HYCOSWIN: A HYBRID TRANSFORMER CONVNET ARCHITECTURE FOR ENHANCED SUB SAMPLED MRI RECONSTRUCTION

Sai Pavan Tadem, Sutanu Bera, Debashis Sen, Subhamoy Mandal, Prabir Kumar Biswas*

Indian Institute of Technology, Kharagpur, India

*sutanu.bera@iitkgp.ac.in

ABSTRACT

In this study, we propose HyCoSwin, a novel encoder-decoder network for subsampled MRI reconstruction. The encoder part of HyCoSwin features Swin transformer blocks, while the decoder incorporates custom convolutional modules (ConvFuUnit and CConvFuUnit) designed for this task. Evaluated on the FastMRI dataset, HyCoSwin's compact design enables consistent outperformance of the existing state-of-the-art, with significant gains in low-data scenarios. This demonstrates HyCoSwin's effectiveness for subsampled MRI reconstruction, particularly in low-data environments typical of medical imaging.

1. INTRODUCTION

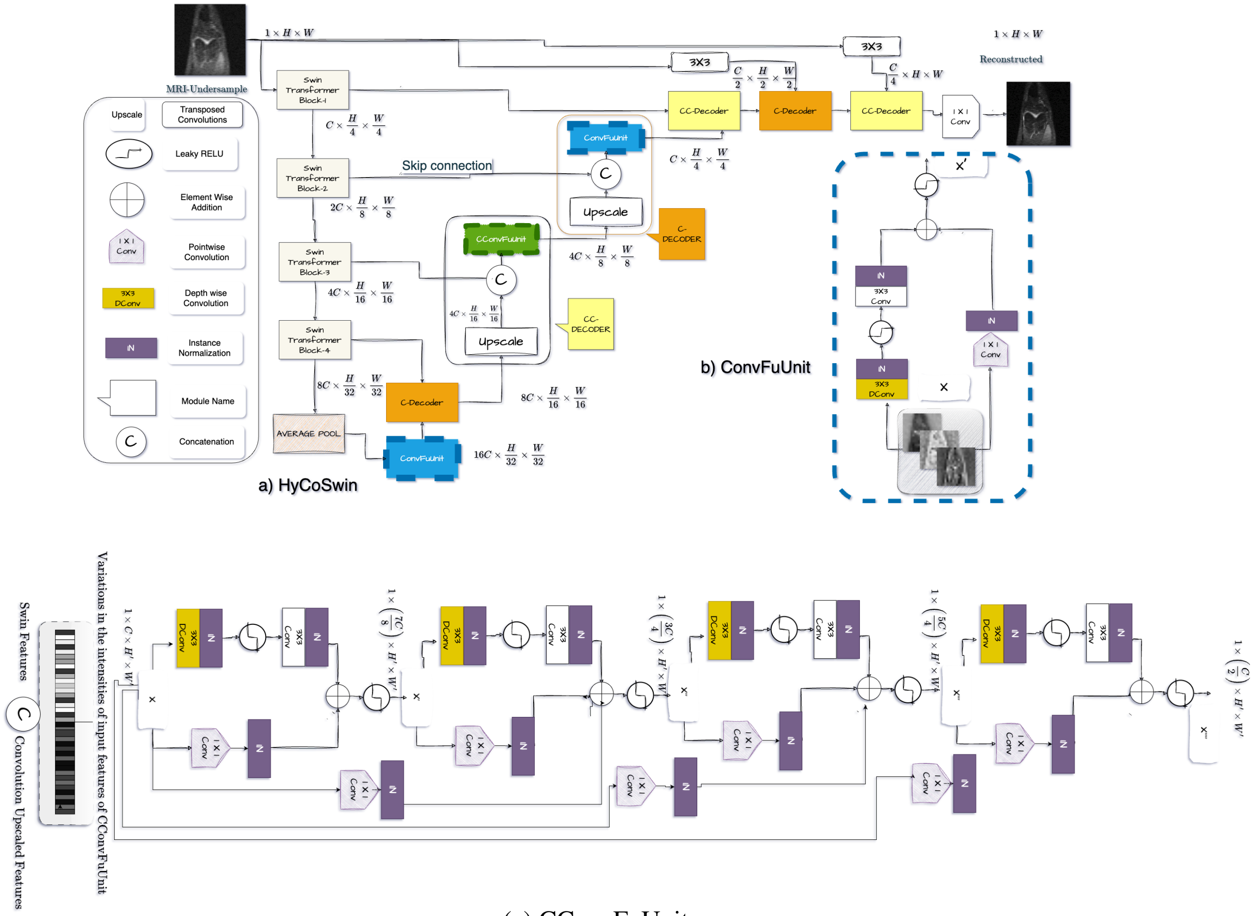
Magnetic Resonance Imaging (MRI) is a powerful, non-invasive imaging technique[1], but long scan times remain a significant drawback[2, 3]. To address this, subsampled MRI reconstruction algorithms are essential for accelerating scans while maintaining image quality. Convolutional neural networks (CNNs) have emerged as the leading method for this task[4, 5, 6, 7, 8, 9], owing to their capacity to effectively capture spatial patterns and reconstruct high-quality images from limited data. However, CNN-based approaches often struggle to model long-range dependencies and global context, which are crucial for accurately reconstructing highly subsampled MR images.

