applied sciences

MDPI

Article

A Multi-Hierarchical Complementary Feature Interaction Network for Accelerated Multi-Modal MR Imaging

Haotian Zhang, Qiaoyu Ma, Yiran Qiu and Zongying Lai *

School of Ocean Information Engineering, Jimei University, Xiamen 361021, China;
202211810011@jmu.edu.cn (H.Z.); maqiaoyu86@gmail.com (Q.M.); 202111810019@jmu.edu.cn (Y.Q.)
* Correspondence: zongyinglai@jmu.edu.cn

Abstract: Magnetic resonance (MR) imaging is widely used in the clinical field due to its non-invasiveness, but the long scanning time is still a bottleneck for its popularization. Using the complementary information between multi-modal imaging to accelerate imaging provides a novel and effective MR fast imaging solution. However, previous technologies mostly use simple fusion methods and fail to fully utilize their potential sharable knowledge. In this study, we introduced a novel multi-hierarchical complementary feature interaction network (MHCFIN) to realize joint reconstruction of multi-modal MR images with undersampled data and thus accelerate multi-modal imaging. Firstly, multiple attention mechanisms are integrated with a dual-branch encoder–decoder network to represent shared features and complementary features of different modalities. In the decoding stage, the multi-modal feature interaction module (MMFIM) acts as a bridge between the two branches, realizing complementary knowledge transfer between different modalities through cross-level fusion. The single-modal feature fusion module (SMFFM) carries out multi-scale feature representation and optimization of the single modality, preserving better anatomical details. Extensive experiments are conducted under different sampling patterns and acceleration factors. The results show that this proposed method achieves obvious improvement compared with existing state-of-the-art reconstruction methods in both visual quality and quantity.

Keywords: magnetic resonance imaging; multi-modal imaging; deep learning; complementary feature fusion

check for updates

Citation: Zhang, H.; Ma, Q.; Qiu, Y.; Lai, Z. A Multi-Hierarchical Complementary Feature Interaction Network for Accelerated Multi-Modal MR Imaging. *Appl. Sci.* **2024**, *14*, 9764. https://doi.org/10.3390/app14219764

Academic Editor: Jose Machado

Received: 4 September 2024
Revised: 19 October 2024
Accepted: 21 October 2024
Published: 25 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Magnetic resonance (MR) imaging is a revolutionary medical imaging technology that plays a crucial role in clinical diagnosis and biomedical research due to its excellent soft tissue contrast, radiation-free safety, and high-resolution imaging of living structures [1]. However, due to the physical nature, the scanning process of MR takes a long time [2]. This may not only cause discomfort to the patient but may also produce motion artifacts due to the patient's small movements, thus hindering its further promotion and application [2]. To address this problem, several solutions for accelerated MR imaging have been established [3–6]. The mainstream is to restore high-quality images from undersampled images with aliasing artifacts. This accelerates MR imaging through undersampling k-space data [7–10]. In this way, data acquisition time can be significantly reduced. However, undersampling causes aliasing artifacts in the image (as shown in Figure 1), affecting image quality. To address this problem, previous studies have proposed compressed sensing (CS) algorithms to recover high-quality MR images from undersampled k-space data. The CS [3,11] methods exploit the sparsity of images and reconstruct high-quality images by imposing sparsity constraints on undersampled data through optimization algorithms. However, the CS reconstruction process usually involves complex optimization algorithms and has high computational costs [12]. In addition, the performance of CS depends on the selection of multiple hyperparameters, and optimizing these parameters requires experience and extensive experiments [13].

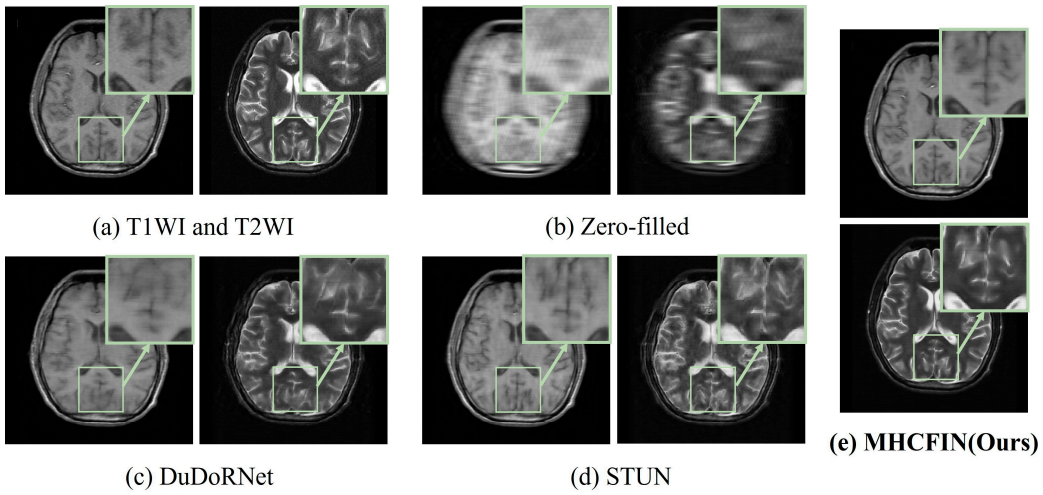


Figure 1. Illustration of different MRI modalities and the reconstruction comparison results on fastMRI brain datasets. (**a**) Fully-sampled T1WI and T2WI. (**b**) Zero-filled T1WI and T2WI with 1D random sampling at 8× accelerate factors. (**c**) Reconstruction results of DuDoRNet [14]. (**d**) Reconstruction results of STUN [15]. (**e**) Reconstruction results of MHCFIN (Ours). Green boxes highlight detailed structures.

More recently, deep learning [12,14,16–24], has excelled in reducing the computational burden and solving nonlinear problems, thus handling inverse problems more efficiently. Typically, in deep learning-based methods [12], the image reconstruction problem is formulated as an end-to-end learning task, directly mapping undersampled images to fully sampled images. In addition, these methods [16,25] can also restore k-space data directly.

Most scanners can generate multi-modal images using different acquisition parameters. The complementary information between these multi-modal images is very beneficial to MR imaging. As shown in Figure 1a, the T1-weighted images (T1WI) and T2-weighted images (T2WI) are different modalities from the same subject. These modalities usually have structural similarities and modality-specific information. However, due to the characteristics of MR imaging, T1WI acquisition has a shorter repetition time (TR) and echo time (TE). Therefore, in many studies, images with shorter acquisition times are used as auxiliary modalities (prior information) to guide the accelerated imaging of the target modality with longer acquisition times [14,26,27]. Joint reconstruction explores complementary information among multi-modal images in varying degrees [12,28,29]. In addition, super-resolution reconstruction [30–32], can enhance image resolution and introduce high-resolution (HR) modalities to guide the restoration of low-resolution (LR) modalities. These methods provide new ideas for joint learning of multi-modal images. However, early studies [18,31,33,34] usually spliced images of different modalities into the network or directly input images of different modalities. Although this improved the image quality to a certain extent, it has not fully utilized the potential complementarity and interaction between different modalities.

In this study, to explore the complementarity between different modalities, multi-hierarchical modality correlations are investigated and a multi-scale approach is adopted to fuse the target modality (e.g., proton density-weighted images—PDs) and auxiliary modality (e.g., fat-suppressed proton density-weighted images—PDFSs) to achieve synchronous acceleration of multi-modal imaging. Specifically, this study proposed a novel multi-hierarchical complementary feature interaction network (MHCFIN). As shown in Figure 2, a dual-branch encoding–decoding learning framework is designed based on a multi-attention mechanism. This dual-branch learning architecture can process undersampled images of different modalities simultaneously. In this architecture, we introduce an innovative triple attention block that can accurately capture key features at multiple levels and identify important features shared across different modalities. In addition, to

effectively preserve important features and reduce information loss, we upgrade the traditional skip connection to a residual learning mechanism. In the decoder, we designed a multi-modal feature interaction module (MMFIM) to merge multi-scale features using a dual attention mechanism, capturing complementary information between modalities. Finally, the convolution-based single-modal feature fusion module (SMFFM) integrates and refines multi-scale features, further enhancing feature representation and establishing long-range dependencies between contexts. The proposed method demonstrates superior performance across various experimental settings, as illustrated in Figure 1c,d,e. Compared to other methods [14,15], it effectively restores details and exhibits better performance. The main contributions are as follows:

- In this study, we introduce an innovative multi-modal reconstruction approach, termed MHCFIN, that facilitates high-fidelity image reconstruction by harnessing the latent, learnable complementary information inherent across modalities.
- We propose a dual-branch encoder–decoder learning structure based on a multi-attention mechanism, which can accurately capture key features at different levels and identify important shared features across modalities, while dynamically optimizing the learning process and using residual learning to reduce information loss.
- We integrate dual cross-attention at the decoder to achieve complementary knowledge transfer across modalities and integrate multi-scale features through a single-modal feature fusion module within the same modality to establish contextual dependencies.
- Experimental results show that the proposed method outperforms the current state-of-the-art reconstruction methods in multi-modal image reconstruction.

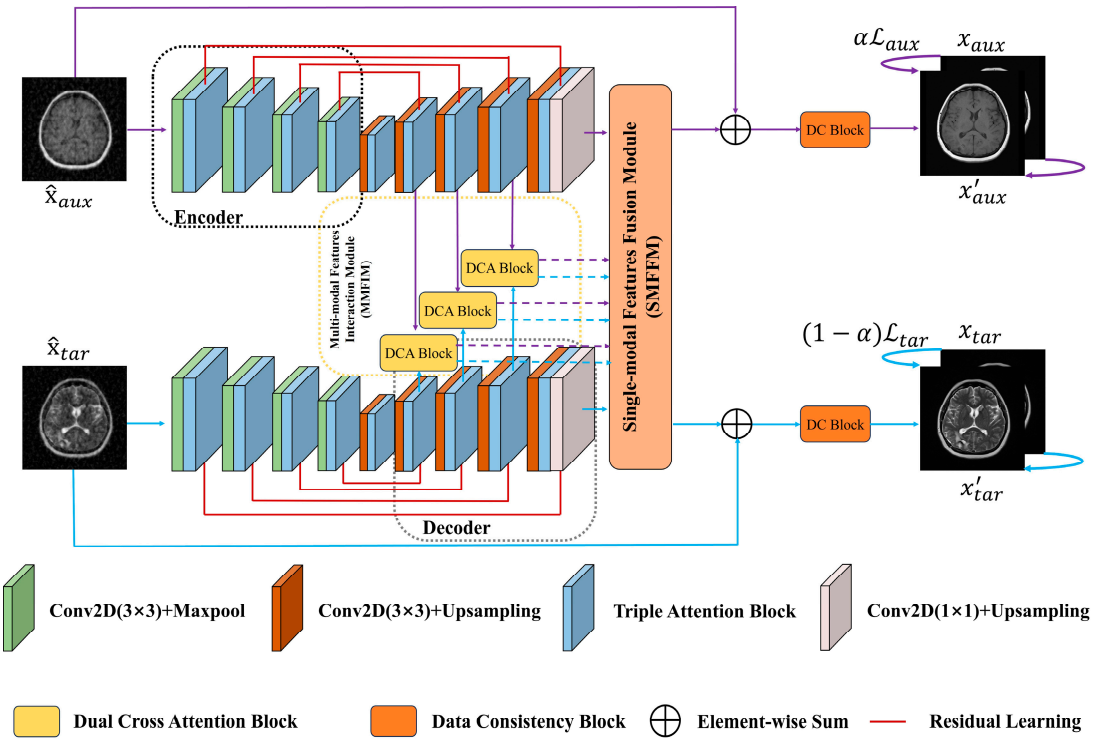


Figure 2. Overall architecture of the proposed multi-hierarchical complementary feature interaction network (MHCFIN).

