Compressible Latent-Space Invertible Networks for Generative Model-Constrained Image Reconstruction

Varun A. Kelkar, Sayantan Bhadra, and Mark A. Anastasio, *Senior Member, IEEE*

***Abstract*—There remains an important need for the development of image reconstruction methods that can produce diagnostically useful images from undersampled measurements. In magnetic resonance imaging (MRI), for example, such methods can facilitate reductions in data-acquisition times. Deep learning-based methods hold potential for learning object priors or constraints that can serve to mitigate the effects of data-incompleteness on image reconstruction. One line of emerging research involves formulating an optimization-based reconstruction method in the latent space of a generative deep neural network. However, when generative adversarial networks (GANs) are employed, such methods can result in image reconstruction errors if the sought-after solution does not reside within the range of the GAN. To circumvent this problem, in this work, a framework for reconstructing images from incomplete measurements is proposed that is formulated in the latent space of invertible neural network-based generative models. A novel regularization strategy is introduced that takes advantage of the multiscale architecture of certain invertible neural networks, which can result in improved reconstruction performance over classical methods in terms of traditional metrics. The proposed method is investigated for reconstructing images from undersampled MRI data. The method is shown to achieve comparable performance to a state-of-the-art generative model-based reconstruction method while benefiting from a deterministic reconstruction procedure and easier control over regularization parameters.**

***Index Terms*—Image reconstruction, compressive sensing, generative neural networks, invertible neural networks**

I. Introduction

Modern imaging systems are typically computed in nature and utilize a reconstruction method to estimate an image from a collection of measurements. In magnetic resonance imaging (MRI) and other medical imaging modalities, there are compelling reasons for reducing data-acquisition times. In certain modalities, one way to achieve this is to simply reduce the number of measurements acquired. This strategy for accelerating data-acquisitions is relevant to MRI, where data-acquisition times are proportional to the number of measured k-space samples. When the acquired measurements are insufficient to uniquely specify the sought-after object, i.e., the measurements are *incomplete*, prior information about the object generally needs to be imposed in the form of regularization in order to recover images that possess potential utility.

The concept of *sparsity* has been widely exploited to develop effective regularization strategies that can mitigate the effects of measurement-incompleteness in inverse problems such as image reconstruction [1]–[4]. Modern sparse image reconstruction methods exploit the fact that many objects of interest can typically be described by use of sparse representations and have proven to be highly effective at estimating images from under-sampled measurement data in MRI and other modalities [5]–[8]. Sparse reconstruction methods are commonly formulated as penalized least squares estimators, where the penalty is specified as an ℓ_1-norm that promotes solutions that are sparse in a specified transform domain. Such reconstruction approaches are prescribed by compressive sensing theories [3], [9], [10] and have enabled design of innovative measurement strategies [7], [8], [11].

Instead of using hand-crafted penalties (i.e., object priors) such as the ℓ_1-norm or total variation semi-norm [12], there has been considerable research aimed at learning object priors from a dataset of representative objects. Some of these techniques such as *dictionary learning* and *transform learning*, involve learning a dictionary that maps the images of interest to sparse vectors [13], [14]. A more detailed review of sparsity and data-driven methods for image reconstruction can be found in reference [15]. More recently, there have emerged numerous deep learning-based approaches for image reconstruction that also seek to capture and exploit information regarding the sought-after object in order to mitigate measurement-incompleteness or noise [16]–[23].

Deep generative models, such as generative adversarial networks (GANs), have shown great promise in learning distributions of objects [24], [25]. An object distribution represented by a generative model can be employed as an object prior in an image reconstruction approach, which can potentially outperform traditional sparsity-based priors. For example, Bora *et al.* proposed an approach in which the solution of a least squares image reconstruction problem is constrained to reside within the range of a generative model [19]. This approach promotes solutions that are consistent with the measured data, with theoretical guarantees on the reconstruction error

This work was supported in part by NIH Awards EB020604, EB023045, NS102213, EB028652, and NSF Award DMS1614305.

Varun A. Kelkar is with the Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: vak2@illinois.edu).

Sayantan Bhadra is with the Department of Computer Science and Engineering, Washington University in Saint Louis, Saint Louis, MO USA (e-mail: sayantanbhadra@wustl.edu).

Mark A. Anastasio is with the Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: maa@illinois.edu).

This paper has supplementary downloadable material available at https://ieeexplore.ieee.org/. The material includes additional results and figures. This material is 4.5 MB in size.

obtained [19]. It was also demonstrated empirically that in the severe undersampling regime, this method could outperform traditional sparsity-based reconstruction methods in terms of mean-squared error. Although this method can perform well when the measured data are produced from an object contained in the range of the generator, in practice this condition can easily be violated. Due to the GAN architecture, dataset size and variability, and other reasons outlined in [26], a high-dimensional object may not exactly reside on the low-dimensional manifold that is the range of the state-of-the-art GANs. This often leads to *representation error* and can result in reconstructed images that look realistic but contain false features. This phenomenon is highly undesirable in medical imaging applications.

Several approaches have been proposed to mitigate representation error in generative model-constrained image reconstruction approaches. One such approach, the *SparseGEN* framework [27], accounts for sparse deviations of the true image from the range of the generative model, and achieves theoretical guarantees for signals that are only sparsely outside the range of a generative model. However, in practice, it might be difficult to find a linear mapping that sparsifies the difference of two realistic images. Another approach, known as the deep image prior (DIP) involves starting out with an untrained neural network and learning the parameters during reconstruction [18], [28]. This method shows impressive performance, due to the fact that the structure of the convolutional neural network layers itself acts as a regularization. However, it has been shown that this approach eventually overfits the measurement noise, and early stopping is needed [28]. A similar approach, known as image-adaptive GAN based reconstruction, starts out with a pretrained network similar to [19], and then adapts the parameters of the GAN along with optimizing over the latent space vector [29]–[31]. These approaches involve optimizing over potentially a large number of parameters, depending upon the complexity of the GAN architecture.

A recent and promising approach to mitigating representation error in generative model-constrained image reconstruction is to employ invertible neural networks (INNs) [26]. In INN-based generative models, referred to here as invertible generative models, the latent space and range have the same dimension and all possible images reside within the range. There also exists a unique latent space representation for every image. Hence, invertible generative models have theoretically zero representation error. While the ability of invertible generative models to eliminate representation error is desirable, an undesirable consequence of this is that they can also describe features that are not contained within the distribution of objects under consideration. In this sense, the flexibility they provide in representing objects comes at the cost of weakening the strength of the prior information employed to constrain the solution of the inverse problem. This can result in reconstruction methods that produce object estimates that are noisy or contain hallucinations [26].

In this work, novel regularization strategies are proposed for reconstruction methods that are constrained by use of invertible generative models. Specifically, to address the limitations described above, the proposed regularization strategies are based upon the multiscale architecture of certain INNs. It is demonstrated that INNs with a multiscale architecture have a compressible latent space that can be exploited to effectively regularize the constrained image reconstruction problem, resulting in images comparable to state-of-the-art image adaptive GAN based approaches while benefiting from a deterministic reconstruction procedure and easier control over regularization parameters. While the proposed method is applicable to a variety of linear inverse problems, in this work, it is systematically investigated by means of stylized undersampled MRI experiments and compared to existing sparsity-based and generative model-based approaches. This includes an investigation of *in-distribution* cases, where a test image belongs to the same probability distribution as the training data, and an *out-of-distribution* case, where the image belongs to a distribution different from the training data. Finally, the proposed method is validated by use of a bias-variance analysis and other standard evaluation metrics such as mean-squared error and structural similarity [32].

The remainder of the article is organized as follows. First, in Section II, the considered problem is formulated and a description of compressed sensing under sparsity priors and using generative models is reviewed. In the same section, a brief introduction to invertible neural network architectures is provided. In Section III, a description of how the latent space of certain INN architectures is compressible is given. A new reconstruction method that exploits the compressibility of the latent space for regularization is described in Section IV. The design of the numerical studies based on stylized MRI experiments is described in Section V, with the results given in Section VI. Finally, a discussion and conclusion is provided in Section VII.

II. Background

Many digital imaging systems, including MRI, are well-approximated by a linear imaging model described as [33]

$$\mathbf{g} = H\mathbf{f} + \mathbf{n}, \tag{1}$$

where $\mathbf{f} \in \mathbb{E}^n$ corresponds to the discretized approximation of the object to-be-imaged, $\mathbf{g} \in \mathbb{E}^m$ corresponds to the measurements taken, $H \in \mathbb{E}^{m\times n}$ corresponds to the linear discrete-to-discrete operator that approximately describes the imaging system, and $\mathbf{n} \in \mathbb{E}^m$ is the measurement noise. Here, the symbol $\mathbb{E}$ is used to denote a Euclidean space, specifically $\mathbb{R}$ or $\mathbb{C}$. In this work, the problem of estimating $\mathbf{f}$ from incomplete measurements $\mathbf{g}$ is considered; namely, H is assumed to be rank deficient.

A. Recovering sparse objects from underdetermined systems

Compressed sensing (CS) has emerged as a popular framework for recovering signals from an underdetermined linear system of equations. In compressed sensing, the prior information about the structure of the object $\mathbf{f}$ is imposed through sparsity in some domain. More specifically, if H satisfies the *restricted isometry property* (RIP) over the set of all $2k$-sparse matrices, then the recovery of any k sparse vectors

can be guaranteed [3], [5]. Several matrices, such as random Gaussian sensing matrices of appropriate column length and independent and identically distributed (i.i.d.) elements, as well as random Fourier sampling matrices relevant for compressive MRI satisfy the RIP [3]. Intuitively, this means that two objects that are sparse in some domain, give rise to measurements that are not close. Hence, if an object is sparse, under certain conditions, it can be recovered uniquely from noiseless measurements [3].

Definition II.1 (Restricted Isometry)**.** Let S_k be the set of all k-sparse vectors in $\mathbb{R}^n$. The *restricted isometry constant* is the smallest constant $\delta_k \in (0,1)$ that satisfies

$$(1-\delta_k)\left\|\mathbf{f}\right\|_2^2 \leq \left\|H\mathbf{f}\right\|_2^2 \leq (1+\delta_k)\left\|\mathbf{f}\right\|_2^2, \tag{2}$$

for all $\mathbf{f} \in S_k$. H is said to satisfy the RIP over S_k if δ_k is not too close to 1 in a prescribed sense [3].

B. Recovering objects using generative priors

A data-driven framework for compressed sensing has been developed [19], where instead of sparsity in some domain, the prior information about the object is expressed in terms of a generative model.

Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a generative model, typically a deep neural network, parametrized by a vector $\boldsymbol{\theta}$. The parameters $\boldsymbol{\theta}$ of the generative model are estimated by training the model on a dataset of images, such that if a $\mathbf{z} \in \mathbb{R}^k$ is sampled from a simple tractable distribution, such as $\mathcal{N}(0, I)$, then $G(\mathbf{z})$ is approximately a sample from the distribution of images that make up the dataset. Here, $\mathbf{z}$ is also called the latent representation of $G(\mathbf{z})$. Since many image data distributions are approximately low dimensional, an architecture with $k \ll n$ can work well for the purpose of image generation. State-of-the-art generative performance is achieved by progressively growing generative adversarial networks (ProGAN) [25] and its variants, such as StyleGAN [34], [35].

Similar to the RIP, Bora *et. al* in [19] introduced the *set-restricted eigenvalue condition* (S-REC) in the context of compressed sensing using generative models (CSGM).

Definition II.2 (Set-restricted eigenvalue condition)**.** Let $S \subseteq \mathbb{R}^n$. For some constants $\gamma > 0$ and $\delta \geq 0$, a matrix $H \in \mathbb{R}^{m \times n}$ satisfies the set-restricted eigenvalue condition S-REC(S, γ, δ) if for any $\mathbf{f}_1, \mathbf{f}_2 \in S$,

$$\left\|H(\mathbf{f}_1 - \mathbf{f}_2)\right\|_2 \geq \gamma \left\|\mathbf{f}_1 - \mathbf{f}_2\right\|_2 - \delta. \tag{3}$$

It has been shown that specific sensing matrices, such as certain i.i.d. random Gaussian matrices, satisfy the S-REC [19]. Note that, similar to the RIP, the interpretation of Definition II.2 is that if S is the range of a generative model, then any two objects $\mathbf{f}_1, \mathbf{f}_2$ in the range of the generative model that are sufficiently far apart in terms of the ℓ_2 distance give rise to imaging measurements that are also far apart (up to an error of δ), if the sensing matrix obeys a suitable S-REC. This is not the case for arbitrary vectors $\mathbf{f}_1, \mathbf{f}_2 \in \mathbb{R}^n$ that may have very different components in the null space of H while giving rise to measurements that are close.

For a generative model G with latent space dimensionality k and Lipschitz constant L, Bora, *et al.* [19] showed that $O(k \log(Lr/\delta))$ measurements suffice to stably recover those signals in $\mathcal{R}(G)$ whose latent representation has an ℓ_2-norm of at most r. This recovery guarantee is applicable to a solution of the following optimization problem:

$$\begin{aligned} \hat{\mathbf{z}} &= \underset{\mathbf{z}, \|\mathbf{z}\| \leq r}{\arg\min} \left\|\mathbf{g} - HG(\mathbf{z}; \boldsymbol{\theta})\right\|_2^2, \\ \hat{\mathbf{f}} &\equiv G(\hat{\mathbf{z}}; \boldsymbol{\theta}), \end{aligned} \tag{4}$$

where $\mathbf{g} = H\tilde{\mathbf{f}} + \mathbf{n}$ is the measurement corresponding to the unknown true object $\tilde{\mathbf{f}}$. Here, $\mathcal{R}(G)$ is the range of G, defined as

$$\mathcal{R}(G) \equiv \{G(\mathbf{z}) \ s.t. \ \mathbf{z} \in \mathbb{R}^k\} \tag{5}$$

.

In [19], this problem is reformulated in the Lagrangian form as:

$$\begin{aligned} \hat{\mathbf{z}} &= \underset{\mathbf{z}}{\arg\min} \left\|\mathbf{g} - HG(\mathbf{z}; \boldsymbol{\theta})\right\|_2^2 + \lambda \left\|\mathbf{z}\right\|_2^2, \\ \hat{\mathbf{f}} &\equiv G(\hat{\mathbf{z}}; \boldsymbol{\theta}), \end{aligned} \tag{6}$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter used to implicitly impose the constraint $\|z\| \leq r$. While this problem is non-convex, it has been empirically observed that gradient descent-based algorithms can find critical points that have a sufficiently low value of the objective to yield a reconstruction with low error [19], [26], [29]. For images that lie in the range of G, this gives a reconstruction for which the ℓ_2 error is only limited by the magnitude of measurement noise and the error due to non-convergence of the gradient-descent type algorithm used to approximately solve Eq. (6).

However, when $\tilde{\mathbf{f}} \notin \mathcal{R}(G)$, the reconstructed estimate of $\tilde{\mathbf{f}}$ contains an additional error, known as the *representation error* [19]. Here, the representation error is defined as

$$\rho_G(\tilde{\mathbf{f}}) \equiv \min_{\mathbf{z}} \left\|G(\mathbf{z}) - \tilde{\mathbf{f}}\right\|_2.$$

In practice, G has limited representational capacity and is trained on limited training data by optimizing a non-convex objective with gradient-based methods. Also, $\mathcal{R}(G)$ is only a k-dimensional manifold in $\mathbb{R}^n$. Hence, there is a significant representation error even for in-distribution images. This, coupled with the fact that the generative models such as the ProGAN produce highly realistic images, can result in plausible but wrong solutions to Eq. (6).

