Accelerated MRI Reconstruction with SwinUNet: Enhancing Image Quality through Transformer-Based Architecture

Feolu Kolawole, Yogesh Seenichamy, Kesavan Ramakrishnan
Stanford University
Stanford, CA 94305
flukol@stanford.edu, yogesh60@stanford.edu, kesavanr@stanford.edu

Abstract

Magnetic Resonance Imaging (MRI) is a critical medical imaging technique that provides detailed anatomical information without harmful radiation. However, the lengthy acquisition time of fully-sampled MRI scans presents significant challenges in clinical settings. In this paper, we address the problem of accelerated MRI reconstruction from undersampled k-space data using deep learning approaches. We propose a hybrid architecture that combines the strengths of U-Net with the Swin Transformer to effectively capture both local features and global dependencies in MRI images. Our approach involves a systematic comparison of multiple architectures, including baseline U-Net, Transformer at bottleneck (BT), and SwinUNet, followed by extensive hyperparameter tuning. Experimental results on the fastMRI single-coil knee dataset demonstrate that our optimized SwinUNet model achieves superior performance with a PSNR of 33.1 dB and SSIM of 0.72, outperforming the baseline U-Net by approximately 4 dB in PSNR and maintaining more stable SSIM values. These improvements translate to enhanced image quality with better preservation of anatomical details, potentially enabling faster MRI acquisition without sacrificing diagnostic value.

1. Introduction

Magnetic Resonance Imaging (MRI) stands as one of the most valuable non-invasive medical imaging techniques, providing exceptional soft tissue contrast and detailed anatomical information without exposing patients to ionizing radiation. Despite these advantages, MRI suffers from inherently long acquisition times, often requiring patients to remain motionless for extended periods. This limitation not only reduces patient comfort and scanner throughput but can also lead to motion artifacts that degrade image quality. Additionally, lengthy scan times increase healthcare costs and limit MRI accessibility in resource-constrained settings.

To address these challenges, accelerated MRI techniques have been developed that undersample k-space (the raw frequency domain data collected by MRI scanners) to reduce acquisition time. However, this undersampling introduces artifacts and distortions in the reconstructed images when conventional reconstruction methods are used. The fundamental challenge lies in reconstructing high-quality, diagnostically valuable images from this incomplete k-space data.

Traditional approaches to this problem include parallel imaging techniques like SENSE and GRAPPA, and compressed sensing methods. While these approaches have shown promise, they often suffer from lengthy reconstruction times, increased noise, or dependence on specific sampling patterns. More recently, deep learning-based methods have emerged as powerful alternatives, demonstrating superior performance in terms of both reconstruction quality and speed.

In this paper, we investigate a hybrid deep learning architecture for accelerated MRI reconstruction that combines the strengths of convolutional neural networks (CNNs) and transformer models. Our approach builds upon the widely-used U-Net architecture, which has proven effective for image-to-image tasks, by incorporating transformer components to better capture long-range dependencies in the image data. Specifically, we explore and compare several architectural variants, including a baseline U-Net, a U-Net with transformer blocks at the bottleneck (BT), and a SwinUNet that integrates the hierarchical Swin Transformer design.

Our contributions can be summarized as follows:

- We systematically evaluate multiple deep learning architectures for accelerated MRI reconstruction, including U-Net, transformer-augmented U-Net, and SwinUNet.
- We demonstrate that the SwinUNet architecture achieves superior performance compared to the base-

line U-Net, with approximately 4 dB improvement in PSNR and more stable SSIM values.

- We conduct extensive hyperparameter tuning and ablation studies to optimize model performance and prevent overfitting.
- We provide a comprehensive analysis of the trade-offs between different architectural choices and their impact on reconstruction quality.

The remainder of this paper is organized as follows: Section 2 discusses related work in accelerated MRI reconstruction. Section 3 describes the fastMRI dataset and our data preprocessing pipeline. Section 4 details our technical approach, including the baseline U-Net and our proposed SwinUNet architecture. Section 5 presents experimental results and comparisons. Finally, Section 6 concludes the paper and suggests directions for future work.

2. Related Work

Accelerated MRI reconstruction has been an active area of research for decades, with approaches evolving from traditional signal processing methods to advanced deep learning techniques. Here, we review key developments in this field, focusing on deep learning-based approaches that are most relevant to our work.

