Deep Attention Encoder and Convolutional Fusion Decoder for MR Image Reconstruction

Abstract

The scarcity of a single-coil fast MRI dataset has resulted in a noticeable absence of transformer-based approaches in the reconstruction field. Despite the rising popularity of transformer research, its integration into MRI reconstruction predominantly favors multi-coil datasets, leaving single-coil data untapped. To address this challenge, we introduce a dedicated hybrid reconstruction model that combines hierarchical attention layers based on Swin transformers in the encoder with an efficient attention-free decoder employing ConvFuUnit (Convolutional Fusion Unit) & CConvFuUnit (Complex Convolutional Fusion Unit). Our model scored 0.1dB & 0.3dB higher PSNR values than Unet and Swin Unet on the Proton Density (PD) dataset and slightly higher scores than all models for Proton Density with Fat Saturation (PDFS) and PD+PDFS datasets. We also present Bi-data Inference, training models on one dataset class and testing on both (PD+PDFS). Remarkably, our hybrid architecture successfully overcomes data limitations, achieving superior reconstruction results with 0.1dB & 0.15dB higher than Unet and SwinUnet on PD Bi-data Inference.This improvement is attributed to the fact that convolutions emphasize local features, while transformers capture long-range dependencies, thereby enhancing global context and visual coherence.Our research paves the way for future advancements, allowing the integration of the latest transformer trends into the field of single coil reconstruction.

1. Introduction

MRI reconstruction, specifically accelerating MRI, is important because it significantly reduces the acquisition time required to obtain high-quality images. Image reconstruction is considered an ill-posed problem in computer vision because it lacks a unique, stable solution. There are typically more unknowns (pixel values) than available information (observations or measurements). This leads to ambiguity and sensitivity to noise, making it challenging to find an exact solution.Datasets are readily available to researchers for both single-coil and multi-coil reconstructions in various anatomical regions within the medical field. In our work, we specifically focused on the largest fastMRI single coil knee dataset provided by Facebook.

The Total Variation (TV) minimization algorithm is a well-known classical compressed sensing (CS) based iterative image reconstruction algorithm, renowned for its accurate image reconstruction. However, recent years have witnessed the emergence of deep learning-based methods that have surpassed traditional reconstruction algorithms. Many deep learning algorithms proposed for single-coil reconstruction tasks rely on Convolutional Neural Networks (CNNs). Notable examples include the Baseline FastMRI UNet and XPD-Net, which are trained on the image domain.

Moreover, researchers have also directed their attention towards reconstruction methods that leverage dual-domain information, encompassing both image data and K-space data. Prominent deep learning models in this category include KIKI Net, MD-Recon-Net, KV-Net, and DIIKnet, which exploit diverse types of information while still employing a CNN as the underlying architecture.

However, it is worth noting that all the aforementioned deep learning approaches are based on the CNN backbone. The main reason for the limited adoption of transformers in these methods is the scarcity of single-coil datasets, whereas transformers typically require a larger amount of data. In the case of multicoil reconstruction, most transformer-based models have outperformed the baseline UNet model and other CNN-based approaches.

While CNN models excel in understanding global information and learning features from MRI scans, they struggle with capturing long-range dependencies and relationships between different anatomical regions. These relationships are crucial for accurate reconstructions. Another observation is that CNN-based designs tend to over-refine the reconstructed images. Transformers, on the other hand, can address this issue but cannot grasp global information as effectively as CNNs. To tackle these challenges, we contributed following.

- We have developed a hybrid model that combines a swin transformer as the encoder and a CNN-based design as the decoder, aiming to leverage the strengths of

both architectures and address their limitations. This hybrid approach offers the advantage of working effectively with limited data. The encoder focuses on capturing the relationships between anatomical regions and passes this information to the decoder. On the other hand, the decoder, being predominantly convolutional, excels at understanding global information and performing accurate image reconstruction.

- Utilizing a conventional convolutional setup in the decoder is not an ideal approach for the hybrid design for the image reconstruction task. Therefore, we have implemented two novel architectures called Convolutional Fusion Unit(ConvFuUnit) & Complex Convolutional Fusion Unit(CConvFuUnit) that integrates the information obtained from the swin transformer-based encoder and applies upscaling techniques.

Our proposed hybrid model not only addresses these discrepancies but also produces visual superior image quality compared to the baseline UNet model even with slightly higher PSNR & SSIM values.

