IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2023, accepted 11 July 2023, date of publication 1 August 2023, date of current version 10 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300789

RESEARCH ARTICLE

GLFormer: An Efficient Transformer Network for Fast Magnetic Resonance Imaging Reconstruction

RONGQING WANG, MENGDIE SONG, (Graduate Student Member, IEEE), JIANTAI ZHOU, AND BENSHENG QIU, (Member, IEEE)

Centers for Biomedical Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China

Corresponding author: Bensheng Qiu (bqiu@ustc.edu.cn)

This work was supported by the Fundamental Research Funds for the Central Universities under Grant WK5290000003.

ABSTRACT Deep learning (DL)-based methods substantially enhance the speed of magnetic resonance imaging (MRI). Recently, transformer network architectures have been increasingly applied to image reconstruction owing to their exceptional ability to model long-range dependencies. However, directly employing a transformer network for MRI reconstruction results in a considerable computational burden because the computational complexity of the transformer is proportional to the square of the image spatial resolution. To alleviate this limitation, this study aims to design a computationally efficient transformer network with improved reconstruction performance. The proposed network, termed the global-local-transformer (GLFormer), is based on a multi-input multi-output architecture consisting of three components. A simplified self-attention, global attention is designed to extract the long-range dependency using a global pooling operator while maintaining linear complexity. Furthermore, depth convolution is incorporated into a feedforward network (FFN) to perform local feature aggregation, and a parallel-gated branch is designed for the FFN, thereby enhancing the effectiveness of representation learning and improving the reconstruction performance. To enhance the ability of the network to perceive frequency information, a deep frequency attention module is proposed to adaptively decompose and adjust frequency domain features, thereby enhancing the reconstruction performance. Experiments conducted on public datasets indicate that GLFormer outperforms state-of-the-art DL-based methods for different undersampling rates and types of undersampling patterns. Furthermore, GLFormer only exploits fewer model parameters and has a lower computational burden (i.e., 2.4 M and 19G) than the previous methods, while maintaining high reconstruction quality.

INDEX TERMS Magnetic resonance imaging, deep learning, deep frequency attention, transformer.

I. INTRODUCTION

Magnetic resonance imaging (MRI) is widely used in clinical settings because of its excellent soft tissue contrast, low radiation levels, and non-invasiveness. Despite these significant advantages, MRI is restricted by the prolonged scanning time. Fast MRI relies heavily on image reconstruction from undersampled k-space data using rapid imaging sequences [1], parallel imaging [2], and compressed sensing [3]. However, these conventional methods are hindered by limited acceleration factors and slow nonlinear optimization processes.

Recently, deep learning (DL)-based methods have been exceptionally effective in MRI reconstruction, and have found extensive applications in commercial systems. Wang et al. [4] were the first to utilize convolutional neural networks (CNN) for MRI reconstruction, which builds a mapping between undersampled MR images and fully sampled reconstructions. Schlemper et al. [5] introduced a deep network called DCCNN, which comprises a cascade of CNNs, to reconstruct MR images from undersampled

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

data. Yang et al. [6] proposed an end-to-end reconstruction model based on conditional generative adversarial networks (DAGAN), that utilize U-Net as the generator.

A vision transformer [7] demonstrated superior performance owing to the global receptive field characteristics of the self-attention mechanism. Consequently, several studies have investigated the potential of transformer models for MRI reconstruction. For instance, Kang et al. [8] added a reconstruction head to a vanilla transformer network and directly applied it to magnetic resonance image reconstruction. Huang et al. [9] proposed a reconstruction framework that utilizes a swin transformer as the backbone for fast MRI reconstruction. They leveraged the advantages of the swin transformer for image recognition and achieved state-of-the-art results in MRI reconstruction. Similarly, Guo et al. [10] introduced a texture transformer module for accelerated MRI reconstruction, which captured the textural information in MRI images and achieved excellent performance in reducing reconstruction errors. Zhao et al. [11] proposed a swin-transformer-based dual-domain generative adversarial network (SwinGAN) consisting of a frequency domain generator and an image-domain generator for accelerated MRI reconstruction. Finally, MTrans [12] is an accelerated multimodal MRI technique with a cross-attention module that extracts and fuses complementary features from an auxiliary modality with the target modality.

Despite the promising results of previous studies, the direct application of transformer networks to MRI reconstruction has several limitations. First, traditional self-attention mechanisms require the computation of attention maps, which entails quadratic complexity, resulting in considerable computational burden and memory usage. This, in turn, hinders the further application of transformer networks in MRI. Second, linear projection layers in feedforward networks (FFN) only aggregate information within channels, which hinders their ability to capture spatially local information, restricting the capacity of the network to represent local features. Finally, the existing neural network methods generally perform poorly in recovering high-frequency details when reconstructing magnetic resonance images. Previous research has suggested that this may be related to the frequency preferences of the neural networks [13], [14].

In this paper, we propose a global-local transformer (GLFormer), a computationally efficient and high-performance transformer network for fast MRI reconstruction. Our approach combines both global and local information from MRI data to improve reconstruction accuracy while maintaining computational efficiency. The main contributions are as follows:

- Obtaining a global feature vector through pooling simplifies the query cost of attention maps, reduces square complexity to linear complexity, and preserves the capture of global features.
- The aggregation of local spatial features through depth convolution while improving the information flow of the network through a gating mechanism improves representation learning.
- A deep frequency attention module adaptively adjusts the frequency information distribution of deep features to enhance the final reconstruction performance.

II. RELATIVE WORKS

A. CLASSIC MRI RECONSTRUCTION

In the classical framework of compressed sensing magnetic resonance imaging (CS-MRI) reconstruction, the acquisition of undersampled MRI signals follows a specific forward process:

$$y = F_u x + \epsilon. \tag{1}$$

where F_u is a subsampled Fourier transform, which is a combination of the Fourier transform and a binary sampling operator, x is the fully sampled MR image, y refers to the acquired undersampled data and ϵ is the noise introduced during the acquisition process. MRI reconstruction can be viewed as an inverse problem, as follows:

$$\min_{x} \ \frac{1}{2}||F_u x - y||_2^2 + \lambda||x - F_{cnn}(x_u|\theta)||_2^2. \tag{2}$$

where $||F_u x - y||_2^2$ is the data confidelity term, that ensures that the reconstructed image x conforms to the forward process, and $||x - F_{cnn}(x_u|\theta)||_2^2$ is the difference between the reconstructed image, denoted as $F_{cnn}(x_u|\theta)$, generated by the neural network, and the ground truth image x. Here, x_u represents the downsampled input image, and θ represents the learnable parameters of the network.

B. TRANSFORMER

The performance of CNNs is frequently limited by two main factors: limited receptive fields and the inability to learn instance-level features [15]. Transformers have undergone significant development in numerous fields because of their capacity for modeling long-range dependencies [16], [17]. Although the transformer model has been shown to be a notable advancement over CNNs by addressing their drawbacks, including limited receptive fields and inadequate adaptability to input content, the computational complexity of the transformer model increases quadratically with spatial resolution. Therefore, applying a transformer network directly to magnetic resonance (MR) reconstruction results in high computational complexity.

