Unsupervised MRI Reconstruction via Zero-Shot Learned Adversarial Transformers

Yilmaz Korkmaz, Salman U. H. Dar, *Member, IEEE*, Mahmut Yurt, Muzaffer Özbey, *Member, IEEE*,
and Tolga Çukur, *Senior Member, IEEE*

Abstract—**Supervised reconstruction models are characteristically trained on matched pairs of undersampled and fully-sampled data to capture an MRI prior, along with supervision regarding the imaging operator to enforce data consistency. To reduce supervision requirements, the recent deep image prior framework instead conjoins untrained MRI priors with the imaging operator during inference. Yet, canonical convolutional architectures are suboptimal in capturing long-range relationships, and priors based on randomly initialized networks may yield suboptimal performance. To address these limitations, here we introduce a novel unsupervised MRI reconstruction method based on zero-Shot Learned Adversarial TransformERs (SLATER). SLATER embodies a deep adversarial network with cross-attention transformers to map noise and latent variables onto coil-combined MR images. During pre-training, this unconditional network learns a high-quality MRI prior in an unsupervised generative modeling task. During inference, a zero-shot reconstruction is then performed by incorporating the imaging operator and optimizing the prior to maximize consistency to undersampled data. Comprehensive experiments on brain MRI datasets clearly demonstrate the superior performance of SLATER against state-of-the-art unsupervised methods.**

Index Terms—**Adversarial, transformers, MRI, unsupervised, reconstruction, zero shot, generative.**

I. INTRODUCTION

MAGNETIC Resonance Imaging (MRI) is a clinical powerhouse due to its excellent soft tissue contrast. Yet, low spin polarization at mainstream field strengths limits signal-to-noise ratio and necessitates prolonged exams. Extensive MRI exams incur added economic costs and lower use efficiency, and they may not be possible in uncooperative patient populations. A dire consequence is administration of exams that are in lower quality and/or diversity, which projects poorly onto diagnostic utility. Accelerated MRI addresses this fundamental limitation by performing undersampled acquisitions, and then solving an inverse problem to reconstruct images from available k-space data [1]–[4].

Given their exceptional performance in inverse problems, supervised neural networks have been quickly adopted for MRI reconstruction [5]. Reconstruction refers to the task of mapping undersampled acquisitions to images that are as consistent as possible with corresponding fully-sampled acquisitions. Supervised models performing this conditional mapping are trained on matched pairs of undersampled and fully-sampled data. The training process involves multiple lines of supervision including pairing of input-output data to learn an indirect prior that reduces aliasing artifacts in MR images [6]–[9], and enforcement of data consistency to embed the imaging operator that reflects k-space undersampling and coil-sensitivity encoding [10]–[14]. Accordingly, supervised models are typically retrained for notable changes in the data distribution (e.g, different MRI contrast) or the imaging operator (e.g., different k-space sampling density) [15]–[17].

Several important approaches have been proposed in literature to reduce supervision requirements. A group of studies have reduced explicit supervision related to raw data, proposing models trained using unpaired sets of input-output data [18]–[20], or using only undersampled data [21]–[27]. That said, these models involve implicit supervision regarding the imaging operator. As such, they are trained for a specific coil-array configuration and k-space sampling density, factors assumed to be consistent across the training and test sets [21]–[24]. To remove other supervision aspects, a second group of studies have built unsupervised models by decoupling the MRI prior from the imaging operator. These models capture an MRI prior via generative networks that are either untrained [28]–[31] or trained to synthesize fully-sampled MR images [32]–[34]. The imaging operator is then conjoined with the MRI prior during inference on test data. This unsupervised approach excludes paired datasets for training, and promises improved generalization against deviations in the imaging operator [33], [34]. Yet, previously proposed models are commonly based on convolutional neural networks (CNNs) that suffer from limited sensitivity in capturing long-range dependencies [35]–[37].
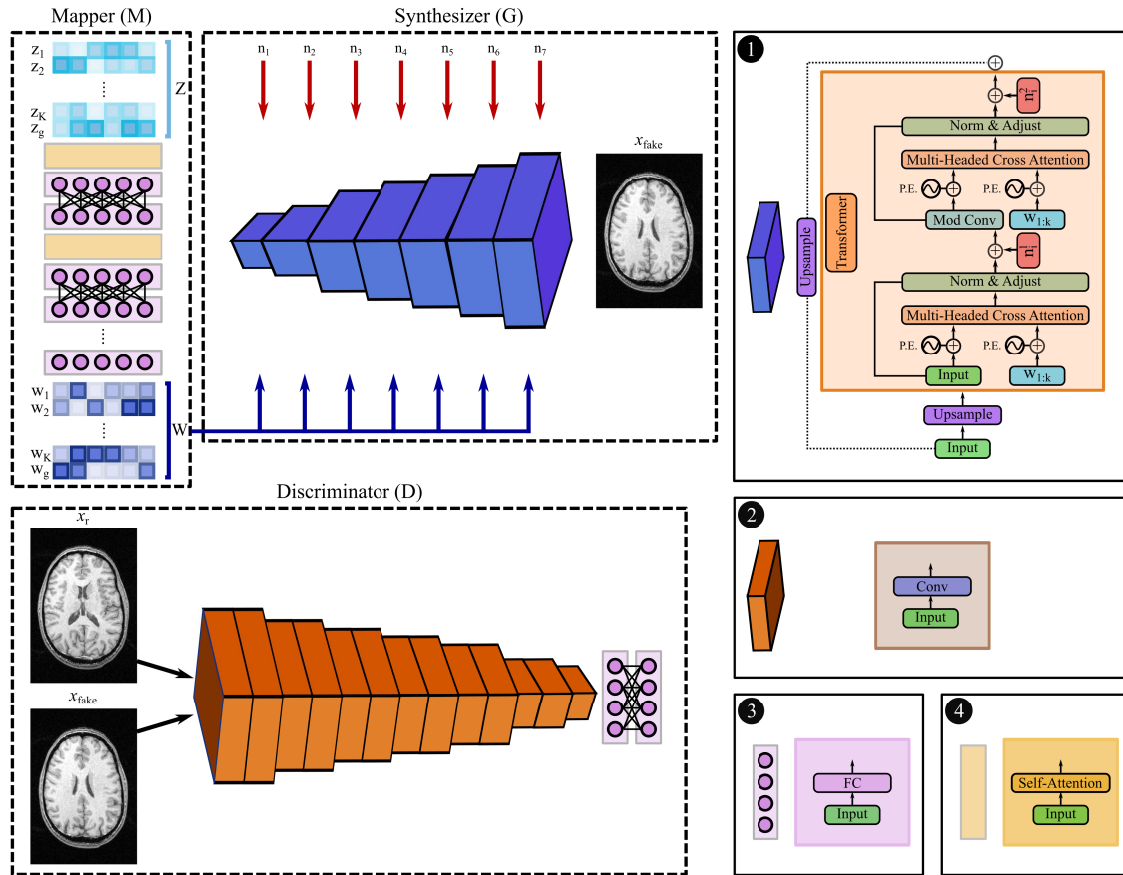
Fig. 1. SLATER is based on an unconditional adversarial architecture with a synthesizer (G), a discriminator (D), and a mapper (M). G is a multi-layer architecture where image resolution is progressively increased across layers. Within intermediate layers, a convolutional upsampling block is followed by a cross-attention transformer block (see 1). The transformer block receives global and local latent variables, noise variables and positional encoding (P.E.) for latents and image features. D enables adversarial learning while it receives as input synthetic and actual MR images. It is composed of convolutional and fully-connected (FC) blocks (see 2, 3). M projects raw latent variables (Z) onto k local ($w_1, \ldots, w_K$) variables and one global (w_g) latent variable. It is composed of self-attention and FC blocks (see 3, 4, and Supp. Fig. 1 for details). An unsupervised generative modeling task is performed using the SLATER model to capture high-quality MRI priors.

Here, we introduce a novel unsupervised MRI reconstruction based on zero-shot learned adversarial transformers (SLATER). SLATER decomposes the reconstruction process to decouple learning of the MRI prior from embedding of the imaging operator.[1] During a pre-training phase, an unconditional adversarial model is used to synthesize high-quality, coil-combined MR images (Fig. 1). To improve capture of long-range spatial context without introducing computational burden, we propose an architecture comprising cross-attention transformer blocks between low-dimensional latent variables and high-dimensional image features. During the inference phase, the learned MRI prior is combined with the imaging operator. This is achieved by optimizing network parameters that reflect the MRI prior to enforce a data-consistency loss on undersampled test data (Fig. 2). To improve inference efficiency, a weight propagation strategy is proposed where the optimized network weights are transferred across consecutive cross-sections.

The proposed method performs an unsupervised generative modeling task on coil-combined images derived from fully-sampled MRI acquisitions. Adapting the generative model to the reconstruction task without any

training samples, a zero-shot reconstruction then maps undersampled data to high-quality MR images during inference. The decoupled reconstruction process and model adaptation during inference contribute to improved generalization performance for SLATER. The source code for SLATER can be found at: https://github.com/icon-lab/SLATER. Supplementary Text, Tables and Figures can be viewed at https://arxiv.org/pdf/2105.08059.pdf.

Contributions:

- For the first time in literature, we introduce an adversarial vision transformer model for MRI reconstruction.
- Our proposed model uses cross-attention transformers to capture long-range spatial dependencies without computational burden.
- Sample-specific model adaptation and cross-sectional weight propagation strategies are introduced that respectively enhance out-of-domain generalizability and inference efficiency.

II. RELATED WORK

Supervised reconstruction models are current state-of-the-art in MRI with numerous successful architectures reported including basic CNNs [6], [39]–[41], residual

---

[1] see [38] for a preliminary version of this work presented at ISMRM 2021

a Inference on test data

Undersampled Reference

b Reconstruction loss

$$L_{\hat{R}}(W, n, \theta_G) = \lambda_1 \left\| F\hat{X}_\Omega - FX_\Omega \right\|_1 + \lambda_2 \left\| F\hat{X}_\Omega - FX_\Omega \right\|_2$$
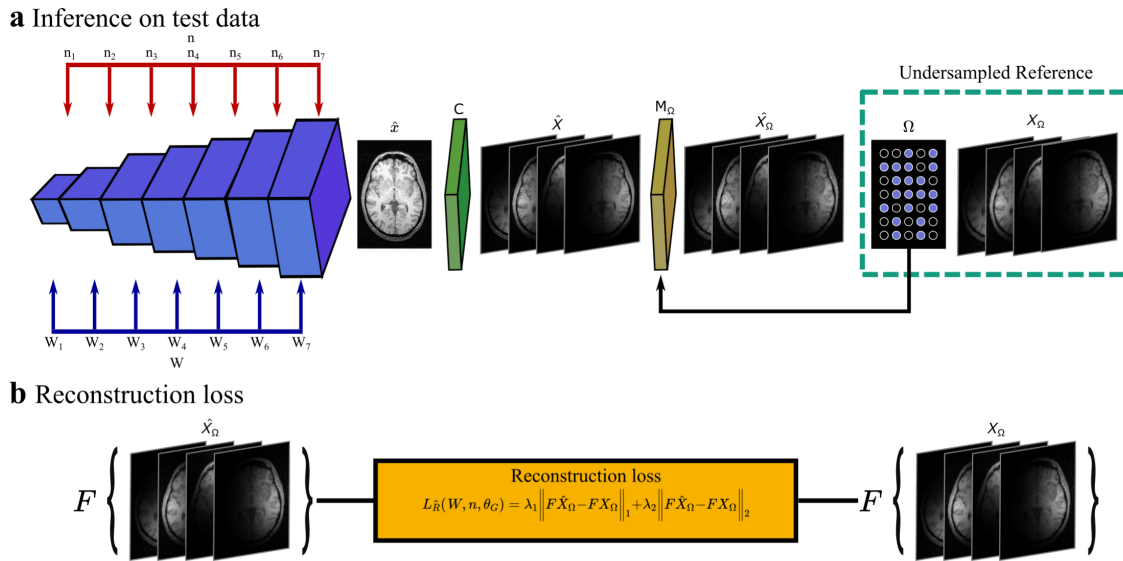
Reconstruction loss

Fig. 2. **(a)** The generative model in SLATER is adapted to perform zero-shot reconstruction for accelerated MRI. To do this, the generator output is back-projected onto individual coils (C), and masked with the same sampling pattern as in the undersampled acquisition (M_Ω). **(b)** The network prior consisting of noise (n), latent variables (W) and weights (θ_G) is then optimized to maximize consistency between the reconstructed and acquired k-space data.

CNNs [42], [43], perceptrons [17], [44], physics-guided unrolled CNN networks [7], [10], [11], [14], [45]–[49], recurrent CNNs [9], [50], [51], generative adversarial networks (GANs) [52]–[55], and variational networks [8], [13], [15], [56]. However, supervised models are trained on paired sets of undersampled and fully-sampled data, along with supervision regarding the imaging operator in the form of a data-consistency term. Compiling large sets of paired data is non-trivial [57], and supervised models often require retraining to cope with deviations in the imaging operator [17].

