Boosting ViT-based MRI Reconstruction from the Perspectives of Frequency Modulation, Spatial Purification, and Scale Diversification

Yucong Meng[1 2*], **Zhiwei Yang**[1 2 3*], **Yonghong Shi**[1 2†], **Zhijian Song**[1 2†]

[1]Digital Medical Research Center, School of Basic Medical Science, Fudan University, Shanghai 200032, China
[2]Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, Shanghai 200032, China
[3]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China
{ycmeng21, zwyang21} @m.fudan.edu.cn, {yonghong.shi, zjsong} @fudan.edu.cn

Abstract

The accelerated MRI reconstruction process presents a challenging ill-posed inverse problem due to the extensive undersampling in k-space. Recently, Vision Transformers (ViTs) have become the mainstream for this task, demonstrating substantial performance improvements. However, there are still three significant issues remain unaddressed: (1) ViTs struggle to capture high-frequency components of images, limiting their ability to detect local textures and edge information, thereby impeding MRI restoration; (2) Previous methods calculate multi-head self-attention (MSA) among both related and unrelated tokens in content, introducing noise and significantly increasing computational burden; (3) The naive feed-forward network in ViTs cannot model the multi-scale information that is important for image restoration. In this paper, we propose FPS-Former, a powerful ViT-based framework, to address these issues from the perspectives of frequency modulation, spatial purification, and scale diversification. Specifically, for issue (1), we introduce a frequency modulation attention module to enhance the self-attention map by adaptively re-calibrating the frequency information in a Laplacian pyramid. For issue (2), we customize a spatial purification attention module to capture interactions among closely related tokens, thereby reducing redundant or irrelevant feature representations. For issue (3), we propose an efficient feed-forward network based on a hybrid-scale fusion strategy. Comprehensive experiments conducted on three public datasets show that our FPS-Former outperforms state-of-the-art methods while requiring lower computational costs.

Introduction

Magnetic Resonance Imaging (MRI) is an essential tool in clinical diagnostics due to its non-radiation, high resolution, and superior contrast (Yang et al. 2022; Chen et al. 2022). However, its long scan times often increase the physical burden on patients. Additionally, involuntary movements like breathing, swallowing, and heartbeats usually blur images, limiting MRI's application. Reducing k-space acquisition can speed up MRI while introduce artifacts (Zeng et al. 2021). Eliminating these artifacts and reconstructing high-quality MRI images remains a significant challenge.

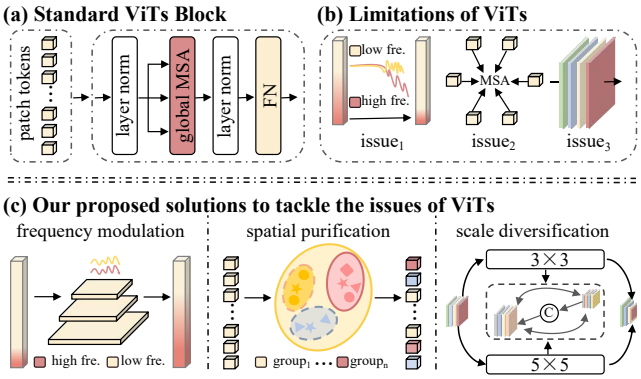

Figure 1: Our main idea. (a) The pipeline of standard ViTs block. (b) ViTs suffer limitations of high-frequency attenuation, irrelevant token interactions, and a lack of multi-scale feature representation. (3) We propose to tackle the above issues from the perspectives of frequency modulation, spatial purification, and scale diversification, thereby enhancing the performance of ViT-based MRI reconstruction.

In recent years, many methods have adopted various CNN architectures for MRI reconstruction (Zeng et al. 2020; Aghabiglou 2021; Aghabiglou and Eksioglu 2021). Due to the powerful non-linearity and feature representation capabilities of CNNs, CNN-MRI outperforms traditional compressed sensing (CS) based methods (Tamir et al. 2016). However, the convolutional operation has intrinsic characteristics such as local receptive fields and independence of input content (Li et al. 2021b; Zheng et al. 2022). Therefore, CNN-based models cannot eliminate long-range degradation perturbation and gain suboptimal MRI reconstruction performance (Zhou and Zhou 2020; Knoll et al. 2020).

To alleviate such limitations, Vision Transformers (ViTs) (Dosovitskiy et al. 2020) have been applied, shedding new light on MRI reconstruction tasks (Huang et al. 2022; Guo et al. 2024). As shown in Figure 1 (a), ViTs stack Multi-head Self-Attention (MSA) blocks, treating each image patch as a semantic token and modeling their interactions globally (Han et al. 2023). Unlike CNNs, which hierarchically enlarge the receptive field from local to global, even a shallow ViT can effectively capture global contexts, resulting in

*These authors contributed equally.
†Corresponding author.

highly competitive performance for various computer vision tasks (Ali et al. 2023; Yang et al. 2024b; Zhao et al. 2023).

However, ViTs still struggle to restore MRI details, facing several critical issues as shown in Figure 1 (b): (1) ViTs are limited in capturing high-frequency information, impairing their ability to detect local textures and edges essential for effective MRI reconstruction. As demonstrated in (Park and Kim 2022; Wang et al. 2022), the MSA inherently amounts to a low-pass filter, which indicates that ViTs will overlook high-frequency information crucial for image restoration when it scales up its depth. (2) Standard ViTs calculate MSA among both related and unrelated tokens, introducing noise and increasing computational burden. Previous ViT-based methods linearly project all patch tokens into query, key, and value, and then perform matrix multiplication for MSA (Zhou et al. 2023; Shen et al. 2024). However, some patches in MRI images are not related in content. Handling all tokens simultaneously introduces content-irrelevant noise and significantly increases computational complexity. (3) The multi-scale representation provides complementary information and plays a vital role in MRI reconstruction, while MSA in standard ViTs fails to effectively model multi-scale features (Chen, Fan, and Panda 2021; Cai et al. 2023).