2. Related Work

2.1. Accelerated Single-Modal MR Imaging

Without replacing or upgrading existing imaging hardware, image reconstruction technology can maximize the potential of existing equipment and achieve better imaging results. In previous studies, traditional techniques such as CS-based iterative algorithms [5,7], low-rank [35], patch-based operators [36], parallel imaging [37], and dictionary learning [38]

have achieved remarkable results in accelerating single-modal MR imaging. Recently, deep learning methods have been increasingly utilized for accelerating MR imaging. These methods can be categorized into two types: end-to-end [39–41], and model-based [42–44] methods. Wang et al. [39] first applied convolutional neural networks (CNN) to accelerate MR imaging. Schlemper et al. [40] proposed a deep cascade network based on CNN that can recover images with high fidelity from zero-filled data. In the pursuit of enhancing the transparency and clinical utility of deep learning models, recent scholarly work has introduced model-based methods to deep learning. Aggarwal et al. [42] combined data-driven deep learning with physics-based models to deal with inverse problems, improving the generalization performance of the model in diverse imaging environments. ADMM-CSNet combines the alternating direction method of multipliers method (ADMM) with CNN and achieves good performance [10]. In addition, the Generative Adversarial Network (GAN) [15], an unsupervised learning framework consisting of a generator and a discriminator, has also been applied to accelerate single-modal imaging. Mardani et al. [41] use adversarial loss and perceptual loss to optimize the reconstruction process of compressed sensing MRI, improving image quality and detail recovery. Although GANs have shown impressive performance, their training instability has motivated the exploration of new models. Recently, the Swin Transformer [45,46] has been widely used in the field of accelerated MR imaging due to its ability to effectively capture long-range dependencies in the data. Huang et al. [46] used the hierarchical feature representation and shift window mechanism of the Swin Transformer to capture image information of different scales and obtain global context information through the self-attention mechanism. Compared with traditional CNN, this method further improved the MR image reconstruction effect. Currently, research on accelerating multi-modal imaging by leveraging the potential learnable complementary knowledge between different modalities has become a hot topic.

2.2. Accelerated Multi-Modal MR Imaging

Accelerated multi-modal imaging can be approached in two ways: one involves using auxiliary modalities as prior information to guide the reconstruction of the target modality time [14,26,27,47,48], and the other involves simultaneously reconstructing multiple modalities [29,49,50]. Zhou et al. [14] proposed a dual-domain recurrent network (DuDoRNet), which effectively improved the reconstruction quality of MRI images by combining iterative restoration in the image domain and k-space domain and incorporating T1-weighted imaging as a priori. Lyu et al. [27] proposed a method that combines dual-domain learning and cross-attention fusion (DuDoCAF) to further fuse long-distance dependencies in feature maps through recurrent transformers to reconstruct performance. In addition, by simultaneously accelerating multi-modal imaging, the complementarity between different modalities can be fully utilized to ensure synchronous acquisition of data, which is crucial for achieving accurate diagnosis and treatment planning. Feng et al. [28] proposed an innovative cross-attention module that can effectively extract and fuse multi-modal features at different scales, combined with a multi-modal transformer (Mtrans) to capture global information, significantly improving the performance of image reconstruction and super-resolution. Li et al. [12] exploited the complementary information in multi-contrast (MCCA) MR protocols to accelerate the imaging process and developed a novel hybrid convolutional module that can simultaneously extract global and local contextual dependencies to improve image reconstruction quality. Wang et al. [29] proposed a model graph-driven transformer (MD-GraphFormer) that combines the physical constraints of MR imaging and uses graph structure and attention mechanism to explore the complementarity between multi-modal MR images, thus achieving simultaneous reconstruction of multi-modal images. The multi-stage integrated network [49] (MINet) deeply explores the mutual relationship between multi-modal features and realizes the similarity integration of features, thereby obtaining richer feature representation and effectively improving image quality.

2.3. U-Net Based Methods

The U-shaped Network [51] (U-Net) was originally designed for medical image segmentation and is particularly suitable for fine boundary recognition tasks such as cell and organ segmentation. With the development of medical imaging technology, U-Net and its derivative models [15,33,47,51] have been shown to have the potential to accelerate the imaging process and improve image quality. Aghabiglou and Eksioglu [47] proposed a cascaded U-Net network fused with a projection-based data consistency layer for the reconstruction of MR images from severely zero-filled k-space data. Xiang et al. [33] designed an innovative Dense-UNet network that integrates dense blocks in the encoding and decoding paths, taking advantage of the complementarity of multi-modal data to significantly improve the quality of image reconstruction. Zhao et al. [15] proposed an innovative Swin Transformer-based encoder–decoder architecture (STUN) in their study, which was designed to replace the convolutional filters in the traditional U-Net network.

3. Methods and Materials

3.1. Background and Problem Definition

Due to the inherent nature of MRI, its scanning process tends to be time-consuming, which can lead to patient discomfort and potential motion artifacts caused by slight movements during the scan. As a result, undersampling the k-space has become a main approach to reduce the scanning time. Let x_f represent the fully sampled MR image and x_u represent the zero-filled image. Then the k-space undersampled process in the real scene can be expressed as:

$$x_u = \mathrm{F}^{-1}\left(\mathrm{M} \circ \mathrm{F}x_f + \epsilon\right) \tag{1}$$

where F and F^{-1} denote the Fourier and the inverse Fourier transformer respectively, M denotes the undersampling mask, '$\circ$' denotes the Hadamard multiplication operation, and ϵ denotes the noise. However, undersampling the k-space leads to aliasing artifacts. The CS leverages the inherent sparsity (in MRI, most of the image information is contained in a few key coefficients, allowing the rest to be ignored or set to zero without losing critical details. This enables data compression and reconstruction from fewer samples) of MR images and applies regularization techniques (regularization adds constraints to the reconstruction process, guiding it towards better results) to recover high-fidelity images from noisy k-space data. This approach enables sampling rates far below the Nyquist limit, significantly accelerating MR imaging. In general, the CS-MRI reconstruction model can be expressed as an optimization problem:

$$x'_f = \underset{x_f}{\operatorname{argmin}} \left\| \mathrm{M} \circ \mathrm{F}x_f - k_u \right\|_2^2 + \lambda \psi\left(x_f\right) \tag{2}$$

where $\psi(\cdot)$ denotes a priori regularization which is used to introduce additional knowledge or assumptions about the image, λ denotes the hyperparameter used to balance the reconstruction quality and the regularization effect, k_u denotes the undersampled k-space data, and x'_f denotes the reconstructed image close to the ground truth. In the deep learning framework, image reconstruction is transformed into a direct end-to-end learning process, which can learn the direct mapping from zero-filled k-space to fully sampled images. This process can be described as:

$$x'_f = \underset{\Theta}{\operatorname{argmin}} \| \mathrm{MF}(\mathcal{N}_\Theta(x_u \,|\Theta)) - k_u \|_2^2 + \lambda \| x_f - \mathcal{N}_\Theta(x_u \,|\Theta) \|_2^2 \tag{3}$$

where $\mathcal{N}_\Theta$ denotes the network MHCFIN proposed in this paper driven by parameter Θ. Based on the above problem description, we will elaborate on the specific details of the proposed network in this section. Specifically, we first outline the overall architecture of MHCFIN, which adopts a dual-branch encoder–decoder network based on the attention mechanism. Then, we deeply explore the different hierarchical multi-modal feature in-

teraction module (MMFIM) and the single-modal feature fusion module (SMFFM). The overall architecture of the proposed method is shown in Figure 2. The network uses a dual-branch learning structure, which can process multi-modal data simultaneously and learn the complementarity between them.

3.2. Overall Architecture

To fully exploit the complementarity between different modalities, we design a dual-branch learning structure to process the zero-filled auxiliary modality $\hat{x}_{aux}$ and the target modality $\hat{x}_{tar}$ of the input, respectively. The dual-branch network uses a unified structural design, with an encoding–decoding structure enhanced by residual learning and multi-attention. As shown in Figure 3, to enhance the network's recognition and emphasis on key features, we integrate a triple attention block (TAB, as shown in the right part of Figure 3) consisting of three types of attention into the basic modules of the encoder and decoder. At the same time, to improve the feature expression ability and optimize the information flow, the TAB is also used to replace the traditional linear operation at the bottleneck. The TAB consists of channel attention, spatial attention, and gate mechanism. Its detailed structure is shown in Figure 4. Combined with local residual learning, the network can automatically focus on key multi-scale features at different hierarchical levels to achieve more efficient feature processing and information transfer. In addition, replacing traditional skip connections with residual learning can alleviate the gradient vanishing problem in deep networks and improve generalization ability.

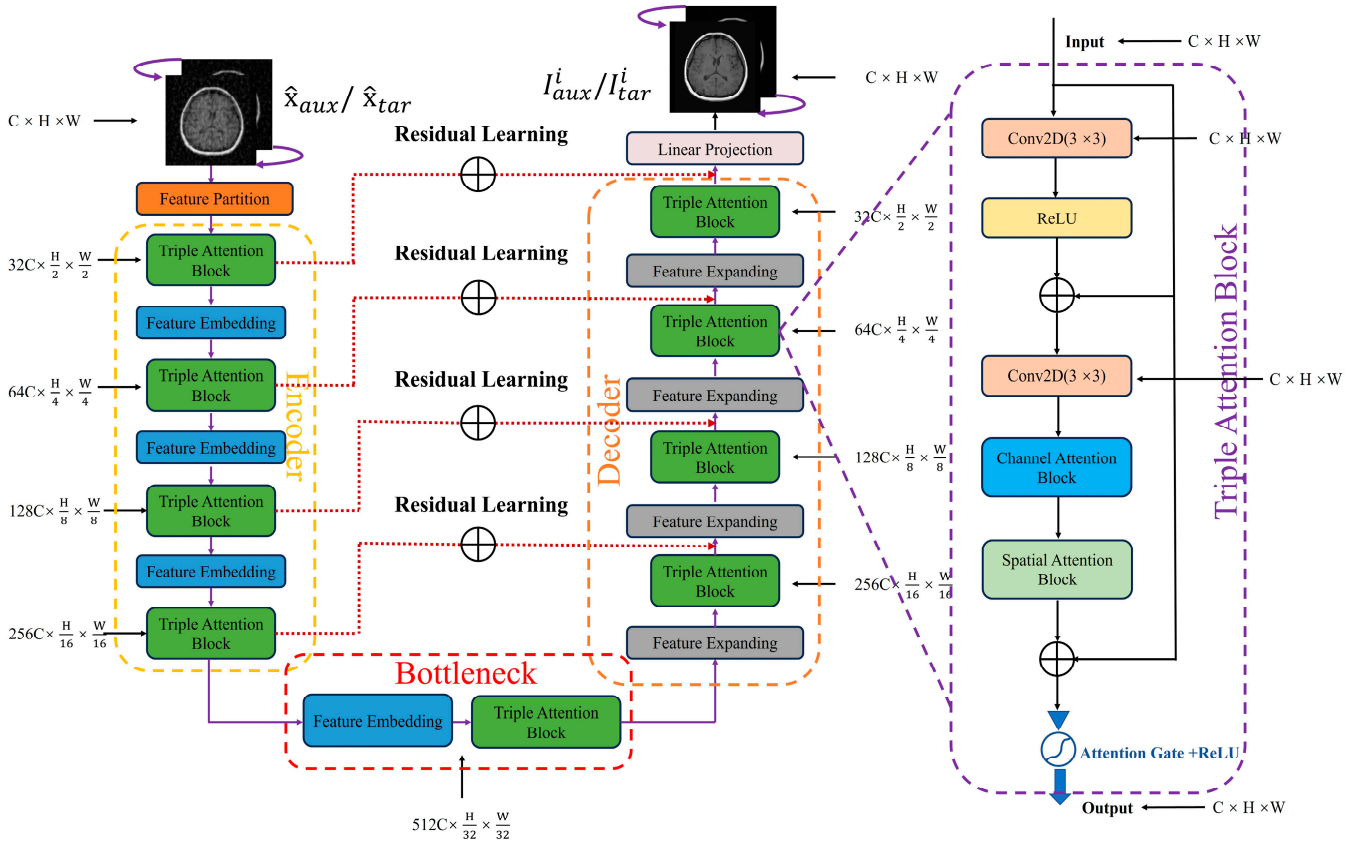


Figure 3. Detailed architecture of the proposed encoder–decoder and detailed construction of triple attention.

