017a), MEDN+ (Ye, 2017b), MESC (Ye *et al.*, 2019) and MESC-Net Sep_Dict (Ye *et al.*, 2020) (abbreviated as MESC_Sep here) are variations of the same class of algorithm, and MEDN+ and MESC_Sep have superior performance than MEDN and MESC, thus we only showed the results of MEDN+ and MESC_Sep. We further tested the networks with number of diffusion directions. Robustness test was also carried out by smearing the signal and adding heavy noise.

3.5.1. Comparison with other algorithms

In the estimation accuracy test, we used downsampled q-space with 30 diffusion directions per shell from 21 test subjects and compared the MSE of the different algorithms. Fig. 9 indicated that the v_{ic} , v_{iso} and OD estimated from AMICO was overall worse than other learning-based methods. The proposed METSC provided the best results compared to other learning algorithms (Table 6). The error maps between the were shown in Fig. 10, and a zoomed view of error maps were illustrated in Fig. 11. All results pointed to that the AMICO results were the worst with reduced q-space data, and our proposed METSC outperformed others. The mean and standard deviations of the average estimation errors across 21 test subjects were shown in Table 6 for all algorithms, which showed that the statistically reduced estimated errors by METSC lower than all other methods via paired Student's t-test.

3.5.2. Effect of the different number of diffusion directions

In the previous experiments, the number of diffusion directions was set to be 30 for each shell ($b = 1000, 2000s/mm^2$). In this part, we further reduced the number of directions to 18 and 12 for each shell. The results in Table 6 demonstrated METSC achieved minimal estimation errors compared to other algorithms for all choices of gradient numbers.

Table 6. Evaluation of estimation errors on NODDI parameters using different methods on different number of diffusion directions. **p<0.01, *p<0.001 by paired t-test between different methods with respect to the METSC is marked.**

		AMICO	q-DL	MEDN+	MESC_Sep	METSC
30 diffusion gradients per shell	v_{ic}	4.9 ± 1.8	0.4 ± 0.09	0.3 ± 0.04	7.0 ± 2.8	0.08 ± 0.01
	$\times 10^{-2}$	(***)	(***)	(**)	(***)	
	v_{iso}	3.0 ± 1.3	1.7 ± 0.2	1.5 ± 0.3	1.8 ± 0.4	0.4 ± 0.02
	$\times 10^{-3}$	(***)	(***)	(***)	(***)	
18 diffusion gradients per shell	OD	31 ± 29	4.2 ± 0.3	3.9 ± 0.2	2.3 ± 2.3	0.9 ± 0.01
	$\times 10^{-3}$	(***)	(***)	(***)	(***)	
	v_{ic}	6.7 ± 2.0	0.5 ± 0.1	0.4 ± 0.03	8.9 ± 8.7	0.09 ± 0.01
	$\times 10^{-2}$	(***)	(***)	(***)	(***)	
12 diffusion gradients per shell	v_{iso}	5.7 ± 1.4	2.9 ± 1.5	2.0 ± 0.8	2.2 ± 0.5	0.5 ± 0.03
	$\times 10^{-3}$	(***)	(***)	(***)	(***)	
	OD	54 ± 31	7.8 ± 1.3	4.9 ± 1.3	11 ± 11	1.5 ± 0.02
	$\times 10^{-3}$	(***)	(***)	(***)	(***)	
12 diffusion gradients per shell	v_{ic}	7.5 ± 2.0	1.0 ± 0.1	0.6 ± 0.3	51 ± 9.2	0.1 ± 0.02
	$\times 10^{-2}$	(***)	(***)	(***)	(***)	
	v_{iso}	7.4 ± 2.4	7.3 ± 1.5	3.5 ± 0.8	2.6 ± 0.5	0.6 ± 0.03
	$\times 10^{-3}$	(***)	(***)	(***)	(***)	
12 diffusion gradients per shell	OD	87 ± 31	8.9 ± 1.3	5.1 ± 0.4	60 ± 7.0	2.0 ± 0.2
	$\times 10^{-3}$	(***)	(***)	(**)	(***)	

3.5.3. Robustness test

Beyond evaluating the estimation accuracy as done in previous studies (Golkov *et al.*, 2016; Ye, 2017a,b; Ye *et al.*, 2019, 2020), this study also investigated the robustness of the network. The robustness test was divided into two parts, the first part tested how the choice of diffusion directions affected the results (Karimi *et al.*, 2021), and the second part tested the robustness of the network in response to the abnormal input signal.