2. Related Work

In the first part, we check out the latest research trends for the single-coil dataset. Then, in the next part, we talk about why vision transformers are important for single coil reconstructions. We use research from multicoil cases to show this. Since there haven't been any notable designs for single coils due to not enough data, this section helps us see why transformer models are really important for dealing with single coil.

2.1. Single-Coil FastMRI Reconstruction

The FastMRI Unet serves as a foundational model for both single and multicoil data image reconstruction tasks. Research within this field primarily centers around two crucial domains: model design and the integration of cross-domain information. For example, MD-Reconstruction achieves image reconstruction by utilizing derived dual-domain latent information from both K-space and spatial data. Another notable method, the KIKI net, comprises two distinct networks KCNN and ICNN tailored for K-space and Image space, respectively. These networks engage in interleaved data consistency operations, contributing to the production of high-quality reconstructed images.

Moreover, a dual-domain reconstruction network known as KV-net has been introduced. This network combines the K-net, responsible for K-space data, with the V-net, designed for image data. The final reconstructed image is achieved through the propagation of undersampled k-space data throughout the entire network. Another innovative technique gaining popularity is the Invertible Recurrent Inference Machine, or i-RIM. It's worth noting that the XPD-Net models also find their foundation in convolutional neural networks, primarily trained on single-coil data.

Despite these advancements, the use of deep convolutional networks oriented toward either the image domain or frequency domain is limited in simultaneously harnessing spatial and K-space features. The methods listed above involve pooling operations that result in information loss. To address this limitation, researchers have proposed a novel architecture capable of working with full-resolution cross-domain features.

However, a drawback persists: the aforementioned models primarily rely on convolutional designs, which lack a comprehensive understanding of the relationships between anatomical regions necessary for accurate reconstructions.

2.2. Vision Transformers for FastMRI Multicoil MRI Reconstruction

3. Method

Our primary objective is to devise a hybrid model that combines a transformer as an encoder with a convolutional approach as a decoder. This fusion aims to leverage the distinctive capabilities of each while mitigating their respective limitations, ultimately enhancing the quality of visual reconstruction. However, a direct strategy involving the use of transformer features with simple scaling for reconstruction through conventional convolutional architectures is suboptimal. To address this, we introduce two innovative architectures: the 'Convolutional Fusion Unit(ConvFuUnit)' and the 'Complex Convolutional Fusion Unit(CConvFuUnit)'. We will elaborate on the design pipeline and then delve into the details of **(a)** the Convolutional Fusion Unit(ConvFuUnit) and **(b)** the Complex Convolutional Fusion Unit(CConvFuUnit).

Overall Pipeline. The input, an undersampled MR Image $I \in \mathbb{R}^{H \times W \times 1}$, undergoes a process involving a 4-layer Swin Transformer-based Encoder followed by a Decoder. Initially, the image is divided into patches, creating $Io \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 16}$, which are then projected as embeddings. Attention is subsequently applied using the Swin Transformers.

Our approach commences with a patch splitting module, inspired by methods such as Swin and Vision Transformer (ViT). This module effectively partitions the input image into non-overlapping patches, each treated as a discrete 'token', with feature representation derived from concatenating raw pixel values. We utilize a patch size of 4×4, similar to the Swin Transformer, resulting in a feature dimension of 16 for each patch ($4 \times 4 \times 1$). A linear embedding layer is then applied to these raw-valued features, projecting them into an arbitrary dimension denoted as 'C'.

Self-attention is computed in four stages for patch tokens of sizes $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$ for stages 1, 2, 3, and 4, respectively. The bottleneck incorporates the Convolutional Fusion Unit(ConvFuUnit), while the decoder is optimized with both the Convolutional Fusion Unit(ConvFuUnit) and the Complex Convolutional Fusion Unit(CConvFuUnit) alternately, aiming to maintain a limited number of parameters. Skip connections are established from the encoder to the decoder to mitigate information loss. In our quest to enhance image quality, we have employed initial convolutional layers on the original image, utilizing this information to inform the decoder.