In recent years, transformer-based architectures have gained significant attention in sub-sampled MRI reconstruction[10, 11, 12], as their self-attention mechanisms allow them to capture both local and global features. Although transformer models address the limitations of CNNs such as restricted receptive fields and limited adaptability to input content, their computational complexity scales quadratically with spatial resolution. This makes them impractical for high-resolution image restoration tasks like subsampled MRI reconstruction. Additionally, transformers require large datasets for effective training, and in cases where sufficient data is unavailable, their performance often falls short. In fact, without extensive training data, transformer-based models can sometimes underperform compared to CNN-based methods. Given these limitations, recent research has shifted towards developing compact architectures that combine transformer blocks with CNN layers to improve performance on low-level image processing tasks such as image restoration[13, 14] and segmentation[15, 16]. However, there is limited literature focused on designing compact networks specifically for subsampled MRI reconstruction. To address this gap, we propose HyCoSwin, a novel architecture that combines specially designed CNN modules with transformer blocks to address the unique challenges of this task.

The proposed HyCoSwin integrates transformer[17] and convolutional models to enhance subsampled MRI reconstruction. By leveraging a transformer for encoding and convolutional layers for decoding, it effectively combines global and local feature extraction. However, straightforward integration of transformer outputs with conventional convolutional methods may prove inadequate, where the convolutional blocks are not designed to exploit the features appropriately. To address this, we introduce two innovative modules: the Convolutional Fusion Unit (ConvFuUnit) and the Complex Convolutional Fusion Unit (CConvFuUnit), which optimize the fusion of transformer and convolutional features for superior reconstruction.

We evaluated the proposed method on the publicly available FastMRI dataset for subsampled MRI reconstruction[4]. Comparisons included state-of-the-art models such as CNN-based UNet, Transformer-based SwinUNet[18], Uformer[14], and the recently proposed CNN-Transformer hybrid model Restormer[13]. Results show that HyCoSwin outperforms these existing approaches. Furthermore, an additional experiment was conducted to assess the performance with limited training data, using only 25% and 50% of the total data for training. Our method demonstrated consistent performance even with reduced training data, while other methods struggled to perform reliably in these low-data scenarios.

We discuss our proposed method in Section 2, provide experimental results and analysis in Sections 3 and 4, and conclude in Section 5. The source code for our project is publicly available at: `https://github.com/SaiPavan-Tadem/HyCoSwin`. (Will be released upon acceptance)



MRI-Undersample
$1 \times H \times W$
Swin Transformer Block-1
Swin Transformer Block-2
Swin Transformer Block-3
Swin Transformer Block-4
$C \times \frac{H}{4} \times \frac{W}{4}$
$2C \times \frac{H}{8} \times \frac{W}{8}$
$4C \times \frac{H}{16} \times \frac{W}{16}$
$8C \times \frac{H}{32} \times \frac{W}{32}$
$16C \times \frac{H}{32} \times \frac{W}{32}$
$8C \times \frac{H}{16} \times \frac{W}{16}$
$4C \times \frac{H}{8} \times \frac{W}{8}$
$C \times \frac{H}{4} \times \frac{W}{4}$
$\frac{C}{2} \times \frac{H}{2} \times \frac{W}{2}$
$\frac{C}{4} \times H \times W$
Skip connection
AVERAGE POOL
ConvFuUnit
CConvFuUnit
C-Decoder
CC-Decoder
C-DECODER
CC-DECODER
Upscale
C
3X3
1 X 1 Conv
$1 \times H \times W$
Reconstructed
Upscale
Transposed Convolutions
Leaky RELU
Element Wise Addition
Pointwise Convolution
3X3 DConv
Depth wise Convolution
IN
Instance Normalization
Module Name
Concatenation
a) HyCoSwin
b) ConvFuUnit
X′
X
IN
3X3 Conv
3X3 DConv
1 X 1 Conv
Swin Features
Convolution Upscaled Features
Variations in the intensities of input features of CConvFuUnit
$1 \times C \times H' \times W'$
$1 \times \left(\frac{7C}{8}\right) \times H' \times W'$
$1 \times \left(\frac{3C}{4}\right) \times H' \times W'$
$1 \times \left(\frac{5C}{4}\right) \times H' \times W'$
$1 \times \left(\frac{C}{2}\right) \times H' \times W'$
x
x′
x″
x‴
x⁗