In the first test, we used three different combinations of diffusion directions ($n=30$) by bootstrap, and the standard deviation of estimated parameters from the three datasets was used to evaluate the robustness. Fig. 12 indicated that METSC resulted the least variation among the bootstraps and thus the highest robustness. When the network took the abnormal inputs, such as the smeared input and the noise added input (Fig. 13), it did not generate unexpected / forged outputs.

4. Discussion

In this study, a model-bias was introduced to facilitate the training of the transformer structure. And in this part, the discussion will be talked about our motivation, how the model-bias work, how we setup our network, and our future work.

4.1. Motivation

Motivated by the feature extraction capacity of the Transformer, in this work, we proposed a *Microstructure Estimation Transformer with Sparse Coding* (METSC) that integrates

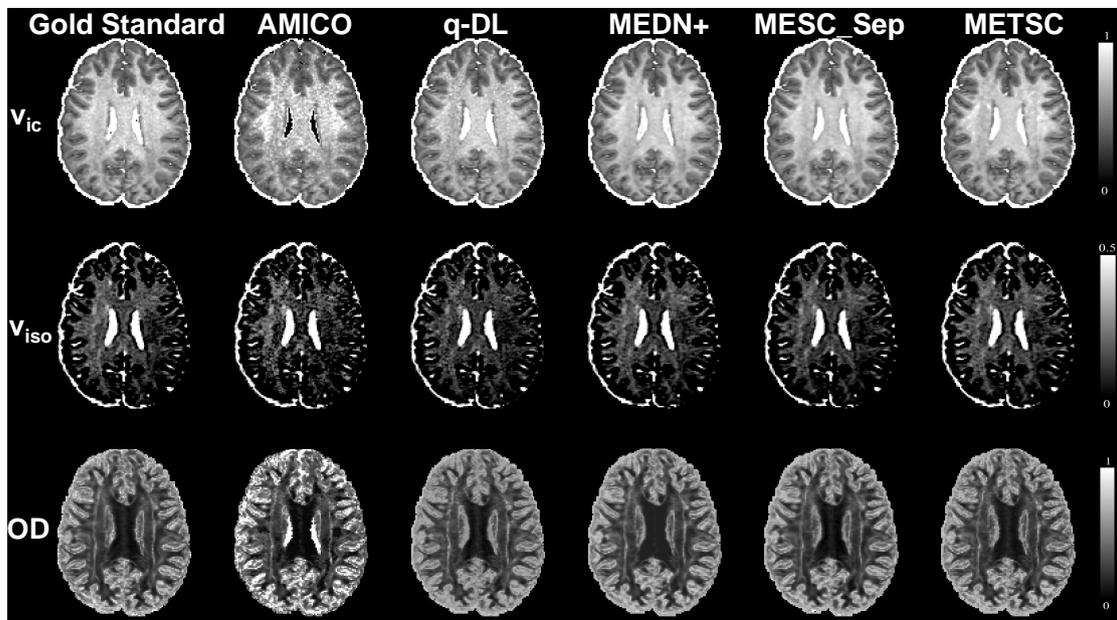


Fig. 9. The gold standard and estimated NODDI parameter v_{ic} , v_{iso} , and OD based on AMICO, q-DL, MEDN+, MESC_Sep, and METSC (ours) in a test subject with 30 diffusion directions per shell at b-value of 1000 and 2000 s/mm^2 .

