

TRFSA-HQS: Transformer-based MRI Compressed Sensing Network with Frequency Domain Self-Attention

Taiping Mo¹, Mingfeng Zheng^{1, a}, Peng Sun^{1, b} and Jiangtao Li¹

¹ School of Electronic Engineering and Automation, Guilin University of Electronic Science and Technology, Guangxi, China;

^azheng41677@gmail.com, ^bsunpeng@guet.edu.cn.

Abstract. Compressive sensing (CS) can reconstruct undersampled magnetic resonance images, but its iterative optimization process tends to be computationally intensive and time-consuming. Convolutional neural networks (CNNs) have demonstrated impressive reconstruction performance through non-linear feature extraction and mapping capabilities. However, CNNs often struggle to effectively learn and capture global dependencies in the data. Therefore, constructing a MRI reconstruction method that meets clinical real-time imaging and captures dynamic global correlation is crucial in fast and high-precision MRI tasks. We propose a deep unfolding network (TRFSA-HQS) for fast and accurate CS reconstruction, which combines frequency domain self-attention (FSA) based Transformer and half-quadratic splitting (HQS) iterative optimization scheme. TRFSA-HQS adopts an iterative scheme based on HQS to effectively decouple and update optimization problems. The decoupled data subproblems are updated by minimizing the objective function with competitive terms, while the regularization subproblems are updated using the TRFSA deep prior network. The TRFSA module employs an asymmetric UNet architecture, where the encoder and decoder utilize a frequency-domain discriminative feedforward network (DFFN) and FSA. The DFFN selectively extracts deep features from different frequency components, while the FSA captures global dependencies in the frequency domain. The experiment demonstrate that the proposed model achieves an average reconstruction PSNR of 35.11dB on a 0.2 sampling rate test set on FastMRI knee joint data, with an inference time of 0.74s. It has good reconstruction performance and noise robustness, meeting the clinical real-time imaging requirements.

Keywords: Compressed sensing MRI; half-quadratic splitting; transformer; deep prior.

1. Introduction

MRI is widely used in clinical imaging diagnosis and treatment. The traditional sampling mode requires repeated application of gradient fields for full sampling scanning, which is time-consuming and does not meet the clinical real-time imaging requirements. Moreover, long scanning durations are susceptible to interference from respiration and heartbeats, impacting image quality. Therefore, improving scanning speed is crucial. CS[1] has succeeded in undersampling imaging. Compared to full sampling, CS can reconstruct high-quality images from limited data through partial sampling. However, the iterative optimization algorithm of CS is computationally complex, hindering faster reconstruction and limiting CS application in real-time MRI. Deep learning (DL), especially data-driven CNN models, has shown promising performance in medical imaging. CNNs can directly learn input-output mapping from large samples through powerful feature extraction and non-linear mapping capabilities. This makes them an important approach for improving MRI reconstruction speed and accuracy.

Combining CS models and DL approaches offers an effective solution for complex MRI reconstruction tasks. However, pure-CNN models have certain constraints in modelling global dependencies relationships. exhibit certain limitations in modeling global spatial relationships^[1]. Global contextual information, including relative positions and overall structures of different regions, is crucial for learning features of key areas. Transformer models[1] have successfully captured global dependencies by learning attention weights through self-attention. Yet, directly applying Transformers to MRI data requires significant computational resources, as MRI features are typically high-dimensional. Integrating Transformers with CS provides a new framework for

learning deep priors^[2]. The standard Transformers capture long-range dependencies by computing the correlations between a token and all other tokens. This computation process can be implemented using convolution operations by rearranging the tokens [34]. Furthermore, the convolution theorem states that spatial-domain convolution is equivalent to Hadamard product in the frequency domain^[3]. This insight inspires us to explore more efficient self-attention mechanisms in the frequency domain for MRI reconstruction.

We propose a deep unfolding network (termed TRFSA-HQS) based on frequency domain self-attention (FSA) combined with model based half-quadratic splitting (HQS)[4, 7] iterations to achieve fast and accurate CS of MRI. The HQS alternating optimization strategy avoids complex equations and operations of matrix, thereby reducing the overall algorithmic complexity. In the HQS iterative backbone, we propose TRFSA based on Transformer for deep feature extraction and global dependency modeling. The TRFSA enhances the structure and details of the image while reducing data requirements and computational complexity. The main contributions are summarized as follows:

1) We devise a half-quadratic splitting (HQS) iterative backbone to decouple the optimization problem in CS reconstruction. During the compressive stage, we employ radial undersampling to sufficiently sample the low-frequency components of MR data, which helps suppress motion artifacts and aliasing. In the sensing stage, we use a UNet [7] encoder-decoder structure to extract shallow local features. Subsequently, the optimization problem is decoupled into alternating iterations of data subproblems and regularization subproblems through HQS. The data subproblem is updated by optimizing the objective function with competitive terms, while the regularization subproblem is solved by designing an unfolding network.

2) We devise a TRFSA network with an asymmetric UNet structure. The network consists of a Frequency-domain Self-Attention (FSA) module and a Discriminative Feedforward Network (DFFN). The FSA module computes frequency-domain attention to extract rich global dependencies, while the DFFN, through a gating mechanism, extracts deep features of key frequency components and further enhances the global correlations through non-linear transformations. Finally, a SENet block is embedded to strengthen channel-wise features.

2. Related works

The reconstruction model integrating CS and DL can be divided into two categories: prior based data-driven networks and model-based deep unfolding networks[6, 7].