One natural extension to the optimization problem formulated in Eq. (6) is to adapt $\mathcal{R}(G)$ based on the measured data. This can be achieved by jointly optimizing over the parameters $\boldsymbol{\theta}$ of the generative model and the latent space vector $\mathbf{z}$:

$$\begin{aligned} \hat{\mathbf{z}}, \hat{\boldsymbol{\theta}} &= \underset{\mathbf{z}, \boldsymbol{\theta}}{\arg\min} \left\|\mathbf{g} - HG(\mathbf{z}; \boldsymbol{\theta})\right\|_2^2, \\ \hat{\mathbf{f}} &\equiv G(\hat{\mathbf{z}}, \hat{\boldsymbol{\theta}}). \end{aligned} \tag{7}$$

This technique is known as image-adaptive GAN-based reconstruction (IAGAN) [29]. Approximate solutions to Eq. (7) can be obtained using a standard gradient-descent based algorithm. However, as shown in [28], the success of such an algorithm

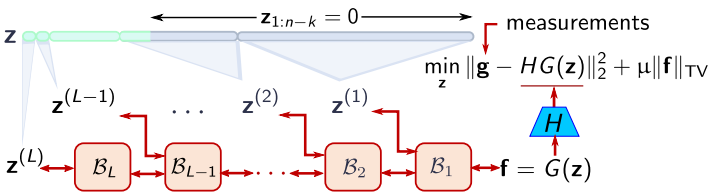


Fig. 1: A schematic of our approach that exploits the multiscale structure of the INN. $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(L)}$, are sections of the latent space that are introduced at different levels in the INN architecture, with L being the number of levels. $\mathcal{B}_i$'s represent the invertible blocks that constitute the INN and H is the forward model.

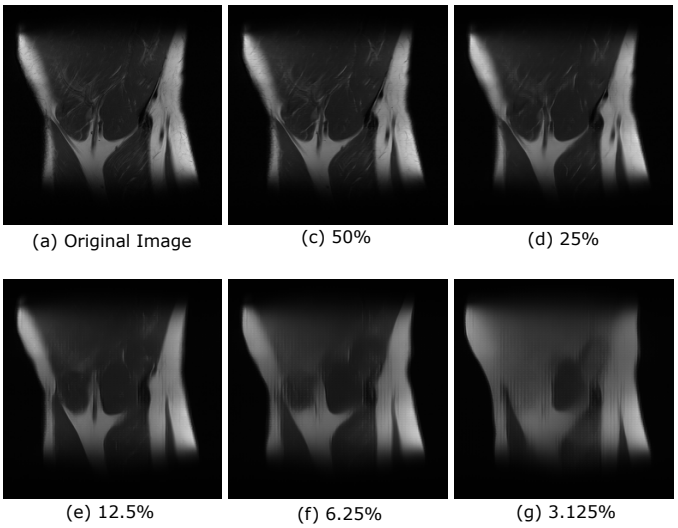


Fig. 2: Low errors when $\mathbf{z}$ coefficients are truncated: a consequence of the compressible latent space. (a) A ground truth image, (b-f) images obtained by keeping a fraction of the $\mathbf{z}$ coefficients.

depends upon early stopping, and convergence results in overfitting the noisy measurements. Moreover, the method is typically initialized with a solution of Eq. (6) [29], [30]. Equation (6) may need to be solved several times with different random initializations to yield an initial estimate that gives competitive performance for Eq. (7), in terms of mean squared error.

C. Invertible generative models

Invertible neural networks (INN) are bijective mappings

$$G_{\text{inn}} : \mathbb{R}^n \to \mathbb{R}^n, \tag{8}$$

constructed via neural networks, with the vector θ again denoting the parameters of the network [36]–[41]. They can be trained as generative models on a dataset of images independently sampled from an image distribution $p_{\mathbf{f}}$, such that a sample $\mathbf{z} \in \mathbb{R}^n$ from a simple tractable distribution $p_{\mathbf{z}}$ produces a sample $\mathbf{f} = G_{\text{inn}}(\mathbf{z}) \in \mathbb{R}^n$ from a distribution approximating $p_{\mathbf{f}}$. Since the mapping is a bijection, every $\mathbf{f}$ has a unique latent-space representation $\mathbf{z}$. Moreover, the probability distributions of the input and the output of G_{inn} are related by [36], [37]

$$p_{\mathbf{f}}(\mathbf{f}) |\det(\nabla_{\mathbf{z}} G_{\text{inn}}(\mathbf{z}))| = p_{\mathbf{z}}(\mathbf{z}), \tag{9}$$

TABLE I: Ensemble RMSE between the 500 test dataset images and images obtained by keeping a fraction of the $\mathbf{z}$ coefficients

% z_i's kept	50	25	12.5	6.25	3.125
Mean RMSE	0.0090	0.0148	0.0244	0.0367	0.0518
Std. dev. of RMSE	0.0024	0.0035	0.0057	0.0086	0.0135

or equivalently,

$$-\log p_{\mathbf{f}}(\mathbf{f}) = -\log p_{\mathbf{z}}(\mathbf{z}) + \log|\det(\nabla_{\mathbf{z}} G_{\text{inn}}(\mathbf{z}))|, \tag{10}$$

where $\mathbf{f} = G_{\text{inn}}(\mathbf{z})$ or, equivalently, $\mathbf{z} = G_{\text{inn}}^{-1}(\mathbf{f})$.

Accordingly, an INN-based generative model can be trained by use of a log-likelihood based objective function:

$$\begin{aligned}\mathcal{L}(\mathcal{D}) &= -\frac{1}{D}\sum_{i=1}^{D} \log p_{\mathbf{f}}(\mathbf{f}^{\{i\}}) \\ &= \frac{1}{D}\sum_{i=1}^{D} \log|\det(\nabla_{\mathbf{z}} G_{\text{inn}}(\mathbf{z}^{\{i\}}))| - \frac{1}{D}\sum_{i=1}^{D} \log p_Z(\mathbf{z}^{\{i\}}),\end{aligned} \tag{11}$$

where $\mathcal{D} = \{\mathbf{f}^{\{i\}}\}_{i=1}^{D}$ is the training dataset of size D.

For training scalable invertible networks via Eq. (11), the following conditions need to be satisfied: (1) for an invertible layer that maps a vector $\mathbf{x} \in \mathbb{R}^n$ to a vector $\mathbf{y} \in \mathbb{R}^n$, computing $\mathbf{x}$ from $\mathbf{y}$ and computing $\mathbf{y}$ from $\mathbf{x}$ must have similar computational costs, and (2) the determinant of the Jacobian of the network is computationally tractable. Several architectures satisfying the above constraints have been proposed [38]–[40], [42]. In many of these architectures, a key enabling factor in satisfying the above constraints is the affine coupling layer. If $\mathbf{x}$ and $\mathbf{y}$ are the input and output of a certain affine coupling layer, the transformation relating $\mathbf{y}$ to $\mathbf{x}$ is given by

$$\begin{aligned}\mathbf{y}_{1:p} &= \mathbf{x}_{1:p}, \\ \mathbf{y}_{p+1:n} &= \mathbf{x}_{p+1:n} \odot \exp(s(\mathbf{x}_{1:p})) + t(\mathbf{x}_{1:p}),\end{aligned} \tag{12}$$

where the notation $\mathbf{x}_{u:v}$ is used to denote the vector formed from components of $\mathbf{x}$ from the u-th index to the v-th index, and $\odot$ denotes the Hadamard product or the element-wise product of two vectors. Here, s and t are functions parametrized by neural networks, and need not be invertible. It can be verified that the above transformation is invertible [39]. Moreover, the determinant of the Jacobian of the above transformation is given by

$$\left|\det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)\right| = \exp\left(\sum_{i=1}^{n-p} s(\mathbf{x}_{1:p})_i\right), \tag{13}$$

which is computationally inexpensive as compared to the usual $O(n^3)$ complexity needed to evaluate a general $n \times n$ Jacobian.

The use of INNs to impose a reconstruction constraint can be achieved by replacing the GAN-based generative model in Eq. (6) with an invertible network [26]. However, in current practice, the state-of-the-art GANs generally model the probability distribution of the data better than the state-of-the-art invertible generative models, both perceptually, as well as in terms of the Fréchet Inception Distance (FID) scores between real and generated datasets [43]. Hence, a naive application of INNs in Eq. (6) can result in a noisy or distorted

image estimate when dealing with high resolution images. Intuitively, Eq. (6) can be thought of as optimizing over the norm ball $\|\mathbf{z}\| \leq r \in \mathbb{R}^+$ which, when high dimensional, contains undesirable solutions to Eq. (6). Hence, a new way of regularizing the problem is needed in order to constrain the solution space while minimizing the representation error.

III. Compressibility of INN Latent Space

In order to reduce the computation and memory during training, Dinh, *et al.* proposed a multiscale invertible architecture [39]. As shown in figure Fig. 1, this results in sections of the latent space vector being introduced into the network at different points in the network, leading to a compressible latent space [39]. This operation can be recursively described as [39]

$$\mathbf{z} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(L)}), \tag{14}$$

$$\mathbf{h}^{(L-1)} = \mathcal{B}_L(\mathbf{z}_L), \tag{15}$$

$$\mathbf{h}^{(l)} = \mathcal{B}_{l+1}(\mathbf{h}^{(l+1)}, \mathbf{z}^{(l+1)}),\ l = 0, \ldots, L-2, \tag{16}$$

$$\mathbf{f} = \mathbf{h}^{(0)}, \tag{17}$$

where $\mathcal{B}_l$ represents the l-th invertible block, constructed out of affine coupling layers, and L is the number of levels in the INN. The schematic in Fig. 1 shows the compressible structure of the latent space. The effect of this compressibility was examined on an ensemble of 500 images from a test dataset of coronal knee images that was kept out of the INN training dataset. This was done as follows. First, for an image $\mathbf{f}_{\text{orig}}$ in the ensemble, the exact latent representation $\mathbf{z}_{\text{orig}}$ was computed. This can be divided into multiple sections $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(L)}$ based on the multilevel architecture of the INN. Next, all sections from $\mathbf{z}^{(1)} \ldots \mathbf{z}^{(i)}$ were progressively set to zero such that only 50%, 25%, 12.5%, 6.25% and 3.125% of the components of $\mathbf{z}$ remain non-zero. For these modified latent space vectors $\mathbf{z}_{50\%}$, $\mathbf{z}_{25\%}$, $\mathbf{z}_{12.5\%}$, $\mathbf{z}_{6.25\%}$ and $\mathbf{z}_{3.125\%}$, the corresponding images $\mathbf{f}_P = G_{\text{inn}}(\mathbf{z}_P)$, $P = 50\%,\ 25\%,\ 12.5\%,\ 6.25\%$ and 3.125% were computed, and the root mean square errors (RMSEs) $\|\mathbf{f}_P - \mathbf{f}_{\text{orig}}\|/\sqrt{n}$ with respect to the original image $\mathbf{f}_{\text{orig}}$ were calculated. The error versus the percentage of $\mathbf{z}$ coefficients kept, averaged over the entire ensemble, was computed and is reported in Table I. Figures 2b-2f display the effect of the truncation on a single ground-truth image Fig. 2a. Thus, in addition to being a generative prior, these results suggest that a multilevel INN architecture also serves to establish a compressible representation for in-distribution images, which is consistent with the findings in [39]. The performance of the INN in terms of compressibility was compared to that of the Haar wavelet transform via RMSE and SSIM, and is included in the supplementary section.

IV. Regularization through compressibility

The fact that the latent space of the INN is compressible can be exploited to design a regularization strategy for inverse problems. Motivated by the compressible structure of the latent space, the recovery of objects $\mathbf{f}$ such that $\left\|G^{-1}(\mathbf{f})\right\|_1/n \leq r/n = 1 + o(1)$ will be examined. Let $S_k = \{\mathbf{f}\ s.t.\ G^{-1}(\mathbf{f})_{1:n-k} = 0\}$, with $k \in \mathbb{N},\ k \leq n$, and let $T_\nu = \{\mathbf{f}\ s.t.\ \|\mathbf{f}\|_{\text{TV}} \leq \nu\}$, $\nu \in \mathbb{R}^+$. The relevant measurement model can be described as

$$\mathbf{g} = H\mathbf{f}, \quad \text{where } \mathbf{f} \in S_k \cap T_\nu. \tag{18}$$

Note that the mapping in Eq. (18) may not be injective. An approximate inverse of the above measurement model can be implicitly defined via a problem similar to the one in Eq. (6), with an added constraint of restricting $\mathbf{z}$ to a k-dimensional subspace corresponding to the most important coefficients. A TV penalty on the image is also included. The considered optimization problem is stated as follows:

$$\begin{aligned}\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} &\|\mathbf{g} - HG_{\text{inn}}(\mathbf{z})\|_2^2 - \lambda \log p_{\mathbf{z}}(\mathbf{z}) \\ &+ \mu \|G_{\text{inn}}(\mathbf{z})\|_{\text{TV}}, \\ \text{subject to} \quad &\mathbf{z}_{1:n-k} = 0, \\ \hat{\mathbf{f}} \equiv\ &G_{\text{inn}}(\hat{\mathbf{z}}),\end{aligned} \tag{19}$$

where k is used to restrict $\mathbf{f}$ to S_k, and λ and μ are used to implicitly impose the constraints $\|\mathbf{z}\|_1 \leq r$, and $\mathbf{f} \in T_\nu$ respectively. All of these are treated as regularization parameters which need to be tuned in order to achieve a suitable restriction. However, it was observed in preliminary studies that λ is not critical to the reconstruction performance. In fact, the best results were achieved if λ is set to 0, as shown in Table II. This reduces the number of explicit regularization parameters to two - the dimensionality k of the latent subspace and the TV penalty μ. Note that similar to Eq. (6), the objective function in Eq. (19) is non-convex, and only approximate solutions are typically obtained when gradient-based methods are employed [44]. Moreover, a medical image may not lie in $S_k \cap T_\nu$; however, Eq. (19) can potentially yield image estimates in $S_k \cap T_\nu$ that are close to the original object.

TABLE II: RMSE values for reconstructions with varying λ

λ	0	10^{-4}	3×10^{-4}	10^{-3}	3×10^{-3}
RMSE (%)	1.342	1.349	1.353	1.390	1.535

Note that, due to the latent space projection, the range of the INN is restricted to S_k. Due to this, overfitting to noise can be avoided. However, this results in representation error, now defined as

$$\rho_k(\tilde{\mathbf{f}}) = \min_{\mathbf{f} \in S_k} \left\|\mathbf{f} - \tilde{\mathbf{f}}\right\|_2. \tag{20}$$

Now, $\rho_k(\tilde{\mathbf{f}})$ is upper bounded by the truncation error,

$$\tau_k(\tilde{\mathbf{f}}) = \left\|\mathbf{f}_t - \tilde{\mathbf{f}}\right\|_2, \tag{21}$$

where

$$\mathbf{f}_t = \arg\min_{\mathbf{f} \in S_k} \left\|G_{\text{inn}}^{-1}(\mathbf{f}) - G_{\text{inn}}^{-1}(\tilde{\mathbf{f}})\right\|. \tag{22}$$

According to the previously shown compressibility results, this error is expected to be minimal for in-distribution images, as compared to the representation error incurred in the approach described in [19].