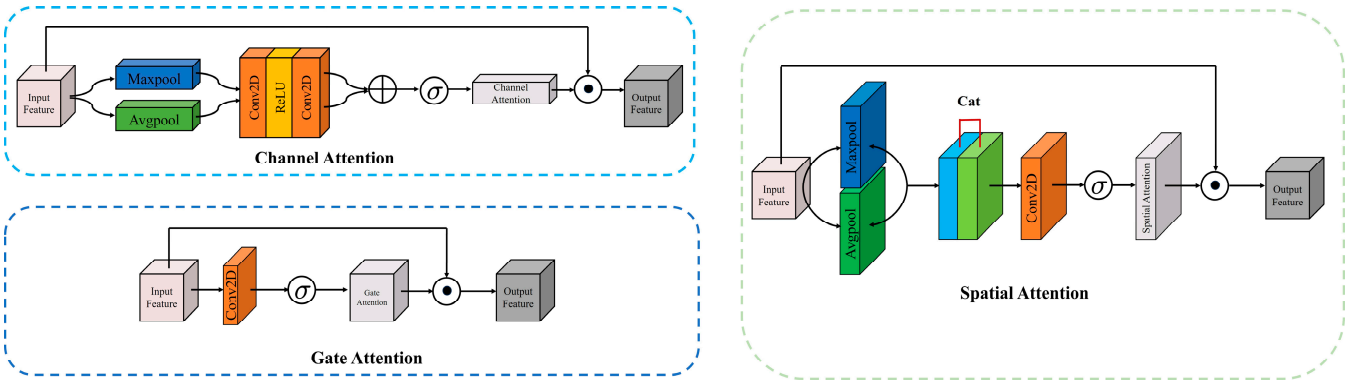


Figure 4. Detailed structure of the three types of attention (channel, spatial, and gate attention).

As shown in Figure 2, in the dual-branch learning architecture, the top branch accepts the input of the zero-filled auxiliary modality $\hat{x}_{aux} \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 2}$ (e.g., T1WI or PDs), and the bottom branch processes the zero-padded target modality $\hat{x}_{tar} \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 2}$ (e.g., T2WI or PDFSs), where H and W are the height and width of the image, and 2 are the real and imaginary parts. Then, during the forward propagation, we can obtain feature maps at different hierarchical levels. In the decoder stage, we interact with the feature maps of the auxiliary modality $I^i_{aux}(i = 1, 2, \cdots, \mathrm{N})$ at each hierarchical with the feature maps of the target modality I^i_{aux} through the proposed MMFIM, where i represents the output feature map of the layer, and N is set to 4 in this paper. After all features in the encoder complete the interaction, we introduce the feature maps of the same modality into the SMFFM to further realize the fusion of feature maps of the same modality at different scales. In addition, we introduce a data consistency layer (DC) to fully utilize the k-space data to enhance the realism of the reconstructed images. Finally, after processing by the global residual learning module, we obtain the restored high-fidelity auxiliary modality x'_{aux} and target modality x'_{tar}.

3.3. Multi-Scale Complementary Features Interaction

These methods [28,52] use cross-attention to fuse features of different modalities to learn complementary information between them. By improving the multi head attention mechanism, the features of the auxiliary modality are considered as the key and value, while the target modality is used as the query, thus enabling the interaction between the target and auxiliary features. Inspired by these methods, we introduce an improved dual cross-attention block (DCAB), which simultaneously processes both auxiliary and target modalities. The auxiliary branch's query Q_{aux} interacts with the target branch's key K_{tar} and value V_{tar}, while the target branch's query Q_{tar} also interacts with the auxiliary branch's key K_{aux} and value V_{aux}. Unlike previous work, our proposed DCAB facilitates mutual interaction between the features of different modalities, enabling the learning of complementary information across modalities and allowing for the simultaneous reconstruction of images for both modalities, as shown in Figure 5. In addition, DCAB is used as the basic building block of MMFIM, i.e., $\mathrm{MMFIM} = (\mathrm{DCAB}_1, \mathrm{DCAB}_2, \cdots, \mathrm{DCAB}_k)$, where $k = 1, 2, \cdots, \mathrm{K}$ and K is set to 3 in this paper. As shown in Figure 2, the input of MMFIM comes from the output of the decoder part in the dual-branch learning architecture, i.e., $I^i_{aux} \in \mathrm{R}^{\mathrm{H}\cdot 2^{i-1}\times \mathrm{W}\cdot 2^{i-1}\times \mathrm{C}/2^{i-1}}$ and $I^i_{tar} \in \mathrm{R}^{\mathrm{H}\cdot 2^{i-1}\times \mathrm{W}\cdot 2^{i-1}\times \mathrm{C}/2^{i-1}}$, where $i = 1, 2, 3$. First, the input feature maps I^i_{aux} and I^i_{tar} are normalized and then, unlike traditional cross-attention with its use of linear projection before each attention block, we use deep convolution to generate query matrices Q_{aux} and Q_{tar}, keyword matrices K_{aux} and K_{tar}, and value matrices V_{aux} and V_{tar} for 2D normalized feature mapping. The process can be expressed as:

$$\mathrm{Q}^i_{aux} = I^i_{aux}\mathrm{W}^q_{aux},\ \mathrm{K}^i_{aux} = I^i_{aux}\mathrm{W}^k_{aux},\ \mathrm{V}^i_{aux} = I^i_{aux}\mathrm{W}^v_{aux} \tag{4}$$

$$\mathrm{Q}^i_{tar} = I^i_{tar}\mathrm{W}^q_{tar},\ \mathrm{K}^i_{tar} = I^i_{tar}\mathrm{W}^k_{tar},\ \mathrm{V}^i_{tar} = I^i_{tar}\mathrm{W}^v_{tar} \tag{5}$$

where $W^{q}_{aux/tar}$, $W^{k}_{aux/tar}$, $W^{v}_{aux/tar}$ are the weight matrices of the convolution operations used to generate query, key, and value matrices for the auxiliary and target modalities, respectively. Different from the traditional cross-attention, in the proposed DCAB, one branch generates a query matrix $Q_{aux/tar}$ and uses $Q_{aux/tar}$ to query the $K_{tar/aux}$ and $K_{tar/aux}$ of the other branch through attention. This method ensures that the features of different branches can effectively interact and fuse, thereby capturing the complementary information between multi-modalities. The process can be expressed as:

$$I^{sx}_{aux} = softmax\left(\frac{Q_{aux}K^{T}_{tar}}{\sqrt{d}}\right)V_{tar} \tag{6}$$

$$I^{sx}_{tar} = softmax\left(\frac{Q_{tar}K^{T}_{aux}}{\sqrt{d}}\right)V_{aux} \tag{7}$$

where I^{sx}_{aux}, I^{sx}_{tar} are the features of the auxiliary modality and the target modality after cross-attention, and d is the embedding dimension of the attention head. In the proposed DCAB, we also have two trainable parameters, α and β, where the model can dynamically adjust the attention weights to optimize the weight configuration during training, i.e., αI^{sx}_{aux} and βI^{sx}_{tar}. In the final stage of DCAB, we integrated LayerNorm (LN) and Multi-layer Perceptron (MLP) to normalize the output features, to ensure that the input data of each layer of the network remains relatively stable while improving the depth and complexity of feature extraction. In addition, residual learning is used to avoid losing fine information and prevent gradient disappearance. The process can be expressed as:

$$z^{l}_{aux} = \mathrm{MLP}\left(\mathrm{LN}\left(I^{\mathrm{RL}}_{aux}\right)\right) + I^{\mathrm{RL}}_{aux} \tag{8}$$

$$z^{l}_{tar} = \mathrm{MLP}\left(\mathrm{LN}\left(I^{\mathrm{RL}}_{tar}\right)\right) + I^{\mathrm{RL}}_{tar} \tag{9}$$

where $I^{\mathrm{RL}}_{aux} = I^{sx}_{aux} + I^{i}_{aux}$ and $I^{\mathrm{RL}}_{tar} = I^{sx}_{tar} + I^{i}_{tar}$ represent the features from the first residual learning, $z^{l}_{aux} \in \mathrm{R}^{\mathrm{H}\cdot 2^{l-1}\times \mathrm{W}\cdot 2^{l-1}\times \mathrm{C}/2^{l-1}}$ and $z^{l}_{tar} \in \mathrm{R}^{\mathrm{H}\cdot 2^{l-1}\times \mathrm{W}\cdot 2^{l-1}\times \mathrm{C}/2^{l-1}}$ represent the features of the auxiliary modality and the target modality from the DCAB, where $l = 1, 2, 3$. Different from the traditional cross-attention, at the end of DCAB, we introduce an LN to stabilize the distribution of activation, which accelerates model training and improves its generalization ability. Additionally, a MLP is employed to enhance the model's ability for extracting complex features through nonlinear transformations, which further refining the capture and comprehension of features across different modalities.

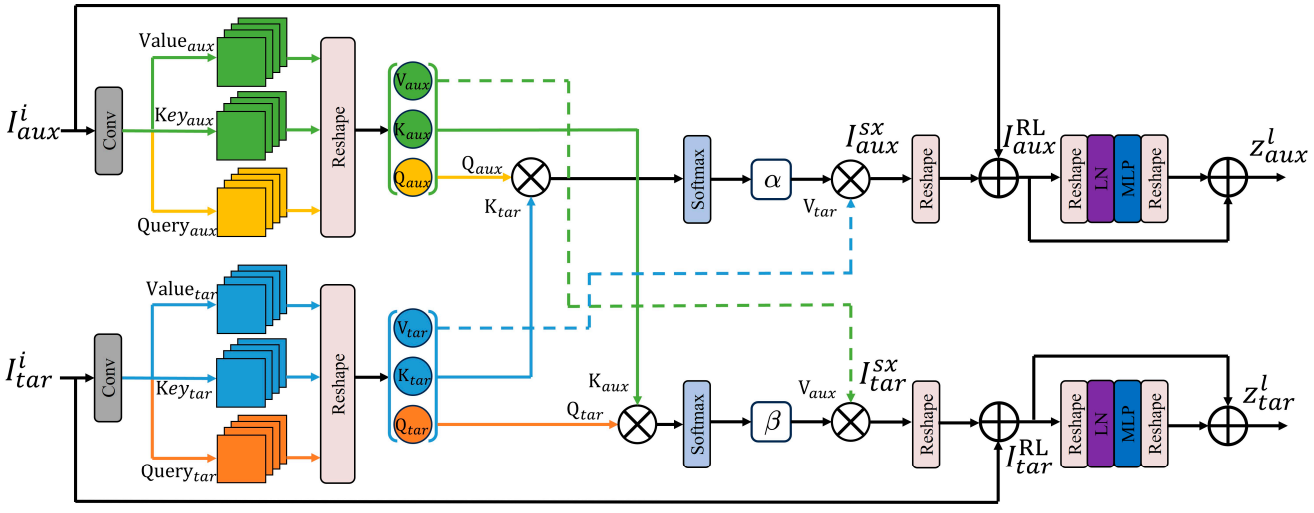


Figure 5. The proposed multi-modal feature interaction module consists of a double cross-attention. To interact with multi-scale features of different modalities at different hierarchical.

3.4. Multi-Scale Feature Fusion of the Same Modality

To achieve in-depth integration of multi-scale features under the same modality, we propose a single-modal feature fusion module (SMFFM). By combining features at different scales, the model is able to capture both detailed information and global context. The SMFFM is designed to synchronize the micro-local details and macro-global context of the image, thereby providing the model with a richer and more comprehensive feature representation. The module consists of a series of convolutions (kernel size = 1, padding = 0), Batch Norm, ReLU activation function, and bilinear interpolation. First, we perform convolution, Batch Norm, ReLU, and interpolation operations on the first output of MMFIM z^1_{aux}/z^1_{tar} to match the feature dimension of the second output z^2_{aux}/z^2_{tar}. Then, we concatenate the two feature maps with the same dimension and repeat this operation. Finally, we combine the fused features with the output features of the decoder I^4_{aux}/I^4_{tar} to generate the final output features. The process can be expressed as:

$$g_{aux} = \mathrm{G}_4(\mathrm{G}_3(\mathrm{G}_2(\mathrm{G}_1\left(z^1_{aux} \oplus z^2_{aux}\right) \oplus z^3_{aux}\right) \oplus I^4_{aux}\left)\right) \tag{10}$$

$$g_{tar} = \mathrm{G}_4(\mathrm{G}_3(\mathrm{G}_2(\mathrm{G}_1\left(z^1_{tar} \oplus z^2_{tar}\right) \oplus z^3_{tar}\right) \oplus I^4_{tar}\left)\right) \tag{11}$$

where $\oplus$ represents concatenate operation, and $\mathrm{G_h}(\cdot)$, $h = 1, 2, 3$ represents a series of operations shown in Figure 6, including convolution, Batch Norm, ReLU, and bilinear interpolation, and $\mathrm{G}_4(\cdot)$, the bilinear interpolation step, is omitted, but convolution, Batch Norm, and ReLU are still retained to ensure that the feature map can maintain the integrity and expressiveness of its spatial features while matching the dimensions. In summary, SMFFM achieves effective multi-scale feature fusion by extracting features at different scales and fusing these features through upsampling, concatenate, and further convolutional processing.