For the optimization of traditional CS, one type of research utilizes prior information to design regularization terms, such as non-local self-similarity or low-rank priors, as demonstrated in methods like PANO [8] and NLR-CS [9]. Another type of research seeks to find appropriate sparse representations, such as DLMRI [10], which utilizes dictionary learning to represent the data and imposes L1 regularization for sparsity constraints. These model-based approaches typically force the results to adhere to observation and image priors by balancing the competition between data fidelity and regularization terms. The iterative algorithms for solving the above optimization problems include Iterative Soft-Threshold Algorithm (ISTA)[11], Approximate Message Passing Method (AMP)[12], and Decoupling Method. ISTA applies sparse constraints through L1 regularization and combines gradient descent with soft threshold operations to seek the optimal sparse representation; The AMP algorithm updates parameter estimates and residuals in each iteration, gradually approaching the true signal through linear transformation and soft threshold processing; Decoupling methods such as ADMM[13] and HQS[4, 7] improve computational efficiency by alternately minimizing data fidelity and regularization terms.

Data-driven models based on pure DL, such as ReconNet [14], can extract features through stacked convolutional layers and achieve end-to-end training to reconstruct images. Although stacking convolutional layers can expand the receptive field to extract more contextual information and abstract features, excessive stacking of convolutional layers poses problems such as

inefficiency, gradient vanishing, and overfitting[15]. ResNet[16] and UNet[5] both introduce skip connections to address this issue. DR2-Net[17] based on ReconNet utilizes ResNet to achieve more effective image reconstruction. These CNN based DL models have achieved significant results, but there are limitations in extracting global dependencies. The DL model based on Transformer can capture global dependency relationships. Zamir et al.[18] proposed a Transformer model based on dot product attention to extract different features along the channel dimension. Wang et al.[19] proposed a UNet based Transformer for image restoration using self-attention based on non-overlapping windows. However, its standard attention has a high computational complexity for MRI data with high dimensionality and small data volume.

Prior based data-driven networks can significantly reduce training data requirements and improve image reconstruction quality. For example, Hyun et al. incorporated structural similarity loss and frequency domain data consistency constraints into the UNet model, optimizing the reconstruction effect[20]. The PD-UNet proposed by Philipp et al.[21] pre-trains undersampling maps through UNet in the Sinogram domain and introduces Primal Dual networks to enhance sparse constraints. The CSNet[22] proposed by Shi et al. based on DR2-Net uses convolution and fully connected layers to map measurement data into sparse representations, and finally uses ResNet for reconstruction. Compared to ReconNet and DR2-Net, CSNet reduces block artifacts through prior constraints. Despite its aforementioned advantages, traditional priors often rely on manually designed constraints or specific sparse representations, and overly simplified assumptions about complex image priors may make it difficult to capture rich and complex image structures.

Model-based deep unfolding networks embed learnable modules into the iterative optimization process of traditional algorithms, adopting deep image priors (DIP) as implicit priors, and extracting statistical characteristics of images through the network structure [2]. The network parameters can be optimized according to specific tasks and observed data, improving the interpretability and reconstruction quality of the model. For instance, ADMM-CSNet[13] unfolds the ADMM algorithm steps of the CS optimization framework into learnable deep network modules, adaptively learning the optimal algorithm parameters and deep prior through end-to-end training, thereby achieving high-quality reconstruction. The ISTA-Net[23] converts the gradient descent and soft threshold operations of each iteration of iterative soft threshold algorithm (ISTA)[24] into learnable convolutional layers and soft threshold layers to achieve networked representation. These methods demonstrate the enormous potential of deep unfolding networks for adaptive learning of deep priors rather than relying on preset sparse transformations or manual priors to optimize image reconstruction. However, the CS deep unfolding model combined with pure CNN is difficult to learn the global dependencies within MRI data, so it cannot adapt to MRI data with different layer structures. The newly proposed GA-HQS[25] combines pyramid attention with multi-scale Transformers to design a prior network embedded in the HQS iteration step, enhancing the global dependency of features. It brings a new paradigm for solving deep priors.

3. Proposed method

We design a model based deep unfolding network TRFSA-HQS, which includes two aspects: 1) Designing the HQS-CS iterative backbone decoupling CS optimization problem and iteratively optimize the data and prior terms, illustrated in the upper parts of Fig. 1; 2) Designing a TRFSA network for deep feature extraction and global dependency modeling to solve for depth priors.

3.1 HQS-CS Iterative Framework

We design a model based deep unfolding network TRFSA-HQS, which includes two aspects: 1) Designing the HQS-CS iterative backbone decoupling CS optimization problem and iteratively optimize the data and prior terms, illustrated in the upper parts of Fig. 1; 2) Designing a TRFSA network for deep feature extraction and global dependency modeling to solve for depth priors. illustrated in the lower parts of Fig. 1.

