

# DSFormer: A Dual-domain Self-supervised Transformer for Accelerated Multi-contrast MRI Reconstruction

Bo Zhou<sup>1,3</sup> Neel Dey<sup>2,3</sup> Jo Schlemper<sup>3</sup> Seyed Sadegh Mohseni Salehi<sup>3</sup>  
Chi Liu<sup>1</sup> James S. Duncan<sup>1</sup> Michal Sofka<sup>3</sup>  
<sup>1</sup>Yale University      <sup>2</sup>New York University      <sup>3</sup>Hyperfine Research

## Abstract

*Multi-contrast MRI (MC-MRI) captures multiple complementary imaging modalities to aid in radiological decision-making. Given the need for lowering the time cost of multiple acquisitions, current deep accelerated MRI reconstruction networks focus on exploiting the redundancy between multiple contrasts. However, existing works are largely supervised with paired data and/or prohibitively expensive fully-sampled MRI sequences. Further, reconstruction networks typically rely on convolutional architectures which are limited in their capacity to model long-range interactions and may lead to suboptimal recovery of fine anatomical detail. To these ends, we present a dual-domain self-supervised transformer (DSFormer) for accelerated MC-MRI reconstruction. DSFormer develops a deep conditional cascade transformer (DCCT) consisting of cascaded Swin transformer reconstruction networks (SwinRN) trained under two deep conditioning strategies to enable MC-MRI information sharing. We further use a dual-domain (image and k-space) self-supervised learning strategy for DCCT to alleviate the costs of acquiring fully sampled training data. DSFormer generates high-fidelity reconstructions which outperform current fully-supervised baselines and approach the performance of full supervision.*

## 1. Introduction

Diagnosticians often capture a series of multi-contrast magnetic resonance images (MC-MRI) of a single subject to acquire complementary tissue information towards more accurate and comprehensive radiological evaluation [21, 2]. However, due to physical constraints, MRI intrinsically requires prolonged acquisition which often leads to patient discomfort and the accumulation of motion artifacts and system imperfections in the image that obfuscate biomedically-relevant anatomical detail. These limitations have led to immense interest in accelerated methods that can reconstruct high-fidelity and artifact-free images from fewer (undersampled) frequency-domain (*k-space*)

MRI measurements and reduced scan time.

While the inverse Fourier transform can reconstruct images from fewer *k-space* measurements, it comes at the cost of strong aliasing and blurring effects in the reconstruction and has thus motivated works which exploit transform-domain data priors to achieve higher quality reconstructions with fewer artifacts [20, 6, 22]. However, these methods may still yield blurred and sub-clinical reconstructions and are generally slow and hyperparameter-sensitive as they are based on iterative *instance-specific* optimization. More recently, deep MRI reconstruction networks have greatly improved MRI reconstruction fidelity under high undersampling rates with prediction times on the order of seconds [32, 27, 24, 8, 1, 7, 43, 38, 3, 40, 31, 37, 34, 28, 42, 25].

However, these works typically achieve their strong results via supervised training on ground-truth *fully-sampled* images and/or *k-space* target data, which is often practically infeasible in both time and cost to acquire. Recently, *self-supervised* reconstruction frameworks have emerged requiring only undersampled *k-space* data [31, 37, 10], yet their performance remains upper-bounded by full supervision. Further, whether supervised or self-supervised, the aforementioned works largely focus on single contrast MRI acceleration, whereas most diagnostics require MC-MRI to visualize disparate anatomical characteristics. Fortunately, in MC-MRI reconstruction, fully sampled MRI modalities requiring shorter acquisitions can be used as a reference to guide target modalities that require longer acquisitions via methods which inject a fully-sampled reference modality as an extra input channel into a reconstruction network [35, 29, 5, 42, 17, 18].

While these MC-MRI methods achieve excellent reconstructions, they have the following major limitations. First, previous MC-MRI reconstruction methods operate directly on the undersampled target MRI image as input (reconstructed via zero-padding and the inverse Fourier transform) and thus suffer from severe aliasing in their starting point. Second, current MC-MRI reconstruction networks ubiquitously employ convolutional architectures [17, 35, 5, 26], such as U-shaped network designs [26] and sequential con-

