Learning-Based Compressive MRI

Baran Gözcü[1], Rabeeh Karimi Mahabadi[1], Yen-Huan Li[1], Efe Ilıcak[2],
Tolga Çukur[2,3], Jonathan Scarlett[4], and Volkan Cevher[1]
[1]Laboratory for Information and Inference Systems (LIONS), EPFL, Switzerland
[2]National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey
[3]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
[4]Department of Computer Science & Department of Mathematics, National University of Singapore, Singapore

In the area of magnetic resonance imaging (MRI), an extensive range of non-linear reconstruction algorithms have been proposed that can be used with general Fourier subsampling patterns. However, the design of these subsampling patterns has typically been considered in isolation from the reconstruction rule and the anatomy under consideration. In this paper, we propose a learning-based framework for optimizing MRI subsampling patterns for a specific reconstruction rule and anatomy, considering both the noiseless and noisy settings. Our learning algorithm has access to a representative set of training signals, and searches for a sampling pattern that performs well on average for the signals in this set. We present a novel parameter-free greedy mask selection method, and show it to be effective for a variety of reconstruction rules and performance metrics. Moreover we also support our numerical findings by providing a rigorous justification of our framework via statistical learning theory.

***Index Terms*—Magnetic resonance imaging, compressive sensing, learning-based subsampling, greedy algorithms**

I. Introduction

Magnetic resonance imaging (MRI) serves as a crucial diagnostic modality for scanning soft tissue in body parts such as the brain, knee, and spinal cord. While early MRI technology could require over an hour of scan time to produce diagnostic-quality images, subsequent advances have led to drastic reductions in the scan time without sacrificing the imaging quality.

The application of MRI has served as a key motivation for compressive sensing (CS), a modern data acquisition technique for sparse signals. The theory and practice of CS for MRI have generally taken very different paths, with the former focusing on sparsity and uniform random sampling of the Fourier space, but the latter dictating the use of *variable-density* subsampling. Common to both viewpoints, however, is the element of *non-linear decoding* via optimization formulations.

In this paper, we propose a *learning-based framework* for compressive MRI that is both theoretically grounded and practical. The premise is to use training signals to *optimize the subsampling specifically for the setup at hand.*

In more detail, we propose a novel greedy algorithm for mask optimization that can be applied to arbitrary reconstruction rules and performance measures. This mask selection algorithm is parameter-free, excluding unavoidable parameters of the reconstruction methods themselves. We use statistical learning theory to justify the core idea of optimizing the empirical performance on training data for the sampling design problem. In addition, we provide numerical evidence that our framework can find good sampling patterns for different performance metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index [1], and for a broad range of decoders, from basis pursuit and total variation to neural networks and BM3D. Since our framework can be applied to arbitrary decoders, we also anticipate that it can benefit *future* decoding rules.

Organization of the paper: In Section II, we introduce the compressive MRI problem and outline the most relevant existing works, as well as summarizing our contributions. In Section III, we introduce our learning-based framework, along with its theoretical justification. In Section IV, we demonstrate the effectiveness of our approach on a variety of data sets, including comparisons to existing approaches. Conclusions are drawn in Section V.

II. Background

A. Signal acquisition and reconstruction

In the *compressive sensing* (CS) problem [2], one seeks to recover a sparse vector via a small number of linear measurements. In the special case of compressive MRI, these measurements take the specific form of subsampled Fourier measurements, described as follows:

$$\mathbf{b} = \mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{\Psi} \in \mathbb{C}^{p\times p}$ is the Fourier transform operator applied to the vectorized image,[1] $\mathbf{P}_\Omega : \mathbb{C}^p \to \mathbb{C}^n$ is a subsampling operator that selects the rows of $\mathbf{\Psi}$ indexed by the set Ω, with $|\Omega| = n$, and $\mathbf{w} \in \mathbb{C}^n$ is possible additive noise. We refer to Ω as the *sampling pattern* or *mask*.

