Deep Learning Method with Integrated Invertible Wavelet Scattering for Improving the Quality of In Vivo Cardiac DTI

Zeyu Deng, Lihui Wang, Zixiang Kuai, Qijian Chen, Chen Ye, Andrew Scott, Sonia Nielles-Vallespin, Yue-Min Zhu

Deep Learning Method with Integrated Invertible Wavelet Scattering for Improving the Quality of In Vivo Cardiac DTI

Zeyu Deng[1],Lihui Wang[1,*],Zixiang Kuai[2],Qijian Chen[1],Chen Ye[1],Andrew D. Scott[3],Sonia Nielles-Vallespin[3],Yuemin Zhu[4]

1 Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis of Guizhou Province, College of Computer Science and Technology, State Key Laboratory of Public Big Data,Guizhou University,Guiyang,China
2 Imaging Center, Harbin Medical University Cancer Hospital,Harbin,China
3 CMR Unit, Royal Brompton Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK and National Heart and Lung Institute, Imperial College London, UK
4 University Lyon, INSA Lyon, CNRS, Inserm, IRP Metislab CREATIS UMR5220, U1206, Lyon 69621, France
* Author to whom any correspondence should be addressed.

E-mail: lhwang2@gzu.edu.cn, wlh1984@gmail.com

January 2024

Abstract.

Respiratory motion, cardiac motion and inherently low signal-to-noise ratio (SNR) are major limitations of *in vivo* cardiac diffusion tensor imaging (DTI). We propose a novel enhancement method that uses unsupervised learning based invertible wavelet scattering (IWS) to improve the quality of *in vivo* cardiac DTI. It starts by extracting nearly transformation-invariant features from multiple cardiac diffusion-weighted (DW) image acquisitions using multi-scale wavelet scattering (WS). Then, the relationship between the WS coefficients and DW images is learned through a multi-scale encoder and a decoder network. Using the trained encoder, the deep features of WS coefficients of multiple DW image acquisitions are further extracted and then fused using an average rule. Using the fused WS features and trained decoder, the enhanced DW images are finally derived. We evaluate the performance of the proposed method by comparing it with several methods on three *in vivo* cardiac DTI datasets in terms of SNR, contrast to noise ratio (CNR), fractional anisotropy (FA), mean diffusivity (MD) and helix angle (HA). Comparing against the best comparison method, SNR/CNR of diastolic, gastric peristalsis influenced, and end-systolic DW images were improved by 1%/16%, 5%/6%, and 56%/30%, respectively. The approach also yielded consistent FA and MD values and more coherent helical fiber structures than the comparison methods used in this work. In addition, the ablation results verify that using the transformation-invariant and noise-robust wavelet scattering features enables us to effectively explore the useful information from the limited data, providing a potential mean to alleviate the dependence of the fusion results on the number of repeated acquisitions, which is beneficial for dealing with the issues of noise and residual motion simultaneously and therefore improving the quality of *in vivo* cardiac DTI.

keywords: *In vivo* cardiac DTI, wavelet scattering, image fusion, convolutional neural network

1. Introduction

Diffusion tensor imaging (DTI) is a promising technique for investigating noninvasively the microstructure of both normal and diseased hearts [1–5]. It has been demonstrated that several cardiac diseases such as myocardial infarction, hypertrophic cardiomyopathy and accurate ischemia are highly related to the change of diffusion metrics in DTI [6–12], showing that DTI may provide meaningful imaging biomarkers for early diagnosis of cardiac diseases. However, DTI of *in vivo* hearts is sensitive to motions, not only can breathing or/and respiratory motions cause significant signal loss in diffusion weighted (DW) images, but some physiologic motions, such as gastric peristalsis, can also cause bulk motion artifacts, thus making it difficult to explore exactly the microstructure of *in vivo* hearts. Therefore, reducing the influence of motions on *in vivo* cardiac DTI is important for research.

Currently, two main kinds of imaging sequences are used for *in vivo* cardiac DTI, which are simulated echo acquisition mode (STEAM) sequence [13] and motion compensated spin echo (MCSE) sequence [14], respectively. STEAM assumes that the heart position and cardiac motion state keep unchanged during two cardiac cycles, in this case, distributing the diffusion encoding/decoding gradients over two consecutive heart beats enables it to minimize the effects of cardiac bulk motion. Since STEAM does not require high-performance gradient hardware, it has been widely used in both healthy and diseased hearts. However, due to the long acquisition time, it suffers from low signal-to-noise ratio (SNR). In addition, even with breath-holding, the influence of cardiac strain on the diffusion weighted signal is still not avoidable [15]. MCSE is an alternative to STEAM, which makes it possible to perform in vivo cardiac DTI during free breathing with the assistance of the first, second or higher order motion compensated diffusion gradient waveforms. The MCSE sequence can produce higher SNR than STEAM. However, since the heart motion is complex, namely the velocity or acceleration of the motion during the diffusion encoding is not constant, the signal attenuation caused by residual motion still exists. Despite the advances in sequence design, achieving a high quality *in vivo* cardiac DTI remains a challenge due to low SNR and the residual motion.

To deal with this issue, postprocessing-based methods for *in vivo* cardiac DTI have been proposed. For instance, PCATMIP integrates the principal component analysis (PCA) and temporal maximum intensity projection (TMIP) methods to recover signal loss in DW images caused by breath motion from multiple acquisitions [16]. WIF uses the wavelet transform to recover lost signals and remove noise by fusing the effective information extracted from multiple DW images under free breathing [17]. With the emergence of deep learning models, using the network to fuse the images for promoting the quality has received intensive attention [18–20]. In the field of cardiac

DTI, Ferreira et al. proposed an automatic *in vivo* cardiac diffusion tensor post-processing framework, in which DW images are firstly denoised and segmented and then fused by registration with UNet. However, due to the varying contrast and the intrinsically low SNR of DW images, registration-based fusion may introduce additional errors in the tensor estimation [21]. To address this problem, Weine et al. proposed a parameterized pipeline to generate synthetic *in vivo* cardiac DW images, and then used the synthetic data to train a residual convolutional neural network for fusing DW images from multiple acquisitions without the prerequisite of registration [22]. As ground-truth about diffusion tensors of *in vivo* hearts is not available, the above-mentioned supervised learning-based methods are not feasible in practice. Accordingly, unsupervised learning-based methods were proposed. For example, Xu et al. [23] presented a U2Fusion network to fuse magnetic resonance images and PET images in an unsupervised learning manner, which can adaptively preserve the useful information of different source images by training the network with structural similarity and continue learning losses [23]. Jung et al. proposed a DIF-NET, in which the intensity fidelity and structure tensor losses are combined to allow the fused image to preserve the overall contrast of different inputs [24]. These fusion based methods have potential to deal with the signal loss or noise problems of in vivo cardiac DW images.

Besides the fusion based methods, several denoising methods dedicated to the diffusion MRI based on single acquisition have also been proposed, such as MPPCA [25, 26], Patch2Self [27] and DDM2 [28]. MPPCA is currently the best traditional DW image denoising method. It first maps the multi-directional DW images onto a set of principal component orthogonal basis, and then uses the properties that the DW signals along multiple directions have a low rank and the eigenvalues of noise conform to the Marchenko-Pastur distribution to threshold the principal components for denoising. When the noise level is higher or the number of diffusion gradient directions is fewer, the low rank property of the DW signal and the noise distribution are not satisfied, thereby its denoising performance decreases. In Patch2Self, it uses patches of DW images along other diffusion gradient directions to fit the DW signal of a certain pixel in the current direction. Since the noise along different directions are not correlated, Patch2Self can denoise using a fitting method. However, when there are fewer gradient directions, the denoising performance of Pacth2Self decreases significantly. DDM2 is a kind of generative method based on diffusion model to denoise DW images, it involves three steps of noise estimation, forward diffusion state matching and reverse diffusion reconstruction. This multi-stage learning method is easily affected by multiple factors and may generate false structures in the denoised images, especially when the sample size is not large enough. The above mentioned methods are designed to denoise, their potential in signal loss compensation is unknown. How to deal with signal loss caused by motion and the noise influence simultaneously in diffusion MRI is still not well explored.

In view of the noise robustness of wavelet scattering and its superiority in multi-scale image decompositions without signal loss, as well as the merits of deep learning for feature extraction and fusion, we propose an invertible wavelet scattering fusion

method based on multi-scale convolutional neural network (WS-MCNN) to obtain high quality *in vivo* human cardiac DW images. The proposed method first calculates the nearly transformation-invariant (hereinafter referred to as transformation-invariant for simplicity) coefficients of *in vivo* cardiac DW images using multi-scale wavelet scattering. Then the deep features of these multi-scale wavelet scattering coefficients of multiple acquisitions are extracted with an encoder and then fused based on an average fusion rule. From the fused feature maps, inverse wavelet scattering transform via CNN is finally performed to recover the high quality images. To evaluate the effectiveness of WS-MCNN, we trained it with systole cardiac DTI acquired from one site, and tested on three datasets acquired from other sites, including end-systolic and end-diastolic cardiac DTI, as well as cardiac DTI affected by gastric peristalsis.

2. Methods

2.1. Basic principle and properties of wavelet scattering

Wavelet transform is a common method for extracting image features via the following transformation

$$Wx = \begin{bmatrix} x*\psi_{j,r}(u) \\ x*\varphi_j(u) \end{bmatrix}, \tag{1}$$

where $*$ is the convolution operator, $\varphi_j(u)$ is a scaling function, and $\psi_{j,r}(u)$ is a direction wavelet function. The $x * \varphi_j(u)$ represents the low-frequency information of image x at the scale j, and $x * \psi_{j,r}(u)$ the high-frequency information at the scale j and in the direction r.