2.1. Traditional Reconstruction Methods

Conventional approaches to accelerated MRI reconstruction include parallel imaging techniques such as SENSE [11] and GRAPPA [3], which leverage data from multiple receiver coils. Compressed sensing methods [10] exploit the sparsity of MRI images in appropriate transform domains to recover images from undersampled data. While these methods have been widely adopted in clinical practice, they often suffer from lengthy reconstruction times, increased noise levels, and reliance on specific sampling patterns or calibration data.

2.2. Deep Learning for MRI Reconstruction

The application of deep learning to MRI reconstruction began with simple CNN architectures. Hammernik et al. [4] proposed a variational network that unrolls the optimization process of a variational model. Schlemper et al. [12] introduced a cascade of CNNs that operate in both image and data domains. The fastMRI challenge [13] accelerated progress in this field by providing a large-scale dataset and standardized evaluation metrics.

U-Net-based architectures have emerged as particularly effective for MRI reconstruction. Hyun et al. [6] adapted the U-Net for k-space completion, while Lee et al. [7] proposed a deep residual learning approach using U-Net. The U-Net's ability to capture multi-scale features through its encoder-decoder structure with skip connections makes it well-suited for preserving both fine details and global context in reconstructed images.

2.3. Transformer-Based Approaches

More recently, transformer models, which were originally developed for natural language processing tasks, have been adapted for computer vision applications, including medical image analysis. Vision Transformers (ViT) [2] demonstrated that pure transformer architectures could achieve competitive performance on image classification tasks. This success has inspired various transformer-based approaches for medical image segmentation and reconstruction.

Lin and Heckel [8] showed that Vision Transformers can match U-Net performance for accelerated MRI while reducing computational requirements. Huang et al. [5] proposed a Swin Deformable Attention U-Net Transformer (SDAUT) that combines the strengths of both CNNs and transformers for explainable fast MRI reconstruction. These hybrid approaches leverage the local feature extraction capabilities of CNNs and the long-range dependency modeling of transformers.

2.4. Swin Transformer and SwinUNet

The Swin Transformer [9] introduced a hierarchical architecture with shifted windows that efficiently models both local and global dependencies. This design addresses the quadratic computational complexity of standard transformers with respect to image size, making it more practical for high-resolution medical images. Building on this, Cao et al. [1] proposed SwinUNet, which adapts the Swin Transformer for medical image segmentation in a U-Net-like encoder-decoder structure.

Our work builds upon these advances by adapting and optimizing the SwinUNet architecture specifically for the task of accelerated MRI reconstruction. Unlike previous approaches that either use transformers as a complete replacement for CNNs or only at specific points in the network, our method systematically evaluates different integration strategies and identifies the optimal configuration for MRI reconstruction.

3. Data

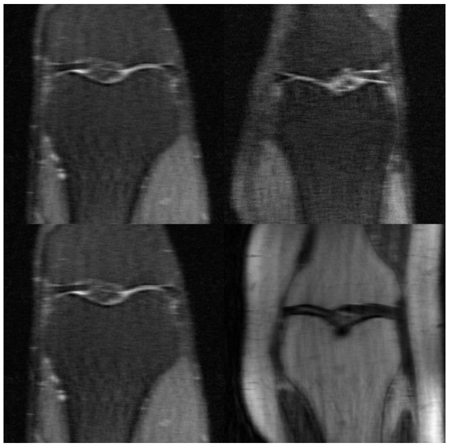

Figure 1. Raw and Undersampled Knee MRI Data

3.1. Dataset Description

Split	Volumes/Middle Slices
Training	973
Validation	199

Table 1. Dataset statistics for the single-coil knee MRI reconstruction task.

For this study, we utilized the fastMRI dataset [13], a large-scale collection of MRI data specifically designed for machine learning approaches to MR image reconstruction. The dataset contains both raw k-space measurements and clinical MR images from knee and brain MRI scans. We focused exclusively on the single-coil knee MRI subset to reduce complexity while still addressing a clinically relevant reconstruction problem.

The fastMRI knee dataset includes two types of contrast mechanisms: Proton Density weighted without fat suppression (PD) and with fat suppression (PDFS). Both acquisition types were included in our experiments to ensure model generalization across different contrast settings. Following the official fastMRI data split, we used 973 volumes for training and 199 volumes for validation, as shown in Table 1.

From each volume, we extracted the middle slice (calculated by dividing the k-space shape by 2). This decision was motivated by several factors:

- The middle slice typically contains substantial anatomical information and is representative of the volume's content
- It provides a consistent anatomical reference point across different scans
- It reduces computational requirements while still preserving the core image reconstruction challenge

3.2. Data Pre-processing Pipeline

Our data pre-processing workflow consisted of the following steps:

1. **K-space Data Loading:** We loaded the complex-valued k-space data from the h5 files using the h5py library.