(c) CConvFuUnit

Fig. 1: (a) The proposed HyCoSwin model, consisting of Swin transformers in the encoder and CConvFuUnit and ConvFuUnit blocks in the decoder. Detailed views of the (b) ConvFuUnit block and (c) the CConvFuUnit block.

2. THE PROPOSED METHOD

HyCoSwin aims to combine the strengths of transformers and convolutional architectures in a unified model for improved sub sampled MRI reconstruction. By using a transformer as the encoder and a convolutional layers as the decoder, the model balances global feature extraction with local processing. However, directly using transformer features with simple scaling in standard convolutional architectures may be insufficient for optimal reconstruction. To address this, we propose two novel architectures: the Convolutional Fusion Unit (ConvFuUnit) and the Complex Convolutional Fusion Unit (CConvFuUnit). These units are designed to facilitate efficient fusion of transformer and convolutional features. In the following sections, we have discussed the design pipeline and provide an in-depth explanation of both the ConvFuUnit and CConvFuUnit.

Overall Pipeline. As shown in Figure 1, the input to HyCoSwin is an undersampled MR image $I \in \mathbb{R}^{H \times W \times 1}$, which is processed by a Swin Transformer-based Encoder followed by a Decoder. The input first passes through a patch merging layer that divides the image into 4×4 patches, downsampling by a factor of 4 and producing $I_o \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 16}$. The raw pixel values of each patch are concatenated channel-wise and projected to a latent embedding of size C using a linear layer. This output is then passed to a Swin Transformer block. Following this, three additional patch merging and Swin Transformer blocks are applied, where the patch merging layer has a downsampling factor of 2. The output of each Swin Transformer block is connected to its corresponding decoder layer via skip connections. Additionally, a 2-layer convolutional module extracts low-level features from the input image, which are forwarded to the final two decoder layer to enhance preservation of low-level details. The encoder's output is passed to a bottleneck layer, implemented using the ConvFuUnit. The decoder in HyCoSwin alternates between ConvFuUnit and CConvFuUnit, as illustrated in Figure 1 (a). Each ConvFuUnit and CConvFuUnit is preceded

