MRI IMAGE RECONSTRUCTION USING VISION TRANSFORMER

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

Master of Technology

IN

Signal Processing

BY

Prabhat Kumar Jharia

SR. No.: 04-03-01-10-51-23-1-22745

Under the guidance of

Prof. Muthuvel Arigovindan

Department of Electrical Engineering
Indian Institute of Science
Bangalore – 560 012 (INDIA)

June, 2025

Declaration of Originality

I, **Prabhat Kumar Jharia**, with SR No. **04-03-01-10-51-23-1-22745** hereby declare that the material presented in the thesis titled

MRI IMAGE RECONSTRUCTION USING VISION TRANSFORMER

represents original work carried out by me in the **Department of Electrical Engineering** at **Indian Institute of Science** during the years **2023-2025**.

With my signature, I certify that:

- I have not manipulated any of the data or results.
- I hereby declare that I have used ChatGPT and Gemini towards generation of refining the language and presentation. My estimate of the percentage contribution of the ChatGPT and Gemini towards this report is 8%-10%.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

. Prabhat Kumar Jharia

Date: 17 June 2025 Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Prof. Muthuvel Arigovindan Advisor Signature

Acknowledgements

I would like to express my deepest gratitude to MTech Project advisor Prof. Muthuvel Arigovindan for giving me this project and for his support and guidance during this work.

I would also like to thank Prof. Chandra Sekhar Seelamantula for his guidance, especially in the project's medical image aspect.

Special thanks to my lab mates, Raksham and Piyush Sharma, for the interesting technical talks that greatly improved my knowledge of our study topic. I would also like to genuinely thank my M.Tech friends and classmates for their constant support, collaborative spirit, and the shared academic experiences that made this trip truly worthwhile.

Finally, I offer my heartfelt thanks to my parents, family, and friends for their unwavering support, patience, and drive throughout this effort.

Abstract

The Vision Transformer (ViT) has demonstrated remarkable accuracy and computational efficiency in image classification when trained on extensive datasets. Motivated by these advantages, this paper explores the application of ViTs to accelerated magnetic resonance imaging (MRI) reconstruction. We show that a ViT specifically adapted for image reconstruction, when trained on the fastMRI dataset (comprising thousands of images), achieves reconstruction accuracy comparable to the U-net, a strong convolutional neural network baseline, while offering superior throughput and reduced memory consumption.

Recognizing that Transformers excel with large-scale pre-training and that MRI data acquisition is costly, we propose a novel pre-training strategy utilizing readily available natural image datasets like ImageNet. Our findings indicate that pre-training drastically enhances the data efficiency of the Vision Transformer for accelerated MRI, significantly improving performance, especially in low-data regimes where only a limited number of MRI training images are available. In such cases, the pre-trained ViT's surpasses both pre-trained CNN's and current state-of-the-art methods in image quality. Furthermore, pre-trained ViT's exhibit increased robustness to anatomical shifts. This work presents the potential of Vision Transformers for fast and robust accelerated MRI, particularly in data-scarce environments.

Contents

Acknowledgements **i**

Abstract **ii**

Contents **iii**

List of Figures **v**

List of Tables **vi**

Introduction **1**

1.1 Background and Motivation . . . 2
1.2 Problem Statement . . . 3
1.3 Contribution . . . 3
1.3.1 Rotary Positional Embeddings (RoPE) . . . 3
1.3.2 Multi-Scale Feature Aggregation . . . 4
1.3.3 Attention Mechanisms . . . 5
1.3.4 Computational Complexity Analysis . . . 6
1.3.4.1 Attention Complexity . . . 6
1.3.4.2 Memory Efficiency . . . 6
1.3.5 Position Encoding Comparison . . . 6
1.3.6 Feature Extraction Enhancement . . . 7
1.3.7 Training Stability Analysis . . . 7
1.3.8 Loss Function Formulation . . . 7
1.3.8.1 Individual Loss Components . . . 8
1.3.8.2 Mathematical Interactions . . . 8

Literature Survey **11**

2.1 Classical and Analytical Reconstruction . . . 11

2.2 Traditional Compressed Sensing Approaches . . . 13
2.3 Deep Learning: CNN-based Methods . . . 14
2.4 Vision Transformers for MRI Reconstruction . . . 14
2.4.1 Multi-Head Self-Attention (MHSA) . . . 14

Problem Formulation and Methodology 16
3.1 Forward Model of 2D MRI . . . 16
3.2 Vision Transformer-Based Reconstruction . . . 17
3.3 Training and Loss Function . . . 20

Experimental Validation 21
4.1 Dataset . . . 21
4.1.1 fastMRI Dataset . . . 21
4.1.2 ImageNet . . . 22
4.2 Experimental Setup . . . 23
4.3 Results . . . 24

Conclusion and Future Work 27
5.1 Conclusion . . . 27
5.2 Future Work . . . 29
5.2.1 Architectural Enhancements and Hybrid Models . . . 29
5.2.2 Advanced Pre-training and Self-supervised Learning Strategies . . . 30
5.2.3 Novel Data Consistency and Physics Integration . . . 30
5.2.4 Towards Clinical Translation and Robustness . . . 31

Bibliography 32

List of Figures

2.1 A series of 35 sagittal MRI slices, displaying the rich anatomical information obtained across different depths of a scanned region of knee. Each slice represents a 2D cross-section, collectively generating a 3D volume. 12

3.2 Block Diagram of the Modified Transformer Architecture 19

4.3 FastMRI Image Sample 21

4.4 Experiments Training Loss Comparision 24

4.5 Experiments Validation Loss Comparision 24

4.6 from left to right 1.undersampled image,2.output of our work SSIM:0.6739, 3.ground truth ,4.output of author work SSIM:0.6553 25

4.7 from left to right 1.undersampled image,2.output of our work SSIM:0.8407, 3.ground truth ,4.output of author work SSIM:0.8321 25

4.8 from left to right 1.undersampled image,2.output of our work SSIM:0.7591, 3.ground truth ,4.output of author work SSIM:0.7362 25

4.9 from left to right 1.undersampled image,2.output of our work SSIM:0.6723, 3.ground truth ,4.output of author work SSIM:0.6542 26

4.10 from left to right 1.undersampled image,2.output of our work SSIM:0.6477, 3.ground truth ,4.output of author work SSIM:0.6313 26

List of Tables

1.1 Computational complexity comparison where N is sequence length, d is embedding dimension, and s is stride. 6

1.2 Mapping of Loss Components to Their Primary Architectural Influence and Mathematical Property 9

4.3 Experimental Results for MRI Reconstruction 23

5.4 SSIM Comparison for 4-fold Accelerated Multi-Coil Knee MRI 28

Introduction

Magnetic Resonance Imaging (MRI) is a necessary tool in modern diagnostics, delivering detailed anatomical and functional insights without ionizing radiation. However, its long acquisition times hinder clinical efficiency, reducing patient throughput and increasing the the chance of motion abnormalities. Addressing these restrictions by speeding MRI scans without impacting picture quality remains a significant research focus.

Traditional methodologies, such as compressed sensing (CS), have used complex signal processing to reclaim images from undersampled data. While effective, they often suffer with reconstruction speed and generalizability. Deep learning approaches, especially Convolutional Neural Networks (CNNs) like U-net, have recently emerged as powerful alternatives, giving faster and more accurate reconstructions in many scenarios.

Vision Transformers (ViTs), originally created for natural language processing, have lately achieved state-of-the-art outcomes in computer vision by capturing long-range dependencies through self-attention. Their capacity to model global spatial linkages and their computation efficiency make them intriguing for medical picture reconstruction. This paper discusses the use of ViTs for expedited MRI, utilizing their strengths while addressing challenges triggered by the limited availability of large-scale medical imaging datasets.

1.1 Background and Motivation

Magnetic Resonance Imaging (MRI) is crucial for medical diagnosis, although its lengthy scan periods cause patient pain, motion aberrations, and limit clinical throughput. Accelerating MRI without losing image quality is a constant challenge. Traditional approaches and compressed sensing offer solutions, but deep learning, particularly Convolutional Neural Networks (CNNs) like the U-net, has revolutionized MRI reconstruction, exhibiting higher performance by learning complicated mappings from undersampled data.