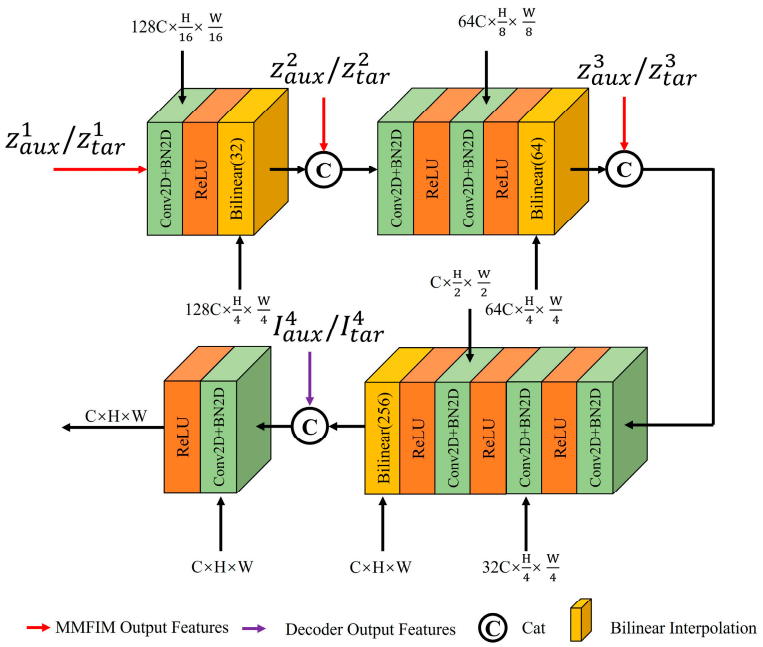


Figure 6. Detailed architecture of the single-modal feature fusion module.

3.5. Loss Function

Following [18,53], we use the l_2 loss to evaluate image domain reconstruction:

$$\mathcal{L} = \frac{1}{\mathrm{M}}\sum_{m=1}^{\mathrm{M}} \alpha \left\| x'^m_{aux} - x^m_{aux} \right\|_2^2 + (1-\alpha)\left\| x'^m_{tar} - x^m_{tar} \right\|_2^2 \tag{12}$$

where M represents the batch size, and $\alpha = 0.6$ is the weighted parameter to balance each loss item.

4. Experiments

In this section, we elaborate on the datasets used in the experiments and the specific implementation details of the experiments. Finally, we introduce the metrics used to evaluate the performance of the model.

4.1. Datasets

We use two raw MR images to evaluate our method. The fastMRI [54,55] dataset is a large, publicly available, repository of raw MR images, provided by the New York University School of Medicine at https://fastmri.med.nyu.edu/ (accessed on 18 August 2024). (1) fastMRI Brain: we filter out 768 and 96 pairs of T1WI and T2WI brain slices for training and validation. The resolution of each slice was 256 × 256. (2) fastMRI Knee: we filter out 160 and 20 pairs of PDs and PDFSs knee slices for training and validation. The resolution of each slice was 256 × 256. In both datasets, each MR image is normalized to the range [0, 1] by dividing each pixel value by the pixel value with the largest absolute value within the image. This process helps to minimize variability across different images and ensures data consistency. T1WI and PDs serve as the auxiliary modal, while T2WI and PDFSs serve as the target modal.

4.2. Experimental Details

The proposed method is running in Pytorch (2.0.0), with an NVIDIA 3090Ti GPU (NVIDIA, Santa Clara, CA, USA) and i9-12900KF CPU (Intel Corporation, Santa Clara, CA, USA) equipped with 24 GB of video memory. The AdamW optimizer is used with an initial learning rate of 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the learning decay is 0.9. The model is trained for 100 epochs with a batch size of 1 until convergence. Also, we adopted three dual cross-attention blocks to construct the MMFIM. In addition, we conducted experiments on the knee and brain datasets of fastMRI [54,55]. On the knee dataset, we first applied a one-dimensional Cartesian random sampling pattern and conducted experiments under the conditions of 10× and 12× acceleration factors. Then, on the brain dataset, we adopted two sampling patterns: one-dimensional Cartesian random and equispaced, and conducted experiments under the acceleration factors of 4×, 8×, and 12×. We also compared our method with six state-of-the-art methods, including three single-modality MRI reconstruction methods: U-Net [51], STUN [15], and PC-UNet [47], and three multi-modality MRI reconstruction methods: MINet [49], DuDoCAF [27], and DuDoRNet [14].

4.3. Evaluation Metrics

The peak signal-to-noise ratio (PSNR) is a widely recognized metric for evaluating image quality, especially effective in comparing the discrepancies between fully sampled reference images and their processed or reconstructed versions. Within the realm of MRI, PSNR serves as a valuable tool for assessing quality of a reconstructed image, and thus assessing the performance of reconstruction methods. Higher PSNR values indicate a higher signal-to-noise ratio in the reconstructed image, corresponding to a reconstruction method with better suppression of impurities and noise. The PSNR is defined as:

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \|x - x\prime\|_2^2 \tag{13}$$

$$\mathrm{PSNR}(x', x) = 10\log_{10} \frac{\max(x')^2}{\mathrm{MSE}(x, x')} \tag{14}$$

where N is the number of pixels, x is the fully-sampled MR images of the auxiliary modality or the target modality, and $x\prime$ is the reconstruction image and PSNR in decibels (dB). The relative l_2 norm error (RLNE) is a standard metric employed in MRI for assessing the

accuracy of reconstructed images against the fully sampled images. The RLNE quantifies the relative error between the reconstructed and fully sampled images, providing a measure of how closely the reconstruction process adheres to the original data. The lower the RLNE, the smaller the difference between the reconstructed image and the fully sampled image. The RLNE is defined as:

$$\mathrm{RLNE}(x', x) = \frac{\|x - x'\|_2}{\|x\|_2} \tag{15}$$

The structural similarity index (SSIM) is a metric designed to assess image quality by focusing on the quantification of structural information within images. In the context of MRI, SSIM plays a crucial role in evaluating the performance of image reconstruction techniques, particularly with respect to the preservation of detail. Higher SSIM indicates greater structural similarity between two images, which usually means better image quality. The SSIM is defined as:

$$\mathrm{SSIM}(x', x) = \frac{(2\mu_x\mu_{x'} + \mathrm{C}_1)(2\sigma_{xx\prime} + \mathrm{C}_2)}{(\mu_x^2 + \mu_{x'}^2 + \mathrm{C}_1)(\sigma_x^2 + \sigma_{x'}^2 + \mathrm{C}_2)} \tag{16}$$

where μ_x and $\mu_{x'}$ denote the mean values of x and $x\prime$, σ_x and $\sigma_{x'}$ denote the local variances, and $\sigma_{xx\prime}$ denote the local covariance between these two MR images. $\mathrm{C}_1 = 0.03$ and $\mathrm{C}_2 = 0.07$ are the constant relaxation terms.

5. Results and Discussion

In this section, we exhaustively describe the results of the comparison of our proposed method with other existing methods in terms of quantitative and qualitative aspects, show the analysis of ablation experiments, and present a discussion of computational complexity.

5.1. Quantitative Results

We use two Cartesian sampling patterns (random and equispaced) and four acceleration factors (4×, 8×, 10×, 12×), and measure the performance based on three evaluation metrics (PSNR, SSIM, RLNE). In Table 1, we show the results of three different acceleration factors using random Cartesian sampling on the fastMRI datasets. The results show that our method performs best in all cases. Specifically, in the T1WI, the corresponding PSNR values of the three acceleration factors are 42.969 dB, 39.896 dB, and 35.571 dB, respectively; while in the T2WI, the corresponding PSNR values are 38.815 dB, 35.235 dB, and 32.607 dB, respectively. This proves that our proposed method can effectively utilize the complementary information between different modalities, thereby achieving effective performance improvement in both modalities. In addition, when using the U-Net [51] for experiments, we simply input images of two different modalities into two independent branches for processing. However, the performance of this method is not ideal. Similarly, DuDoRNet [14] combines images of different modalities by alternating learning in the k-space domain and the image domain, and performing multi-cascade operations. Although the results are better than those of a single fusion method, they are still slightly inferior to those of our proposed method. Compared with DuDoRNet, our method improves PSNR by 3.052 dB, 5.613 dB, and 4.185 dB, SSIM by 0.011, 0.056, and 0.049, and RLNE by 0.011, 0.033, and 0.038, respectively, under different acceleration factors (4×, 8×, and 12×) in the T1WI. In the T2WI, the PSNR gains are 2.856 dB, 4.691 dB, and 4.706 dB, the SSIM increases by 0.013, 0.041, and 0.064, and the RLNE decreases by 0.017, 0.048, and 0.064, respectively. In addition, as the acceleration factor increases, the performance of DuDoRNet shows a gradually weakening trend. In contrast, our method exhibits better robustness.

Table 1. Quantitative comparison of different reconstruction methods on fastMRI brain datasets in 1D random undersampling pattern with three acceleration factors.

Methods	fastMRI Brain					
	T1WI			T2WI		
	PSNR ↑ [3]	SSIM ↑	RLNE ↓ [4]	PSNR ↑	SSIM ↑	RLNE ↓
1D Random 4× (Mean (Standard))						
ZF [1]	27.204 (1.763)	0.735 (0.060)	0.158 (0.039)	25.801 (1.268)	0.790 (0.040)	0.195 (0.024)
U-Net [51]	36.619 (1.103)	0.932 (0.020)	0.053 (0.010)	32.789 (1.042)	0.929 (0.023)	0.087 (0.009)
STUN [15]	38.195 (1.326)	0.944 (0.018)	0.044 (0.009)	34.325 (1.815)	0.940 (0.030)	0.074 (0.017)
PC-UNet [47]	41.078 (1.172)	0.966 (0.012)	0.031 (0.006)	37.237 (1.133)	0.960 (0.017)	0.052 (0.006)
MINet [49]	37.327 (1.352)	0.936 (0.019)	0.049 (0.011)	33.264 (1.230)	0.932 (0.027)	0.083 (0.011)
DuDoCAF [27]	40.915 (0.998)	0.962 (0.011)	0.032 (0.005)	36.937 (1.059)	0.956 (0.018)	0.054 (0.006)
DuDoRNet [14]	39.917 (1.146)	0.963 (0.012)	0.036 (0.007)	35.959 (1.137)	0.955 (0.018)	0.061 (0.008)
MHCFIN [2]	**42.969 (1.238)**	**0.974 (0.010)**	**0.025 (0.005)**	**38.815 (1.183)**	**0.968 (0.015)**	**0.044 (0.006)**
1D Random 8× (Mean (Standard))						
ZF	22.572 (1.057)	0.620 (0.065)	0.265 (0.045)	21.764 (1.054)	0.673 (0.041)	0.310 (0.033)
U-Net [51]	33.965 (1.239)	0.910 (0.024)	0.072 (0.016)	30.485 (0.978)	0.906 (0.031)	0.113 (0.011)
STUN [15]	34.297 (1.349)	0.893 (0.028)	0.069 (0.017)	30.634 (1.284)	0.905 (0.037)	0.112 (0.017)
PC-UNet [47]	37.845 (1.125)	0.951 (0.016)	0.046 (0.009)	34.515 (1.073)	0.939 (0.024)	0.080 (0.009)
MINet [49]	33.070 (1.522)	0.896 (0.028)	0.081 (0.022)	28.933 (1.418)	0.889 (0.044)	0.137 (0.023)
DuDoCAF [27]	37.443 (0.888)	0.945 (0.017)	0.050 (0.008)	33.675 (0.965)	0.937 (0.022)	0.079 (0.008)
DuDoRNet [14]	34.283 (1.229)	0.908 (0.023)	0.069 (0.016)	30.544 (1.009)	0.907 (0.031)	0.113 (0.011)
MHCFIN	**39.896 (1.073)**	**0.964 (0.014)**	**0.036 (0.006)**	**35.235 (1.049)**	**0.948 (0.025)**	**0.065 (0.007)**
1D Random 12× (Mean (Standard))						
ZF	20.805 (0.999)	0.548 (0.071)	0.324 (0.050)	20.400 (1.005)	0.602 (0.041)	0.362 (0.036)
U-Net [51]	25.881 (1.025)	0.741 (0.050)	0.181 (0.031)	22.564 (0.967)	0.726 (0.054)	0.282 (0.025)
STUN [15]	31.697 (1.427)	0.857 (0.038)	0.094 (0.025)	28.155 (1.485)	0.864 (0.053)	0.150 (0.027)
PC-UNet [47]	34.951 (1.493)	0.912 (0.030)	0.064 (0.015)	31.752 (1.067)	0.922 (0.032)	0.098 (0.011)
MINet [49]	30.157 (1.680)	0.842 (0.043)	0.113 (0.033)	26.684 (1.806)	0.851 (0.066)	0.179 (0.041)
DuDoCAF [27]	34.077 (1.017)	0.910 (0.020)	0.070 (0.012)	31.404 (0.879)	0.915 (0.025)	0.102 (0.009)
DuDoRNet [14]	31.386 (1.253)	0.865 (0.032)	0.097 (0.022)	27.901 (0.971)	0.865 (0.039)	0.153 (0.015)
MHCFIN	**35.571 (0.963)**	**0.914 (0.028)**	**0.059 (0.011)**	**32.607 (0.903)**	**0.929 (0.029)**	**0.089 (0.008)**