By addressing the aforementioned issues from the perspectives of Frequency modulation, spatial Purification, and Scale diversification, we propose FPS-Former, a powerful ViT-based framework that significantly enhances the performance of MRI reconstruction, as shown in Figure 1 (c). Specifically, for issue (1), we propose the Frequency Modulation Attention Module (FMAM). FMAM recalibrates features in a Laplacian pyramid, enabling the retrieval of high-frequency information. This approach suppresses the low-pass filtering characteristic of ViTs, allowing the retention of more high-frequency details, which is beneficial for restoring local textures and edges. For issue (2), we design the Spatial Purification Attention Module (SPAM). Instead of processing all projected tokens simultaneously as standard ViTs, SPAM clusters tokens into different groups by identifying similar elements that yield the maximum inner product. Tokens within each group are considered closely related in content. The MSA operation is then applied within each group, reducing the noise impact of content-irrelevant tokens and significantly lowering computational complexity. For issue (3), we introduce a Scale Diversification Feed-forward Network (SDFN) that explores multi-scale feature representation by inserting two multi-scale deep convolution paths during feature transmission. Finally, observing that undersampled MRI images exhibit various types and degrees of degradation artifacts, we incorporate Hybrid Experts Feature Refinement (HEFR) into our model. HEFR comprises several convolutional layers and provides collaborative refinement for MRI reconstruction.

The main contributions of our work are listed as follows:

- We propose the Frequency Modulation Attention Module to enhance the self-attention map by recalibrating frequency information in a Laplacian pyramid, selectively strengthening the contributions of shape and texture features, thereby overcoming the low-pass filtering of ViTs.
- We introduce the Spatial Purification Attention Module to capture interactions among closely related tokens, thereby reducing redundant or irrelevant feature representations for precise self-similarity capturing.
- We propose the Scale Diversification Feed-forward Network to effectively model multi-scale information.
- Extensive experiments on both single-coil and multi-coil datasets under various undersampling patterns show that our method outperforms state-of-the-art (SoTA) competitors while requiring lower computational costs.

Related Work

CNN-based MRI Reconstruction

MRI reconstruction techniques can enhance image quality with less dependency on physiology and hardware, making them more accessible for accelerated MRI. Recent advances in deep learning have spurred the development of CNN-based MRI reconstruction. CMRNet pioneered the application of deep learning in MRI reconstruction by creating an offline CNN to map zero-filled to fully-sampled MRI images (Wang et al. 2016). D5C5 proposed a CNN cascade for dynamic cardiac MRI (Schlemper et al. 2018). DuDoRNet incorporated T1 priors for simultaneous k-space and image restoration (Zhou and Zhou 2020). Dual-OctConv learned multi-scale spatial-frequency features from both real and imaginary components for parallel MRI (Feng et al. 2021). Despite these successes, CNNs exhibit a limited receptive field and struggle to model long-range dependencies. Therefore, CNNs are suboptimal for restoring various image regions and cannot achieve satisfactory reconstruction performance (Khan et al. 2020; Sarvamangala and Kulkarni 2022).

ViT-based MRI Reconstruction

Vision Transformers (ViTs) treat images as sequences of patches and use self-attention to capture global context (Yang et al. 2024a). Compared to CNNs, ViTs have advantages such as capturing global patterns and have been used for MRI reconstruction. As demonstrated in (Lin and Heckel 2022), a ViT tailored for image reconstruction can achieve performance comparable to U-net while providing higher throughput and reduced memory consumption. SLATER addressed unsupervised MRI reconstruction by using a cross-attention module to capture correlations between latent variables and image features (Korkmaz et al. 2022). SwinMR designed a parallel imaging coupled swin transformer-based model for fast CS-MRI (Huang et al. 2022). ReconFormer incorporated a local pyramid and global columnar ViT structure to learn multi-scale features at any stage, enabling enhanced reconstruction performance (Guo et al. 2024).

However, these methods still failed to achieve precise MRI reconstruction because they overlooked inherent issues of ViTs such as loss of high-frequency information, interference among unrelated patches, and the inability to model multi-scale features. By addressing these issues from three perspectives respectively, we propose FPS-Former to boost the performance of ViT-based MRI reconstruction.

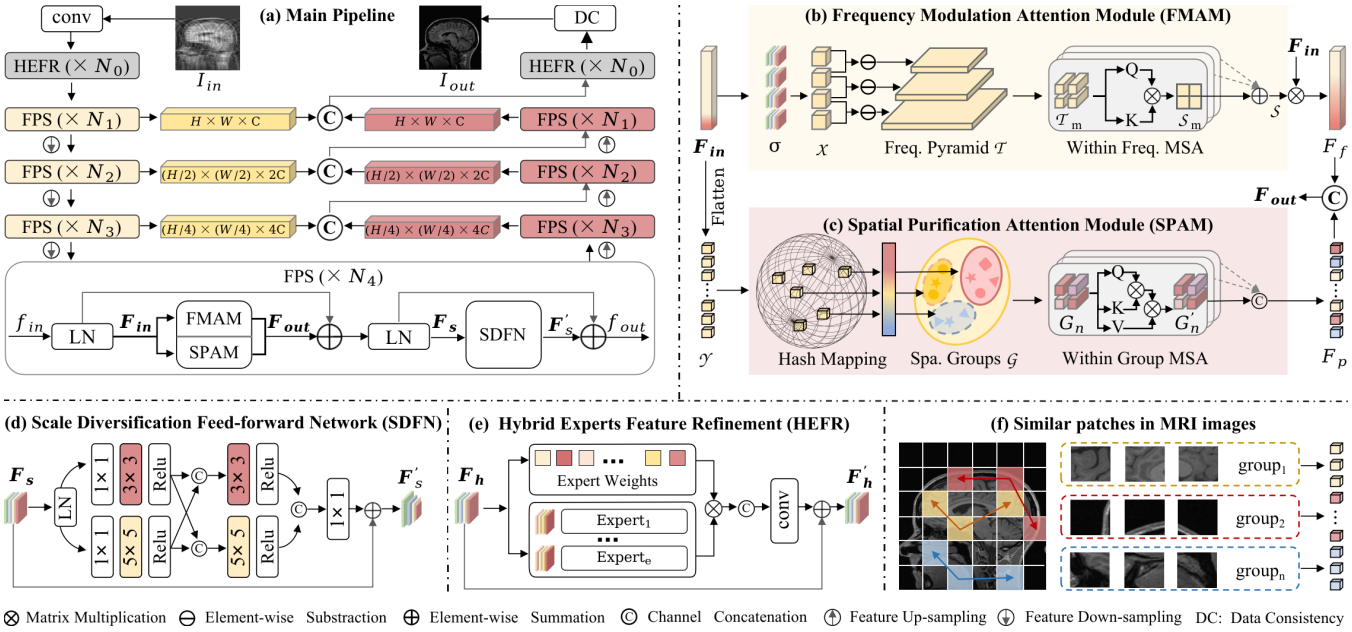

Figure 2: (a) The overall architecture of the proposed FPS-Former. Given an input image I_{in}, we first apply a 3×3 convolution to obtain patch tokens. In the network backbone, we stack multiple FPS blocks to extract hierarchical features. FPS, consisting of FMAM (b), SPAM (c), and SDFN (d), is designed to tackle the issues of ViT-based MRI reconstruction. Besides, at both early and final stages of the network, we design HEFR (e) to provide refined features, ensuring the reconstruction of high-quality output I_{out}. (f) The motivation of our SPAM. MRI images contain widely distributed, similar patches that appear in groups.