Recently, Vision Transformers (ViTs) have emerged as powerful alternatives to CNNs in computer vision. Inspired by their success in image classification—especially with huge datasets—and their computational efficiency (better throughput, less memory), we study ViTs for faster MRI reconstruction. ViTs' self-attention mechanism allows them to capture global image dependencies, which could be particularly advantageous for complex reconstruction tasks.

A primary reason is addressing the limited availability of big MRI datasets, which often hinder the training of data-hungry ViTs. We suggest a novel pre-training strategy: utilizing enormous natural picture datasets (e.g., ImageNet) to train the ViT before fine-tuning it for MRI. This strategy intends to substantially enhance data efficiency, enabling ViTs to perform admirably even with minimal MRI training data. Furthermore, we evaluate if this pre-training boosts the ViT's resistance to anatomical differences. Our goal is to demonstrate that ViTs can offer accurate, faster, and more robust accelerated MRI, particularly in data-constrained clinical contexts.

1.2 Problem Statement

Accelerated Magnetic Resonance Imaging (MRI) is extremely important for improving patient comfort, avoiding motion artifacts, and enhancing clinical throughput by minimizing scan times. While deep learning methods, particularly Convolutional Neural Networks (CNNs), have significantly advanced image reconstruction from undersampled MRI data, they still face limitations in achieving optimal computational efficiency and robustness, especially when confronted with the inherent scarcity of large-scale medical imaging datasets required for extensive training. Vision Transformers (ViTs) offer a promising alternative given their higher performance and efficiency in general image processing tasks; yet, their data-intensive nature offers a considerable hurdle for direct use in data-constrained MRI situations. Therefore, the problem is to develop a highly efficient and robust deep learning framework for accelerated MRI reconstruction that can overcome the limitations of existing methods and effectively leverage the capabilities of Vision Transformers, particularly by addressing the challenge of limited MRI training data through innovative pre-training strategies, thereby enabling faster, higher-quality, and more reliable clinical imaging.

1.3 Contribution

The proposed architecture differentiates itself from the base Vision Transformer (ViT) model primarily through its advanced training objective and the potential integration of refined architectural components that enhance its ability to process and reconstruct intricate image details. Here's a breakdown of how proposed model is more advanced:

1.3.1 Rotary Positional Embeddings (RoPE)

The original model uses **learnable absolute positional embeddings**:

$$\mathbf{x}' = \mathbf{x} + \mathbf{PE}_{abs} \tag{1.1}$$

where $\mathbf{PE}_{abs} \in \mathbb{R}^{N \times d}$ is a learnable parameter matrix.

but we uses **Rotary Positional Embeddings (RoPE)** for relative position encoding:

$$\mathbf{q}_m = \mathbf{q} \odot \cos(m\boldsymbol{\theta}) - \mathbf{q} \odot \sin(m\boldsymbol{\theta}) \tag{1.2}$$

$$\mathbf{k}_n = \mathbf{k} \odot \cos(n\boldsymbol{\theta}) + \mathbf{k} \odot \sin(n\boldsymbol{\theta}) \tag{1.3}$$

where:

$$\boldsymbol{\theta} = \left\{ \frac{1}{10000^{2i/d}} : i \in [0, 1, \ldots, d/2] \right\} \tag{1.4}$$

Advantages:

- Preserves relative positional relationships: $\langle \mathbf{q}_m, \mathbf{k}_n \rangle = f(m-n)$
- Enables length extrapolation beyond training sequences
- Geometrically meaningful attention patterns

1.3.2 Multi-Scale Feature Aggregation

The original model uses Simple patch embedding with single-scale processing:

$$\mathbf{F} = \text{Conv2D}(\mathbf{X}, \mathbf{W}) \tag{1.5}$$

But we uses **Dilated Convolutions** with multiple dilation rates $r \in \{1, 2, 4\}$:

$$\mathbf{y}[i] = \sum_k \mathbf{x}[i + r \cdot k] \cdot \mathbf{w}[k] \tag{1.6}$$

The multi-scale features are aggregated as:

$$\mathbf{F}_{multi} = \text{Proj}\left(\text{Concat}\left[\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_4\right]\right) \tag{1.7}$$

Benefits:

- Receptive field expansion: $RF = k + (k-1)(r-1)$
- Multi-scale context capture
- Parameter efficiency through weight sharing

1.3.3 Attention Mechanisms

The original model uses **Gated Positional Self-Attention (GPSA)** combines patch-based and position-based attention:

$$\mathbf{A}_{patch} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \tag{1.8}$$

$$\mathbf{A}_{pos} = \text{softmax}\left(\text{PosProj}(\mathbf{R})\right) \tag{1.9}$$

$$\mathbf{A} = (1-\sigma(\mathbf{g})) \odot \mathbf{A}_{patch} + \sigma(\mathbf{g}) \odot \mathbf{A}_{pos} \tag{1.10}$$

But we uses **Cosine Attention** with L2 normalization:

$$\mathbf{Q}_{norm} = \frac{\mathbf{Q}}{\|\mathbf{Q}\|_2} \tag{1.11}$$

$$\mathbf{K}_{norm} = \frac{\mathbf{K}}{\|\mathbf{K}\|_2} \tag{1.12}$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}_{norm}\mathbf{K}_{norm}^T}{\tau}\right) \tag{1.13}$$

where $\tau = 0.1$ is the temperature parameter.

Advantages:

- Gradient stability through normalization
- Scale-invariant similarity computation
- Controlled attention sharpness via temperature scaling

1.3.4 Computational Complexity Analysis

1.3.4.1 Attention Complexity

Model	Original	With Downsampling
Original Model	$\mathcal{O}(N^2 d)$	-
Our Model	$\mathcal{O}(N^2 d)$	$\mathcal{O}\left(\left(\frac{N}{s}\right)^2 d\right)$

Table 1.1: Computational complexity comparison where N is sequence length, d is embedding dimension, and s is stride.

1.3.4.2 Memory Efficiency

Conv-based downsampling in Our model:

$$\mathbf{X}_{down} = \text{Conv2D}(\mathbf{X}, \text{kernel} = 3, \text{stride} = 2) \tag{1.14}$$

reduces memory footprint by factor of s^2.

1.3.5 Position Encoding Comparison

Original Model uses:

$$\mathbf{X}_{final} = \mathbf{X}_{patch} + \mathbf{PE}_{learnable} \tag{1.15}$$

Our Model uses:

$$\mathbf{X}_{pos} = \mathbf{X}_{patch} + \mathbf{PE}_{learnable} \quad \text{(Absolute)} \tag{1.16}$$

$$\mathbf{Q}, \mathbf{K} = \text{RoPE}(\mathbf{Q}, \mathbf{K}) \quad \text{(Relative)} \tag{1.17}$$

1.3.6 Feature Extraction Enhancement

Multi-Scale Dilated Convolution The dilated convolution operation for dilation rate r is:

$$(\mathbf{f} *_r \mathbf{g})[n] = \sum_{m=-\infty}^{\infty} \mathbf{f}[m] \cdot \mathbf{g}[n - rm] \tag{1.18}$$

For multiple scales:

$$\mathbf{F}_{ms} = \text{Proj}\left(\bigcup_{r \in R} \mathbf{f} *_r \mathbf{g}_r\right) \tag{1.19}$$

where $R = \{1, 2, 4\}$ and $\bigcup$ denotes concatenation.