Due to the restriction of the measurement operator on S_k, Lemma 4.1 in [19] implies that a random measurement matrix with $m = O(k/\alpha^2 \log(Lr/\delta))$ rows and i.i.d. elements drawn

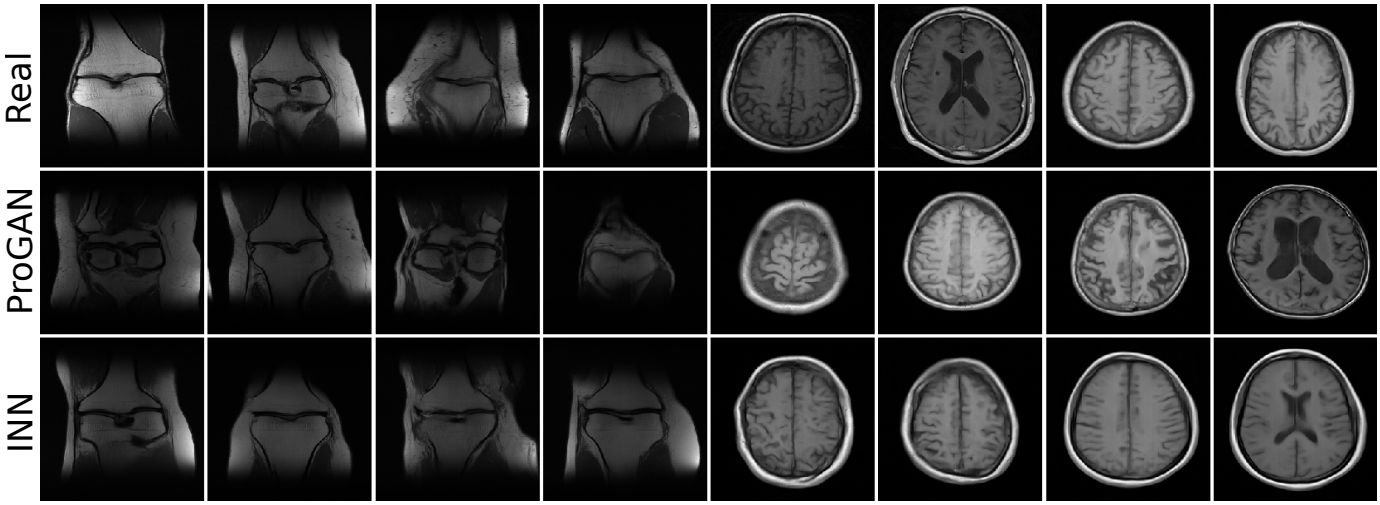


Fig. 3: Generated samples from the progressive GAN and the INN, alongside the experimentally acquired "real" images.

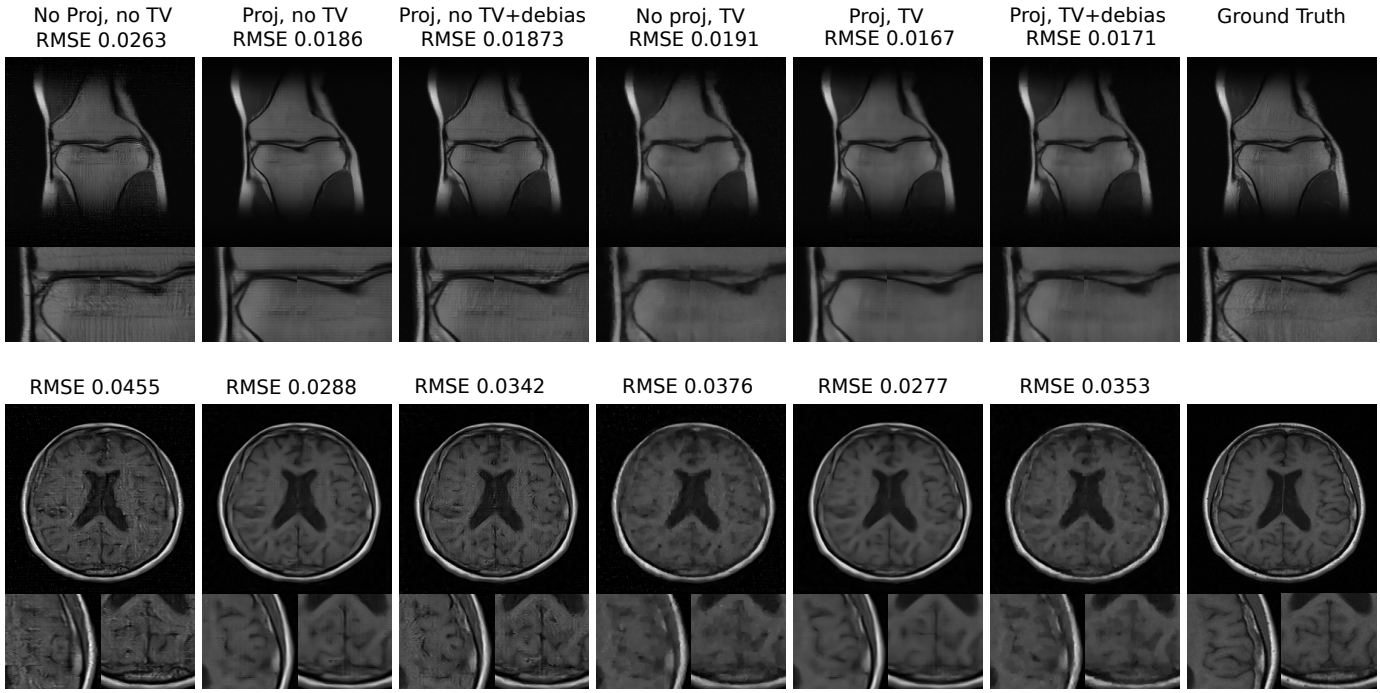


Fig. 4: Comparison of images reconstructed by use of various regularization combinations in the INN reconstruction framework.

from $\mathcal{N}(0, 1/m)$ satisfies the S-REC$(S_k, 1-\alpha, \delta)$ with high probability, where L is now the Lipschitz constant of G_{inn}, with $0 < \alpha < 1$ and $\delta > 0$. A large L and the requirement of uniform recovery implies that this bound on the number of measurements needed is pessimistic for the type of images and measurement operators examined in this manuscript for both the invertible generative model, as well as the GAN. This is analogous to similar observations about RIP-based guarantees in general [45], [46]. However, in this work, this theoretical result is relevant because it provides intuitive understanding on how the number of measurements scale with respect to the dimensionality k of the latent subspace that is used to restrict the domain of H.

A approximate solution of Eq. (19) can be found by a regularized projected Adam technique, which includes iterative steps of the Adam algorithm [47] with a proximal step on $\mathbf{z}$ followed by a projection onto the convex subspace $\{\mathbf{z} \mid \mathbf{z}_{1:n-k} = 0\}$ after each gradient step. The procedure for finding an approximate solution of Eq. (19) is shown in Algorithm 1.

Debiasing. It was observed that the reconstruction accuracy of the proposed approach can be further improved by debiasing an approximate solution of Eq. (19). This can be achieved by removing the constraint of $\mathbf{z}$ lying in a k-dimensional subspace. Debiasing can be performed by finding an approximate

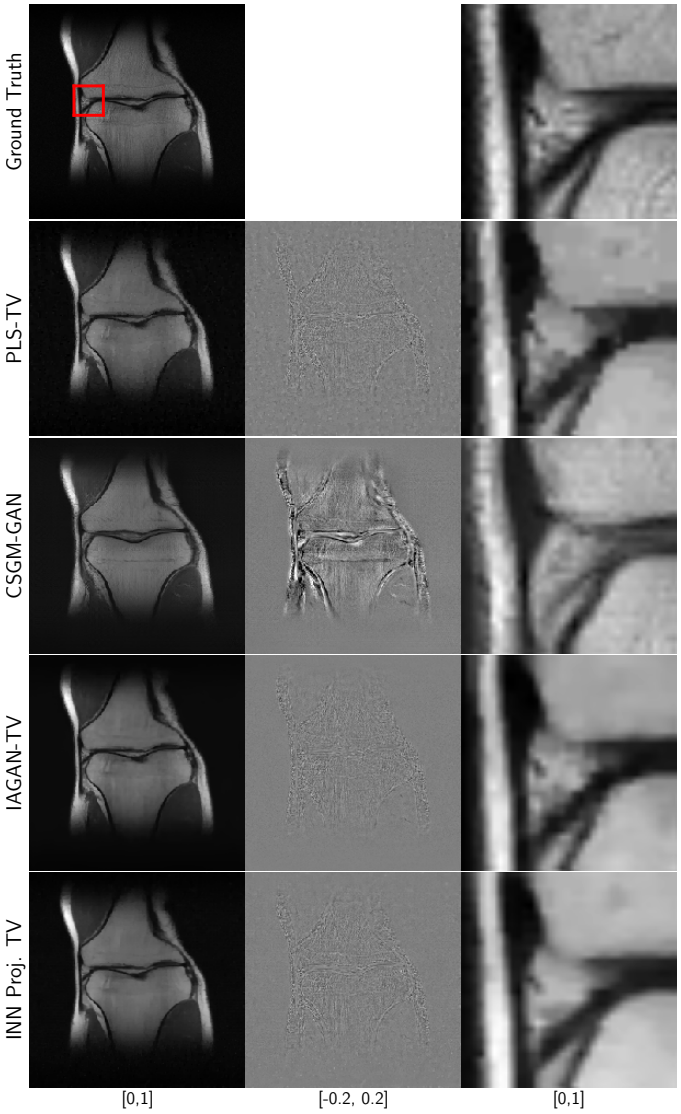


Fig. 5: Ground truth, difference plots and reconstruction results for a coronal PD weighted knee image without fat suppression, with 8-fold undersampling and 20 dB measurement SNR. The RMSE and SSIM values are displayed in Table IV.

solution to the following:

$$\hat{\mathbf{z}}_{\text{deb}} = \arg\min_{\mathbf{z}} \|\mathbf{g} - HG_{\text{inn}}(\mathbf{z})\|_2^2 + \mu \|G_{\text{inn}}(\mathbf{z})\|_{\text{TV}}$$
$$\hat{\mathbf{f}}_{\text{deb}} = G_{\text{inn}}(\hat{\mathbf{z}}_{\text{deb}}), \tag{23}$$

where the above problem is initialized with an approximate solution of Eq. (19). Improvement via debiasing can be obtained through iterative minimization of Eq. (23) by use of early stopping. However, similar to other techniques that require early stopping [18], [29], use of an appropriate stopping criterion requires additional tuning parameters [48]. Hence, debiasing with early stopping was not employed in the studies described below.

The proposed reconstruction method was evaluated as described below.

Algorithm 1 The proposed Projected Adam algorithm for finding an approximate solution to Eq. (19).

Given: Measurements $\mathbf{g}$.
Pick the regularization parameters k and μ. Fix $\lambda = 0$.
Set $\mathbf{p}^{(k)} \in \mathbb{R}^n$ such that $\mathbf{p}^{(k)}_{1:n-k} = 0$ and $\mathbf{p}^{(k)}_{n-k+1:n} = 1$.
Set the Adam optimizer parameters $(\alpha, \beta_1, \beta_2)$ from [47]. Default parameters recommended in Algorithm 1 of [47] are used.
$\mathbf{z}_{\text{init}} \leftarrow \mathbf{0}$. (Initialize the latent space vector)
$t \leftarrow 0$. (Initialize the iteration number)
while $\mathbf{z}_t$ not converged **do**
    Perform an Adam iteration from Algorithm 1 in [47],

$$\mathbf{z}_t \leftarrow \texttt{ADAM}_{\alpha,\beta_1,\beta_2}(\mathbf{z}_{t-1}; \mu),$$

    Perform projection onto the latent subspace:

$$\mathbf{z}_t \leftarrow \mathbf{z}_t \odot \mathbf{p}^{(k)}$$
$$t \leftarrow t + 1$$

end while

V. Numerical Studies

Numerical studies were conducted to assess the effectiveness of the proposed method, especially in terms of recovering fine object features. The reduction in the appearance of realistic but false features and oversmoothing artifacts was studied. Our studies were divided into two parts - (1) reconstruction from stylized, simulated undersampled single-coil MRI measurements (henceforth referred to as the simulation study), and (2) reconstruction from emulated experimental undersampled single-coil MRI measurements (henceforth referred to as the emulated experimental study). For the simulation study, in-distribution images, i.e. the images that come from the same distribution as the training dataset, as well as out-of-distribution images were considered. The proposed method was compared to traditional sparsity-based, as well as recent GAN-based reconstruction methods. For the comparisons, traditional image quality metrics such as the root mean squared error defined as the discrete error norm $\|\hat{\mathbf{f}} - \tilde{\mathbf{f}}\|_2$, as well as structural similarity (SSIM) index [32] were utilized. Where applicable, bias-variance tradeoff calculations were carried out to assess the robustness of the algorithms.

A. Datasets and sensing system

1) Simulation study: The generative models were trained on single channel 2D MRI images of size 256 $\times$ 256. The following two datasets were employed for training, as well as evaluation in the case of in-distribution images:

- *FastMRI knee dataset*: 15000 non-fat suppressed, proton density (PD) weighted coronal knee images from the NYU fastMRI Initiative database [49].
- *FastMRI brain dataset*: 12000 T1-weighted axial adult brain images from the NYU fastMRI Initiative database.

For evaluation of reconstruction performance on out-of-distribution images, images from a pediatric epilepsy resection MRI dataset containing anomalies [50], [51] were used, along

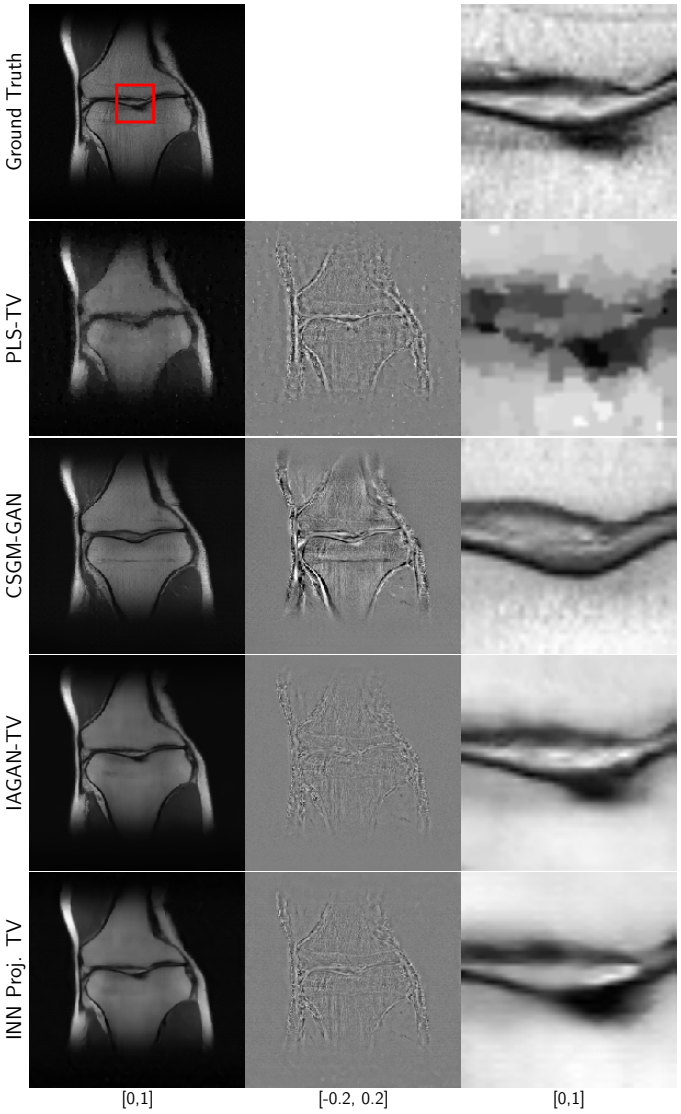


Fig. 6: Ground truth, difference plots and reconstruction results for a coronal PD weighted knee image without fat suppression, with 20-fold undersampling and 20 dB measurement SNR. The RMSE and SSIM values are displayed in Table IV.