To improve utility of deep MRI reconstruction, a common strategy has been to target explicit supervision on raw data in order to lower reliance on large, paired training datasets. Domain-transferred models are trained in a data-abundant source domain and then adopted for reconstruction in the target domain [16], [58]. Residual models are trained to predict residual error between ground truth and the output of a conventional MRI reconstruction [59], [60]. Both approaches permit training with relatively few samples, but training often requires paired datasets. Unpaired models are instead trained on input-output data collected from separate groups of subjects. This can be achieved with cycle-consistent models that learn bidirectional mappings between undersampled and fully-sampled data in order to enforce self-consistency of network inputs or outputs [12], [18], [61]. However, unpaired models can require substantially larger datasets for training compared to paired models [20]. Lastly, several prominent approaches were proposed to train unsupervised models in the absence of ground truth. Scan-specific models learn non-linear interpolation kernels from an auto-calibration region in undersampled acquisitions, and then recover missing k-space samples via interpolation during inference [62]–[64]. These models do not require a priori training, but their performance relies on the assumption that local dependencies between

samples are largely invariant across k-space. Self-supervised models are trained via proxy loss terms substituted for the true reconstruction loss [21]. Specifically, data-consistency loss calculated on samples available in undersampled acquisitions can serve as a proxy [22], [23]. An alternative is to leverage statistical estimators for mean-squared reconstruction loss [27]. Such estimators are often analytically derived, so their use might be limited to specific types of loss functions. Recent studies have also proposed to mask a subset of k-space samples in undersampled acquisitions, where the masked subset is used to define a data-consistency loss [24], [25]. Based on this loss, self-supervised models are trained to recover masked-out samples from remaining samples. A related image-domain strategy has been proposed for multi-image MRI data [26], [65]. In this case, models can be trained to predict a masked subset of undersampled frames in dynamic MRI from remaining frames [66], or to map between differently undersampled versions of a given fully-sampled acquisition [67]. Because the indirect MRI prior to reduce aliasing artifacts is learned from undersampled data, self-supervised models might show suboptimal performance at high acceleration factors. Moreover, unsupervised methods that used fixed model weights during inference might suffer from suboptimal generalization to test data [68], [69].

A collective attribute of the aforementioned methods to limit supervision on raw data is that they are based on conditional models that map undersampled acquisitions to MR images. Thus, they involve inherent supervision regarding the imaging operator, and they are typically retrained for varying coil-array configurations and sampling densities [24]. Removing instead supervision related to the imaging operator, a fundamentally different approach decomposes the reconstruction process to decouple the MRI prior from the imaging operator. The deep image prior (DIP) approach employs an unconditional model

that maps relatively low-dimensional latent variables onto images as a native MRI prior [28]. The imaging operator is embedded during inference, and the prior is adapted to minimize a data-consistency loss on undersampled test data. This decoupled approach introduces flexibility in employing the same prior for various different imaging operators. Yet, DIP methods pervasively use untrained CNN architectures with randomly initialized parameters [28]–[31], [70]. In turn, untrained priors might be suboptimal in capturing the distribution of MR images [32], [33], and CNN models can generally suffer from limited sensitivity toward long-range spatial interactions [35]–[37].

III. THEORY

Inspired by the DIP framework, here we introduce a reconstruction method for accelerated MRI based on a deep generative model that maps noise and latent variables onto MR images. Unlike previous methods, this mapping is based on a style-generative adversarial network with transformer blocks. The network learns an MRI prior in pre-training phase, followed by a zero-shot inference phase where it is adapted to reconstruct undersampled acquisitions. We first overview the inverse problem formulation in accelerated MRI and DIP. We then describe the fundamental building blocks of SLATER.

A. Accelerated MRI Reconstruction

In accelerated MRI, undersampled k-space acquisitions are performed to speed up scans, typically with variable-density random sampling patterns:

$$F_p C m = y_s \qquad (1)$$

F_p is the partial Fourier operator defined by the sampling pattern, C denotes coil sensitivities, m is the target MR image and y_s are the acquired multi-coil k-space samples. The target m must be computed given the available data y_s. However, the linear system in Eq.1 is underdetermined, so MRI reconstruction is an ill-posed inverse problem. To obtain high-quality reconstruction, the solution has to be regularized with additional prior information on MR images:

$$\widehat{m} = \underset{m}{\mathrm{argmin}} \| y_s - F_p C m \|_2^2 + H(m) \qquad (2)$$

where $\widehat{m}$ is the reconstructed image, and $H(m)$ is the regularization term. The regularization term can be designed based on sparsity priors [3], [4], structured low-rank priors [71], [72], or network priors [6].

B. Deep Image Prior

The DIP framework has recently been introduced for unsupervised learning in computer vision tasks, including super resolution, inpainting and denoising [73]. DIP observes that local filtering operations in CNNs constrain the set of images that can be generated, so untrained CNNs can serve as native image regularizers. DIP performs random initialization of network inputs and weights without any pretraining. During inference, network parameters are optimized by enforcing consistency with the available corrupted image. Thus, DIP projects the corrupted image onto the space of CNN-generatable images to filter out corruptions such as blur or noise. DIP can be adopted for MRI reconstruction as follows:

$$\theta^* = \underset{\theta}{\mathrm{argmin}} \| F_p C d_\theta(z) - y_s \|_1 \qquad (3)$$

where θ are network weights, θ^* are optimized network weights, z are latent variables, $d_\theta(z)$ is the network mapping from latents onto the reconstructed image [14], [28]. Both network weights and latents are randomly initialized, and the optimization in Eq.3 is performed over θ, while z is fixed. In contrast to mainstream learning-based methods, DIP inverts a random network prior to identify weights that are most consistent with the corrupted image. The reconstruction can then be expressed as:

$$\widehat{m} = d_{\theta^*}(z) \qquad (4)$$

Despite its prowess in MRI reconstruction, models with randomly initialized parameters setup a generic image prior that may not be as strongly-tuned towards the distribution of MR data as trained models [32]. Furthermore, DIP is traditionally based on CNN architectures, where local kernels introduce suboptimal sensitivity in capture of long-range spatial interactions [35].

C. Adversarial Transformer Model

SLATER is based on an unconditional adversarial network that receives noise and latent variables to generate MR images [38]. Here we adopt a style-generative architecture given the success of this model family in computer vision tasks [74]. We further propose to build the network layers with cross-attention transformer blocks as inspired by a recent study on natural image synthesis [75]. Self-attention transformers that compute interactions among all image pixels are prohibitive at relatively high spatial resolutions [36], [76], [77]. Instead, cross-attention transformers enable efficient capture of long-range context based on attentional interactions between low-dimensional latent variables and high-dimensional image features. Our model contains three sub-networks (Fig.1): a synthesizer that generates MR images; a mapper that prepares the set of latent variables input to the synthesizer; and a discriminator. The synthesizer aims to generate realistic images, while the discriminator aims to distinguish actual images from synthesized images. Thus, the discriminator aids the synthesizer in capturing an MRI prior that reflects the distribution of high-quality MR images.

1) Synthesizer (G): The synthesizer contains a total of N_L layers, each comprising a convolutional upsampling block to progressively increase image resolution followed by a cross-attention transformer block. Prior style-generative models typically use a global latent variable at each layer to control high-level image features related to style [74]. In SLATER, in addition to a global latent ($w_g \in \mathbb{R}^{L_s}$, L_s: dimensionality), the synthesizer receives as input K local ($W_l \in \mathbb{R}^{K \times L_s}$) latent variables at each layer. The global latent is still used to perform a spatially-uniform affine transformation of convolutional feature maps, and so it modulates high-level image features. Meanwhile, local latents are used to
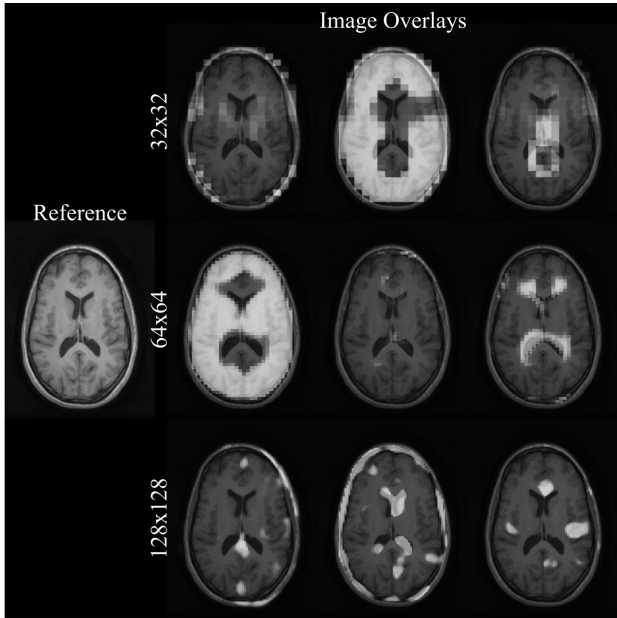
Fig. 3. Cross-attention maps in SLATER for a T_1-weighted acquisition. Sample maps from the first CA sub-block are displayed by overlay onto the respective MR image across three resolutions (i.e., at network layers 4-6). Attention maps for separate latents typically show segregated spatial distribution, and tend to group tissue clusters with similar signal intensity and texture broadly distributed across the image.

perform spatially-selective modulation of feature maps via cross-attention mechanisms. Each local latent variable focuses on a learned group of feature-map locations, so local latents serve to modulate relatively lower-level image features. For enabling cross-attention mechanism, a sinusoidal position encoding is used for spatial feature maps, whereas a learnable position encoding is assumed for local latent variables as their positions are unknown prior to training. Each cross-attention transformer block contains a serial cascade of the following sub-blocks: (cross-attention (CA^1), noise injection (NI^1), style-modulated convolution (SC), cross-attention (CA^2), noise injection (NI^2)). These sub-blocks are detailed below.

a) Cross-attention (CA^1): CA derives contextual representations by mediating attentional interaction between W_l and input feature maps $X_i^0 \in \mathbb{R}^{h_1 \times h_2 \times u}$, where h_1 and h_2 denote height and width depending on the resolution at the ith layer, and u is the number of feature channels. Let $X_{i,vec}^0 \in \mathbb{R}^{(h_1 \times h_2) \times u}$ be the vectorized form of X_i^0 along the spatial dimensions. Attention maps $\tilde{A}_{i,maps}^1 \in \mathbb{R}^{(h_1 \times h_2) \times K}$ that characterize the relation between W_l and $X_{i,vec}^0$ are as follows (see Fig. 3 and Supp. Fig. 2 for sample maps):

$$\tilde{A}_{i,maps}^1 = smax\left(\frac{\tilde{q}_i^{\,1}(X_{i,vec}^0 + PE_{i,X}^1)\tilde{k}_i^{\,1}(W_l + PE_{W_l})^T}{\sqrt{u}}\right) \tag{5}$$

where $smax$ is the softmax function, $\tilde{q}_i^{\,1}(.) \in \mathbb{R}^{(h_1 \times h_2) \times u}$ are queries that receive $X_{i,vec}^0$ added with layer-specific position encoding variables $PE_{i,X}^1 \in \mathbb{R}^{(h_1 \times h_2) \times u}$ and perform a learnable linear projection, $\tilde{k}_i^{\,1}(.) \in \mathbb{R}^{K \times u}$ are keys that

receive W_l added with position encoding variables $PE_{W_l} \in \mathbb{R}^{K \times L_s}$ and perform a learnable linear projection. $PE_{i,X}^1$ are taken as learnable linear projections of pre-defined sinusoidal encoding variables (see Supp. Text I-A), while PE_{W_l} are randomly initialized learnable position encoding variables. Note that $PE_{i,X}^1$ and $A_{i,maps}^1$ are vectorized along the spatial dimensions. Attention feature maps ($\tilde{A}_{i,feat}^1 \in \mathbb{R}^{(h_1 \times h_2) \times u}$) are then obtained by multiplying $A_{i,maps}^1$ with values that are learnable linear projections ($\tilde{v}_i^{\,1}(.) \in \mathbb{R}^{K \times u}$).

$$\tilde{A}_{i,feat}^1 = \tilde{A}_{i,maps}^1 \, \tilde{v}_i^{\,1}(W_l + PE_{W_l}) \tag{6}$$

Lastly, input feature maps are modulated via a single attention head that scales and shifts $X_{i,vec}^0$ with learnable linear projections of $\tilde{A}_{i,feat}^1$ to output $X_{i,vec}^1 \in \mathbb{R}^{(h_1 \times h_2) \times u}$.

$$X_{i,vec}^1 = \tilde{\gamma}_1(\tilde{A}_{i,feat}^1) \odot \left(\frac{X_{i,vec}^0 - \mu(X_{i,vec}^0)}{\sigma(X_{i,vec}^0)}\right) + \tilde{b}_1(\tilde{A}_{i,feat}^1) \tag{7}$$

where $\tilde{\gamma}_1(.) \in \mathbb{R}^{(h_1 \times h_2) \times u}$ and $\tilde{b}_1(.) \in \mathbb{R}^{(h_1 \times h_2) \times u}$ are learnable linear projections, μ denotes mean, σ denotes variance, and $\odot$ is the Hadamard product. The mapping through CA^1 can thus be summarized as: $X_i^1 = CA^1(X_i^0)$, where $X_i^1 \in \mathbb{R}^{h_1 \times h_2 \times u}$ is the matrix form of $X_{i,vec}^1$.