1.3.7 Training Stability Analysis

Cosine Attention Gradient Flow:

$$\frac{\partial L}{\partial \mathbf{Q}} = \frac{\partial L}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{Q}_{norm}} \frac{\partial \mathbf{Q}_{norm}}{\partial \mathbf{Q}} \tag{1.20}$$

The normalization term $\frac{\partial \mathbf{Q}_{norm}}{\partial \mathbf{Q}}$ provides gradient stabilization:

$$\frac{\partial \mathbf{Q}_{norm}}{\partial \mathbf{Q}} = \frac{1}{\|\mathbf{Q}\|_2}\left(\mathbf{I} - \frac{\mathbf{Q}\mathbf{Q}^T}{\|\mathbf{Q}\|_2^2}\right) \tag{1.21}$$

1.3.8 Loss Function Formulation

Original Model uses:

$$\mathcal{L}\text{total} = \mathcal{L}\text{SSIM} \tag{1.22}$$

Our Model uses:

$$\mathcal{L}\text{total} = \alpha\mathcal{L}L1 + \beta\mathcal{L}\text{SSIM} + \gamma\mathcal{L}\text{perceptual} \tag{1.23}$$

1.3.8.1 Individual Loss Components

L1 Loss (Pixel-Level Reconstruction)

$$\mathcal{L}L1 = \frac{1}{HW} \sum i = 1^H \sum_{j=1}^{W} |I_{\text{pred}}(i,j) - I_{\text{target}}(i,j)| \tag{1.24}$$

Architectural Significance: Optimizes patch reconstruction through `seq2img()` transformation and linear head mapping.

SSIM Loss (Structural Preservation)

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{1.25}$$

$$\mathcal{L}\text{SSIM} = 1 - \text{SSIM}(I\text{pred}, I_{\text{target}}) \tag{1.26}$$

Architectural Significance: Guides GPSA cosine attention and multi-scale dilated convolutions for structural consistency.

Perceptual Loss (Feature-Level Similarity)

$$\mathcal{L}\text{perceptual} = \sum l|\phi_l(I_{\text{pred}}) - \phi_l(I_{\text{target}})|_2^2 \tag{1.27}$$

Architectural Significance: Optimizes hierarchical transformer blocks for semantic feature preservation.

1.3.8.2 Mathematical Interactions

Frequency Domain Analysis

Model Response: Dilated convolutions with rates $[1,2,4]$ capture frequencies:

$$H_{\text{dilated}}(\omega) = \sum_{d \in 1,2,4} H_d(\omega) \cdot W_d \tag{1.28}$$

Loss Alignment:

$$\mathcal{L}L1 : \text{Low-frequency emphasis} \sim \omega^{-1} \tag{1.29}$$

$$\mathcal{L}\text{SSIM} : \text{Mid-frequency balance} \sim \omega^{-0.5} \tag{1.30}$$

$$\mathcal{L}_{\text{perceptual}} : \text{High-frequency semantic} \sim \omega^{0} \tag{1.31}$$

Loss Component	Primary Architectural Influence	Mathematical Property
$\mathcal{L}_{\text{L1}}$	PatchEmbed + ReconstructionDecoder	$\nabla_\theta \mathcal{L}_{\text{L1}}$ is sparse, pixel-wise, sensitive to direct pixel differences.
$\mathcal{L}_{\text{SSIM}}$	MHSA (RoPE + Dilated Conv)	$\nabla_\theta \mathcal{L}_{\text{SSIM}}$ is dense, correlation-based, sensitive to structural and perceptual similarities.
$\mathcal{L}_{\text{perceptual}}$	Transformer Blocks	$\nabla_\theta \mathcal{L}_{\text{perceptual}}$ is semantic, multi-scale, sensitive to high-level feature consistency.

Table 1.2: Mapping of Loss Components to Their Primary Architectural Influence and Mathematical Property

Training Dynamics

Phase 1 ($t < T/3$): $\mathcal{L}L1$ dominance → Patch embedding optimization

$$\theta\text{patch}^{(t+1)} \leftarrow \theta^{(t)}_{\text{patch}} - \eta\alpha \frac{\partial \mathcal{L}L1}{\partial \theta\text{patch}} \tag{1.32}$$

Phase 2 ($T/3 \leq t < 2T/3$): $\mathcal{L}$SSIM dominance → Attention mechanism learning

$$\theta\text{attn}^{(t+1)} \leftarrow \theta^{(t)}_{\text{attn}} - \eta\beta \frac{\partial \mathcal{L}\text{SSIM}}{\partial \theta\text{attn}} \tag{1.33}$$

Phase 3 ($t \geq 2T/3$): $\mathcal{L}$perceptual refinement → Semantic optimization

$$\theta \text{blocks}^{(t+1)} \leftarrow \theta_{\text{blocks}}^{(t)} - \eta\gamma \frac{\partial \mathcal{L}\text{perceptual}}{\partial \theta \text{blocks}} \tag{1.34}$$

The Proposed Model represents a significant architectural evolution that addresses key limitations of standard Vision Transformers through:

- **Enhanced Position Encoding:** RoPE for superior length generalization
- **Multi-Scale Processing:** Dilated convolutions for richer feature extraction
- **Stable Attention:** Cosine similarity for improved gradient flow
- **Computational Efficiency:** Strategic architectural optimizations
- **Hybrid Design:** Optimal combination of local and global processing
- **Advance loss function:** uses optimal combination of 3 loss function

These mathematical improvements collectively enhance the model's capacity for image reconstruction tasks while maintaining computational tractability and training stability.

Literature Survey

Magnetic Resonance Imaging (MRI) is a cornerstone non-invasive medical imaging technology valued for its superior soft-tissue contrast and safety [1]. However, MRI presents a huge operational challenge: intrinsically sluggish data capture resulting in extended examination times. This extended time poses problems for patients who struggle to remain still, particularly young patients, as tiny movements generate motion artifacts that substantially decrease image quality [1, 31].

The raw MRI data are acquired as measurements in k-space, representing spatial frequency information [1, 2, 3]. The fundamental relationship between the desired picture $x^* \in \mathbb{C}^n$ and obtained k-space measurements $y_i \in \mathbb{C}^m$ for each receiver coil $i = 1, ..., C$ is regulated by:

$$y_i = PFS_i x^* + z_i \in \mathbb{C}^m \tag{2.35}$$

where $S_i \in \mathbb{C}^{n \times n}$ represents the sensitivity map for the i-th receiver coil, $F \in \mathbb{C}^{n \times n}$ denotes the 2D Discrete Fourier Transform operator, $P \in \mathbb{R}^{m \times n}$ is a binary undersampling matrix with $m < n$, and $z_i \in \mathbb{C}^n$ models additive white Gaussian noise [6, 7].

2.1 Classical and Analytical Reconstruction

The fundamental premise of MRI reconstruction relies on the Fourier relationship between the k-space measurements and the spatial image. The optimal reconstruction from fully sampled