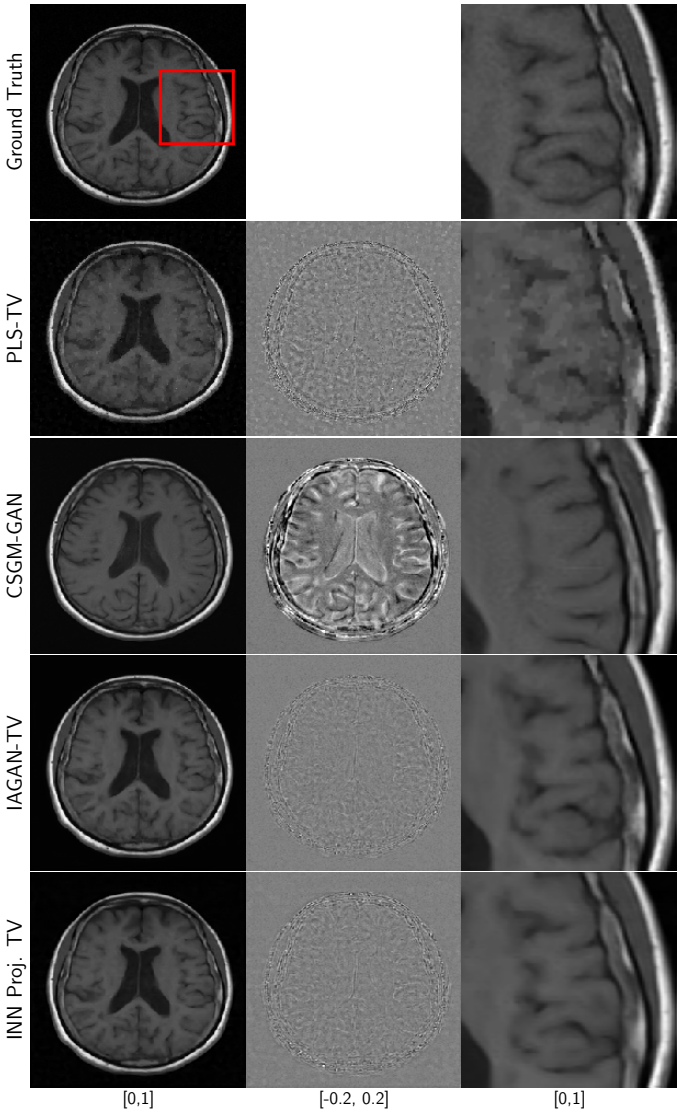


Fig. 7: Ground truth, difference plots and reconstruction results for an axial T1 weighted brain image, with 8-fold undersampling and 20 dB measurement SNR. The RMSE and SSIM values are displayed in Table IV.

with generative models trained on the FastMRI brain dataset. Evaluating the robustness of a reconstruction method on out-of-distribution images is relevant because (i) in practice, test images may not exactly correspond to the training data distribution and a practitioner might be oblivious to these small differences, and (ii) it is of interest to examine the scenario of *transfer compressed sensing*, where learned priors from one dataset are employed to recover images from a closely related but different test distribution, due to the unavailability of sufficient data to learn the priors (for example, data including rare anomalies.)

Simulated undersampled single-coil MR measurements were employed as a proxy for experimental MRI k-space measurements. Variable density Poisson disc sampling patterns shown in Fig. 8 corresponding to R = 8 and R = 20 undersampling ratios were utilized, which retain low frequencies and randomly sample higher frequencies with a variable density [7], [52].

2) Emulated experimental study: Data for training the generative models were prepared in the following way. The fastMRI initiative database provides *emulated single-coil k-space measurements*, each of which is a complex-valued linear combination of responses from multiple coils of raw multi-coil k-space data [53]. These fully sampled k-space measurements were used to generate complex-valued images via the inverse fast Fourier transform (IFFT). They were divided into a training dataset for training the generative models, and a test dataset. The complex-valued images were converted to two-channel real images for training, and generative models with two-channel output were trained. Image reconstruction was performed directly from retrospectively undersampled emulated single-coil measurements, and the image estimates were compared with the reconstructions from the corresponding fully sampled k-space measurements in the test dataset. A

Cartesian random undersampling mask with R = 4 was used for the retrospective undersampling.

For evaluating the reconstruction performance on the above-described image types, a validation image and a test dataset was used for each of the image types. These images were kept unseen during training. The regularization parameters for all the reconstruction methods were tuned on the respective validation image for each image type, and the parameter setting showing the best RMSE performance was chosen. Images corresponding to the regularization parameter sweeps are shown in the supplementary section. The tuned parameters were used in the reconstruction of images from the unseen test datasets. Performance metrics and their statistical significance were reported on these test datasets for all the image types.

B. Network architecture and training

The employed INN architecture was adapted from Kingma and Dhariwal [40]. It consisted of 6 levels in the multilevel architecture, with this choice being empirically determined. The primary invertible layers used were affine coupling layers [39] with invertible 1x1 convolutions [40]. The functions s and t described in Eq. (12) were parametrized by 3 layer convolutional neural networks, with SoftPlus activation functions as the nonlinearity between the convolutional layers [54]. Also, similar to the official implementation by Kingma and Dhariwal [55], the exponential function in the affine coupling layer was replaced with a sigmoid function, which stabilizes the training and makes the invertible transformation Lipschitz stable. In order to gain Lipschitz stability in the reverse direction, the output of the sigmoid function was rescaled to lie in the range $(c, 1]$ with c being a positive, tunable parameter less than 1. Lastly, it was determined that using a standard i.i.d. Laplacian distribution as the latent space prior $p_{\mathbf{z}}(\mathbf{z})$ improves the performance during image generation and reconstruction. Loosely speaking, the Laplacian prior, being more "compressible" than the Gaussian prior, seems to help in learning a closer approximation to the near-low dimensional real data-distribution. The INN was trained on a system with a 2x 20-core IBM POWER9 Central Processing Unit (CPU) @ 2.4GHz, and 4 16 GB NVIDIA V100 Graphical Processing units (GPUs) for a period of about 2.5 days [56].

The progressive GANs (ProGANs) were trained using the original implementation provided by Karras *et al.* [57]. The default settings for the training parameters were employed in this study. As implemented elsewhere [29], [31], the default latent space dimensionality of 512 was maintained. The training was performed on a system with an Intel Xeon E5-2620v4 CPU @ 2.1 GHz and 4 NVIDIA TITAN X GPUs. The algorithms are implemented in Python 3.6/Tensorflow 1.14. Random i.i.d. draws from the generative models are shown in Fig. 3, along with the real samples from the training dataset. The Fréchet Inception Distance (FID) scores for the invertible generative model and the state-of-the-art progressive GAN were calculated by use of the official Python implementation [58] and are shown in Table III. Further details regarding FID scores can be found in the literature [43], [59].

Dataset	ProGAN	INN
Knee	22.72	75.06
Brain	10.67	82.41

TABLE III: FID scores of the generative models. A lower FID is correlated with improved visual quality of generated images [43].

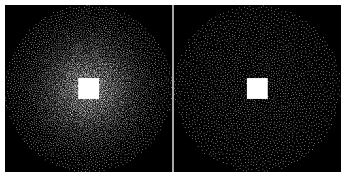


Fig. 8: 8-fold (left) and 20-fold (right) undersampling masks

C. Image reconstruction

In order to assess the effect of each of the two regularization parameters associated with the INN-based reconstruction method as well as debiasing, an ablation study was performed. Simulated measurements corresponding to the R = 20 undersampling ratio were used. Also, complex i.i.d Gaussian noise was added to the measurements, such that the per-pixel SNR of the measurements (with respect to signal power) was 20 dB. The images were reconstructed with (i) no regularization, (ii) only the latent subspace projection, (iii) only the latent subspace projection followed by debiasing, (iv) only the TV penalty, (v) the latent subspace projection and the TV penalty, and (vi) the latent subspace projection with the TV penalty followed by debiasing.

Next, a coronal knee image was reconstructed from simulated fully sampled, noiseless k-space data so that the forward operator is bijective, and the loss decay was analyzed as the iterative optimization progresses. For analyzing how accurate our approach gets to actually solving the inverse problem corresponding to the measurement model in Eq. (18), a knee image $\tilde{\mathbf{f}} \in S_k \cap T_\nu$ for k = 16384 and $\left\|\tilde{\mathbf{f}}\right\|_{\mathrm{TV}} = 922.8$, referred to as the *latent-projected* image, was considered. Measurements corresponding to the $R = 8$ undersampling ratio were simulated and images were reconstructed from these noiseless measurements.

Next, the performances of the following reconstruction methods were qualitatively and quantitatively compared - (i) penalized least squares with TV regularization (PLS-TV) solved with the fast iterative shrinkage and thresholding algorithm (FISTA) [60], (ii) the method proposed by Bora, *et. al* [19], i.e. the problem stated in Eq. (6), with a ProGAN [25] trained as described in Section VB as the generative model (henceforth referred to as CSGM-GAN), (iii) Image-adaptive GAN-based reconstruction with TV regularization described in equation Eq. (7) (IAGAN-TV) [29], [30], and (iv) INN-based reconstruction using latent space projection and TV regularization, described in equation Eq. (19) (INN Proj. TV). For the simulation study, coronal knee and axial brain images were reconstructed from simulated measurements corresponding to the R = 8 and R = 20 undersampling ratios for this comparison. This was done with noiseless measurements, as well as measurements with i.i.d Gaussian noise with 20 dB per-pixel SNR.

Next, the approaches described above were employed to reconstruct anomalous pediatric brain images from 8-fold simulated undersampled measurements with 20 dB SNR [51]. The generative models used to reconstruct the pediatric brain image were trained on axial adult brain images from the previously described NYU fastMRI initiative database.

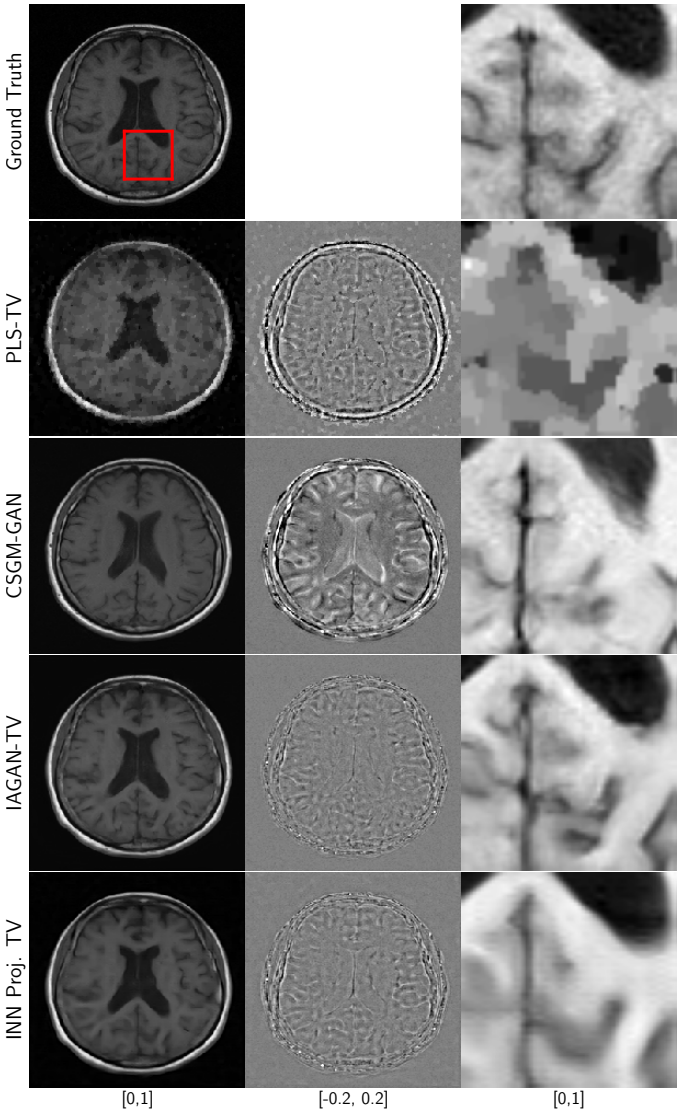


Fig. 9: Ground truth, difference plots and reconstruction results for an axial T1 weighted brain image, with 20-fold undersampling and 20 dB measurement SNR. The RMSE and SSIM values are displayed in Table IV.

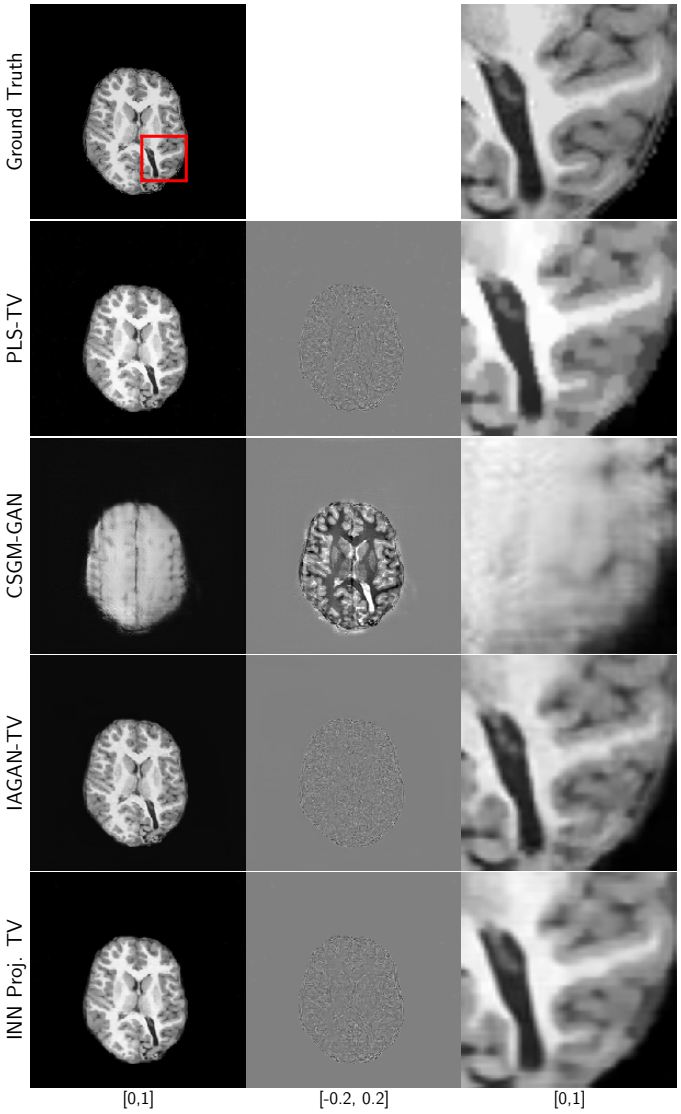


Fig. 10: Ground truth, difference plots and reconstruction results for an axial T1 weighted pediatric brain image with anomaly, with 8-fold undersampling and 20 dB measurement SNR. The RMSE and SSIM values are displayed in Table V.

Finally, for the emulated experimental study, image reconstruction was performed from four-fold retrospectively undersampled emulated single-coil measurements. The image estimates were compared to IFFT-based reconstructions from fully sampled measurements.

VI. Results and Evaluation

This section is organized as follows. First, the results of the ablation study are described. The results for the fully sampled, noiseless reconstruction and the associated RMSE and SSIM values and convergence analysis is deferred to the supplementary section. This is followed by the results of the stylized study where the object lies in $S_k \cap T_\nu$, and thus perfectly satisfies the measurement model. Next, the results for the test images that do not lie in $S_k \cap T_\nu$ are shown, after which the RMSE and SSIM comparisons, statistical significance tests, and the bias-variance trade-off calculations are described.

A. The ablation study

The results of the ablation study are shown in Fig. 4. It was observed that the best RMSE performance was achieved by a combination of latent subspace projection and TV penalty, without debiasing. Hence, in the rest of the manuscript, results using this particular combination of regularization parameters will be described. It should be noted, however, that a combination of latent subspace projection and TV penalty followed by debiasing was able to improve upon the performance of the chosen method if early stopping was performed during debiasing. However, this adds an additional tunable parameter. In the interest of simplicity, the use of debiasing was avoided.