Although the wavelet transform can restore the details of the image, it allows obtaining only high-frequency components in three directions (vertical, diagonal, longitudinal) and does not have translation invariance due to convolution operation of wavelet. To deal with these issues, wavelet scattering was proposed by Mallat [29], which which allows us to extract transformation-invariant and noise-robust features at multi-scale without information loss. It can be expressed as

$$\widetilde{Wx} = \begin{bmatrix} U_m \\ S_m \end{bmatrix}, \tag{2}$$

where $\widetilde{Wx}$ means the wavelet scattering transform for image x. The subscript m designates the wavelet scattering level that indicates the number of decomposition operations on the high-frequency information. U_m represents the scattering propagation operator and S_m the scattering coefficients, which are calculated as

$$\begin{aligned} U_m &= \begin{cases} x & m = 0 \\ \left\{U^m_{j,r},\ {}_{j=1,2,\ldots,J,\ r=1,2,\ldots,L}\right\} & m \geq 1 \end{cases}, \\ S_m &= U_m * \varphi_J(u) \end{aligned} \tag{3}$$

where J indicates the maximum wavelet decomposition scale, L the number of directions for directional wavelet transform, $\varphi_J(u)$ is a low-pass filter at the maximum scale J,

formulated as $\varphi_J(u) = 2^{-2J}\varphi(2^{-J}u)$, and $\left\{U_{j,r}^m,\ _{j=1,2,\ldots,J,\ r=1,2,\ldots,L}\right\}$ is the set of scatter propagation operators at m-th scattering level. When $m \geq 1$, $U_{j,r}^m$ is formulated as:

$$U_{j,r}^m = \left|\left|\left|x * \psi_{j,r}^1(u)\right| * \psi_{j,r}^2(u)\right| \cdots * \psi_{j,r}^m(u)\right| \tag{4}$$

where $|\cdot|$ represents the modulus operation, $\psi_{j,r}^m(u)$ a directional wavelet function at the scale j in the direction r used in the m^{th} scattering level. For any m, the directional wavelet function is the same and expressed as

$$\psi_{j,r}^1(u) = \psi_{j,r}^2(u) = \cdots = \psi_{j,r}^m(u) = \psi_{j,r}(u) = 2^{-2j}\psi_r(2^{-j}u). \tag{5}$$

where $\psi_r(\cdot)$ indicates implementing the function $\psi(\cdot)$ along the direction r. In Equations (3) and (5), $\psi(\cdot)$ and $\varphi(\cdot)$ can be any wavelet function and scaling function used in wavelet transform. From Equation (4), we can summarize that the scattering coefficients are calculated by smoothing the modulus of high-frequency information decomposed at different scattering levels.

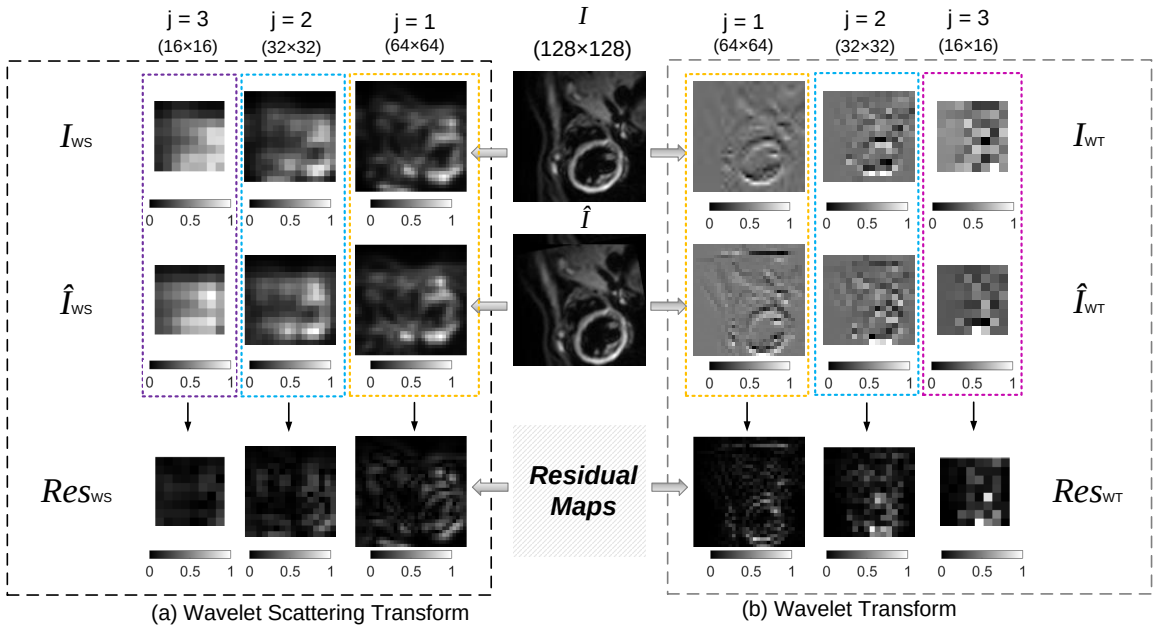

Figure 1. Comparing the multi-scale features derived from wavelet scattering and wavelet transform for the images before and after deformation. I and $\hat{I}$ represent the original and deformed image, respectively. I_{WS} and $\hat{I}_{\mathrm{WS}}$ are the wavelet scattering coefficient maps at three scales for images I and $\hat{I}$, while I_{WT} and $\hat{I}_{\mathrm{WT}}$ are the corresponding wavelet decomposition coefficient maps. Res_{WS} and Res_{WT} show respectively the residuals of wavelet scattering coefficient maps and wavelet decomposition coefficient maps at three scales before and after the transformation.

To gain a more intuitive understanding of the translation- and rotation-invariant of WS coefficients, Figure 1 compares the changes of wavelet transform coefficients and those of WS coefficients at different scales when the image deformed slightly. Here, I represents the original image, and $\hat{I}$ the image obtained by rotating I clockwise by

10 degrees and then shifting it downward by 10 pixels. We performed discrete wavelet decomposition and WS decomposition on both I and $\hat{I}$ at three scales, and computed the residuals of the coefficient maps of I and $\hat{I}$. It can be observed that the residual maps obtained from WS have smaller values than those obtained from the discrete wavelet transform when $j = 2$ and $j = 3$, indicating the stability of WS to rotation and translation.

2.2. Principle of WS-MCNN for improving the quality of in vivo cardiac DTI

The overall workflow of the proposed WS-MCNN for improving the quality of *in vivo* cardiac DTI is depicted in Figure 2. During the training stage, the cardiac DW images were used as the input of WS-MCNN. Then, the multi-scale WS coefficient maps of the input DW images were extracted using wavelet scattering. After that, both WS coefficient maps and original DW images were fed into the Encoder of WS-MCNN to further extract deep features. Finally, the deep features were concatenated and input into the Decoder of WS-MCNN to reconstruct the original DW images. During the fusion or test stage, the *in vivo* cardiac DW images acquired with multiple trigger delays (TDs) or multiple repetitions, and the corresponding wavelet scattering coefficient maps were respectively input into the trained Encoder of WS-MCNN. Following that, the deep features of multiple DW images extracted by Encoder are fused based on an average fusion rule. Finally, the fused features were input into a trained Decoder of WS-MCNN to generate the final fused image with high quality.

In the following subsections, the feature extraction with wavelet scattering, the WS-MCNN training and the image fusion with trained WS-MCNN will be described in detail.

2.2.1. Extracting WS coefficient maps In the present study, the maximum scattering level $m = 1$, wavelet decomposition scales J =1, 2 and 3, respectively, and the number of directions $L = 10$. When $m = 0$, wavelet scattering outputs three transformation-invariant coefficient maps $S_0 = \left\{S_0^{J=1}, S_0^{J=2}, S_0^{J=3}\right\}$, with $S_0^{J=1} = x * \phi_1(u)$, $S_0^{J=2} = x * \phi_2(u)$ and $S_0^{J=3} = x * \phi_3(u)$; when $m = 1$, using Equation (3) and (4) results in 10 scattering coefficients $S_1^{J=1}$, 20 scattering coefficients $S_1^{J=2}$, and 30 scattering coefficients $S_1^{J=3}$. Thus, if setting the maximum scattering level as $m = 1$, 11 scattering coefficient maps will be obtained when the wavelet decomposition scale is $J = 1$, 21 coefficients when $J = 2$, and 31 coefficients when $J = 3$. In other words, wavelet scattering yields a total of 63 multi-scale deformation-invariant and noise-robust coefficient maps, in which, not only low-frequency information extracted by traditional wavelet transformation are included, but also low-frequency components encoded in high-frequency information along multiple directions are taken into account. This provides more useful texture features for image fusion. If the original input image size is $w \times h$, the size of scattering coefficient maps at scales of $J = 1$, $J = 2$ and $J = 3$ is $\frac{w}{2} \times \frac{h}{2}$, $\frac{w}{4} \times \frac{h}{4}$, and $\frac{w}{8} \times \frac{h}{8}$, respectively.

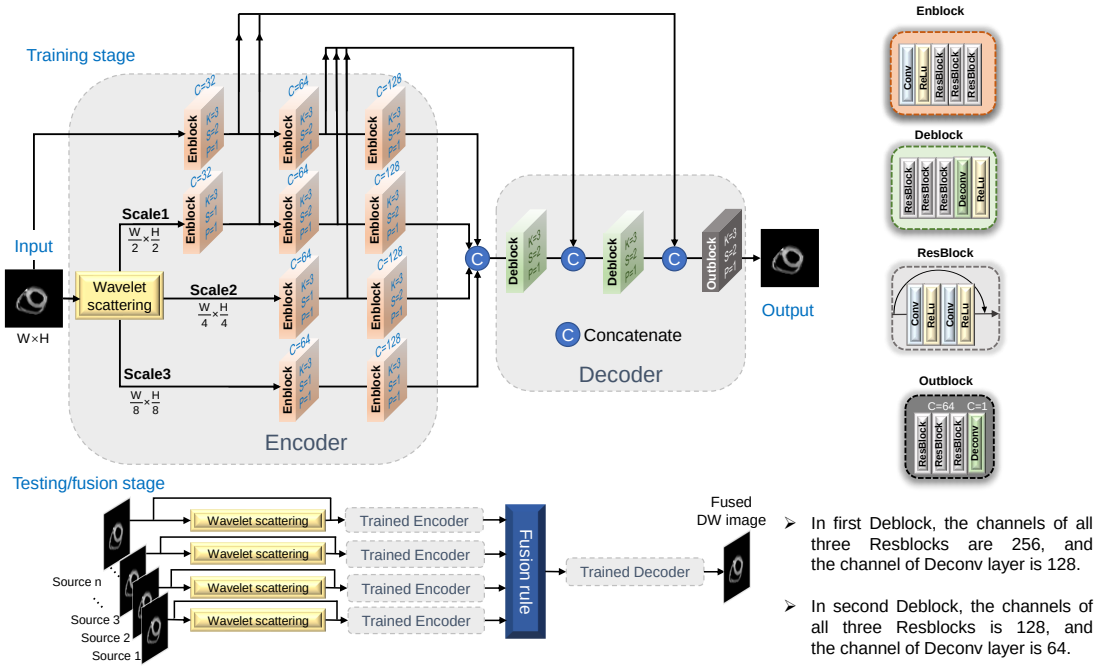

Figure 2. Overall workflow of the proposed WS-MCNN for improving the quality of in *in vivo* cardiac DTI. During the training stage, WS-MCNN uses an encoder-decoder architecture to learn the feature representations of a given DW image. The input and target of WS-MCNN during the training is the same DW image. During the test stage, the multiple acquisitions of one subject are input into the trained encoder to extract their features; these features are then fused with an average rule. The fused features are finally input into the trained decoder to obtain the high-quality DW image.