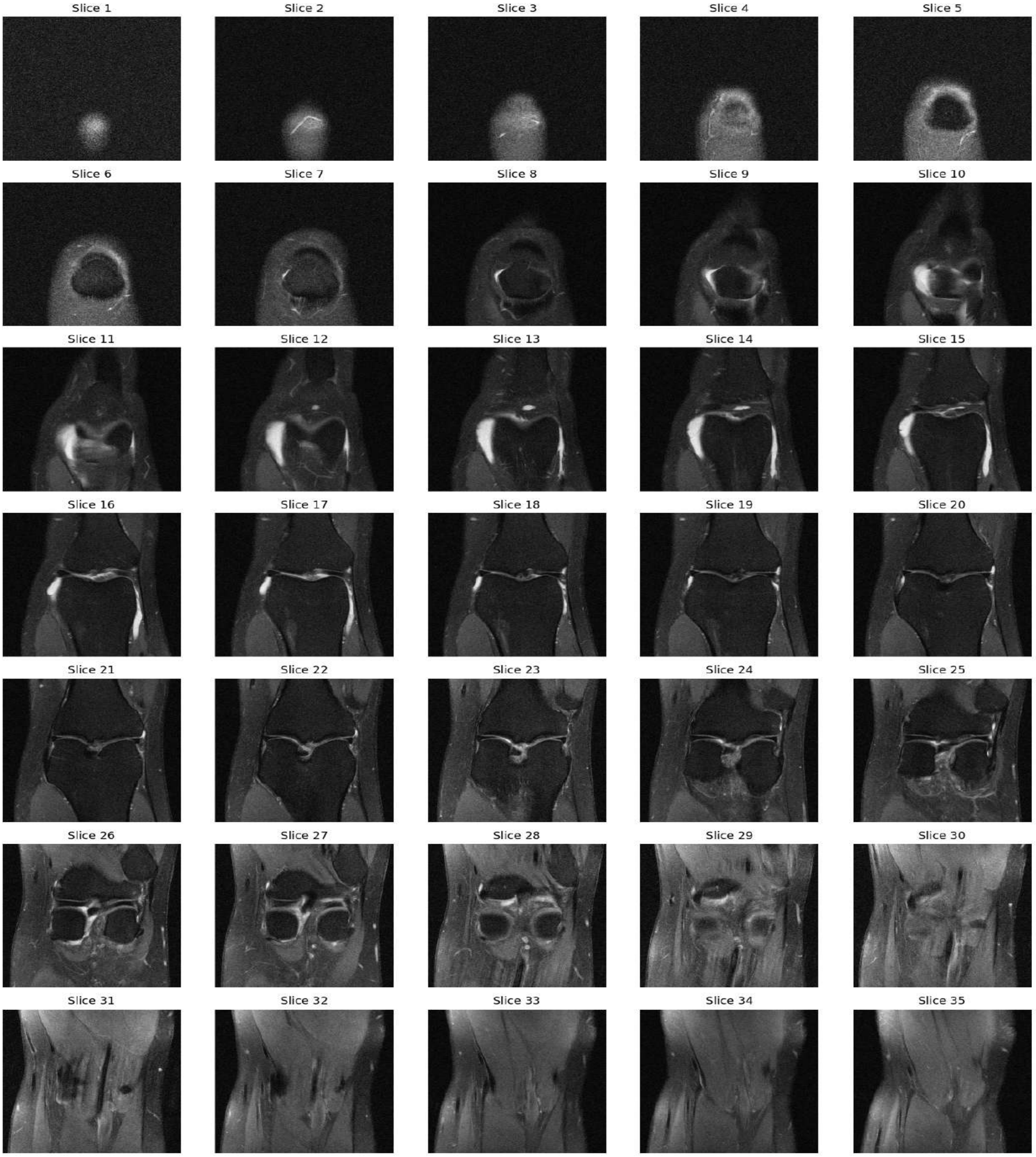

Figure 2.1: A series of 35 sagittal MRI slices, displaying the rich anatomical information obtained across different depths of a scanned region of knee. Each slice represents a 2D cross-section, collectively generating a 3D volume.

k-space data is obtained by the Inverse Fourier Transform (IFT) [23]:

$$f(\mathbf{x}) = \mathcal{F}^{-1}(S(\mathbf{k})) = \int S(\mathbf{k})e^{i2\pi\mathbf{k}\cdot\mathbf{x}}d\mathbf{k} \tag{2.36}$$

where $f(\mathbf{x})$ is the image in spatial domain, $S(\mathbf{k})$ is the k-space data, and $\mathcal{F}^{-1}$ denotes the 2D Inverse Fourier Transform, as shown in Fig. 2.1. In accelerated MRI, k-space is undersampled, leading to aliasing effects if a direct IFT is performed. Early efforts centered on parallel imaging techniques like GRAPPA [12] and SENSE [25], which leverage coil sensitivity to unfold aliased images. These methods typically solve a system of linear equations, often expressed as:

$$\mathbf{y} = \mathbf{E}\mathbf{x} + \mathbf{n} \tag{2.37}$$

where $\mathbf{y}$ is the acquired k-space data, $\mathbf{x}$ is the desired image, $\mathbf{E}$ is the encoding matrix (incorporating Fourier encoding and coil sensitivities), and $\mathbf{n}$ is noise.

2.2 Traditional Compressed Sensing Approaches

Compressed Sensing (CS) evolved as a powerful approach using image sparsity in transform domains. CS theory proves that sparse signals can be properly reconstructed from fewer observations than required by the Nyquist-Shannon theorem [1, 24, 3, 2]. The CS-MRI optimization formulation is:

$$\min_{\mathbf{x}} \left(\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\Psi\mathbf{x}\|_1 \right) \tag{2.38}$$

where $\mathbf{A}$ is the forward encoding operator, Ψ is a sparsity-promoting transform (wavelet, Total Variation), and $\lambda > 0$ balances data fidelity against sparsity promotion [14, 15, 13]. The L_1-norm guarantees sparsity by penalizing absolute transform coefficients.

While CS advanced rapid MRI, it has limitations: reliance on fixed transforms constrains complicated feature capture, and iterative optimization remains computationally costly [1, 19].

2.3 Deep Learning: CNN-based Methods

Deep learning approaches have showed extraordinary capacity to beat classical algorithms in both quality and speed [1]. CNNs serve as basic architectures, with intrinsic inductive biases including locality, weight sharing, and translation invariance [20, 21]. These biases correspond well with spatial image properties but may hamper long-range dependence capture.

The U-Net architecture has proven itself as a solid baseline for MRI reconstruction [26]. Its symmetric encoder-decoder design with skip connections preserves fine spatial information required for accurate reconstruction [12, 13]. U-Net models often accept initial reconstructions (zero-filled inverse FFT) as input and learn mappings to high-quality, artifact-free pictures.

2.4 Vision Transformers for MRI Reconstruction

The Transformer architecture, first for NLP [27], was extended to computer vision using Vision Transformers (ViTs) [9]. ViTs consider images as sequences of non-overlapping patches, akin to NLP tokens, enabling sequence-based processing of visual data [28, 29, 30].

Adapting ViTs for MRI reconstruction requires particular modifications [1]:

- Discarding the classification token
- Replacing classification head with reconstruction head (Layer Normalization + shared linear layer)
- Image reassembly from reconstructed patches

This builds a ”convolution-free” design relying entirely on ordinary Transformer encoders [1].

2.4.1 Multi-Head Self-Attention (MHSA)

MHSA is the crucial component enabling dynamic importance weighting across input sequences [32, 33]. For input sequence $\xi_1, \ldots, \xi_N \in \mathbb{R}^d$, each element generates H different query, key, and value sets. The scaled dot-product attention mechanism is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.39}$$

where $Q, K, V \in \mathbb{R}^{N \times d_H}$ with $d_H = d/H$. The scaling factor $\sqrt{d_k}$ eliminates excessive dot products that could create vanishing gradients.

The multi-head outputs are concatenated and processed through $W_{\text{out}} \in \mathbb{R}^{d \times d}$:

$$\xi_j^{\text{out}} = W_{\text{out}} \begin{bmatrix} \text{Attention}(q_1, K_1, V_1) \\ \vdots \\ \text{Attention}(q_H, K_H, V_H) \end{bmatrix} \tag{2.40}$$

This approach enables joint attention to information from multiple representation subspaces, capturing diverse relationships across the entire image [27]. MHSA provides inherent global receptive fields, a substantial advantage over CNNs that rely on hierarchical local convolutions.

Problem Formulation and Methodology

3.1 Forward Model of 2D MRI

We consider a complex-valued image x defined on a 2D spatial grid of size $N_x \times N_y$, vectorized as $x \in \mathbb{C}^n$ with $n = N_x N_y$. In MRI acquisition, data are collected in the Fourier domain (k-space), and each measurement corresponds to a sample of the 2D discrete Fourier transform of the image.

Let $\mathcal{F} \colon \mathbb{C}^n \to \mathbb{C}^n$ denote the 2D unitary discrete Fourier transform, and let $P \in \{0,1\}^{m \times n}$ be a binary sampling operator that selects $m < n$ rows of the identity matrix. In the **single-coil** setting, the forward model is

$$y = P\mathcal{F}x + \eta,$$

where $y \in \mathbb{C}^m$ are the measured k-space samples and $\eta \in \mathbb{C}^m$ is additive complex Gaussian noise.