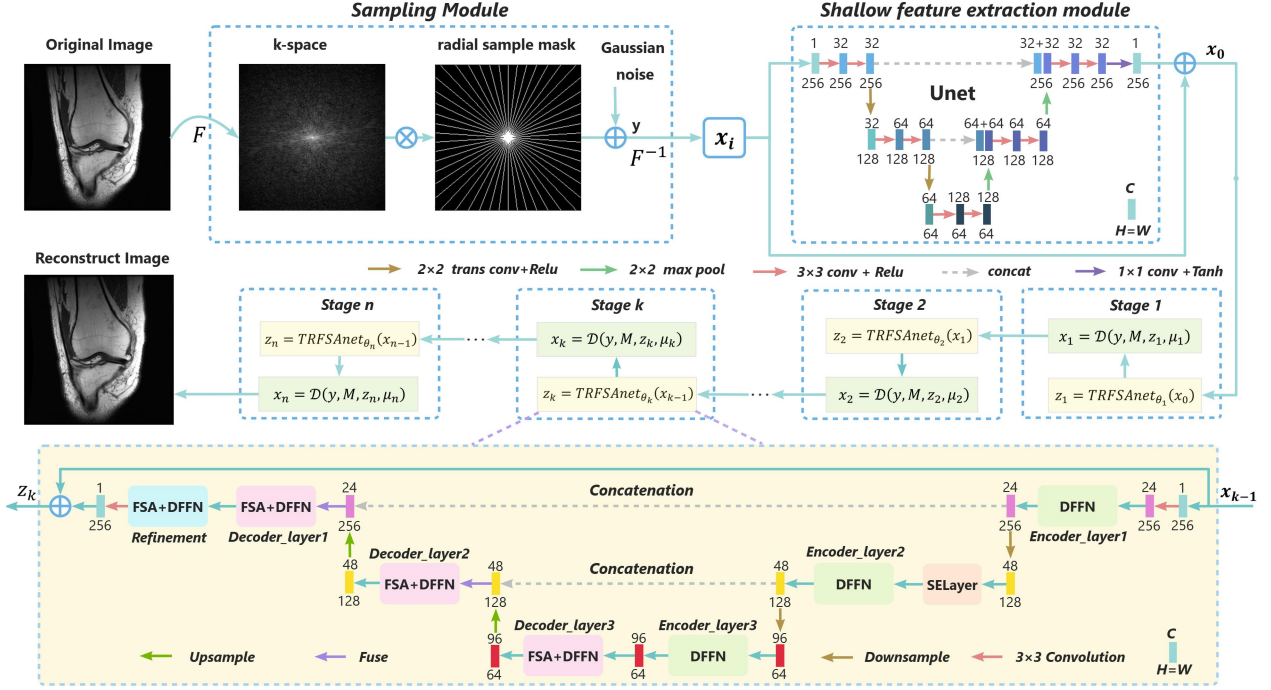


Fig. 1 Overall data flow diagram of TRFSA-HQS iterative unfolding framework

3.1.1 Radial Undersampling

Radial undersampling oversamples the low-frequency portion that determines the main contour and structural information of the image at the center, which helps to suppress artifacts and aliasing. Therefore, CS obtains undersampled data through radial zero filling, represented as:

$$y = F_u x + \varepsilon \#(1)$$

Where $x \in \mathbb{R}^{H \times W}$ represents the fully sampled MR image, $y \in \mathbb{R}^{H \times W}$ represents the undersampling data in k-space, F_u represents the undersampling operator in k-space, $F_u = \text{Mask} \cdot F$, Mask represents the radial mask in k-space, and F represents the Fourier transform matrix, ε Represent the Gaussian white noise accompanying data collection, as shown in the sampling module in Fig. 1.

3.1.2 Shallow Feature Extraction

The shallow feature extraction is shown in the UNet structure in Fig. 1. It uses UNet to extract abstract features and preserve details through encoders and decoders, respectively. Its encoder compresses the resolution through downsampling, thereby increasing the receptive field to extract high-level abstract features. Its decoder gradually restores resolution through upsampling, while fusing high-dimensional features of the same scale encoder through skip connections to extract advanced features and preserve details. The convolutional and transposed convolution in the UNet use Relu activation functions to enhance non-linear mapping and fitting capabilities. Finally, residual connections are used to compensate for lost image details. It can be formally expressed as:

$$x_0 = \mathcal{U}(x_i) + x_i \#(2)$$

Where $\mathcal{U}(\cdot)$ denotes Uet, x_i and x_0 respectively denote the input and output of the UNet.

3.1.3 HQS Iterative Optimization

For the above CS model, the model-based approach reconstructs it by solving the following optimization problem:

$$\hat{x} = \arg \min_x \frac{1}{2} \|y - F_u x\|_2^2 + \lambda R(x) \#(5)$$

where $R(\cdot)$ is the regularization term that introduces a prior constraint, λ is a regularization parameter, and $\|y - F_u x\|_2^2$ is fitting term for the data. HQS reduces the complexity of the

algorithm by decoupling the minimization of data fidelity and regularization terms. By introducing the auxiliary variable z , Eq. (5) can be written as the following constrained optimization problem:

$$\hat{x} = \arg \min_x \frac{1}{2} \|y - F_u x\|_2^2 + \lambda R(z), \text{ s.t. } z = x \quad (6)$$

The above problem is solved by constructing an augmented Lagrangian function using the Lagrangian multiplier method [7, [26] 27]:

$$\mathcal{L}_\mu(x, z) = \frac{1}{2} \|y - F_u x\|_2^2 + \lambda R(z) + \frac{\mu}{2} \|z - x\|_2^2 \quad (7)$$

where μ is the penalty parameter for the data smoothing term. Eq.(7) minimizes the two competing terms by alternating iterations. For the k -th iteration, taking z and x as constants respectively, the problem is decoupled into:

$$x_k = \arg \min_x \frac{1}{2} \|y - F_u x\|_2^2 + \frac{\mu}{2} \|x - z_k\|_2^2 \quad (8)$$

$$z_k = \arg \min_z \frac{\mu}{2} \|z - x_{k-1}\|_2^2 + \lambda R(z) \quad (9)$$

Eq.(8) and (9) represent the data subproblem and regularization subproblem, respectively [28]. Eq.(8) is solved by minimizing the L2 norm of the residual and gradient terms, and its closed form solution is as follows:

$$x_k = (F_u^H F_u + \mu I)^{-1} (F_u^H y + \mu z_k) \quad (10)$$

where F_u^H is the Hermitian conjugate transpose matrix of the Fourier sampling matrix F_u , and I is the identity matrix. Eq.(8) is further simplified by the Sherman-Morrison-Woodbury (SMW) matrix inversion formula [29], as shown below:

$$x_k = z_k + \frac{1}{1 + \mu} F_u^H (y - F_u z_k) \quad (11)$$

Eq. (9) is a convex quadratic optimization problem, which includes a fidelity term associated with the quadratic regularized least squares problem. It can be transformed to Eq. (12):

$$z_k = \arg \min_z \frac{1}{2(\sqrt{\lambda/\mu})^2} \|x_{k-1} - z\|_2^2 + R(z) \quad (12)$$

where the minimization of z represents the proximal operator of the regularization term. According to Bayesian probability theory, Eq. (12) can be interpreted as z_k being the denoised result of a Gaussian denoiser with noise level $\sigma = \sqrt{\lambda/\mu}$. Based on this theory, a series of Gaussian denoisers trained by CNN can serve as modules for image reconstruction [30]. Therefore, the proximal operator Eq. (12) can be simplified to the following form:

$$z_k = \text{Denoiser}(x_{k-1}, \sigma) \quad (13)$$

we uses the TRFSA network as a prior denoiser. Therefore, during the iteration process, Eq. (11) and (13) can be expressed as:

$$z_k = \text{TRFSAnet}(x_{k-1}) \quad (12a)$$

$$x_k = \mathcal{D}(y, M, z_k, \mu_k) \quad (12b)$$

where $\text{TRFSAnet}(\cdot)$ represents a prior subproblem, $\mathcal{D}(\cdot)$ represents a data subproblem. The complete data flow is shown in Fig. 1.