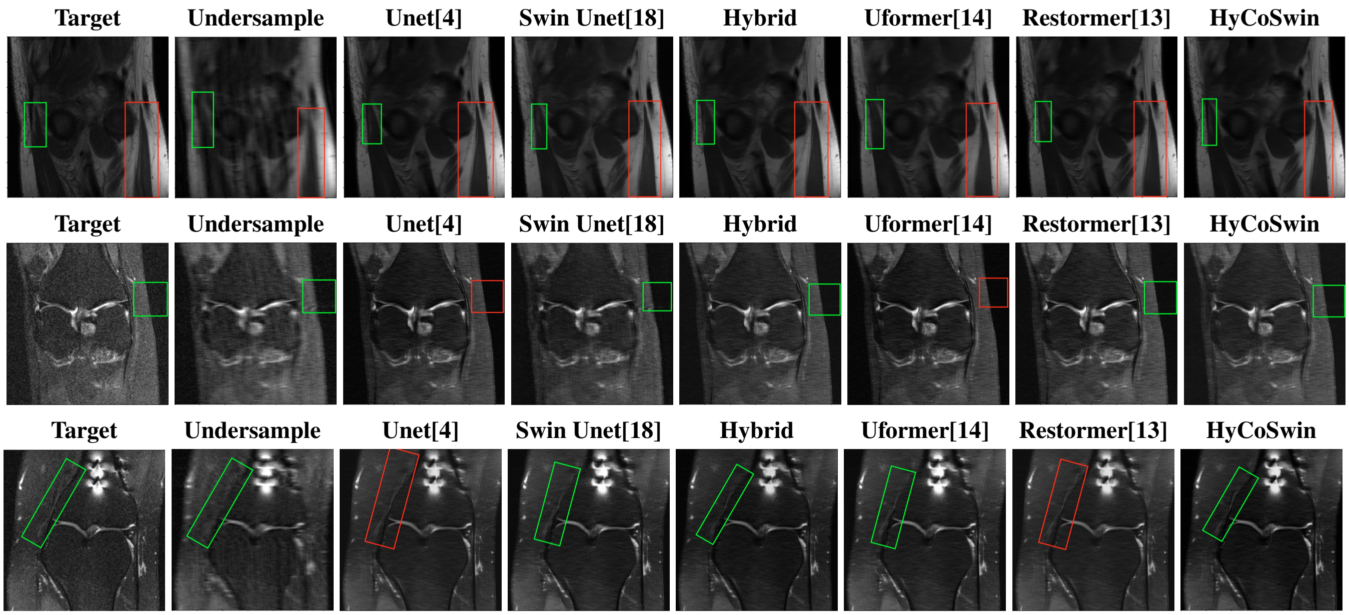


Fig. 2: Visual comparisons with state-of-the-art methods on PD(Top Row), PDFS(Middle Row), and PD+PDFS(Bottom Row)(full data) knee datasets. Note that the baseline **"Hybrid"** model uses a Swin Transformer encoder with a standard CNN-based decoder, whereas our **"HyCoSwin"** incorporates Swin Transformer encoder with *ConvFuUnit* and *CConvFuUnit* components. HyCoSwin presents superior reconstruction quality on the fastMRI dataset. The visuals demonstrate that in the PD case, the model effectively captures texture (green box) and structural details (red box) that are entirely absent in the undersampled data. In the PDFS case, the U-Net produce a oversmooth reconstruction (highlighted in the red box), whereas the other transformer-based designs, except Uformer, manage to reproduce texture. In the PDPDFS case, both U-Net and Restormer extend the texture beyond the actual, as shown in the red box, while the other models maintain it, as indicated by the green box.

by an upsampling layer implemented using transpose convolution. The details about ConvFuUnit and CConvFuUnit is given below.

2.1. ConvFuUnit - Convolutional Fusion Unit

Our primary objective with the ConvFuUnit is to reduce the computational burden of the encoder in the HyCoSwin network while enhancing long-range feature interactions. Instead of relying on computationally intensive convolutions with large kernel sizes, we introduce a novel Convolutional Fusion Unit designed to achieve extended feature interactions efficiently. This unit consists of two parallel sub-paths: the first path extracts channel-wise features using a 1×1 convolutional layer, followed by Instance Normalization (IN), while the second path begins with spatial feature extraction via a 3×3 depthwise separable convolution, followed by IN, a conventional 3×3 convolution, and another IN layer. The outputs from both paths are concatenated and processed through a leaky ReLU activation function. By combining 1×1 and 3×3 depthwise convolutions, we reduce computational requirements and, through the use of two consecutive 3×3 convolutions, achieve superior long-range interactions compared to traditional convolutional methods. A visual illustration of ConvFuUnit is given in Figure 1 (b).

2.2. CConvFuUnit - Complex Convolutional Fusion Unit

The function of the CConvFuUnit is to integrate skip connection features from the encoder with upsampled features from the previous layer. To achieve optimal performance, a multistep fusion scheme is introduced. Given the large number of channels in transformer-derived features, an all-at-once channel and feature fusion approach is avoided. Instead, a multistep fusion process is employed, utilizing a series of depthwise separable convolutions and conventional convolutions for spatial feature fusion, alongside a parallel 1×1 convolution for channel-wise fusion. Detailed layer-wise information about the CConvFuUnit can be found in Figure 1 (c). The core block is similar to the ConvFuUnit module, providing a trade-off between optimal performance and computational efficiency.