2.2.2. High-quality DW image reconstruction based on WS-MCNN Image fusion-based quality improvment is usually implemented by feature fusion and feature-to-image reconstruction. Since the features used in this work are derived from WS that is not exactly invertible [30], we propose to use WS-MCNN to achieve the inverse wavelet scattering transform by mapping feature maps into original DW images.

As shown in Figure 2, WS-MCNN consists of an encoder and a decoder. The encoder adopts the idea of Laplacian pyramid [31], which processes the inputs with different sizes through four different streams. The first stream is composed of three Enblocks (the stride of the Conv layer is 2), which is responsible for extracting features from original DW images. The second stream is also composed of three Enblocks, but the strides for three Conv layers are 1, 2 and 2 respectively; it is used for extracting features from the 1^{st} scale wavelet scattering coefficients. The third stream is composed of two Enblocks (the strides for two Conv layers are 1 and 2 respectively); it extracts features from the 2^{nd} scale wavelet scattering coefficients. The fourth stream is also composed of two Enblocks with strides of Conv layers equal to 1; it is used to extract deep features from the 3^{rd} scale wavelet scattering coefficients. Such pyramid-like encoder enables us to extract more diverse texture information at multi-scales. These multi-scale deep

features are concatenated and input into the decoder. The latter is composed of 2 Deblocks and 1 Outblock, which are employed to double the spatial size of feature maps and reduce the number of channels gradually. Notice that, in our network, the kernel size of convolutions in all the layers is set as 3, and the channel number is noted above each block.

Once the WS-MCNN was trained, we first extracted the deep features of the DW images acquired with multiple TDs or multiple repetitions along a given gradient direction using the trained encoder of WS-MCNN. Subsequently, the fused feature maps were obtained by averaging the encoded feature maps derived from multiple acquisitions (average fusion rule). Finally, these fused feature maps were fed into the trained decoder of WS-MCNN, yielding the desired DW image, which is recalibrated by a linear intensity transform to recover its original intensity range.

3. Experiments

3.1. In vivo cardiac datasets and preprocessings

3.1.1. Training dataset End-systolic cardiac DTI of 6 healthy volunteers and early systolic cardiac DTI of 8 healthy volunteers were downloaded from `https://med.stanford.edu/cmrgroup/data/myofiber_data.html` [32]. The DW images were acquired on 3T MRI scanner (Prisma, Siemens). The second-order (M1-M2) motion compensated diffusion encoding gradients were incorporated into the spin-echo echo-planar imaging (SE-EPI) to minimize signal dropout induced by cardiac motion. To precisely determine which cardiac phases were suited for cardiac diffusion imaging, a prospective trigger delay (TD) scout acquisition was implemented. At each trigger TD, 12 diffusion directions with a b-value of 350 s/mm^2 were acquired. For each subject and a given diffusion gradient direction, the acquisition was repeated 5$\sim$ 8 times for one mid-ventricular short-axis slice during diastole. The acquisition parameters are: TE/TR=61/120 ms, spatial resolution=$1.6 \times 1.6 \times 8$ mm^3, acceleration rate=2 (GRAPPA), partial Fourier=6/8, and matrix size = 128×104. A total of 1152 DW images (not including images with b=0) were used for training WS-MCNN.

3.1.2. Testing dataset We used three kinds of datasets to test the proposed method, detailed as follows.

(1) End-diastolic cardiac DTI of 6 healthy volunteers downloaded also from `https://med.stanford.edu/cmrgroup/data/myofiber_data.html` was used as a test dataset. The acquisition parameters were the same as those in the training set but with different trigger delays. A total of 504 DW images were included.

(2) End-systolic cardiac DW images of 10 subjects acquired from the Royal Brompton Hospital of London in UK were used as another test set. They were acquired using a Skyra 3T MRI scanner (Siemens AG) with STEAM echo planar imaging (STEAM-EPI) sequence under breath hold. DTI of a short-axis slice in the mid-left

ventricle was performed with prescribed b-values of 50, 150, 350, 550, 750, and 950 s/mm^2 along 6 diffusion gradient directions (8 repetitions). The acquisition parameters are as follows: TR = 2 cardiac cycles, FOV =360×135 mm^2, in-plane spatial resolution = 2.8×2.8 mm^2, slice thickness =8 mm, matrix size = 256×96. When b-value was 950 s/mm^2, TE=24 ms, the diffusion gradient magnitude was 35 mT/m, ramp time was 660 μs and the flat top time was 1680 μs. Lower b-values were achieved by reducing the gradient magnitude while keeping the other timing parameters unchanged. Note that, all frames were examined visually to identify and reject frames corrupted by motion, the detailed acquisition strategies can be found in the work of Scott et al. [33]. This dataset contains a total of 1560 selected DW images.

(3) To deal with the influence of gastric peristalsis on cardiac DTI while minimizing the effects of cardiac and respiratory motions, we also used the same dataset as in the work of [34] for testing, which were acquired on a 3T MRI scanner (Philips Ingenia system) from Harbin Medical University Cancer Hospital in China with trigger delay at end-diastolic. One midventricular short-axis slice of 8 subjects was scanned using a single-short SE-EPI sequence with monopolar diffusion-encoding gradients along 6 directions. At the same time, spectral pre-saturation with inversion recovery technique was used for fat suppression. The detailed acquisition parameters are as follows: TR=2 heart beats, TE=66.87 ms, flip angle = 90°, FOV= 260×200 mm^2, voxel size=$3.13 \times 3.41 \times 10$ mm^3, matrix size = 224×224, and b-value=400 s/mm^2. A baseline b-value of 50 s/mm^2 was used instead of b = 0. For each subject and a given diffusion gradient direction, the DTI acquisition was executed repeatedly 8 times, and each of the 8 acquisitions was achieved in 2 cardiac cycles. The total acquisition time was about 25 mins when the heart rate of the subject was about 60 beats/min. In this dataset, there are 228 DW images.

Before training and testing, the DFT-based sub-pixel registration algorithm [35] was applied to align DW images from different acquisitions. Specifically, for each diffusion gradient direction, taking its corresponding DW image from the first acquisition as the fixed image, and then the DW images of subsequent acquisitions were registered to the fixed image one by one to ensure that the DW images from multiple acquisitions along a given direction are spatially aligned. The registration results can be found Figure A5 in the supplementary file. After registration, all the DW images were resized to 128×128 with crop or padding operations.

3.2. Experimental settings

To validate the superiority of the proposed method, we compared it with several methods, including PCATMIP [16], MPPCA [25, 26], WIF [17], U2Fusion [23] and DIF-net [24], Patch2Self [27] and DDM2 [28]. All the comparative models maintained their default settings. Since MP-PCA, Patch2Self, and DDM^2 were designed to denoise the DW images from one single acquisition, to keep the fairness of comparison, their processing results for multiple acquisitions were averaged as the final results. U2Fusion

and DIF were trained using PyTorch on an NVIDIA Tesla P40 GPU. Where U2Fusion was trained using the RMSProp optimizer with a learning rate 10^{-4} and converged after 40 epochs. DIF was trained using the Adam optimizer with a learning rate 10^{-3} and converged after 150 epochs. DDM^2 was trained using PyTorch on an NVIDIA RTX A6000 GPU for 100,000 iterations with an Adam optimizer and a learning rate 10^{-4}. The proposed model was implemented with Tensorflow and the mean square error (MSE) between input DW images and output DW images was used as the loss function of WS-MCNN. The Adam optimizer was applied to train our network, its parameters were set as: learning rate=10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$, respectively. After 45 epochs, the network was converged.

3.3. Analysis of potential influencing factors

In WS-MCNN, we used wavelet scattering to extract multi-scale features. To investigate the influence of different wavelet scattering settings on the fusion results, we compared the models without using wavelet scattering (w/o WS), with 1-scale, 2-scale and 3-scale (WS-MCNN) wavelet scatterings, respectively. In the model without using wavelet scattering, the vanilla convolution layer (kernel=3, and stride=2) was used to replace the wavelet scattering at different scales of the encoder, the remaining structures in decoder are the same as those in WS-MCNN.

In addition, WS-MCNN is a kind of fusion method, its performance may depend on the acquisition times. To further investigate the influence of the number of acquisitions on the fusion results, we have varied the number of acquisitions from at least of 2 to the possible maximum acquisitions and then compared their fusion results with the simple averaging method.

3.4. Evaluation criteria

DW images were evaluated using SNR and contrast to noise ratio (CNR), which are calculated with several selected small ROIs within the myocardium and several homogeneous regions (without any texture information) in the background, formulated as:

$$\begin{aligned} \text{SNR} &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\frac{mean(S_{ROI_i})}{\sqrt{\frac{2}{4-\pi}}\sigma(S_{back_j})} \\ \text{CNR} &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\frac{mean(S_{ROI_i})-mean(S_{back_j})}{\sigma(S_{back_j})} \end{aligned} \tag{6}$$

where $m = n = 6$ in this work, indicating the number of selected small ROIs in the left ventricle and the number of homogenous background regions. S_{ROI_i} and S_{back_j} represent the set of signals in the i-th selected ROI and j-th background region, $mean(\cdot)$ and $\sigma(\cdot)$ indicates the averaging and standard deviation operation, respectively, $\sqrt{\frac{2}{4-\pi}}$ is a correction factor.

To further evaluate the proposed method in terms of in vivo cardiac DTI, the binary mask of left ventricular myocardium was first delineated by an experienced radiologist,

and then the resulted DW images from different methods were segmented and registered to the segmented b0 image for the following diffusion tensor calculation (the registration results can be found in FigureA5 (b) in the supplementary file). Note that, the diffusion tensor images of myocardium were fitted using weighted least square (WLS) method provided by DIPY [36] library in python, from which the fractional anisotropy (FA), mean diffusivity (MD) and helix-angle (HA) [37] maps for myocardium were calculated.

In addition, to verify whether the DW images along different directions obtained by different methods are consistent with the DTI model, based on the diffusion tensor images obtained by different method, the corresponding DW images along different directions were fitted with a physical DTI model. After that, the residual maps and root-mean-square error (RMSE) between DW images obtained by different methods and the corresponding fitted DW images are calculated.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f_i)^2} \tag{7}$$

where y_i denotes the processed DW image signal with different methods and f_i is the fitted signal from the diffusion tensors obtained by different methods, N denotes the number of voxels.

4. Results

4.1. Inverse wavelet scattering reconstruction results

To verify the effectiveness of the proposed encoder-decoder architecture for inverse wavelet scattering reconstruction, Figure 3 compares the input and reconstructed DW images from wavelet scattering features using the proposed model.