Methodology

Overall Pipeline

As shown in Figure 2 (a), our FPS-Former is a hierarchical encoder-decoder framework. Given a low-quality $I_{in} \in \mathbb{R}^{H \times W \times C}$ with spatial resolution $H \times W$ and channel dimension C, we first perform overlapped image patch embedding with a convolution layer. Next, the embedding results are sent to the designed backbone, which stacks N_0 HEFR blocks and $N_{i \in [1,2,3,4]}$ FPS blocks. HEFR block is introduced to provide fine-grained information, as shown in Figure 2 (e). The FPS block consists of the Frequency Modulation Attention Module (FMAM), Spatial Purification Attention Module (SPAM), and Scale Diversification Feed-forward Network (SDFN), as shown in Figure 2 (b), (c), and (d), respectively. It is designed to amend the mentioned issues of ViTs and extract hierarchical features with different spatial resolutions and channel dimensions. Then, the extracted features are sent to decoder, which also includes $N_{i \in [1,2,3,4]}$ FPS and N_0 HEFR blocks. Skip connections are adopted to hierarchically bridge intermediate features between the encoder and decoder. Finally, a Data Consistency (DC) layer is added to reconstruct high-quality I_{out}.

The above reconstruction process can be formulated as: $I_{out} = \mathcal{N}(I_{in})$, where $\mathcal{N}(\cdot)$ is the overall network and is trained by minimizing the following loss function:

$$\mathcal{L} = \left\| I_{out} - I_{gt} \right\|_1, \tag{1}$$

where I_{gt} denotes the ground-truth image, and $\|\cdot\|_1$ is the L_1-norm. The proposed FPS and HEFR blocks will be specifically introduced in the following sections.

FPS Block

ViTs in MRI reconstruction struggle with high-frequency loss, irrelevant interactions, and limited multi-scale representation. To tackle these, we propose FPS block consisting of three modules: Frequency Modulation Attention Module (FMAM), Spatial Purification Attention Module (SPAM), and Scale Diversification Feed-forward Network (SDFN). Given the input f_{in}, the produces of FPS is defined as:

$$f^{'} = f_{in} + \mathcal{F} \circledast \mathcal{P}(LN(f_{in})), f_{out} = f^{'} + \mathcal{S}(LN(f^{'})), \tag{2}$$

where $LN(\cdot)$ denotes the layer normalization, $\mathcal{F} \circledast \mathcal{P}$ represents the combined effect of FMAM and SPAM, and $\mathcal{S}(\cdot)$ denotes the operation of SDFN.

Frequency Modulation Attention Module Standard ViTs struggle with the loss of high-frequency details. To address this, we propose the Frequency Modulation Attention Module (FMAM) to recalibrate the importance of frequency at each level. Specifically, as shown in Figure 2 (b), we first use Gaussian functions with different variances to extract multiple Gaussian representations $\mathcal{X}$ as:

$$\mathcal{X} = \{\mathcal{X}_m\}_{m=1}^{M+1}, \mathcal{X}_m = F_{in} \circledast \frac{1}{\sigma_m \sqrt{2\pi}} e^{-\frac{i^2+j^2}{2\sigma_m^2}}, \tag{3}$$

where $F_{in} \in \mathbb{R}^{H \times W \times C}$ is normalized from f_{in}. (i, j) corresponds to the spatial location, $\sigma_{m \in [1,2,\ldots,M+1]}$ denotes the variance of the Gaussian function for the m-th scale, and $\circledast$ is the convolution operator. Then we construct the frequency pyramid $\mathcal{T}$ by subtracting adjacent elements in $\mathcal{X}$:

$$\mathcal{T} = \{\mathcal{T}_m\}_{m=1}^{M}, \mathcal{T}_m = \mathcal{X}_{m+1} - \mathcal{X}_m \tag{4}$$

The frequency pyramid $\mathcal{T}$ is composed of multiple layers, each containing distinct types of frequency information. To achieve a balanced distribution of low and high-frequency components within the model, we conduct Within Frequency MSA operation and effectively aggregate features from each frequency level. Specifically, we first calculate the attention scores $\mathcal{S}$ for each level of $\mathcal{T}$ as follows:

$$\mathcal{S} = \{\mathcal{S}_m\}_{m=1}^{M}, \mathcal{S}_m = \sum_{i=1}^{I} \mathrm{softmax}((Q_m^i K_m^i)/\sqrt{d}), \tag{5}$$

where I is the number of attention heads, Q_m and K_m are derived from $\mathcal{T}_m$ using linear transformations, and $d = (C/I)$ denotes the dimension of each head. Finally, we sum the attention scores in $\mathcal{S}$ and multiply the result by the Value (V, derived from F_{in}) to obtain the result of FMAM F_f:

$$F_f = (\sum_{m=1}^{M}(\mathcal{S}_m \in \mathcal{S}))V. \tag{6}$$

Spatial Purification Attention Module As shown in Figure 2 (f), MRI images contain clusters of image patches that are similar within each group but distinctly different from those outside the group. Previous ViT-based methods perform a dense MSA operation on all patch tokens simultaneously. This operation leads to noisy interactions among unrelated features, hampering MRI reconstruction. To address this, we propose the Spatial Purification Attention Module (SPAM), which applies a sparsity constraint by computing self-attention only between contextually related tokens to reduce noise and computational complexity.