b) Noise injection (NI^1): To improve control over variability in fine details of feature maps, noise variables are injected onto modulated feature maps from CA^1. Given input $X_i^1 \in \mathbb{R}^{h_1 \times h_2 \times u}$ to NI^1, the output ($X_i^2 \in \mathbb{R}^{h_1 \times h_2 \times u}$) can be expressed as:

$$X_i^2 = \begin{bmatrix} X_i^{1,1} + \alpha_i^1 n_i^1 \\ \vdots \\ X_i^{1,u} + \alpha_i^1 n_i^1 \end{bmatrix} \tag{8}$$

where $X_i^{1,e} \in \mathbb{R}^{h_1 \times h_2}$ denotes e^{th} channel of X_i^1, $n_i^1 \in \mathbb{R}^{h_1 \times h_2}$ is noise added to each eth channel and α_i^1 is a learnable scalar. Note that the learnable noise variables in n_i^1 are initiated via random sampling from a standard normal distribution for each spatial location. During the course of learning, mean and standard deviation of noise variables are normalized to $(0, 1)$ across the spatial dimensions. The final mapping through NI^1 is expressed as: $X_i^2 = NI^1(X_i^1)$.

c) Style modulated convolution (SC): Transformer blocks characteristically contain a feed-forward neural network (FFNN) sub-block following the attention sub-block to extract hidden features of attention-based contextual representations. For computational efficiency, here a convolutional FFNN is utilized to locally refine contextual representations while modulating feature maps to control high-level style features [74]. Analogous to adaptive instance normalization (AdaIN) in style-transfer models, modulation is achieved via an affine transformation that controls the scale of feature maps [78]. Yet, we opted for a more compact implementation based on style-modulated convolution, with comparable complexity to convolution augmented with

AdaIN and only requiring a trivial scaling operation over basic convolution [74]. Accordingly, feature maps are scaled by modulating convolution kernels in SC, where kernel weights are multiplied via a learnable linear projection $w_{i,s} \in \mathbb{R}^u$ of the global latent variable w_g:

$$
\theta' = \begin{bmatrix}
w_{i,s}^1 \theta_i^{1,1} & w_{i,s}^1 \theta_i^{1,2} & \dots & w_{i,s}^1 \theta_i^{1,u} \\
w_{i,s}^2 \theta_i^{2,1} & w_{i,s}^2 \theta_i^{2,2} & \dots & w_{i,s}^2 \theta_i^{2,u} \\
\vdots & \vdots & \ddots & \vdots \\
w_{i,s}^u \theta_i^{u,1} & w_{i,s}^u \theta_i^{u,2} & \dots & w_{i,s}^u \theta_i^{u,u}
\end{bmatrix} \quad (9)
$$

where $\theta_i^{e,h} \in \mathbb{R}^{r \times r}$ denotes the two-dimensional (2D) kernel for e^{th} input and h^{th} output channel, $w_{i,s}^e$ is the scaling coefficient for e^{th} input channel. Note that $\theta_i' \in \mathbb{R}^{r \times r \times u \times u}$ is a 4D tensor of modulated kernels, so the matrix expression in Eq. 9 depicts formation along the third and fourth tensor dimensions. While convolution with the modulated kernels changes the relative scaling of input feature channels for controlling style, it can also alter the overall scale of the output feature map for each channel. To restore output feature maps to unit standard deviation, the modulated kernels are normalized to unit-norm across the output channel dimension. Following the notation in Eq. 9, this normalization can be expressed as:

$$
\theta_i'' = \begin{bmatrix}
\dfrac{\theta_i'^{1,1}}{\sqrt{\sum_c (\theta_i'^{c,1})^2}} & \dfrac{\theta_i'^{1,2}}{\sqrt{\sum_c (\theta_i'^{c,2})^2}} & \dots & \dfrac{\theta_i'^{1,u}}{\sqrt{\sum_c (\theta_i'^{c,u})^2}} \\
\dfrac{\theta_i'^{2,1}}{\sqrt{\sum_c (\theta_i'^{c,1})^2}} & \dfrac{\theta_i'^{2,2}}{\sqrt{\sum_c (\theta_i'^{c,2})^2}} & \dots & \dfrac{\theta_i'^{2,u}}{\sqrt{\sum_c (\theta_i'^{c,u})^2}} \\
\vdots & \vdots & \ddots & \vdots \\
\dfrac{\theta_i'^{u,1}}{\sqrt{\sum_c (\theta_i'^{c,1})^2}} & \dfrac{\theta_i'^{u,2}}{\sqrt{\sum_c (\theta_i'^{c,2})^2}} & \dots & \dfrac{\theta_i'^{c,u}}{\sqrt{\sum_c (\theta_i'^{c,u})^2}}
\end{bmatrix} \quad (10)
$$

where $\theta_i'' \in \mathbb{R}^{r \times r \times u \times u}$ denotes de-modulated kernels and c corresponds to channel index. Finally, the output feature maps ($X_i^3 \in \mathbb{R}^{h_1 \times h_2 \times u}$) can be computed as:

$$
X_i^3 = \begin{bmatrix}
\sum_c X_i^{2,c} \circledast \theta_i''^{c,1} \\
\vdots \\
\sum_c X_i^{2,c} \circledast \theta_i''^{c,u}
\end{bmatrix} \quad (11)
$$

where $X_i^{2,c} \in \mathbb{R}^{h_1 \times h_2}$ are input feature maps, $\circledast$ is convolution. The mapping through the convolution sub-block SC^1 is expressed as: $X_i^3 = SC(X_i^2)$.

d) Cross-attention (CA^2): A second cross-attention sub-block is used following the SC sub-block to further boost sensitivity of the model to global context. This maps the output of the SC sub-block to attention-modulated feature maps ($X_i^4 \in \mathbb{R}^{h_1 \times h_2 \times u}$) as: $X_i^4 = CA^2(X_i^3)$.

e) Noise-injection (NI^2): Finally, noise variables are injected to control variability in the fine details of feature maps following the second CA sub-block: $X_i^5 = NI^2(X_i^4)$. where $X_i^5 \in \mathbb{R}^{h_1 \times h_2 \times u}$ is the output of NI^2.

2) Mapper (M): The mapper projects independent and identically distributed random variables ($Z = z_1, z_2, \dots, z_K, z_g; z \in \mathbb{R}^{1 \times L_s}$) onto a refined set of local and global latent variables ($w_1, w_2, \dots, w_K, w_g; w \in \mathbb{R}^{1 \times L_s}$) expected by the synthesizer [74]. M is a multi-layered architecture comprising a first stream dedicated to the global latent and a second stream of dedicated to local latents (see Supp. Fig. 1). The global latent is processed with a cascade of fully-connected layers [74]. Local latents are instead processed with self-attention blocks to enable interactions among these variables. Self-attention blocks contain a cascade of self-attention (SA), fully-connected (FC^1), fully-connected (FC^2) sub-blocks as detailed below.

a) Self-attention (SA): At the ith self-attention block, intermediate activations for the K latent variables are concatenated as input:

$$
Z_i^0 = (z_1 \oplus_1 \cdots \oplus_1 z_K)_i^0 \quad (12)
$$

where $Z_i^0 \in \mathbb{R}^{K \times L_s}$ and $\oplus_1$ is the concatenation operator along the first dimension. Attention maps ($\hat{A}_{i,maps} \in \mathbb{R}^{K \times K}$) are first obtained:

$$
\hat{A}_{i,maps} = smax \left(\frac{\hat{q}_i(Z_i^0 + PE_{Z^0})\hat{k}_i(Z_i^0 + PE_{Z^0})^T}{\sqrt{L_s}} \right) \quad (13)
$$

where $PE_{Z^0} \in \mathbb{R}^{K \times L_s}$ denotes learnable position encoding variables, $\hat{q}_i(.) \in \mathbb{R}^{K \times L_s}$ and $\hat{k}_i(.) \in \mathbb{R}^{K \times L_s}$ are queries and keys respectively. Attention feature maps ($\hat{A}_{i,feat} \in \mathbb{R}^{K \times L_s}$) are given as:

$$
\hat{A}_{i,feat} = \hat{A}_{i,maps}\, \hat{v}_i(Z_i^0 + PE_{Z^0}) \quad (14)
$$

where $\hat{v}_i(.) \in \mathbb{R}^{K \times L_s}$ are values. $\hat{A}_{i,feat}$ is then used to scale and shift Z_i^0:

$$
Z_i^1 = \hat{\gamma}(\hat{A}_{i,feat}) \odot \left(\frac{Z_i^0 - \mu(Z_i^0)}{\sigma(Z_i^0)} \right) + \hat{b}(\hat{A}_{i,feat}) \quad (15)
$$

where $Z_i^1 \in \mathbb{R}^{K \times L_s}$ is the output, $\hat{\gamma}(.)$ and $\hat{b}(.)$ are learnable linear projections. Lastly, Z_i^1 is decomposed into individual latents, $Z_i^1 \to \{z_1, \dots, z_K\}_i^1$.

b) Fully connected (FC): Next, each latent is separately processed with two fully-connected sub-blocks (FC^1 and FC^2):

$$
\{z_j\}_i^2 = FC^2(FC^1(\{z_j\}_i^1)) \quad (16)
$$

where $\{z_j\}_i^2 \in \mathbb{R}^{1 \times L_s}$ is the output FC^2 for the jth local latent.

3) Discriminator (D): Adversarial models involve an interplay between the synthesizer that generates images and a separate discriminator sub-network [79]. In SLATER, D aims to accurately distinguish images generated by the synthesizer from actual MR images. A feed-forward architecture is employed here with a cascade of convolutional layers

augmented with several fully-connected layers. The mapping through D can be compactly expressed as:

$$x_D = D_{\theta_D}(x) \qquad (17)$$

where $x_D \in \mathbb{R}^1$ is the output of the discriminator, and x is either an actual MR image (x_r) or an image generated by the synthesizer $G(M(Z))$.

D. Self- Versus Cross-Attention Transformers

The main components of vanilla transformers are a self-attention (SA) sub-block followed by a feed-forward neural network (FFNN) sub-block [36]. Self-attention mechanisms serve to explicitly relate different positions within an image to compute contextual representations at each image pixel. Thus, SA is the primary component that learns long-range dependencies, while FFNN performs nonlinear transformations to extract hidden features of attention-based contextual representations. While vanilla transformers with fully-connected FFNN are common in natural language processing [80], computer vision studies often introduce vision-specific modifications for computational efficiency [75]. Cross-attention transformers in SLATER involve two key modifications: cross-attention sub-blocks that use a compact set of latents implicitly relating image pixels to learn contextual representations, and a convolutional FFNN to improve computational efficiency. Here, we overview self-attention and cross-attention transformers in terms of their ability to capture long-range spatial relationships and model complexity. For simplicity, remaining operations in transformers such as normalization or skip connections are ignored, and only a single feature channel and a single attention head are considered.

1) Self-Attention Transformer: Given an input feature map $X_{vec}^0 \in \mathbb{R}^{(h_1 \times h_2)}$, let a, b be two distant pixels whose $r \times r$ neighborhoods $\Delta a, \Delta b$ do not spatially overlap. A convolution sub-block with kernel size r would perform localized processing for each pixel in its $r \times r$ neighborhood:

$$X_{vec}^1 = Conv(X_{vec}^0) \qquad (18)$$

where $X_{vec}^1 \in \mathbb{R}^{(h_1 \times h_2)}$. Intensities of a, b in X_{vec}^1 will thus be conditionally independent given intensities of X_{vec}^0 in $\Delta a, \Delta b$:

$$E\{X_{vec}^1[a]X_{vec}^1[b] \mid X_c\} = E\{X_{vec}^1[a] \mid X_c\}E\{X_{vec}^1[b] \mid X_c\} \qquad (19)$$

where $E\{.\}$ denotes expectation, $X_c = X_{vec}^0[\Delta a \cup \Delta b]$ and $\cup$ is the union operator. As a result, convolutional processing does not leverage long-range relationships in feature maps.