Given the measurements $\mathbf{b}$ (along with knowledge of Ω), a *reconstruction algorithm* (also referred to as the *decoder*) forms an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$. This algorithm is treated as a general function, and is written as follows:

$$\hat{\mathbf{x}} = g(\Omega, \mathbf{b}). \tag{2}$$

A wide variety of decoding techniques have been proposed for compressive MRI; here we present a few of the most widely-used and best-performing techniques, which we will pursue in the numerical experiments in Section IV.

[1]The original image may be 2D or 3D, but we express it in its vectorized form for convenience.

In the general CS problem, decoders based on convex optimization have received considerable attention, both due to their theoretical guarantees and practical performance. In the noiseless setting (i.e., $\mathbf{w} = 0$), a particularly notable choice is *basis pursuit* (BP) [2]:

$$\hat{\mathbf{x}} = \underset{\mathbf{z}\,:\,\mathbf{b}=\mathbf{P}_\Omega\boldsymbol{\Psi}\mathbf{z}}{\arg\min} \|\boldsymbol{\Phi}\mathbf{z}\|_1 \tag{3}$$

where $\boldsymbol{\Phi}$ is a sparsifying operator such as the wavelet or shearlet transform. A similar type of convex optimization formulation that avoids the need for the sparsifying operator is *total variation* (TV) minimization:

$$\hat{\mathbf{x}} = \underset{\mathbf{z}\,:\,\mathbf{b}=\mathbf{P}_\Omega\boldsymbol{\Psi}\mathbf{z}}{\arg\min} \|\mathbf{z}\|_{\mathrm{TV}}, \tag{4}$$

where $\|\mathbf{z}\|_{\mathrm{TV}}$ is the total variation norm.

For the specific application of MRI, heuristic reconstruction algorithms have recently arisen that can outperform methods such as BP and TV, despite their lack of theoretical guarantees. A state-of-the-art reconstruction algorithm was recently proposed in [3] based on the *block matching and 3D filtering (BM3D)* denoising technique [4] that applies principal component analysis (PCA) to patches of the image. At a high level, the algorithm of [3] alternates between denoising using BM3D, and reconstruction using regularized least squares formulations. We refer the reader to [3] for further details, and to Section IV for our numerical results.

Following their enormous success in machine learning applications, *deep neural networks* have also been proposed for MRI reconstruction. We consider the approach of [5], which uses a cascade of convolution neural networks (CNNs) interleaved with data consistency (DC) units. The CNNs serve to perform de-aliasing, and the DC units serve to enforce data consistency in the reconstruction. The deep network is trained by inputing the subsampled signals and treating the full training signal as the desired reconstruction. We refer the reader to [5] for further details, and to Section IV for our numerical results.

Among the extensive existing literature, other relevant works include [6] and [7], where compressive sensing is unified with parallel MRI. In [8], a matrix completion framework is proposed for the parallel MRI setting. In [9], [10], [11] and [12], dictionary learning and faster transform learning methods are shown to provide considerably better quality of reconstructions compared to nonadaptive methods. In [13], [14], [15] and [16], low rank models are used for improved results in dynamic MRI setting, for which dictionary-based approaches are also presented in [17] and [18]. Another notable work in the dynamic setting is [19], which, in a compressive sensing framework, generalizes the previous work that exploits spatiotemporal correlations for improved frame rate [20]. Patient-adaptive methods for dynamic MRI also exist in the literature [21], [22]. A Bayesian approach is taken in [23] and [24] for compressive MRI applications, whereas in [25], Bayesian and dictonary learning approaches are combined.

In [26], a deep convolutional network is used to learn the aliasing artefacts, providing a more accurate reconstruction in the case of uniform sampling. In [27], this approach is applied to the radial acquisition setting. In [28], a deep network is used to train the transformations and parameters present in an regularized objective function. Moreover, in [29], a framework based on generative adversarial networks is applied for an improved compressive MRI performance, whereas in [30], a convolutional network is trained for faster acquisition and reconstruction for dynamic MRI setting.

B. Subsampling pattern design

Generally speaking, the most popular approaches to designing Ω for compressive MRI make use of *random variable-density sampling* according to a non-uniform probability distribution [31]. The random sampling is done in a manner that favors taking more samples at low frequencies. Some examples include variable-density Poisson disk sampling [32], multi-level sampling schemes [33], [34], pseudo-2D random sampling [35], and variable density with continuous and block sampling models [36]–[38].