Specifically, as shown in Figure 2 (c), given the input feature map F_{in}, we first flatten it into $\mathcal{Y} = \left\{f_j \in \mathbb{R}^C\right\}_{j=1}^{J}$, where J represents the number of tokens. Subsequently, we use the hash function to aggregate the information and map the C-dimensional tokens f_j into integer hash codes $\mathcal{Z}$. This hash mapping can be formulated as:

$$\mathcal{Z} = \{\mathcal{Z}_j \in \mathbb{Z}\}_{j=1}^{J}, \mathcal{Z}_j = \lfloor (a \cdot f_j + b)/r \rfloor, \tag{7}$$

where $a \in \mathbb{R}^C$ and $b \in \mathbb{R}$ are random variables satisfying $a \sim \mathcal{N}(0,1)$ and $b \sim \mathcal{U}(0,r)$, $r \in \mathbb{R}$ is a constant, $\lfloor \cdot \rfloor$ is the floor function. Next, we sort all elements in $\mathcal{Y}$ based on their hash code in $\mathcal{Z}$. The $j\text{-}th$ sorted element is denoted as $f_j^{'}$. Then we split them into groups $\mathcal{G}$, which is expressed as:

$$\mathcal{G} = \{\mathcal{G}_n\}_{n=1}^{N}, \mathcal{G}_n = \left\{f_j^{'} : ng+1 \le j \le (n+1)g\right\}, \tag{8}$$

where N denotes the number of groups, and each group has g elements. With such a scheme, closely related tokens are grouped together. Subsequently, we apply the Within Group MSA operation for each G_n to obtain updated groups $\mathcal{G}^{'}$:

$$\mathcal{G}^{'} = \left\{G_n^{'}\right\}_{n=1}^{N}, G_n^{'} = \sum_{i=1}^{I} W_i head_i\left(G_n\right), \tag{9}$$

where $head_i(\cdot)$ represents the self-attention operation of the $i\text{-}th$ head, I is the number of attention heads, and $W_i \in R^{C\times d}$ represents the learnable parameters.

Next, we take out all the elements from each $G_n^{'}$ and unsort them according to their original positions in $\mathcal{Y}$. We then concatenate the elements to obtain the purified features F_p.

Finally, we concatenate $[\cdot]$ the purified features F_p with the frequency result F_f from FMAM. A depthwise convolution $f(\cdot)$ is further applied to aggregate the information. In this way, we retain high-frequency information and achieve spatial purification. The above process is formulated as:

$$F_{out} = f([F_p, F_f]). \tag{10}$$

Scale Diversification Feed-forward Network Multi-scale representations have been proven effective in enhancing MRI reconstruction. However, previous methods often focus on integrating single-scale components into feed-forward networks, overlooking the importance of multi-scale feature representations. To address this, we design a Scale Diversification Feed-forward Network (SDFN) by inserting two multi-scale depth-wise convolution paths in the transmission process as shown in Figure 2 (d). Specifically, given an input F_s, which is normalized from the above aggregated F_{out}, we first expand its channel dimension with 1×1 convolution in the ratio of r. Then, the obtained feature is sent into two parallel branches. During feature transformation, we use 3×3 and 5×5 depthwise convolutions to enhance multi-scale local information extraction. The entire feature fusion process of SDFN can be described as follows:

$$\begin{aligned}
&\hat{F_s} = f_{1\times1}(LN(F_s),\\
&F_{p_1} = \sigma(f_{3\times3}(\hat{F_s})), F_{s_1} = \sigma(f_{5\times5}(\hat{F_s})),\\
&F_{p_2} = \sigma(f_{3\times3}\left[F_{p_1}, F_{s_1}\right]), F_{s_2} = \sigma(f_{5\times5}\left[F_{s_1}, F_{p_1}\right]),\\
&F_s^{'} = f_{1\times1}\left[F_{p_2}, F_{s_2}\right] + F_s,
\end{aligned} \tag{11}$$

where $f_{1\times1}(\cdot)$ denotes the 1×1 convolution, $\sigma(\cdot)$ is a ReLU activation, $f_{3\times3}(\cdot)$ and $f_{5\times5}(\cdot)$ denote 3×3 and 5×5 depthwise convolutions, and $[\cdot]$ is the channel-wise concatenation.

Hybrid Experts Feature Refinement

Inspired by (Chen et al. 2023), we introduce Hybrid Experts Feature Refinement (HEFR) to provide fine-grained information, as shown in Figure 2 (e). Specifically, we extract fine-grained knowledge by carefully selecting multiple CNN operations, referred to as experts. These include average pooling, separable convolution layers, and dilated convolution layers with different kernel sizes. Unlike the traditional approach of combining experts with an external gating network, we employ a self-attention mechanism as a switcher among different experts, adaptively emphasizing the importance of various feature representations based on the input. Specifically, given the input feature $F_h \in \mathbb{R}^{H\times W\times C}$, we first apply the channel-wise average to generate a C-dimensional channel descriptor $\mathcal{K} \in \mathbb{R}^C$:

$$\mathcal{K} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_h(i,j), \tag{12}$$

where $F_h(i,j)$ is the value of feature F_h at spatial location (i,j). Then, the coefficient vector $\mathcal{V}$ of each expert is allocated corresponding to the learnable weight matrices $W_1 \in \mathbb{R}^{D\times C}$ and $W_2 \in \mathbb{R}^{E\times D}$, i.e., $\mathcal{V} = W_2\sigma(W_1\mathcal{K})$. Here, D is the dimension of the weight matrix, E is the number of experts, and $\sigma(\cdot)$ is a ReLU function. Finally, denoting the expert operations as $f_{exp}(\cdot)$, the output $F_h^{'}$ is obtained as:

$$F_h^{'} = f_{1\times1}(\sum_{e=1}^{E} f_{exp}(F_h, \mathcal{V})) + F_h. \tag{13}$$