C. FREQUENCY LEARNING

Frequency learning, which involves filtering and image compression methods, is a well-established technique in traditional image processing [18], [19]. More recently, frequency learning has been applied to deep learning tasks. For example, zhang et al. proposed a frequency-suppression module that removes high-frequency components to enhance the robustness of classification [20]. Similarly, Chen et al. [21] designed a frequency enhancement module that incorporated

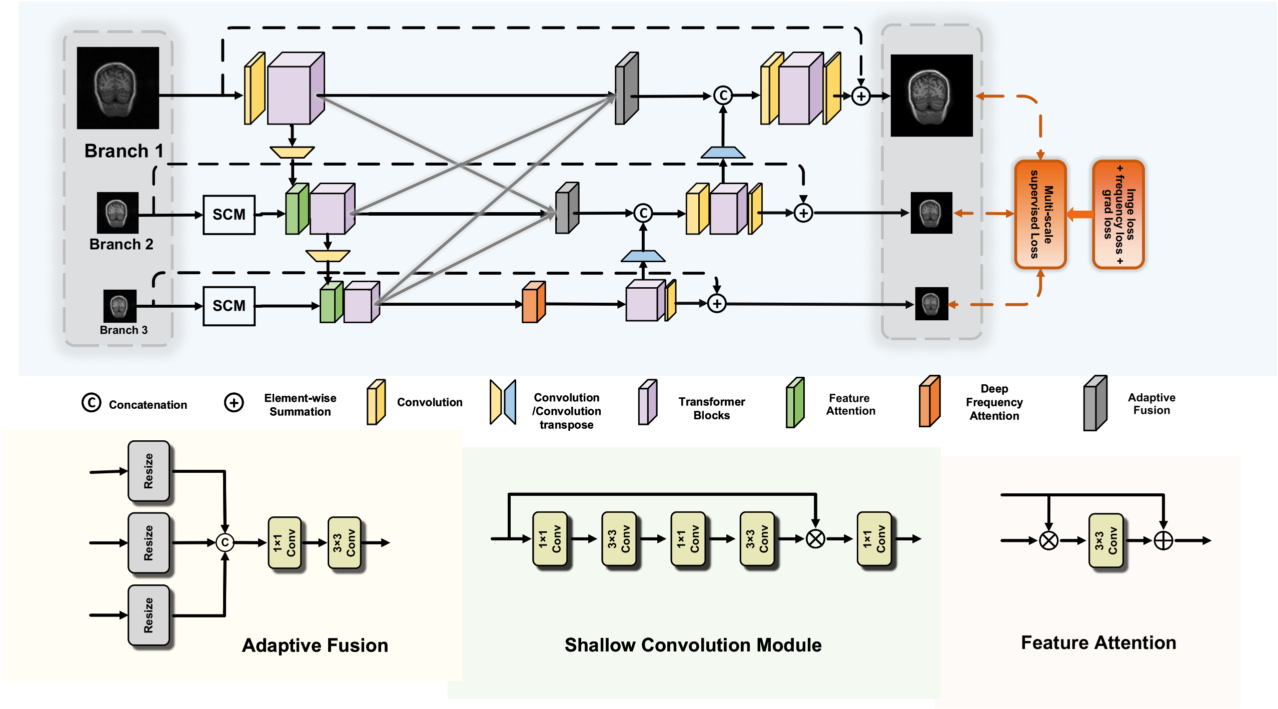


FIGURE 1. Architecture of the proposed GLFormer for MRI reconstruction.

frequency-aware cues into CNN models. In F3-Net [22], the input image is adaptively partitioned by a learnable frequency filter, revealing a preference for specific frequencies in both the CNN and transformer models. Bai et al. [23] divided the features into high and low frequencies, and learn different frequencies independently. Dar et al. [24] introduced high-frequency information priors to aid in the MRI reconstruction process. In contrast to the existing methods, we propose a novel approach called the deep frequency attention (DFA) module. The DFA module permits flexible adjustments of the frequency domain representations of deep features, resulting in a superior recovery of high-frequency details.

III. METHODS

A. MULTI-INPUT-MULTI-OUTPUT ARCHITECTURE

Inspired by MIMO-Unet [29], our architecture follows a MIMO design by integrating multiple-resolution inputs and outputs. The term ''multi-resolution inputs'' refers to the original input image and downscaled images obtained through bilinear interpolation downsampling at two times and four times lower resolutions. In the input stage, the encoder initially extracts shallow features from the input image, which then undergoes n transformer blocks. Next, we extract the low-resolution features from the downsampled image using a shallow convolutional module [29]. Subsequently, the downsampled deep features and low-resolution features are fused together by a feature attention module [29] and passed through the transformer block once again, and this process continues iteratively. The decoder receives deep information from the same branch and combines it with the deep features from other branches by an adaptive fusion module [29] to obtain a more comprehensive set of features. In the output stage, each decoder generates a reconstructed image at a specific resolution. Since the output of each decoder consists of a set of feature maps, we employ a convolutional operation to map the feature maps to an intermediate output image. The intermediate outputs enable the model to better learn image features and details at different resolutions, facilitating its adaptation to diverse scales and structures in the image. An overview of the proposed network is shown in Fig. 1.

B. GLOBAL ATTENTION

The traditional self-attention layer has $O(H^2W^2)$ complexity for a $W \times H$ image [25]. Therefore it is difficult to apply it to a reconstruction task. To address this limitation, we propose a global attention module with linear complexity, as shown in Fig. 2. Considering the sparse nature of magnetic resonance images, an attention map with quadratic complexity is unnecessary. The main component of our module is a global context feature generated by a pooling operator, which requires only linear complexity for attention map queries.

From the layer-normalized tensor $X \in R^{C\times W\times H}$, the global attention module first generates a query (Q), key (K), value (V) by applying 1×1 convolutions. Next, a global context feature is obtained through average pooling.

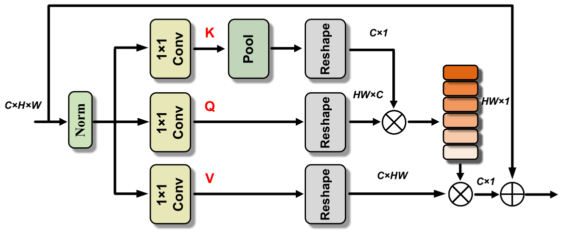


FIGURE 2. The structure of global attention module.

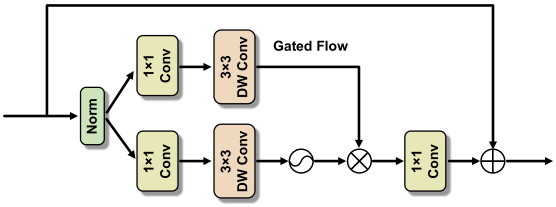


FIGURE 3. The details of local gated feedfoward network.