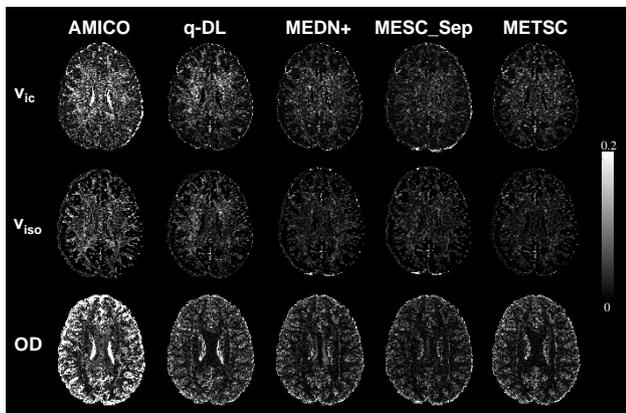


Fig. 10. Estimation errors of v_{ic} , v_{iso} , and OD in a representative test subject using AMICO, q-DL, MEDN+, MESC_Sep, and METSC (ours) in a test subject with 30 diffusion directions per shell at b-value of 1000 and 2000 s/mm^2 .

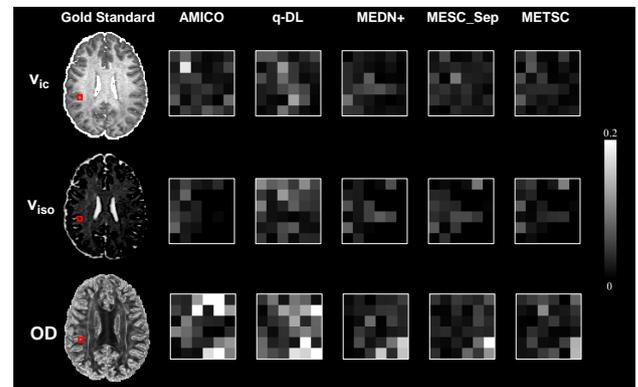


Fig. 11. Zoom-in views of estimation errors of v_{ic} , v_{iso} , and OD in a representative test subject using AMICO, q-DL, MEDN+, MESC_Sep, and METSC (ours) in a test subject with 30 diffusion directions per shell at b-value of 1000 and 2000 s/mm^2 .

the Transformer encoder with a model-based sparsity decoder to enhance the model estimation and enabled the Transformer to be efficiently trained with limited data. To our best knowledge, this is the first time the Transformer structure is applied to a regression task in the medical imaging area, especially for dMRI-based microstructural parameter estimation. Meanwhile, it is also the first time a physiological model bias is introduced into the Transformer structure via an iterative optimization technique, which not only improves the model interpretability but also the training efficiency of the Transformer.

To demonstrated the generalizability of METSC, we chose the IVIM and NODDI models to testing the network performance as they are representative of various types of dMRI models, e.g., IVIM model is a clinically useful model that estimates microcapillary flow from multiple b-values and NODDI emphasizes high angular resolution for resolving neurite orientation and is heavily studied with deep learning techniques. By modifying stage 3 in Fig. 1 to a specific model configuration, the suggested METSC framework can be adapted to a variety of signal models beyond dMRI models, such as T1 and T2 mapping.

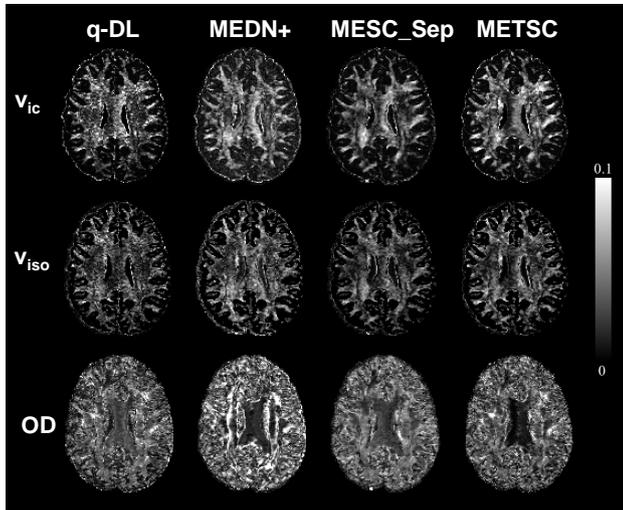


Fig. 12. Evaluation of estimation precision on four learning-based algorithms. The maps showed the standard deviation of estimated parameters between bootstrapped results with shuffled diffusion directions. The METSC (ours) method showed higher precision than the other model-based methods MEDN+, and MESC_Sep, indicated by the lower standard deviation from bootstrap.