Method	Type	CC359						fastMRI					
		NMSE ↓		SSIM ↑		PSNR ↑		NMSE↓		SSIM ↑		PSNR ↑	
		AF=4	AF=8	AF=4	AF=8	AF=4	AF=8	AF=4	AF=8	AF=4	AF=8	AF=4	AF=8
CS (Tamir et al. 2016)		0.0483	0.1066	0.7510	0.6424	26.34	22.85	0.0583	0.0903	0.5736	0.4870	29.54	26.99
KIKI-Net (Eo et al. 2018)	$\mathcal{C}$	0.0221	0.0417	0.8415	0.7773	28.97	26.24	0.0353	0.0546	0.7172	0.6355	31.87	29.27
UNet-32 (Zbontar et al. 2018)	$\mathcal{C}$	0.0197	0.0385	0.8898	0.8348	31.54	28.66	0.0337	0.0477	0.7248	0.6570	31.99	30.02
D5C5 (Schlemper et al. 2018)	$\mathcal{C}$	0.0177	0.0428	0.8977	0.8267	31.59	28.20	0.0332	0.0512	0.7256	0.6457	32.25	29.65
DCRCN (Aghabiglou 2021)	$\mathcal{C}$	0.0119	0.0291	0.9100	0.8649	32.01	29.49	0.0351	0.0443	0.7332	0.6635	32.18	30.76
VIT_Base (Lin and Heckel 2022)	$\mathcal{T}$	0.0207	0.0446	0.8903	0.8254	31.33	28.03	0.0342	0.0460	0.7206	0.6578	32.10	30.28
SwinMR (Huang et al. 2022)	$\mathcal{T}$	0.0109	0.0260	0.9298	0.8695	34.14	30.36	0.0342	0.0476	0.7213	0.6537	32.14	30.21
ReconFormer (Guo et al. 2024)	$\mathcal{T}$	0.0108	0.0276	0.9297	0.8650	34.16	30.11	0.0320	0.0431	0.7327	0.6672	**32.53**	30.76
Restormer (Zamir et al. 2022)	$\mathcal{T}$	0.0164	0.0367	0.9093	0.8445	32.36	28.86	0.0339	0.0450	0.7223	0.6597	32.20	30.46
AST (Zhou et al. 2024)	$\mathcal{T}$	0.0149	0.0322	0.9115	0.8544	32.78	29.45	0.0335	0.0445	0.7234	0.6620	32.26	30.52
Ours		**0.0103**	**0.0217**	**0.9321**	**0.8828**	**34.38**	**31.15**	**0.0316**	**0.0408**	**0.7337**	**0.6692**	32.51	**31.03**

Table 1: Performance comparison of MRI reconstruction under $4\times$ and $8\times$ Acceleration Factor (AF) on the single-coil datasets, including CC359 and fastMRI. $\mathcal{C}$: CNN-based methods. $\mathcal{T}$: transformer-based methods.

Method	SKM-TEA					
	NMSE ↓		SSIM ↑		PSNR ↑	
	AF=4	AF=8	AF=4	AF=8	AF=4	AF=8
KIKI-Net	0.0196	0.0271	0.8577	0.7941	34.26	31.42
UNet-32	0.0204	0.0270	0.8469	0.7904	33.91	31.44
D5C5	0.0188	0.0257	0.8648	0.8030	34.63	31.89
SwinMR	0.0192	0.0256	0.8597	0.8022	34.45	31.94
ReconFormer	0.0179	0.0239	0.8730	0.8158	35.06	32.51
AST	0.0172	0.0297	0.8914	0.8407	35.17	32.61
Ours	**0.0158**	**0.0200**	**0.8975**	**0.8527**	**35.64**	**32.85**

Table 2: Performance comparison under $4\times$ and $8\times$ Acceleration Factor (AF) on the multi-coil SKM-TEA dataset.

Experiments

Experimental Settings

Datasets The proposed FPS-Former is evaluated on three datasets: CC359 (Warfield, Zou, and Wells 2004), fastMRI (Zbontar et al. 2018), and SKM-TEA (Desai et al. 2022). The CC359 dataset is a raw brain MRI dataset acquired from clinical MR scanners (Discovery MR750; GE Healthcare, Waukesha, WI, USA). Following the official dataset split, we randomly selected a training set comprising $4,524$ slices from 25 subjects, and a test set consisting of $1,700$ slices from an additional 10 subjects. The acquisition matrix size is 256×256; The fastMRI dataset contains $1,172$ complex-valued single-coil coronal proton density (PD)-weighted knee MRI scans. We partition this dataset into 973 scans for training and 199 scans (fastMRI validation dataset) for testing. The acquisition matrix size is 320×320; The SKM-TEA raw data provides 155 complex-valued multi-coil T2-weighted knee MRI scans. $124, 10$, and 21 coil-combined volumes are used for training, validation, and testing. Each subject provides approximately 160 cross-sectional knee images with the matrix of size 512×512. In comparison experiments, the input images are generated by randomly undersampling the k-space data using the 1D cartesian function similar to the fastMRI challenge (Zbontar et al. 2018).

Training Details In our model, $(N_0, N_1, N_2, N_3, N_4)$ are set to $(4, 1, 2, 2, 1)$, and the number of attention heads for (N_1, N_2, N_3, N_4) FPS blocks are set to $(1, 2, 4, 8)$. For each FPS block, the number of frequency pyramid levels M in FMAM, the number of groups N in SHAM, and the channel expansion factor r in SDFN are set to $(3, 4, 2)$, respectively. For the HEFR module, we set the number of experts E to 8 and the dimension of the weight matrix D to 32. During training, we used the AdamW optimizer with a batch size of 4 and patch size of 8, for a total of $300K$ iterations. The initial learning rate is fixed at 1×10^{-4} for the first $92K$ iterations, then reduced to 1×10^{-6} using a cosine annealing schedule over the remaining $208K$ iterations. The entire framework is implemented on PyTorch using RTX 3090.

Comparison with State-of-the-arts

Single-coil datasets We compared the proposed FPS-Former with recent MRI reconstruction approaches, including CNN-based and Transformer-based methods. Additionally, we evaluated it against two state-of-the-art natural image restoration methods, Restormer (Zamir et al. 2022) and AST (Zhou et al. 2024), which were equipped with a DC layer with the same settings of MRI reconstruction for a fair comparison. Table 1 shows the comparison results of our FPS-Former with other methods under different acceleration factors (AF) on single-coil datasets, including CC359 and fastMRI. As shown in this table, FPS-Former demonstrates significant improvements over CNN-based methods and consistently surpasses other Transformer-based approaches across different acceleration rates on both datasets. For example, our method shows the superiority of 2.37 dB over the CNN-based SoTA DCRCN and 0.22 dB over Transformer-based counterpart ReconFormer under $4\times$ AF on CC359. Notably, our approach shows greater performance improvement as the acceleration factor increases, particularly in more challenging scenarios. Specifically, for the CC359 and fastMRI datasets, our model outperforms the leading method AST, by 1.70 dB and 0.51 dB at $8\times$ AF, and 1.60 dB and 0.25 dB at $4\times$ AF, respectively.