3. EXPERIMENTAL DETAILS

We evaluated our method on the FastMRI[4] knee dataset, which comprises Proton Density (PD) and Proton Density Fat-Suppressed (PDFS) data. We compared our approach with several baseline methods, including U-Net[4], Swin-UNet[19], Uformer[14], and Restormer[13]. Additionally, we considered a baseline variant (Hybrid) of our proposed Hy-

Method	PD Data			PDFS Data			PD+PDFS Data		
	PSNR ↑	SSIM ↑	NMSE ↓	PSNR ↑	SSIM ↑	NMSE ↓	PSNR ↑	SSIM ↑	NMSE ↓
Zero filled	32.17329	0.76119	0.03973	29.38531	0.61454	0.08882	30.61113	0.67902	0.06724
UNET[4]	35.31228	0.84180	0.02051	30.88685	0.66595	0.05896	32.85532	0.74342	0.04196
SWINUNET[18]	35.11608	0.83785	0.02101	30.67839	0.65892	0.06095	32.73314	0.74129	0.04247
Uformer[14]	34.75781	0.81491	0.02214	30.76356	0.63966	0.06050	32.85075	0.72327	0.04227
Restomer[13]	35.36035	0.84179	0.02043	30.81592	0.66231	0.05883	32.99089	0.75081	0.03970
Hybrid	35.28752	0.84128	0.02066	30.89446	0.66635	0.05889	32.81662	0.74282	0.04212
HyCoSwin	35.42413	0.84348	0.02020	30.91861	0.66702	0.05868	32.92233	0.74502	0.04164

Table 1: Quantitative results on the single coil fastMRI dataset, including PD, PDFS, and PD+PDFS, are shown. The input undersampled image sequences were generated by randomly undersampling the k-space data using a Cartesian undersampling function, targeting a 4x acceleration. HyCoSwin showed best performance in the PD and PDFS, the cases with lower data than PD+PDFS.

PD dataset:	25% data			50% data		
Model	PSNR ↑	SSIM ↑	NMSE ↓	PSNR ↑	SSIM ↑	NMSE ↓
Zero Filled	32.17329	0.76119	0.03973	32.17329	0.76119	0.03973
Unet[4]	35.21072	0.833738	0.020820	35.27622	0.839379	0.020707
Uformer[14]	34.45981	0.801371	0.023118	34.69625	0.811321	0.021920
Restormer[13]	35.20935	0.838320	0.021428	35.31879	0.840423	0.020956
HyCoSwin	35.30013	0.840119	0.021113	35.38701	0.841722	0.020601

Table 2: Performance comparison of models, including HyCoSwin, is presented on 25% and 50% of the randomly sampled PD dataset.

CoSwin, where we substituted the ConvFuUnit and CConvFuUnit with traditional CNN layers. Our experiments employed the Adam optimizer with a learning rate of 0.0001 and weight decay of 0.1, with Mean Absolute Error (L1 loss) as the training loss function.

4. RESULTS AND ANALYSIS

Table 1 presents a detailed comparison of the proposed HyCoSwin model against other related approaches. As seen in the results, HyCoSwin significantly surpasses the performance of the competing methods across most metrics. Notably, when compared to our baseline Hybrid model, our HyCoSwin achieved a 0.14 dB increase in PSNR, demonstrating the clear advantage of the proposed Convolutional Fusion Unit (ConvFuUnit) and Complex Convolutional Fusion Unit (CConvFuUnit) over the conventional ConvUnit.

Figure 2 further illustrates the visual superiority of HyCoSwin, particularly in the highlighted green boxes. In these regions, HyCoSwin successfully restores missing features from undersampled data, a challenge that both CNN and transformer-based methods struggle to address. In contrast, competing methods fail to recover these fine details, leading to suboptimal reconstructions. Moreover, the red box highlights HyCoSwin's superior preservation of structural information, particularly when compared to the Hybrid model and Uformer, where noticeable degradation occurs.

To evaluate the robustness of each method in low-data regimes, we conducted an experiment using only 25% and 50% of the available training data. The results, summarized in Table 2, reveal that HyCoSwin consistently outperforms both UNet and Restormer under these conditions. In particular, HyCoSwin experienced only a 0.04 dB and 0.12 dB PSNR drop when trained on 50% and 25% of the data, respectively. It is especially evident that Uformer and SwinUNet suffer significant performance drops in this scenario, underscoring the limitations of purely transformer-based architectures when faced with restricted training data. This suggests that HyCoSwin's hybrid architecture, which combines the strengths of CNNs and transformers, is more resilient and effective in low-data environments.