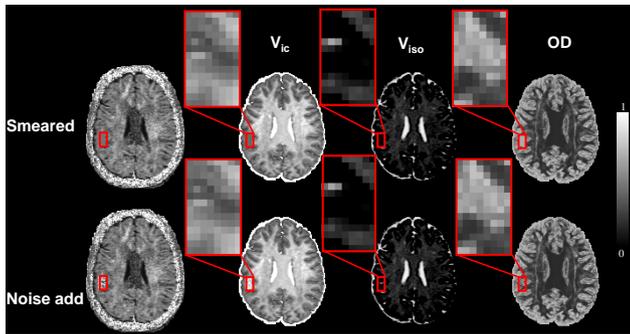


Fig. 13. Evaluation of robustness by testing whether the METSC (ours) method will generate fake output that is not supported by the data. The signal patch in the red rectangle is smeared (a) or noised corrupted (b). Both results demonstrated the network will not produce the unsupported estimation.

4.2. Model-bias

A key contribution of this work is that we brought up and validated the hypothesis that the model-bias can partially solve the data-hungry of the ViT. Compared with the ViT structure, incorporating the model-bias into METSC significantly improved its performance with the small amount of training samples. As demonstrated in the experiments (Section 3.1.3 and Section 3.5.1), the proposed METSC framework was compared with the other networks, with no more than 0.3M training samples in the IVIM model and 1.5M in the NODDI model. Because the nonlinearity of NODDI is much higher than the IVIM model, it is expected that NODDI needs more data to learn the dictionary.

Meanwhile, since the model-bias is a kind of sparse representation, the network could be examined by the sparsity of the representation signals. For instance, using the NODDI model, Fig. 14 showed the distribution of nonzero entries in the dictionary coefficients for a test subject and the sparsity was over 84%, which supported the validity of our sparsity hypothesis.

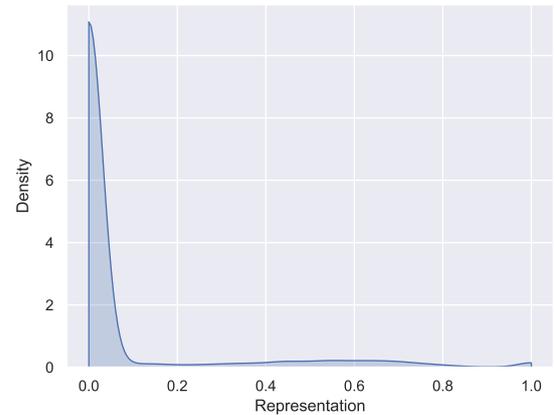


Fig. 14. The distribution of dictionary coefficients in the sparse representation x given by METSC for a representative test subject, indicating high sparsity in the dictionary coefficients.

4.3. Network setup

To access the architecture of the METSC framework, various ablation experiments on hyper-parameters and the structure of METSC were investigated. The METSC framework contains a large number of hyper-parameters. Therefore, we only focused on the major hyper-parameters in the iterative decoder phase, while the hyper-parameters in the Transformer-based encoder were fixed as (Dosovitskiy et al., 2020). We further investigated the encoder and decoder performance with respect to input patch size and dictionary size, which were not fully addressed in previous studies. In terms of data input, we explored the effect of patch size, that has not been investigated before and determined an optimal patch size of 3 could give the lowest loss and best computational efficiency. We also tested the dictionary size N of the decoder on the IVIM model and found that the results for $N < 400$ were the same as (Ye et al., 2020), which however, did not explore $N > 400$. In this work, we tested the whole spectrum of dictionary size from 200-800 and determined dictionary size of 600 was optimal for both accuracy and

efficiency.