3.2 TRFSA unfolding networks

The TRFSA network adpots UNet with asymmetric encoder-decoder structure, and its encoder and decoder are composed of DFFN and FSA. The data flow is shown in the TRFSAnet unfolded in Fig. 1. The features between the same scale encoder and decoder layers are fused using Fusion. Finally, this article embeds SEnet [34] to extract channel attention in the encoder layer, forming an end-to-end trainable network. Its various components are shown in Fig. 2.

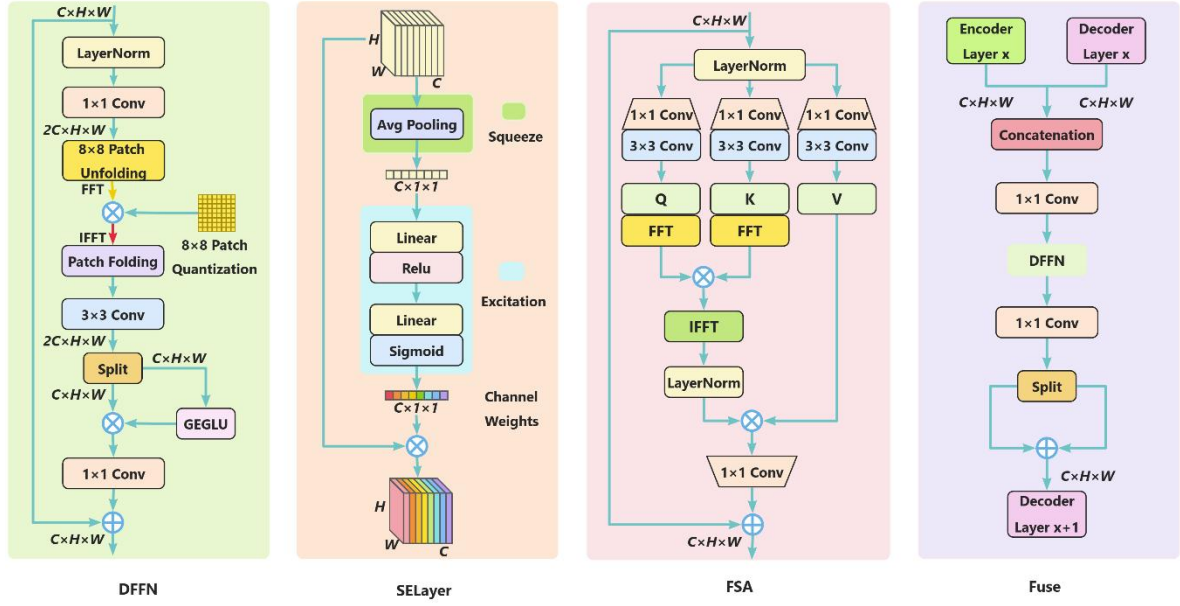


Fig. 2 Module diagram of TRFSA network components

3.2.1 Asymmetric UNet

Most existing UNet methods typically use symmetric architectures in encoder and decoder modules. We noticed that the features extracted by the encoder module often contain blurring effects compared to the deep global features extracted by the decoder module, which can alter the correct calculation of attention weights in the feature map and thus affect image restoration. Therefore, we only embed FSA into the decoder module to form an asymmetric structure for better image reconstruction, as shown in TRFSAnet in Fig. 1.

3.2.2 Discriminative Feedforward Network (DFFN)

The feedforward network mainly scales channels and increases receptive fields through convolutional operations to extract deep abstract features. It introduces a non-linear activation function to enhance the fitting ability of the model and to some extent alleviate the overfitting problem caused by the self-attention module's excessive attention to details. Collaborating with the self-attention module, the feedforward network can scale the dot product attention, enabling the Transformer to have powerful feature representation and semantic modeling capabilities. Given that there may be invalid frequency information for image reconstruction in k-space, we adopt a discriminative feedforward network (DFFN), as shown in Fig. 2. The network assigns learnable weights to the unfolded patches through a learnable quantization matrix, thereby making the extracted deep features more focused on key frequency regions. In terms of non-linear activation functions, we replace the Relu activation function with GEGLU[32] and utilize interaction with features to learn richer comprehensive features.

3.2.3 Frequency Domain Self-Attention (FSA)

The current Transformer first linearly maps a feature map x with a given spatial resolution of $C \times H \times W$ to the Query, Key, and Value feature spaces through 1×1 convolution, and then rearranges it to an $N \times N$ form feature map patch, i.e. $Q, K, V \in \mathbb{R}^{D \times N \times N}$. In this representation, the scaled dot product attention can be expressed as:

$$Attention(Q, K, V) = Softmax\left(\frac{Q \cdot K^T}{\sqrt{D}}\right) \cdot V \# (15)$$