2. **Middle Slice Extraction:** We identified and extracted the middle slice from each volume.

3. **K-space to Tensor Conversion:** We converted the NumPy complex arrays to PyTorch tensors using fastMRI's transformation utilities (T.to_tensor).

4. **Undersampling:** To simulate accelerated MRI acquisition, we retrospectively undersampled the k-space data using the fastMRI RandomMaskFunc with a center fraction of 0.08 and an acceleration factor of 4. This ensures the central 8% of k-space lines are fully sampled (preserving low-frequency information critical for image contrast), while the remaining lines are randomly sampled.

5. **Image Reconstruction:** For undersampled data, we performed an inverse Fast Fourier Transform (fastmri.ifft2c) followed by computing the absolute value (fastmri.complex_abs) to obtain the zero-filled reconstructions.

6. **Normalization:** All images were normalized by dividing by the maximum value provided in the metadata for each scan to ensure consistent intensity ranges across the dataset.

7. **Center Cropping:** Both input and target images were center-cropped to 320×320 pixels to remove readout and phase oversampling, matching the standard protocol in the fastMRI dataset.

3.3. Efficient Data Handling

To efficiently handle the data during training, we implemented a custom PyTorch Dataset class that processes the fastMRI data on-the-fly. Additionally, we created a pre-processing pipeline that saved processed datasets to disk, significantly reducing data loading time during training while maintaining the flexibility to apply different undersampling patterns.

3.4. Data Visualization

Figure 1 illustrates the core challenge addressed in our work: the recovery of high-quality MRI images from undersampled k-space data. The undersampled data demonstrates coherent artifacts that significantly impair diagnostic quality. These artifacts are particularly problematic as they can obscure important anatomical boundaries and small pathological features.

4. Methods

4.1. Problem Formulation

The accelerated MRI reconstruction problem can be formulated as recovering a high-quality image x from undersampled k-space measurements $y = M \odot \mathcal{F}(x)$, where $\mathcal{F}$ represents the Fourier transform, M is a binary mask indicating which k-space points are sampled, and $\odot$ denotes element-wise multiplication. The goal is to learn a mapping function f_θ parameterized by θ such that $\hat{x} = f_\theta(y')$ approximates the fully-sampled image x as closely as possible, where y' is the zero-filled reconstruction obtained by applying the inverse Fourier transform to the undersampled k-space data.

4.2. Baseline U-Net Architecture

Our baseline approach uses a standard U-Net architecture, which has been widely adopted for medical image reconstruction tasks. The U-Net consists of an encoder path that captures context and a decoder path that enables precise localization, with skip connections between corresponding encoder and decoder layers to preserve spatial information.

The encoder is composed of a series of convolutional blocks followed by max-pooling operations. We use 4 pool layers for the downsampling stages, where the initial number of feature channels is set to 32, and this number doubles after each downsampling step. Each convolutional block consists of two 3×3 convolutional layers followed by InstanceNorm2d for feature normalization and a LeakyReLU activation function.

At the bottleneck, a convolution block further processes the features, doubling the channel count from the last layer. The decoder path symmetrically mirrors the encoder, using 4 max unpooling layers. Each stage begins with a 2×2 ConvTranspose2d layer to upsample the feature maps, halving the number of channels. The upsampled feature maps are then concatenated with the corresponding feature maps from the encoder path via skip connections.

After the final upsampling stage, a 1×1 convolutional layer maps the feature channels to a single output channel, producing the reconstructed grayscale MRI. This architecture effectively learns to remove undersampling artifacts while preserving anatomical details.

4.3. Transformer at Bottleneck (BT) Architecture

Building upon the baseline U-Net, we explored a hybrid architecture that incorporates transformer blocks at the bottleneck of the U-Net. This approach aims to leverage the global context modeling capabilities of transformers while maintaining the efficient local feature extraction of CNNs.

The architecture is structured as follows:

1. **U-Net Encoder:** Standard convolutional blocks and max-pooling layers process the input undersampled MR image, producing a bottleneck feature map $F_{enc} \in \mathbb{R}^{C \times H' \times W'}$.

2. **Tokenization & Positional Encoding:** F_{enc} is flattened into $N = H' \times W'$ tokens, each of dimension C. Learnable 1D positional embeddings are added to these tokens to retain spatial information before input to the Transformer.