In addition to the residual map between the input and reconstructed DW images, the p-value between each input-reconstructed DW image pair is also calculated using a paired t-test and given below the corresponding residual map in Figure 3. It can be noticed that the p-values of all the DW image pairs are greater than 0.05, indicating that there is no statistically significant difference between the input and reconstructed DW images. Moreover, to further demonstrate the differences between the input and reconstructed DW signals, we also plotted the signal profiles along one line (yellow line) in input and reconstructed DW images in Figure 3, we can clearly see that they are almost the same, illustrating the feasibility of WSMCNN in approximating the inverse wavelet scattering reconstruction.

4.2. Comparisons against the existing methods

4.2.1. Comparisons in terms of DW images Figure 4(a) shows the original and corrected DW images obtained by different methods (where "original" shows DW image randomly selected from any one acquisition). The top row gives the free-breathing (diastolic) short-axis DW images for b=350 s/mm^2. Although the M1-M2 motion-compensated diffusion gradient can decrease the sensitivity to bulk motion, its long TE

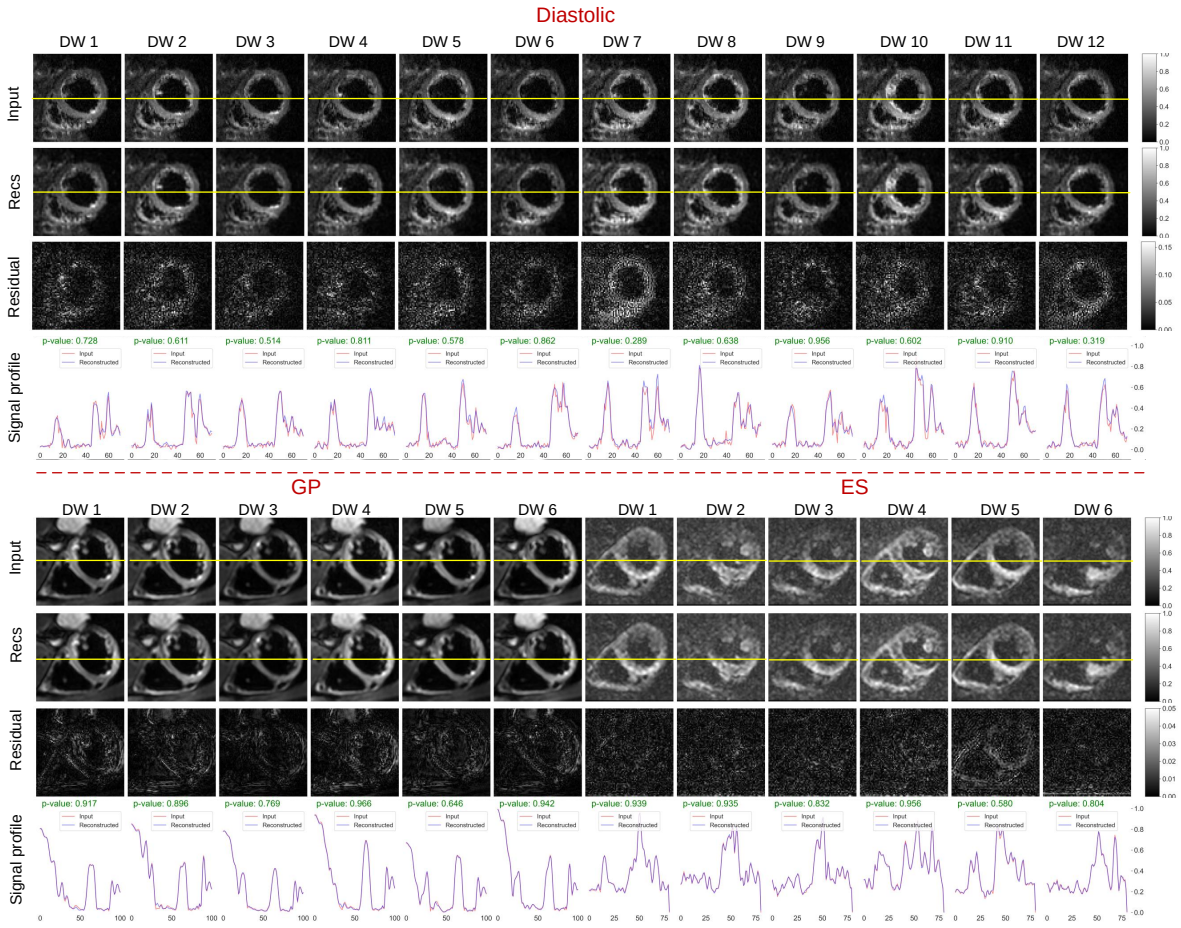

Figure 3. Input and reconstructed DW images of WS-MCNN network (without fusion) on three test datasets, as well as the residual maps between them.

increases also the noise level. By comparing the zoomed-in regions in the rectangular box of the left ventricle, we found that the noise in the DW images obtained by WS-MCNN appears to be the lowest. The third row shows the corrected results for DW images acquired with b=400 s/mm^2 with the influence of gastric peristalsis. A heterogeneous signal was obviously found in the inferior segment of the LV (indicated by the yellow rectangular box) due to the effect of gastric peristalsis. Visually, the lost DW signal can be recovered to some extent after WS-MCNN processing. The fifth row shows the DW images acquired at the end-systolic with b=750 s/mm^2. We notice that the original signal does not present severe loss since it was acquired with breath-hold. However, it had a low SNR due to the STEAM sequence acquisition [14, 38], especially in the cyan rectangular box. Clearly, the WS-MCNN method reduced indeed the effect of noise.

To quantitatively assess the performance of different methods in enhancing the quality of *in vivo* cardiac DW images, we computed the SNRs for cardiac DW images of multiple subjects in different datasets. In Figure 4(b), the height of each rectangular bar represents the mean SNR of DW images along multiple diffusion directions for a given subject, while error bar denotes the standard deviation of SNRs. Our proposed method

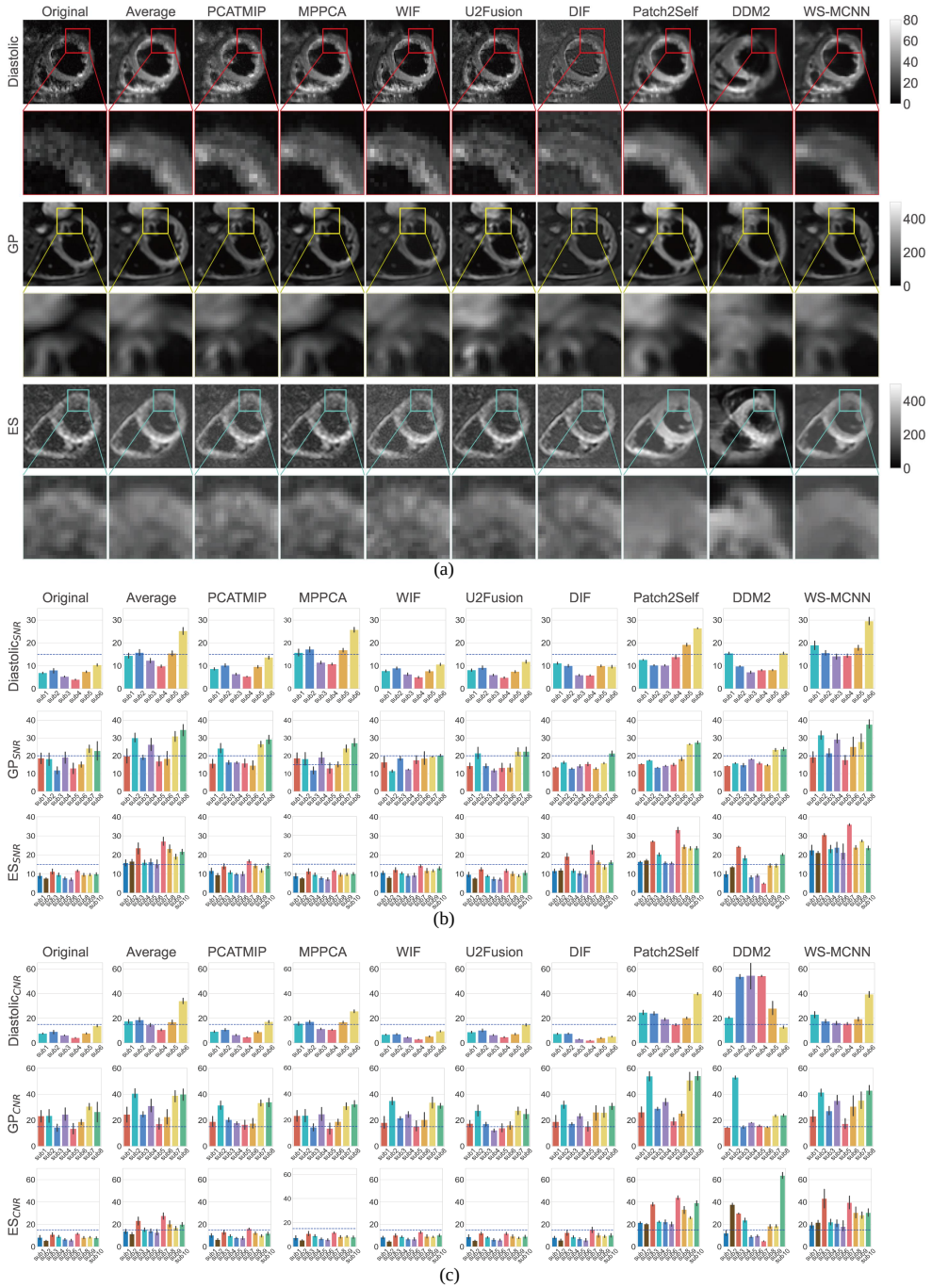

Figure 4. (a) The DW images and local zoom-in maps obtained with different methods on three datasets. SNR (b) and CNR (c) plot of DW images along different diffusion directions for all subjects on three datasets. Different colored bars represent different subjects; the height of the rectangular bars indicates the mean value of SNRs or CNRs; the length of error bars designates the standard deviations of SNRs or CNRs along different directions. GP: gastric peristalsis. ES: end systole.

consistently achieves the highest average SNR for almost all subjects across the three datasets. Furthermore, the standard deviations of the SNRs across different diffusion directions are comparatively low (especially on end diastolic and end systolic datasets), showing the excellent intra-subject (along different directions) and inter-subject stability of the proposed method. Regarding the CNR (Figure 4(c)), except for Patch2Self and DDM2 methods, our method achieves almost the best CNR on all datasets. On GP dataset, in the DW images restored by Patch2Self, the intensity of myocardium region is much higher while the intensity of background much lower (Figure 4(a)), accordingly, the CNR of Patch2Self on this dataset is extremely high. As to the method DDM2, it generates many false structures and its restored DW image experiences severe distortion (the shape of myocardium changes in Figure 4(a)), consequently, its CNR cannot be considered as a solid measure.