Where N represents the number of patches and C represents the number of channels in the feature map, with a computational complexity of $O(N^2C)$. As the number of patches increases, the ability to capture global structural information will also be enhanced, but this will lead to an increase in computational complexity. In the calculation of dot product attention, each element of $Q \cdot K^T$ is

obtained by computing the inner product of vectors Q_i , K_j , which are the i -th row and j -th column vectors of matrices Q and K . By applying convolutional mapping to all Q_i and K_j , the convolution operation of $q_i * k_j$ can be used to obtain all elements of $Q \cdot K^T$ [33]. According to the convolution theorem, the convolution of two spatial signals is equivalent to the Hadamard product in the frequency domain[34]. Compared with matrix multiplication, the frequency domain components calculated by FSA contain global positional information, with a spatial complexity of $O(NC \log N)$. Therefore, calculating frequency domain attention can extract richer global correlation information. Based on this, we use 1×1 convolution to map to the Query, Key, and Value feature spaces, and then use 3×3 convolution to map to q_i and k_j . Subsequently, attention is calculated in the frequency domain through Fourier transform, and then transformed into the spatial domain to obtain correlation weights through layer normalization. The frequency domain attention output is obtained by multiplying it with V . Finally, the original feature channel is matched through 1×1 convolution, and residual connections are made to the input, as shown in FSA in Fig. 2.

3.2.4 Squeeze and Excitation Network (SEnet)

We inserted SENet between the encoder layers (see SELayer in Fig. 2). In this process, the Squeeze operation compresses the features of all channels into a 1×1 form. Then, through the Excitation operation, two linearly fully connected layers and an activation function are used to learn the contribution of each channel to the features, thereby obtaining the attention weights of each channel. The first linear layer uses Relu activation function to compress information, while the second linear layer uses sigmoid activation function to map features to the 0-1 range as channel weights. By weighting the channels of the original feature map, the output focuses more on channels with higher weights[32].

4. Experiments and Results

4.1 Experimental Settings

Full sampled T1WI coronal images of a single coil knee joint in the FastMRI dataset were used to validate the reconstruction performance of the proposed method. Randomly select 4000, 500, and 30 sub images in the dataset to divide into training, validation, and testing sets. Crop and sample all images to a grid pixel size of 256×256 , and normalize pixel values from $[0, 255]$ to between $[0, 1]$.

The hardware platform model used in the experiment is Intel (R) Xeon (R) Platinum 8255C CPU@2.50GHz and GeForce RTX 3080 GPU, and the software platform is Pycharm and PyTorch 1.7.0 framework. The sampling matrix M is set to a radiation mask with uniformly distributed angles in the complex frequency domain space of 256×256 , and the sampling noise standard deviation is set to $\sigma = 0.1$. The HQS iteration backbone is set to $n=4$ stages to represent the depth of unfolding, hyperparameters μ Initialize to 0.03 at each stage. The network is unfolded using a 3-layer asymmetric UNet. The number of patches for the learnable quantization matrix in the DFFN module is set to 8×8 . During the training phase, L2 loss is used, ADAM is used as the optimizer, momentum parameters (0.9, 0.99) are used, and the learning rate is set to 0.0002. The training epoch is set to 80, and due to the small dataset samples used in the experiment, the batch size is set to 1. The model adopts commonly used evaluation criteria: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate the reconstruction quality of different methods.

4.2 Result Comparison

In this section, TRFSA-HQS is compared with other competitive methods in both subjective visual and objective measurement aspects. Competitive methods include DLMRI, NLR, PANO, BM3D-AMP, and BM3D-MRI. At a sampling rate of 0.2, this article visually compares the above

methods in terms of details and errors in areas such as bones, joints, and muscles in the test set, as shown in Fig. 3.

In the figure, DLMRI results in more blurry edges and loss of details due to its inability to fully express all the detailed structures of the image. NLR generates inaccurate textures and contours due to its use of non-local regularization assumption that residuals have global dependencies, which weaken in image edges or complex texture regions. PANO, like NLR, has contour and detail errors, but does not have positional inaccuracies in detail textures. Because it guides reconstruction based on block similarity and non-local operation sorting, it retains more accurate structural information than NLR and thus has higher SSIM. However, it is difficult to handle the expression of detailed features within the block, resulting in significant detail errors. BM3D-MRI and BM3D-AMP have smaller contour and detail errors than NLR and PANO, as BM3D block matching and 3D transform filtering can effectively remove noise and preserve image details, but they produce many block artifacts. However, BM3D-AMP produces vertical artifacts in muscle texture, as AMP cannot directly sparsely represent signals through AMP when dealing with non-uniform radial undersampling. TRFAS-HQS reconstructs more accurate and sharp edge details, while exhibiting lower noise levels.