3. **Transformer Encoder Block:** The sequence of tokens is processed by L_T standard Transformer encoder layers (e.g., $L_T = 2 - 6$). Each layer consists of Multi-Head Self-Attention (MHSA) using torch.nn.MultiheadAttention (e.g., $N_H = 4 - 8$ heads) and a position-wise Feed-Forward Network (FFN), with residual connections and layer normalization.

4. **De-Tokenization & U-Net Decoder:** The Transformer's output sequence is reshaped back to $F_{trans} \in \mathbb{R}^{C \times H' \times W'}$. This globally-aware feature map is then fed into the U-Net's decoder, which upsamples it and combines it with encoder features via skip connections to reconstruct the final image.

This hybrid approach allows the network to capture long-range dependencies at the bottleneck while maintaining the computational efficiency of the U-Net architecture for the majority of the processing.

4.4. SwinUNet Architecture

After initial experiments with the BT architecture, we identified the SwinUNet as a more promising approach. SwinUNet adapts the hierarchical Swin Transformer design to a U-Net-like encoder-decoder structure, offering several advantages for MRI reconstruction:

1. **Hierarchical Feature Representation:** The Swin Transformer's hierarchical design naturally aligns with the multi-scale feature extraction paradigm of U-Net.

2. **Shifted Window Attention:** Instead of global self-attention, which is computationally expensive, Swin Transformer uses shifted window-based self-attention. This approach computes self-attention within local windows and shifts the window partitioning between

consecutive layers, enabling connections across windows while maintaining computational efficiency.

3. **Linear Complexity:** The window-based attention mechanism reduces the computational complexity from quadratic to linear with respect to image size, making it feasible to process high-resolution medical images.

Our SwinUNet implementation consists of:

1. **Patch Embedding:** The input image is divided into non-overlapping patches and projected to a higher-dimensional feature space.

2. **Encoder:** A series of Swin Transformer blocks with patch merging layers that progressively reduce spatial resolution while increasing feature dimension.

3. **Bottleneck:** Swin Transformer blocks that process the most abstract features.

4. **Decoder:** A series of Swin Transformer blocks with patch expanding layers that progressively increase spatial resolution while decreasing feature dimension.

5. **Skip Connections:** Feature maps from the encoder are concatenated with corresponding decoder features to preserve spatial details.

6. **Output Projection:** The final feature map is projected back to the image space to produce the reconstructed MRI.

4.5. Training Strategy

All models were trained using the following strategy:

1. **Loss Function:** We used a combination of L1 loss and SSIM loss to optimize both pixel-wise accuracy and structural similarity: $\mathcal{L} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 (1 - \mathcal{L}_{SSIM})$ where λ_1 and λ_2 are weighting factors.

2. **Optimizer:** We employed the Adam optimizer with a learning rate ranging from 1e-5 to 1e-3, depending on the specific architecture.

3. **Learning Rate Scheduling:** A cosine annealing learning rate scheduler was used to gradually reduce the learning rate during training.

4. **Regularization:** To prevent overfitting, we applied weight decay (1e-4 to 1e-5) and dropout in transformer layers.

5. **Data Augmentation:** Random flips and rotations were applied to increase the effective size of the training dataset and improve model generalization.

6. **Batch Size:** We used batch sizes ranging from 4 to 16, depending on model complexity and available GPU memory.

7. **Training Duration:** Models were trained for 50-60 epochs, with early stopping based on validation loss to prevent overfitting.

5. Experiments

5.1. Experimental Setup

We conducted a series of experiments to evaluate and compare different architectural variants and hyperparameter configurations. All experiments were performed using PyTorch on NVIDIA GPUs. The models were evaluated using three metrics:

1. **Peak Signal-to-Noise Ratio (PSNR):** Measures the pixel-wise accuracy of the reconstruction.

2. **Structural Similarity Index (SSIM):** Assesses the preservation of structural information.

3. **Validation Loss:** The combined L1 and SSIM loss on the validation set.

5.2. Baseline U-Net Results

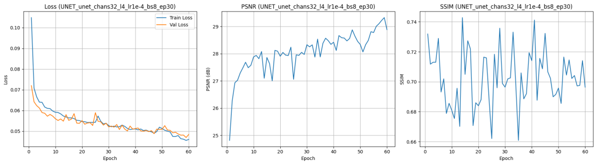

Figure 2. Training progression of the baseline U-Net model over 60 epochs, showing: (a) Training and Validation Loss curves (left); (b) Validation PSNR (center); and (c) Validation SSIM (right).