[1] zero-filled, [2] multi-hierarchical complementary feature interaction network, [3] the higher the better, [4] the lower the better. The best results are labeled in bold.

Similarly, we also adopted a 1D equispaced undersampling pattern and conducted experiments on three acceleration factors on the brain dataset. From the comparison between Tables 1 and 2, when using the one-dimensional Cartesian sampling pattern, the performance difference between the two is not significant. At a 12× acceleration factor, our method outperforms DuDoCAF [27] in multi-modal reconstruction performance, with PSNR improving from 34.481 dB to 36.277 dB and from 31.828 dB to 32.302 dB. At the same time, SSIM improves from 0.911 to 0.913 and from 0.917 to 0.922, respectively, while RLNE decreases from 0.067 to 0.055 and from 0.097 to 0.092, corresponding to the T1WI and the T2WI, respectively. DuDoCAF [27] is slightly lower than our method in PSNR and RLNE but is fairly close to ours in SSIM. We speculate that this may be due to the advantage of its dual-domain alternating learning architecture.

Table 2. Quantitative comparison of different reconstruction methods on fastMRI brain datasets in 1D equispaced undersampling pattern with three acceleration factors.

Methods	fastMRI Brain					
	T1WI			T2WI		
	PSNR ↑ [3]	SSIM ↑	RLNE ↓ [4]	PSNR ↑	SSIM ↑	RLNE ↓
1D Equispaced 4× (Mean (Standard))						
ZF [1]	23.598 (1.064)	0.618 (0.062)	0.236 (0.040)	22.837 (1.028)	0.682 (0.036)	0.274 (0.028)
U-Net [51]	36.155 (1.248)	0.909 (0.028)	0.056 (0.012)	32.448 (1.041)	0.915 (0.030)	0.091 (0.010)
STUN [15]	38.428 (1.372)	0.937 (0.020)	0.043 (0.010)	35.267 (1.322)	0.940 (0.025)	0.066 (0.010)
PC-UNet [47]	40.659 (1.153)	0.959 (0.013)	0.033 (0.006)	36.963 (1.130)	0.955 (0.019)	0.054 (0.007)
MINet [49]	36.370 (1.727)	0.909 (0.028)	0.055 (0.015)	32.679 (1.380)	0.920 (0.033)	0.089 (0.014)
DuDoCAF [27]	39.313 (0.881)	0.942 (0.016)	0.039 (0.006)	35.500 (0.970)	0.941 (0.021)	0.064 (0.007)
DuDoRNet [14]	39.868 (1.293)	0.956 (0.014)	0.036 (0.008)	36.029 (1.220)	0.951 (0.021)	0.060 (0.008)
MHCFIN [2]	**42.985 (1.246)**	**0.970 (0.010)**	**0.025 (0.005)**	**38.924 (1.172)**	**0.966 (0.016)**	**0.043 (0.006)**
1D Equispaced 8× (Mean (Standard))						
ZF	20.666 (1.025)	0.533 (0.069)	0.329 (0.050)	20.414 (0.978)	0.597 (0.039)	0.066 (0.007)
U-Net [51]	32.389 (1.774)	0.868 (0.045)	0.087 (0.022)	28.497 (1.740)	0.868 (0.057)	0.145 (0.029)
STUN [15]	33.840 (1.455)	0.872 (0.036)	0.073 (0.019)	30.605 (1.256)	0.891 (0.041)	0.112 (0.016)
PC-UNet [47]	37.363 (1.144)	0.940 (0.019)	0.048 (0.010)	33.523 (1.069)	0.934 (0.026)	0.080 (0.009)
MINet [49]	32.504 (1.683)	0.867 (0.040)	0.086 (0.026)	28.310 (1.678)	0.874 (0.053)	0.148 (0.030)
DuDoCAF [27]	36.858 (0.880)	0.931 (0.016)	0.051 (0.009)	33.227 (0.930)	0.927 (0.024)	0.083 (0.008)
DuDoRNet [14]	33.866 (1.404)	0.885 (0.030)	0.073 (0.018)	30.404 (1.006)	0.898 (0.034)	0.114 (0.012)
MHCFIN	**38.242 (0.964)**	**0.948 (0.017)**	**0.044 (0.007)**	**35.146 (0.981)**	**0.948 (0.019)**	**0.066 (0.007)**
1D Equispaced 12× (Mean (Standard))						
ZF	19.783 (1.024)	0.511 (0.072)	0.365 (0.055)	19.504 (0.925)	0.560 (0.041)	0.401 (0.035)
U-Net [51]	26.172 (1.070)	0.732 (0.052)	0.175 (0.031)	22.314 (1.256)	0.722 (0.068)	0.292 (0.040)
STUN [15]	31.699 (1.396)	0.835 (0.045)	0.094 (0.024)	28.170 (1.483)	0.859 (0.058)	0.149 (0.027)
PC-UNet [47]	34.906 (1.479)	0.911 (0.032)	0.065 (0.015)	32.020 (1.027)	0.921 (0.031)	0.095 (0.010)
MINet [49]	30.079 (1.765)	0.833 (0.047)	0.114 (0.036)	26.182 (1.870)	0.837 (0.072)	0.190 (0.046)
DuDoCAF [27]	34.481 (1.017)	0.911 (0.020)	0.067 (0.012)	31.828 (0.835)	0.917 (0.025)	0.097 (0.008)
DuDoRNet [14]	31.753 (1.336)	0.858 (0.034)	0.093 (0.023)	28.449 (0.911)	0.869 (0.039)	0.143 (0.013)
MHCFIN	**36.277 (1.013)**	**0.913 (0.031)**	**0.055 (0.010)**	**32.302 (0.890)**	**0.922 (0.030)**	**0.092 (0.008)**

[1] zero-filled, [2] multi-hierarchical complementary feature interaction network, [3] the higher the better, [4] the lower the better. The best results are labeled in bold.

In addition, we also applied a 1D Cartesian random sampling pattern with two acceleration factors (10×, 12×) to conduct experiments on the knee dataset. As can be seen from Table 3, MINet [49] does not perform satisfactorily among the multi-modal reconstruction methods. We replaced LR with zero-filled images and omitted the upsampling step. Although it performs well on HR, it does not perform well when processing the fastMRI dataset. Compared to the second-best DuDoCAF [27], at an acceleration factor of 10×, we achieved a 1.6% and 2.6% increase in PSNR, and a 2.7% and 1.8% increase in SSIM on PDs and PDFSs, respectively.

Table 3. Quantitative comparison of different reconstruction methods on fastMRI knee datasets in 1D random undersampling pattern with two acceleration factors.

Methods	fastMRI Knee					
	PDs			PDFSs		
	PSNR ↑ [3]	SSIM ↑	RLNE ↓ [4]	PSNR ↑	SSIM ↑	RLNE ↓
1D Random 10× (Mean (Standard))						
ZF [1]	23.822 (1.271)	0.648 (0.047)	0.383 (0.065)	27.492 (1.725)	0.734 (0.033)	0.314 (0.046)
U-Net [51]	32.409 (1.312)	0.849 (0.039)	0.142 (0.023)	33.731 (1.863)	0.871 (0.024)	0.153 (0.021)
STUN [15]	33.927 (1.372)	0.856 (0.039)	0.120 (0.024)	37.427 (1.933)	0.905 (0.021)	0.100 (0.019)
PC-UNet [47]	32.464 (1.653)	0.847 (0.050)	0.142 (0.028)	34.758 (1.568)	0.893 (0.022)	0.136 (0.021)
MINet [49]	29.454 (1.376)	0.773 (0.043)	0.201 (0.038)	32.367 (1.941)	0.834 (0.032)	0.179 (0.027)
DuDoCAF [27]	35.327 (1.363)	0.897 (0.027)	0.102 (0.020)	37.297 (2.148)	0.926 (0.019)	0.102 (0.019)
DuDoRNet [14]	33.915 (1.206)	0.873 (0.033)	0.119 (0.015)	34.962 (1.516)	0.874 (0.024)	0.132 (0.014)
MHCFIN [2]	**35.877 (1.709)**	**0.921 (0.025)**	**0.096 (0.023)**	**38.256 (2.236)**	**0.943 (0.015)**	**0.092 (0.018)**
1D Random 12× (Mean (Standard))						
ZF	22.077 (1.069)	0.614 (0.041)	0.464 (0.052)	25.535 (1.651)	0.682 (0.037)	0.391 (0.040)
U-Net [51]	30.853 (1.205)	0.828 (0.046)	0.170 (0.024)	32.285 (1.622)	0.852 (0.023)	0.180 (0.020)
STUN [15]	32.466 (1.523)	0.790 (0.045)	0.142 (0.026)	35.886 (1.943)	0.888 (0.024)	0.120 (0.021)
PC-UNet [47]	31.113 (1.544)	0.820 (0.051)	0.166 (0.031)	33.459 (1.812)	0.856 (0.030)	0.157 (0.016)
MINet [49]	28.830 (1.814)	0.777 (0.042)	0.218 (0.052)	31.799 (1.797)	0.837 (0.030)	0.191 (0.029)
DuDoCAF [27]	34.093 (1.472)	0.879 (0.035)	0.118 (0.025)	36.896 (2.560)	0.925 (0.021)	0.108 (0.022)
DuDoRNet [14]	32.365 (1.266)	0.859 (0.032)	0.142 (0.022)	33.772 (1.633)	0.865 (0.025)	0.152 (0.018)
MHCFIN	**34.132 (1.748)**	**0.892 (0.024)**	**0.117 (0.027)**	**37.771 (2.790)**	**0.936 (0.018)**	**0.099 (0.028)**

[1] zero-filled, [2] multi-hierarchical complementary feature interaction network, [3] the higher the better, [4] the lower the better. The best results are labeled in bold.