3.1. ConvFuUnit-Convolutional Fusion Unit

This block consists of two parallel sub-networks: Image Quality Adjustment (IQA) and Channel Information Reading (CDR). In the IQA sub-network, a series of operations are performed, including initial 3×3 depthwise separable convolutions, followed by InstanceNorm (IN) normalization and the application of a LeakyReLU activation function. Subsequently, a regular 3×3 convolution is employed, along with another InstanceNorm (IN) normalization step. On the contrasting side, the CDR sub-network employs a 1×1 pointwise convolution, followed by InstanceNorm (IN) normalization. Finally, the outcomes of both IQA and CDR sub-networks are fused and passed through the LeakyReLU activation function. Collectively, these processes enhance image quality and extract vital channel information for subsequent stages of the network. The purpose of QDA is swin transformer features have a large varience in the distribution along the channel wise, depth wise convolutions will recieve that information individually and adjust them into to a desired distribution and then allowed to pass through regular convolutions.since the regular convolutions can see only the modified information it cant see the variations along the channel wise so we have implimented a CDR to read and reduce the information.This CDR also helps in keeping the information to the later stages with minimal loss.

3.2. CConvFuUnit-Complex Convolutional Fusion Unit

The design of the Complex Convolutional Fusion Unit(CConvFuUnit) involves a sequence of four straightforward convblocks. Each of these stages acquires compacted data from the root, as depicted in the illustration, utilizing pointwise convolutions. Swin Transformers typically operate with a significant number of channels, such as 96, 128, and 144, for various sizes like small, large, and big Swin Transformers. Rather than an all-at-once channel reduction in the decoder, the strategy here is to perform channel reduction across the four stages. This approach contributes to the reconstruction of top-notch information, resulting in enhanced information quality.

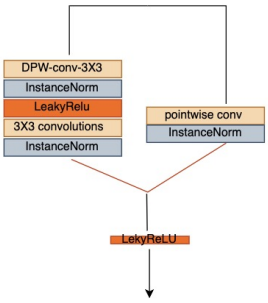

Figure 1. ConvFuUnit-Convolutional Fusion Unit

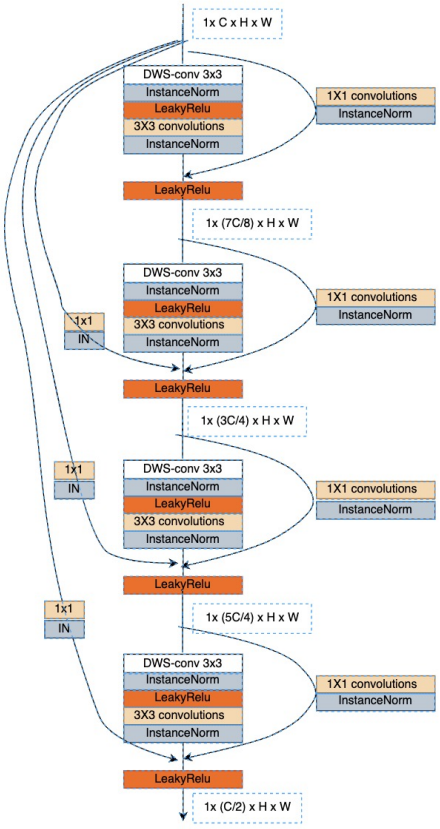

Figure 2. CConvFuUnit-Complex Convolutional Fusion Unit

4. Experiments and Analysis

In this section, we begin by delving into the details of our experimental setup. Subsequently, we proceed to assess the superiority of the reconstructed images in comparison to alternative methods.We conducted our experiments on the Facebook FastMRI single-coil open dataset. The dataset comprises both Proton Density (PD) and Proton Density Fat-Suppressed (PDFS) variants. In this study, we present our reconstruction results for these variants individually, as well as for the combined data $PD + PDFS$ and Bi-data inference for the PD and PDFS data.

4.1. Experimental Setup

Experimental Setup: We conducted our model training over a fifteen epochs. For the initial 12 epochs, a learning rate of 0.001 was employed, followed by a reduced rate of 0.0001 for the remaining epochs. The AdmW optimizer was utilized,with a weight decay of 0.1, while the optimization process was guided by the L1 loss function. Our data preprocessing adhered to the standards set by Facebook's fastMRI framework. The training of our models was facilitated by NVIDIA A6000 and NVIDIA A100 GPUs.

Evaluation Metrics: In the evaluation phase, we employed a set of commonly accepted metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Normalized Mean Squared Error (NMSE). These metrics collectively provide a comprehensive assessment of the quality of the reconstructed images. Ensuring consistency with Facebook's FastMRI guidelines, we consider these three metrics sufficient for conducting a robust comparative analysis of our results.