We first established a baseline using the standard U-Net architecture. Figure 2 shows the training progression of the baseline U-Net model over 60 epochs.

The learning curves show a steady decrease in both training and validation loss, with the validation loss closely tracking the training loss, indicating effective learning and good generalization without significant overfitting. The validation loss stabilized near 0.0526 by epoch 50.

Concurrently, validation PSNR increased from approximately 23 dB to a final value of 28.03 dB, while SSIM improved from around 0.62 to 0.6935. These trends confirm the model's ability to enhance reconstruction fidelity and structural similarity, with the best performance achieved at the final epoch.

5.3. SwinUNet Hyperparameter Tuning

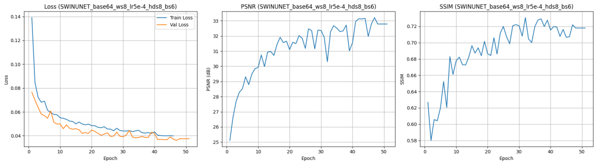

Figure 3. Training progression of the optimized SwinUNet model over 50 epochs, showing: (a) Training and Validation Loss curves (left); (b) Validation PSNR (center); and (c) Validation SSIM (right).

Based on the promising results of the SwinUNet architecture, we conducted extensive hyperparameter tuning to optimize its performance. We explored variations in:

1. **Base Feature Dimension:** We tested base feature dimensions of 64 and 80.

2. **Window Size:** We experimented with window sizes of 7 and 8 for the shifted window attention.

3. **Learning Rate:** We tried learning rates ranging from 5e-5 to 8e-5.

4. **Head Dimension:** We varied the dimension per attention head from 8 to 10.

5. **Batch Size:** We tested batch sizes of 4 and 6.

Figure 4 shows the training progression of our best-performing SwinUNet model.

The optimized SwinUNet model demonstrated more stable training dynamics compared to the baseline U-Net, with smoother learning curves and less fluctuation in validation metrics. The model achieved a final PSNR of 33.10 dB and SSIM of 0.7274, representing substantial improvements over the baseline.

5.4. Architecture Comparison

After establishing the baseline, we compared the performance of different architectural variants: the baseline U-Net, the Transformer at Bottleneck (BT) model, and the SwinUNet. Table 2 summarizes the results.

Architecture	Val. Loss	PSNR (dB)	SSIM
U-Net Baseline	0.0496	28.03	0.6935
BT-UNet	0.0412	29.87	0.7102
SwinUNet	0.0352	33.10	0.7274

Table 2. Performance comparison of different architectural variants.

The SwinUNet architecture significantly outperformed both the baseline U-Net and the BT-UNet, achieving approximately 4 dB higher PSNR and better SSIM values. This improvement can be attributed to the SwinUNet's ability to effectively capture both local and global image features through its hierarchical structure and shifted window attention mechanism.

5.5. Ablation Studies

To understand the contribution of different components and design choices, we conducted several ablation studies:

1. **Effect of Window Size:** Increasing the window size from 7 to 8 improved performance by allowing the model to capture slightly larger contextual regions in each attention operation.

2. **Impact of Base Feature Dimension:** Increasing the base feature dimension from 64 to 80 slightly decreased performance while significantly increasing computational requirements, suggesting that 64 provides a good balance between model capacity and efficiency.

3. **Learning Rate Sensitivity:** We found that the SwinUNet was more sensitive to learning rate than the baseline U-Net, with optimal performance achieved at a learning rate of 8e-5.

4. **Data Augmentation:** Removing data augmentation led to faster initial convergence but poorer generalization, confirming the importance of augmentation for preventing overfitting.

5.6. Model Efficiency

Model	Parameters (M)	Inference Time (ms)
U-Net Baseline	7.8	18.5
SwinUNet-64	27.3	42.7
SwinUNet-80	42.6	56.3

Table 3. Comparison of model size and inference time.

While the SwinUNet achieves superior reconstruction quality, it comes with increased computational requirements compared to the baseline U-Net. Table 3 compares the model sizes and inference times.

Despite the increased computational cost, the SwinUNet's inference time remains practical for clinical applications, where reconstruction quality is often prioritized over speed once a certain threshold of efficiency is met.

5.7. Qualitative Results

Beyond quantitative metrics, we visually assessed the reconstruction quality of different models. Figure 4 shows example reconstructions from the validation set.