In contrast, the SA sub-block in a transformer characterizes the relation between all spatial locations in $X_{vec}^0 \in \mathbb{R}^{(h_1 \times h_2)}$ via an attention map:

$$A_{maps} = smax\left(q(X_{vec}^0 + PE_X)k(X_{vec}^0 + PE_X)^T\right) \qquad (20)$$

where $PE_X, q, k \in \mathbb{R}^{(h_1 \times h_2)}$ are positional encoding, query, and key, and $A_{maps} \in \mathbb{R}^{(h_1 \times h_2) \times (h_1 \times h_2)}$. Attention-modulated feature map is then derived:

$$A_{feat} = A_{maps}v(X_{vec}^0 + PE_X) \qquad (21)$$

where $A_{feat} \in \mathbb{R}^{(h_1 \times h_2)}$, and $v \in \mathbb{R}^{(h_1 \times h_2)}$ is value. Intensities of A_{feat} at pixels a, b are:

$$A_{feat}[a] = A_{maps}[a, 1 : (h_1 \times h_2)] \, v[1 : (h_1 \times h_2)]$$
$$A_{feat}[b] = A_{maps}[b, 1 : (h_1 \times h_2)] \, v[1 : (h_1 \times h_2)] \qquad (22)$$

As seen here, both $A_{feat}[a]$ and $A_{feat}[b]$ are functions of all pixels in v and thereby X_{vec}^0. Next, the FFNN sub-block processes A_{feat} to extract hidden representations, typically via a fully-connected architecture:

$$X_{vec}^1 = FC(A_{feat}) \qquad (23)$$

where $X_{vec}^1 \in \mathbb{R}^{(h_1 \times h_2)}$. Note that $A_{feat}[a]$ and $A_{feat}[b]$ are readily dependent on all pixels. Intensities of pixels a, b in output feature maps are then statistically dependent even when conditioned on $X_c = X_{vec}^0[\Delta a \cup \Delta b]$:

$$E\{X_{vec}^1[a]X_{vec}^1[b] \mid X_c\} \neq E\{X_{vec}^1[a] \mid X_c\}E\{X_{vec}^1[b] \mid X_c\} \qquad (24)$$

Thus, vanilla transformers utilize dependencies across distant pixels to compute contextual representations.

The SA sub-block examines interactions among all possible pairs of pixels as described in Eqs. 20 and 22, so it has a computational complexity of $O((h_1 \times h_2)^2)$. Likewise, the FFNN sub-block as described in Eq. 23 exhaustively considers inter-pixel interactions with a complexity of $O((h_1 \times h_2)^2)$. This quadratic complexity with respect to image size limits the applicability of self-attention transformers at relatively high resolutions encountered in MRI [36].

2) Cross-Attention Transformer: Instead of exhaustively modeling inter-pixel interactions in high-dimensional feature maps, the CA sub-block uses a small set of K latent variables to implicitly characterize these interactions:

$$A_{maps} = smax\left(q(X_{vec}^0 + PE_X)k(W_l + PE_{W_l})^T\right) \qquad (25)$$

where $A_{maps} \in \mathbb{R}^{(h_1 \times h_2) \times K}$, $W_l \in \mathbb{R}^{K \times L_s}$: local latent variables, $PE_{W_l} \in \mathbb{R}^{K \times L_s}$: positional encoding for W_l. Attention-modulated feature map can be expressed as:

$$A_{feat} = A_{maps}v(W_l + PE_{W_l}) \qquad (26)$$

where $A_{feat} \in \mathbb{R}^{(h_1 \times h_2)}$, and $v \in \mathbb{R}^K$ is value. In turn, intensities of pixels a, b are:

$$A_{feat}[a] = A_{maps}[a, 1 : K] \, v[1 : K]$$
$$A_{feat}[b] = A_{maps}[a, 1 : K] \, v[1 : K] \qquad (27)$$

As seen above, both $A_{feat}[a]$ and $A_{feat}[b]$ are functions of all latent variables in W_l, so they are statistically dependent.

Next, a convolutional FFNN processes A_{feat} to extract hidden representations:

$$X_{vec}^1 = Conv(A_{feat}) \qquad (28)$$

where $X_{vec}^1 \in \mathbb{R}^{(h_1 \times h_2)}$. Although convolutional processing is local, the dependency introduced in the CA sub-block carries over to the output feature maps:

$$E\{A_{feat}[a]A_{feat}[b]|X_c\} \neq E\{A_{feat}[a]|X_c\}E\{A_{feat}[b] \mid X_c\}$$
$$E\{X_{vec}^1[a]X_{vec}^1[b] \mid X_c\} \neq E\{X_{vec}^1[a] \mid X_c\}E\{X_{vec}^1[b] \mid X_c\} \qquad (29)$$

Therefore, cross-attention transformers can model long-range dependencies to compute contextual representations.

The CA sub-block examines interactions among image pixels and local latents as described in Eqs. 25 and 27, so it has a computational complexity of $O((h_1 \times h_2) \times K)$. The convolutional FFNN considers local interactions with a complexity of $O((h_1 \times h_2) \times r^2)$. Because $K \ll (h_1 \times h_2)$ and $r^2 \ll (h_1 \times h_2)$ typically, the cross-attention transformer achieves notably lower complexity to permit use at higher spatial resolutions.

E. Learning Procedures

SLATER uses a two-stage strategy towards MRI reconstruction with a pre-training phase to learn the MRI prior, followed by a zero-shot reconstruction phase to embed the imaging operator. These two phases are detailed below.

1) Pre-Training of the MRI Prior: Since SLATER completely decouples the MRI prior from the imaging operator, pre-training assumes no prior knowledge on the imaging operator such as undersampling patterns or coil sensitivity encoding. Instead, the adversarial transformer model is trained to capture a prior on coil-combined, complex MR images, derived from fully-sampled acquisitions. Note that the synthesizer in this unconditional model maps noise and latent variables onto MR images, unlike conditional models with explicitly defined input-output relationships (i.e. undersampled versus fully-sampled data). Therefore, SLATER's pre-training is categorized as an unsupervised generative modeling task where the distribution of MR images is learned so that new, random samples can be drawn from the distribution [81].

Adversarial models commonly involve synthesizer and discriminator sub-networks that are trained with inter-linked loss functions to improve quality of synthesized images [79]. In SLATER, the synthesizer along with the mapper that provides latent variables are trained to minimize a common adversarial loss based on non-saturating logistic function:

$$L_{G,M}(\theta_G, \theta_M) = -\mathrm{E}_{p(Z)}\{log(f(D(G_{\theta_G}(M_{\theta_M}(Z)))))\} \quad (30)$$

where $\mathrm{E}_{p(.)}$ is expectation with probability density p, $f(.)$ is the sigmoid function, θ_G are parameters of the synthesizer, θ_M are parameters of the mapper. This particular loss is preferred in order to prevent the saturation problem in adversarial learning, where the discriminator starts outperforming the synthesizer by a significant margin and learning stops prematurely [79].

Meanwhile, the discriminator is trained to minimize an adversarial loss based on non-saturating logistic function augmented with a gradient penalty term:

$$
\begin{aligned}
L_D(\theta_D) = &-\mathrm{E}_{p(Z)}\{log(1 - f(D_{\theta_D}(G(M(Z)))))\} \\
&- \mathrm{E}_{p(x_r)}\{log(f(D_{\theta_D}(x_r)))\} \\
&+ \frac{\eta}{2}\mathrm{E}_{p(x_r)}\{\|\nabla D_{\theta_D}(x_r)\|^2\}
\end{aligned}
\quad (31)
$$

where x_r denotes coil-combined, complex MR images derived from actual scans, η is the regularization parameter, and θ_D are the parameters of the discriminator. The first two terms define the adversarial loss, whereas the third term has been suggested to improve adversarial learning by enforcing limited gradients in x_r according to the learned distribution [82].

2) Zero-Shot Reconstruction: The learned MRI prior does not contain any information regarding the conditional mapping from undersampled to fully-sampled data. Thus, the prior is conjoined with the imaging operator during inference to reconstruct test data. To adapt the pre-trained generative model to the reconstruction task, a data-consistency loss is employed on undersampled acquisitions to optimize synthesizer parameters (Fig. 2). Note that zero-shot learning is an unsupervised task-adaptation approach, where a model trained for an initial task is later transferred to a different target task without using additional training samples [83]. Analogously, SLATER adapts its adversarial model pre-trained to perform generative modeling of MR images to perform MRI reconstruction without any extra training samples. Therefore, the inference phase of SLATER is categorized as zero-shot reconstruction.

During inference, we optimize all components of the synthesizer including noise (n), latent variables (W) and weights (θ_G) to minimize the data-consistency loss. The synthesizer outputs a coil-combined, complex MR image, which is back-projected onto individual coils given sensitivity estimates, and then Fourier transformed to select available k-space coefficients according to the undersampling pattern [55]. Consistency of acquired and reconstructed k-space coefficients in the test data is then computed:

$$
\begin{aligned}
L_{\hat{R}}(W, n, \theta_G) = &\lambda_1 \|F_p C G(W, n, \theta_G) - k_x\|_1 \\
&+ \lambda_2 \|F_p C G(W, n, \theta_G) - k_x\|_2
\end{aligned}
\quad (32)
$$

where $\hat{R}$: reconstructed image, k_x: acquired k-space coefficients, F_p: partial Fourier operator, G: synthesizer. Data consistency is taken as a weighted $\ell_1 - \ell_2$-norm loss, where (λ_1, λ_2): weightings of (ℓ_1, ℓ_2) loss components. This loss function is considered to offer a more balanced weighting of errors across k-space compared to ℓ_2-norm that can be over-sensitive to lower spatial frequencies [24]. During pre-training, a common set of latent variables produced by the mapper are input uniformly across all CA and SC sub-blocks, whereas sub-block specific noise is included. To improve performance during zero-shot reconstruction, latent variables for each sub-block within each layer are instead segregated for independent optimization. In Eq. 32, $W = W_{1,..,N_L}$ where W_i denotes the collection of latent variables for the ith layer and contains the local latents for the two CA sub-blocks ($W_{l,i}^{1,2}$) and the global latent for the SC sub-block ($w_{g,i}$); $n = n_{1,..,N_L}$ where n_i denotes the collection of noise components for the ith layer and contains noise for the two NI sub-blocks ($n_i^{1,2}$).

IV. METHODS

A. Network Architecture

Architecture of SLATER's synthesizer, mapper and discriminator are described below (see also Supp. Text I-B).

1) Synthesizer: The synthesizer is a multi-layer architecture where image resolution is progressively increased. Each layer comprises a convolutional upsampling block to increase image resolution by a factor of 2 followed by a cross-attention transformer block, and a skip connection to add the upsampled

input. The first layer receives a constant input randomly drawn from a standard normal distribution. As the input layer, the first layer does not contain the upsampling blocks and the first attention sub-block is replaced with an identity transformation. The last layer calculates the final synthesizer output. Attention sub-blocks are omitted in the last layer to preserve precise localization in high-resolution images with convolution operators, as contextual representations have been extracted in previous layers [36]. Two separate channels are used to output real and imaginary parts of images. The upsampling block uses transpose convolution, and upsampling and modulated convolutions have a kernel size of 3×3. The cross-attention transformer block contains a cascade of cross-attention and convolutional sub-blocks. The fully-connected sub-block in vanilla transformers is replaced with a convolutional sub-block to improve computational efficiency and permit use at high resolutions. Since the convolutional sub-block inherently focuses on local relationships, a second cross-attention block is used to reinforce long-range interactions. Note that the imaging matrix sizes differ between the IXI and fastMRI datasets analyzed here. The number of synthesizer layers was adjusted accordingly. The resolution of the final layer was set to 256×256 for IXI with a total of seven layers, whereas the resolution of the final layer was set to 512×512 for fastMRI with a total of eight layers. During unsupervised pretraining, MR images were zero-padded to the resolution of the final layer. During inference, the output of the synthesizer was cropped to match the matrix size of the MRI acquisition.

2) Mapper: The mapper comprises two streams of multi-layer architectures for processing local and global latent variables. The local stream has a total of five layers, with the first four containing a self-attention sub-block and the last containing a fully-connected sub-block. The global stream has a total of nine layers, each containing fully-connected sub-blocks.

3) Discriminator: The discriminator aggregates information across multiple spatial scales in a multi-layered architecture. Real and imaginary parts of synthesized and actual MR images are represented in separate channels. Each layer comprises a convolution block followed by downsampling by a factor of 2 and a skip connection to add the downsampled input. The downsampling block uses convolution with a kernel size of 3×3. The resolution of the first layer was set to 256×256 for IXI with a total of seven layers, and 512×512 for fastMRI with a total of eight layers.

B. Competing Methods

SLATER was comparatively demonstrated against state-of-the-art techniques based on supervised and unsupervised models, as well as a traditional method. For each technique, hyperparameter optimization was performed via cross-validation on a three-way split of subjects. Optimization was performed for number of epochs, number of inference iterations, and weights for regularization terms based on performance on the validation set. For supervised models, performance in the validation set was quantified as ℓ_2-norm difference between the reconstructed

and fully-sampled ground-truth images. For unsupervised models, validation performance was instead quantified as the ℓ_2-norm difference between reconstructed and available k-space samples in undersampled acquisitions. A single set of hyperparameters yielding near-optimal results in all tasks were selected for each technique. Please see Supp. Fig. 3 for sample performance curves in the validation set versus number of training epochs where training was continued up to 1500 for methods that perform generative modeling of MR images. During inference on test data, strict data consistency was enforced to the network outputs. Codes were run on an eight-core Intel Xeon E5-2690v3 CPU for LORAKS, and in parallel on five nVidia 2080 Ti GPUs for all network models.