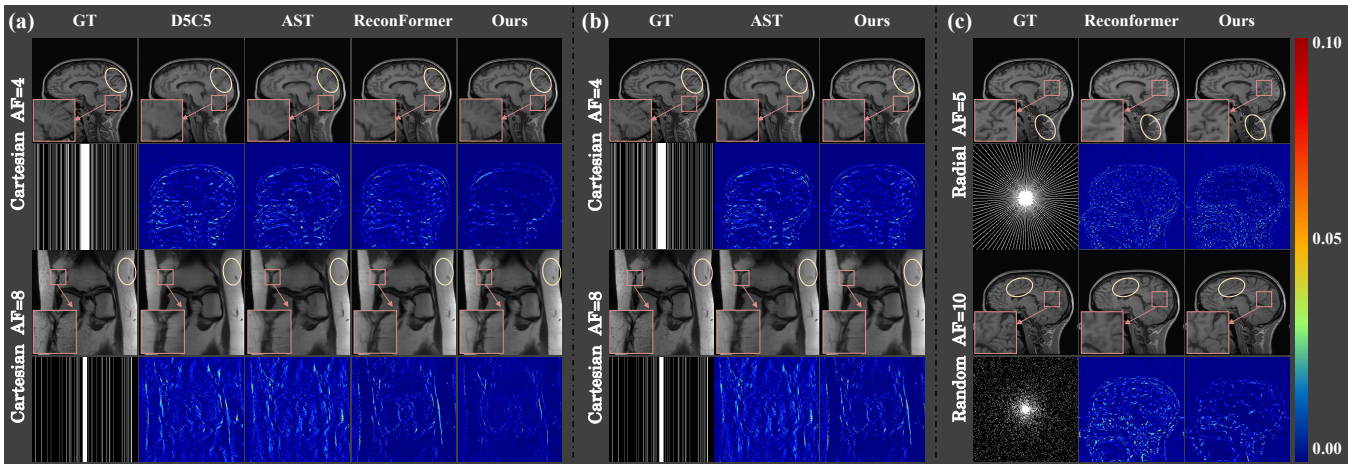

Figure 3: Qualitative comparison of different methods on (a) the single-coil dataset including CC359 and fastMRI, (b) the multi-coil dataset SKM-TEA, and (c) the CC359 dataset using different undersampling masks. The second row of each subplot shows the corresponding error maps. The red boxes and yellow ellipses highlight the details in the reconstruction results.

Method	Mask	NMSE ↓		SSIM ↑		PSNR ↑	
		AF=5	AF=10	AF=5	AF=10	AF=5	AF=10
ReconFormer	$\mathcal{I}$	0.0070	0.0170	0.9450	0.8971	36.18	32.17
AST		0.0068	0.0173	0.9448	0.8988	36.12	32.10
Ours		**0.0060**	**0.0163**	**0.9467**	**0.9017**	**36.40**	**32.36**
ReconFormer	$\mathcal{R}$	0.0125	0.0176	0.9164	0.8918	33.52	32.02
AST		0.0127	0.0178	0.9142	0.8924	33.48	31.98
Ours		**0.0119**	**0.0173**	**0.9170**	**0.8943**	**33.74**	**32.11**

Table 3: Performance comparison of MRI reconstruction under $5\times$ and $10\times$ Acceleration Factors (AF) on the CC359 using more masks. $\mathcal{I}$: Radial mask. $\mathcal{R}$: Random mask.

Model	FMAM	SPAM	SDFN	HEFR	NMSE	SSIM	PSNR
(a)		✓	✓	✓	0.0109	0.9294	34.14
(b)	✓		✓	✓	0.0117	0.9246	33.83
(c)	✓	✓		✓	0.0114	0.9260	33.95
(d)	✓	✓	✓		0.0113	0.9274	33.98
Ours	✓	✓	✓	✓	**0.0103**	**0.9321**	**34.38**

Table 4: Ablation results on FPS-Former on CC359 (AF=4).

Multi-coil datasets Table 2 gives comparison results of MRI reconstruction on the multi-coil SKM-TEA dataset. We achieved 35.64 and 32.85 PSNR under $4\times$ and $8\times$ AF respectively. Our FPS-Former significantly outperforms previous CNN-based solutions and shows the superiority of 0.47 dB and 0.24 dB over AST at $4\times$ AF and $8\times$ AF, respectively. This further demonstrates the superiority of our method.

Experiments on different masks To further demonstrate the robustness of our FPS-Former, we conducted experiments using radial and random undersampling patterns under $5\times$ and $10\times$ acceleration factors on CC359 dataset. As shown in Table 3, FPS-Former consistently outperforms other methods, highlighting its ability to effectively reconstruct MRI images from various undersampling masks.

Visualization Results The qualitative results for the single-coil dataset, multi-coil dataset, and mask experiments are shown in Figure 3 (a), (b), and (c), respectively. In (a), the CNN-based SoTA, D5C5, suffers from severe edge blurring and substantial detail loss. Although the Transformer-based AST and ReconFormer partially alleviate these issues, they still lose crucial anatomical details in challenging tasks with high acceleration factors. In contrast, our FPS-Former demonstrates robustness to various anatomical structures and acceleration factors. By addressing the issues of ViT models, our method better preserves important anatomical details, as highlighted by the zoomed boxes and ellipses. In (b), our FPS-Former can restore more abundant details than other counterparts. This further demonstrates that our method can effectively reconstruct not only single-coil but also multi-coil MRI images. In (c), our method demonstrates great robustness across various undersampling patterns and acceleration rates, further validating its effectiveness.

Ablation Studies and Analysis

Efficacy of Key Components We performed a breakdown ablation to investigate the effect of each component and their interactions. As the results in Table 4 show, (a) When FMAM is removed, the performance dramatically degrades by 0.24 in PSNR and 0.0027 in SSIM. This can be attributed to FMAM's ability to preserve high-frequency details, which are crucial for restoring textures and edges. (b) Excluding SPAM leads to reductions in PSNR and SSIM by 0.55 and 0.0075, respectively. This demonstrates that SPAM effectively mitigates the noise impact from irrelevant tokens, thereby enhancing performance. (c) Replacing SDFN with a conventional feed-forward network in the standard ViT results in a decrease in PSNR from 34.38 to 33.95. This highlights SDFN's effectiveness in representing multi-scale features, which is essential for MRI reconstruction. (d) The absence of HEFR results in a decline of 0.40 in PSNR and 0.0047 in SSIM, showing its significant contribution.