Note that in the NODDI experiments, compared with the previous model-based network MESC_Sep (Ye et al., 2020), METSC did not use LSTM to incorporate the historical information. Although incorporating historical information can improve the accuracy of estimation, we found it made the network unstable e.g., a slight change in the diffusion direction may reduce the fitting accuracy and it is known the preprocessing steps of registration and motion correction could easily change the actual gradients. This can be partially compensated by interpolating the diffusion directions to match the target dataset (Qin et al., 2020). The results with the interpolated diffusion directions were shown in Fig 15(a c), which showed the performance of MESC_Sep after interpolation on shore basis (Merlet and Deriche, 2013) were improved with this strategy but still cannot beat METSC. Combining the IVIM and NODDI results in Fig. 12, Fig. 15, Table A1, and Table 6, it can be concluded that our proposed method has the best robustness against both the directions and magnitudes (b-value) of diffusion gradients.

4.4. Future work

Training of the learning-based network requires densely sampled diffusion signals with high quality (Golkov et al., 2016; Ye, 2017a,b; Ye et al., 2019, 2020; Chen et al., 2020), which are often difficult to obtain in practice. Recent studies have demonstrated the pre-trained model can help the microstructural estimation using an auxiliary dataset (Li et al., 2021). Similar the great success of the pre-trained Transformer model in NLP (Brown et al., 2020; Devlin et al., 2018; Radford et al., 2018), our pre-trained METSC can probably transfer the knowledge to other domains that are not limited to the dMRI models but also T1 mapping, T2 mapping, other multi-pool models, which will be investigated in future work.

5. Summary and Conclusion

In this work, we proposed a novel model-driven Transformer with sparse coding to estimate microstructural parameters in dMRI models with reduced q-space data. The proposed METSC framework integrated the strength of the Trans-

former and also address the large training data requirement of ViT by introducing model-based inductive bias. Compared with the conventional optimization methods and the state-of-the-art learning-based methods, METSC achieved the highest accuracy in estimating model parameters for both IVIM and NODDI models. The network also showed good interpretability, generalizability and robustness, and thus, is potentially useful for fast dMRI acquisition with undersampled q-space, which may be particularly important for dMRI of the moving subjects.

Acknowledgments

This work is supported by Ministry of Science and Technology of the People’s Republic of China (2018YFE0114600), National Natural Science Foundation of China (61801424, 81971606, 82122032), and Science and Technology Department of Zhejiang Province (202006140, 2022C03057).

Appendix

The following Appendix describes the comparison of three supervised learning based methods in estimating IVIM model parameters using different combinations of b-values, in terms of MSE.

Table A1. Comparison of three supervised learning based methods in estimating IVIM model parameters using different combinations of b-values, in terms of MSE. The combinations with top performance were highlighted in bold.

		f ($\times 10^{-4}$)	D ($\times 10^{-4} \mu\text{m}^2/\text{ms}$)	D^* ($\times 10^{-2} \mu\text{m}^2/\text{ms}$)
3 b-values (20, 150, 500) s/mm^2	q-DL	2.3	12	27
	SCDNN	2	37	14
	METSC	0.41	2.8	8.4
5 b-values Comb1 (20, 50, 150, 300, 500) s/mm^2	q-DL	0.68	5.1	38
	SCDNN	0.072	3.1	1.8
	METSC	0.063	2.2	1.4
5 b-values Comb2 (20, 50, 150, 200, 500) s/mm^2	q-DL	4	4.6	22
	SCDNN	7.1	2.1	6.2
	METSC	0.14	2.1	1.8
5 b-values Comb3 (20, 50, 200, 300, 500) s/mm^2	q-DL	0.39	8.8	45
	SCDNN	0.4	2.1	9
	METSC	0.063	5.6	1.7
5 b-values Comb4 (20, 100, 150, 300, 500) s/mm^2	q-DL	0.43	8.8	46
	SCDNN	0.27	9	2.2
	METSC	0.097	1.7	1.7
5 b-values Comb2 (20, 50, 150, 200, 500) s/mm^2	q-DL	4.1	5	29
	SCDNN	0.19	3.2	2.7
	METSC	0.094	1.7	2
7 b-values (20, 50, 100, 150, 200, 300, 500) s/mm^2	q-DL	0.37	9.7	2.8
	SCDNN	0.1	3.2	2.8
	METSC	0.086	1.4	2.6