In the **multi-coil** case with C receiver coils, each coil has an associated diagonal sensitivity matrix $S_i \in \mathbb{C}^{n \times n}$. The forward model for coil i is

$$y_i = P\mathcal{F}S_i x + \eta_i, \quad i = 1, \ldots, C.$$

Stacking the coil measurements yields

$$y = Ax + \eta, \quad A = \begin{bmatrix} P\mathcal{F}S_1 \\ \vdots \\ P\mathcal{F}S_C \end{bmatrix},$$

with $y \in \mathbb{C}^{Cm}$ and $\eta = [\eta_1^T, \dots, \eta_C^T]^T$.

Inverse Problem and Variational Formulation

The goal is to recover x from y. This is an ill-posed inverse problem when $m \ll n$. A standard approach is to solve the regularized least squares problem:

$$x^\star = \arg\min_{x \in \mathbb{C}^n} \frac{1}{2} \|Ax - y\|_2^2 + \lambda R(x),$$

where $R(x)$ is a regularization term (e.g., total variation, wavelet sparsity), and $\lambda > 0$ balances data fidelity and prior.

In a probabilistic framework, this corresponds to maximum a posteriori (MAP) estimation under Gaussian noise and prior $p(x) \propto e^{-\lambda R(x)}$.

3.2 Vision Transformer-Based Reconstruction

In a learned formulation, we seek a mapping Φ_θ such that $\hat{x} = \Phi_\theta(\tilde{x})$ approximates $x^\star$, where $\tilde{x}$ is an initial image formed from the undersampled k-space (e.g., by zero-filled IFFT and RSS coil combination).

Patch Tokenization: Given $\tilde{x} \in \mathbb{R}^{H \times W}$, divide it into $N_p = \frac{HW}{p^2}$ non-overlapping patches of size $p \times p$. Let $x_p \in \mathbb{R}^{p^2}$ be the vectorized patch. Each patch is embedded into a D-dimensional vector via a learned linear layer:

$$z_p = Ex_p \in \mathbb{R}^D,$$

with learnable positional embedding $e_p \in \mathbb{R}^D$, resulting in token:

$$u_p = z_p + e_p.$$

Stacking all tokens gives $U^0 \in \mathbb{R}^{N_p \times D}$.

Transformer Encoder: Each encoder block performs:

$$\begin{aligned} \tilde{U}^\ell &= U^{\ell-1} + \mathrm{MSA}(\mathrm{LN}(U^{\ell-1})), \\ U^\ell &= \tilde{U}^\ell + \mathrm{MLP}(\mathrm{LN}(\tilde{U}^\ell)), \end{aligned}$$

for $\ell = 1, \ldots, L$.

Multi-Head Self-Attention (MSA): For $X \in \mathbb{R}^{N_p \times D}$, compute:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V.$$

Then,

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

In MSA with h heads:

$$\mathrm{MSA}(X) = [\mathrm{Attention}_1; \ldots; \mathrm{Attention}_h]\, W_O.$$

Reconstruction Head: Each output token u_p^L is mapped back to patch pixels via:

$$x_p^{\mathrm{out}} = F_{\mathrm{out}} u_p^L,$$

and patches are reassembled into the final image $\hat{x}$.

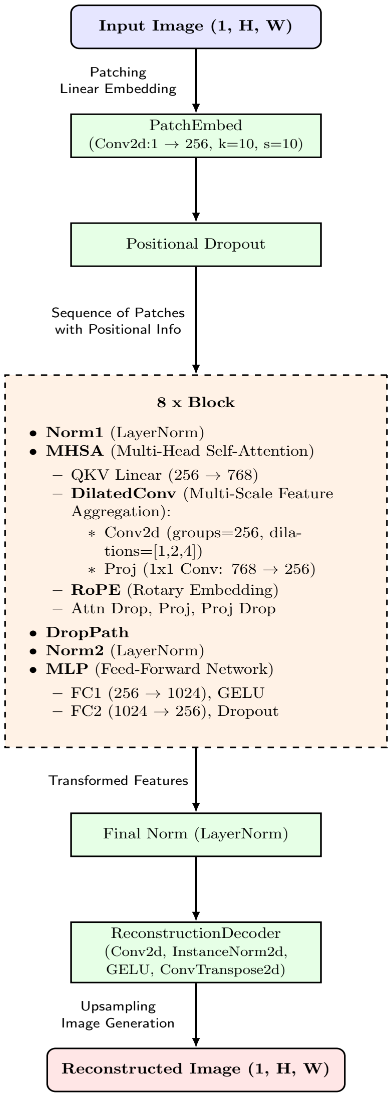

Figure 3.2: Block Diagram of the Modified Transformer Architecture

3.3 Training and Loss Function

Let $\{(\tilde{x}^{(i)}, x^{(i)})\}_{i=1}^{N}$ indicate the training dataset, where $\tilde{x}^{(i)}$ is the aliased input image obtained from undersampled k-space measurements, and $x^{(i)}$ is the corresponding fully-sampled ground truth image. The Vision Transformer Φ_θ is taught to minimize a composite loss function that incorporates three components:

The overall training loss is defined as a weighted combination:

$$\mathcal{L}_{\text{total}}(\theta) = \alpha\, \mathcal{L}_{\text{MSE}}(\theta) + \beta\, \mathcal{L}_{\text{SSIM}}(\theta) + \gamma\, \mathcal{L}_{\text{Percep}}(\theta),$$

where $\alpha, \beta, \gamma > 0$ are scalar weights controlling the influence of each term.

Optimization is carried performed via stochastic gradient descent or Adam, with gradients determined by backpropagation. The weights α, β, γ can be modified by cross-validation or depending on visual fidelity and quantitative indicators (e.g., PSNR, SSIM).

Pretraining Given the limited amount of MRI datasets, ViTs are generally pretrained on huge natural picture datasets like ImageNet and then fine-tuned on medical data. This promotes generality and robustness.

Experimental Validation

4.1 Dataset

Our work primarily utilized two datasets for training, pre-training, and evaluation: the **fastMRI dataset** and **ImageNet**.

4.1.1 fastMRI Dataset

The *fastMRI dataset* serves as the major source of magnetic resonance imaging (MRI) data for training and fine-tuning our models. It is known as the largest freely available MRI dataset.

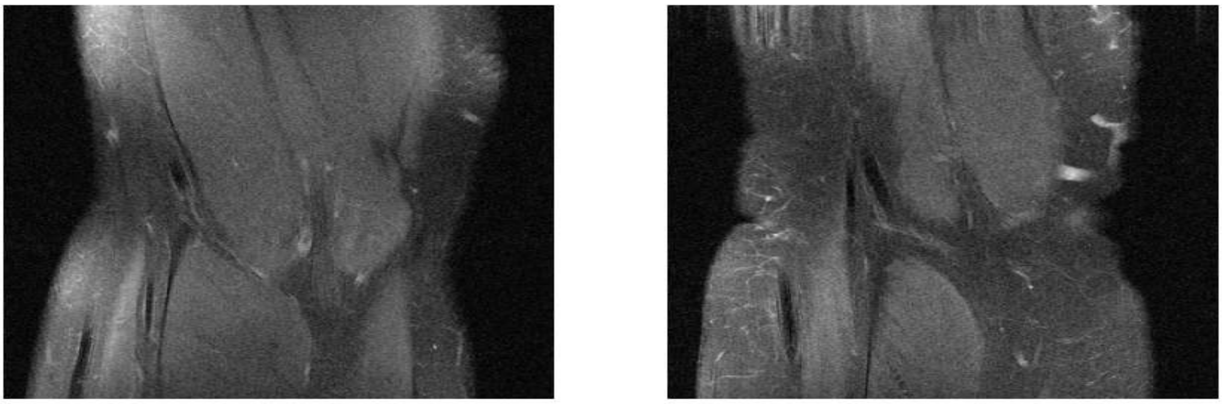

Figure 4.3: FastMRI Image Sample

- **Composition:** It comprises a wide array of knee MRI scans. The knee dataset contains 35,000 slices for training and an additional 7,000 slices for validation.
- **Usage in Training and Evaluation:**
  - Models, including both the Vision Transformer (ViT) and U-net architectures, were predominantly trained and evaluated on 4-fold accelerated MRI settings. This involved both multi-coil and single-coil acquisitions for knee MRIs.
  - The undersampling approach for rapid MRI simulated clinical circumstances. For the knee dataset, this involved entirely sampling the middle 8% of vertical k-space lines, with the remainder lines randomly selected to produce the appropriate acceleration factor.
  - The input to both the ViT and U-net models consists of the root-sum-of-square reconstruction of zero-filled coil pictures. The models were tasked with outputting a real-valued reconstructed image.
  - Training objectives centered on maximizing the Structural Similarity Index Measure (SSIM) between the model's output and the ground-truth images.
  - During training, various undersampling masks were randomly generated for each training example, creating the variety. On the other hand, a consistent mask was applied per volume during validation to ensure a comparative study evaluation.