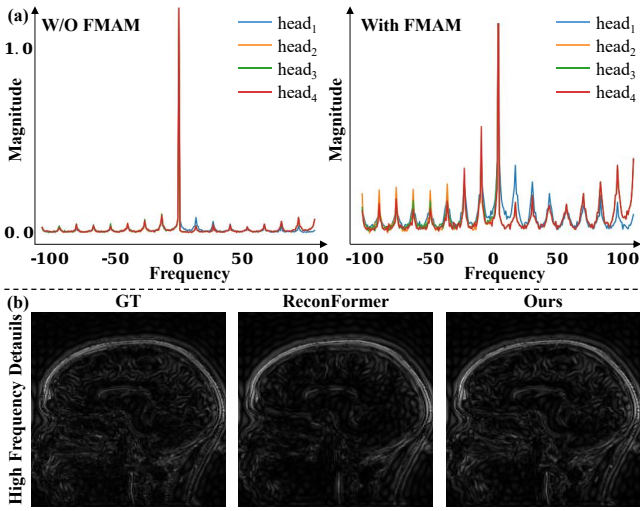

Figure 4: (a) Frequency response analysis. (b) Visualization results of high-frequency details in reconstructed images.

Model	Com.	NMSE	SSIM	PSNR
2 layered pyramid	$\mathcal{F}$	0.0104	0.9312	34.30
4 layered pyramid		0.0103	0.9320	34.36
Global-MSA (Alexey 2020)	$\mathcal{P}$	0.0117	0.9246	33.83
Window-MSA (Liu et al. 2021)		0.0115	0.9261	33.89
FN (Alexey 2020)	$\mathcal{S}$	0.0114	0.9260	33.95
DFN (Li et al. 2021a)		0.0111	0.9277	34.05
GDFN (Zamir et al. 2022)		0.0117	0.9247	33.82
Ours (3 layered+SPAM+SDFN)		**0.0103**	**0.9321**	**34.38**

Table 5: Ablation study for variants of FMAM ($\mathcal{F}$), SPAM ($\mathcal{P}$), and SDFN ($\mathcal{S}$) on CC359 under $4\times$ acceleration factor.

Method	FLOPs	Param.	SSIM	PSNR
ReconFormer (Guo et al. 2024)	342G	1.14M	0.9297	34.16
AST (Zhou et al. 2024)	155G	26.10M	0.9115	32.78
Ours	152G	12.51M	**0.9321**	**34.38**

Table 6: Efficiency comparison of FPS-Former and others.

Analysis of FPS block To further analyze the effectiveness of our FPS block, we compared various variants for FMAM, SPAM, and SDFN. As shown in Table 5: (1) we investigated the impact of the number of pyramid layers M in FMAM. Our results indicate that both too small and large M negatively affect performance, with the optimal reconstruction results achieved when $M = 3$. Fewer layers struggle to capture diverse frequency, while excessive layers lead to confusion in feature aggregation. (2) We compared our SPAM with several self-attention mechanisms, including global MSA and local window-based MSA. SPAM shows the most significant improvement because it performs MSA calculations among tokens with closely related content, effectively reducing noisy interactions. (3) We evaluated SDFN against three methods: the conventional feed-forward network (FN), the depth-wise convolution feed-forward network (DFN), and the gated-depth-wise convolution feed-forward network (GDFN). Although GDFN employs a gating mechanism to enhance performance, it does not leverage multi-scale feature integration. Our SDFN incorporates local feature extraction and fusion across different scales, achieving a PSNR gain of 0.56 dB over GDFN.

Efficiency of Frequency Modulation Attention Module To further validate the effectiveness of our FMAM, we follow (Wang et al. 2022) and present a spectral response comparison between network variants with and without FMAM at the last encoder layer, as shown in Figure 4 (a). The frequency response of the network without FMAM exhibits greater attenuation of high frequency compared to FPS-Former. Additionally, we extracted high-frequency structures from the reconstructed images of different methods using a high-pass filter, as illustrated in Figure 4 (b). Due to FMAM's effective preservation of high-frequency details, our method reconstructs edges and textures more accurately and completely. This visual evidence underscores FMAM's superior ability to tackle ViT's low-pass filter issues.

Analysis of Training Efficiency The training efficiency comparison is reported in Table 6. The recent ViT-based ReconFormer employs a recurrent structure to maintain a few trainable parameters. However, its significantly higher computational complexity (FLOPs=342G) substantially increases both training difficulty and inference time. On the other hand, AST addresses both spatial and channel redundancy and achieves a notable reduction in FLOPs. Nevertheless, AST has a larger parameter count of 26.10 M and a noticeable drop in performance. Compared to these methods, our FPS-Former achieves significant performance improvements while maintaining both low computational complexity and a minimal number of trainable parameters.

Analysis of Hyper-parameters The analysis of key hyper-parameters, such as the number of experts in HEFR, the number of frequency pyramid levels in FMAM, the expansion ratio in SDFN, and the number of FPS blocks and HEFR modules, etc., is specifically discussed in the *Supplementary Materials*. FPS-Former demonstrates consistent performance across different hyper-parameter variations.

Conclusion

In this work, we propose to boost ViT-based MRI reconstruction by tackling three issues, including loss of high-frequency information, redundancy interactions among irrelevant tokens, and challenges in multi-scale feature modeling. To achieve this, we propose the frequency modulation attention module for frequency information correction, the spatial purification attention module for grouped token interactions, and the scale diversification feed-forward network for multi-scale feature transmission, respectively. Extensive experiments and analysis are conducted on CC359, fastMRI, and SKM-TEA datasets, validating the efficiency of FPS-Former in tackling the issues of ViT-MRI and significantly improving the performance of MRI reconstruction.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.82372097.

References

Aghabiglou, A. 2021. MR image reconstruction using densely connected residual convolutional networks. *Computers in Biology and Medicine*, 139: 105010.

Aghabiglou, A.; and Eksioglu, E. M. 2021. Projection-Based cascaded U-Net model for MR image reconstruction. *Computer Methods and Programs in Biomedicine*, 207: 106151.

Alexey, D. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Ali, A. M.; Benjdira, B.; Koubaa, A.; El-Shafai, W.; Khan, Z.; and Boulila, W. 2023. Vision transformers in image restoration: A survey. *Sensors*, 23(5): 2385.