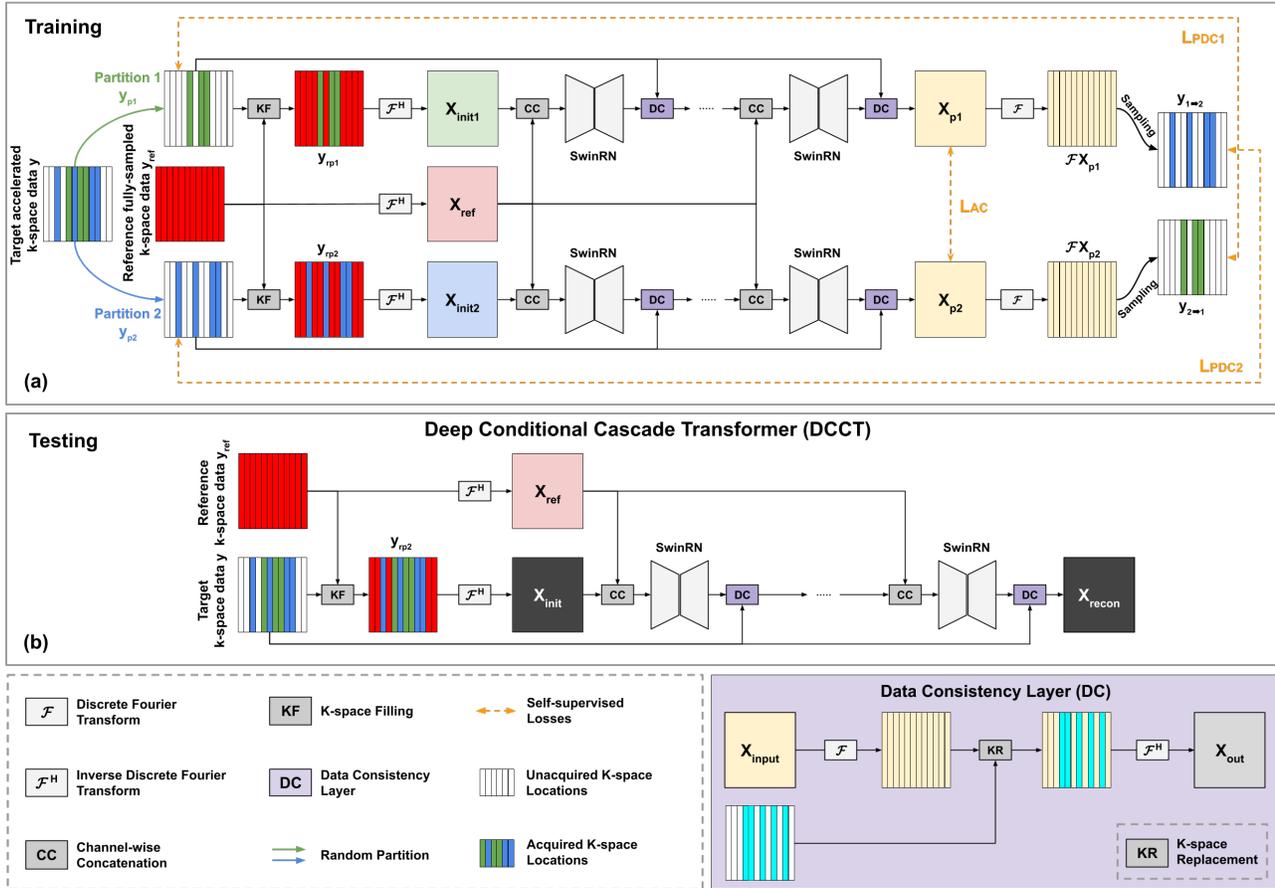


Figure 1. An overview of the **Dual-domain Self-supervised TransFormer (DSFormer)**. The Deep Conditional Cascade Transformer (DCCT) is trained in a self-supervised fashion with randomly partitioned undersampled k-space data sets fed into DCCT in parallel. The partition data consistency loss ( $\mathcal{L}_{PDC}$ ) and appearance consistency loss ( $\mathcal{L}_{AC}$ ) are used for self-supervised training. At testing, arbitrarily undersampled data is reconstructed by DCCT. SwinRN (Fig. 3) is used as the backbone network for DCCT.

volitional layers with residual connections [17, 27], both of which are limited in modeling long-range interactions and may recover reduced fine image detail due to the lack of *non-local* contextual information. Third, existing MC-MRI reconstruction methods require fully-supervised and fully-sampled training data from large-scale paired data, which is prohibitively expensive to obtain. The fully-sampled data of target contrasts demanding longer acquisitions are also prone to motion and other accumulating errors. Therefore, self-supervised learning operating on undersampled data with shorter acquisitions would be less susceptible to non-ideal imaging conditions.

To these ends, we present DSFormer, a dual-domain self-supervised transformer for accelerated MC-MRI reconstruction, with the following contributions:

1. *Multi-contrast information sharing.* We develop a deep MC-MRI conditioning method for efficient usage of multi-contrast information in MC-MRI reconstruction. Briefly, as opposed to the zero-padded

and aliased initial reconstruction used in most works, our framework leverages fully-sampled *reference* MRI data by grafting its k-space data into the unacquired k-space bins of the undersampled/accelerated *target* modality, whose inversion provides a sharp, de-aliased, and anatomically-correct starting point for the network to operate on (Figure 2). To further reinforce reference information in the undersampled reconstruction, we also channel-wise concatenate the reference MRI alongside the network inputs.

2. *Vision Transformers for MRI reconstruction.* Inspired by recent advances in vision transformers showing improved image restoration over CNNs [33, 4, 16, 19] by using non-local processing to recover fine detail, we develop a Swin transformer Reconstruction Network (SwinRN) to be used as a backbone in a cascaded framework. By combining MC-MRI conditioning with SwinRN-cascades, we propose a Deep Condi-

tional Cascade Transformer (DCCT) for high-fidelity MRI reconstruction.

3. *Dual-domain self-supervised learning.* To train DCCT in a self-supervised fashion using only undersampled target MRI data, we further use a dual image and k-space domain self-supervised learning approach, achieving reconstruction quality comparable to fully supervised training.

Extensive experiments on MC-MRI data with different acceleration protocols demonstrate that DSFormer trained with either full supervision or only self-supervision generates superior reconstructions over previous architectures and conditioning mechanisms with fully supervised training strategies.

## 2. Related work

**Fully-supervised MRI reconstruction.** Convolutional neural networks (CNNs) have been extensively studied to reconstruct images from undersampled k-space data. For example, Wang *et al.* [32] recover fully-sampled MRIs from undersampled acquisitions using supervised CNN training on paired data. Schlemper *et al.* [27, 24] develop a deep cascade of CNNs with intermediate data consistency layers which ensure that the originally-sampled k-space in the input is consistent with the reconstruction. Hammernik *et al.* [8] develop variational networks to solve reconstruction optimization using gradient descent with CNNs. Similarly, Aggarwal *et al.* [1] use a conjugate gradient algorithm within the reconstruction network.

In addition to methods operating in the image domain, dual image and k-space methods have also been explored. Eo *et al.* [7] add an additional k-space reconstruction network to [27] to enable cross-domain MRI reconstruction. Similarly, Singh *et al.* [28] show that combining frequency and image feature representation learning using two-task-independent layers can improve reconstruction performance over single-domain methods. Zhu *et al.* [43] directly map the undersampled k-space data to its image reconstruction using manifold learning. Moreover, reinforcement learning-aided reconstruction networks were also found to improve the reconstruction quality [34, 23]. While achieving promising performance, these methods require fully supervised training data from large-scale paired undersampled and fully sampled k-space scans [38, 3, 14]. Moreover, these methods only focus on single-contrast MRI reconstruction instead of MC-MRI reconstruction.