Figure 5 shows the residual maps (a) and curves of RMSE (b) between the original/processed DW images and the fitted DW images from their corresponding diffusion tensors with DTI model. We notice that, our WS-MCNN achieves the smallest RMSE on three datasets and the residual maps obtained by WS-MCNN contain less structural information, which suggests that DW images processed by WS-MCNN method are more consistent with DTI model than the others.

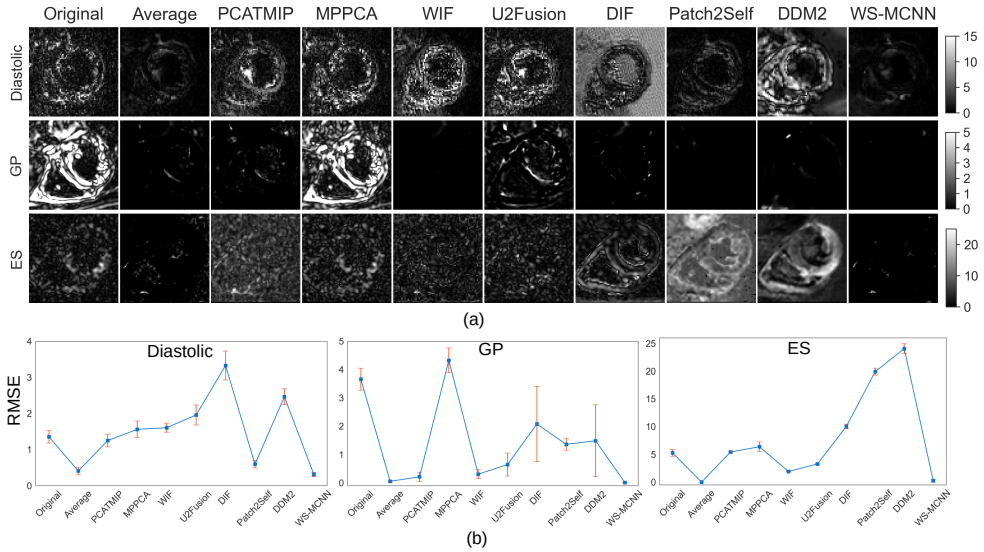

Figure 5. Residual maps (a) and curves of RMSE (b) between the original/processed DW images and the fitted DW images from their corresponding diffusion tensors with DTI model. In the plots of RMSE curves, the circle indicates the average RMSE between original/processed and fitted DW images along all diffusion gradient direction from all subjects, and the orange error bars indicate the standard deviation of RMSE. GP: gastric peristalsis. ES: end systole.

4.2.2. Comparisons in terms of diffusion metrics Figure 6 provides FA and MD maps for different methods (in all the figures related with diffusion metrics, "Original" represents the diffusion metrics fitted using original noisy DW images and b0 images from all repeated acquisitions). For the diastolic dataset (b=350s/mm^2), the FA or MD maps from the original DW images present more outliers (i.e, FA values close to 1 and MD values larger than 3×10^{-3} mm^2/s). After processing with PCATIP, WIF, DIF and U2Fusion, these outliers are not effectively removed. Regarding Patch2Self and DDM2, although they can remove the influence of noise, their resulted FA and MD maps are not normal, with FA much smaller in Patch2Self and MD much higher in DDM2. As to the proposed method WS-MCNN, it achieves comparable performance with averaging and MPPCA methods but with less grainy FA and MD maps. For dataset affected by gastric peristalsis (b=400s/mm^2), most of the original FA values in the left ventricle are close to 1, and the MD value in the region affected by gastrointestinal peristalsis was very low. Among comparison methods, only the averaging approach can solve these problems to a certain extent, but it is observed that the FA values at inferior segment are still much larger. In contrast, WS-MCNN can successfully correct FA values at this region. As for the end-systolic dataset (b=750 s/mm^2), most comparison methods cannot handle outliers in FA maps, while with WS-MCNN, the FA and MD maps are much smoother and less impacted by noise.

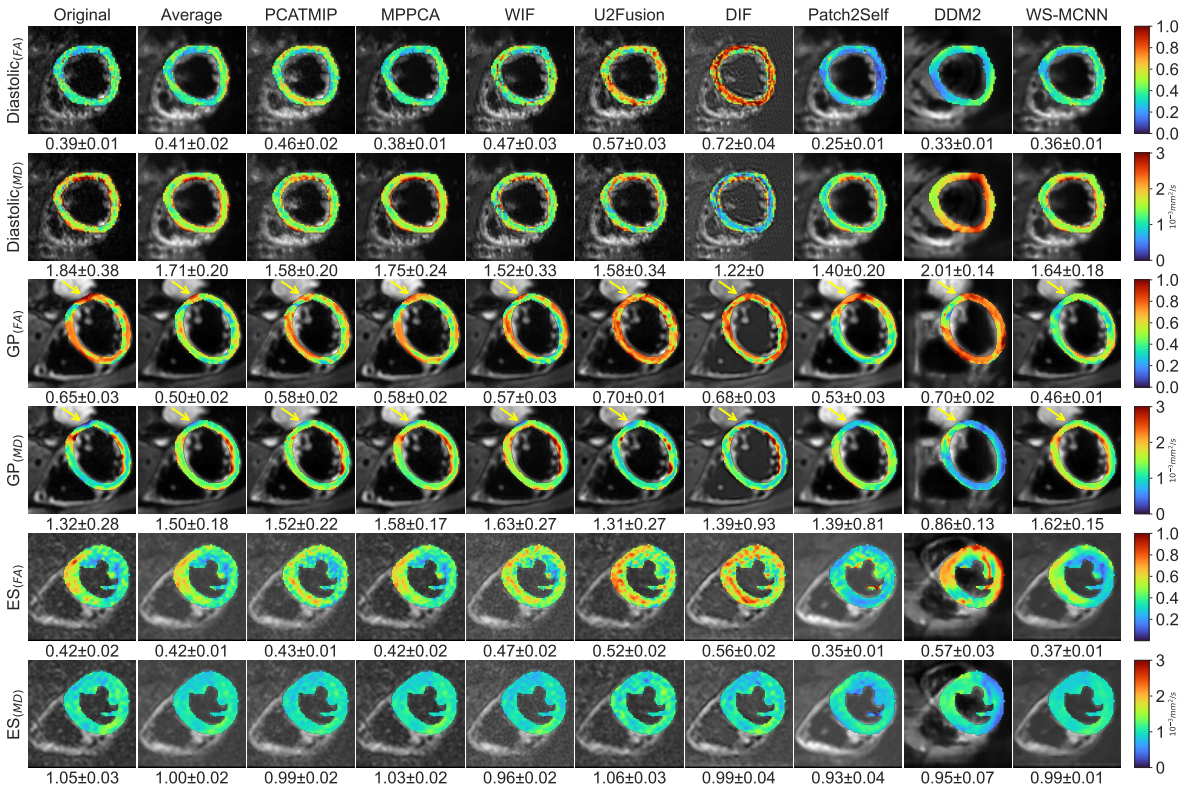

Figure 6. FA and MD maps obtained from *in vivo* DW images using different methods. GP: gastric peristalsis. ES: end systole. The number below each subfigure indicates the mean ± std of FA or MD values.

To further quantitatively assess the effectiveness of different methods for DTI quality enhancement, we also compared the inter-subject distribution of FA and MD values on three datasets, as shown in Figure 7. For each line plot, the horizontal axis indicates the transmural locations from endocardium to epicardium, the vertical axis indicates the values of FA or MD, the red line is the average of FA or MD values of all voxels from all subjects in the region of interest (different transmural regions), and the red shadow region around the red line shows the interquartile range. It can be observed that in all datasets, FA values corrected with WS-MCNN peaks in the midmyocardium compared with endocardium and epicardium, while there is no such trend in FA values obtained by other methods. Regarding MD values, our method generates fewer outliers, showing that our method can effectively remove the influence of motion or noise on MD maps. In addition, our method achieves the narrower shadowed regions, implying that the performance of WS-MCNN is stable for all the subjects.

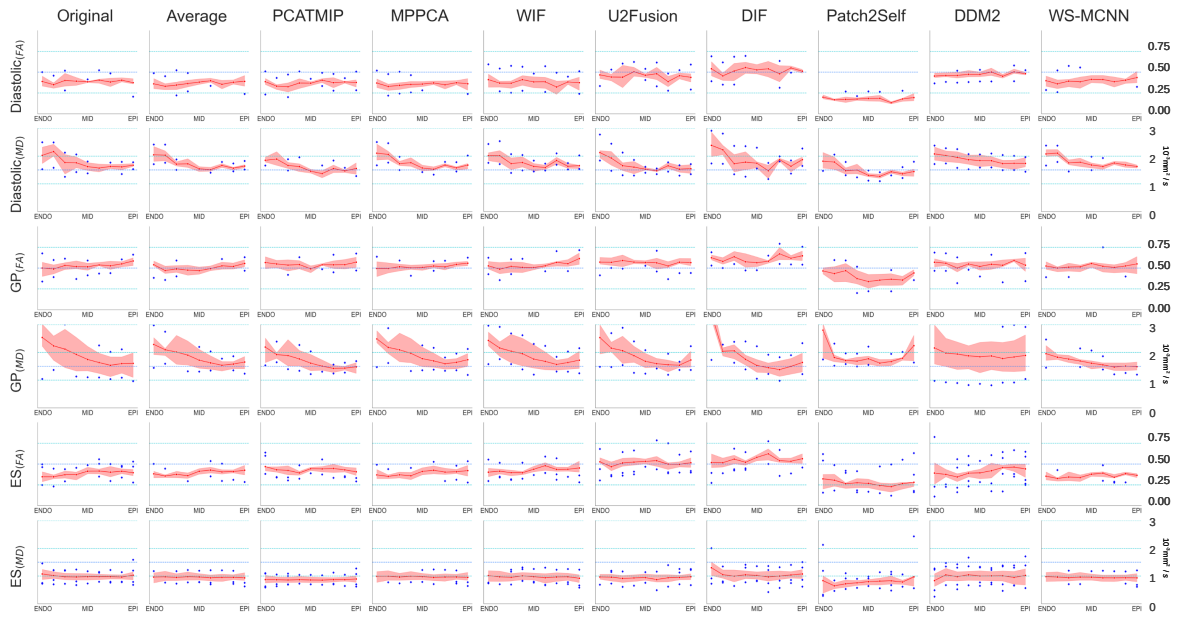

Figure 7. Line plots of FA and MD values in LV obtained with different methods. The horizontal axis indicates the transmural locations changing from endocardium through mid-myocardium to epicardium. The vertical axis indicates FA or MD values. The red line is the average of FA or MD values of all voxels from all subjects in the region of interest (different transmural regions), the red shadow region around the red line shows the interquartile range (IQR, between the first quartile (Q1) and the third quartile (Q3)), and the blue points indicate the outliers which out of the range of [Q1-1.5×IQR, Q3+1.5×IQR]. GP: gastric peristalsis. ES: end systole.