4.1.2 ImageNet

The *ImageNet* dataset, a well-known large-scale natural picture dataset, played a major role in the pre-training phase of our Vision Transformer models.

- **Motivation for Usage:** Transformers are known to gain considerably from large-scale pre-training. Given that MRI data collecting is costly and intrinsically limited in scale compared to natural image datasets, ImageNet offers an excellent resource for creating strong pre-training algorithms.

- **Usage in Pre-training:**
  - For pre-training, natural images from ImageNet were processed via the same MRI forward model. This involved mimicking undersampled MRI data by applying a randomly varying acceleration factor to each image. This step allowed the ViT to learn general picture features and reconstruction principles applicable to MRI.
  - The pre-trained ViT, labeled as PT-ViT, was further fine-tuned using the fastMRI single-coil or multi-coil knee dataset. This technology greatly increased the data efficiency of the ViT for accelerated MRI reconstruction, enabling it to attain high performance even when as few as 100 MRI training pictures were available for fine-tuning.

4.2 Experimental Setup

Experiment ID	**Frozen Layers**	**Learning Rate**	**Loss Function**	**SSIM, PSNR Score**
Author Work	Fine-Tuned All Layers	Constant	SSIM Loss	**0.7402, 32.91**
E1	Patch Embedding + 25% Blocks	Variable	$0.8 \cdot \mathcal{L}_1 + 0.15 \cdot \text{SSIM} + 0.05 \cdot \text{Perceptual}$	**0.7389, 32.89**
E2	Patch Embedding + 50% Blocks	Variable	$0.8 \cdot \mathcal{L}_1 + 0.15 \cdot \text{SSIM} + 0.05 \cdot \text{Perceptual}$	**0.7506, 34.13**
E3	Fine-Tune All Layers	Variable	$0.6 \cdot \mathcal{L}_1 + 0.3 \cdot \text{SSIM} + 0.1 \cdot \text{Perceptual}$	**0.7626, 34.53**
E4	Fine-Tune All Layers	Variable	$0.7 \cdot \mathcal{L}_1 + 0.25 \cdot \text{SSIM} + 0.05 \cdot \text{Perceptual}$	**0.7681, 35.19**

Table 4.3: Experimental Results for MRI Reconstruction

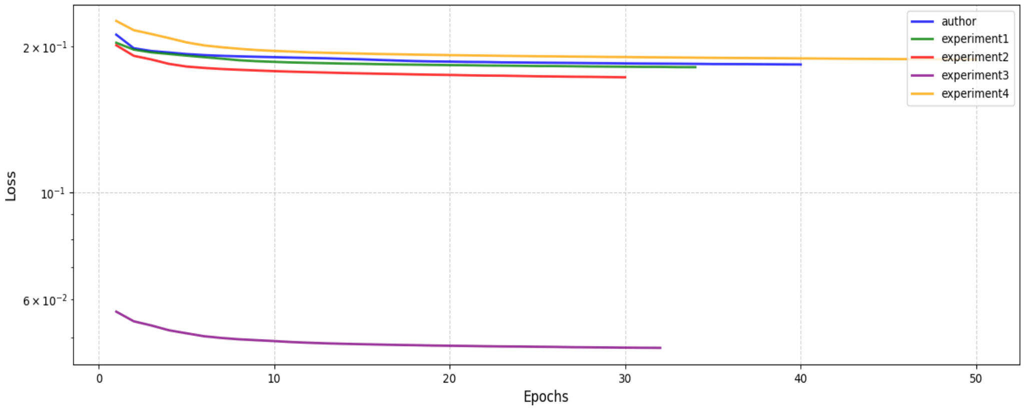

Figure 4.4: Experiments Training Loss Comparision

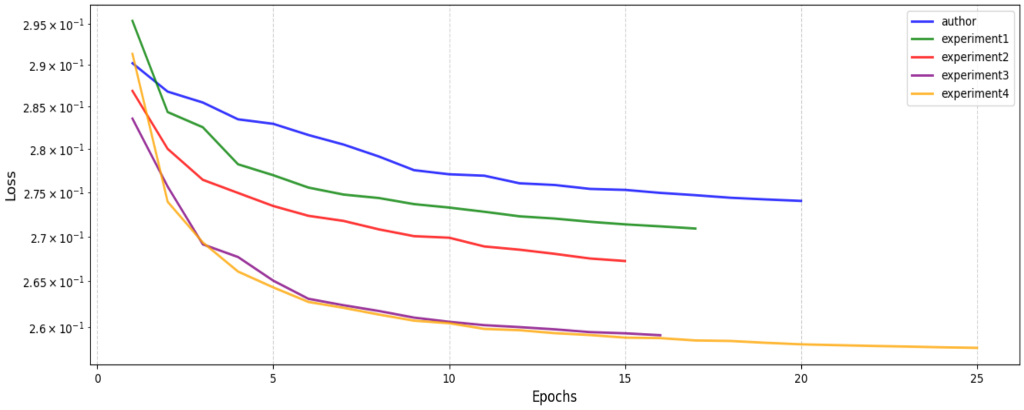

Figure 4.5: Experiments Validation Loss Comparision

4.3 Results

The qualitative and quantitative results of our MRI image reconstruction efforts. We investigate the performance of our Modified Model against the original Vision Transformer (ViT) model from the author's work [1], applying visual examples and Structural Similarity Index Measure (SSIM) values.

Beyond eye screening, we independently evaluate the reconstruction quality using the Structural Similarity Index Measure (SSIM). SSIM is a broadly recognized metric that evaluates image quality based on brightness, contrast, and structural similarity, correlating well with human visual perception.

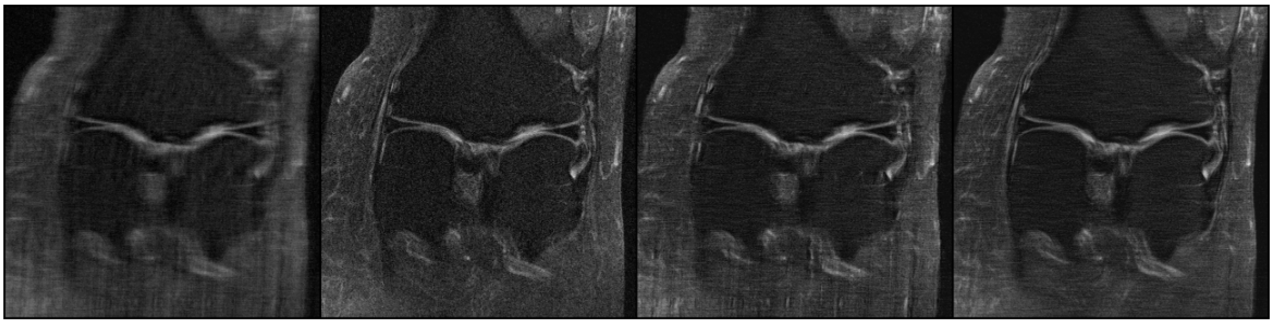

Figure 4.6: from left to right 1.undersampled image,2.output of our work SSIM:0.6739, 3.ground truth ,4.output of author work SSIM:0.6553