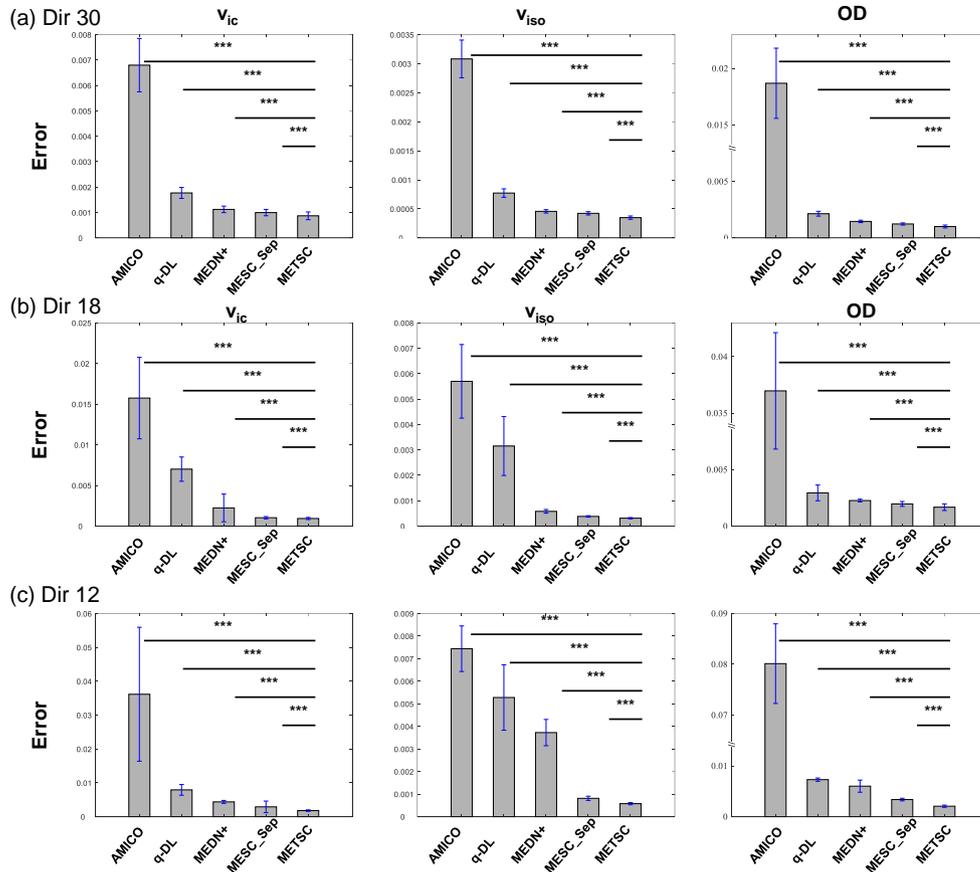


Fig. 15. The means and standard deviations of the whole-brain average estimation errors of NODDI parameters in test subjects ($n=21$) using different estimation algorithms at downsampled diffusion directions of 30, 18, and 12 per shell at b-values of 1000 and 2000 s/mm^2 . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ by paired t-test.