This global feature represents the expected information of the features and is characterized by its global nature. Subsequently, we multiply it by Q to obtain an attention map with linear complexity $O(HW)$, instead of a large regular attention map of size $O(HW \times HW)$. Overall, the global attention process is defined as:

$$X_{out} = X_{in} + Attention(Q, K, V). \tag{3}$$
$$Attention(Q, K, V) = softmax(Q, Pooling(K))V. \tag{4}$$

where X_{in} and X_{out} are the input and output feature, $Q \in R^{C\times HW}$, $K \in R^{C\times 1}$, $V \in R^{HW\times C}$

C. LOCAL-GATED FEED FORWARD NETWORK

The traditional FFN performs a nonlinear transformation of a feature representation into a high-dimensional space at each position via fully connected layers and a nonlinear activation function. Subsequently, this high-dimensional representation is remapped back to the feature dimension. Nonetheless,the linear mapping present in the FFN is equivalent to a 1 × 1 convolution, which causes inadequacies in the ability of the network to extract local features in the transformer structure. The design of the local-gated feedforward network (LGFFN) is illustrated in Fig. 3. In this paper, we propose two modifications to enhance the feature extraction capabilities of the FFN architecture. First, we introduce a 3 × 3 depth convolution [26] after a 1 × 1 convolution layer to capture interpixel information in the vicinity, resulting in an increased ability to extract superior local image structures during the reconstruction process. Second, the gating selection mechanism consists of parallel branches of a 1 × 1 convolution and a 3 × 3 depth convolution applied to high-dimensional features, followed by an element-wise multiplication operation. This mechanism effectively enables the network to learn the relevant information flow required for the current level, resulting in a highly competitive learning process. The entire LGFFN is formulated as follows:

$$X_{out} = W_p Gating(X_{in}) + X_{in}. \tag{5}$$
$$Gating(X_{in}) = \phi(W_p W_d LN(X_{in})) \odot W_p W_d((LN(X_{in}))). \tag{6}$$

where X_{in} denotes the input features. W_p is the 1×1 convolution used for mixing channels, and W_d is the 3×3 depth convolution used for aggregating local features. ϕ represents the non-linear activation function Gelu, and $\odot$ denotes the element-wise multiplication used to implement the gate mechanisms.

D. DEEP FREQUENCY ATTENTION

To facilitate the network's ability to perceive a variety of frequency information and precisely reconstruct the image content, especially high-frequency details, we propose a deep-frequency attention module based on frequency domain learning, as illustrated in Fig. 4. Previous methods for learning in the frequency domain typically involved decomposing images or features into different frequency spaces to address issues, such as smoothing or making explicit adjustments to certain frequency components. However, these approaches have limited flexibility and adaptability. Our proposed method involves both frequency domain feature extraction and adaptive frequency recombination, which enables the decoupling of all the frequency components and implicitly allows for flexible adjustments. To acquire a comprehensive representation of the frequency space, we utilized a Fast Fourier Transform (FFT) [27] to convert the image features into the frequency domain, as expressed by Eq. (3):

$$\mathbf{X}_F(x, y) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{X}(h, w) e^{-j2\pi\left(x\frac{h}{H} + y\frac{w}{W}\right)}. \tag{7}$$

Given a spatial feature X with dimensions (CHW), we apply a Fourier transformation to each channel of the feature to obtain the corresponding frequency domain representation. We utilize the torch.rfft function from the PyTorch framework, which separates the real and imaginary components of the complex numbers and takes advantage of the conjugate symmetry property in the frequency domain. As a result, we only utilized half of the input size. The resulting transformed feature, denoted by X_F, has dimensions of (2C, H/2, W/2), which allows for the analysis of image characteristics in the frequency domain. Our method provides two benefits for the acquisition of frequency domain representations. First, the implementation of FFT effectively decouples the image space into all frequencies, thereby producing a more comprehensive frequency representation. Secondly, each frequency component in the frequency representation is the summation of all image points, resulting in operations in the frequency space being more global in nature. Next, we introduce the frequency attention operation, that adaptively modulates different frequency components by generating two masks along the channel and spatial dimensions of X_F. Specifically, we applied a 1×1 convolution to project X_F onto a spatial

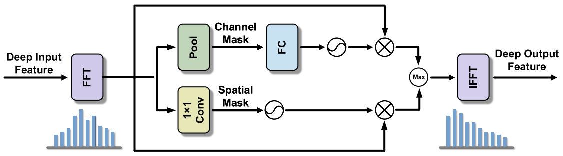


FIGURE 4. The details of deep frequency attention module.

embedding space, enabling a spatial frequency modification. Channel-dimension embedding was simultaneously implemented using a pooling operator. This allows the adaptive generation of two masks that can effectively modulate the contribution of each frequency component and enhance the performance of the network. The proposed deep frequency attention can be formulated as:

$$X'_F = Maxout(X_F \otimes S_{mask}(X_F), X_F \otimes C_{mask}(X_F)). \tag{8}$$

S_{mask} applies a 1×1 convolution to aggregate pixels channel-by-channel, followed by a sigmoid function. C_{mask} uses a pooling layer to spatially aggregate the pixels, followed by a fully connected layer and a sigmoid function. These masks allow for effective modulation of the contribution of each frequency component, thereby enhancing network performance. Our frequency attention embedding operation was primarily inspired by CBAM [28]. Finally, we utilized an inverse Fourier transform to convert the adjusted frequency features into the image domain. These features were subsequently inputted into the following network for further processing. The corresponding mathematical expression for the inverse Fourier transform is given by Eq. (9).

$$\mathbf{X}(h, w) = \frac{1}{H \times W} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X'_F(x, y) e^{j2\pi\left(x\frac{h}{H}+y\frac{w}{W}\right)}. \tag{9}$$

To maintain consistency with the lightweight objective of this study, we incorporate the deep frequency attention module into the third branch of the network, which exhibits reduced parameter count and computational burden on deep-level features.

E. MULTI-SCALE SUPERVISED LOSS

The selection of an appropriate loss function is critical for low-level vision tasks because it can significantly affect model performance. The use of different loss functions during training can result in widely divergent outcomes within the same model [30]. In this study, we employed the L1 loss, which has been demonstrated to be more effective for image restoration. Previous research has shown that this loss function exhibits superior convergence characteristics compared to the L2 loss [30].

To maintain consistency in our multi-input multi-output structure, we designed a multiscale loss function. By incorporating intermediate outputs, this function provides additional loss signals to the model, thereby expediting the convergence and mitigating the risk of overfitting. We used three types of multiscale supervised L1 losses: image, frequency, and gradient losses.

The first loss function we introduce is content loss, which is defined as follows:

$$L_{content} = \sum_{k=1}^{K} \frac{1}{t_k} ||\hat{x}_k - x_k||. \tag{10}$$

where $\hat{x}_k$ represents the reconstructed image, while x_k denotes the ground truth label. k is the number of branches. We divide the loss by the number of total elements t_k for normalization.

TABLE 1. Parameters of compared methods.

Method	DCCNN	DAGAN	SwinMR	GLFormer
Params(M)	**1.2**	171.88	11.4	2.4
FLOPs(G)	78.4	28.54	113	**19**

Because the primary objective of MRI reconstruction is to restore lost frequency components, it is crucial to minimize the differences in the frequency space. To achieve this, we propose a multiscale frequency reconstruction loss function, which can be expressed as:

$$L_{freq} = \sum_{k=1}^{K} \frac{1}{t_k} ||F(\hat{x}_k) - F(x_k)||. \tag{11}$$

where F denotes the fast Fourier transform (FFT) that transfers the image signal to the frequency domain.