4.2.3. Comparisons in terms of myocardial fiber structure Figure 8(a) shows the corrected fiber orientations accompanied by HA maps obtained with different

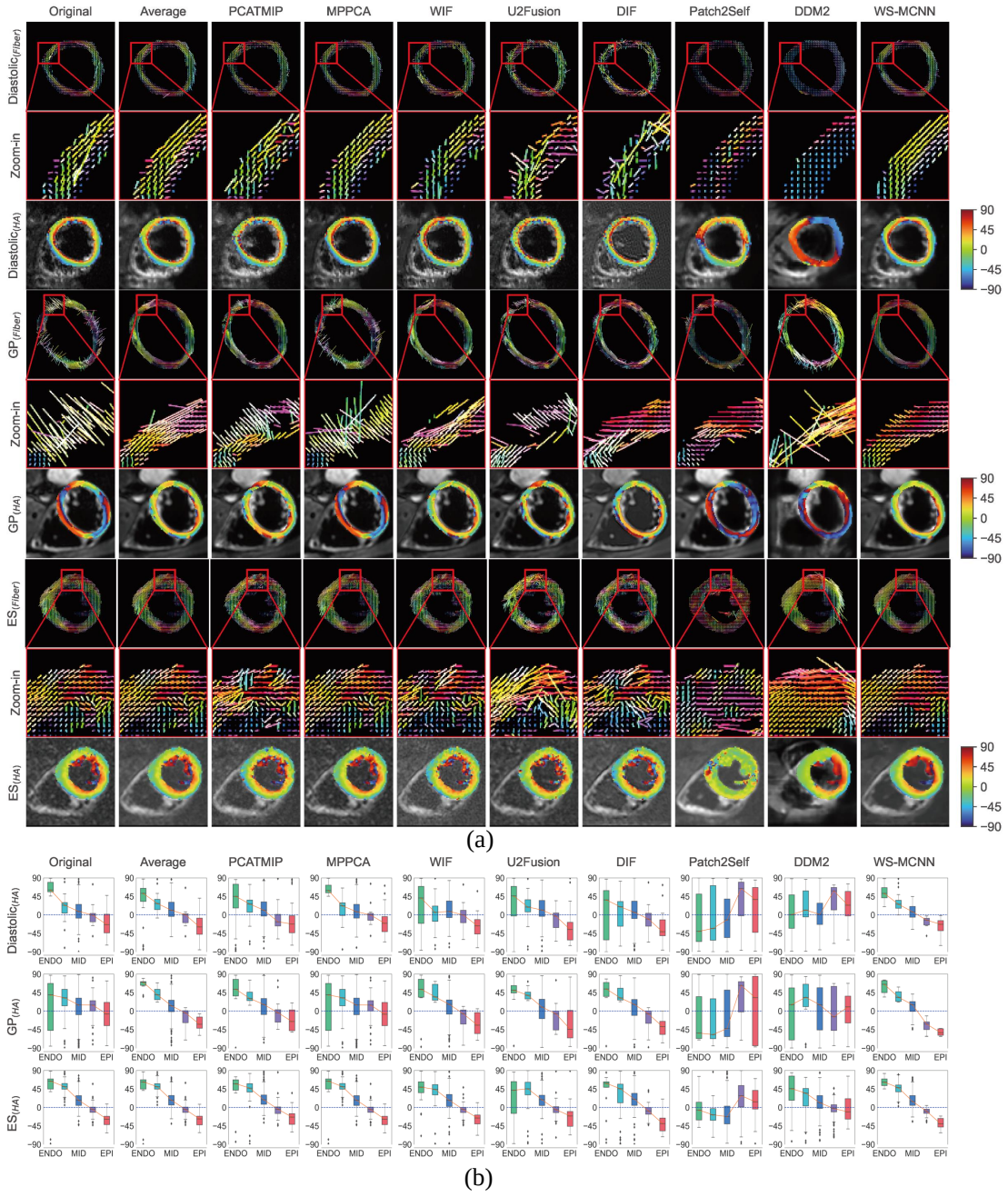

Figure 8. (a) Comparing the fiber orientation and HA maps obtained from one of original acquisitions against those corrected ones with different methods. (b) The boxplots of HA values for different subjects at different transmural locations. The red curves in (b) reflect the variations of the median helix angles from endocardium to epicardium. ENDO: endocardium, MID: midmyocardium, EPI: epicardium; GP: gastric peristalsis. ES: end systole.

methods. For diastolic dataset, Patch2Self and DDM2 generate totally erroneous fiber orientations, while PCATMIP, WIF, U2Fusion and DIF cannot overcome the influence

of noise, resulted in the disarranged fiber orientations at inferior segment. Although averaging multiple acquisitions can restore the helical structure, WS-MCNN method produces a better HA distribution in the endocardium, that means the HA values in endocardium derived from our method do not have outliers (comparing the second column against the last column in Figure 8 (b)). As to MPPCA method, it achieves comparable performance as our method. For the GP dataset, due to the influence of gastric peristalsis and breath motion, the helical structure of myocardium fibers derived from the original acquisitions were disrupted. Although all the methods can restore the fiber orientations at the free lateral segment, most of them are not able to restore the fiber orientations in the inferoseptal segment (red rectangles). As to our proposed WS-MCNN method, it yields the best corrected HA map with the HA values varying almost continuously and smoothly through myocardial walls. For the ES dataset, since it was acquired with breath-hold, the noise instead of the motion is the main factor for image quality degradation. Accordingly, averaging simply the multiple acquisitions also leads to desirable result. However, the fiber orientations obtained with our WS-MCNN are more coherent and HA maps are much smoother. In addition, except for DDM2 and Patch2Self, all the other comparison methods can restore the fiber orientations to a certain level, but a little noisy. More quantitatively (Figure 8(b)), with WS-MCNN, the mean HA values at midmyocardium for all the datasets are close to zero, and the variation trend from endocardium to epicardium is more coherent with respect to the other methods. In addition, the interquartile range of HAs at all the locations are almost the smallest, illustrating that HA values obtained by our method did not change too much across different subjects.

4.3. Evaluating the results obtained with different settings

4.3.1. Influence of different wavelet scattering settings In Figure 9 are shown the original DW images, fiber orientations, FA, MD, and HA maps from multiple acquisitions (Acquisition 1 to Acquisition 5), as well as those obtained from averaging and WSMCNN models with different wavelet scattering settings (the last four columns). Noise and signal loss caused by motion in the original DW images can be clearly seen (indicated by yellow arrows), even though using traditional CNN (w/o WS) to fuse multiple acquisitions are able to deal with motion and noise problems to a certain level, they are not better than simply averaging and there are still some residual noise and outliers in FA (extremely high FA values) and HA maps. Using wavelet scattering features of multiple scales (from scale1_WS to scale3_WS (WS-MCNN)) can improve gradually the fusion quality, with the increment of mean SNR and CNR for multiple directional DW images being up to 99% and 93%, respectively. In terms of fiber orientation, increasing wavelet scattering scales can make the fiber orientations more coherent and the transitions of helix angles much smoother.

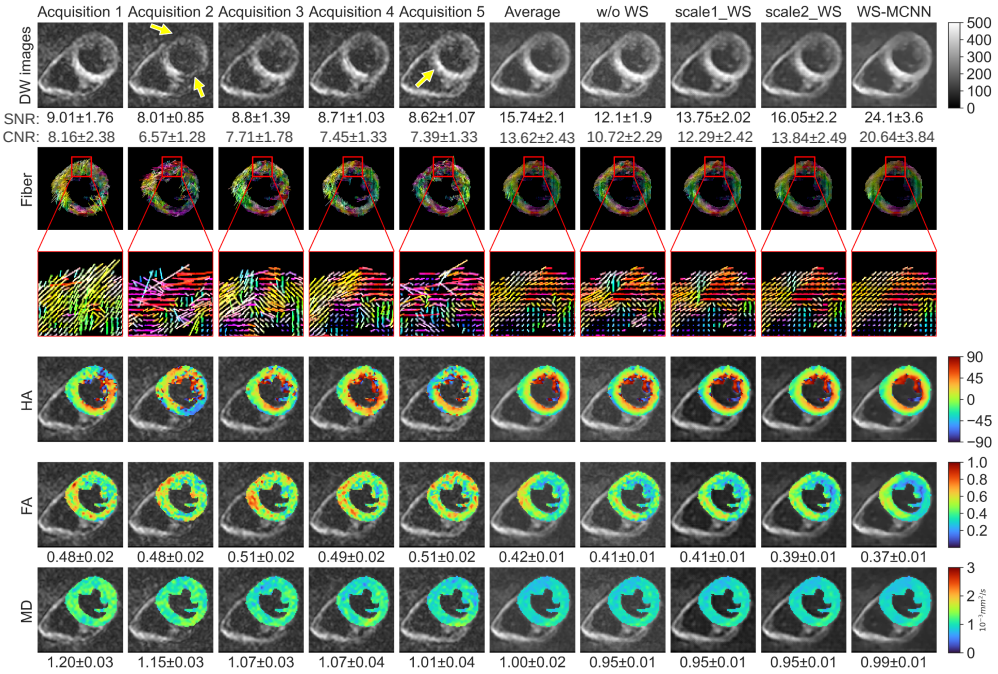

Figure 9. Original DW images from multiple acquisitions and the fusion results obtained with averaging and WSMCNN models with different wavelet scattering settings, as well as the corresponding diffusion metric maps. w/o WS means using the convolution layers to replace the multi-scale wavelet scattering in WS-MCNN; scale1_WS and scale2_WS represent using one-scale and two-scale wavelet scattering, respectively; WS-MCNN is the proposed model where wavelet scattering at three scales is used.

4.3.2. Influence of the number of acquisitions From Figures 6-9, we notice that, when the repeated acquisition number is large (8, 6 and 5 times for diastolic, GP and ES datasets, respectively), there is no significant difference between averaging and our method WS-MCNN. To test the influence of acquisition times on their performance, Figure 10 compares the results derived from averaging and our method when varying the acquisition times from 2 to 5. We can observe that the performance of averaging method is sensitive to the number of repeated acquisitions. When the number of repeated acquisition is less than 3, it cannot suppress noise effectively. In contrast, the proposed WS-MCNN method is more robust to the number of the repeated acquisitions, when the acquisition number changes from 2 to 5, its fused DW images, FA and MD maps are almost the same, illustrating the superiority of the proposed method. It means that using the transformation-invariant and noise-robust wavelet scattering features enables us to get the better fusion results from the limited data.