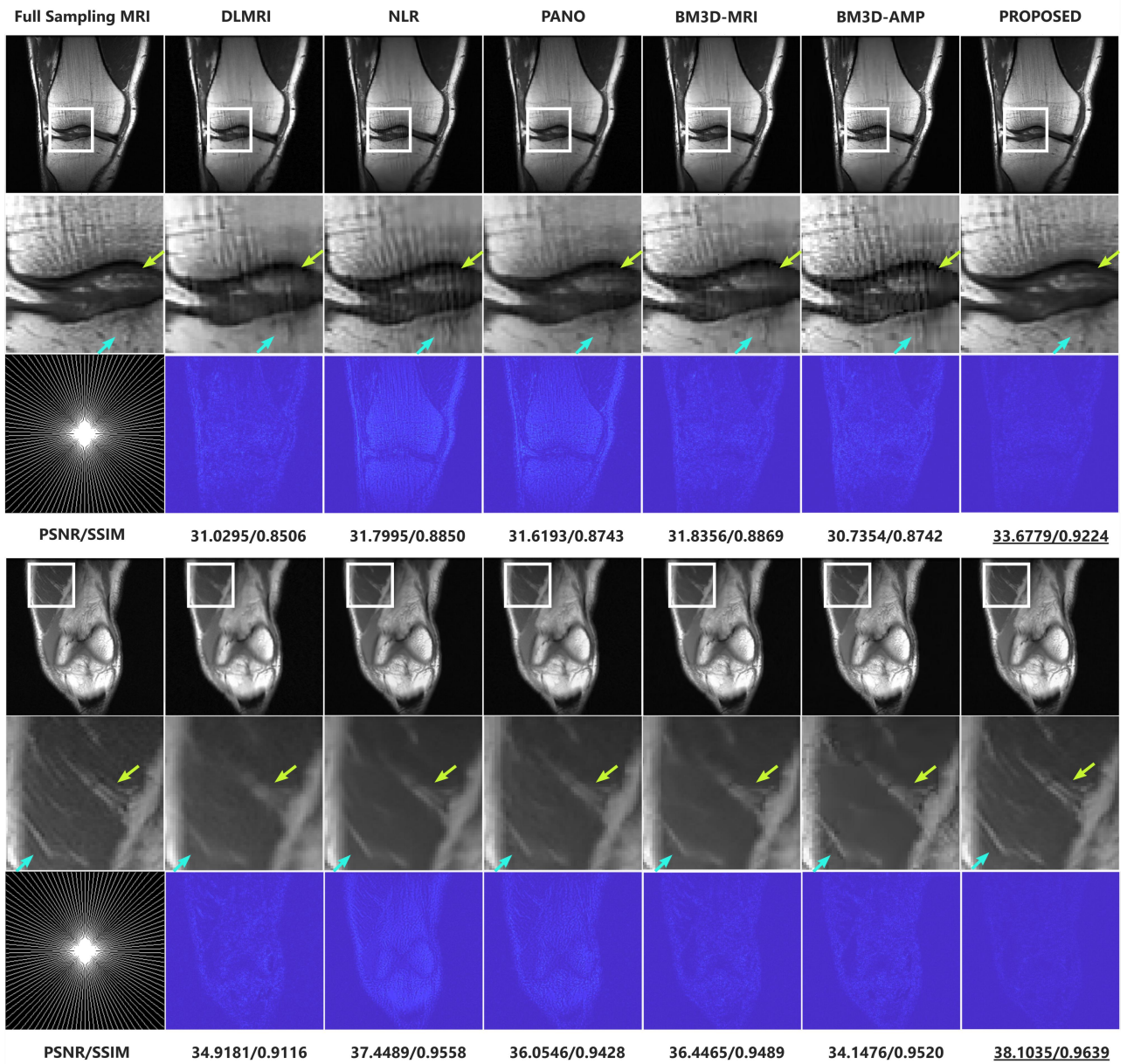


Fig. 3 Comparison of Reconstruction Results of Competitive Algorithm in Bone Joint and Muscle Textures

In terms of objective measurement, we selected four sampling rates (0.05, 0.10, 0.2, 0.30) to evaluate the reconstruction performance of each model. Table 1 quantitatively compare their effects in terms of PSNR, SSIM and reconstruction time. Among them, NLR has better reconstruction performance, but each iteration requires the calculation of gradient and second-order derivative matrix, so the inference time is longer. DLMRI requires training a dictionary and has a high computational load. In contrast, PANO and BM3D-MRI effectively utilize repetitive information through non-local block similarity sorting and block matching 3D filtering, respectively, with lower computational complexity. Compared to BM3D-MRI, BM3D-AMP requires iterative calculation of update rules for message passing, so the computational complexity is slightly higher than BM3D-MRI. Finally, compared with the aforementioned method, TRFSA-HQS significantly improves the average PSNR and inference speed of its reconstruction at four sampling rates.

Table 1. Reconstruction performance of competitive methods at different sampling rates

Sampling Rate / Radial lines		0.05/15	0.10/25	0.20/55	0.30/85	Time(s)
DLMRI	PSNR/ SSIM	24.29/0.6815	27.77/0.7733	32.66/0.8715	35.25/0.9102	284.93
BM3D-AMP	PSNR/ SSIM	24.89/0.7164	28.22/0.7924	33.85/0.8969	36.00/0.9362	183.24
PANO	PSNR/ SSIM	23.74/0.6328	26.95/0.7351	32.08/0.8647	35.16/0.9157	6.77
BM3D-MRI	PSNR/ SSIM	24.86/0.6795	28.72/0.8044	33.85/0.9069	36.58/0.9399	6.95
NLR	PSNR/ SSIM	24.73/0.6646	28.10/0.7549	33.99/0.9070	37.27/0.9446	53.09
TRFSA-HQS	PSNR/ SSIM	27.67/0.7894	30.39/0.8505	35.15/0.9301	37.90/0.9553	0.74

5. Discussions

5.1 Ablation Experiment

In this subsection, the effectiveness of FSA, DFFN, SENet and Asymmetric UNet (AMU) is discussed through ablation experiments. The experiments involve ablating the aforementioned modules in the alternating iterative backbone of HQS. Under the condition of one training epoch, the performance of the remaining parts of the model is evaluated. Table 2 represents the reconstruction performance of the remaining models on the validation set after ablating each structure.

Table 2. Comparison of model reconstruction performance after ablation

Sampling Rate	0.05	0.10	0.20	0.30	Avg.
TRFSA-HQS	25.59/0.700	28.60/0.791	32.52/0.885	35.68/0.930	30.60/0.826
Ablate FSA	25.39/0.696	28.29/0.781	32.37/0.872	35.36/0.926	30.35/0.818
Ablate DFFN	25.29/0.695	27.84/0.774	32.07/0.861	34.84/0.921	30.01/0.813
Ablate SENet	25.51/0.698	28.42/0.786	32.44/0.877	35.39/0.927	30.44/0.821
Ablate AMU	24.44/0.658	27.56/0.761	31.88/0.841	34.58/0.917	29.62/0.794

The experimental data demonstrates that the reconstruction performance of the models decreases after ablating each module at different sampling rates. Ablation of these modules resulted in an average decrease of 0.25/0.0081, 0.59/0.0135, 0.16/0.0046, and 0.98/0.0323 in PSNR/SSIM performance at four sampling rates, respectively. Asymmetric UNet directly affects the encoder-decoder layer as the Transformer backbone, showing the largest decrease in model reconstruction performance after alignment ablation. DFFN is responsible for deep feature extraction in the encoder and non-linear enhancement and feature transformation in the decoder, contributing significantly to the model's reconstruction performance. FSA highlights the key frequency features of MRI by capturing global dependencies in the frequency domain, making a contribution to the model's reconstruction performance secondary to DFFN. Meanwhile, SENet shows the smallest decrease in model reconstruction performance, confirming the effectiveness of each structure.