**Self-supervised MRI reconstruction.** Recently, self-supervised reconstruction methods requiring only undersampled k-space data have been proposed for single-contrast MRI reconstruction. HQS-Net [31] decouples the minimization of the data consistency term and

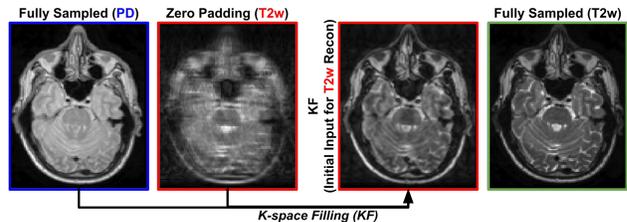


Figure 2. K-space filling (KF) conditioning for  $\times 4$  accelerated T2w reconstruction. Undersampled T2w data is combined with fully sampled PD data via KF to generate an initial DCCT input generating much fewer artifacts as compared to zero-padding.

regularization term in [27] based on a neural network, such that network training relies only on undersampled measurements. Yaman *et al.* [37] proposed a physically-guided self-supervised learning method that trains the deep cascade reconstruction network [27] by predicting one undersampled k-space data partition using the other data partition, with a similar approach used in Yaman *et al.* [36] for subject-specific zero-shot MRI reconstruction. Concurrently to our work, Korkmaz *et al.* [15] propose a self-supervised transformer-GAN for zero-shot *instance-specific* optimization and is not comparable to this submission as it focuses on latent noise-to-image GAN mapping and needs to be trained on each new input slice. Hu *et al.* [10] also propose to use ISTA-Net [39] with a parallel training framework for self-supervised single-contrast MRI reconstruction. Furthermore, Zhou *et al.* [41] devise a triple branch-based dual-domain self-supervised reconstruction framework, achieving promising performance on single-contrast low-field MRI. However, to our knowledge, self-supervised *multi-contrast* MRI reconstruction remains unexplored and is the subject of this work.

**MC-MRI reconstruction.** Currently, there are few deep learning-based fast MC-MRI reconstruction methods [42, 35, 29, 5, 17, 18]. Xiang *et al.* [35] use fully sampled T1w images as an additional CNN channel input to facilitate accelerated T2w reconstructions. Similarly, Dar *et al.* [5] add adversarial learning and a perceptual loss [12] to further improve performance. More recently, Liu *et al.* [17] and Zhou *et al.* [42] feed the fully sampled reference data as an additional channel input into a deep cascade network [27]. Similar strategies have also been proposed for variational reconstruction [8, 18].

## 3. Methods and Materials

The overall DSFormer pipeline is illustrated in Figure 1 and consists of two major parts: (1) the deep conditional cascade transformer architecture (Fig. 1b) and (2) the dual-domain self-supervised learning strategy used for training DCCT (Fig. 1a).

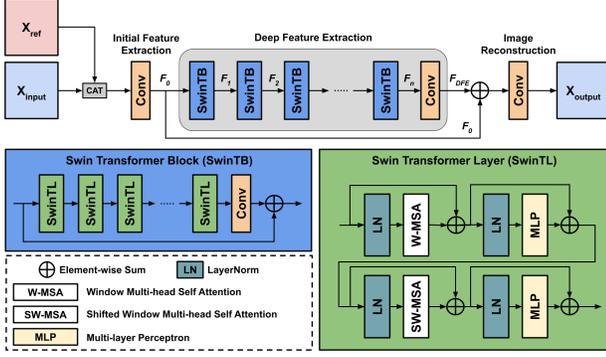


Figure 3. The architecture of Swin Transformer Reconstruction Network (SwinRN). It consists of initial feature extraction, deep feature extraction, and image reconstruction modules, and is used as the backbone reconstruction network in Figure 1.

### 3.1. Deep Conditional Cascade Transformer

DCCT uses a cascaded network design with interleaved data consistency (DC) layers [27]. To efficiently exploit multi-contrast information for reconstruction learning, we develop two deep multi-contrast network conditioning mechanisms to better leverage fully-sampled reference acquisitions. To further enable high-quality reconstruction, we propose a Swin Transformer Reconstruction Network (SwinRN) as the backbone network in DCCT.

**Deep MC-MRI Conditioning.** With MC-MRI, we use two conditioning methods for sharing reference MRI with target MRI in DCCT: K-space Filling (KF) conditioning and Channel-wise (CC) conditioning. First, we use KF, because multi-contrast MRI depicts distinct physiological properties of imaged tissues, resulting in different image contrast, but the multi-contrast MRI images share the same anatomy. While target contrast zero-padded reconstruction with undersampled k-space data could result in severe artifacts, filling the unacquired k-space with reference contrast k-space data (assuming no motion between the target and reference) can produce alias-reduced reconstruction with the same anatomy and altered contrast. This KF reconstruction can be used as initial DCCT input, so that it can focus on learning contrast conversion instead of de-aliasing. An example of KF is shown in Fig. 2.

In addition to KF, we also use CC. As illustrated in Figure 1b, the first input to the cascade is the channel-wise concatenation of KF target and the reference contrast MRI image, while the following cascade inputs are the channel-wise concatenations of the previous cascade output and the reference contrast MRI image.

**Swin Transformer Reconstruction Network.** SwinRN is used as the backbone network for DCCT with its architec-

ture shown in Figure 3. SwinRN consists of three modules: initial feature extraction (IFE) using a  $3 \times 3$  convolutional layer, deep feature extraction (DFE) using multiple Swin Transformer Blocks (SwinTB), and image reconstruction using global residual learning and a  $3 \times 3$  convolution layer. The workflow is described as  $F_0 = P_{IFE}(X_{init}|X_{ref})$ , where  $P_{IFE}$  denotes the IFE operation and  $| \cdot$  denotes conditional input. The IFE feature  $F_0$  is then used for residual learning in the reconstruction step and is fed into multiple SwinTBs for DFE. If there are  $n$  SwinTBs, the  $n$ -th output  $F_n$  is  $F_n = P_{SwinTB_n}(F_{n-1})$ .