Moreover, to restore the high-frequency details of the image better, we introduce a multilevel gradient loss, defined as follows:

$$L_{grad} = \sum_{k=1}^{K} \frac{1}{t_k} ||\nabla\hat{x}_k - \nabla x_k||. \tag{12}$$

The symbol ∇ denotes the gradient operator. The overall loss function for training the network is determined as follows:

$$L = \alpha L_{grad} + \beta L_{content} + \gamma L_{freq}. \tag{13}$$

where α, β, and γ are the weights that control the balance each loss.

IV. EXPERIMENT

First, we tested our model on the MICCAI 2013 Grand Challenge,[1] IXI[2] and fastmri datasets [31] to verify its reconstruction performance. To validate the effectiveness of the proposed module, ablation experiments were conducted on each module.

A. DATASETS

From the MICCAI 2013 Grand Challenge dataset, 100 T1-weighted brain MRI datasets were chosen and split into two groups:16,095 2D images for training and 5,033 valid 2D images for validation. In the testing phase, we

[1] http://masiweb.vuse.vanderbilt.edu/workshop2013/index
[2] http://brain-development.org/ixi-dataset/

TABLE 2. Quantitative assessment of PSNR, SSIM, NMSE ($\times 10^{-2}$) of the comparison methods in the MICCAI 2013 brain dataset, using different undersampling rates of Spiral 2D undersampling mask.

Sample rates	Metrics	ZF	DAGAN	DCCNN	SwinMR	GLFormer
10%	NMSE	0.132	0.083	0.082	0.052	**0.059**
	PSNR	36.06	42.12	43.06	43.98	**44.23**
	SSIM	0.352	0.901	0.922	0.944	**0.952**
20%	NMSE	0.146	0.027	0.027	0.020	**0.019**
	PSNR	35.19	42.53	43.47	50.12	**50.92**
	SSIM	0.609	0.980	0.981	0.994	**0.995**
30%	NMSE	0.089	0.035	0.023	0.023	**0.021**
	PSNR	38.00	45.03	46.69	50.12	**50.70**
	SSIM	0.652	0.990	0.990	0.998	**0.999**

TABLE 3. Quantitative assessment of PSNR, SSIM, NMSE ($\times 10^{-2}$) of the comparison methods in the MICCAI 2013 brain dataset, using different undersampling rates of Gaussian 1D undersampling mask.

Sample rates	Metrics	ZF	DAGAN	DCCNN	SwinMR	GLFormer
10%	NMSE	0.063	0.032	0.033	0.024	**0.011**
	PSNR	24.01	29.75	29.68	32.43	**33.03**
	SSIM	0.574	0.912	0.893	0.938	**0.945**
20%	NMSE	0.035	0.014	0.015	0.008	**0.002**
	PSNR	29.09	36.98	36.39	41.98	**42.21**
	SSIM	0.688	0.964	0.963	0.989	**0.990**
30%	NMSE	0.031	0.010	0.011	0.006	**0.001**
	PSNR	30.24	39.83	39.04	42.52	**43.26**
	SSIM	0.717	0.976	0.969	0.992	**0.994**

TABLE 4. Quantitative assessment of PSNR, SSIM, NMSE ($\times 10^{-2}$) of the comparison methods in the IXI brain dataset, using different undersampling rates of Radial 2D undersampling mask.

Sample rates	Metrics	ZF	DAGAN	DCCNN	SwinMR	GLFormer
20%	NMSE	7.8	1.02	0.81	0.36	**0.34**
	PSNR	21.37	32.22	33.39	37.22	**37.89**
	SSIM	0.480	0.921	0.922	0.950	**0.952**
30%	NMSE	4.10	0.50	0.42	0.18	**0.14**
	PSNR	24.81	35.30	35.42	41.21	**42.10**
	SSIM	0.603	0.974	0.933	0.974	**0.975**

TABLE 5. Comparison experiments in the fastmri dataset.

Method	PSNR	SSIM
SwinMR	30.25	0.81
GLFormer	**32.69**	**0.85**

used 50 independent datasets containing 9,854 2D images. From the IXI dataset, 373, 92, and 116 T1-weighted brain images were randomly selected for training, validation, and testing, respectively. The fastmri dataset consists of raw k-space data, including real acquired noise. From the knee data, we randomly selected 7000 slices as training data, 2000 slices as validation data, and 4000 slices as testing data. The input images were cropped to a size of 256×256 and fed into the network for reconstruction. For each dataset, images from a single subject were kept separate for training, testing, and validation, ensuring the independence of training and testing procedures. All settings were consistent with those used in previous studies.

B. IMPLIMENTATION DETAILS

For model training, we used the Adam optimizer [32]. The learning rate was initially set to 10e-4 and decreased by a factor of 0.5 every 30 epochs. There are eight transformer blocks. To prevent overfitting, training was terminated when the validation loss did not decrease over 5 epochs. A uniform set of hyperparameters was adopted for different sampling rates and markers: batch size $= 8$, $\alpha = 1$, $\beta = 0.1$, $\gamma = 1$. All experiments were conducted on a GTX 3090 GPU with 24 GB of memory using the PyTorch framework.

To simulate undersampling, we obtained undersampled k-space data using one-dimensional (1D) Gaussian, Cartesian, radial, and spiral masks. Each used different undersampling rates of 10%, 20%, and 30%, representing accelerations of 10×, 5×, and 3.3×, respectively. To ensure a fair comparison, we chose zero-filled images as inputs throughout the whole training process.

C. EVALUATION METRICS

The most important evaluation criterion for image restoration is the image quality [33], so three experimental metrics are used: the normalized mean square error (NMSE), peak signal-to-noise ratio (PSNR) [34], and structural similarity index (SSIM) [35]. The PSNR and SSIM indices are commonly used metrics for assessing image reconstruction quality. Both involve pixel-wise comparisons between the fully sampled