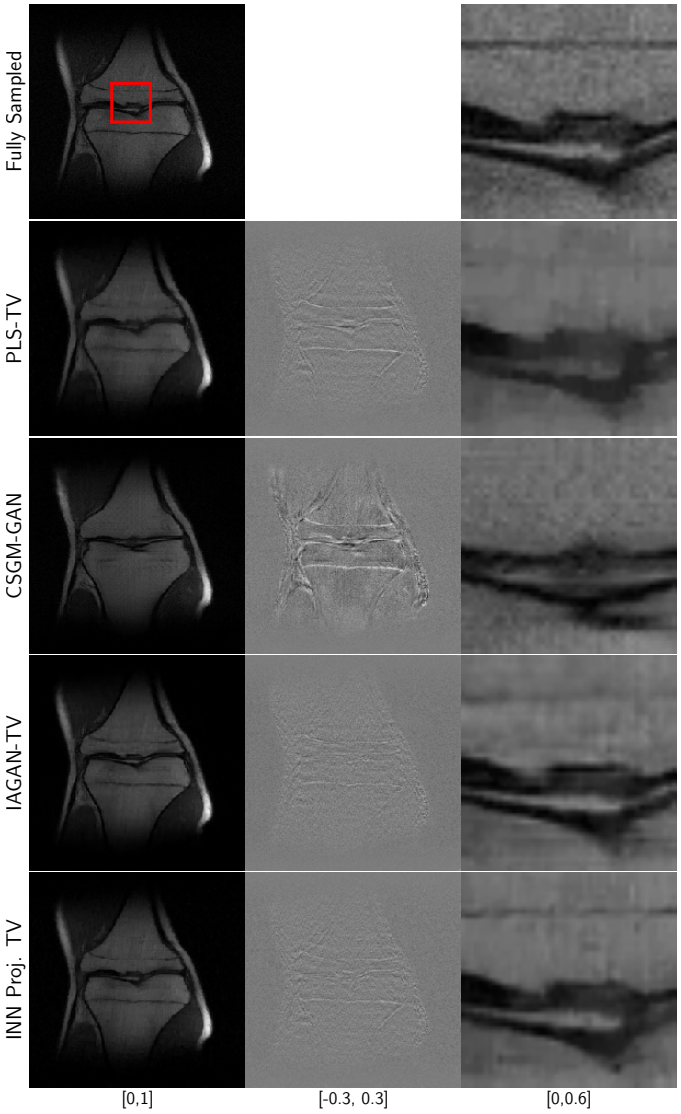


Fig. 11: The absolute value of coronal PD weighted knee images reconstructed from emulated single-coil measurements with Cartesian four-fold retrospective undersampling. The RMSE and SSIM values are displayed in Table V.

B. Reconstruction of latent-projected images

Estimates of 20 latent-projected images were obtained from noiseless, 8-fold undersampled measurements. One of the reconstructed images is shown along with the ground truth in the supplementary section. The mean and standard deviation of RMSE and SSIM values obtained over the ensemble were 0.0046 ± 0.0007 and 0.9956 ± 0.0012. This indicates that the measurement model was not exactly inverted. Here, although the proposed method performs worse than the noiseless, fully sampled case, it performs significantly better as compared to the noiseless undersampled case with the test data. This suggests that the inverse problem when restricted to $S_k \cap T_\nu$ is less ill-conditioned than the case where $\tilde{\mathbf{f}} \notin S_k \cap T_\nu$.

C. Reconstruction of knee and brain test images

1) Simulation study: Reconstruction from noiseless undersampled measurements: The RMSE and SSIM evaluation metrics for the test knee and brain images reconstructed from noiseless undersampled data, and the corresponding images are displayed in the supplementary section.

2) Simulation study: Reconstruction from undersampled measurements with 20 dB measurement SNR: Figures 5 and 6 display reconstructed images of a coronal knee test image from 8-fold and 20-fold noisy undersampled measurements respectively. One key observation is that for 8-fold subsampling, all algorithms except for CSGM-GAN performed well, in terms of RMSE and SSIM. This was because the 8-fold variable density Poisson disc undersampling mask is designed in order to keep the low frequency information intact, and randomly sample only the high frequency information with a variable density. It should be noted that due to the representation error, the CSGM-GAN reconstruction retained highly realistic features, some of which, were false. Further, it should be noted that the IAGAN-TV and the INN-based method seem to have performed the best in terms of recovering the finer features of the image. As shown in Fig. 6, for 20-fold undersampling, it was seen that the PLS-TV reconstruction has characteristic smoothing artifacts due to the TV regularization. Choosing lower regularization values led to noisier images, as shown in the supplementary section.

Similar observations can be made for the results of reconstruction of an axial brain image from 8-fold and 20-fold undersampled measurements, as shown in Fig. 7 and Fig. 9, respectively. In addition, it should be noted that for the 8-fold undersampling case, some of the finer features, such as the folds in the brain, are difficult to recover using PLS-TV, but were successfully recovered with both IAGAN-TV and the INN-based reconstruction. For the 20-fold undersampling case, all the methods face challenges in recovering finer features such as the folds of the brain, and may produce oversmoothened features or even realistic hallucinations. Finally, the results for the reconstruction of the pediatric brain image are shown in Fig. 10. Here, the out-of-distribution image was accurately recovered by the IAGAN-TV and the proposed method, but not in the case of CSGM-GAN. Here, the poor performance of CSGM-GAN could also be due to domain shifts unrelated to anatomical features.

3) Emulated experimental study: The absolute value of the image reconstructed by use of the proposed INN-based method from four-fold retrospectively undersampled emulated single-coil measurements is shown in Fig. 11, along with the IFFT-based reconstruction from fully sampled data and difference plots. Here, we see that the proposed method demonstrates superior performance among all the examined methods.

4) Root Mean square error and structural similarity: Root mean-squared error (RMSE) and structural similarity (SSIM) index values over an ensemble of 50 test images of each category described above were calculated. Ensemble mean and standard deviation of these values are displayed in Tables IV and V. It can be noted that, across several image categories, the performance of the IAGAN-TV and the proposed method was comparable and the best among all the methods compared, although IAGAN-TV outperformed the proposed method for some of the image categories by a small margin. For the emulated experimental study, RMSE

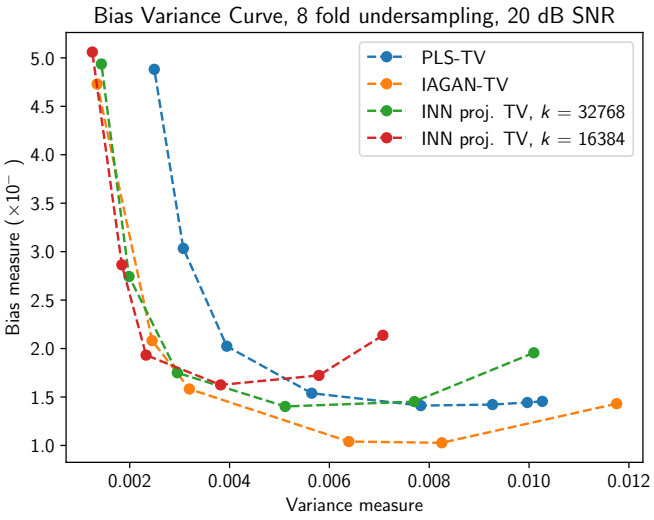


Fig. 12: Bias variance tradeoff analysis for 8-fold undersampling comparing PLS-TV, IAGAN-TV and the proposed method, while sweeping μ.

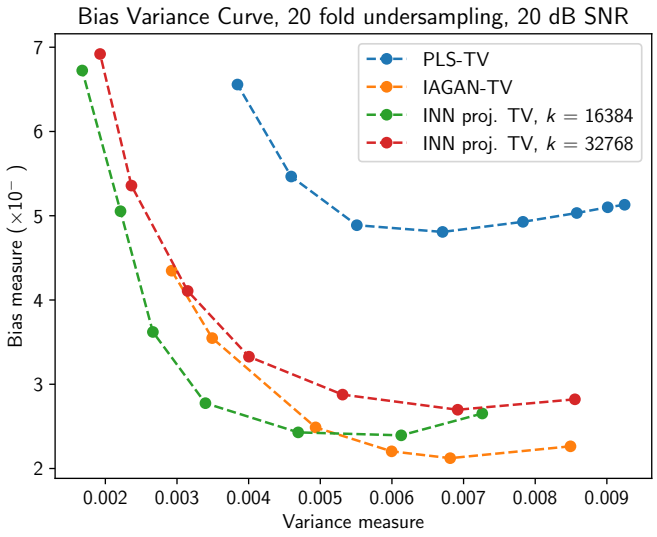


Fig. 13: Bias variance tradeoff analysis for 20-fold undersampling comparing PLS-TV, IAGAN-TV and the proposed method, while sweeping μ.

and SSIM values were computed with respect to the fully-sampled IFFT-based reconstruction. The performance of the proposed method outperformed all other examined methods for this study.

The statistical significance of the differences between the reconstruction methods was tested using the one way repeated measures ANOVA test, followed by post-hoc paired samples t-tests between pairs of algorithms, with the Bonferroni correction. Since the metrics obtained from CSGM-GAN violated some of the assumptions of the ANOVA test, it was left out of the statistical significance study. It was observed that IAGAN-TV and the proposed approach are both statistically significantly better than PLS-TV (with p-value $< 10^{-17}$ for the in-distribution images, p-value $< 10^{-12}$ for the out-of-distribution images, and p-value $< 10^{-6}$ for the emulated experimental study). For the in-distribution images in the simulation study, there is a small but statistically significant difference between the performance of IAGAN-TV and the proposed method, with IAGAN performing better (with p-value $< 10^{-5}$). However, for the out-of-distribution image, no statistically significant difference was observed between the two approaches. For the emulated experimental study, the proposed method significantly outperformed IAGAN-TV (with p-value $< 10^{-8}$).

5) Bias-Variance tradeoff: Although the evaluation of perceptual quality and quantitative evaluation in terms of RMSE and SSIM indicate the superiority of the INN-based reconstruction as compared to more traditional approaches, a task-based assessment of reconstruction algorithms is necessary to determine the superiority of one reconstruction algorithm to the other [33]. However, such a detailed task-based assessment of generative models based reconstruction algorithms is a substantial task in itself, and remains a topic for future study. Here, an analysis of the bias-variance trade-off is provided.

As described in equation Eq. (19), the INN-based reconstruction method involves the use of two explicit regularization parameters - (i) k, the dimensionality of the latent subspace containing the most important $\mathbf{z}$ components, and (ii) μ, the weight of the TV regularization. In the presented analysis, for a fixed value of k, μ was swept to obtain different values of bias. This entire procedure was repeated for another value of k. The results of both parts were compared with PLS-TV and IAGAN-TV, where the TV regularization weight was swept.

Bias-variance analysis was performed on images reconstructed from simulated measurements corresponding to both 8-fold and 20-fold undersampling patterns, with 20 dB measurement SNR. The ground truth used for this study was an image from the fastMRI knee dataset. Stylized, simulated undersampled single-coil MRI measurements were used. A dataset of reconstructed images $\{\hat{f}^{(i)}\}_{i=1}^{d}$ from measurements with $d = 100$ independent noise realizations was considered for every regularization setting. The bias $\mathbf{b}$ and the variance σ_i of a pixel i were calculated as:

$$\mathbf{b} = \frac{1}{d}\sum_{i=1}^{d}\mathbf{f}^{(i)} - \tilde{\mathbf{f}} \tag{24}$$

$$\sigma_j^2 = \frac{1}{d-1}\left(\mathbf{f}_j^{(i)} - \frac{1}{d}\sum_{i=1}^{d}\mathbf{f}_j^{(i)}\right)^2, \tag{25}$$

where $\tilde{\mathbf{f}}$ is the ground truth image. As a summary measure, the average squared bias $\frac{1}{n}\|\mathbf{b}\|_2^2$ versus the average variance $\frac{1}{n}\sum_{j=1}^{n}\sigma_j^2$ was plotted. Figures 12 and 13 show the bias-variance curves for 8 and 20-fold undersampling, respectively.

As can be seen, the bias and variance curves for the INN-based method lie below the curves for PLS-TV, which is indicative of superior performance over a range of regularization values. This also indicates that, while the transition from an over-smoothed image to a noisy image is such that intermediate images could be both noisy and oversmoothed, this trade-off is better for the proposed reconstruction approach.

TABLE IV: Comparison of RMSE and SSIM for different algorithms for undersampled data with 20 dB measurement SNR, for the simulation study, computed on an ensemble of 50 images. The values outside the parentheses denote the ensemble mean values of the metric, where as the values inside the parentheses denote the standard deviation (SD) of the metric.

Algorithm	Knee (in dist.) 8x		Knee (in dist.) 20x		Brain (in dist.) 8x		Brain (in dist.) 20x	
	RMSE mean (RMSE SD)	SSIM mean (SSIM SD)	RMSE mean (RMSE SD)	SSIM mean (SSIM SD)	RMSE mean (RMSE SD)	SSIM mean (SSIM SD)	RMSE mean (RMSE SD)	SSIM mean (SSIM SD)
PLS-TV	0.0122 (0.0033)	0.9736 (0.0108)	0.0178 (0.0050)	0.9556 (0.0172)	0.0228 (0.0033)	0.9609 (0.0093)	0.0473 (0.0098)	0.8798 (0.0337)
CSGM-GAN	0.0381 (0.0157)	0.8808 (0.0485)	0.0389 (0.0154)	0.8753 (0.0470)	0.0721 (0.0318)	0.8174 (0.0633)	0.0725 (0.0249)	0.8153 (0.0598)
IAGAN-TV	**0.0099** (0.0026)	**0.9844** (0.0064)	**0.0140** (0.0041)	**0.9705** (0.0135)	**0.0148** (0.0024)	**0.9794** (0.0061)	**0.0246** (0.0043)	**0.9483** (0.0146)
INN Proj. TV	0.0102 (0.0027)	0.9829 (0.0070)	0.0147 (0.0042)	0.9678 (0.0137)	0.0163 (0.0028)	0.9723 (0.0086)	0.0262 (0.0049)	0.9414 (0.0177)

TABLE V: Comparison of RMSE and SSIM for different algorithms over an ensemble of 50 out-of-distribution brain images for the simulation study and 40 knee images for the emulated experimental study. The values inside the parentheses denote the standard deviation (SD) of the metric.

Algorithm	Brain (out of dist.)		Emulated experimental	
	RMSE mean (RMSE SD)	SSIM mean (SSIM SD)	RMSE mean (RMSE SD)	SSIM mean (SSIM SD)
PLS-TV	0.0124 (0.0012)	0.9813 (0.0046)	0.0148 (0.0046)	0.9151 (0.0459)
CSGM-GAN	0.0516 (0.0088)	0.7932 (0.022)	0.0283 (0.0072)	0.8815 (0.0455)
IAGAN-TV	0.0119 (0.0013)	**0.9847** (0.0041)	0.0142 (0.0049)	0.9098 (0.0590)
INN Proj. TV	**0.0118** (0.0011)	0.9846 (0.0038)	**0.0135** (0.0044)	**0.9320** (0.0360)

VII. Discussion and conclusion

Due to the design of the measurement operator in the case of variable density Poisson disc undersampling, PLS-TV outperformed CSGM-GAN in most cases in terms of MSE and SSIM. The discrepancies in the image estimate from CSGM-GAN are also visually evident from the difference plot. This is consistent with the observation made by Bora *et. al* [19], where they report that CSGM-GAN outperformed sparsity-based methods (with respect to MSE) only in the cases of severe undersampling. Here, it can be seen that if the measurement operator is well designed, it is possible to have formally severe undersampling scenarios where the PLS-TV outperforms CSGM-GAN. Moreover, it was observed that that the INN-based methods, as well as IAGAN-TV were successful in removing the plausible but false features that could be present in images reconstructed by CSGM-GAN.