References

- Alexander, D.C., Zikic, D., Ghosh, A., Tanno, R., Wotschel, V., Zhang, J., Kaden, E., Dyrby, T.B., Sotiropoulos, S.N., Zhang, H., et al., 2017. Image quality transfer and applications in diffusion mri. *NeuroImage* 152, 283–298.
- Alexander, D.C., Zikic, D., Zhang, J., Zhang, H., Criminisi, A., 2014. Image quality transfer via random forest regression: applications in diffusion mri, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 225–232.
- Arun, K., Huang, T., Blostein, S., 1987. Least-squares fitting of 2 3-d point set. *IEEE Trans. Pattern Anal. Mach. Intell* 9, 699–700.
- Assaf, Y., Blumenfeld-Katzir, T., Yovel, Y., Basser, P.J., 2008. Axcaliber: a method for measuring axon diameter distribution from diffusion mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 59, 1347–1354.
- Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Barbieri, S., Gurney-Champion, O.J., Klaassen, R., Thoeny, H.C., 2020. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted mri. *Magnetic resonance in medicine* 83, 312–321.
- Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al., 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.
- Blumensath, T., Davies, M.E., 2009. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis* 27, 265–274.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901.
- Chen, G., Hong, Y., Zhang, Y., Kim, J., Huynh, K.M., Ma, J., Lin, W., Shen, D., Yap, P.T., Consortium, U.B.C.P., et al., 2020. Estimating tissue microstructure with undersampled diffusion data via graph convolutional neural networks, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 280–290.
- Daducci, A., Canales-Rodríguez, E.J., Zhang, H., Dyrby, T.B., Alexander, D.C., Thiran, J.P., 2015. Accelerated microstructure imaging via convex optimization (amico) from diffusion mri data. *Neuroimage* 105, 32–44.
- Daducci, A., Van De Ville, D., Thiran, J.P., Wiaux, Y., 2014. Sparse regularization for fiber odF reconstruction: from the suboptimality of ℓ_2 and ℓ_1 priors to ℓ_0 . *Medical Image Analysis* 18, 820–833.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Federau, C., O’Brien, K., Meuli, R., Hagmann, P., Maeder, P., 2014. Measuring brain perfusion with intravoxel incoherent motion (ivim): initial clinical experience. *Journal of Magnetic Resonance Imaging* 39, 624–632.
- Gibbons, E.K., Hodgson, K.K., Chaudhari, A.S., Richards, L.G., Majersik, J.J., Adluru, G., DiBella, E.V., 2019. Simultaneous noddI and gfa parameter map generation from subsampled q-space imaging using deep learning. *Magnetic resonance in medicine* 81, 2399–2411.
- Golkov, V., Dosovitskiy, A., Sperl, J.I., Menzel, M.I., Czisch, M., Sämann, P., Brox, T., Cremers, D., 2016. Q-space deep learning: twelve-fold shorter and model-free diffusion mri scans. *IEEE transactions on medical imaging* 35, 1344–1351.
- Gregor, K., LeCun, Y., 2010. Learning fast approximations of sparse coding, in: *Proceedings of the 27th international conference on international conference*