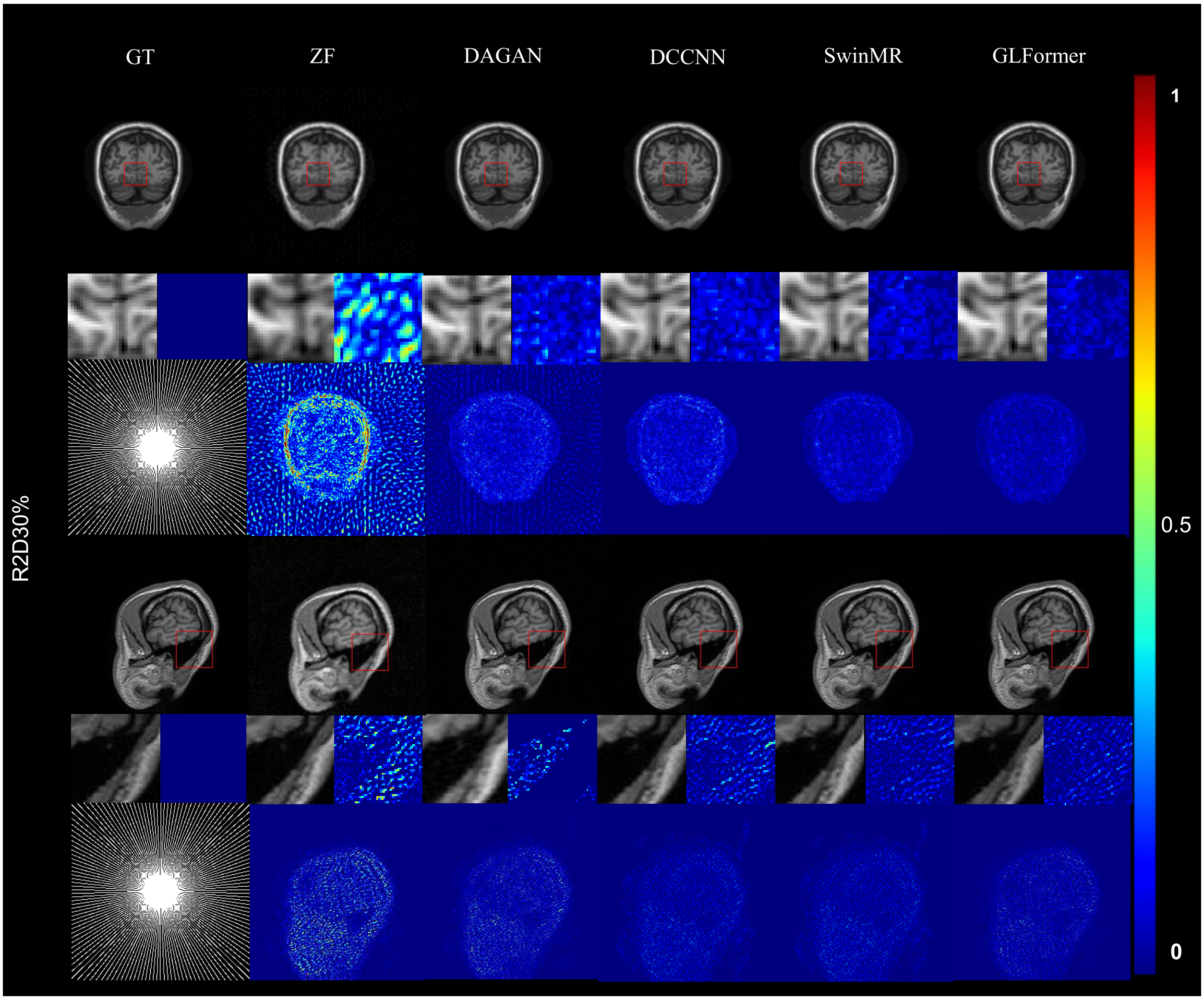


FIGURE 5. Qualitative visualization of different methods using Radial 30% mask (R2D30%), For every three rows of figure, row 1: GT(ground truth image), ZF(zero-filled image) and reconstructed images of different methods; row 2: The zoomed-in images and corresponding error maps which are pointed out in the first row by a red rectangle; row 3:The error maps of reconstructed images (10×). The upper part shows the quantitative results of the MICCAI2013 dataset, and the lower part shows the quantitative results of the IXI dataset.

MR image and the reconstructed image. PSNR calculates the ratio between the maximum possible signal power and the power of the corrupting noise, as measured by the mean squared error (MSE). The SSIM index is a symmetrical measure that considers the interdependence between pixels in an image and the mean and variance of the image intensities.

D. COMPARISON WITH OTHER METHODS

To validate the effectiveness of GLFormer, we evaluated four established models: DAGAN (a GAN-based model) [6], DCCNN (a CNN-based model) [5], and SwinMR (a transformer-based model) [9].

Table 1 lists the parameters of the four methods. It is evident that the proposed network has fewer model parameters and fewer FLOPs especially compared to SwinMR. This is attributed to the design of an efficient global attention module and LGFFN module.

We performed the quantitative experiments on the MICCAI 2013 dataset using a 1D Gaussian (shown in Table 2) and a spiral undersampling mask (shown in Table 3). Additionally, we performed experiments on the IXI dataset with a radial undersampling mask (shown in Table 4). The best results are indicated in bold. The results in Table 2 demonstrate that the proposed method generated better numerical results than DAGAN, DCCNN, and SwinMR. Similarly, Table 3 shows that the proposed method achieved the best performance in terms of PSNR and SSIM compared to the other models. GLFormer outperformed several other reconstruction methods on the IXI dataset, as shown in Table 4.

The visualization results for the radial 2D 30% and spiral 2D 30% undersampling settings on the IXI and MICCAI2013 datasets are shown in Figs. 5 and 6. The error maps were magnified ten fold and displayed within the range of [0, 1]. The error maps demonstrate that the proposed method had the