Better trained networks give better performance, since they impose a better generative prior - a fact that was also observed when testing reconstruction using INNs trained with poorer hyperparameter settings. The current state-of-the-art GANs possess superior generative performance compared to state-of-the-art for invertible generative models, as evidenced by literature [25], [40], [42] as well as the FID scores calculated in Table III. This explains why in several cases, IAGAN-TV outperformed the proposed method. However, one key thing to note here is that for the INN-based reconstruction, the parameters of the network were *not* adapted, and still a performance very close to IAGAN-TV was achieved. Also, the IAGAN-TV performance was achieved using early stopping.

Moreover, the number of parameters that need to be optimized for the IAGAN-TV would increase as the complexity of GANs increases, where for the INN, optimization over only the latent-space vector was needed in order to achieve comparable performance. For instance, in this study, IAGAN-TV requires optimization over more than 23 million parameters whereas the proposed approach require optimization over less than 33000 parameters for image reconstruction from simulated measurements, and around 65000 parameters for the experimental measurements. Finally, the IAGAN-TV optimization was carried out by initializing with the CSGM-GAN solution, which itself required about 10 independent random restarts to achieve a reasonable reconstruction. Including good initialization via CSGM-GAN and the several random initializations needed thereof, the net runtime for IAGAN-TV is around 120 minutes on a single NVIDIA 1080 Ti GPU, whereas the proposed method requires around 20 minutes.

It is important to note how the studies conducted here relate to a real experimental MRI scenario. The forward operator used in the iterative reconstruction in all our studies was the discrete Fourier transform followed by undersampling in the Fourier domain based on a binary mask. This is a stylized simulation and does not reflect all the complexities of the true MRI forward model, such as the single-coil sensitivity. For the in-distribution simulation studies, the ground truth images were real-valued floating point numbers. For the out-of-distribution simulation studies, however, the ground truth images were originally real-valued, skull-stripped and compressed to 8 bit integers. The noise model for the simulated study was iid Gaussian. All these factors may simplify the image reconstruction task substantially. For the emulated experimental study, the IFFT of the fully sampled k-space corresponds to complex-valued floating point numbers and has non-trivial phase variations. Since image reconstruction was performed directly from the emulated undersampled single-coil measurements, the implicit noise model is expected to be more realistic than the simulation studies. The undersampling, however, was retrospective and the measurement data were generated as a linear combination of responses from raw multi-coil data.

In conclusion, a new method of image reconstruction from incomplete measurements using invertible generative priors was proposed, based on a novel regularization strategy for INNs with a multiscale architecture. This method was eval-

uated and compared with other competing methods on the problem of estimating images from simulated and emulated experimental undersampled MRI measurements. Some important extensions of this work include comparisons with regularizaton methods adapted to continuous signal variation such as total generalized variation (TGV) [61], developing strategies for multi-coil MRI and 3D image reconstruction, as well as a task-based evaluation of generative model-based image reconstruction methods. Another interesting avenue for future research is examining different strategies for penalizing the latent vector, such as constraining the latent vector to lie on a surface on which a random latent vector concentrates [62].

REFERENCES

[1] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 6, pp. 797–829, 2006.

[2] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[3] E. Candes, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements arXiv: math/0503066v2 [math. NA] 7 Dec 2005," *Science*, vol. 40698, pp. 1–15, 2005.

[4] E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Physics in Medicine & Biology*, vol. 53, no. 17, p. 4777, 2008.

[5] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[6] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[7] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[8] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[9] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[10] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[11] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE Journal of selected topics in signal processing*, vol. 4, no. 2, pp. 375–391, 2010.

[12] J. A. Clarkson and C. R. Adams, "On definitions of bounded variation for functions of two variables," *Transactions of the American Mathematical Society*, vol. 35, no. 4, pp. 824–854, 1933.

[13] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE transactions on medical imaging*, vol. 30, no. 5, pp. 1028–1041, 2010.

[14] B. Wen, S. Ravishankar, L. Pfister, and Y. Bresler, "Transform learning for magnetic resonance image reconstruction: From model-based learning to building neural networks," *arXiv preprint arXiv:1903.11431*, 2019.

[15] S. Ravishankar, J. C. Ye, and J. A. Fessler, "Image reconstruction: From sparsity to data-adaptive methods and machine learning," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 86–109, 2019.

[16] B. Kelly, T. P. Matthews, and M. A. Anastasio, "Deep learning-guided image reconstruction from incomplete data," *arXiv preprint arXiv:1709.00584*, 2017.

[17] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[18] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[19] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 537–546.

[20] J. Sun, H. Li, Z. Xu *et al.*, "Deep ADMM-Net for compressive sensing MRI," in *Advances in neural information processing systems*, 2016, pp. 10–18.

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

[22] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated MRI data," *Magnetic resonance in medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.

[23] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, 2017.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[25] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[26] M. Asim, G. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: mitigating representation error and dataset bias," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 4577–4587.

[27] M. Dhar, A. Grover, and S. Ermon, "Modeling sparse deviations for compressed sensing using generative models," *arXiv preprint arXiv:1807.01442*, 2018.

[28] D. Van Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis, "Compressed sensing with deep image prior and learned regularization," *arXiv preprint arXiv:1806.06438*, 2018.

[29] S. A. Hussein, T. Tirer, and R. Giryes, "Image-adaptive GAN based reconstruction," *arXiv preprint arXiv:1906.05284*, 2019.

[30] S. Bhadra, W. Zhou, and M. A. Anastasio, "Medical image reconstruction with image-adaptive priors learned by use of generative adversarial networks," in *Medical Imaging 2020: Physics of Medical Imaging*, vol. 11312. International Society for Optics and Photonics, 2020, p. 113120V.

[31] D. Narnhofer, K. Hammernik, F. Knoll, and T. Pock, "Inverse GANs for accelerated MRI reconstruction," in *Wavelets and Sparsity XVIII*, vol. 11138. International Society for Optics and Photonics, 2019, p. 111381A.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[33] H. H. Barrett and K. J. Myers, *Foundations of image science*. John Wiley & Sons, 2013.

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[35] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *arXiv preprint arXiv:1912.04958*, 2019.

[36] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing Flows for Probabilistic Modeling and Inference," *arXiv preprint arXiv:1912.02762*, 2019.

[37] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

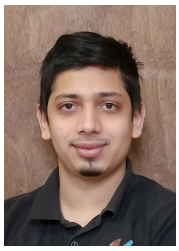
[38] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[39] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[40] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[41] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," *arXiv preprint arXiv:1811.00995*, 2018.

[42] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving flow-based generative models with variational dequantization and architecture design," *arXiv preprint arXiv:1902.00275*, 2019.
[43] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.
[44] E. Y. Sidky, I. Lorente, J. G. Brankov, and X. Pan, "Do CNNs solve the CT inverse problem?" *arXiv preprint arXiv:2005.10755*, 2020.
[45] A. Bastounis and A. C. Hansen, "On the absence of the RIP in real-world applications of compressed sensing and the RIP in levels," *arXiv preprint arXiv:1411.4449*, 2014.
[46] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," in *Forum of Mathematics, Sigma*, vol. 5. Cambridge University Press, 2017.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[48] Q. Zhou, C. Zhou, H. Hu, Y. Chen, S. Chen, and X. Li, "Towards the Automation of Deep Image Prior," *arXiv preprint arXiv:1911.07185*, 2019.
[49] J. Zbontar, F. Knoll, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana *et al.*, "fastMRI: An open dataset and benchmarks for accelerated MRI," *arXiv preprint arXiv:1811.08839*, 2018.
[50] A. M. S. Maallo, E. Freud, T. T. Liu, C. Patterson, and M. Behrmann, "Effects of unilateral cortical resection of the visual cortex on bilateral human white matter," *NeuroImage*, vol. 207, p. 116345, 2020.
[51] A. Maallo, T. Liu, E. Freud, C. Patterson, and M. Behrmann, "Pediatric epilepsy resection MRI dataset," 2019, https://doi.org/10.1184/R1/9856205.
[52] M. Uecker, F. Ong, J. I. Tamir, D. Bahri, P. Virtue, J. Y. Cheng, T. Zhang, and M. Lustig, "Berkeley advanced reconstruction toolbox," in *Proc. Intl. Soc. Mag. Reson. Med*, vol. 23, no. 2486, 2015.
[53] M. Tygert and J. Zbontar, "Simulating single-coil mri from the responses of multiple coils," *arXiv preprint arXiv:1811.08026*, 2018.
[54] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.
[55] "Calculation of scale term #62, Github issues," 2018, https://github.com/openai/glow/issues/62.
[56] "Hal Cluster," https://wiki.ncsa.illinois.edu/display/ISL20/HAL+cluster.
[57] "Progressive Growing of GANs for Improved Quality, Stability, and Variation – Official TensorFlow implementation of the ICLR 2018 paper," 2018, https://github.com/tkarras/progressive_growing_of_gans.
[58] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Two time-scale update rule for training gans: Fréchet inception distance (fid)," *Github*, 2017. [Online]. Available: https://github.com/bioinf-jku/TTUR/
[59] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study." arxiv e-prints," *arXiv preprint arXiv:1711.10337*, 2017.
[60] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[61] F. Knoll, K. Bredies, T. Pock, and R. Stollberger, "Second order total generalized variation (tgv) for mri," *Magnetic resonance in medicine*, vol. 65, no. 2, pp. 480–491, 2011.
[62] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2437–2445.



Varun A. Kelkar is a Ph.D. candidate in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, USA. He received the M.S. degree in Electrical and Computer Engineering from UIUC in 2019, and the B.Tech. degree in Engineering Physics from the Indian Institute of Technology Madras, TN, India in 2017. His research interests include computational imaging, inverse problems, signal processing, optics and machine learning. He is a member of the SPIE, and was a recipient of the SPIE Optics and Photonics Education Scholarship in 2019.

Sayantan Bhadra received the B.E. degree in Electronics and Telecommunication Engineering from Jadavpur University, West Bengal, India, in 2016. He is currently working toward the Ph.D degree in Computer Science and Engineering at the Washington University in St. Louis, MO, USA. He is also a visiting research scholar in the Computational Imaging Science Laboratory, Department of Bioengineering, University of Illinois at Urbana-Champaign, IL, USA. His research interests include image reconstruction and machine learning for medical imaging applications.

Mark A. Anastasio is the Donald Biggar Willett Professor in Engineering and the Head of the Department of Bioengineering UIUC. Before joining UIUC in 2019, he was a professor of biomedical engineering and the founding director of one of the nation's first stand-alone PhD programs in imaging science at Washington University in St. Louis. Dr. Anastasio's research addresses the computational aspects of image formation, modern imaging science, and machine learning. He has conducted research in the fields of diffraction tomography, X-ray phase-contrast imaging, and ultrasound tomography. He is also a leading authority on photoacoustic computed tomography. His research has been continuously funded by the NIH and NSF and he was the recipient of an NSF CAREER Award. He is a Fellow of the American Institute for Medical and Biological Engineering and the SPIE and served as the Chair of the NIH BMIT-B and EITA Study Sections.

Compressible Latent-Space Invertible Networks for Generative Model-Constrained Image Reconstruction – Supplementary Information

Varun A. Kelkar, Sayantan Bhadra, and Mark A. Anastasio, *Senior Member, IEEE*

Index Terms

Image reconstruction, compressive sensing, generative neural networks, invertible neural networks

I. Experiments for testing the compressibility of latent space of the INN

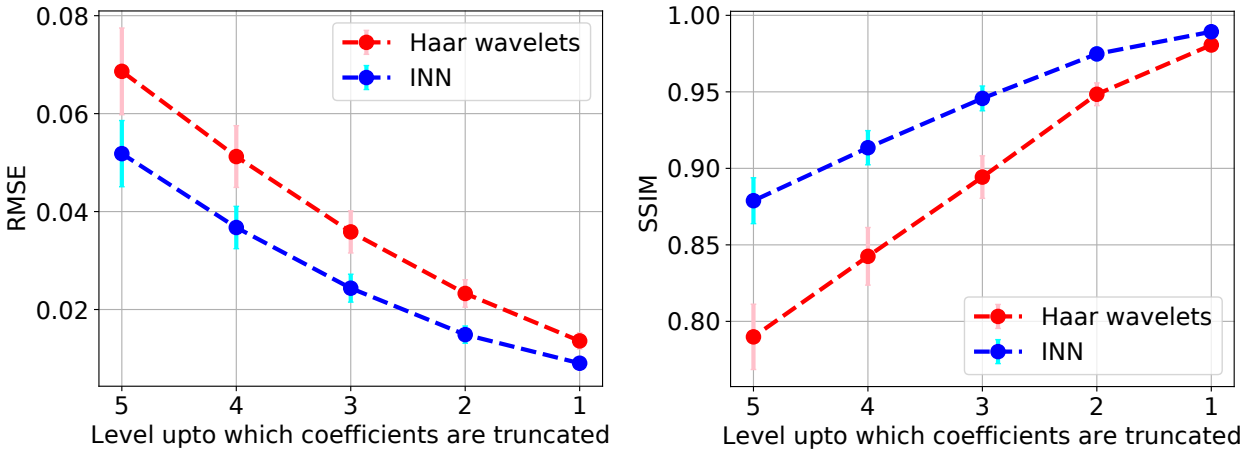
As described in Section III of the manuscript, the effect of the latent-space compressibility for the INN was examined on an ensemble of 500 images from a test dataset of coronal knee images that was kept out of the INN training dataset. This was done as follows. First, for an image $\mathbf{f}_{\text{orig}}$ in the ensemble, the exact latent representation $\mathbf{z}_{\text{orig}}$ was computed. This can be divided into multiple sections $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(L)}$ based on the multilevel architecture of the INN. Next, all sections from $\mathbf{z}^{(1)} \ldots \mathbf{z}^{(i)}$ were progressively set to zero such that only 50%, 25%, 12.5%, 6.25% and 3.125% of the components of $\mathbf{z}$ remain non-zero. For these modified latent space vectors $\mathbf{z}_{50\%}$, $\mathbf{z}_{25\%}$, $\mathbf{z}_{12.5\%}$, $\mathbf{z}_{6.25\%}$ and $\mathbf{z}_{3.125\%}$, the corresponding images $\mathbf{f}_P = G_{\text{inn}}(\mathbf{z}_P),\ P = 50\%,\ 25\%,\ 12.5\%,\ 6.25\%$ and 3.125% were computed, and the root mean square errors (RMSEs) $\|\mathbf{f}_P - \mathbf{f}_{\text{orig}}\| / \sqrt{n}$ with respect to the original image $\mathbf{f}_{\text{orig}}$ were calculated.

This work was supported in part by NIH Awards EB020604, EB023045, NS102213, EB028652, and NSF Award DMS1614305.

Varun A. Kelkar is with the Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: vak2@illinois.edu).

Sayantan Bhadra is with the Department of Computer Science and Engineering, Washington University in Saint Louis, Saint Louis, MO USA (e-mail: sayantanbhadra@wustl.edu).

Mark A. Anastasio is with the Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: maa@illinois.edu).



(a) Truncation RMSE values

(b) Truncation SSIM values

Fig. 1: Truncation RMSE and SSIM values for INN and Haar wavelet transform.