5.2 Convergence

In the convergence experiment, we verified the effects of the model's unfolding depth and training epoch on the reconstruction performance of the validation set. The noise standard deviation and undersampling rate used in the experiment were 0.1 and 0.2, respectively. Fig. 4(a) shows the trend of PSNR and SSIM of the validation set as the unfolding depth increases and epoch. Firstly, to verify the convergence of the model, we gradually increased the unfolding depth from 1 to 12 and fixed the training epoch to 1. The results indicate that as the unfolding depth increases, the reconstruction performance of the model continuously improves. After reaching a depth of 4, the performance gradually converges. In order to balance reconstruction speed and model complexity, the author ultimately set the unfolding depth to 4.

Next, with a fixed depth of 4, we gradually increase the training epoch from 1 to 80 and observe the trend of the model's reconstruction performance on the validation set, as shown in Fig. 4(b). As the number of epochs increases, PSNR and SSIM first rapidly increase, gradually converge after 20 epochs, and even slightly decrease. This indicates that excessive training can cause the model to learn some noise information, leading to overfitting.

The experimental data shows that selecting network unfolding depth and training epoch can enhance the model's expression ability, with good fitting effect and generalization ability.

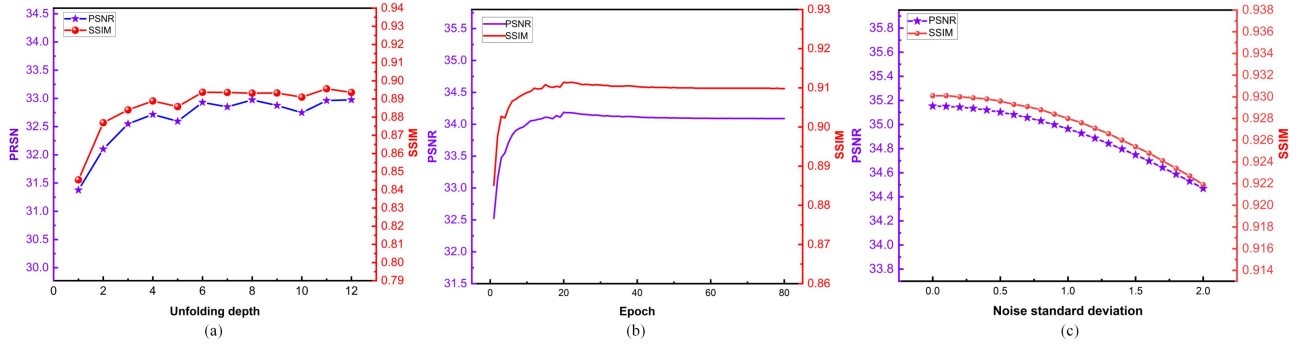


Fig. 4 The convergence and noise robustness of the model. (a) Reconstruction effects with different unfolding depths. (b) Reconstruction effects of different training epochs. (c) Reconstruction effect on data with different noise levels

5.3 Noise Robustness

MRI signals often face the issue of unstable noise conditions, primarily due to hardware electromagnetic interference and internal thermal noise, both of which follow Gaussian distributions. In the experiment, the noise intensity can be adjusted by setting the standard deviation of Gaussian distribution noise. Generally, in low noise scenarios, the standard deviation of noise is typically between 0.1 and 0.5; in moderate noise environments, it ranges from 0.5 to 1.0; while above 1.0 indicates high noise conditions. To test the reconstruction performance of the model under different noise levels, undersampled data with a noise standard deviation of 0.1 and a sampling rate of 0.2 are used for training and validation. Subsequently, the noise robustness of the model trained with a noise standard deviation of 0.1 is evaluated on different test datasets within the [0-2.0] noise standard deviation range.

As shown in Fig. 4(c), as the noise level of the test set increases, the reconstruction performance of the model decreases: When the standard deviation of noise is 0.5, the PSNR decreases by 0.05 dB; When the standard deviation reaches 1.0, the PSNR decreases by 0.19 dB; When the standard deviation reached 2.0, the PSNR decreased by 0.68 dB. From the results, it can be seen that the reconstruction performance of the model is relatively stable in low noise environments, and even in high noise data, the reconstruction performance of the model is still relatively high. This indicates that the model has strong noise robustness and can adapt to data with different noise levels.

6. Summary

This paper introduces a novel architecture based on the frequency-domain Transformer, referred to as TRFSA-HQS. It adopts a customized HQS optimization approach to iteratively decouple competitive terms for complexity reduction. In the unfolding network, we adopted asymmetric Uet, which embeds customized frequency domain patch discriminative feedforward networks and frequency domain self-attention in its encoder and decoder to extract abstract features and enhance the perception ability of key frequency information. Additionally, an SENet is integrated within the encoder layers to extract essential channel features. Conclusions drawn from ablation experiments, convergence experiments, and noise robustness experiments are as follows: 1. Within the model-based iterative backbone, each module of TRFSA-HQS demonstrates varying degrees of contribution. 2. Under the same sampling rate conditions, TRFSA-HQS outperforms other competing methods in both reconstruction performance and reconstruction time. 3. TRFSA-HQS exhibits high robustness when dealing with new samples in different noise environments. This paper aims to further expand the scope of MRI rapid reconstruction methods based on hybrid architectures of models and Transformers, thus enhancing the performance and applicability of magnetic resonance imaging technology.