1) SLATER: SLATER was first pretrained to map random noise and latent codes onto high-quality MR images. In the IXI dataset, the model was trained to map onto single-coil magnitude images. In fastMRI dataset, the model was instead trained to map onto coil-combined complex images, with real and imaginary channel outputs. To do this, coil sensitivity maps were derived via ESPIRiT with default parameters [84]. Using these estimates, an optimal linear combination on multi-coil complex images was then performed [85]. Pre-training was performed via the Adam optimizer with β_1:0.0, β_2:0.99, $\eta = 10$ and a learning rate of 0.001 as adopted from [74]. The dimensionality of latent variables were set as $K=16$ and $L_s=32$. Network weights were randomly initialized using a standard normal distribution. Inference was performed via the RMSprop optimizer with learning rate 0.1, momentum parameter 0.9, early-stopping and learning rate schedule as adopted from [74]. Cross-validation indicated 1000 iterations as a favorable early-stopping point for maintaining near-optimal performance and computational efficiency. For IXI, 470th pretraining epoch, $\lambda_1 = 1.0, \lambda_2 = 0.0$ were selected. For fastMRI, 1280th pretraining epoch, $\lambda_1 = 0.5, \lambda_2 = 0.5$ were selected. Separate SLATER models were trained to build MRI priors for each tissue contrast within each dataset. Reconstruction was then performed via inference on individual test data conditioned by the MRI prior.

2) LORAKS: A traditional parallel-imaging reconstruction based on low-rank modeling of local k-space neighborhoods was performed [86] via libraries in the LORAKS V2.1 toolbox [87]. Here, an autocalibrated reconstruction was performed where the structured low-rank matrix was formed based on limited image support assumption [86]. Accordingly, the k-space neighborhood radius and the rank of the resultant matrix were selected via cross-validation as: (2,6) for IXI, and (2,30) for fastMRI.

3) GAN_{sup}: A fully-supervised conditional generative adversarial network (GAN_{sup}) was trained using paired ground-truth and undersampled acquisitions. Network architecture and loss functions were adopted from [55]. Training was performed via the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, dropout regularization rate 0.5, and a learning rate of 0.0002. Training was continued over 100 epochs, with learning rate schedule from [55]. Network weights were randomly initialized using a normal distribution with zero mean and 0.02 standard deviation. Regularization parameters for (pixel-wise, perceptual, adversarial) losses were selected as (100,100,1). A separate

GAN$_{sup}$ model was trained for each contrast within each dataset and acceleration rate.

4) SSDU: A self-supervised version of the conditional GAN model in GAN$_{sup}$ was trained on undersampled data [24]. Acquired k-space samples were split into two sets of nonoverlapping points, where 60% of samples were used to estimate model weights and 40% were used to define the network loss. Analogous to Eq. 32, the network loss was taken as a weighted sum of ℓ_1-, ℓ_2-norm differences between recovered and acquired k-space samples. For both datasets, $\lambda_1 = 1.0, \lambda_2 = 0.0$ were selected. All other procedures were identical to GAN$_{sup}$. Separate SSDU models were trained for each contrast within each dataset and acceleration rate.

5) GAN$_{prior}$: Following [32], unsupervised pretraining on fully-sampled MRI data was performed using an unconditional GAN. Network architecture was adopted from [74] for fair comparison against SLATER. Training and inference procedures were identical to SLATER with minor modifications for enhanced performance. The synthesizer in GAN$_{prior}$ was trained to minimize the same loss as in Eq. 30 with an additional path length regularization parameter adopted from [74]. The discriminator was trained to minimize Eq. 31. A matching number of latents to SLATER were prescribed with L_s=512. For IXI, 60th pretraining epoch, $\lambda_1 = 1.0, \lambda_2 = 0.0$ were selected. For fastMRI, 180th pretraining epoch, $\lambda_1 = 1.0, \lambda_2 = 0.0$ were selected. Separate GAN$_{prior}$ models were trained for each contrast within each dataset.

6) SAGAN: Zero-shot learned reconstructions were also implemented using a self-attention GAN model. The network architecture was adopted from [74] for fair comparison. Training and inference procedures were identical to SLATER. Optimization of network weights was not performed as it was observed to degrade reconstruction performance. Instead contrast-specific epoch selection was adopted for SAGAN since in this case it yielded enhanced performance. For T$_1$ and T$_2$ reconstructions, 798th, 399th epochs in IXI, and 967th, 2661th epoch in fastMRI, along with $\lambda_1 = 1.0, \lambda_2 = 0.0$ were selected. Separate models were trained for each contrast within each dataset.

7) GAN$_{DIP}$, SAGAN$_{DIP}$, SLATER$_{DIP}$: DIP reconstructions were performed via untrained GAN$_{prior}$, SAGAN and SLATER models respectively. The inference procedures were identical to pre-trained counterparts. Network weights were randomly initialized using a normal distribution with zero mean and unit standard deviation.

C. Datasets

Demonstrations were performed on single-coil brain MRI data from IXI (http://brain-development.org/ixi-dataset/) and multi-coil brain MRI data from fastMRI [88]. T$_1$-weighted and T$_2$-weighted acquisitions were considered. In IXI, 25 subjects were used for training, 5 for validation and 10 for testing. Parameters for T$_1$-weighted scans are: repetition time (TR)=9.813 ms, echo time (TE)=4.603 ms, flip angle=8°, matrix size=256 × 256×150, voxel size=0.94 × 0.94 × 1.2 mm^3; and those for T$_2$-weighted scans are: TR=8178 ms, TE=100 ms, flip angle=90°, matrix size=256 × 256×130,
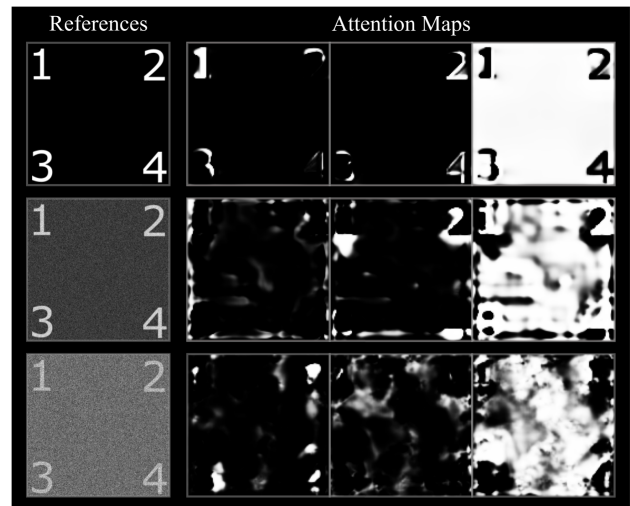

Fig. 4. Cross-attention maps in SLATER for a simulated phantom with varying levels of normally-distributed white noise. Sample maps from the first CA sub-block are displayed at 128 × 128 resolution (network layer 6). Relative to a peak signal intensity of 1, top, middle and bottom rows show results for no noise, noise variance of 0.01, and noise variance of 0.1, respectively.

voxel size=0.94 × 0.94 × 1.2 mm^3. In fastMRI, 100 subjects were used for training, 10 for validation and 40 for testing. Data from multiple sites are included with no common protocol. For consistency, only volumes with at least 10 cross-sections and acquired with at least 5 coils were selected. To reduce computational complexity, GCC [89] was used to decrease the number of coils to 5. For both datasets, subject selection and splitting was done sequentially. Data were retrospectively undersampled using variable-density random patterns [3]. Undersampling masks were generated based on a 2D normal distribution with covariance adaptively adjusted to obtain the desired acceleration rates of R=[4, 8].

D. Quantitative Assessments

To assess reconstruction quality, quantitative comparisons were performed against reference images Fourier-reconstructed from fully-sampled acquisitions. Both reconstructed and reference images were normalized to a maximum of 1 prior to measurement. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) were calculated between the reconstructed and reference images. In Tables, summary statistics for quantitative metrics were provided as mean ± standard deviation across test subjects. Statistical significance of differences between methods was assessed via nonparametric Wilcoxon signed-rank tests.

V. RESULTS

A. Cross-Attention Transformers

We first examined the utility of cross-attention transformers in capturing contextual representations via experiments on a simulated phantom. The phantom had numerical digits of unit intensity for four corners against a zero-intensity background (Fig. 4). This design creates a strong dependency among digit pixels in distant corners, and separately among background
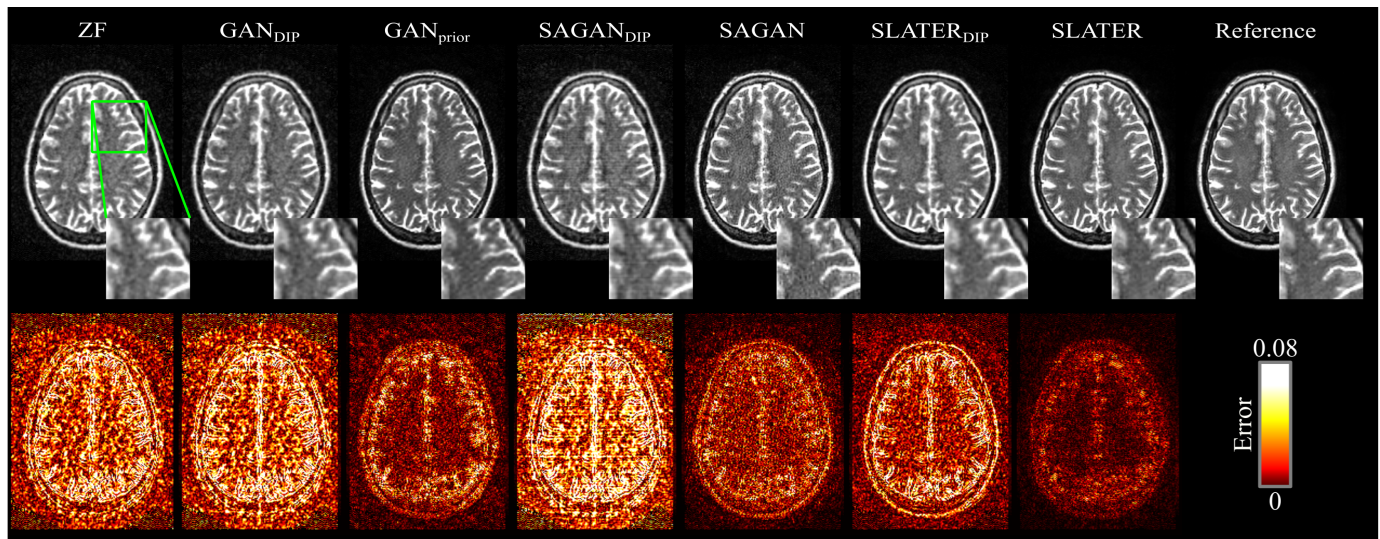
Fig. 5. Reconstructions of a representative T_2-weighted acquisition at R=4 are shown for the Fourier method (ZF), DIP methods (GAN$_{DIP}$, SAGAN$_{DIP}$, SLATER$_{DIP}$) and zero-shot reconstructions (GAN$_{prior}$, SAGAN, SLATER) along with the reference image. Zoom-in display windows are added to aid visualization of performance differences. Corresponding error maps are underneath the images for each method.

pixels. Normally-distributed white noise was added to vary the degree of contextual information present in phantom images by dampening inter-pixel correlations. DIP reconstructions at R=4 were then performed using SLATER. The spatial distribution of cross-attention maps in Eq. 5 characterize learned groupings of correlated pixels. Thus, we reasoned that the attention maps for local latent variables should span over distant albeit correlated image pixels for relatively limited noise levels, and the maps should degrade for higher noise levels due to weakened dependencies. To test this prediction, we inspected the attention maps for the phantom image in Fig. 4 (see Supp. Fig. 4 for details). As expected, attention maps clearly span across digit pixels or across background pixels for limited noise. Note that attention maps learned on brain images also manifest similar grouping of correlated albeit spatially-distant pixels (see Fig. 3 and Supp. Fig. 2 for representative maps). Yet, towards substantially higher noise levels in the simulated phantom, attention maps show a less clear grouping structure as contextual information is gradually weakened. Taken together, these results suggest that SLATER can capture long-range spatial interactions among distant image pixels.

B. Model Invertability

During inference on test data, SLATER inverts its generative model to identify noise, latents and network weights that are most consistent with the undersampled MRI acquisition. To evaluate model invertability, we compared SLATER against CNN-based and self-attention GAN models. Each model was used in both DIP and zero-shot reconstructions. Representative results on T_1- and T_2-weighted acquisitions from IXI at R=4 are displayed in Fig. 5 and Supp. Fig. 5 respectively. DIP reconstructions tend to suffer from visible loss of spatial resolution, and GAN$_{prior}$ and SAGAN have elevated noise and artifacts. In contrast, SLATER yields low residual errors and

high visual acuity. Performance metrics are listed in Table I. SLATER achieves superior performance against GAN$_{prior}$ and SAGAN in both DIP and zero-shot reconstructions ($p < 0.05$). Compared to the second-best method, SLATER yields 4.4dB higher PSNR and 7.7% higher SSIM in DIP, and 4.1dB higher PSNR and 5.9% higher SSIM in zero-shot reconstruction. Furthermore, SLATER yields 5.1dB higher PSNR and 7.0% higher SSIM over SLATER$_{DIP}$. These results indicate that the cross-attention transformer blocks in SLATER enhance model invertability compared to CNN architectures with or without self-attention and that the unsupervised pretraining stage in SLATER improves reconstruction performance.