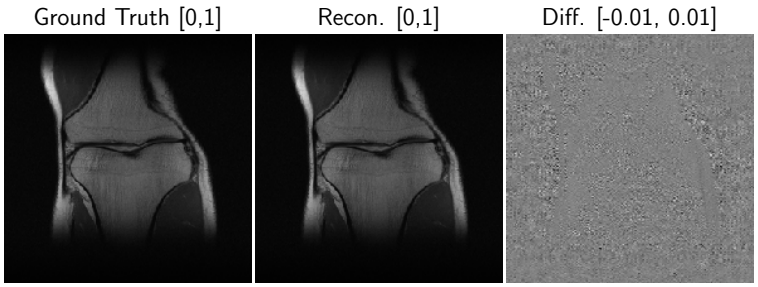
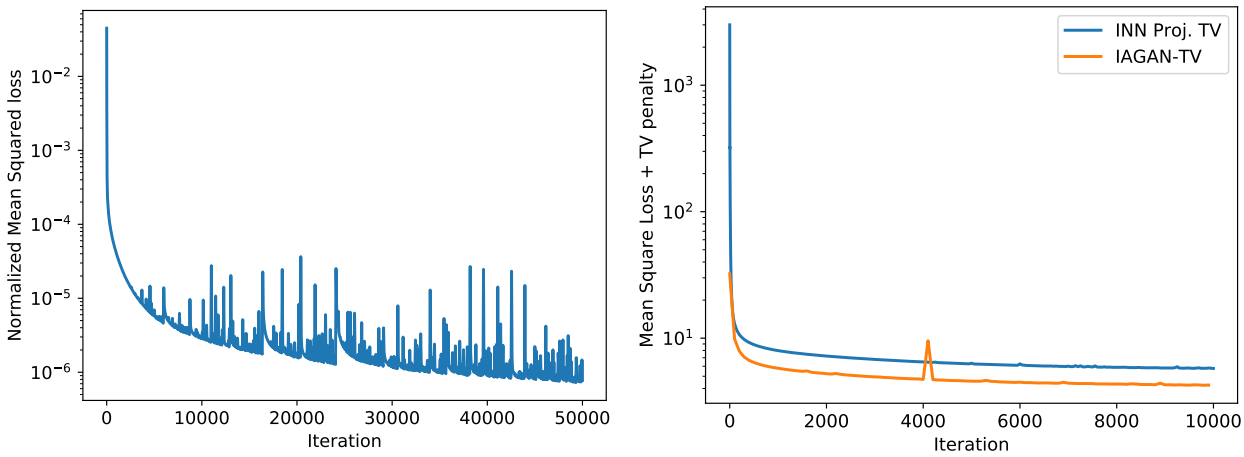


Fig. 2: Ground truth, reconstruction and difference plot for reconstruction of a coronal knee image from fully sampled, noiseless measurement data

The error versus the percentage of $\mathbf{z}$ coefficients kept, averaged over the entire ensemble, was computed. This was then compared with the Haar wavelet transform as follows. The Haar wavelet transform with L levels (same as that of the INN) was computed for each image in the ensemble. If the multilevel Haar wavelet coefficients are divided into $(\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(L)})$, then the i-th truncation level was computed by setting $\mathbf{c}^{(1)} \ldots \mathbf{c}^{(i)}$ to zero. This was then compared with the i-th truncation level of the INN. The RMSE and SSIM truncation errors for both the INN and the Haar wavelet transform are shown in Figures 1a and 1b respectively.



(a) Normalized mean square loss plotted against the iteration for the inverse crime study.

(b) Mean squared loss + TV penalty versus iteration, for 8 fold noisy undersampled measurements.

Fig. 3: Loss function profiles for the inverse crime study and for reconstruction from noisy undersampled simulated measurements.

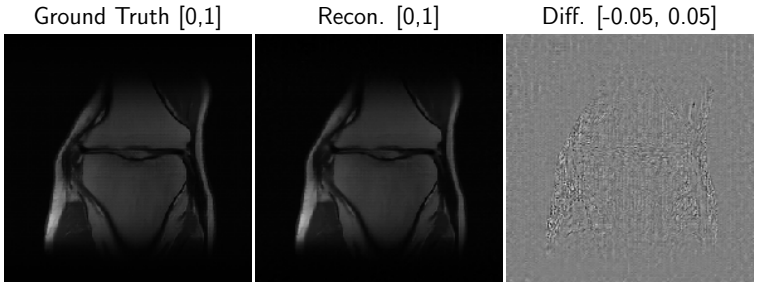


Fig. 4: Ground truth, reconstruction and difference plot for reconstruction of a latent-projected image, reconstructed from noiseless, 8 fold undersampled measurements

II. FISTA FOR SOLVING THE PLS-TV OPTIMIZATION PROBLEM

The PLS-TV optimization problem – formulated as

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\mathbf{g} - H\mathbf{f}\|_2^2 + \lambda \|\mathbf{f}\|_{\mathrm{TV}}\,, \tag{1}$$

where

$$\|\mathbf{f}\|_{\mathrm{TV}} = \sum_{i,j} (|\mathbf{f}_{i,j} - \mathbf{f}_{i,j+1}| + |\mathbf{f}_{i,j} - \mathbf{f}_{i+1,j}|) \tag{2}$$

and $\mathbf{f}_{i,j}$ represents the (i,j)-th pixel of the image $\mathbf{f}$ – was solved using FISTA [1]. The basic step in FISTA, given by Eq. (4.1) in the original paper by Beck, *et al.*, is a composition of a

TABLE I: Comparison of RMSE and SSIM for different algorithms for reconstruction from noiseless undersampled data for the simulation study. The metrics are evaluated on a single test image shown in Figures 5, 6, 7 and 8. The metrics in brackets are evaluated on a region of interest (ROI) selected as shown in Figures 5, 6, 7 and 8.

	Knee (in dist.) 8x		Knee (in dist.) 20x		Brain (in dist.) 8x		Brain (in dist.) 20x	
	RMSE full (ROI)	SSIM full (ROI)	RMSE full (ROI)	SSIM full (ROI)	RMSE full (ROI)	SSIM full (ROI)	RMSE full (ROI)	SSIM full (ROI)
PLS-TV	0.0114 (0.021)	0.9793 (0.9577)	0.0216 (0.049)	0.949 (0.8516)	0.0195 (0.0232)	0.9672 (0.9539)	0.0547 (0.0631)	0.8421 (0.7689)
CSGM-GAN	0.0298 (0.0699)	0.9266 (0.7589)	0.0327 (0.0738)	0.9165 (0.7446)	0.0545 (0.0733)	0.8342 (0.7381)	0.0553 (0.0723)	0.8293 (0.7543)
IAGAN-TV	**0.0101 (0.0185)**	**0.9833 (0.9645)**	**0.0139 (0.0285)**	**0.9727 (0.9289)**	**0.0114 (0.0138)**	**0.9836 (0.9785)**	**0.0238 (0.0293)**	**0.9474** (0.9267)
INN Proj. TV	0.0113 (0.0206)	0.9789 (0.9589)	0.0156 (0.0325)	0.9666 (0.9157)	0.014 (0.0165)	0.9757 (0.9708)	0.0256 **(0.0293)**	0.9398 **(0.9289)**

gradient update and a proximal update. The gradient update was performed using Python, with automatic gradient computation from Tensorflow. The proximal update was performed using the Python package ProxTV [2]. For the emulated experimental study, where $\mathbf{f}$ is complex valued, the real and imaginary parts were considered separately for the computation of the proximal update.

III. Reconstruction from noiseless, fully sampled k-space measurements

This study was conducted to analyze the loss decay during the iterative optimization, when the forward operator H is bijective. Let $\mathbf{g} = H\tilde{\mathbf{f}}$ be the measurement corresponding to the unknown true object $\tilde{\mathbf{f}}$. Figure 2 displays the results of this study. Figure 3a shows the normalized mean square loss versus the iteration. As can be seen in the figure, the loss does not go down to zero to machine precision, but the mean square reconstruction error goes down to around 10^{-6}. Thus, similar to other deep learning based methods for image reconstruction [3], we obtain only an

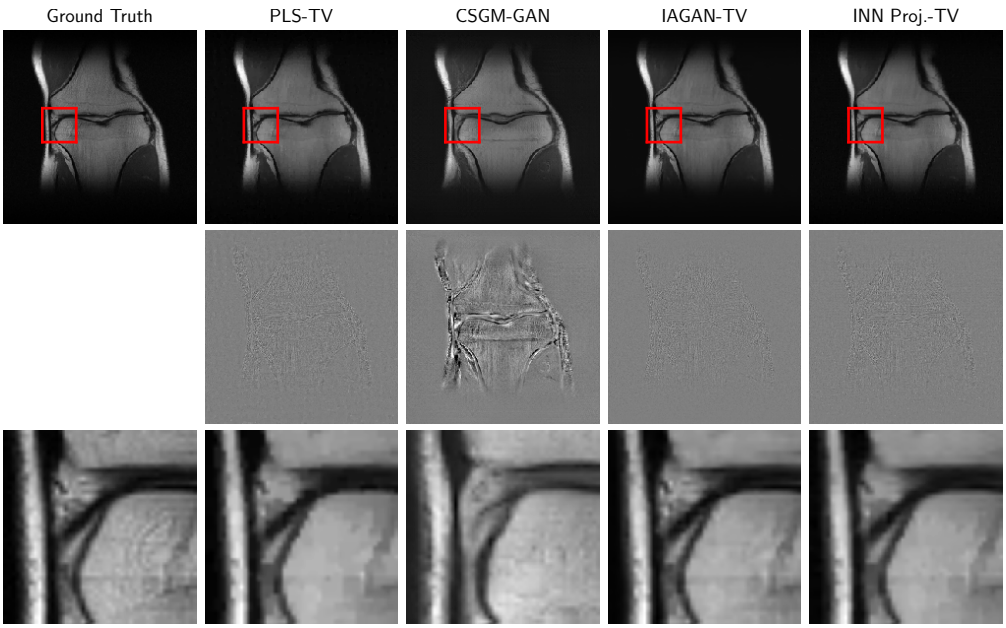


Fig. 5: Ground truth, difference plots and reconstruction results for a coronal PD weighted knee image without fat suppression from noiseless 8 fold undersampled measurements.

approximate solution to

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \|\mathbf{g} - HG_{\text{inn}}(\mathbf{z})\|_2^2 - \lambda \log p_Z(\mathbf{z}) + \mu \|G_{\text{inn}}(\mathbf{z})\|_{\text{TV}}\,,$$
$$\text{subject to} \quad \mathbf{z}_{1:n-k} = 0,$$
$$\hat{\mathbf{f}} \equiv G_{\text{inn}}(\hat{\mathbf{z}}), \tag{3}$$

even with complete measurements, which can be attributed to the fact that a non-convex optimization problem is solved by use of gradient-based methods.

IV. Reconstruction of latent-projected images from noiseless, simulated undersampled measurements

Here, a ground-truth image $\tilde{\mathbf{f}} \in S_k \cap T_\nu$ is referred to as a latent-projected image, where $S_k = \{\mathbf{f}\ s.t.\ \left\|G_{\text{inn}}^{-1}(\mathbf{f})\right\| \leq k\}$, and $T_\nu = \{\mathbf{f}\ s.t.\ \|\mathbf{f}\|_{\text{TV}} \leq \nu\}$. Since the measurements are

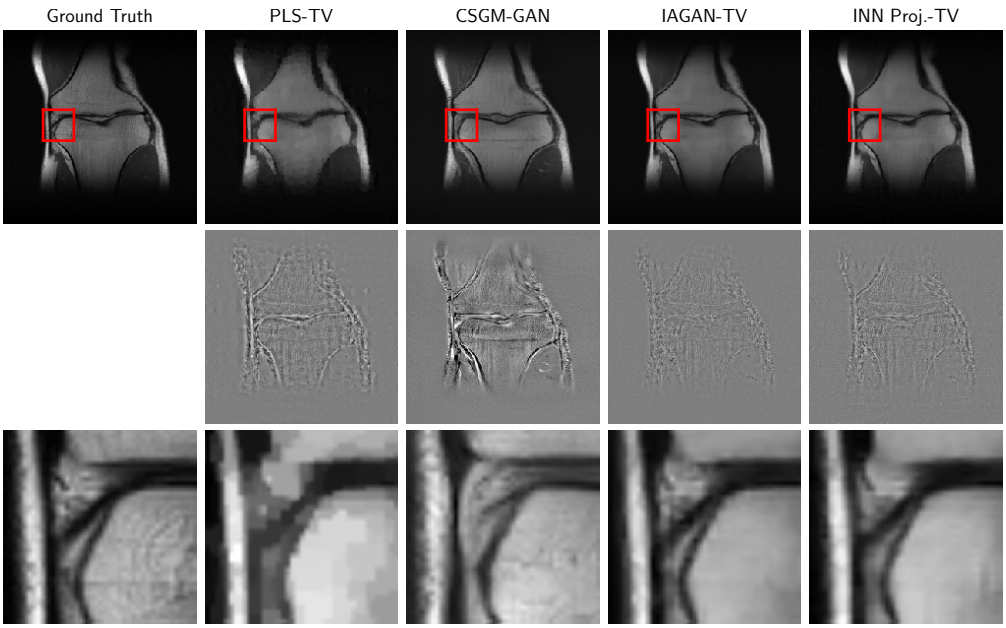


Fig. 6: Ground truth, difference plots and reconstruction results for a coronal PD weighted knee image without fat suppression from noiseless 20 fold undersampled measurements.

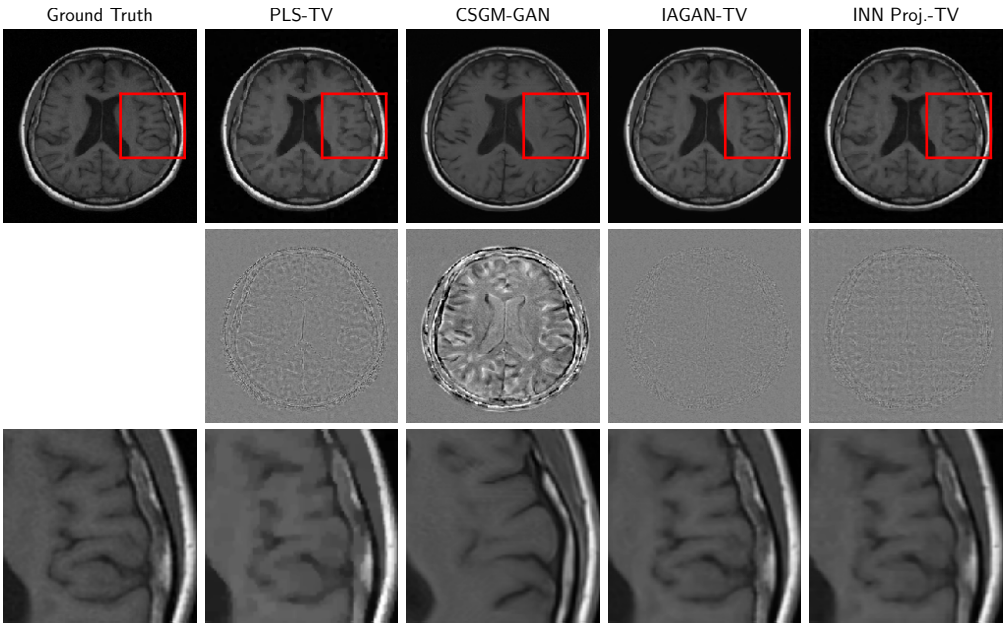


Fig. 7: Ground truth, difference plots and reconstruction results for a T1 weighted axial brain image from noiseless 8 fold undersampled measurements.

noiseless, this image perfectly satisfies the measurement model. Image reconstruction on an ensemble of 20 images. One of the reconstructed images is shown in Fig. 4.