Then, the output of DFE is given by  $F_{DFE} = P_{DFE}(F_n)$ , where  $P_{DFE}$  is a  $3 \times 3$  convolutional layer for final feature fusion in DFE. Given  $F_{DFE}$  and the global residual connection of  $F_{IFE}$ , the final reconstruction can be generated via

$$X_{output} = P_{IR}(F_{DFE} + F_{IFE}), \quad (1)$$

where  $P_{IR}$  is another  $3 \times 3$  convolutional layer for generating a one-channel image reconstruction output.

**Swin Transformer Block.** Each SwinTB (Fig. 3) consists of multiple Swin transformer layers (SwinTL), a convolution layer for local feature fusion, and a residual connection for local residual learning. Given the input feature  $F_{i,0}$  of the  $i$ -th SwinTB, the intermediate feature is written as:

$$F_{i,j} = P_{SwinTL_{i,j}}(F_{i,j-1}), \quad (2)$$

where  $P_{SwinTL_{i,j}}(\cdot)$  is the  $j$ -th SwinTL in the  $i$ -th SwinTB. Then, local feature fusion and local residual learning is applied to generate the SwinTB output:

$$F_i = P_{LFF_i}(F_{i,K}) + F_{i,0}, \quad (3)$$

where  $K$  is the number of SwinTL in SwinTB and  $P_{LFF_i}$  is a convolutional layer for SwinTB's  $i$ -th local feature fusion.

SwinTL [19] consists of layer normalization (LN), multi-layer perceptrons (MLP), and multi-head self attention (MSA) modules [30] with regular windowing (W-MSA) and shifted windowing (SW-MSA) configurations. Given an input with feature size  $H \times W \times C$ , SwinTL first reshapes the input into  $(M \times M) \times (\frac{H}{M} \times \frac{W}{M}) \times C$  by partitioning it into non-overlapping  $\frac{H}{M} \times \frac{W}{M}$  windows with each window containing  $M \times M$  patches. Then, self-attention can be computed for each window [30] and can formulate the attention output of W-MSA. To enable cross-window connection, self-attention is also computed for each window by shifting the feature by  $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$  before partitioning. SwinTL processing (Fig. 3) can thus be summarized as,

$$\bar{z} = W-MSA(LN(z)) + z, \quad (4)$$

$$\tilde{z} = MLP(LN(\bar{z})) + \bar{z}, \quad (5)$$

$$\hat{z} = SW-MSA(LN(\tilde{z})) + \tilde{z}, \quad (6)$$

$$\check{z} = MLP(LN(\hat{z})) + \hat{z}, \quad (7)$$

where  $MLP$  is a 2-layer and 30-60 neuron wide MLP with a GELU activation [9].

In summary, SwinRN with SwinTB blocks is embedded in the cascaded framework of DCCT for MRI reconstruction. We use three SwinRNs by default in our cascade, with each SwinRN sharing the same parameters as the default setting. The number of SwinTB in each SwinRN is set to four, with each SwinTB containing four SwinTLs.

### 3.2. Dual-Domain Self-Supervised Learning

To train DCCT in a self-supervised fashion without using any fully-sampled ground truth data in the target domain, we use dual-domain self-supervision, as illustrated in Figure 1. Let  $f_{dcct}(y_{tag}, y_{ref})$  denote DCCT, where  $y_{tag}$  is the target contrast’s undersampled data and  $y_{ref}$  is the reference contrast’s fully sampled data. During training, we first randomly partition  $y_{tag}$  into two disjoint sets via,

$$y_{p_1} = y_{tag} \odot M_1 \quad (8)$$

$$y_{p_2} = y_{tag} \odot M_2, \quad (9)$$

where  $\odot$  is element-wise multiplication and  $M_1$  and  $M_2$  are binary k-space masks for partition 1 and partition 2. Note that  $M_1 + M_2 = M_{tag}$ , where  $M_{tag}$  is the binary mask indicating all under-sampled locations. The partitions  $y_{p_1}$  and  $y_{p_2}$  are then fed into DCCT for parallel reconstruction,

$$X_{p_1} = f_{dcct}(y_{p_1}, y_{ref}) \quad (10)$$

$$X_{p_2} = f_{dcct}(y_{p_2}, y_{ref}), \quad (11)$$

where the networks share the same weights. As the reconstructions of  $y_{p_1}$  and  $y_{p_2}$  should be consistent with each other, our first loss is an Appearance Consistency (AC) loss operating in the image domain as,

$$\mathcal{L}_{AC} = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_{grad}, \quad (12)$$

where,

$$\mathcal{L}_{img} = \|X_{p_1} - X_{p_2}\|_1 \quad (13)$$

and,

$$\mathcal{L}_{grad} = \|\nabla_v X_{p_1} - \nabla_v X_{p_2}\|_1 + \|\nabla_h X_{p_1} - \nabla_h X_{p_2}\|_1, \quad (14)$$

where  $\nabla_v$  and  $\nabla_h$  are vertical and horizontal intensity gradient operators, respectively. We empirically found  $\lambda_1 = 1$  and  $\lambda_2 = 0.1$  to achieve optimal performance.

Our second loss corresponds to a Partition Data Consistency (PDC) loss which operates in k-space. If DCCT can generate a high-quality image from any undersampled k-space measurement, the k-space data of the image predicted from the first partition  $y_{p_1}$  should be consistent with

the other partition  $y_{p_2}$  and vice versa. The predicted k-space partition can be written as,

$$y_{2 \rightarrow 1} = \mathcal{F}(X_{p_2}) \odot M_1 \quad (15)$$

$$y_{1 \rightarrow 2} = \mathcal{F}(X_{p_1}) \odot M_2, \quad (16)$$

Therefore, the PDC loss is formulated as,

$$\mathcal{L}_{PDC} = \|y_{2 \rightarrow 1} - y_{p_1}\|_1 + \|y_{1 \rightarrow 2} - y_{p_2}\|_1, \quad (17)$$

where the first and second term are the partial data consistency losses for partitions 1 and 2, respectively. Combining the AC loss in the image domain and the PDC loss in k-space, our total loss can be written as,

$$\mathcal{L}_{tot} = \mathcal{L}_{AC} + \lambda_3 \mathcal{L}_{PDC} \quad (18)$$

where  $\lambda_3 = 0.1$  is used to balance the scale between k-space and image domain losses.