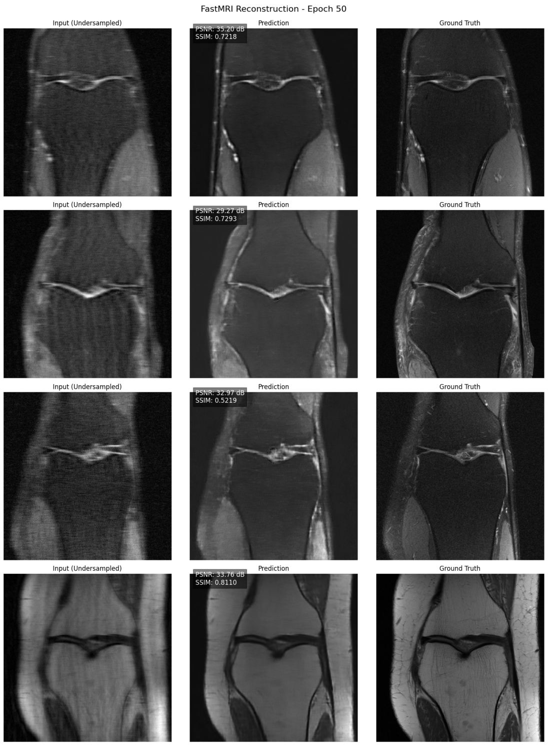

Figure 4. Visual results of the optimized SwinUNet model over 50 epochs, showing: (a) Raw Under-sampled Data(left); (b) SwinUNet reconstruction (center); and (c) Ground Truth Fully-sampled Data (right).

The SwinUNet reconstructions demonstrate superior preservation of fine anatomical details compared to the baseline U-Net. Particularly notable is the improved definition of cartilage boundaries, meniscal structures, and bone margins. The SwinUNet also more effectively removes the coherent streaking artifacts that are characteristic of undersampled MRI reconstruction, resulting in images that more closely resemble the fully-sampled ground truth.

6. Conclusion

In this paper, we presented a comprehensive study of deep learning architectures for accelerated MRI reconstruction, with a focus on integrating transformer components to enhance reconstruction quality. Our experiments demonstrate that the SwinUNet architecture significantly outperforms the baseline U-Net, achieving approximately 4 dB higher PSNR and more stable SSIM values.

The superior performance of SwinUNet can be attributed to its ability to effectively model both local and global image features through its hierarchical structure and shifted window attention mechanism. This enables better removal of undersampling artifacts while preserving fine anatomical details that are crucial for diagnostic purposes.

Our work contributes to the growing body of evidence supporting the effectiveness of transformer-based architectures for medical image analysis tasks. By systematically comparing different architectural variants and conducting extensive hyperparameter tuning, we provide valuable insights for researchers and practitioners working on accelerated MRI reconstruction.

6.1. Limitations and Future Work

Despite the promising results, our study has several limitations that point to directions for future work:

1. **Single-Coil Focus:** We focused exclusively on single-coil MRI reconstruction. Extending our approach to multi-coil data would be a natural next step.

2. **Fixed Acceleration Factor:** We used a fixed acceleration factor of 4. Future work could explore the model's performance across different acceleration factors and sampling patterns.

3. **Computational Efficiency:** While the SwinUNet achieves superior reconstruction quality, its computational requirements are higher than the baseline U-Net. Further optimization of the architecture for improved efficiency would be valuable.

4. **Clinical Validation:** Although we used standard quantitative metrics for evaluation, clinical validation with radiologists would provide more insight into the diagnostic value of the reconstructed images.

In future work, we plan to address these limitations and explore additional architectural innovations, such as integrating frequency-domain learning and incorporating uncertainty estimation to provide confidence measures for the reconstructions.

References

[1] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

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

[3] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase. Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.

[4] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

[5] J. Huang, X. Xing, Z. Gao, and G. Yang. Swin deformable attention u-net transformer (sdaut) for explainable fast mri. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, pages 538–548. Springer, 2022.

[6] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, and J. K. Seo. Deep learning for undersampled mri reconstruction. *Physics in Medicine & Biology*, 63(13):135007, 2018.

[7] D. Lee, J. Yoo, and J. C. Ye. Deep residual learning for accelerated mri using magnitude and phase networks. *IEEE Transactions on Biomedical Engineering*, 65(9):1985–1995, 2018.

[8] K. Lin and R. Heckel. Vision transformers enable fast and robust accelerated mri. *Proceedings of Machine Learning Research*, 172:1–22, 2022.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[10] M. Lustig, D. Donoho, and J. M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[11] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999.

[12] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2017.

[13] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.