C. Within-Domain Reconstructions

Next, we assessed within-domain reconstruction performance when the training and testing domains matched (e.g., T_1 reconstruction based on a T_1-prior for SLATER). SLATER was compared against LORAKS, GAN$_{sup}$, SSDU, GAN$_{prior}$ and SAGAN at R=4 and 8. Representative reconstructions are shown for IXI in Supp. Figs. 6 and 7, and for fastMRI in Fig. 6 and Supp. Fig. 8. SLATER yields lower residual errors and higher acuity in depicting detailed tissue structure than competing methods. Quantitative assessments are listed in Supp. Table I and Table II. SLATER achieves significantly enhanced reconstruction quality against all competing unsupervised methods ($p < 0.05$) offering 1.1dB higher PSNR and 1.1% higher SSIM compared to the second-best method. Furthermore, it offers 4.4dB higher PSNR and 5.5% higher SSIM compared to GAN$_{prior}$. It also yields higher performance than GAN$_{sup}$ in all tasks ($p < 0.05$), except at R=8 in IXI where the two methods perform similarly and R=8 in fastMRI where GAN$_{sup}$ yields higher SSIM. These results indicate that SLATER offers enhanced reconstruction quality for within-domain tasks compared to unsupervised baselines, while offering on par performance to a supervised baseline.

TABLE I

RECONSTRUCTION PERFORMANCE OF DIP AND ZERO-SHOT RECONSTRUCTIONS FOR T_1- AND T_2-WEIGHTED ACQUISITIONS IXI AT R=4 AND 8.
PERFORMANCE METRICS ARE PRESENTED AS MEAN ± STANDARD DEVIATION ACROSS TEST SUBJECTS. RESULTS ARE LISTED
FOR GAN_{PRIOR}, SAGAN, SLATER, THEIR DIP VARIANTS

	GAN_{DIP}		$SAGAN_{DIP}$		$SLATER_{DIP}$		GAN_{prior}		SAGAN		SLATER	
	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)
T_1, R=4	26.7±1.2	87.4±1.0	26.6±1.1	86.5±1.0	32.6±1.8	94.6±0.8	34.4±0.8	94.4±0.7	32.1±0.9	92.1±0.7	38.8±0.8	97.9±0.5
T_1, R=8	23.5±0.9	83.0±1.1	23.4±0.9	82.5±1.2	30.3±1.7	91.6±1.4	29.3±1.2	89.7±1.4	28.6±0.9	88.3±1.2	33.2±0.9	95.2±0.9
T_2, R=4	30.2±0.5	80.8±1.2	29.9±0.4	79.4±1.4	32.7±0.7	87.7±1.7	33.4±0.9	87.5±1.0	34.9±0.6	91.6±1.1	40.0±0.8	97.7±0.5
T_2, R=8	27.6±0.4	76.2±1.3	27.3±0.4	74.1±1.6	29.9±0.7	84.0±1.9	31.2±0.7	85.3±1.0	30.7±0.5	86.4±1.4	34.1±0.8	94.8±0.7

TABLE II

WITHIN-DOMAIN RECONSTRUCTION PERFORMANCE FOR T_1- AND T_2-WEIGHTED ACQUISITIONS IN FASTMRI AT R=4 AND 8

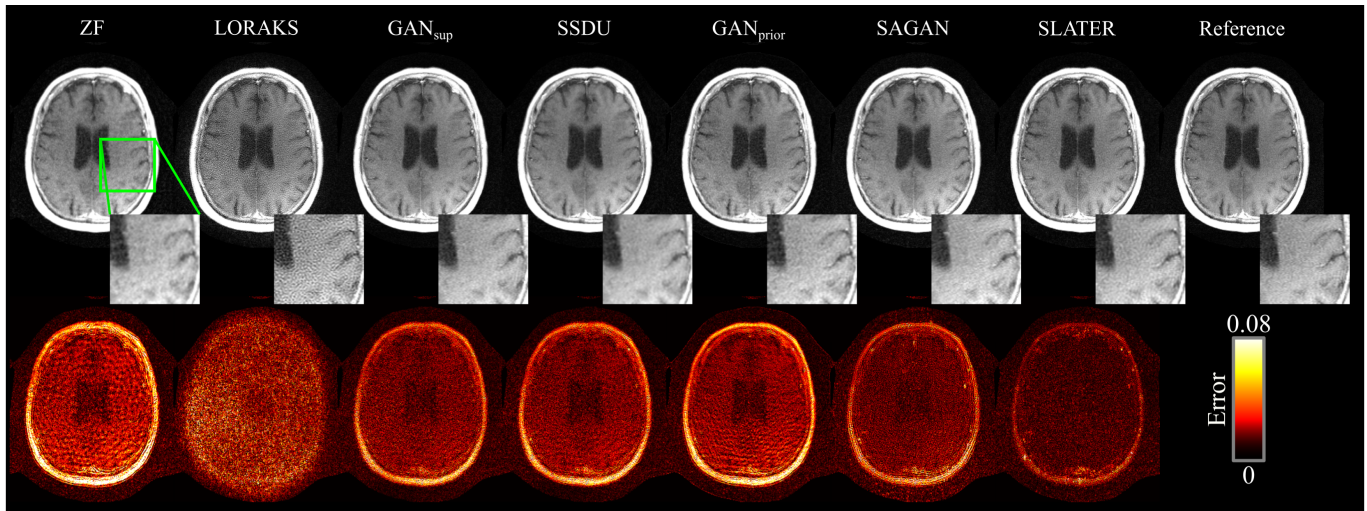|	LORAKS		GAN_{sup}		SSDU		GAN_{prior}		SAGAN		SLATER	
	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)
T_1, R=4	33.4±2.7	82.2±7.7	37.2±2.6	94.4±5.7	37.2±3.3	94.2±7.5	32.8±2.0	92.5±5.2	36.1±2.6	94.1±5.1	37.6±3.2	93.9±9.5
T_1, R=8	32.5±2.4	83.0±9.0	34.6±2.4	92.0±7.0	33.9±2.6	90.6±8.5	29.8±1.9	88.3±6.7	33.1±2.1	90.4±6.7	34.2±2.4	90.7±7.9
T_2, R=4	34.3±1.0	90.8±1.6	35.4±1.2	95.5±0.5	33.0±2.9	94.6±1.3	33.5±1.1	91.5±1.8	33.5±1.3	94.1±0.8	36.3±1.2	95.5±0.7
T_2, R=8	33.1±1.0	91.7±1.1	32.7±1.3	93.0±0.8	31.3±1.3	91.0±1.5	28.0±1.4	85.1±2.5	30.9±1.3	91.2±1.1	33.4±1.1	93.0±1.0

Fig. 6. Within-domain reconstructions of a T_1-weighted acquisition in fastMRI at R=4. Results are shown for ZF, LORAKS, GAN_{sup}, SSDU, GAN_{prior}, SAGAN and SLATER along with the reference image, and error maps in the bottom row.

D. Across-Domain Reconstructions

We then evaluated across-domain reconstruction performance when the tissue contrast in the test domain was different than that in the domain of the trained MRI prior (e.g., T_2 reconstructions based on a T_1-prior). SLATER was compared against LORAKS, GAN_{sup}, SSDU, GAN_{prior} and SAGAN at R=4. Representative results are shown for IXI in Supp. Figs. 9 and 10, and for fastMRI in Fig. 7 and Supp. Fig. 11. Quantitative assessments are listed in Table III. Again, SLATER yields lower residual errors and higher acuity in depicting detailed tissue structure. It also achieves superior reconstruction quality against all competing supervised and unsupervised methods ($p < 0.05$), offering 2.5dB higher PSNR and 1.5% higher SSIM compared to the second-best method. Static models that are not adapted during inference such as GAN_{sup} and SSDU can yield suboptimal performance when the data distribution differs between the training and testing domains. Compared to within-domain reconstruction,

we find that GAN_{sup} and SSDU suffer from 1.7dB PSNR, 1.7% SSIM loss on average in across-domain reconstruction. In contrast, adaptive models such as SLATER can attain more similar performance for across-domain and within-domain cases. We find that SLATER's average performance differs less than 0.01dB PSNR and 0.11% SSIM between the two cases. Note that SLATER reconstructions based on within-domain priors generally yield on par or better performance than those based on across-domain priors, yet there are few exceptions with higher across-domain performance that might be attributed to empirical limitations in gradient-descent optimization (see Discussion).

We also assessed across-domain reconstructions when the acceleration rate of the imaging operator differed between training and testing (i.e., trained at R=8 and tested at R=4). Performance measurements are listed in Supp. Table II. SLATER achieves superior reconstruction quality against all competing supervised and unsupervised methods ($p < 0.05$), offering 4.0dB PSNR, 1.5% SSIM improvement over the
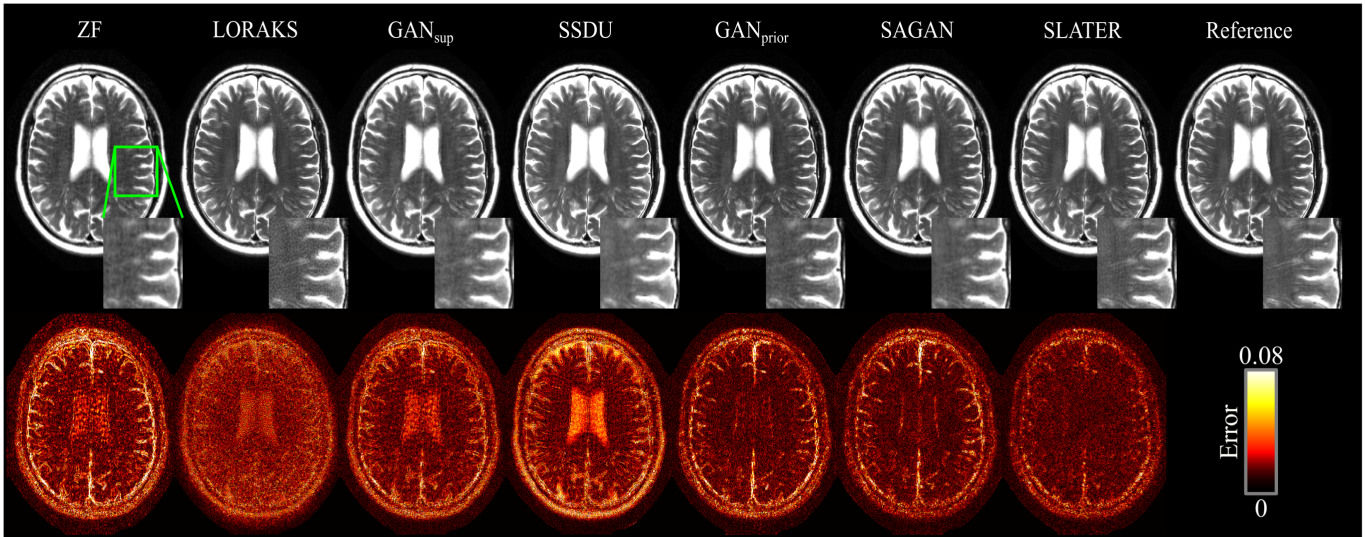
Fig. 7. Across-domain reconstructions of a T_2-weighted acquisition in fastMRI at R=4. Results are shown for ZF, LORAKS, GAN_{sup}, SSDU, GAN_{prior}, SAGAN and SLATER along with the reference image, and error maps in the bottom row.

TABLE III
ACROSS-DOMAIN RECONSTRUCTION PERFORMANCE FOR T_1- AND T_2-WEIGHTED ACQUISITIONS IN THE IXI AND FASTMRI DATASETS AT R=4. IN A->B, A IS THE TRAINING DOMAIN AND B IS THE TEST DOMAIN

	LORAKS		GAN_{sup}		SSDU		GAN_{prior}		SAGAN		SLATER	
	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)	PSNR	SSIM(%)
IXI T_1->T_2	35.4±0.5	92.3±1.2	36.8±0.7	95.3±0.6	34.8±1.0	90.3±1.4	36.4±0.7	93.2±0.6	32.3±0.7	86.1±1.3	39.9±0.8	97.7±0.5
IXI T_2->T_1	30.7±1.2	91.7±1.0	35.3±0.6	96.4±0.3	35.7±0.4	96.3±0.3	32.5±1.0	91.6±1.1	32.8±0.7	93.1±0.6	38.7±0.9	97.9±0.5
fastMRI T_1->T_2	34.3±1.0	90.8±1.6	33.8±1.2	94.1±0.7	34.8±1.1	95.1±0.7	33.5±1.5	94.3±0.8	34.5±1.2	94.4±0.9	36.2±1.1	94.6±1.0
fastMRI T_2->T_1	33.4±2.7	82.2±7.7	36.9±2.4	93.6±5.8	34.3±3.3	92.9±9.2	35.9±2.1	90.8±5.4	34.4±2.1	92.2±4.9	38.0±2.6	95.3±6.2

second-best method. Note that pre-training for zero-shot methods including SLATER is agnostic to the imaging operator, so these methods yield equivalent performance for within-domain and across-domain reconstructions. In contrast, GAN_{sup} and SSDU that are explicitly trained for a specific acceleration rate suffer from 3.3dB PSNR, 2.1% SSIM loss on average in across-domain reconstructions. Collectively, these results demonstrate that SLATER has improved generalization capabilities compared to static supervised and unsupervised models with fixed weights during inference, while still outperforming zero-shot reconstructions based on pure CNN architectures.