### 3.3. Data Preparation

We use 578 MC-MRI subjects with both T2-weighted and Proton Density (PD)-weighted acquisitions from IXI<sup>1</sup> for our experiments. The registered MC-MRI data consisting of 11808 pairs of T2 and PD weighted axial slices are split subject-wise into 8376 pairs for training, 1080 for validation, and 2352 for testing, with no slices from any subject overlapping. We consider two MC-MRI scenarios in our experiments: accelerating T2-weighted acquisition (the target protocol) by utilizing a fully sampled PD-weighted acquisition (the reference protocol), and accelerating PD-weighted target acquisition with a fully sampled T2-weighted reference. Here, we consider the Cartesian sampling pattern with the acceleration factor (R) set to a value between 2 and 8 corresponding to acceleration in acquisition time for the target protocol.

### 3.4. Evaluation Metrics and Baselines Comparisons

Benchmark results are presented on 2352 test slices from 114 patients. We evaluate the target reconstruction results using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) computed against ground truth. For baseline comparison, we first compare our results against previous fully-supervised MC-MRI reconstruction methods that require ground truth fully-sampled data, including UFNet [35], MCNet [17], and VarNet [18]. To further benchmark against previous self-supervised MRI reconstruction methods originally designed for single-contrasts, we also extend SSDU [37], HQSNet [31], and SSISTA [10] to the MC-MRI setting by using the reference contrast as an extra input channel. As an upper bound, we also compare self-supervised DSFormer against a supervised-variant where the DCCT of DSFormer was trained in a fully supervised fashion with ground truth available.

<sup>1</sup><https://brain-development.org/ixi-dataset/>, CC BY-SA 3.0 license

Table 1. Quantitative comparison of T2 (left sub-table) and PD (right sub-table) reconstructions under three different acceleration settings for the target contrast MRI. Fully supervised methods and self-supervised methods are marked in **bold** and underlined, respectively. Best results are marked in **red**.

PSNR/SSIM Methods	Target: T2w   Reference: PD				Target:PD   Reference:T2w				Runtime (ms)	Number of Param
	×2	×4	×6	×8	×2	×4	×6	×8		
Zero-padding	24.86/.761	22.72/.679	21.48/.623	19.11/.597	23.92/.744	21.68/.663	20.47/.610	18.78/.579	-	-
CS-TV[11]	30.18/.890	29.04/.862	26.13/.789	25.67/.762	30.09/.887	29.18/.861	26.05/.787	25.21/.750	3023.8	-
<b>SSDU</b> [37]	42.54/.981	38.47/.976	34.21/.972	31.72/.961	41.78/.983	37.63/.979	33.99/.968	31.61/.960	49.9	0.11M
<b>HQSNet</b> [31]	40.13/.980	37.79/.973	33.38/.969	31.12/.958	40.82/.981	35.97/.974	33.37/.967	31.03/.956	49.9	0.11M
<b>SSISTA</b> [10]	42.09/.980	38.29/.974	34.13/.971	31.51/.960	41.29/.982	37.43/.977	33.67/.968	31.43/.959	40.8	0.38M
UFNet[35]	32.30/.970	32.07/.969	31.85/.967	30.02/.950	32.22/.971	32.06/.969	31.88/.968	29.99/.948	10.6	7.6M
VarNet[18]	33.01/.973	32.71/.971	32.43/.970	30.65/.957	33.08/.974	32.86/.972	32.67/.970	30.55/.955	48.7	8.2M
MCNet[17]	43.79/.989	39.14/.983	35.61/.972	32.12/.963	42.90/.988	38.56/.979	35.45/.971	32.03/.961	49.9	0.11M
<b>DSFormer</b>	<b>45.05/.993</b>	<b>40.31/.985</b>	<b>37.04/.977</b>	<b>33.65/.969</b>	<b>45.07/.993</b>	<b>40.52/.987</b>	<b>37.45/.982</b>	<b>33.48/.967</b>	51.3	0.18M

### 3.5. Implementation Details

We implement our method in Pytorch and perform experiments using an NVIDIA Quadro RTX 8000 GPU with 48GB memory. The Adam solver [13] was used to optimize our models with  $lr = 2 \times 10^{-4}$ ,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ . We use a batch size of 3 during training. In DSFormer, the number of cascades can be flexibly adjusted and is set to three as the default setting in the main experiments and is swept over in Fig. 6. The SwinRN shares the same parameters in each cascade. The number of SwinTB in each SwinRN is set to four, with each SwinTB containing four SwinTLs. During training, the data partitioning rate is randomly generated between [0.2, 0.8] on-the-fly which separates the undersampled k-space data into two disjoint k-space data and augments the training data. For baseline implementations, we compare with SSDU, HQSNet, and SSISTA, UFNet, MCNet, and VarNet. The hyperparameters of each method are tuned on the validation set with test data held-out for final evaluation.

## 4. Experimental Results

### 4.1. Image Quality Evaluation and Comparison

Quantitative evaluations on two different MC-MRI scenarios under three different acceleration settings are summarized in Table 1. The left sub-table summarizes MC-MRI reconstruction with T2 target contrast and PD reference contrast (T2 reconstructions were evaluated here). Among fully supervised methods, MCNet [17] achieves the best T2 reconstruction performance with PSNR up to 43.79 dB and SSIM up to 0.989 when using  $\times 2$  acceleration. It can also be observed that MCNet [17] consistently outperforms the previous self-supervised MRI reconstruction methods modified to operate on multi-contrast data. In the last row of Table 1, we see that DSFormer trained with self-supervision alone outperforms supervised baselines and increases PSNR from 43.79 dB to 45.05 dB and SSIM from 0.989 to 0.993. Similar observations are made for the  $\times 4$

accelerated T2 experiments where DSFormer outperforms MCNet, with PSNR increasing from 39.14 dB to 40.31 dB and SSIM increasing from 0.983 to 0.985.