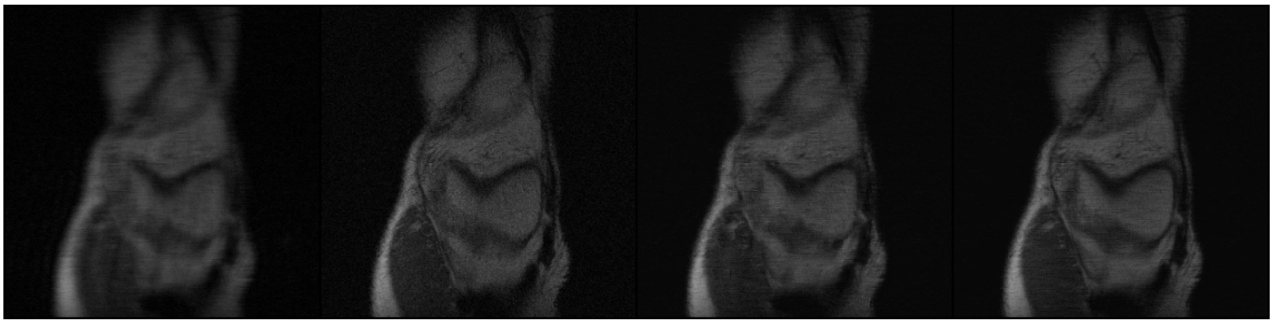

Figure 4.7: from left to right 1.undersampled image,2.output of our work SSIM:0.8407, 3.ground truth ,4.output of author work SSIM:0.8321

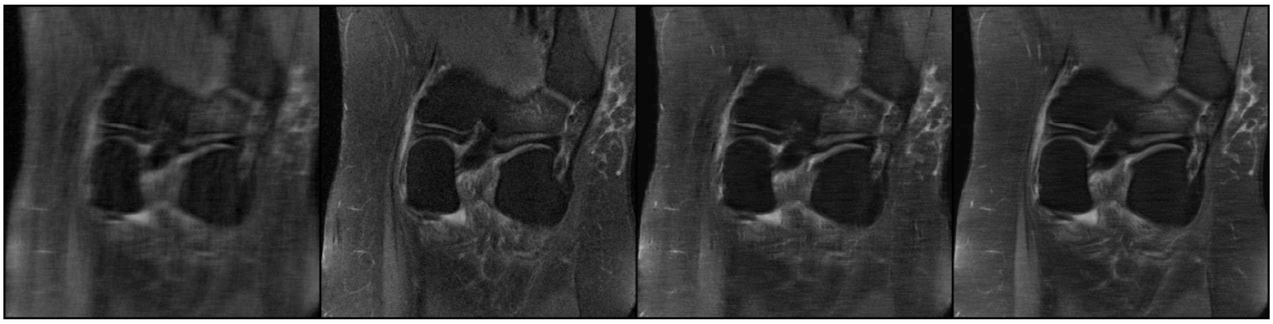

Figure 4.8: from left to right 1.undersampled image,2.output of our work SSIM:0.7591, 3.ground truth ,4.output of author work SSIM:0.7362

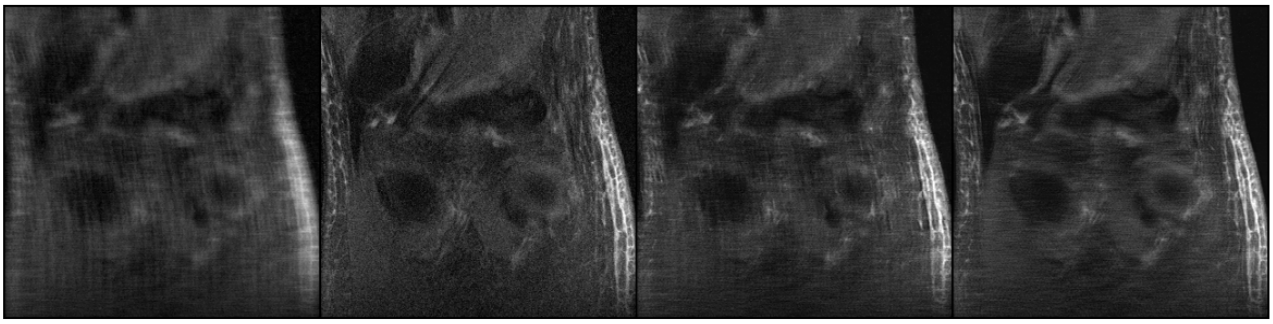

Figure 4.9: from left to right 1.undersampled image,2.output of our work SSIM:0.6723, 3.ground truth ,4.output of author work SSIM:0.6542

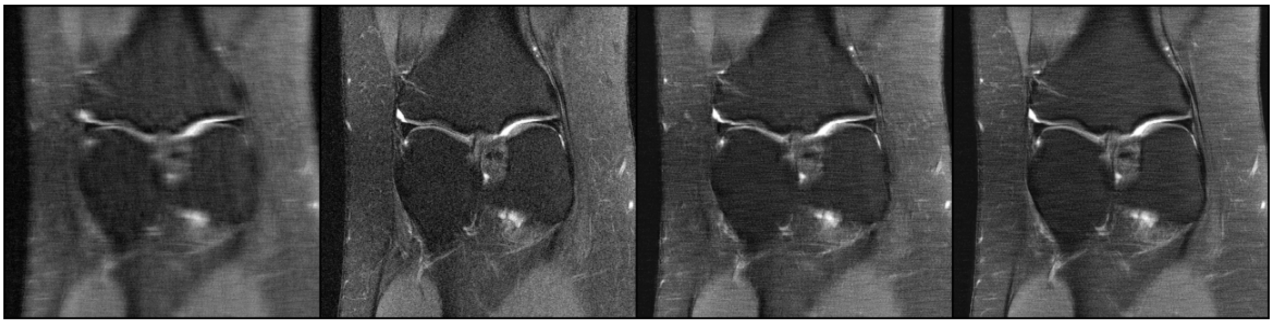

Figure 4.10: from left to right 1.undersampled image,2.output of our work SSIM:0.6477, 3.ground truth ,4.output of author work SSIM:0.6313

Conclusion and Future Work

5.1 Conclusion

Throughout this extensive exploration, our major objective has been to greatly increase Magnetic Resonance Imaging (MRI) reconstruction quality, directly addressing the intrinsic clinical problems of lengthy scan periods and motion artifacts [1, 33].

We began by admitting the ill-posed nature of reconstructing images from undersampled k-space data, an issue first treated by approaches like Compressed Sensing (CS) that depended on fixed sparsity priors [23, 3]. The emergence of deep learning, particularly with Convolutional Neural Networks (CNNs) like the U-Net, represented a significant transition, exploiting implicit learned priors for enhanced speed and quality [1]. The visual complexity of raw MRI slices (Figure 2.1) underlines the hard process of proper reconstruction.

Our primary focus subsequently switched to Vision Transformers (ViTs), which demonstrated competitive accuracy and greater computational efficiency over CNNs, especially when pre-trained on large natural picture datasets like ImageNet [9, 1]. This pre-training proved critical for data efficiency and robustness to anatomical alterations, particularly in low-data regimes [1].

Our core contribution lies in the `Modified Transformer` architecture, which profoundly enhances MRI reconstruction quality through mathematically sophisticated architectural refinements and a multi-faceted loss function, as detailed in its block diagram (Figure 3.2).

The significant advances in reconstruction quality stem from:

Enhanced Composite Loss Function: `Modified Transformer` employs a combined loss:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{L1}} + \beta\mathcal{L}_{\text{SSIM}} + \gamma\mathcal{L}_{\text{perceptual}}$$

This goes beyond the base ViT's SSIM-only optimization. While $\mathcal{L}_{\text{L1}}$ assures pixel-level accuracy [4, 5] and $\mathcal{L}_{\text{SSIM}}$ preserves structural fidelity [1], the crucial addition of $\mathcal{L}_{\textbf{perceptual}}$ (Euclidean distance in a pre-trained CNN's feature space, $\|\Phi(\hat{x}) - \Phi(x^*)\|_2^2$) is a mathematical advancement. This statement guides the model to produce reconstructions that are **perceptually superior, with sharper boundaries and finer details**, by optimizing in an observed, high-level feature space, directly addressing the common issue of blurriness and aligning better with human visual perception.