E. Ablation Experiments

We examined the contributions of individual parameter sets in SLATER that are optimized during inference. Variant models were built by progressively introducing optimization for latent variables, noise and network weights. Supp. Table III lists performance metrics for experiments on T_1- and T_2-weighted acquisitions in the IXI dataset at R=4. On average, the incurred performance gain in (PSNR, SSIM) is (4.1dB, 7.2%) with latent optimization, (6.5dB, 11.6%) with latent and noise optimization, and (10.8dB, 14.0%) with latent, noise and weight optimization. These results indicate the importance of each component in SLATER.

F. Weight Propagation

Average inference times for competing methods are listed in Table IV (see Supp. Table IV for training times in the IXI

TABLE IV
AVERAGE INFERENCE TIMES IN SEC PER CROSS SECTION. RUN TIMES WITHOUT AND WITH WEIGHT PROPAGATION (WP) ARE LISTED FOR ZERO-SHOT RECONSTRUCTIONS. SAGAN DOES NOT PERFORM WEIGHT OPTIMIZATION, SO IT IS UNAFFECTED BY WP

	LORAKS	GAN_{sup}	SSDU	GAN_{prior}	SAGAN	SLATER
No WP	4.10	0.002	0.002	12.73	7.74	14.80
With WP	–	–	–	2.52	7.74	2.63

dataset). Model adaptation to specific test samples in zero-shot reconstructions leads to prolonged inference. In principle, neighboring cross-sections with structural correlations should have similar reconstructions. Thus we reasoned that propagating model parameters across consecutive cross-sections should increase efficiency by accelerating the progression onto high-quality reconstructions. Accordingly, the network weights for a given cross-section at the end of its inference optimization were stored, and then used to initialize inference optimization for the next cross-section within each subject. Latent variables and noise were not shared as they control context and fine details that could vary between cross-sections. Table IV lists inference times with this weight propagation procedure (see Supp. Table V for reconstruction performance). While methods that do not perform model adaptation still provide faster inference, weight propagation substantially increases the inference efficiency for SLATER to improve its practicality.

VI. DISCUSSION

Here we propose zero-shot learned adversarial transformers for unsupervised reconstruction in accelerated MRI. To our

knowledge, this is the first study to introduce a transformer network for MRI reconstruction. Traditional GANs contain cascades of convolutional layers, which might limit ability to capture long-range spatial dependencies. We instead employed cross-attention transformer blocks to efficiently capture contextual image features. Note that self-attention among all feature map locations leads to excessive computational burden [36]. As such, self-attention modules following convolutional blocks have only been leveraged in layers with modest resolution to prevent quadratic complexity [35], [90]–[93]. In contrast, the cross-attention mechanism between low-dimensional latent variables and image features permits use at higher resolutions [75].

Our demonstrations clearly indicate the superiority of the proposed method over a fully-supervised GAN model (GAN_{sup}), a self-supervised model (SSDU), DIP based on CNN and self-attention GANs (GAN_{DIP}, $SAGAN_{DIP}$), and zero-shot reconstructions (GAN_{prior}, SAGAN). SLATER offers on par performance with GAN_{sup} for within-domain tasks, while it outperforms GAN_{sup} for across-domain tasks. SLATER also outperforms DIP by a substantial margin, which can be attributed to its MRI prior. Lastly, SLATER yields superior performance to GAN_{prior} that indicates the importance of transformer blocks in learning a high-fidelity prior.

Recent studies aiming to reduce supervision requirements related to raw data have proposed unpaired or unsupervised learning strategies for MRI reconstruction. A successful approach is to train models on unpaired samples of undersampled and fully-sampled acquisitions [12], [18], [61]. Another promising approach is to train models via self-supervision on undersampled data [21], [23]–[25]. Both approaches perform model training in conjunction with the imaging operator to map undersampled acquisitions to MR images. Trained model weights that reflect an indirect prior to reduce artifacts are then fixed during inference. Thus, reconstructions based on an across-domain prior can suffer from suboptimal generalization due to inconsistencies in the data distribution (e.g., tissue contrast) or the imaging operator (e.g., acceleration rate) between training and testing domains.

SLATER instead learns an MRI prior agnostic to the imaging operator that is later adapted to test data during inference. Prior adaptation involves an inference optimization with considerable computational burden compared to static models with fixed priors. Yet, our experiments indicate that it mitigates potential performance losses during across-domain reconstructions. Ideally, a within-domain prior would be expected to initialize the inference optimization closer to a desirable local minimum, resulting in on par or higher performance than an across-domain prior. Given complex loss surfaces for deep models, however, gradient-descent optimizers can occasionally show non-monotonic behavior approaching less preferred minima despite favorable initialization [94]. The few cases where an across-domain prior yielded relatively higher performance are likely attributed to such behavior, since we observed that the performance gap between the priors diminished with substantially prolonged optimizations in unreported experiments. That said, we opted for early stopping during inference to maintain a desirable compromise between performance

and inference time while also mitigating potential risks for over-fitting [30], [73].

The inference optimization performed during zero-shot reconstruction is closely related to DIP methods [73]. Compared to DIP with randomly initialized networks, here we show that SLATER yields enhanced reconstruction quality via a pre-trained prior that is more strongly tuned towards the distribution of MR images. Yet, it remains an important research topic to examine the convergence behavior of both untrained and pre-trained models. Several recent studies on accelerated imaging report theoretical and empirical validation of convergence for iterative optimization with traditional recovery methods [95] as well as untrained generative priors [96], [97]. Further theoretical and empirical research is warranted to investigate whether these results generalize to pre-trained priors.

Among recent efforts, the closest to the presented approach are prior-adaptation methods in [33] that uses a variational auto-encoder to learn patch priors and in [32] that uses a GAN to learn image priors. Our work differs from these efforts in the following aspects: (i) Compared to [33] that uses a patch-based implementation, we leverage an image-based implementation that can improve performance in leveraging non-local context. (ii) [33] uses a fixed patch prior during inference, whereas we perform test-sample specific adaptation of the prior. (iii) Compared to [32] that uses a CNN-based GAN, here we use cross-attention transformers to better capture long-range spatial dependencies. (iv) SLATER includes noise and latents at each synthesizer layer to better control image features.

Flow-based models have also been proposed to learn priors for inverse problems [98]. These models use a composition of fully-invertible flow steps that transform latent variables onto data samples, and they offer exact estimates of the maximum likelihood of data samples and improved immunity against representation errors due to bias. In comparison, adversarial models such as SLATER implicitly minimize discrepancy between generated and actual data distributions without explicit estimation of probability densities. Note, however, that flow-based models characteristically require high-dimensional latent spaces to retain a target level of expressiveness, which might result in a less favorable trade-off between computational burden and quality of generated samples [99]. It remains important future work to compare flow-based and adversarial methods as well as their hybrids in the context of MRI reconstruction.

Several lines of development can be pursued for the proposed technique. First, zero-shot reconstructions can be initialized with latent estimates based on conventional parallel imaging/compressed sensing reconstructions. This can increase computational efficiency by shortening the inference procedure. Second, the reconstruction loss in SLATER can be combined with SURE-type estimates or regularization terms on network weights [27]. Here we did not explicitly regularize the weights, latents or noise to allow higher degree of consistency to acquired data. When desired, network regularization and on-line error estimates can be introduced to reduce potential for overfitting. Third, SLATER learns a coil-combined

MRI prior and subsequently incorporates coil-sensitivity information during zero-shot reconstruction. This is achieved by back-projecting the synthesizer output onto individual coils. A powerful alternative is to build a synthesizer with a consistent prior across coils to generate multi-coil images [31]. Similar computational efficiency might be expected from both approaches that use a common MRI prior across coils. That said, effects of the decoupling strategy in SLATER against the channel-consistence prior on reconstruction quality remain to be investigated.

SLATER pre-training performs unsupervised generative modeling of coil-combined MR images derived from fully-sampled acquisitions. Combined with the imaging operator during inference, the learned prior is then used to perform reconstruction of undersampled acquisitions via unsupervised model adaptation. This decoupled approach bypasses the need for paired training datasets. Moreover, the MRI prior is agnostic to the imaging operator and flexibly adapted to the test domain. Thus, a learned prior can be used to reconstruct undersampled acquisitions at varying contrasts or acceleration rates. To lower reliance on fully-sampled datasets, pre-training can be instead performed on undersampled acquisitions. While the resultant prior will not entirely reflect the distribution of high-quality MR images, the model adaptation procedures in SLATER might limit potential performance losses.

Conventional supervised models perform poorly when training data are scarce, and domain shifts between the training and test sets are prominent. Recent studies have proposed supervised models of low complexity trained with limited or synthetic data to partly mitigate losses in generalization performance [100], [101]. While compact models are less prone to over-fitting, data-driven priors learned from limited or synthetic data can still elicit a biased representation of MRI data, increasing risk of inaccurate reconstruction for atypical anatomy [15], [16]. Future studies are warranted to assess the utility of supervised compact versus unsupervised models in domain generalization. Supervised and unsupervised procedures can be combined to facilitate model training on larger datasets by mixing paired and unpaired data. Unsupervised adaptation of such semi-supervised models during inference might further enhance reliability against atypical anatomy.

VII. CONCLUSION

Here we introduced a novel unsupervised MRI reconstruction based on an unconditional deep adversarial network. SLATER leverages cross-attention transformers to improve capture of contextual image features. Benefits of SLATER over state-of-the-art supervised and unsupervised methods were demonstrated in brain MRI. SLATER can also be adopted for structural and dynamic MRI in other anatomies, or other imaging modalities such as CT. Reduced supervision requirements and subject-specific adaptation render SLATER a promising candidate for high-performance accelerated MRI.

REFERENCES

[1] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, no. 5, pp. 952–962, 1999.

[2] M. A. Griswold *et al.*, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Magn. Reson. Med.*, vol. 47, no. 6, pp. 1202–1210, 2002.

[3] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, Oct. 2007.

[4] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, "An efficient algorithm for compressed MR imaging using total variation and wavelets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[5] C. M. Sandino, J. Y. Cheng, F. Chen, M. Mardani, J. M. Pauly, and S. S. Vasanawala, "Compressed sensing: From research to clinical practice with deep neural networks: Shortening scan times for magnetic resonance imaging," *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 117–127, Jan. 2020.

[6] S. Wang *et al.*, "Accelerating magnetic resonance imaging via deep learning," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 514–517.

[7] Y. Yang, J. Sun, H. Li, and Z. Xu, "ADMM-CSNet: A deep learning approach for image compressive sensing," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 42, no. 3, pp. 521–538, Mar. 2020.

[8] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, 2017.

[9] M. Mardani *et al.*, "Deep generative adversarial neural networks for compressive sensing MRI," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 167–179, Jan. 2019.

[10] S. Ramanarayanan, B. Murugesan, K. Ram, and M. Sivaprakasam, "DC-WCNN: A deep cascade of wavelet based convolutional neural networks for MR image reconstruction," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 647–658.

[11] H. K. Aggarwal, M. P. Mani, and M. Jacob, "Modl: Model-based deep learning architecture for inverse problems," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 394–405, Feb. 2019.

[12] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing MRI reconstruction with cyclic loss in generative adversarial Networks," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1488–1497, Aug. 2018.

[13] A. Sriram *et al.*, "End-to-end variational networks for accelerated MRI reconstruction," in *Proc. MICCAI*, 2020, pp. 64–73.

[14] S. Biswas, H. K. Aggarwal, and M. Jacob, "Dynamic MRI using model-based deep learning and SToRM priors: MoDL-SToRM," *Magn. Reson. Med.*, vol. 82, no. 1, pp. 485–494, 2019.

[15] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson, "Assessment of the generalization of learned image reconstruction and the potential for transfer learning," *Magn. Reson. Med.*, vol. 81, no. 1, pp. 116–128, 2019.

[16] S. U. H. Dar, M. Özbey, A. B. Çatlı, and T. Çukur, "A transfer-learning approach for accelerated MRI using deep neural networks," *Magn. Reson. Med.*, vol. 84, no. 2, pp. 663–685, Aug. 2020.

[17] B. Zhu, J. Z. Liu, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[18] G. Oh, B. Sim, H. Chung, L. Sunwoo, and J. C. Ye, "Unpaired deep learning for accelerated MRI using optimal transport driven CycleGAN," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1285–1296, 2020.

[19] D.-I. Eun, R. Jang, W. S. Ha, H. Lee, S. C. Jung, and N. Kim, "Deep-learning-based image quality enhancement of compressed sensing magnetic resonance imaging of vessel wall: Comparison of self-supervised and unsupervised approaches," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 13950.

[20] H. Chung, E. Cha, L. Sunwoo, and J. C. Ye, "Two-stage deep learning for accelerated 3D time-of-flight MRA without matched training data," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102047.