Cai, H.; Li, J.; Hu, M.; Gan, C.; and Han, S. 2023. EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 17302–17313.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 357–366.

Chen, X.; Pan, J.; Lu, J.; Fan, Z.; and Li, H. 2023. Hybrid cnn-transformer feature fusion for single image deraining. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 378–386.

Chen, Y.; Schönlieb, C.-B.; Liò, P.; Leiner, T.; Dragotti, P. L.; Wang, G.; Rueckert, D.; Firmin, D.; and Yang, G. 2022. AI-based reconstruction for fast MRI—A systematic review and meta-analysis. *Proceedings of the IEEE*, 110(2): 224–245.

Desai, A. D.; Schmidt, A. M.; Rubin, E. B.; Sandino, C. M.; Black, M. S.; Mazzoli, V.; Stevens, K. J.; Boutin, R.; Ré, C.; Gold, G. E.; et al. 2022. Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. *arXiv preprint arXiv:2203.06823*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eo, T.; Jun, Y.; Kim, T.; Jang, J.; Lee, H.-J.; and Hwang, D. 2018. KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic resonance in medicine*, 80(5): 2188–2201.

Feng, C.-M.; Yang, Z.; Chen, G.; Xu, Y.; and Shao, L. 2021. Dual-octave convolution for accelerated parallel MR image reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 116–124.

Guo, P.; Mei, Y.; Zhou, J.; Jiang, S.; and Patel, V. M. 2024. ReconFormer: Accelerated MRI Reconstruction Using Recurrent Transformer. *IEEE Transactions on Medical Imaging*, 43(1): 582–593.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; Yang, Z.; Zhang, Y.; and Tao, D. 2023. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 87–110.

Huang, J.; Fang, Y.; Wu, Y.; Wu, H.; Gao, Z.; Li, Y.; Del Ser, J.; Xia, J.; and Yang, G. 2022. Swin transformer for fast MRI. *Neurocomputing*, 493: 281–304.

Khan, A.; Sohail, A.; Zahoora, U.; and Qureshi, A. S. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53: 5455–5516.

Knoll, F.; Hammernik, K.; Zhang, C.; Moeller, S.; Pock, T.; Sodickson, D. K.; and Akcakaya, M. 2020. Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues. *IEEE signal processing magazine*, 37(1): 128–140.

Korkmaz, Y.; Dar, S. U.; Yurt, M.; Özbey, M.; and Cukur, T. 2022. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. *IEEE Transactions on Medical Imaging*, 41(7): 1747–1763.

Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; and Van Gool, L. 2021a. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*.

Li, Z.; Liu, F.; Yang, W.; Peng, S.; and Zhou, J. 2021b. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12): 6999–7019.

Lin, K.; and Heckel, R. 2022. Vision transformers enable fast and robust accelerated MRI. In *International Conference on Medical Imaging with Deep Learning*, 774–795. PMLR.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Park, N.; and Kim, S. 2022. How do vision transformers work? *arXiv preprint arXiv:2202.06709*.

Sarvamangala, D.; and Kulkarni, R. V. 2022. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, 15(1): 1–22.

Schlemper, J.; Caballero, J.; Hajnal, J. V.; Price, A. N.; and Rueckert, D. 2018. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE transactions on Medical Imaging*, 37(2): 491–503.

Shen, G.; Li, M.; Anderson, S.; Farris, C. W.; and Zhang, X. 2024. Magnetic Resonance Image Processing Transformer for General Reconstruction. arXiv:2405.15098.

Tamir, J. I.; Ong, F.; Cheng, J. Y.; Uecker, M.; and Lustig, M. 2016. Generalized magnetic resonance image reconstruction using the Berkeley advanced reconstruction toolbox. In

ISMRM Workshop on Data Sampling & Image Reconstruction, Sedona, AZ, volume 7, 8.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*.

Wang, S.; Su, Z.; Ying, L.; Peng, X.; Zhu, S.; Liang, F.; Feng, D.; and Liang, D. 2016. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, 514–517. IEEE.

Warfield, S. K.; Zou, K. H.; and Wells, W. M. 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7): 903–921.

Yang, G.; Zhang, L.; Zhou, M.; Liu, A.; Chen, X.; Xiong, Z.; and Wu, F. 2022. Model-guided multi-contrast deep unfolding network for mri super-resolution reconstruction. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3974–3982.

Yang, Z.; Fu, K.; Duan, M.; Qu, L.; Wang, S.; and Song, Z. 2024a. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3606–3615.

Yang, Z.; Meng, Y.; Fu, K.; Wang, S.; and Song, Z. 2024b. Tackling Ambiguity from Perspective of Uncertainty Inference and Affinity Diversification for Weakly Supervised Semantic Segmentation. *ArXiv*, abs/2404.08195.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zbontar, J.; Knoll, F.; Sriram, A.; Murrell, T.; Huang, Z.; Muckley, M. J.; Defazio, A.; Stern, R.; Johnson, P.; Bruno, M.; et al. 2018. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*.

Zeng, G.; Guo, Y.; Zhan, J.; Wang, Z.; Lai, Z.; Du, X.; Qu, X.; and Guo, D. 2021. A review on deep learning MRI reconstruction without fully sampled k-space. *BMC Medical Imaging*, 21(1): 195.

Zeng, W.; Peng, J.; Wang, S.; and Liu, Q. 2020. A comparative study of CNN-based super-resolution methods in MRI reconstruction and its beyond. *Signal Processing: Image Communication*, 81: 115701.

Zhao, H.; Gou, Y.; Li, B.; Peng, D.; Lv, J.; and Peng, X. 2023. Comprehensive and Delicate: An Efficient Transformer for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14122–14132.

Zheng, M.; Xu, J.; Shen, Y.; Tian, C.; Li, J.; Fei, L.; Zong, M.; and Liu, X. 2022. Attention-based CNNs for image classification: A survey. In *Journal of Physics: Conference Series*, volume 2171, 012068. IOP Publishing.

Zhou, B.; Dey, N.; Schlemper, J.; Salehi, S. S. M.; Liu, C.; Duncan, J. S.; and Sofka, M. 2023. DSFormer: A Dual-Domain Self-Supervised Transformer for Accelerated Multi-Contrast MRI Reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4966–4975.

Zhou, B.; and Zhou, S. K. 2020. 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*, 4273–4282.

Zhou, S.; Chen, D.; Pan, J.; Shi, J.; and Yang, J. 2024. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2952–2963.