5.2. Qualitative Results

Figures 7 and 8 present T1WI and T2WI reconstructions from the fastMRI brain dataset under different undersampling conditions: an 8× acceleration factor with a one-dimensional random sampling pattern and a 12× factor with a one-dimensional equispaced sampling pattern. These illustrations compare the performance of various reconstruction methods. At the same time, the error map between the real image and the reconstructed image and the enlarged image of the reconstructed details is also shown in the figure. The error map has a numerical range of $[-0.4, 0.4]$, and the color depth represents the quality of the reconstruction. The results show that the zero-filled image leads to a large error and a serious loss of details. It is obvious from the figures that our proposed method achieves the lowest error in T1WI and T2WI reconstruction and effectively preserves the key details.

Figure 9 shows the reconstruction results of knee images using different methods with a 10× acceleration factor using a one-dimensional Cartesian random undersampling pattern. Among the multi-modal reconstruction methods, MINet achieved unsatisfactory results in the reconstruction results of PDs and PDFSs, and it seems that its effect on processing fastMRI data is not as good as HR. From observation of the figure, we can find that our method achieves the lowest error in PDs and PDFSs reconstruction, which verifies the effectiveness of our use of multi-hierarchical feature interaction to learn complementary information.

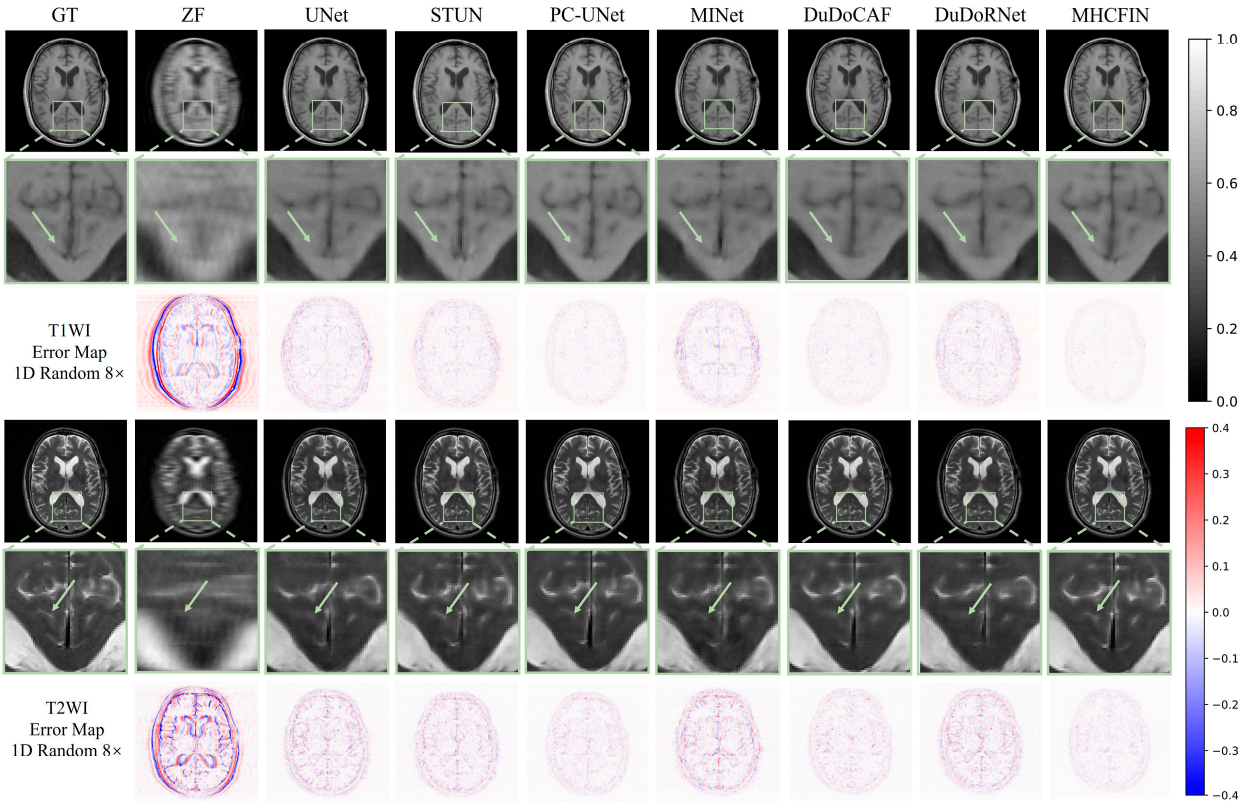


Figure 7. Qualitative comparison with different reconstruction methods using 1D random undersampling pattern with acceleration factor 8× on fastMRI brain datasets. Ground truth (GT), zero-filled (ZF), reconstructed MR images (T1WI and T2WI), error maps, and zoomed-in details are provided.

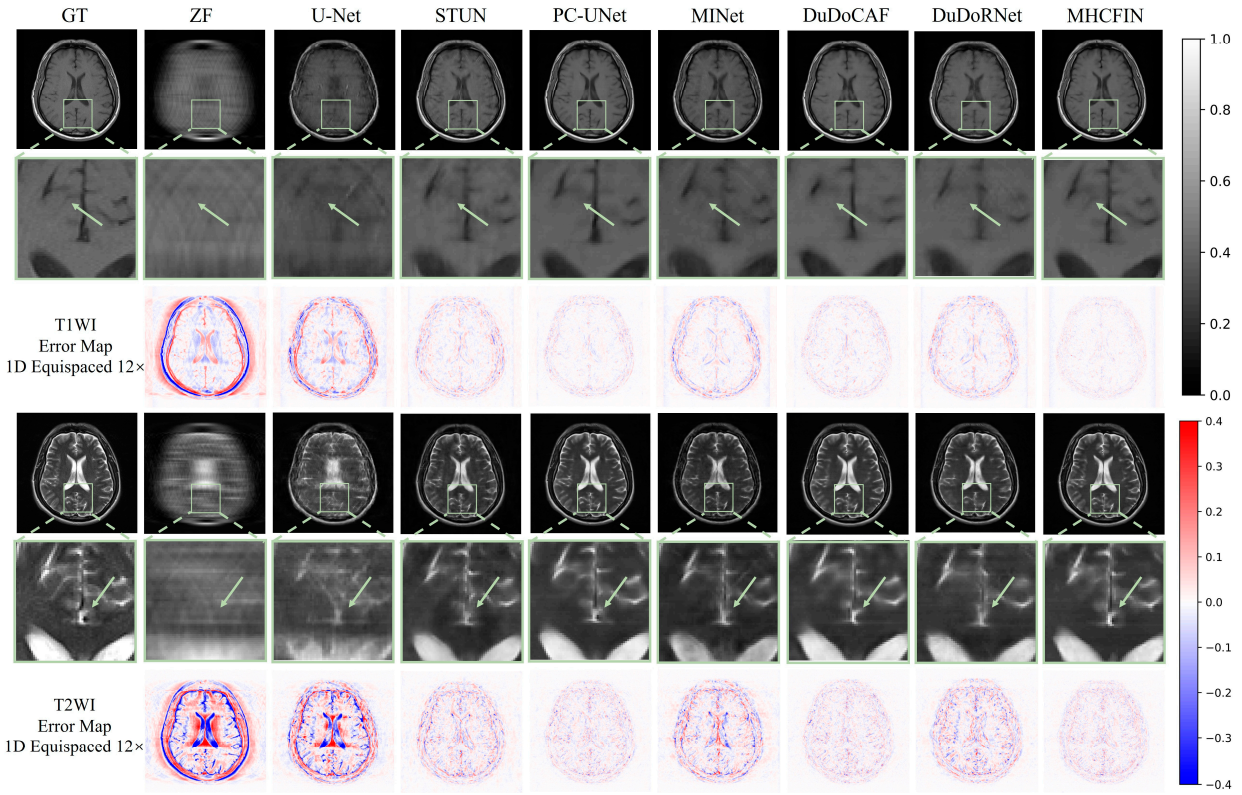


Figure 8. Qualitative comparison with different reconstruction methods using 1D equispaced undersampling pattern with acceleration factor 12× on fastMRI brain datasets.

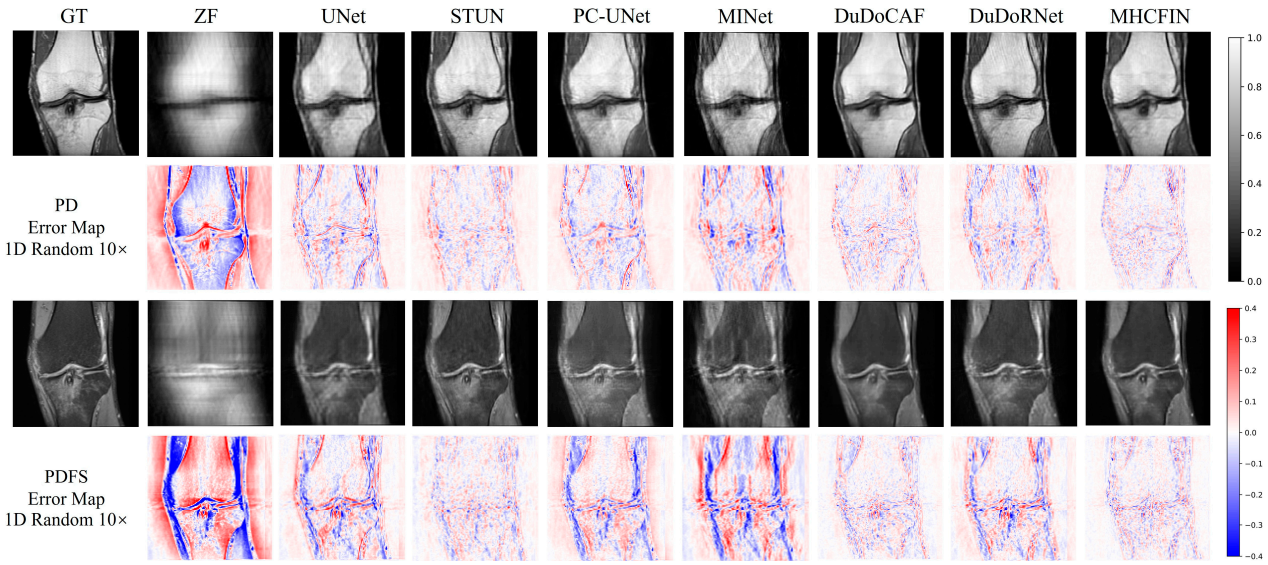


Figure 9. Qualitative comparison with different reconstruction methods using 1D random undersampling pattern with acceleration factor 10× on fastMRI knee datasets.

Figure 10 shows the performance evaluation of different reconstruction methods on the fastMRI brain dataset using an 8× acceleration factor in the form of a bar graph. Part (a) shows the PSNR value, part (b) shows the SSIM value and part (c) reflects the RLNE value. The standard deviation of the performance of each method is represented by a black arrow, attached to the corresponding bar graph. The observation results show that, in parts (a) and (b), our proposed reconstruction method shows good performance on T1WI and T2WI data, outperforming other methods. In addition, in part (c), the RLNE value of our method is the lowest, further confirming its high accuracy in data reconstruction. These results verify the effectiveness of the MHCFIN method in learning transferable complementary knowledge through multi-hierarchical feature interactions.

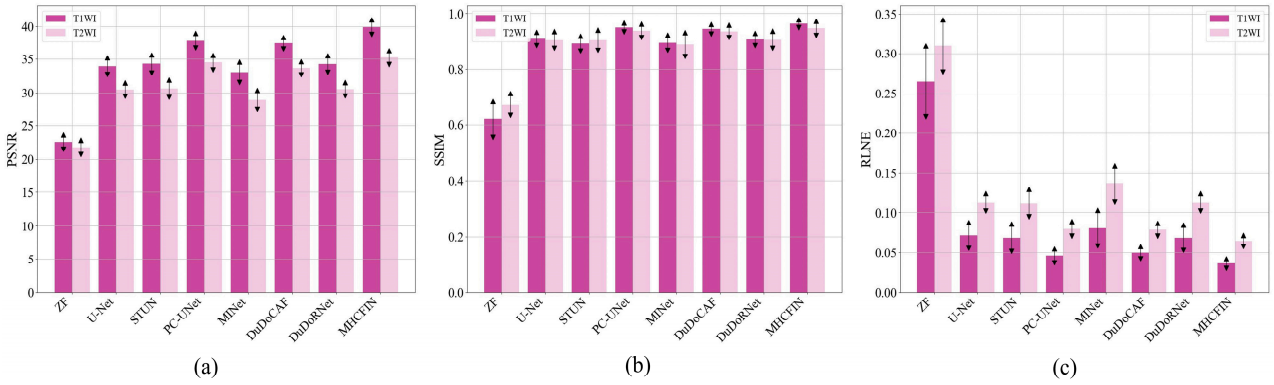


Figure 10. Bar charts for (**a**) PSNR, (**b**) SSIM, and (**c**) RLNE depicting the performance of various reconstruction methods using an 8× acceleration factor mask on the fastMRI brain dataset. The black arrows represent the standard deviation (mean ± standard deviation) for each method.