[21] J. I. Tamir, S. X. Yu, and M. Lustig, "Unsupervised deep basis pursuit: Learning reconstruction without ground-truth data," in *Proc. ISMRM*, 2019, p. 0660.

[22] A. Q. Wang, A. V. Dalca, and M. R. Sabuncu, "Neural network-based reconstruction in compressed sensing MRI without fully-sampled training data," in *Proc. MLMIR*, 2020, pp. 27–37.

[23] E. K. Cole, J. M. Pauly, S. S. Vasanawala, and F. Ong, "Unsupervised MRI reconstruction with generative adversarial networks," 2020, *arXiv:2008.13065*.

[24] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magn. Reson. Med.*, vol. 84, no. 6, pp. 3172–3191, Dec. 2020.

[25] P. Huang *et al.*, "Deep MRI reconstruction without ground truth for training," in *Proc. ISMRM*, 2019, p. 4668.

[26] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "RARE: Image reconstruction using deep priors learned without groundtruth," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1088–1099, Oct. 2020.

[27] H. K. Aggarwal, A. Pramanik, and M. Jacob, "ENSURE: A general approach for unsupervised training of deep image reconstruction algorithms," 2020, *arXiv:2010.10631*.

[28] J. Yoo, K. H. Jin, H. Gupta, J. Yerly, M. Stuber, and M. Unser, "Time-dependent deep image prior for dynamic MRI," 2019, *arXiv:1910.01684*.

[29] S. Arora, V. Roeloffs, and M. Lustig, "Untrained modified deep decoder for joint denoising and parallel imaging reconstruction," in *Proc. ISMRM*, 2020, p. 3585.

[30] Q. Zou, A. H. Ahmed, P. Nagpal, S. Kruger, and M. Jacob, "Deep generative SToRM model for dynamic imaging," 2021, *arXiv:2101.12366*.

[31] M. Z. Darestani and R. Heckel, "Accelerated MRI with un-trained neural networks," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 724–733, 2021.

[32] D. Narnhofer, K. Hammernik, F. Knoll, and T. Pock, "Inverse GANs for accelerated MRI reconstruction," *Proc. SPIE*, vol. 11138, Sep. 2019, Art. no. 111381A.

[33] K. C. Tezcan, C. F. Baumgartner, R. Luechinger, K. P. Pruessmann, and E. Konukoglu, "MR image reconstruction using deep density priors," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1633–1642, Jul. 2019.

[34] Q. Liu, Q. Yang, H. Cheng, S. Wang, M. Zhang, and D. Liang, "Highly undersampled magnetic resonance imaging reconstruction using autoencoding priors," *Magn. Reson. Med.*, vol. 83, no. 1, pp. 322–336, Jan. 2020.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. ICML*, 2019, pp. 7354–7363.

[36] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[37] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," 2021, *arXiv:2106.16031*.

[38] Y. Korkmaz, S. U. Dar, M. Yurt, M. Ozbey, and T. Çukur, "Zero-shot learning for unsupervised reconstruction of accelerated MRI acquisitions," in *Proc. ISMRM*, 2021, p. 1965.

[39] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, and J. K. Seo, "Deep learning for undersampled MRI reconstruction," *Phys. Med. Biol.*, vol. 63, no. 13, Jun. 2018, Art. no. 135007.

[40] J. Yoon *et al.*, "Quantitative susceptibility mapping using deep neural network: QSMnet," *NeuroImage*, vol. 179, pp. 199–206, Oct. 2018.

[41] J. C. Ye, Y. Han, and E. Cha, "Deep convolutional framelets: A general deep learning framework for inverse problems," *SIAM J. Imag. Sci.*, vol. 11, no. 2, pp. 991–1048, Jul. 2017.

[42] D. Lee, J. Yoo, S. Tak, and J. Ye, "Deep residual learning for accelerated MRI using magnitude and phase networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1985–1995, Sep. 2018.

[43] A. Hauptmann, S. Arridge, F. Lucka, V. Muthurangu, and J. A. Steeden, "Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease," *Magn. Reson. Med.*, vol. 81, no. 2, pp. 1143–1156, Feb. 2019.

[44] K. Kwon, D. Kim, and H. Park, "A parallel MR imaging method using multilayer perceptron," *Med. Phys.*, vol. 44, no. 12, pp. 6209–6224, Dec. 2017.

[45] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-Net for compressive sensing MRI," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 10–18.

[46] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang, "KIKI-Net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2188–2201, 2018.

[47] J. Cheng, H. Wang, L. Ying, and D. Liang, "Model learning: Primal dual networks for fast MR imaging," in *Proc. MICCAI*, 2019, pp. 21–29.

[48] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, Jun. 2018.

[49] S. Wang *et al.*, "DIMENSION: Dynamic MR imaging with both k-space and spatial prior knowledge obtained via multi-supervised network training," *NMR Biomed.*, Sep. 2019, Art. no. e4131.

[50] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Convolutional recurrent neural networks for dynamic MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 280–290, Jan. 2019.

[51] S. A. H. Hosseini, B. Yaman, S. Moeller, M. Hong, and M. Akcakaya, "Dense recurrent neural networks for accelerated MRI: History-cognizant unrolling of optimization algorithms," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1280–1291, Oct. 2020.

[52] Y. Chen, D. Firmin, and G. Yang, "Wavelet improved GAN for MRI reconstruction," *Proc. SPIE*, vol. 11595, Feb. 2021, Art. no. 1159513.

[53] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[54] M. Mardani, H. Monajemi, V. Papyan, S. Vasanawala, D. Donoho, and J. Pauly, "Recurrent generative adversarial networks for proximal learning and automated compressive image recovery," 2017, *arXiv:1711.10046*.

[55] S. U. H. Dar, M. Yurt, M. Shahdloo, M. E. Ildiz, B. Tinaz, and T. Cukur, "Prior-guided image reconstruction for accelerated multi-contrast MRI via generative adversarial networks," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1072–1087, Oct. 2020.

[56] D. Polak *et al.*, "Joint multi-contrast variational network reconstruction (jVN) with application to rapid 2D and 3D imaging," *Magn. Reson. Med.*, vol. 84, no. 3, pp. 1456–1469, Sep. 2020.

[57] Q. Chang *et al.*, "Synthetic learning: Learn from distributed asynchronized discriminator GAN without sharing medical image data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13853–13863.

[58] Y. Han, J. Yoo, H. H. Kim, H. J. Shin, K. Sung, and J. C. Ye, "Deep learning with domain adaptation for accelerated projection-reconstruction MR," *Magn. Reson. Med.*, vol. 80, no. 3, pp. 1189–1205, Sep. 2018.

[59] S. U. Dar, M. Yurt, and T. Çukur, "A few-shot learning approach for accelerated MRI via fusion of data-driven and subject-driven priors," in *Proc. ISMRM*, 2021, p. 1949.

[60] Y. Arefeen, O. Beker, J. Cho, H. Yu, E. Adalsteinsson, and B. Bilgic, "Scan specific artifact reduction in K-space (SPARK) neural networks synergize with physics-based reconstruction to accelerate MRI," 2021, *arXiv:2104.01188*.

[61] K. Lei, M. Mardani, J. M. Pauly, and S. S. Vasanawala, "Wasserstein GANs for MR imaging: From paired to unpaired training," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 105–115, Jan. 2021.

[62] M. Akçakaya, S. Moeller, S. Weingärtner, and K. Uğurbil, "Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction: Database-free deep learning for fast imaging," *Magn. Reson. Med.*, vol. 81, pp. 439–453, Jan. 2019.

[63] S. A. H. Hosseini *et al.*, "Accelerated coronary MRI with sRAKI: A database-free self-consistent neural network k-space reconstruction for arbitrary undersampling," *PLoS ONE*, vol. 15, no. 2, pp. 1–13, 2020.

[64] T. Hyung Kim, P. Garg, and J. P. Haldar, "LORAKI: Autocalibrated recurrent neural networks for autoregressive MRI reconstruction in k-space," 2019, *arXiv:1904.09390*.

[65] P. Huang *et al.*, "Unsupervised deep learning reconstruction using the MR imaging model," in *Proc. ISMRM*, 2020, p. 3617.

[66] C. Eldeniz, W. Gan, S. Chen, J. Liu, U. S. Kamilov, and H. An, "Phase2Phase: Reconstruction of free-breathing MRI into multiple respiratory phases using deep learning without a ground truth," in *Proc. ISMRM*, 2020, p. 807.

[67] J. Liu *et al.*, "RED-N2N: Image reconstruction for MRI using deep CNN priors trained without ground truth," in *Proc. ISMRM*, 2020, p. 993.

[68] S. A. H. Hosseini, B. Yaman, S. Moeller, and M. Akcakaya, "High-fidelity accelerated MRI reconstruction by scan-specific fine-tuning of physics-based neural networks," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1481–1484.

[69] H. K. Aggarwal and M. Jacob, "Model adaptation for image reconstruction using generalized Stein's unbiased risk estimator," 2021, *arXiv:2102.00047*.

[70] Z. Ke *et al.*, "Assessment of the generalization of learned unsupervised deep learning method," in *Proc. ISMRM*, 2020, p. 3630.

[71] K. H. Jin, D. Lee, and J. C. Ye, "A general framework for compressed sensing and parallel MRI using annihilating filter based low-rank Hankel matrix," *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 480–495, Dec. 2016.

[72] X. Zhang *et al.*, "Image reconstruction with low-rankness and self-consistency of k-space data in parallel MRI," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101687.

[73] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[74] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[75] D. A. Hudson and C. L. Zitnick, "Generative adversarial transformers," 2021, *arXiv:2103.01209*.

[76] J. Yang *et al.*, "Focal self-attention for local-global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[77] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, Oct. 2021, pp. 10012–10022.

[78] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," 2017, *arXiv:1703.06868*.

[79] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[80] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.

[81] C. M. Bishop, *Pattern Recognition and Machine Learning*. Cham, Switzerland: Springer, 2006.

[82] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge," in *Proc. ICML*, vol. 80, 2018, pp. 3481–3490.

[83] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[84] M. Uecker *et al.*, "ESPIRiT—An eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA," *Magn. Reson. Med.*, vol. 71, no. 3, pp. 990–1001, Mar. 2014.

[85] M. Bydder, D. J. Larkman, and J. V. Hajnal, "Combination of signals from array coils using image-based estimation of coil sensitivity profiles," *Magn. Reson. Med.*, vol. 47, no. 3, pp. 539–548, 2002.

[86] J. P. Haldar and J. Zhuo, "P-LORAKS: Low-rank modeling of local k-space neighborhoods with parallel imaging data," *Magn. Reson. Med.*, vol. 75, no. 4, p. 1499, 2016.

[87] T. Kim and J. Haldar, "LORAKS software version 2.0: Faster implementation and enhanced capabilities," Univ. Southern California, Los Angeles, CA, USA, Tech. Rep. USC-SIPI-443, 2018.

[88] F. Knoll *et al.*, "FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190007.

[89] T. Zhang, J. M. Pauly, S. S. Vasanawala, and M. Lustig, "Coil compression for accelerated imaging with Cartesian sampling," *Magn. Reson. Med.*, vol. 69, no. 2, pp. 571–582, 2013.

[90] Y. Wu, Y. Ma, J. Liu, J. Du, and L. Xing, "Self-attention convolutional neural network for improved MR image reconstruction," *Inf. Sci.*, vol. 490, pp. 317–328, Jul. 2019.

[91] H. Lan, A. W. Toga, and F. Sepehrband, "SC-GAN: 3D self-attention conditional GAN with spectral normalization for multi-modal neuroimaging synthesis," *BioRxiv*, pp. 1–35, Art. no. bioRxiv:2020.06.09.143297.

[92] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[93] Z. Yuan *et al.*, "SARA-GAN: Self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction," *Frontiers Neuroinform.*, vol. 14, p. 58, Nov. 2020.

[94] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[95] X. Zhang *et al.*, "A guaranteed convergence analysis for the projected fast iterative soft-thresholding algorithm in parallel MRI," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101987.

[96] W. Huang, P. Hand, R. Heckel, and V. Voroninski, "A provably convergent scheme for compressive sensing under random generative priors," *J. Fourier Anal. Appl.*, vol. 27, no. 2, p. 19, Apr. 2021.

[97] R. Heckel and M. Soltanolkotabi, "Compressive sensing with untrained neural networks: Gradient descent finds the smoothest approximation," 2020, *arXiv:2005.03991*.

[98] M. Asim, M. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: Mitigating representation error and dataset bias," in *Proc. ICML*, vol. 119, 2020, pp. 399–409.

[99] A. Grover, M. Dhar, and S. Ermon, "Flow-GAN: Combining maximum likelihood and adversarial learning in generative models," 2017, *arXiv:1705.08868*.

[100] Z. Wang *et al.*, "One-dimensional deep low-rank and sparse network for accelerated MRI," 2021, *arXiv:2112.04721*.

[101] Z. Wang *et al.*, "A sparse model-inspired deep thresholding network for exponential signal reconstruction—Application in fast biological spectroscopy," 2020, *arXiv:2012.14830*.