As expected, the reconstruction performance of all methods decreases as the acceleration rate increases. However, DSFormer is still able to widely outperform previous supervised methods and keep PSNR at 37.04 and SSIM at 0.977 with  $\times 6$  accelerated T2 reconstruction. The inference run time and the number of model parameters of different methods are also summarized in Table 1, with the deep learning methods achieving orders of magnitude faster reconstruction over iterative methods like CS-TV [11]. As compared to the previous best results of MCNet, DSFormer requires only a slightly increased number of parameters and run-time to achieve improved reconstruction performance.

The qualitative comparison of various T2 reconstructions is shown in Fig. 4, illustrating  $\times 4$  and  $\times 6$  acceleration settings. Reconstructions with zero padding create significant aliasing artifacts and lose anatomical details. While both VarNet [18] and MCNet [17] significantly reduce the aliasing artifacts with decreased reconstruction error, they require fully supervised training from paired data. On the other hand, DSFormer, using only self-supervision and multi-contrast conditioning, further reduces the residual error between the reconstruction and the ground truth and yields superior reconstruction quality. The full qualitative comparison of T2 reconstruction is visualized in Figure S1 in the supplemental materials.

The right sub-table of Table 1 summarizes MC-MRI with PD target contrast and T2 reference contrast, where PD reconstructions were evaluated. Similar observations are made for PD reconstruction, where DSFormer still achieves the best reconstruction under all three acceleration settings over previous fully supervised methods [35, 18, 17] and modified previous self-supervised methods [37, 31, 10]. The qualitative comparison of several PD reconstruction methods is shown in the supplementary materials.

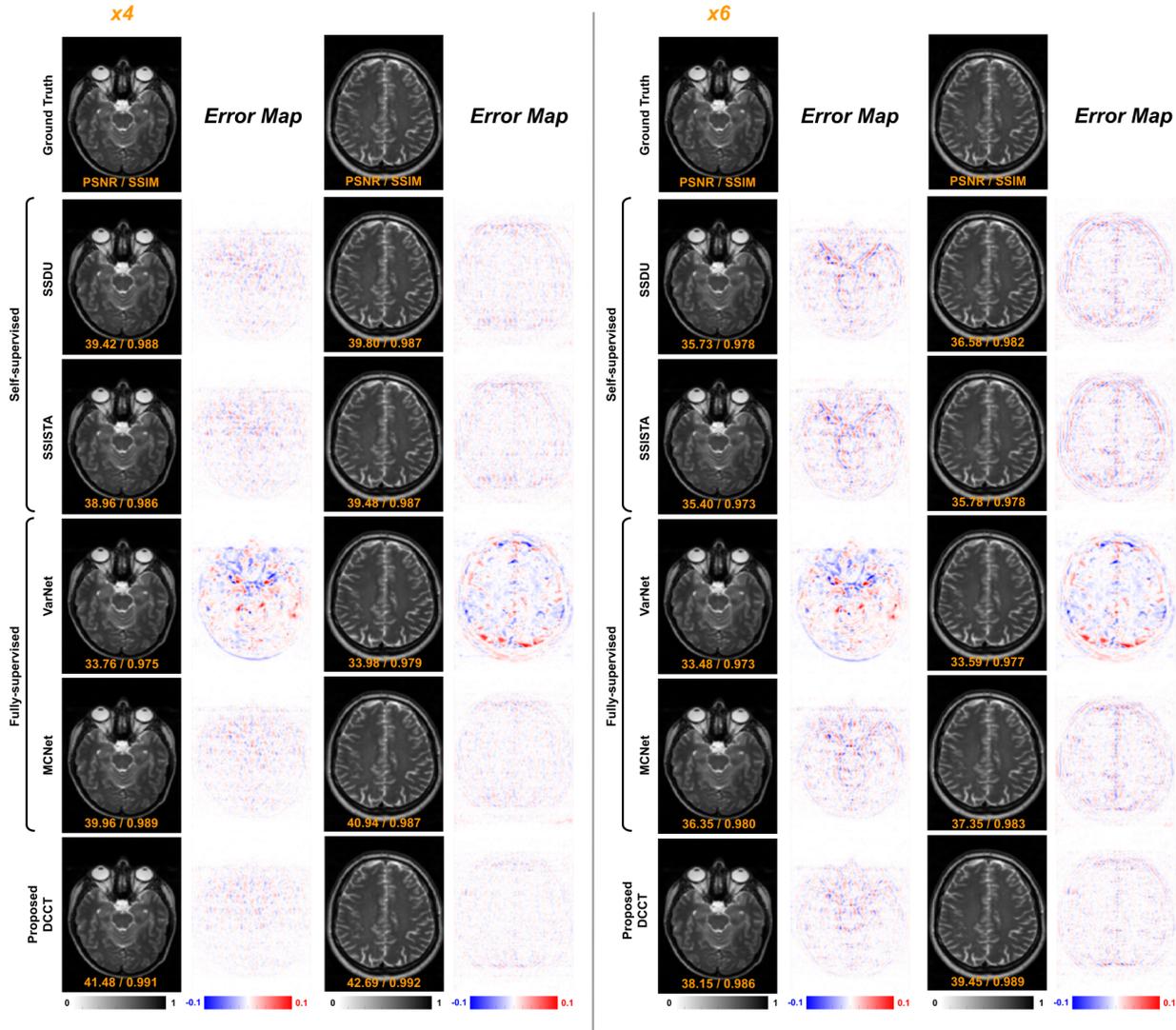


Figure 4. Qualitative comparisons of T2 reconstructions using  $\times 4$  and  $\times 6$  acceleration. For T2 reconstruction, PD is used as the reference contrast. The corresponding error maps between ground truth images and the reconstructions are illustrated in BWR colormaps. Across both supervised and self-supervised methods, DSFormer achieves the highest-fidelity reconstructions due to its improved architecture, dual-domain self-supervision, and conditioning mechanisms.