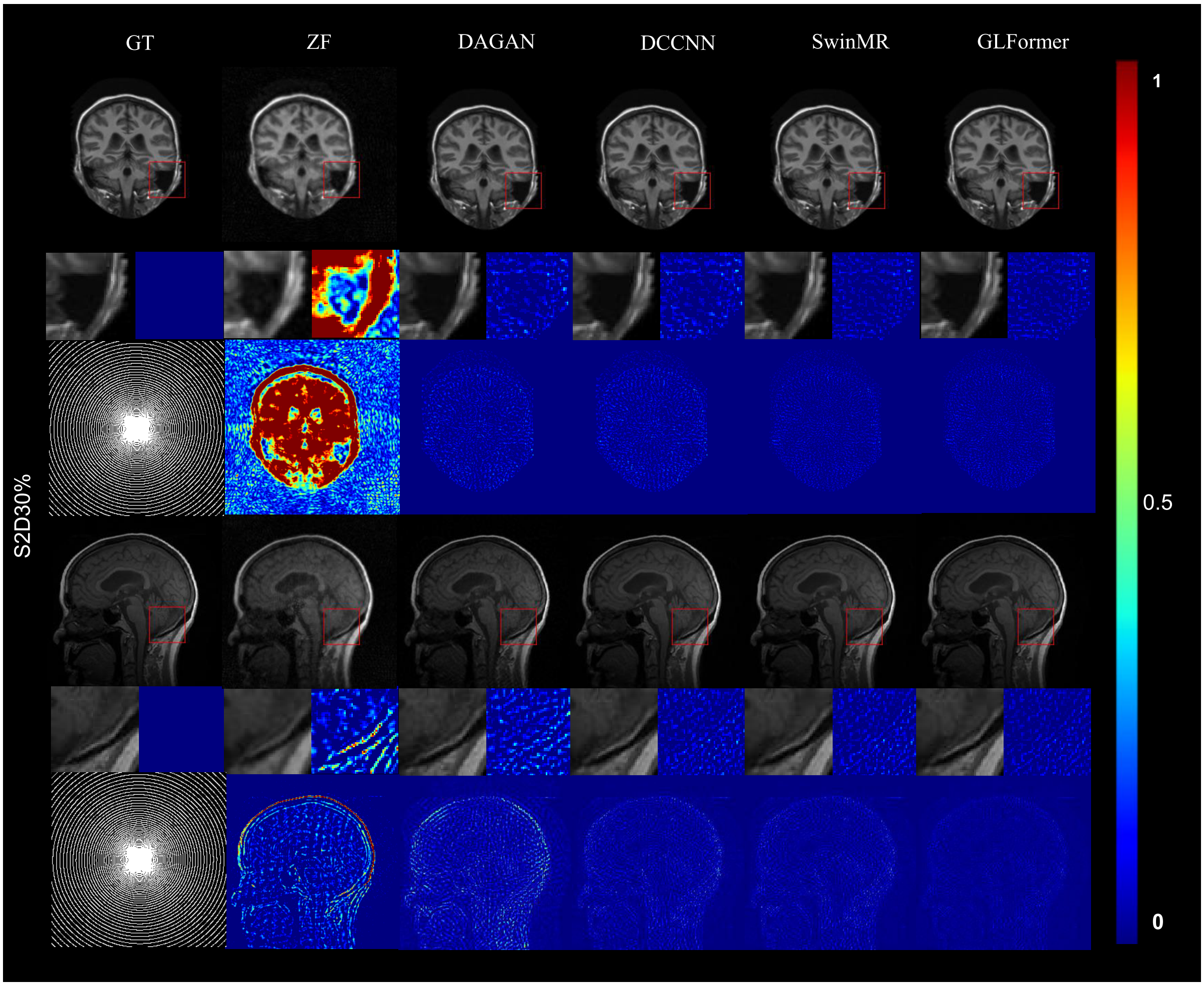


FIGURE 6. Qualitative visualization of different methods using Spiral 30% mask (S2D30%), For every three rows of figure, row 1: GT(ground truth image), ZF(zero-filled image) and reconstructed images of different methods; row 2: The zoomed-in images and corresponding error maps which are pointed out in the first row by a red rectangle; row 3:The error maps of reconstructed images (10×) The upper part shows the quantitative results of the MICCAI2013 dataset, and the lower part shows the quantitative results of the IXI dataset.

lowest reconstruction error for both datasets. Additionally, the zoomed-in images demonstrate that the proposed method reconstructed details and textures more effectively. The visualization results for the Guassian 1D 30% and radial 2D 20% undersampling settings on the fastmri dataset are shown in Fig. 7. Table 5 provides the quantitative results for the fastmri dataset. In comparison to the SwinMR method, our approach demonstrates a lower reconstruction error.

E. ABLATION STUDY

Table 6 shows the results of ablation experiments conducted on the overall network. Residual-Baseline model refers to the use of residual modules instead of transformer modules, and it does not include the DFA module.

1) IMPROVEMENTS IN GLOBAL ATTENTION

Table 6(c) shows that our global attention mechanism resulted in a significant improvement of 1.19 dB over the baseline. To further verify the global modeling ability of the mechanism, we visualized the output results of the first attention block for each branch. The results presented in Fig. 8 show that our global attention mechanism efficiently captures global features at a low computational cost. Moreover, each branch focuses on different image details, indicating that the multi-input multi-output structure results in more efficient reconstruction.

A quantitative comparison was conducted on the IXI dataset using a 20% Cartesian sampling pattern to assess the performance of the self-attention mechanism and the global attention mechanism. The results in Table 7 demonstrate that the global attention mechanism exhibited superior reconstruction quality compared to the self-attention mechanism. This performance improvement may be attributed to the presence of considerable redundancy in the computations involved in the self-attention mechanism.

2) IMPROVEMENTS IN LGFNN

Table 6(d) demonstrates the effectiveness of the parallel gating mechanism in controlling the information flow within the

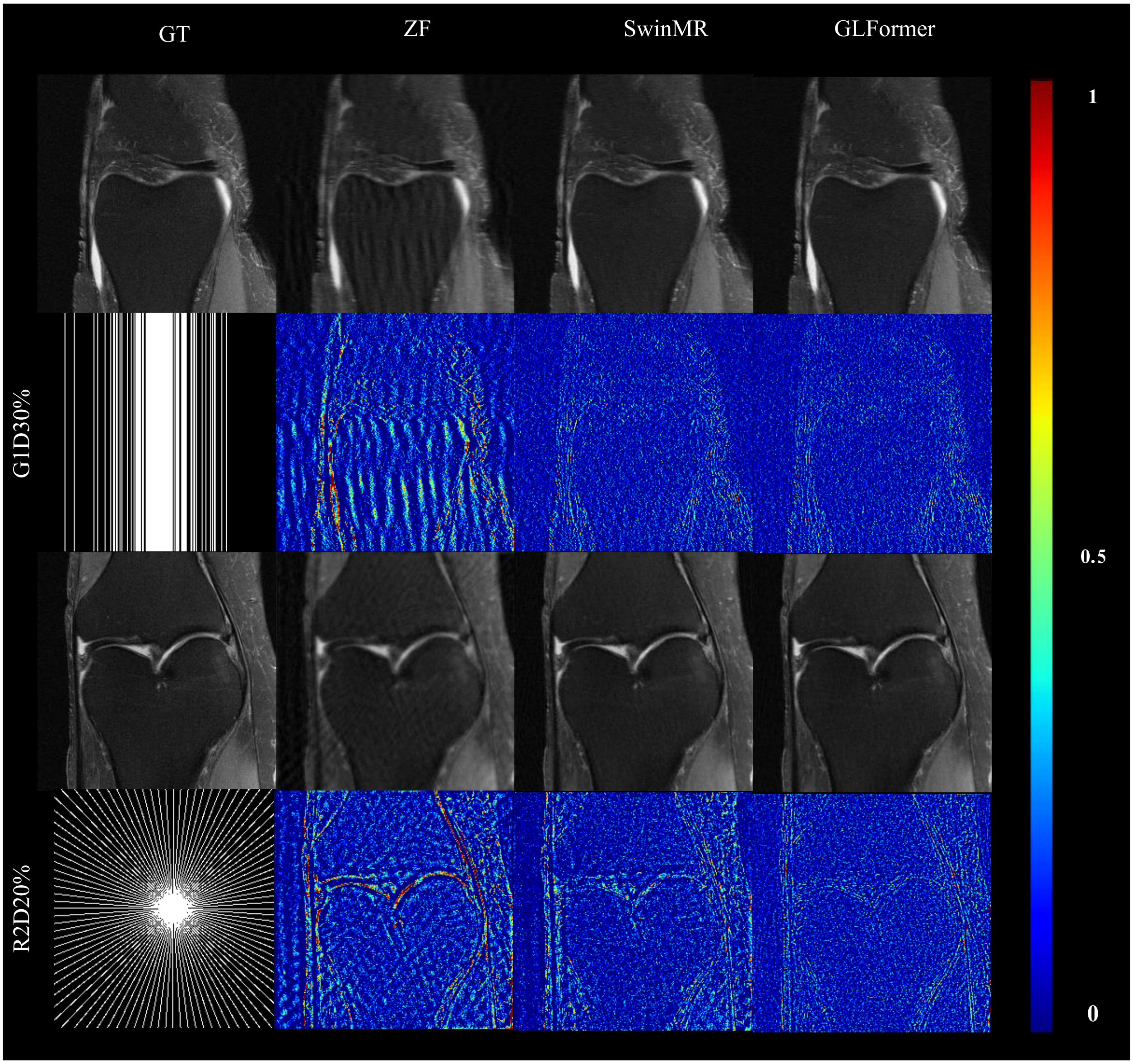


FIGURE 7. Qualitative visualization of different methods in fastmri dataset using Gaussian 1D 30 % mask (G1D30%) and radial 20% mask (R2D20%). For every three rows of figure, row 1: GT(ground truth image), ZF(zero-filled image), and reconstructed images of different methods; row 2:The error maps of reconstructed images (10×).