Figure 11 shows the trend that the loss values of each reconstruction method gradually decrease as the number of epochs increases during the training process. The analysis results show that our proposed method first achieved the lowest loss value and showed stable convergence characteristics. It was followed by the DuDoCAF method, which had a slightly higher loss value than our method but also showed good performance. In contrast, the loss value of MINet fluctuated greatly during the training process, which reflects its instability when processing the fastMRI dataset, and thus shows that its application effect on this dataset may not be as ideal as other methods.

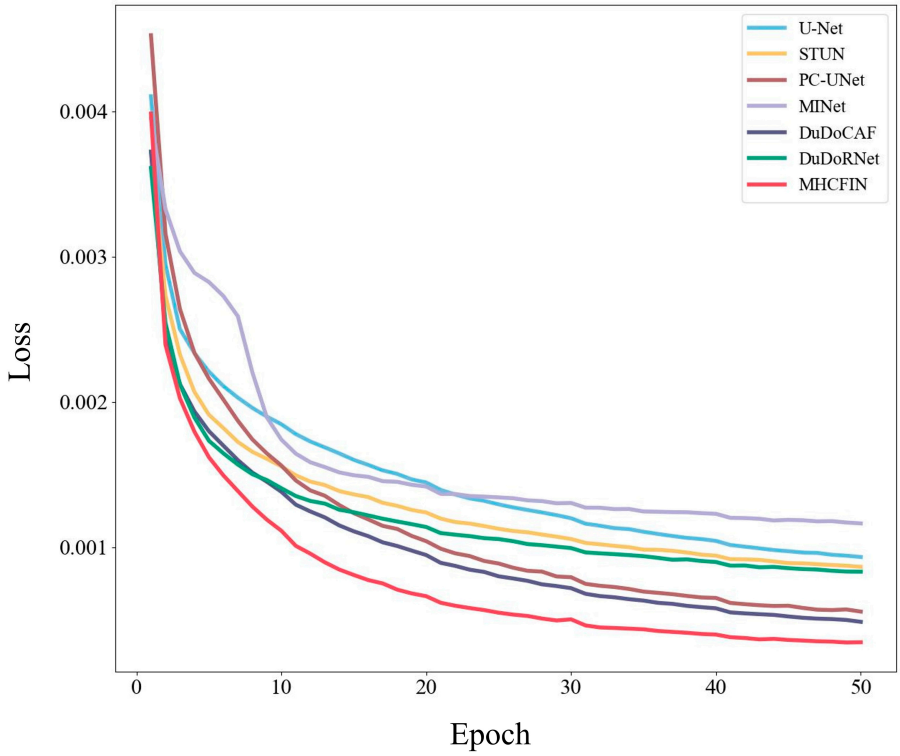


Figure 11. The training loss of different reconstruction methods on the fastMRI brain dataset.

5.3. Discussion

(a) **Discussion on encoder and decoder based on residual multi-attention:** To verify the efficacy of the key components in the proposed method, we evaluate three models in an ablation study. First, we evaluate the proposed residual multi-attention-based decoder–encoder model, in which we replace residual learning with traditional skip connections and remove the triple-attention mechanism. From the data in Table 4, we can see that, compared with configurations 1 and 2, without residual multi-attention (RMA), the PSNR of T1WI decreases by 2.974 dB and the SSIM decreases by 0.037; similarly, the PSNR of T2WI decreases by 5.964 dB and the SSIM decreases by 0.08. It can also be observed in Figure 12 that the reconstruction error of configuration 2 is reduced compared to configuration 1 which does not integrate any components, and the restored image details are clearer.

Table 4. Ablation study on different variant methods on fastMRI brain datasets with the 1D random undersampling pattern at acceleration rate 8×.

Variant	RMA	MMFIM	SMFFM	fastMRI Brain					
				T1WI			T2WI		
				PSNR ↑ [3]	SSIM↑	RLNE↓ [4]	PSNR ↑	SSIM↑	RLNE↓
				1D Random 8× (Mean)					
ZF	× [1]	×	×	22.572	0.646	0.300	21.764	0.673	0.310
1	×	×	×	33.599	0.902	0.074	27.261	0.856	0.164
2	√ [2]	×	×	36.573	0.939	0.053	33.225	0.936	0.083
3	×	√	√	34.381	0.907	0.068	30.599	0.908	0.112
MHCFIN	√	√	√	**39.896**	**0.964**	**0.036**	**35.235**	**0.948**	**0.065**

[1] add the module, [2] remove the module. [3] the higher the better, [4] the lower the better. The best results are labeled in bold.

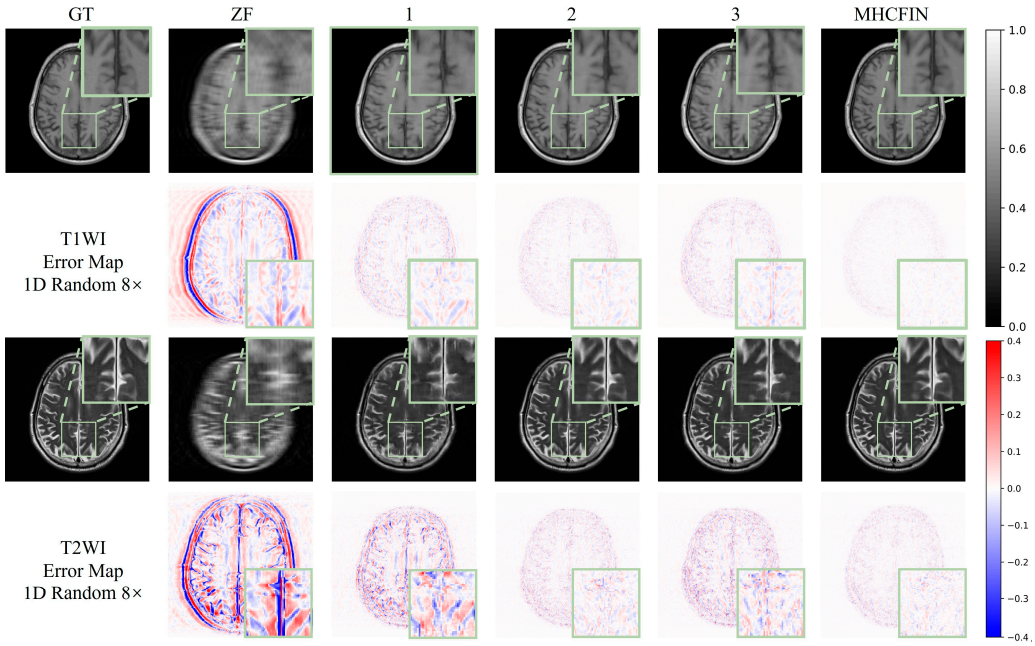


Figure 12. Ablation study of the key components in the proposed method using 1D random undersampling pattern with acceleration factor 8× on fastMRI brain datasets.

(b) **Discussion on the multi-modal feature interaction module and single-modal feature fusion module:** Since the multi-modal fusion interaction module and the single-modal feature fusion module are used jointly, we discuss their effects together in the ablation study. As can be seen from the data in Table 4, compared with configuration 1 without using these feature fusion technologies, configuration 3 has the PSNR of T1WI increased by 0.782 dB, SSIM increased by 0.005, and RLNE decreased with the help of multi-modal feature interactive learning by 0.006; the improvement of T2WI is more significant, as PSNR increases by 3.338 dB, SSIM increases by 0.052, and RLNE decreases by 0.052. Figure 12 also clearly shows that, after the introduction of MMFIM and SMFFM, the overall reconstruction error is reduced; especially, the improvement on T2WI is particularly obvious. Based on the introduction of RMA, the difference between MMFIM and SMFFM is particularly significant. By comparing configuration MHCFIN with configuration 2, it can be seen that the PSNR of T1WI has increased by 3.323 dB, and the PSNR of T2WI has increased by 2.01 dB. This result verifies the effectiveness of our proposed components, and when these components are used jointly, they can bring better results than when used alone. At the same time, learning complementary information between different modalities can also further improve the quality of reconstruction.

(c) **Discussion of the computational complexity of the proposed method:** When evaluating the computational complexity of our proposed dual-branch learning network, we recognize that, while the network significantly enhances image quality through the use of TAB, MMFIM, SMFFM, these advanced feature extraction techniques—such as attention mechanisms, convolution operations, and residual connections—also introduce notable computational overhead. Our tests show that, on a system equipped with an NVIDIA 3090Ti GPU and 24 GB of memory, processing 96 brain images with a resolution of 256 × 256 pixels take approximately 18 s, averaging 0.2 s per MR image. While this performance is sufficient for many clinical applications, there is still room for improvement in scenarios requiring real-time feedback. To address these concerns, we plan to explore techniques like Linformer [56], which approximates the attention mechanism to reduce computational complexity, or model pruning to reduce model size. Our ultimate goal is to develop a lightweight model that can meet the real-time processing demands of clinical applications.

6. Conclusions

In this study, we introduce an innovative multi-modal reconstruction approach, termed MHCFIN, that facilitates high-fidelity image reconstruction by harnessing the latent, learnable complementary information inherent across modalities. The network adopts a dual-branch architecture and integrates a multi-attention mechanism to focus on the key channels and areas of the image, accurately capture the core features and identify important shared features between different modalities. A gating mechanism is used to dynamically adjust the learning process to effectively reduce the interference of irrelevant information. Then, with a dual cross-attention interactions block, complementary information between multi-modality is learned and fused. Experimental results show that the proposed joint reconstruction network outperforms the state-of-the-art reconstruction network in reconstruction performance, with better preservation of details and smaller reconstruction errors.

Author Contributions: Conceptualization, H.Z. and Z.L.; methodology, H.Z.; software, H.Z.; validation, Q.M., Y.Q. and Z.L.; formal analysis, H.Z.; investigation, Z.L.; data curation, Y.Q.; writing—original draft preparation, H.Z.; writing—review and editing, Q.M. and Z.L.; visualization, H.Z.; supervision, Z.L.; project administration, Z.L.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 61901188, the Natural Science Foundation of Fujian Province of China under Grant 2022J05163, and the Science and Technology Fund of Fujian Education Department under Grant JT180280.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used for the proposed method is the publicly available fastMRI dataset (https://fastmri.med.nyu.edu/, accessed on 18 August 2024), provided by New York University.