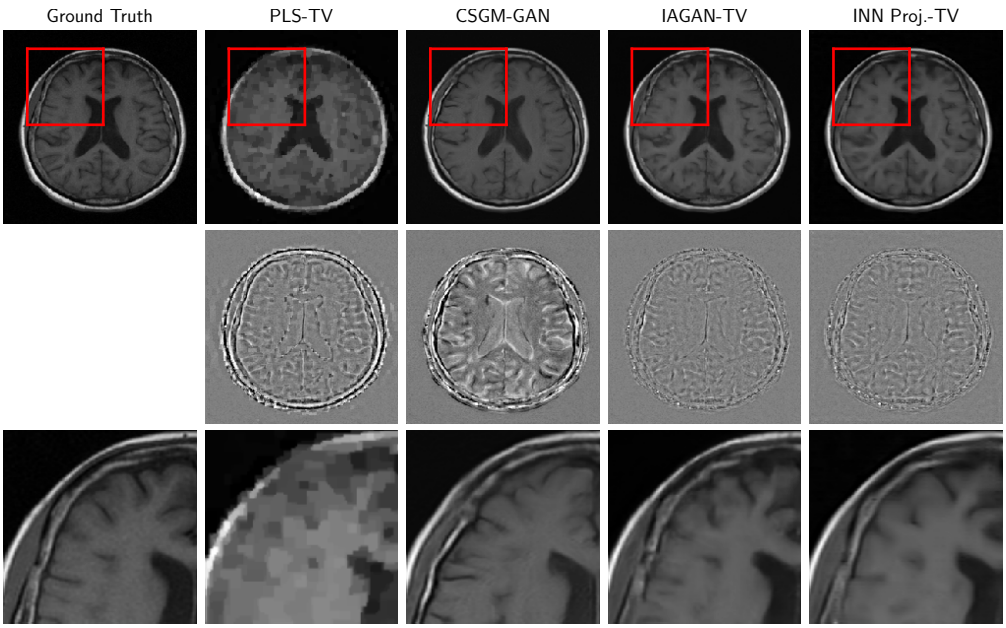


Fig. 8: Ground truth, difference plots and reconstruction results for a T1 weighted axial brain image from noiseless 20 fold undersampled measurements.

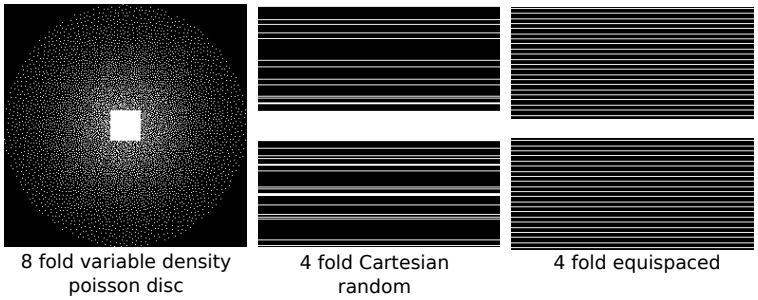
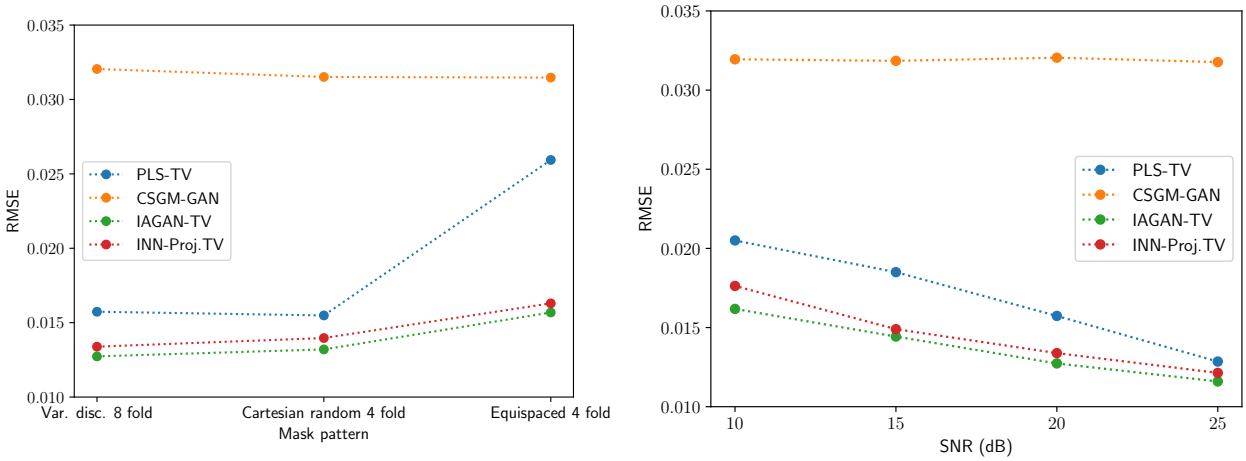


Fig. 9: Undersampling masks used for the mask variation study.

V. Simulation study: reconstruction of images from undersampled measurements

Figures 5 and 6 display reconstructed images of a coronal knee image from *noiseless* 8 fold and 20 fold undersampled measurements respectively. Figures 7 and 8 display reconstructed images of an axial brain image from noiseless 8 fold and 20 fold undersampled measurements respectively. The corresponding RMSE and SSIM values are shown in Table I.



(a) RMSE performance of the algorithms for different undersampling masks

(b) RMSE performance of the algorithms for different measurement noise levels

Fig. 10: RMSE performance of the algorithms for different undersampling masks and different measurement noise levels.

A. *Convergence Analysis*

For the reconstruction of a coronal knee image from the 8 fold undersampled measurements with 20 dB measurement SNR, the total loss at iteration i, defined by

$$\ell^{(i)} = \frac{1}{n}\left\|\mathbf{g} - HG_{\text{inn}}(\mathbf{z}^{(i)})\right\|_2^2 + \mu\left\|G_{\text{inn}}(\mathbf{z}^{(i)})\right\|_{\text{TV}} \tag{4}$$

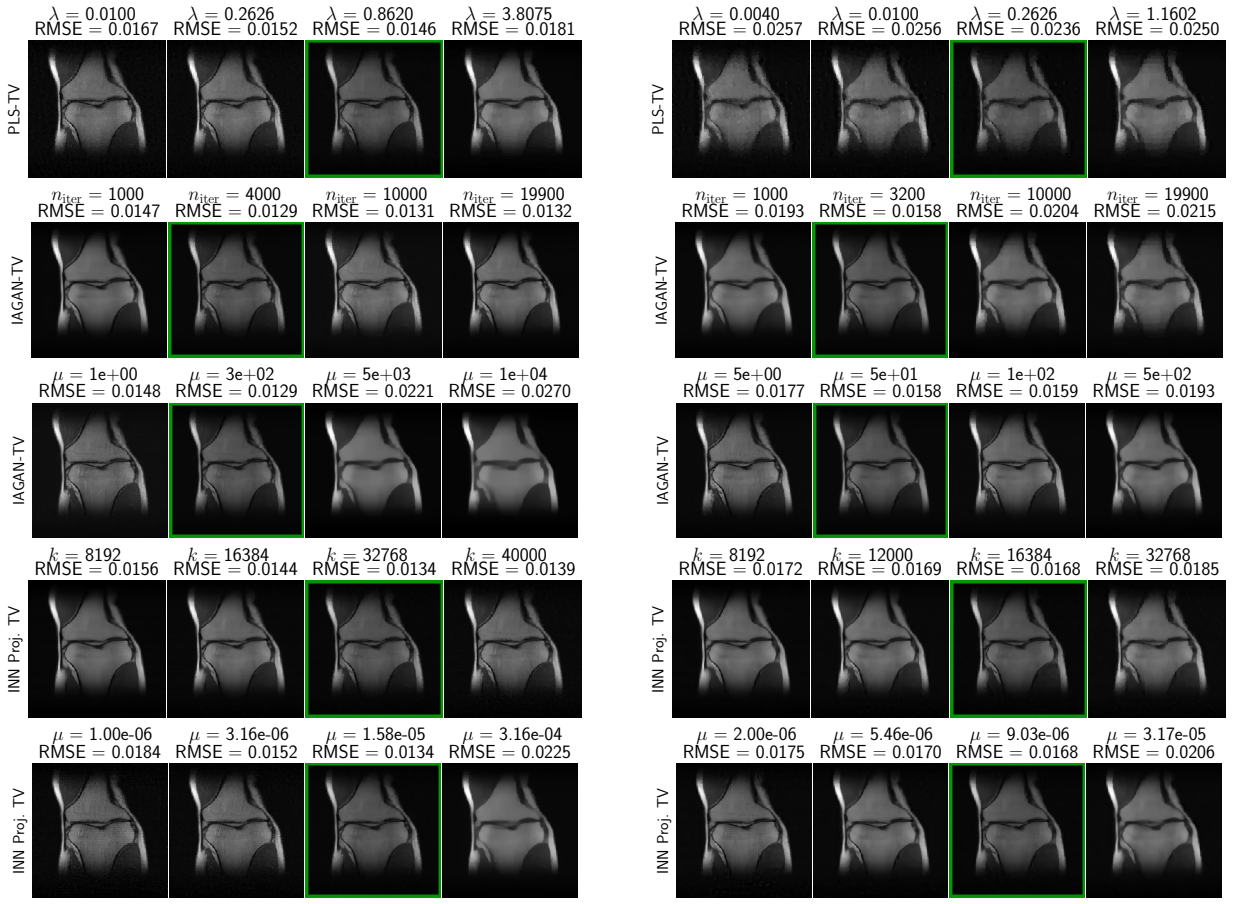
was plotted versus the iteration i, in Fig. 3b.

VI. Mask variation study

The manuscript describes the performance of the PLS-TV, CSGM-GAN, IAGAN-TV and INN Proj.TV methods for reconstruction of images from measurements generated by use of R = 8 and R = 20 variable density Poisson disc undersampling masks for the simulation study, and 4 fold Cartesian random undersampling mask for the emulated experimental study. Here, an estimate of a single knee image is obtained from simulated stylized MRI measurements that are undersampled using different masks. The masks used for this study are shown in Fig. 9. The corresponding RMSE values for the various reconstruction methods are shown in Fig. 10a.

VII. Noise variation study

Here, an estimate of a single knee image is obtained from simulated stylized MRI measurements that are undersampled using the 8 fold variable density poisson disc undersampling



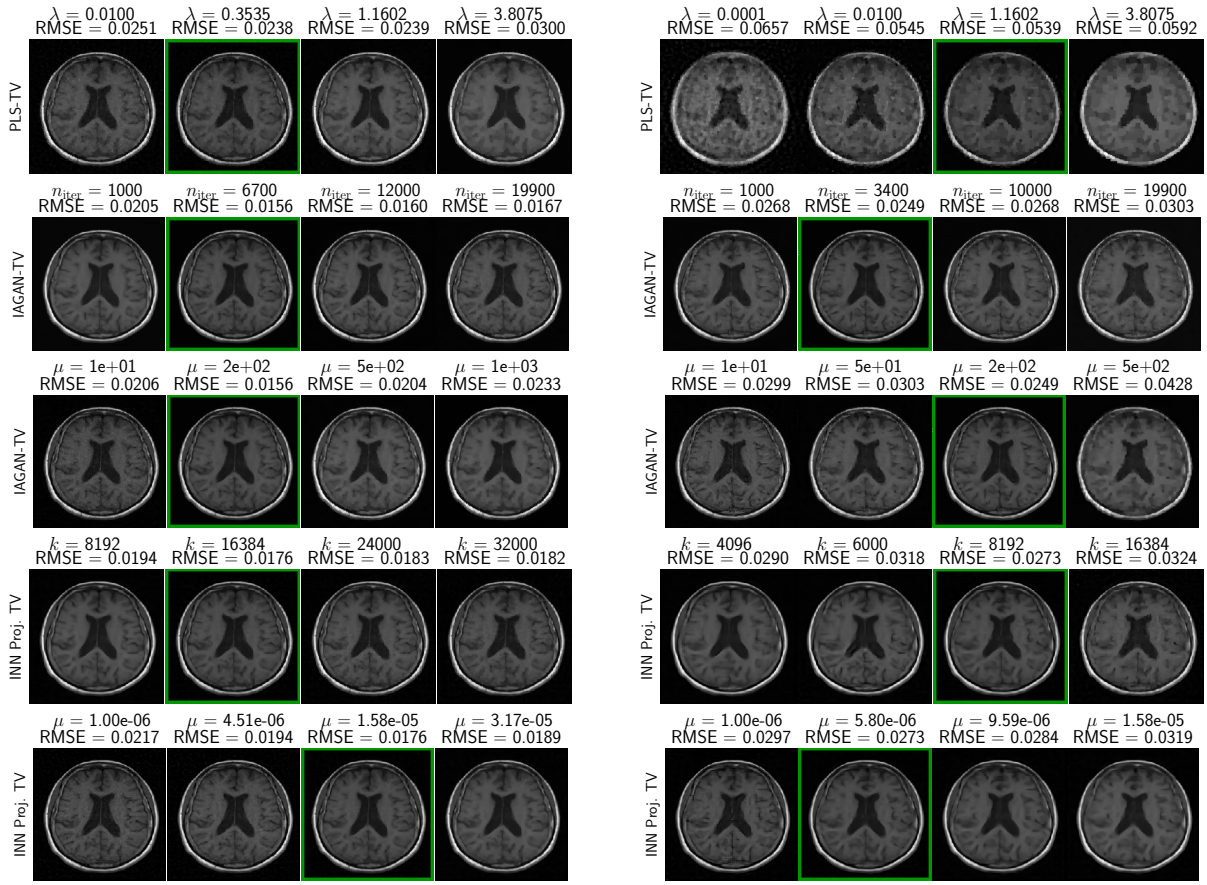
(a) Reconstruction from 8 fold undersampled measurements **(b)** Reconstruction from 20 fold undersampled measurements

Fig. 11: Reconstructed coronal PD weighted knee images from 8 and 20 fold undersampled measurements in the simulation study, for various regularization parameter settings for different reconstruction methods. λ denotes the TV regularization weight for PLS-TV, n_{iter} denotes the number of iterations for IAGAN-TV (regularization via early stopping), k denotes the dimensionality of the latent subspace for INN Proj. TV, and μ is used to denote the TV regularization weight for IAGAN-TV and INN Proj. TV.

mask, with varying levels of noise added to the measurements. The RMSE performance of the four reconstruction methods for measurement SNRs ranging from 25 dB to 10 dB is shown in Fig. 10b.

VIII. Regularization parameter selection

The regularization parameter λ for PLS-TV was chosen by line search, and the parameter giving the best RMSE value was chosen. The regularization parameters for IAGAN-TV (i.e. the stopping iteration n_{iter} and the TV regularization weight μ), and those for the proposed method



(a) Reconstruction from 20 fold undersampled measurements **(b)** Reconstruction from 20 fold undersampled measurements

Fig. 12: Reconstructed axial T1 weighted brain images from 8 and 20 fold undersampled measurements in the simulation study, for various regularization parameter settings for different reconstruction methods. λ denotes the TV regularization weight for PLS-TV, n_{iter} denotes the number of iterations for IAGAN-TV (regularization via early stopping), k denotes the dimensionality of the latent subspace for INN Proj. TV, and μ is used to denote the TV regularization weight for IAGAN-TV and INN Proj. TV.

(i.e. the latent subspace dimension k and the TV regularization weight μ) were chosen with the help of a 2D grid search, and the parameters giving the best RMSE value was chosen. Figures 11a, 11b, 12a and 12b show the selection of the regularization parameters. From the 2D grid search for IAGAN-TV and the proposed method, only the 1D line-search images through the optimal parameter point are shown in Figures 11a, 11b, 12a and 12b.

REFERENCES

[1] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[2] A. Barbero and S. Sra, "Modular proximal optimization for multidimensional total-variation regularization," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2232–2313, 2018.

[3] E. Y. Sidky, I. Lorente, J. G. Brankov, and X. Pan, "Do CNNs solve the CT inverse problem?" *arXiv preprint arXiv:2005.10755*, 2020.