5. Discussion

In the present work, we have proposed to combine multi-scale wavelet scattering coefficients and CNN model (WS-MCNN) to improve the quality of *in vivo* cardiac

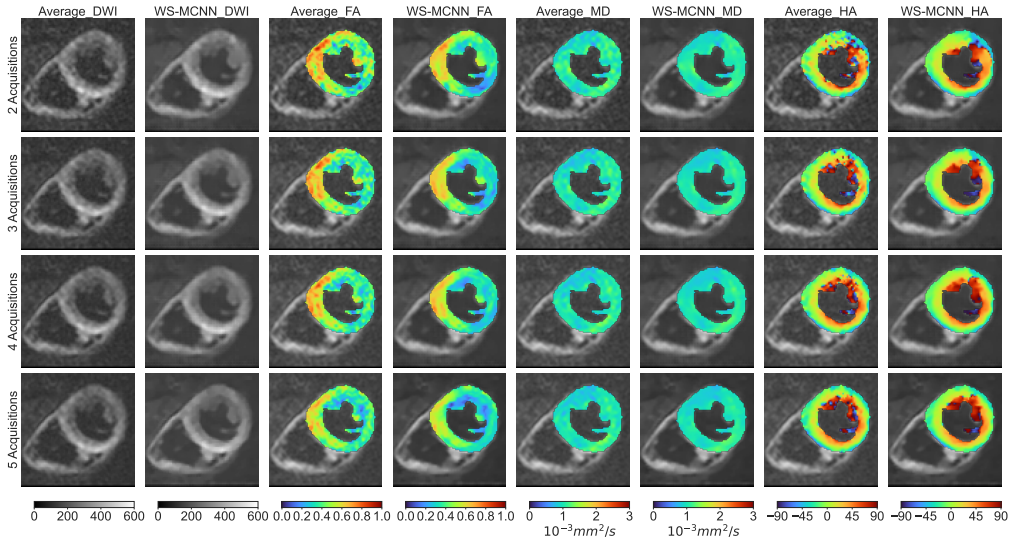

Figure 10. Influence of the number of repeated acquisition on the results of averaging and WS-MCNN methods on ES dataset. The acquisition number varies from 2 to 5.

DTI. Through the experiments on three test datasets, we demonstrated that WS-MCNN allowed us to clearly restore the helical structure of myocardial fibers from the DW images degraded by residual motion and noise, with a circumferential zero contour at mid-myocardium and a smooth positive-negative transmural transition from endocardium to epicardium.

We used two kinds of cardiac DW images acquired at diastole under free breathing. The one was acquired with b =350 s/mm^2 and low SNR, and the other with b =400 s/mm^2 without severe signal loss but with obvious influence of gastric peristalsis. For these two datasets, after the correction with WS-MCNN, DW image details were well restored, with the highest SNR and CNR (Figure 4(b) and (c)), and the helical structure of cardiac fibers were also recovered. The corrected mean FA ($\text{FA}_{\text{b350}} = 0.36 \pm 0.01$) and mean MD values ($\text{MD}_{\text{b350}} = 1.64 \pm 0.18\ 10^{-3}\text{mm}^2/\text{s}$) were similar to those obtained in the previous studies in diastole using M1-M2 motion-compensated gradients (FA = 0.35 and MD = $1.64\ 10^{-3}\text{mm}^2/\text{s}$) [32]. However, FA ($\text{FA}_{\text{b400}} = 0.46 \pm 0.01$) and MD ($\text{MD}_{\text{b400}} = 1.62 \pm 0.15\ 10^{-3}\text{mm}^2/\text{s}$) obtained by WS-MCNN were different from those calculated from IVIM data ($\text{FA}_{\text{IVIM}} = 0.37 \pm 0.03, \text{MD}_{\text{IVIM}} = 1.78 \pm 0.21\ 10^{-3}\text{mm}^2/\text{s}$), acquired using the same acquisition protocol by Zhang et al [34]. This may be due to the fact that DTI model takes into account direction information in the calculation of diffusion coefficient but not in the case of IVIM bi-exponential model [39].

For the DW images acquired at end systole with breath-hold, the main problem is noise influence rather than motion effect. Although almost all the post-processing methods can reduce the noise to a certain level, our proposed method achieved the best performance in terms of SNR of corrected DW images (Figure 3) and HA maps (Figure 8(a)), which validates the superiority of the WS-MCNN. Regarding the corrected mean FA and MD values, the range of FA and MD values ($\text{FA}_{\text{b750}} =$

$0.37 \pm 0.01, \mathrm{MD}_{\mathrm{b750}} = 0.99 \pm 0.01\ 10^{-3}\mathrm{mm}^2/\mathrm{s}$) was also consistent with the previous findings ($\mathrm{FA}_{\mathrm{b750}} = 0.39 \pm 0.02, \mathrm{MD}_{\mathrm{b750}} = 0.93 \pm 0.037\ 10^{-3}\mathrm{mm}^2/\mathrm{s}$) [33]. Since this dataset was acquired with several b-values, the correction results for the cardiac DTI with other b-values can be found in the supplementary files. The ranges of the corrected FA values for other b-values were also consistent with the previous studies [33]. This demonstrates the reliability of our proposed method.

To the best of our knowledge, it is the first time that multi-scale transformation-invariant features and deep learning model were combined to improve the quality of *in vivo* cardiac DW images. Although there are a few acquisition techniques aimed at reducing the effects of bulk motion in *in vivo* cardiac DW images, such as bipolar diffusion encoding gradient pulses [40], simulated echoes over two cardiac cycles or a single short acquisition with navigator-based gating [41], and diffusion encoding schemes that compensate for the first-, second- or high-order motion [40, 42, 43], our method provided a post-processing alternative to deal with the residual motion and noise issues that have not yet been solved with current acquisition sequences. Compared to existing post-processing methods of compensating for motions in *in vivo* cardiac DTI, such as PCATMIP [16] and WIF [17], the proposed WS-MCNN method worked on transformation-invariant and noise-robust wavelet scattering coefficient maps extracted from multiple repeated DW images, which enabled us to overcome influence of irrelevant information in multiple DW images on the fusion result, accordingly, achieved the best performance, even compared with several CNN-based fusion methods (U2Fusion and DIF). Comparing against the state-of-the-art denoising methods for diffusion MRI, such as MPPCA [25, 26], Patch2Self [27] and DDM2 [28], our method still obtained superior performance. As can be seen from the Figure 4, the MPPCA and Patch2Self methods generate satisfactory results on the diastolic dataset. On the GP and ES datasets, the results of MPPCA are similar to the original data, with no appreciable improvement. This can be attributed to the inherent limitations of the MPPCA method, which uses principal component analysis for denoising and relies on the redundancy of DW images. When the number of diffusion gradient directions is insufficient, DW images no longer exhibit redundancy Similarly, in the Patch2Self method, it uses the neighboring signal along other diffusion gradient directions to fit the clean signal of a given voxel along current diffusion gradient direction. When the number of the diffusion gradient direction is lower, the fitting error becomes larger. This is why Patch2Self does not perform well on GP and ES datasets. As to the diffusion model based method DDM2, it generated the worst results on all datasets, with severe artifacts and distortions. This can be caused by the influences of both sample size and processing flow of DDM2. DDM2 is kind of generative diffusion models which involves three steps for denoising, including noise estimation, forward diffusion state matching and reverse diffusion reconstruction. This multi-stage learning method is easily affected by multiple factors and may generate false structures in the denoised images, especially when the sample size is not large enough.

The superiority of wavelet scattering was further verified by ablation experiments, which demonstrated the importance of multi-scale invariant features for image fusion.

Since the wavelet scattering features are robust to noise and transformations, they are beneficial for image restoration. Simply using CNN without the wavelet scattering did not allow us to effectively compensate for motion or noise effects (Figure 9 (w/o ws)). This explains why the DIF-net and U2Fusion did not restore well cardiac fiber orientations (Figure 8). In addition, the wavelet scattering transform itself being non-invertible, to reconstruct the fused DW images from scattering coefficient maps, we used a CNN-based network to learn the relationship between the wavelet scattering coefficients and the corresponding DW images. In the present work, a total of 1152 samples acquired early-systole and end-systole were used for training. Such a dataset size was enough to guarantee that our network can learn accurately the relationship between the nearly-invariant wavelet scattering coefficients and DW images, this can be verified by the Figure 3, where the difference between the input and reconstructed DW images is not significant. Accordingly, using such network to restore DW images from the fused features extracted from wavelet scattering maps is reliable. Moreover, the ablation results in Figure 10 verify that the performance of the proposed WS-MCNN method is more robust to the number of the repeated acquisitions, it can achieve the comparable or even better results than averaging method only with fewer repeated acquisitions, illustrating that using the transformation-invariant and noise-robust wavelet scattering features enables us to effectively explore the useful information from the limited data to fuse, providing a potential mean to alleviate the dependence of the fusion results on the number of repeated acquisitions.

It was commonly recognized that respiratory and cardiac motions are two main sources of signal loss, even with good cardiac triggering and motion compensation, there may be slight cardiac deformation or rotation which can cause signal loss. Meanwhile, the presence of gastric peristalsis may also be a source of signal loss. In our present study, for each subject in GP dataset and a given diffusion gradient direction, the DTI acquisition was executed repeatedly 8 times, and each of the 8 acquisitions was achieved in 2 cardiac cycles. Consequently, gastric peristalsis was certainly present. In the future, it would be interesting to design appropriate acquisition strategies to avoid the impact of any organ motion including gastric peristalsis. In addition, in the dataset acquired with STEAM sequence, we did not adjust the b-values on heart rate to calculate the subsequent diffusion metrics, which may bring some influence on FA and MD values. Moreover, the acquisition protocols involved in the present study allowed acquiring only one slice along a few diffusion directions. If the few diffusion gradient directions are not optimally distributed, the fitted diffusion metrics will be influenced (such as on GP datasets, all the methods cannot generate the satisfactory helix angle map, which may be caused by diffusion gradient directions.) Besides, only one slice is not sufficient for expressing the diffusion properties of the entire heart, which is a limitation of our method. In the future, it would be interesting to develop post-processing methods allowing augmenting dataset through obtaining more slices or more directions, which could enable us to recover possible missing fiber structure details and compensate for motion and noise effect simultaneously. Furthermore, bulk motion also influences the

phase image. The proposed motion compensation and noise removing method was implemented only on magnitude images. Performing motion compensation and noise reduction on complex data (if available in advanced sequences) containing both phase and magnitude would constitute an interesting future work. Besides, we validated the performance of WS-CNN on end-systolic and diastolic datasets, further validations on datasets acquired at other cardiac phases would also be required. Lastly, to guarantee the performance of all methods, the acquisitions with severe signal loss (such as ROI is invisible or most of ROI disappears) were excluded from all the methods. Therefore, even though our method can achieve the best performance on the dataset with exclusion, it should be noted that all the existing methods, including the proposed one, are still not capable of recovering images with complete or severe signal loss. Restoring the DW images completely or highly corrupted by motion remains a great challenge.