- on machine learning, pp. 399–406.
- Gustafsson, O., Montelius, M., Starck, G., Ljungberg, M., 2018. Impact of prior distributions and central tendency measures on bayesian intravoxel incoherent motion model fitting. *Magnetic Resonance in Medicine* 79, 1674–1683.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- Jalnefjord, O., Andersson, M., Montelius, M., Starck, G., Elf, A.K., Johanson, V., Svensson, J., Ljungberg, M., 2018. Comparison of methods for estimation of the intravoxel incoherent motion (ivim) diffusion coefficient (d) and perfusion fraction (f). *Magnetic Resonance Materials in Physics, Biology and Medicine* 31, 715–723.
- Jiang, X., Li, H., Xie, J., Zhao, P., Gore, J.C., Xu, J., 2016. Quantification of cell size using temporal diffusion spectroscopy. *Magnetic resonance in medicine* 75, 1076–1085.
- Kaden, E., Kruggel, F., Alexander, D.C., 2016. Quantitative mapping of the per-axon diffusion coefficients in brain white matter. *Magnetic resonance in medicine* 75, 1752–1763.
- Kärger, J., Pfeifer, H., Heink, W., 1988. Principles and application of self-diffusion measurements by nuclear magnetic resonance, in: *Advances in Magnetic and optical resonance*. Elsevier, volume 12, pp. 1–89.
- Karimi, D., Jaimes, C., Machado-Rivas, F., Vasung, L., Khan, S., Warfield, S.K., Gholipour, A., 2021. Deep learning-based parameter estimation in fetal diffusion-weighted mri. *NeuroImage* 243, 118482.
- Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. *Nature Reviews Physics* 3, 422–440.
- Koppers, S., Bloy, L., Berman, J.I., Tax, C.M., Edgar, J.C., Merhof, D., 2019. Spherical harmonic residual network for diffusion signal harmonization, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 173–182.
- Le Bihan, D., 2019. What can we see with ivim mri? *Neuroimage* 187, 56–67.
- Le Bihan, D., Breton, E., Lallemand, D., Aubin, M., Vignaud, J., Laval-Jeantet, M., 1988. Separation of diffusion and perfusion in intravoxel incoherent motion mr imaging. *Radiology* 168, 497–505.
- Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H., 2001. Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 13, 534–546.
- Li, Y., Qin, Y., Liu, Z., Ye, C., 2021. Pretraining improves deep learning based tissue microstructure estimation, in: *Computational Diffusion MRI*. Springer, pp. 173–185.
- Liang, D., Cheng, J., Ke, Z., Ying, L., 2019. Deep mri reconstruction: Unrolled optimization algorithms meet neural networks. *arXiv preprint arXiv:1907.11711*.
- Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems* 29.
- Merlet, S.L., Deriche, R., 2013. Continuous diffusion signal, eap and odf estimation via compressive sensing in diffusion mri. *Medical image analysis* 17, 556–572.
- Mori, S., Zhang, J., 2006. Principles of diffusion tensor imaging and its applications to basic neuroscience research. *Neuron* 51, 527–539.
- Nedjati-Gilani, G.L., Schneider, T., Hall, M.G., Wheeler-Kingshott, C.A., Alexander, D.C., 2014. Machine learning based compartment models with permeability for white matter microstructure imaging, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 257–264.
- Neil, J.J., Bretthorst, G.L., 1993. On the use of bayesian probability theory for analysis of exponential decay date: an example taken from intravoxel incoherent motion experiments. *Magnetic resonance in medicine* 29, 642–647.
- Novikov, D.S., Fieremans, E., Jespersen, S.N., Kiselev, V.G., 2019. Quantifying brain microstructure with diffusion mri: Theory and parameter estimation. *NMR in Biomedicine* 32, e3998.
- Palombo, M., Ianus, A., Guerreri, M., Nunes, D., Alexander, D.C., Shemesh, N., Zhang, H., 2020. Sandi: a compartment-based model for non-invasive apparent soma and neurite imaging by diffusion mri. *NeuroImage* 215, 116835.
- Qin, Y., Li, Y., Liu, Z., Ye, C., 2020. Knowledge transfer between datasets for learning-based tissue microstructure estimation, in: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, IEEE. pp. 1530–1533.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training.
- Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al., 2013. The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems* 30.
- Wang, N., Sun, J., 2020. Model meets deep learning in image inverse problems. *learning* 2, 10.
- Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803.
- Wang, Y., Wang, Q., Haldar, J.P., Yeh, F.C., Xie, M., Sun, P., Tu, T.W., Trinkaus, K., Klein, R.S., Cross, A.H., et al., 2011. Quantification of increased cellularity during inflammatory demyelination. *Brain* 134, 3590–3601.
- Xu, Z., Sun, J., 2018. Model-driven deep-learning. *National Science Review* 5, 22–24.
- Yang, Y., Sun, J., Li, H., Xu, Z., 2018. Admm-csnet: A deep learning approach for image compressive sensing. *IEEE transactions on pattern analysis and machine intelligence* 42, 521–538.
- Ye, C., 2017a. Estimation of tissue microstructure using a deep network inspired by a sparse reconstruction framework, in: *International Conference on Information Processing in Medical Imaging*, Springer. pp. 466–477.
- Ye, C., 2017b. Tissue microstructure estimation using a deep network inspired by a dictionary-based framework. *Medical image analysis* 42, 288–299.
- Ye, C., Li, X., Chen, J., 2019. A deep network for tissue microstructure estimation using modified lstm units. *Medical image analysis* 55, 49–64.
- Ye, C., Li, Y., Zeng, X., 2020. An improved deep network for tissue microstructure estimation with uncertainty quantification. *Medical image analysis* 61, 101650.
- Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C., 2012. Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* 61, 1000–1016.
- Zheng, T., Sun, C., Wang, G., Zheng, W., Shi, W., Sun, Y., Zhang, Y., Ye, C., Wu, D., 2021. A model-driven deep learning method based on sparse coding to accelerate ivim imaging in fetal brain, in: *ISMRM 2021: The 29th International Society for Magnetic Resonance in Medicine*.