TABLE 6. Ablation experiments for GLFormer."Baseline" refers to the use of residual blocks as the basic module. × means that the component is not used, instead, √ means that the component is used.

Methods	*Baseline*	*GradientLoss*	*GA*	*LGFFN*	*DFA*	PSNR	SSIM	Param
(*a*)	√	×	×	×	×	49.32	0.990	6.80M
(*b*)	√	√	×	×	×	49.63	0.991	6.80M
(*c*)	×	√	√	×	×	50.82	0.994	7.44M
(*d*)	×	√	×	√	×	51.50	0.994	1.70M
(*e*)	×	√	√	√	×	52.08	0.997	2.34M
(*f*)	×	√	√	√	√	**54.45**	**0.999**	2.40M

FFN. The addition of a local pixel-aggregating mechanism also provides performance benefits. Our LGFFN module shows a PSNR gain of 1.87 dB over the baseline model. Furthermore, our contribution to the transformer block led to a substantial improvement of 2.45 dB over the baseline as shown in Table 6(e)

To compare the reconstruction performances of the proposed LGFFN and traditional FFN, we evaluated the

TABLE 7. Comparison experiments with Self attention.

Method	Complexity	PSNR
Self attention	O(HW*HW)	32.19
Global Attention	O(HW)	33.08

TABLE 8. Comparison experiments between LGFFN, FFN, and residual block.

Method	PSNR	SSIM	Param	FLOPs
residual block	49.63	0.991	6.8	66.84
FFN	50.05	0.989	3.3	34.6
LGFFN	**51.50**	**0.994**	**1.7**	**19.58**

TABLE 9. Comparison experiments for DFA.

Method	PSNR	SSIM
C_{mask}	52.98	0.998
S_{mask}	52.39	0.998
C_{mask}+S_{mask}(Serial)	53.28	**0.999**
C_{mask}+S_{mask}(Parallel)	**54.45**	**0.999**

TABLE 10. Comparison experiments with Dual domain network.

Method	PSNR	SSIM
KIKI-Net	35.46	0.97
SwinGAN	37.47	0.98
GLFormer	**37.81**	**0.98**

reconstruction results using the baseline model (using residual blocks), FFN, and the LGFFN module proposed in this study. As shown in Table 8, the results suggest that the LGFFN module not only reduces the parameter and computational burden but also achieves superior reconstruction results compared to the original FFN.

3) IMPROVEMENTS IN DFA

We proposed a deep frequency domain attention module to improve the frequency deviation of the reconstruction network. The effectiveness of our frequency attention module, which showed improved reconstruction results, is presented in Table 6(f). To further investigate the module's effectiveness, we compared the performance of frequency adjustment using only channel embeddings and only spatial embeddings and analyzed the performance differences between parallel and serial learning for both types of embeddings. The results presented in Table 9. demonstrate that channel embeddings alone are superior to spatial embeddings in terms of frequency adjustment. Furthermore, the simultaneous parallel learning of both types of embeddings leads to maximum performance gain.

To gain a deeper understanding of the mechanism of deep-frequency attention, we visualized the channel and spatial masks. The visualization results are shown in Fig. 8. We found that adaptive decoupling of different spatial frequencies was achieved by the channel and spatial masks. Moreover, the channel and spatial masks paid more attention to low-frequency information and high-frequency information, respectively.

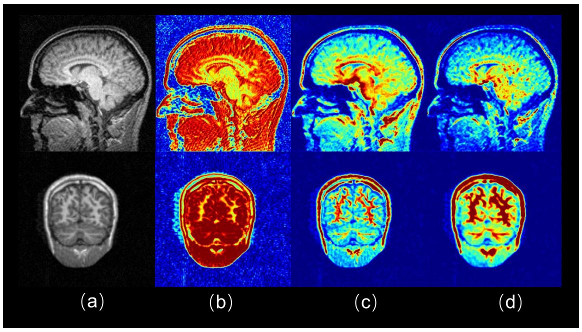


FIGURE 8. Visualizing first global attention module of different network branches. The first row presents images from the MICCAI 2013 dataset, and the second row presents images from the IXI dataset. (a) input images (b) visualization results of the first branch, (c) visualization results of the second branch, (d) visualization results of the third branch.

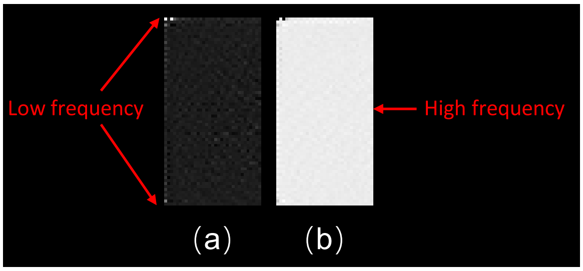


FIGURE 9. Visualization of the spatial mask and channel mask of DFA. (a) The left side of the figure represents a channel mask, which mainly learns low-frequency information, (b) the right side of the figure represents a spatial mask which mainly learns a large amount of high-frequency information.

We further compared our proposed network with dual-domain networks [11], [36]. As shown in Table 10, GLFormer outperforms the two most recent dual-domain networks in terms of reconstruction performance while requiring only a lightweight frequency domain module.

V. DISCUSSION

While the transformer model overcomes the limitations of CNNs, such as a limited receptive field, its quadratic computational complexity limits its applicability in magnetic resonance imaging reconstruction. To address this issue, we propose an efficient transformer-based network, termed GLFormer. The experimental results demonstrate that GLFormer has smaller reconstruction errors than those of several other methods and significantly reduces both computational and parameter costs. The visualization results demonstrate that the proposed global attention module captures global features at a linear computational cost. The ablation experiments listed in Table 6 demonstrate that both the global attention and LGFFN modules significantly improve the reconstruction results. Additionally, the proposed DFA module was adapted to decouple the learning of different frequency distributions based on the visualization results shown in Fig. 9, leading to improved reconstruction details.

However, considering the potential presence of noise and pathological data in real-world scenarios, further exploration is required to enhance the network's noise robustness and ability to reconstruct pathological data. In addition, we only tested the undersampled reconstruction of brain images, and the distribution differences between different parts may affect the reconstruction performance of the network.

In future work, we will extend the network to utilize multi-contrast magnetic resonance images for improved reconstruction. Furthermore, owing to the small computational burden of GLFormer, we can apply this lightweight transformer to high-resolution imaging scenarios such as cardiac imaging [37], [38], [39].

VI. CONCLUSION

In this study, we propose a novel efficient transformer network based on MIMO architecture. The proposed GLFormer includes two improvements over conventional models. The first is a simplified self-attention mechanism that obtains global feature representations through pooling and attention queries. This effectively reduces the computational complexity while preserving the representation ability of global features. The second improvement is the introduction of local feature aggregation and gating mechanisms in the feedforward network, which emphasizes the spatially local context and improves information flow. Additionally, we propose a deep frequency attention operation that learns the embedding of frequency domain features obtained from Fourier transforms in both the channel and spatial dimensions and adjusts the distribution of frequency domain features accordingly. Qualitative and quantitative experiments conducted on public datasets demonstrated that GLFormer surpasses existing methods on public datasets and exhibits superior reconstruction performance. Our study presents an innovative solution for efficient MRI reconstruction with transformer networks, providing significant potential for clinical applications.