6. Conclusion

We have proposed a novel method WS-MCNN to compensate for the effects of motion and noise in *in vivo* cardiac DTI. The WS-MCNN consists of extracting and fusing multi-scale wavelet scattering features of DW images acquired at different times or different repetitions, and reconstructing high quality DW images using a CNN-based invertible wavelet scattering. The experimental results on three kinds of *in vivo* cardiac DTI datasets, including diastolic, gastric peristalsis influenced, and end-systolic cardiac DTI, showed that the proposed WS-MCNN method effectively compensates for motion-induced signal loss and removes the noise, producing better DW image quality and more coherent fiber structures. Compared to three traditional methods (PCATMIP, MPPCA and WIF) and several learning-based methods (U2Fusion, DIF-net, Patch2Self and DDM2), WS-MCNN outperforms them for dealing with the effects of motion and noise in *in vivo* cardiac DTI.

Acknowledgments

This work was supported by the National Nature Science Foundations of China (Grant No.62161004), Guizhou Provincial Science and Technology Projects (QianKeHe ZK [2021] Key 002), Nature Science Foundations of Guizhou Province (QianKeHe [2020]1Y255), Guizhou Provincial Basic Research Program (QianKeHe ZK [2023] Key 058) and International Research Project METISLAB of CNRS. Andrew Scott and Sonia Nielles-Vallespin acknowledge funding from British Heart Foundation grant RG/19/1/34160. We would like to thank Reviewers for taking the time and effort necessary to review the manuscript. We sincerely appreciate that one Reviewer point out one mistake about the data preprocessing, which helped us to improve the quality and solidity of the manuscript.

Appendices

Figure A1 FA and MD maps obtained from *in vivo* DW images with b=150, 350, 550 s/mm^2 using different methods.
Figure A2 HA maps obtained from *in vivo* DW images with b=150, 350, 550 s/mm^2 using different methods.
Figure A3 FA and MD lineplots obtained from *in vivo* DW images with b=150, 350, 550 s/mm^2 for all subjects using different methods.
figure A4 HA boxplots obtained from *in vivo* DW images with b=150, 350, 550 s/mm^2 for all subjects using different methods.
figure A5 (a) Align multiple DW image acquisitions to the first acquisition before network training and testing. This alignment is performed for each given diffusion gradient direction separately. (b) After network testing, align the processed DW images along different diffusion gradient directions to b0 image for calculating the diffusion metrics. Note that only the left ventricular myocardium was considered during this registration.

References

[1] Edelman R R, Gaa J, Wedeen V J, Loh E, Hare J M, Prasad P and Li W 1994 *J Magnetic Resonance in Medicine* **32** 423–428
[2] Frindel C, Robini M, Schaerer J, Croisille P and Zhu Y M 2010 *J Magnetic Resonance in Medicine* **64** 1215–1229
[3] Nielles-Vallespin S, Scott A, Ferreira P, Khalique Z, Pennell D and Firmin D 2020 *J Journal of Magnetic Resonance Imaging* **52** 348–368
[4] Mekkaoui C, Reese T G, Jackowski M P, Bhat H and Sosnovik D E 2017 *J NMR in Biomedicine* **30** e3426
[5] Afzali M, Mueller L, Coveney S, Fasano F, Evans C J, Engel M, Szczepankiewicz F, Teh I, Dall'Armellina E, Jones D K *et al.* 2024 *Magnetic Resonance in Medicine*
[6] Wu M T, Tseng W Y I, Su M Y M, Liu C P, Chiou K R, Wedeen V J, Reese T G and Yang C F 2006 *J Circulation* **114** 1036–1045
[7] Mekkaoui C, Huang S, Dai G, Reese T G, Ruskin J, Hoffmann U, Jackowski M P and Sosnovik D E 2013 *J Journal of Cardiovascular Magnetic Resonance* **15** 1–3
[8] Sosnovik D E, Mekkaoui C, Huang S, Chen H H, Dai G, Stoeck C T, Ngoy S, Guan J, Wang R, Kostis W J *et al.* 2014 *J Circulation* **129** 1731–1741
[9] Das A, Kelly C, Teh I, Sharrack N, Stoeck C T, Kozerke S, Schneider J E, Plein S and Dall'Armellina E 2022 *Journal of Magnetic Resonance Imaging* **56** 1171–1181
[10] Das A, Chowdhary A, Kelly C, Teh I, Stoeck C T, Kozerke S, Maxwell N, Craven T P, Jex N J, Saunderson C E *et al.* 2021 *Journal of Magnetic Resonance Imaging* **53** 73–82
[11] Farzi M, Coveney S, Afzali M, Zdora M C, Lygate C A, Rau C, Frangi A F, Dall'Armellina E, Teh I and Schneider J E 2023 *Magnetic Resonance in Medicine* **90** 2144–2157
[12] Das A, Kelly C, Teh I, Nguyen C, Brown L A, Chowdhary A, Jex N, Thirunavukarasu S, Sharrack N, Gorecka M *et al.* 2022 *European Heart Journal-Cardiovascular Imaging* **23** 352–362
[13] Tseng W Y I, Reese T G, Weisskoff R M and Wedeen V J 1999 *J Magnetic Resonance in Medicine* **42** 393–403
[14] von Deuster C, Stoeck C T, Genet M, Atkinson D and Kozerke S 2016 *J Magnetic resonance in medicine* **76** 862–872

[15] Stoeck C T, von Deuster C, Fleischmann T, Lipiski M, Cesarovic N and Kozerke S 2018 *Magnetic resonance in medicine* **79** 2265–2276
[16] Pai V, Rapacchi S, Kellman P, Croisille P and Wen H 2011 *J Magnetic Resonance in Medicine* **65** 1611–1619
[17] Wei H, Viallon M, Delattre B M, Moulin K, Yang F, Croisille P and Zhu Y 2014 *J IEEE transactions on medical imaging* **34** 306–316
[18] Ghodrati V, Bydder M, Ali F, Gao C, Prosper A, Nguyen K L and Hu P 2021 *J NMR in Biomedicine* **34** e4433
[19] Pawar K, Chen Z, Shah N J and Egan G F 2022 *J NMR in Biomedicine* **35** e4225
[20] Gadjimuradov F, Benkert T, Nickel M D, Führes T, Saake M and Maier A 2022 *J Magnetic Resonance in Medicine* Doi:10.1002/mrm.29380
[21] Ferreira P F, Martin R R, Scott A D, Khalique Z, Yang G, Nielles-Vallespin S, Pennell D J and Firmin D N 2020 *J Magnetic Resonance in Medicine* **84** 2801–2814
[22] Weine J, van Gorkum R J, Stoeck C T, Vishnevskiy V and Kozerke S 2022 *J Computerized Medical Imaging and Graphics* **99** 102075 ISSN 0895-6111 URL `https://www.sciencedirect.com/science/article/pii/S0895611122000489`
[23] Xu H, Ma J, Jiang J, Guo X and Ling H 2020 *J IEEE Transactions on Pattern Analysis and Machine Intelligence* **44** 502–518
[24] Jung H, Kim Y, Jang H, Ha N and Sohn K 2020 *J IEEE Transactions on Image Processing* **29** 3845–3858
[25] Veraart J, Novikov D S, Christiaens D, Ades-Aron B, Sijbers J and Fieremans E 2016 *Neuroimage* **142** 394–406
[26] Mosso J, Simicic D, Şimşek K, Kreis R, Cudalbu C and Jelescu I O 2022 *NeuroImage* **263** 119634
[27] Fadnavis S, Batson J and Garyfallidis E 2020 *Advances in Neural Information Processing Systems* **33** 16293–16303
[28] Xiang T, Yurt M, Syed A B, Setsompop K and Chaudhari A 2023 *arXiv preprint arXiv:2302.03018*
[29] Mallat S 2012 *J Communications on Pure and Applied Mathematics* **65** 1331–1398
[30] Bruna J and Mallat S 2013 *J IEEE transactions on pattern analysis and machine intelligence* **35** 1872–1886
[31] Burt P J and Adelson E H 1987 *J Readings in computer vision* 671–679
[32] Moulin K, Verzhbinsky I A, Maforo N G, Perotti L E and Ennis D B 2020 *J PloS one* **15** e0241996
[33] Scott A D, Ferreira P F, Nielles-Vallespin S, Gatehouse P, McGill L A, Kilner P, Pennell D J and Firmin D N 2015 *J Magnetic Resonance in Medicine* **74** 420–430
[34] Zhang X S, Sang X Q, Kuai Z X, Zhang H X, Lou J, Lu Q and Zhu Y M 2021 *Magnetic Resonance in Medicine* **85** 1414–1426
[35] Guizar-Sicairos M, Thurman S T and Fienup J R 2008 *J Optics letters* **33** 156–158
[36] Garyfallidis E, Brett M, Amirbekian B, Rokem A, Van Der Walt S, Descoteaux M, Nimmo-Smith I and Contributors D 2014 *Frontiers in neuroinformatics* **8** 8
[37] Khalique Z, Ferreira P F, Scott A D, Nielles-Vallespin S, Firmin D N and Pennell D J 2020 *Cardiovascular Imaging* **13** 1235–1255
[38] Ferreira P F, Nielles-Vallespin S, Scott A D, de Silva R, Kilner P J, Ennis D B, Auger D A, Suever J D, Zhong X, Spottiswoode B S *et al.* 2018 *J Magnetic Resonance in Medicine* **79** 2205–2215
[39] Moulin K, Croisille P, Feiweier T, Delattre B M, Wei H, Robert B, Beuf O and Viallon M 2016 *Magnetic resonance in medicine* **76** 70–82
[40] Gamper U, Boesiger P and Kozerke S 2007 *J Magnetic Resonance in Medicine* **57** 331–337
[41] Nielles-Vallespin S, Mekkaoui C, Gatehouse P, Reese T G, Keegan J, Ferreira P F, Collins S, Speier P, Feiweier T, De Silva R *et al.* 2013 *J Magnetic Resonance in Medicine* **70** 454–465
[42] Stoeck C T, Von Deuster C, Genet M, Atkinson D and Kozerke S 2016 *J Magnetic Resonance in Medicine* **75** 1669–1676
[43] Welsh C L, DiBella E V and Hsu E W 2015 *J IEEE transactions on medical imaging* **34** 1843–1853