References

- [1] Minghe S, Hongping G, Chao N, et al. TransCS: A Transformer-based Hybrid Architecture for Image Compressed Sensing.[J]. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, 2022, PP.
- [2] D. Ulyanov, A. V. Edaldi, and V. Lempitsky, "Deep image prior," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2018, pp. 9446–9454.
- [3] Pan, H., Zhu, X., Atici, S.F., Cetin, A.: A hybrid quantum-classical approach based on the hadamard transform for the convolutional layer. In: International Conference on Machine Learning. PMLR.p. 26891–26903 (2023)
- [4] Xin B, Phan T, Axel L, et al. Learned Half-Quadratic Splitting Network for MR Image Reconstruction[C]//International Conference on Medical Imaging with Deep Learning. PMLR, 2022: 1403-1412.
- [5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.
- [6] Li R, Luo L, Zhang Y. Convolutional Neural Network Combined with Half-Quadratic Splitting Method for Image Restoration[J]. Journal of Sensors, 2020, 2020:1-12. DOI:10.1155/2020/8813413.
- [7] Sun Y, Yang Y, Liu Q, et al. Learning Non-Locally Regularized Compressed Sensing Network With Half-Quadratic Splitting[J]. IEEE Transactions on Multimedia, 2020, PP(99):1-1. DOI:10.1109/TMM.2020.2973862.
- [8] Qu X, Hou Y, Lam F, et al. Magnetic resonance image reconstruction from undersampled measurements using a patch-based nonlocal operator[J]. Medical Image Analysis, 2014, 18(6):843-856. DOI:10.1016/j.media.2013.09.007.
- [9] Dong W, Shi G, Li X, et al. Compressive sensing via nonlocal low-rank regularization[J]. IEEE transactions on image processing, 2014, 23(8): 3618-3632.
- [10] Ravishanker S, Bresler Y. MR image reconstruction from highly undersampled k-space data by dictionary learning[J]. IEEE transactions on medical imaging, 2010, 30(5): 1028-1041.
- [11] Zibetti M V W, Helou E S, Pipa D R. Accelerating Over-Relaxed and Monotone Fast Iterative Shrinkage-Thresholding Algorithms with Line Search for Sparse Reconstructions[J]. IEEE Transactions on Image Processing, 2017, PP(99):1-1. DOI:10.1109/TIP.2017.2699483.
- [12] Eksioğlu E M, Tanc A K. Denoising amp for mri reconstruction: Bm3d-amp-mri[J]. SIAM Journal on Imaging Sciences, 2018, 11(3): 2090-2109.
- [13] Y. Yang, J. Sun, H. Li, and Z. Xu, "ADMM-CSNet: A deep learning approach for image compressive sensing," IEEE Trans. Pattern Anal. Mach. Intell. vol. 42, no. 3, pp. 521–538, Mar. 2020.

- [14] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "ReconNet: Non-iterative reconstruction of images from compressively sensed measurements," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 449–458.
- [15] Prakash A, Storer J, Florencio D, et al. RePr: Improved Training of Convolutional Filters[C]//Computer Vision and Pattern Recognition.IEEE, 2019.DOI:10.1109/CVPR.2019.
- [16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [17] Yao H, Dai F, Zhang D, et al. DR2-Net: Deep Residual Reconstruction Network for Image Compressive Sensing[J]. 2017.DOI:10.1016/j.neucom.2019.05.006.
- [18] Zamir, Syed Waqas, et al. "Restormer: Efficient transformer for high-resolution image restoration." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [19] Wang, Zhendong, et al. "Uformer: A general u-shaped transformer for image restoration." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [20] Hyun C M, Kim H P, Lee S M, et al. Deep learning for undersampled MRI reconstruction[J]. Physics in Medicine & Biology, 2018, 63(13): 135007.
- [21] Philipp E, Soumick C, Georg R, et al. Sinogram upsampling using Primal-Dual UNet for undersampled CT and radial MRI reconstruction.[J]. Neural networks : the official journal of the International Neural Network Society, 2023, 166.
- [22] W. Shi, F. Jiang, S. Liu, and D. Zhao, "Image compressed sensing using convolutional neural network," IEEE Trans. Image Process., vol. 29, pp. 375–388, 2020.
- [23] Zhang J, Ghanem B. ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2017.DOI:10.1109/CVPR.2018.00196.
- [24] Lee B., Ku B., Kim W.-J., Kim S., Ko H.. Denoising ISTA-Net: Learning based compressive sensing with reinforced non-linearity for side scan sonar image denoising[J]. Journal of the Acoustical Society of Korea, 2020, 39(4)
- [25] Jiang, Jiawei, et al. "GA-HQS: MRI reconstruction via a generically accelerated unfolding approach." 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023.
- [26] W. Abdul, Hameed, et al. Augmented Lagrange Multiplier Method to Solve Quadratic Programming Problems in Standard Form: A Neural Network Approach[J]. Research journal of pharmacy and technology, 2016, 9(10): 1727-1731.
- [27] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," IEEE Trans. Image Process., vol. 20, no. 3, pp. 681–695, Mar. 2011.
- [28] Zhang K, Gool L V, Timofte R. Deep unfolding network for image super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3217-3226.
- [29] Afonso M V, Bioucas-Dias J M, Figueiredo M A T. Fast image recovery using variable splitting and constrained optimization[J]. IEEE transactions on image processing, 2010, 19(9): 2345-2356.
- [30] K. Zhang, W. Zuo, S. Gu and L. Zhang, "Learning Deep CNN Denoiser Prior for Image Restoration," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2808-2817, doi: 10.1109/CVPR.2017.300.
- [31] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
- [32] Shazeer, Noam. "GLU Variants Improve Transformer." (2020).
- [33] Kong, Lingshun, et al. "Efficient frequency domain-based transformers for high-quality image deblurring." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
- [34] Song, Tianyu, et al. "Exploring an efficient frequency-guidance transformer for single image deraining." Signal, Image and Video Processing (2023): 1-10.