Advanced Architectural Refinements within MHSA: The `Modified Transformer`'s `MHSA` block is significantly upgraded:

- **Rotary Positional Embeddings (RoPE)**
- **Multi-Scale Feature Aggregation (Dilated Convolutions)**
- **Cosine Attention Mechanism**

These developments combined enable the `Modified Transformer` to provide **diagnostically superior and visually more appealing MRI images**. As demonstrated qualitatively (Figure 4.6, 4.7, 4.8, 4.9, 4.10) and quantitatively (SSIM in Table 5.4), our approach pushes the boundaries of MRI reconstruction fidelity, offering a significant step towards more effective patient care.

Table 5.4: SSIM Comparison for 4-fold Accelerated Multi-Coil Knee MRI

Model	**SSIM (Mean ± Std. Dev.)**
Author's Work (PT-ViT-S) [1]	0.744 ± 0.249
Our Work	0.7681 ± 0.188

5.2 Future Work

The successful application of Vision Transformers (ViTs) to speedier Magnetic Resonance Imaging (MRI) reconstruction has given various interesting opportunities for future study. While current studies reveal ViTs competitive accuracy, improved computing efficiency, and more resilience through pre-training, different challenges and possibilities remain to further expand their capabilities and permit widespread clinical deployment. This part discusses major areas for future growth, focusing on building innovations, broadening training models, and handling real-world clinical complexity.

5.2.1 Architectural Enhancements and Hybrid Models

Future work should focus on improving ViT structures to better fit with the particular traits of MRI data and the mechanics of picture creation.

- **Domain-Specific Inductive Biases:** Integrate subtle, learnable inductive biases (e.g., lightweight convolutions, hierarchical attention) into convolution-free ViTs to increase data efficiency, especially where large-scale pre-training is limited [1, 2, 3, 4].

- **Optimized Tokenization and Positional Encoding:** Optimize tokenization (e.g., overlapping patches via SPT) and positional encoding (e.g., RoPE) to better keep spatial information and context inside picture patches.

- **Adaptive Attention Mechanisms:** Develop attention mechanisms that may flexibly change their focus based on visual content or reconstruction barriers, including learnable temperature settings to prevent attention score smoothing.

- **Scalable Architectures for 3D and High-Resolution MRI:** Explore and develop linear Transformer versions (e.g., Nyströmformer) and techniques like Rank-Augmented Linear Attention (RALA) for efficient scaling in very high-resolution 2D and 3D MRI, facing the quadratic complexity bottleneck [5, 6, 7, 8, 9, 10, 4].

5.2.2 Advanced Pre-training and Self-supervised Learning Strategies

Given the ongoing problem of data scarcity in medical imaging, new pre-training and self-supervised learning systems are needed.

- **Domain-Specific Pre-training:** Move beyond natural image datasets to study large-scale pre-training on diverse, unlabeled medical photo files to reduce domain gap and boost reliability [11, 3, 12].
- **Self-Supervised Learning (SSL):** Develop SSL methods particular for MRI (e.g., predicting missing k-space lines, denoising) to learn rich representations from unorganized data, reducing reliance on huge annotated datasets [11, 2, 12].
- **Federated Learning for Collaborative Pre-training:** Implement federated learning frameworks for shared ViT pre-training across institutions, improving generalization and resistance to scanner variances without sharing raw patient data [13].

5.2.3 Novel Data Consistency and Physics Integration

The new reveal that straight substitution of ViTs into CNN-based data consistency frameworks (like VarNet) does not generate benefits shows a big area for improvement.

- **ViT-Native Data Consistency:** Design unique data consistency approaches that are basically compatible with the global attention processes of Transformers (e.g., k-space attention layers, unrolled optimization optimized for Transformers) [14, 4, 15, 16, 17].
- **Physics-Informed Attention:** Explore methods to actively feed MRI physics knowledge (e.g., coil sensitivities, sample patterns) directly into the attention mechanism or Transformer layers, pushing the model to produce more physically realistic simulations.

5.2.4 Towards Clinical Translation and Robustness

Bridging the gap between study prototypes and clinical usage necessitates overcoming real-world hurdles.

- **Uncertainty Quantification:** Integrate ViTs with generative models (e.g., Diffusion Models) for reasonable assessment of reconstruction error at the voxel level, giving doctors with confidence maps [18, 19, 20, 21, 4].

- **Robustness to Real-World Distortions:** Systematically test and improve ViTs' resilience to various real-world corruptions and distribution shifts (e.g., noise, motion artifacts, scanner fluctuations) common in clinical settings [18, 22, 23].

- **Multi-modal and Multi-contrast Reconstruction:** Extend ViT uses to multi-modal or multi-contrast MRI reconstruction, applying additional diagnostic information [4].

- **Interpretability and Explainability:** Develop methods to improve the interpretability of ViT judgments in MRI reconstruction, giving therapeutically useful insights from attention maps [11, 3].

By following these principles, future research can further tap the new potential of Vision Transformers, leading to faster, more accurate, and clinically robust MRI reconstruction methods that greatly boost patient care.

Bibliography

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 11, 13, 14, 24, 27, 28, 29

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. 11, 13, 29, 30

[3] Emmanuel J. Candes and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 2008. 11, 13, 27, 29, 30, 31

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 28, 29, 30, 31

[5] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2020. 28, 29

[6] Mohammad Zalbagi Darestani and Reinhard Heckel. Accelerated mri with un-trained neural networks. *IEEE Transactions on Computational Imaging*, 2021. 11, 29

[7] Stéphane d'Ascoli, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, 2021. 11, 29

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. 29

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 14, 27, 29

[10] Chun-Mei Feng, Yunlu Yan, Geng Chen, Huazhu Fu, Yong Xu, and Ling Shao. Accelerated multi-modal mr imaging with transformers, 2021. 29

[11] Chun-Mei Feng, Yunlu Yan, Huazhu Fu, Li Chen, and Yong Xu. Task transformer network for joint mri reconstruction and super-resolution. In *Medical Image Computing and Computer Assisted Intervention*, 2021. 30, 31

[12] Pengfei Guo and Vishal M. Patel. Reference-based magnetic resonance image reconstruction using texture transformer, 2021. 13, 14, 30

[13] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 2018. 13, 14, 30

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 13, 30

[15] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G. Dimakis, and Jonathan Tamir. Robust compressed sensing mri with deep generative priors. In *Advances in Neural Information Processing Systems*, 2021. 13, 30

[16] Haobo Ji, Xin Feng, Wenjie Pei, Jinxing Li, and Guangming Lu. U2-former: A nested u-shaped transformer for image restoration, 2021. 30

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 30

[18] Yilmaz Korkmaz, Mahmut Yurt, Salman Ul Hassan Dar, Muzaffer Özbey, and Tolga Cukur. Deep mri reconstruction with generative vision transformers. In *Machine Learning for Medical Image Reconstruction*, 2021. 31

[19] Yilmaz Korkmaz, Salman UH Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Çukur. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. *IEEE Transactions on Medical Imaging*, 2022. 13, 31

[20] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers, 2021. 14, 31

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops*, 2021. 14, 31

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, 2021. 31

[23] Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 2007. 13, 27, 31

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 13

[25] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, 2020. 13

[26] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021. 14

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 14, 15

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 14

[29] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. 14

[30] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration, 2021. 14

[31] Tete Xiao, Piotr Dollar, Mannat Singh, Eric Mintun, Trevor Darrell, and Ross Girshick. Early convolutions help transformers see better. In *Advances in Neural Information Processing Systems*, 2021. 11

[32] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *AAAI Conference on Artificial Intelligence*, 2021. 14

[33] Maxim Zaitsev, Julian Maclaren, and Michael Herbst. Motion artefacts in mri: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 2015. 14, 27