Acknowledgments: The authors would like to thank sponsoring organizations.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Abbreviation

MRI	Magnetic resonance imaging
MHCFIN	Multi-hierarchical complementary feature interaction network
MMFIM	Multi-modal feature interaction module
SMFFM	Single-modal feature fusion module
TAB	Triple attention block
DCAB	Dual cross-attention block
T1WI	T1-weighted images
T2WI	T2-weighted images
PDs	Proton density-weighted images
PDFSs	Fat-suppressed proton density-weighted images
PSNR	Peak signal-to-noise ratio
SSIM	Structural similarity index measure
RLNE	Relative l_2 norm error

References

1. Song, P.; Weizman, L.; Mota, J.F.; Eldar, Y.C.; Rodrigues, M.R. Coupled dictionary learning for multi-contrast MRI reconstruction. *IEEE Trans. Med. Imaging* **2019**, *39*, 621–633. [CrossRef] [PubMed]
2. Lei, P.; Fang, F.; Zhang, G.; Zeng, T. Decomposition-based variational network for multi-contrast MRI super-resolution and reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 21296–21306.
3. Lustig, M.; Donoho, D.; Pauly, J.M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med.* **2007**, *58*, 1182–1195. [CrossRef] [PubMed]
4. Fessler, J.A. Model-based image reconstruction for MRI. *IEEE Signal Process. Mag.* **2010**, *27*, 81–89. [CrossRef] [PubMed]

5. Qu, X.; Zhang, W.; Guo, D.; Cai, C.; Cai, S.; Chen, Z. Iterative thresholding compressed sensing MRI based on contourlet transform. *Inverse Probl. Sci. Eng.* **2010**, *18*, 737–758. [CrossRef]
6. Liang, D.; Liu, B.; Wang, J.; Ying, L. Accelerating SENSE using compressed sensing. *Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med.* **2009**, *62*, 1574–1584. [CrossRef]
7. Liu, Y.; Zhan, Z.; Cai, J.-F.; Guo, D.; Chen, Z.; Qu, X. Projected iterative soft-thresholding algorithm for tight frames in compressed sensing magnetic resonance imaging. *IEEE Trans. Med. Imaging* **2016**, *35*, 2130–2140. [CrossRef]
8. Bilgic, B.; Goyal, V.K.; Adalsteinsson, E. Multi-contrast reconstruction with Bayesian compressed sensing. *Magn. Reson. Med.* **2011**, *66*, 1601–1615. [CrossRef]
9. Chen, E.Z.; Wang, P.; Chen, X.; Chen, T.; Sun, S. Pyramid convolutional RNN for MRI image reconstruction. *IEEE Trans. Med. Imaging* **2022**, *41*, 2033–2047. [CrossRef]
10. Yang, Y.; Sun, J.; Li, H.; Xu, Z. ADMM-CSNet: A deep learning approach for image compressive sensing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 521–538. [CrossRef]
11. Lustig, M.; Donoho, D.L.; Santos, J.M.; Pauly, J.M. Compressed sensing MRI. *IEEE Signal Process. Mag.* **2008**, *25*, 72–82. [CrossRef]
12. Li, B.; Hu, W.; Feng, C.-M.; Li, Y.; Liu, Z.; Xu, Y. Multi-Contrast Complementary Learning for Accelerated MR Imaging. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1436–1447. [CrossRef] [PubMed]
13. Hammernik, K.; Klatzer, T.; Kobler, E.; Recht, M.P.; Sodickson, D.K.; Pock, T.; Knoll, F. Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* **2018**, *79*, 3055–3071. [CrossRef] [PubMed]
14. Zhou, B.; Zhou, S.K. DuDoRNet: Learning a dual-domain recurrent network for fast MRI reconstruction with deep T1 prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4273–4282.
15. Zhao, X.; Yang, T.; Li, B.; Zhang, X. SwinGAN: A dual-domain Swin Transformer-based generative adversarial network for MRI reconstruction. *Comput. Biol. Med.* **2023**, *153*, 106513. [CrossRef] [PubMed]
16. Korkmaz, Y.; Dar, S.U.; Yurt, M.; Özbey, M.; Cukur, T. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. *IEEE Trans. Med. Imaging* **2022**, *41*, 1747–1763. [CrossRef]
17. Falvo, A.; Comminiello, D.; Scardapane, S.; Scarpiniti, M.; Uncini, A. A multimodal dense u-net for accelerating multiple sclerosis mri. In Proceedings of the 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, PA, USA, 13–16 October 2019; pp. 1–6.
18. Sun, L.; Fan, Z.; Fu, X.; Huang, Y.; Ding, X.; Paisley, J. A deep information sharing network for multi-contrast compressed sensing MRI reconstruction. *IEEE Trans. Image Process.* **2019**, *28*, 6141–6153. [CrossRef]
19. Zeng, K.; Zheng, H.; Cai, C.; Yang, Y.; Zhang, K.; Chen, Z. Simultaneous single-and multi-contrast super-resolution for brain MRI images based on a convolutional neural network. *Comput. Biol. Med.* **2018**, *99*, 133–141. [CrossRef]
20. Ma, Q.; Lai, Z.; Wang, Z.; Qiu, Y.; Zhang, H.; Qu, X. MRI reconstruction with enhanced self-similarity using graph convolutional network. *BMC Med. Imaging* **2024**, *24*, 113. [CrossRef]
21. Peng, C.; Lin, W.-A.; Chellappa, R.; Zhou, S.K. Towards multi-sequence MR image recovery from undersampled k-space data. In Proceedings of the Medical Imaging with Deep Learning, Montreal, QC, Canada, 6–8 July 2020; pp. 614–623.
22. Kim, K.H.; Do, W.J.; Park, S.H. Improving resolution of MR images with an adversarial network incorporating images with different contrast. *Med. Phys.* **2018**, *45*, 3120–3131. [CrossRef]
23. Lyu, Q.; Shan, H.; Steber, C.; Helis, C.; Whitlow, C.; Chan, M.; Wang, G. Multi-contrast super-resolution MRI through a progressive network. *IEEE Trans. Med. Imaging* **2020**, *39*, 2738–2749. [CrossRef]
24. Dhengre, N.; Sinha, S. Multiscale U-net-based accelerated magnetic resonance imaging reconstruction. *Signal Image Video Process.* **2022**, *16*, 881–888. [CrossRef]
25. Sriram, A.; Zbontar, J.; Murrell, T.; Zitnick, C.L.; Defazio, A.; Sodickson, D.K. GrappaNet: Combining parallel imaging with deep learning for multi-coil MRI reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14315–14322.
26. Li, G.; Lyu, J.; Wang, C.; Dou, Q.; Qin, J. Wavtrans: Synergizing wavelet and cross-attention transformer for multi-contrast mri super-resolution. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; pp. 463–473.
27. Lyu, J.; Sui, B.; Wang, C.; Tian, Y.; Dou, Q.; Qin, J. Dudocaf: Dual-domain cross-attention fusion with recurrent transformer for fast multi-contrast mr imaging. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; pp. 474–484.
28. Feng, C.-M.; Yan, Y.; Chen, G.; Xu, Y.; Hu, Y.; Shao, L.; Fu, H. Multimodal transformer for accelerated MR imaging. *IEEE Trans. Med. Imaging* **2022**, *42*, 2804–2816. [CrossRef] [PubMed]
29. Wang, J.; Yang, Y.; Yang, H.; Lian, C.; Xu, Z.; Sun, J. MD-GraphFormer: A Model-Driven Graph Transformer for Fast Multi-Contrast MR Imaging. *IEEE Trans. Comput. Imaging* **2023**, *9*, 1018–1030. [CrossRef]
30. Feng, C.-M.; Yan, Y.; Yu, K.; Xu, Y.; Fu, H.; Yang, J.; Shao, L. Exploring separable attention for multi-contrast MR image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 12251–12262. [CrossRef] [PubMed]
31. Feng, C.-M.; Yan, Y.; Fu, H.; Chen, L.; Xu, Y. Task transformer network for joint MRI reconstruction and super-resolution. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; pp. 307–317.

32. Li, G.; Lv, J.; Tian, Y.; Dou, Q.; Wang, C.; Xu, C.; Qin, J. Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast MRI super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20636–20645.
33. Xiang, L.; Chen, Y.; Chang, W.; Zhan, Y.; Lin, W.; Wang, Q.; Shen, D. Deep-learning-based multi-modal fusion for fast MR reconstruction. *IEEE Trans. Biomed. Eng.* **2018**, *66*, 2105–2114. [CrossRef]
34. Dar, S.U.; Yurt, M.; Shahdloo, M.; Ildız, M.E.; Tınaz, B.; Çukur, T. Prior-guided image reconstruction for accelerated multi-contrast MRI via generative adversarial networks. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1072–1087. [CrossRef]
35. Otazo, R.; Candes, E.; Sodickson, D.K. Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components. *Magn. Reson. Med.* **2015**, *73*, 1125–1136. [CrossRef]
36. Bustin, A.; Voilliot, D.; Menini, A.; Felblinger, J.; de Chillou, C.; Burschka, D.; Bonnemains, L.; Odille, F. Isotropic reconstruction of MR images using 3D patch-based self-similarity learning. *IEEE Trans. Med. Imaging* **2018**, *37*, 1932–1942. [CrossRef]
37. Deshmane, A.; Gulani, V.; Griswold, M.A.; Seiberlich, N. Parallel MR imaging. *J. Magn. Reson. Imaging* **2012**, *36*, 55–72. [CrossRef]
38. Huang, Y.; Paisley, J.; Lin, Q.; Ding, X.; Fu, X.; Zhang, X.-P. Bayesian nonparametric dictionary learning for compressed sensing MRI. *IEEE Trans. Image Process.* **2014**, *23*, 5007–5019. [CrossRef]
39. Wang, S.; Su, Z.; Ying, L.; Peng, X.; Zhu, S.; Liang, F.; Feng, D.; Liang, D. Accelerating magnetic resonance imaging via deep learning. In Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, 13–16 April 2016; pp. 514–517.
40. Schlemper, J.; Caballero, J.; Hajnal, J.V.; Price, A.; Rueckert, D. A deep cascade of convolutional neural networks for MR image reconstruction. In Proceedings of the Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, 25–30 June 2017; pp. 647–658.
41. Mardani, M.; Gong, E.; Cheng, J.Y.; Vasanawala, S.S.; Zaharchuk, G.; Xing, L.; Pauly, J.M. Deep generative adversarial neural networks for compressive sensing MRI. *IEEE Trans. Med. Imaging* **2018**, *38*, 167–179. [CrossRef]
42. Aggarwal, H.K.; Mani, M.P.; Jacob, M. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Trans. Med. Imaging* **2018**, *38*, 394–405. [CrossRef] [PubMed]
43. Zeng, W.; Peng, J.; Wang, S.; Liu, Q. A comparative study of CNN-based super-resolution methods in MRI reconstruction and its beyond. *Signal Process. Image Commun.* **2020**, *81*, 115701. [CrossRef]
44. Aggarwal, H.K.; Jacob, M. J-MoDL: Joint model-based deep learning for optimized sampling and reconstruction. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1151–1162. [CrossRef]
45. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
46. Huang, J.; Fang, Y.; Wu, Y.; Wu, H.; Gao, Z.; Li, Y.; Del Ser, J.; Xia, J.; Yang, G. Swin transformer for fast MRI. *Neurocomputing* **2022**, *493*, 281–304. [CrossRef]
47. Aghabiglou, A.; Eksioglu, E.M. Projection-Based cascaded U-Net model for MR image reconstruction. *Comput. Methods Programs Biomed.* **2021**, *207*, 106151. [CrossRef]
48. Luo, Y.; Wei, M.; Li, S.; Ling, J.; Xie, G.; Yao, S. An effective co-support guided analysis model for multi-contrast MRI reconstruction. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2477–2488. [CrossRef]
49. Feng, C.-M.; Fu, H.; Yuan, S.; Xu, Y. Multi-contrast MRI super-resolution via a multi-stage integration network. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; pp. 140–149.
50. Güngör, A.; Dar, S.U.; Öztürk, Ş.; Korkmaz, Y.; Bedel, H.A.; Elmas, G.; Ozbey, M.; Çukur, T. Adaptive diffusion priors for accelerated MRI reconstruction. *Med. Image Anal.* **2023**, *88*, 102872. [CrossRef]
51. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
52. Qiu, Y.; Zhang, H.; Ma, Q.; Yang, G.; Lai, Z. Multi-contrast MRI reconstruction based on frequency domain separation and cross-self-attention. *IEEE Access* **2024**, *12*, 55062–55076. [CrossRef]
53. Liu, X.; Wang, J.; Lin, S.; Crozier, S.; Liu, F. Optimizing multicontrast MRI reconstruction with shareable feature aggregation and selection. *NMR Biomed.* **2021**, *34*, e4540. [CrossRef]
54. Zbontar, J.; Knoll, F.; Sriram, A.; Murrell, T.; Huang, Z.; Muckley, M.J.; Defazio, A.; Stern, R.; Johnson, P.; Bruno, M. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv* **2018**, arXiv:1811.08839.
55. Knoll, F.; Zbontar, J.; Sriram, A.; Muckley, M.J.; Bruno, M.; Defazio, A.; Parente, M.; Geras, K.J.; Katsnelson, J.; Chandarana, H. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiol. Artif. Intell.* **2020**, *2*, e190007. [CrossRef] [PubMed]
56. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv* **2020**, arXiv:2006.04768.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.