## 4.2. Ablation Studies

**Dual-domain self-supervision.** To isolate the individual utility of the various components of dual-domain self-supervision, we evaluate performance using either only image-domain self-supervision ( $\mathcal{L}_{AC}$ ) or only k-space self-supervision ( $\mathcal{L}_{PDC}$ ). The quantitative comparison is summarized in Table 2. We observe that using either only k-space self-supervision or only image-domain self-supervision still yields strong reconstruction quality, with a PSNR of 39.95 under  $\times 4$  acceleration, which indicates self-supervision in both domains can help with reconstruction. Combining both image-domain and k-space self-

supervision yields the best reconstruction performance. A visual comparison of reconstructions from different self-supervision settings is illustrated in Figure 5.

Table 2. Quantitative comparison of T2 reconstruction performance from different self-supervision settings, including image-domain only self-supervision, k-space only self-supervision, and the proposed dual-domain self-supervision. Higher is better.

PSNR/SSIM	$\times 4$	$\times 6$
Only k-space self-supervision	40.23/0.982	36.89/0.975
Only image self-supervision	39.95/0.981	36.55/0.973
DSFormer (proposed)	40.31/0.985	37.04/0.977

**Deep MC-MRI conditioning.** To understand the impact of

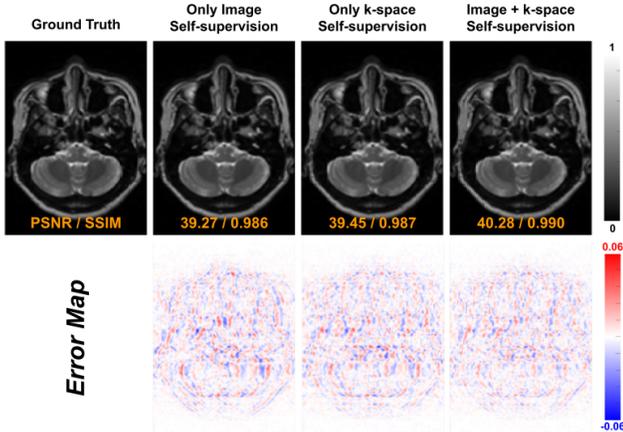


Figure 5. Further ablation studies comparing MRI reconstruction from different self-supervision settings. Lower error is better.

KF and CC at the initial network input, we evaluate DSFormer performance with or without KF and CC, with results summarized in Table 3. DSFormer without both KF and CC implies no multi-contrast data is used and results in only 38.24 PSNR under  $R = 4$  setting. Under the same setting, DSFormer with either KF-only or CC-only can integrate multi-contrast information and improves PSNR from 38.24 to 39.06 with KF-only and 40.22 with CC-only. Using both KF and CC leads to the best performance where PSNR is further improved to 40.31. Similar trends are observed for  $6\times$  acceleration.

Table 3. Quantitative comparison of T2 reconstruction performance when using DSFormer with or without KF and CC and when training DSFormer in a fully supervised manner similar to [17, 35] without our proposed consistency losses. † means fully supervised training.

PSNR/SSIM	$\times 4$	$\times 6$
DSFormer w/o KF	40.22/0.982	36.93/0.972
DSFormer w/o CC	39.06/0.979	35.22/0.970
DSFormer w/o KF and CC	38.24/0.971	34.51/0.961
DSFormer	40.31/0.985	37.04/0.977
†DSFormer (Upper Bound)	40.34/0.989	37.12/0.981
†DSFormer w/o KF	40.25/0.983	36.99/0.974
†DSFormer w/o CC	39.12/0.981	35.31/0.972
†DSFormer w/o KF and CC	38.42/0.976	34.70/0.965

**Fully-supervised vs. Self-supervised DSFormer.** In order to understand the performance gap between full supervision and dual-domain self-supervision, we compare the reconstruction performance of DCCT trained without partitioning the input and replacing its consistency losses with direct image-domain reconstruction losses against the target ground truth, similar to the supervised training in [35, 17]. Quantitative comparisons are summarized in Table S1. As an upper bound, fully supervised DSFormer achieved reconstruction PSNR of 40.34 dB under  $\times 4$  accel-

eration, which is only  $\sim 0.03$  higher than self-supervised DSFormer in terms of PSNR. SwinRN effectiveness can be further evaluated by comparing supervised DSFormer w/o KF with MCNet (Table 1) where both methods share the same cascade framework except the difference in the backbone network. We can observe our supervised method based on SwinRN achieves PSNR of 40.25 which is significantly better than MCNet based on simple sequential convolutional layers with residual connection with PSNR of 39.14 under  $\times 4$  setting. However, at  $6\times$  acceleration we see a gap emerging between supervised DSFormer and self-supervised DSFormer, each achieving 37.12 and 37.04 dB PSNR, respectively, indicating that there is still a performance benefit to using fully sampled training data under higher acceleration factors.

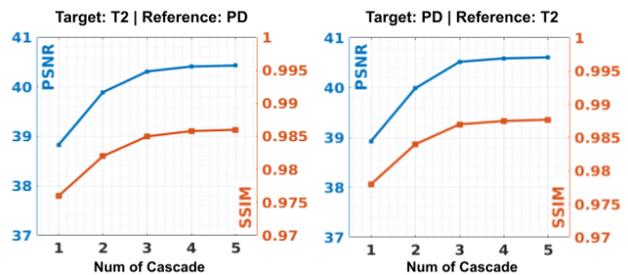


Figure 6. The effect of increasing the number of cascaded SwinRNs in DSFormer at  $\times 4$  acceleration.