REFERENCES

[1] Q. Chen, K. W. Stock, P. V. Prasad, and H. Hatabu, "Fast magnetic resonance imaging techniques," *Eur. J. Radiol.*, vol. 29, no. 2, pp. 90–100, 1999.

[2] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, no. 5, pp. 952–962, Nov. 1999.

[3] M. Lustig, L. D. Donoho, M. J. Santos, and M. J. Pauly, "Compressed sensing MRI," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2007.

[4] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang, "Accelerating magnetic resonance imaging via deep learning," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 514–517.

[5] S. Ramanarayanan, B. Murugesan, K. Ram, and M. Sivaprakasam, "DC-WCNN: A deep cascade of wavelet based convolutional neural networks for MR image reconstruction," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Boone, NC, USA, Apr. 2020, pp. 647–658.

[6] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, and D. Firmin, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[8] K. Lin and R. Heckel, "Vision transformers enable fast and robust accelerated MRI," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 774–795.

[9] J. Huang, Y. Fang, Y. Wu, H. Wu, Z. Gao, Y. Li, J. D. Ser, J. Xia, and G. Yang, "Swin transformer for fast MRI," *Neurocomputing*, vol. 493, pp. 281–304, Jul. 2022.

[10] P. Guo and V. M. Patel, "Reference-based MRI reconstruction using texture transformer," in *Medical Imaging With Deep Learning*. Nashville, TN, USA: OpenReview.net, 2023.

[11] X. Zhao, T. Yang, B. Li, and X. Zhang, "SwinGAN: A dual-domain Swin transformer-based generative adversarial network for MRI reconstruction," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106513.

[12] C.-M. Feng, Y. Yan, G. Chen, Y. Xu, Y. Hu, L. Shao, and H. Fu, "Multimodal transformer for accelerated MR imaging," *IEEE Trans. Med. Imag.*, early access, Jun. 15, 2022, doi: 10.1109/TMI.2022.3180228.

[13] Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[14] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8681–8691.

[15] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–12.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[18] G. A. Baxes, *Digital Image Processing: Principles and Applications*. Hoboken, NJ, USA: Wiley, 1994.

[19] A. M. Raid, W. M. Khedr, M. A. El-Dosuky, and W. Ahmed, "JPEG image compression using discrete cosine transform—A survey," 2014, *arXiv:1405.6147*.

[20] Z. Zhang, C. Jung, and X. Liang, "Adversarial defense by suppressing high-frequency components," 2019, *arXiv:1908.06566*.

[21] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1081–1088.

[22] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 86–103.

[23] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, and W. Liu, "Improving vision transformers by revisiting high-frequency components," in *Proc. 17th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 1–18.

[24] S. U. H. Dar, M. Yurt, M. Shahdloo, M. E. Ildiz, B. Tinaz, and T. Çukur, "Prior-guided image reconstruction for accelerated multi-contrast MRI via generative adversarial networks," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1072–1087, Oct. 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–16.

[26] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1138–1150, 2020.

[27] W. T. Cochran, J. W. Cooley, D. L. Favin, H. D. Helms, R. A. Kaenel, W. W. Lang, G. C. Maling, D. E. Nelson, C. M. Rader, and P. D. Welch, "What is the fast Fourier transform?" *Proc. IEEE*, vol. 55, no. 10, pp. 1664–1674, Oct. 1967.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[29] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4621–4630.

[30] J. Zbontar et al., "fastMRI: An open dataset and benchmarks for accelerated MRI," 2018, *arXiv:1811.08839*.

[31] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[34] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *Proc. 4th Int. Workshop Quality Multimedia Exper.*, Jul. 2012, pp. 37–38.

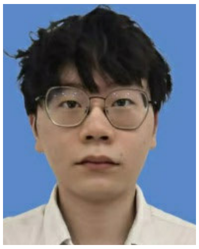
[35] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1488–1499, Apr. 2012.

[36] T. Eo, Y. Jun, T. Kim, J. Jang, H. Lee, and D. Hwang, "KIKI-Net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2188–2201, Nov. 2018.

[37] R. I. Pettigrew, J. N. Oshinski, G. Chatzimavroudis, and W. T. Dixon, "MRI techniques for cardiovascular imaging," *J. Magn. Reson. Imag.*, vol. 10, no. 5, pp. 590–601, Nov. 1999.

[38] A. E. Campbell-Washburn, M. A. Tavallaei, M. Pop, E. K. Grant, H. Chubb, K. Rhode, and G. A. Wright, "Real-time MRI guidance of cardiac interventions," *J. Magn. Reson. Imag.*, vol. 46, no. 4, pp. 935–950, Oct. 2017.

[39] H. Akimoto, T. Nagaoka, T. Nariai, Y. Takada, K. Ohno, and N. Yoshino, "Preoperative evaluation of neurovascular compression in patients with trigeminal neuralgia by use of three-dimensional reconstruction from two types of high-resolution magnetic resonance imaging," *Neurosurgery*, vol. 51, no. 4, pp. 956–962, Oct. 2002.



RONGQING WANG received the B.E. degree in electrical engineering from Anhui University, Hefei, China, in 2019. He is currently pursuing the M.E. degree in biomedical engineering with the University of Science and Technology of China, Hefei. His current research interest includes deep learning fast MR reconstruction.

MENGDIE SONG (Graduate Student Member, IEEE) received the B.E. degree in the Internet of Things from Anhui University, Hefei, China, in 2020. She is currently pursuing the Ph.D. degree in biomedical engineering with the University of Science and Technology of China, Hefei. Her current research interest includes deep learning fast MR reconstruction.

JIANTAI ZHOU received the B.E. degree in applied physics from the Huazhong University of Science and Technology, Wuhan, China, in 2020. He is currently pursuing the Ph.D. degree in biomedical engineering with the University of Science and Technology of China, Hefei. His current research interests include EPI-DWI and constrained image reconstruction algorithm.

BENSHENG QIU (Member, IEEE) received the bachelor's degree in electrical engineering from the College of Electronic Engineering, in 1987, the master's degree in acoustic engineering from Northwestern Polytechnic University, Fremont, CA, USA, in 1990, and the Ph.D. degree in computer science from the Hefei University of Technology, Hefei, China, in 1995.

From 1997 to 2001, he was an Associate Professor of radiology with the PLA General Hospital. From 2001 to 2005, he was with the Department of Radiology, The Johns Hopkins University School of Medicine. From 2006 to 2012, he was an Assistant Professor with the University of Washington. He is currently a Professor and the Vice Chairperson of the Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, and the Director of the Medical Imaging Center, University of Science and Technology of China. He has conducted many projects on MRI-guided gene/stem cell therapy and interventions supported by NIH, RSNA, NSF, and the National Basic Research Program of China. He received many awards from the Radiology Society of North America, the American Heart Association, and the International Society for Magnetic Resonance in Medicine.