5. CONCLUSION

In this paper, we propose a HyCoSwin model designed with novel architectures for MRI reconstruction, incorporating the new Convolutional Fusion Unit (ConvFuUnit) and Complex Convolutional Fusion Unit (CConvFuUnit) to handle datasets of various sizes. We evaluated the HyCoSwin model on the fastMRI Knee dataset and compared it with other methods, demonstrating its stability and robustness across both small and large-scale data scenarios.

6. REFERENCES

[1] Edwin JR van Beek, Christiane Kuhl, Yoshimi Anzai, Patricia Desmond, Richard L Ehman, Qiyong Gong, Garry Gold, Vikas Gulani, Margaret Hall-Craggs, Tim Leiner, et al., “Value of mri in medicine: More than just another test?,” *Journal of Magnetic Resonance Imaging*, vol. 49, no. 7, pp. e14–e25, 2019.

[2] Anni Copeland, Eero Silver, Riikka Korja, Satu J. Lehtola, Harri Merisaari, Ekaterina Saukko, Susanne Sinisalo, Jani Saunavaara, Tuire Lähdesmäki, Riitta Parkkola, Saara Nolvi, Linnea Karlsson, Hasse Karlsson, and Jetro J. Tuulari, “Infant and child mri: A review of scanning procedures,” *Frontiers in Neuroscience*, vol. 15, 2021.

[3] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al., “fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning,” *Radiology: Artificial Intelligence*, vol. 2, no. 1, pp. e190007, 2020.

[4] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui, “fastmri: An open dataset and benchmarks for accelerated mri,” 2019.

[5] Maosong Ran, Wenjun Xia, Yongqiang Huang, Zexin Lu, Peng Bao, Yan Liu, Huaiqiang Sun, Jiliu Zhou, and Yi Zhang, “Md-recon-net: A parallel dual-domain convolutional neural network for compressed sensing mri,” 2020.

[6] Zaccharie Ramzi, Philippe Ciuciu, and Jean-Luc Starck, “Xpdnet for mri reconstruction: an application to the 2020 fastmri challenge,” 2021.

[7] Matthew J. Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll, “Results of the 2020 fastmri challenge for machine learning mr image reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2306–2317, 2021.

[8] Xiaohan Liu, Yanwei Pang, Ruiqi Jin, Yu Liu, and Zhenchang Wang, “Dual-domain reconstruction network with v-net and k-net for fast mri,” *Magnetic Resonance in Medicine*, vol. 88, no. 6, pp. 2694–2708, 2022.

[9] Yu Liu, Yanwei Pang, Xiaohan Liu, Yiming Liu, and Jing Nie, “Diik-net: A full-resolution cross-domain deep interaction convolutional neural network for mr image reconstruction,” *Neurocomputing*, vol. 517, pp. 213–222, 2023.

[10] Jiahao Huang, Yingying Fang, Yinzhe Wu, Huanjun Wu, Zhifan Gao, Yang Li, Javier Del Ser, Jun Xia, and Guang Yang, “Swin transformer for fast mri,” *Neurocomputing*, vol. 493, pp. 281–304, 2022.

[11] Xiang Zhao, Tiejun Yang, Bingjie Li, and Xin Zhang, “Swingan: A dual-domain swin transformer-based generative adversarial network for mri reconstruction,” *Computers in Biology and Medicine*, vol. 153, pp. 106513, 2023.

[12] Pengfei Guo, Yiqun Mei, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel, “Reconformer: Accelerated mri reconstruction using recurrent transformer,” *IEEE transactions on medical imaging*, 2023.

[13] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, “Restormer: Efficient transformer for high-resolution image restoration,” 2022.

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

[15] Hui Tang, Yuanbin Chen, Tao Wang, Yuanbo Zhou, Longxuan Zhao, Qinquan Gao, Min Du, Tao Tan, Xinlin Zhang, and Tong Tong, “Htc-net: A hybrid cnn-transformer framework for medical image segmentation,” *Biomedical Signal Processing and Control*, vol. 88, pp. 105605, 2024.

[16] Mo Zhao, Gang Cao, Xianglin Huang, and Lifang Yang, “Hybrid transformer-cnn for real image denoising,” *IEEE Signal Processing Letters*, vol. 29, pp. 1252–1256, 2022.

[17] 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,” 2021.

[18] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation,” in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[19] FirstName Alpher, “Frobnication,” *IEEE TPAMI*, vol. 12, no. 1, pp. 234–778, 2002.