**Impact of the number of cascades.** As the number of cascades can be flexibly adjusted in DSFormer, we analyze the effect of increasing the number of cascaded blocks in our framework, with the result summarized in Figure 6 using  $\times 4$  acceleration. Using a higher number of cascaded blocks boosts the reconstruction performance, with gains asymptotically stabilizing on further increases beyond three blocks. In T2 reconstruction, increasing the number of cascaded blocks from 3 to 4 only increases PSNR by less than 0.002 dB. Similar observations can be made from the PD reconstructions.

## 5. Conclusion

We developed DSFormer, a dual-domain self-supervised transformer for accelerated multi-contrast MRI reconstruction. DSFormer proposed a deep conditional cascaded transformer architecture trained under both k-space and image domain self-supervision. Benchmarks against established baselines demonstrate that DSFormer outperformed previous *fully supervised methods* that require training with paired data (Table 1) and that DSFormer achieves nearly the same performance when trained with either full supervision or with our proposed dual-domain self-supervision (Table 3), almost closing the gap between supervised and self-supervised methods for accelerated MRI reconstruction.

## References

- [1] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.
- [2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.
- [3] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018.
- [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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.
- [5] Salman UH Dar, Mahmut Yurt, Mohammad Shahdloo, Muhammed Emrullah Ildız, Berk Tınaz, and Tolga Çukur. Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1072–1087, 2020.
- [6] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.
- [7] Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic resonance in medicine*, 80(5):2188–2201, 2018.
- [8] 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*, 79(6):3055–3071, 2018.
- [9] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [10] Chen Hu, Cheng Li, Haifeng Wang, Qiegen Liu, Hairong Zheng, and Shanshan Wang. Self-supervised learning for mri reconstruction with a parallel network training framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–391. Springer, 2021.
- [11] Junzhou Huang, Chen Chen, and Leon Axel. Fast multi-contrast mri reconstruction. *Magnetic resonance imaging*, 32(10):1344–1352, 2014.
- [12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [14] 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*, 2(1):e190007, 2020.
- [15] Yilmaz Korkmaz, Salman UH Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Cukur. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. *IEEE Transactions on Medical Imaging*, 2022.
- [16] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.
- [17] Xinwen Liu, Jing Wang, Jin Jin, Mingyan Li, Fangfang Tang, Stuart Crozier, and Feng Liu. Deep unregistered multi-contrast mri reconstruction. *Magnetic Resonance Imaging*, 2021.
- [18] Xinwen Liu, Jing Wang, Hongfu Sun, Shekhar S Chandra, Stuart Crozier, and Feng Liu. On the regularization of feature fusion and mapping for fast mr multi-contrast imaging via iterative networks. *Magnetic resonance imaging*, 77:159–168, 2021.
- [19] 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. *arXiv preprint arXiv:2103.14030*, 2021.
- [20] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.
- [21] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
- [22] Ricardo Otazo, Daniel Kim, Leon Axel, and Daniel K Sodickson. Combination of compressed sensing and parallel imaging for highly accelerated first-pass cardiac perfusion mri. *Magnetic resonance in medicine*, 64(3):767–776, 2010.
- [23] Luis Pineda, Sumana Basu, Adriana Romero, Roberto Calandra, and Michal Drozdal. Active mr k-space sampling with reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 23–33. Springer, 2020.
- [24] Chen Qin, Jo Schlemper, Jose Caballero, Anthony N Price, Joseph V Hajnal, and Daniel Rueckert. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE transactions on medical imaging*, 38(1):280–290, 2018.

- [25] 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. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(1):120–135, 2020.
- [26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [27] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.
- [28] Nalini M Singh, Juan Eugenio Iglesias, Elfar Adalsteinsson, Adrian V Dalca, and Polina Golland. Joint frequency and image space learning for fourier imaging. *arXiv preprint arXiv:2007.01441*, 2020.
- [29] Liyan Sun, Zhiwen Fan, Xueyang Fu, Yue Huang, Xinghao Ding, and John Paisley. A deep information sharing network for multi-contrast compressed sensing mri reconstruction. *IEEE Transactions on Image Processing*, 28(12):6141–6153, 2019.
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [31] Alan Q Wang, Adrian V Dalca, and Mert R Sabuncu. Neural network-based reconstruction in compressed sensing mri without fully-sampled training data. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 27–37. Springer, 2020.
- [32] Shanshan Wang, Zhenghang Su, Leslie Ying, Xi Peng, Shun Zhu, Feng Liang, Dagan Feng, and Dong Liang. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 514–517. IEEE, 2016.
- [33] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021.
- [34] Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Carola-Bibiane Schönlieb, and Hua Huang. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *International Conference on Machine Learning*, pages 10158–10169. PMLR, 2020.
- [35] Lei Xiang, Yong Chen, Weitang Chang, Yiqiang Zhan, Weili Lin, Qian Wang, and Dinggang Shen. Ultra-fast t2-weighted mr reconstruction using complementary t1-weighted information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 215–223. Springer, 2018.
- [36] Burhaneddin Yaman, Seyed Amir Hossein Hosseini, and Mehmet Akcakaya. Zero-shot self-supervised learning for MRI reconstruction. In *International Conference on Learning Representations*, 2022.
- [37] Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic resonance in medicine*, 84(6):3172–3191, 2020.
- [38] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.
- [39] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1828–1837, 2018.
- [40] Zizhao Zhang, Adriana Romero, Matthew J Muckley, Pascal Vincent, Lin Yang, and Michal Drozdal. Reducing uncertainty in undersampled mri reconstruction with active acquisition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2019.
- [41] Bo Zhou, Jo Schlemper, Neel Dey, Seyed Sadegh Mohseni Salehi, Kevin Sheth, Chi Liu, James S Duncan, and Michal Sofka. Dual-domain self-supervised learning for accelerated non-cartesian mri reconstruction. *Medical Image Analysis*, page 102538, 2022.
- [42] Bo Zhou and S Kevin Zhou. Dudornet: Learning a dual-domain recurrent network for fast mri reconstruction with deep t1 prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4282, 2020.
- [43] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.