While such variable-density approaches often perform well, they have notable limitations. First, they typically require parameters to be tuned (e.g., the rate of decay of probability away from the center). Second, it is generally unclear which particular sampling distribution will be most effective for a given decoding rule and anatomy. Finally, the very idea of randomization is questionable, since in practice one would like to design a fixed sampling pattern to use across many subjects.

Recently, alternative design methods have been proposed that make use of fully sampled training data (i.e., training signals). In [39]–[41], the training data is used to construct a sampling distribution, from which the samples are then drawn randomly. In [42], [43], a single training image is used at a time to choose a row to sample, and in [44] the rows are chosen based on a mutual information criterion. Much like the above-mentioned randomized variable-density sampling approaches, these existing adaptive algorithms contain parameters for their mask selection whose tuning is non-trivial. Moreover, to our knowledge, none of these works have provided theoretical justifications of the mask selection method. On the other hand, except for [43], these algorithms do not optimize the sampling pattern for a given general decoder. We achieve this via a parallelizable greedy algorithm that implements the given decoder on multiple training images at each iteration of the algorithm until the desired rate is attained.

A particularly relevant prior work is that of [45], in which we proposed the initial learning-based framework that motivates the present paper. However, the focus in [45] is on a simple linear decoder and the noiseless setting, and the crucial aspects of non-linear decoding and noise were left as open problems.

An alternative approach to optimizing subsampling based on prior information is given in [46]. The idea therein is that if a subject requires multiple similar scans, then the previous scans can be used to adjust both the sampling and the decoding of future scans. This is done using the randomized variable-density approach, with the probabilities adjusted to favor locations where the previous scans had more energy. In [47], a proposed informative random sampling approach

optimizes the sampling of subsequent frames of dynamic MRI data based on previous frames in real time. In [48], a highly undersampled pre-scan is used to learn the energy distribution of the image and design the sampling prior to the main scan.

A recent comparative study [49] showed that the approaches that directly use training data perform better than the purely parametric (e.g., randomized variable-density) methods.

Other subsampling design works include the following: In [50], a generalized Rosetta shaped sampling pattern is used for compressive MRI, and in [51] a random-like trajectory based on higher order chirp sequences is proposed. Radial acquisition designs have been proposed to improve the performance of compressive MRI in the settings of dynamic MRI [52] and phase contrast MRI [53]. In addition, a recent work [54] considered non-Cartesian trajectory design for high resolution MRI imaging at 7T (Tesla).

C. Theory of compressive sensing

The theory of CS has generally moved in very different directions to the above practical approaches. In particular, when it comes to subsampled Fourier measurements, the vast majority of the literature has focused on guarantees for recovering sparse signals with *uniform random sampling* [55], which performs very poorly in practical imaging applications.

A recent work [33] proposed an alternative theory of CS based on *sparsity in levels*, along with variable-density random sampling and BP decoding. As we outline below, we adopt an entirely different approach that avoids making *any* specific structural assumptions, yet can exploit even richer structures beyond sparsity and its variants.

D. Our contributions

In this paper, we propose a novel *learning-based framework* for designing subsampling patterns, based on the idea of directly maximizing the empirical performance on training data. We adopt an entirely different theoretical viewpoint to that of the existing CS literature; rather than placing structural assumptions (e.g., sparsity) on the underlying signal, we simply think of the training and test signals as coming from a common *unknown* distribution. Using connections with statistical learning theory, we adopt a learning method that automatically extracts the structure inherent in the signal, and optimizes Ω specifically for the decoder at hand.

While our framework is suited to general CS scenarios, we focus on the application of MRI, in which we observe several advantages over the above existing approaches:

- While our previous work [45] exclusively considered a simple linear decoder, in this paper we consider targeted optimization for general non-linear decoders;
- We present a non-trivial extension of our theory and methodology to the noisy setting, whereas [45] only considered the noiseless case;
- We directly optimize for the performance measure at hand (e.g. PSNR), as opposed to less direct measures such as mutual information. Similarly, our framework permits the direct optimization of bottom-line costs (e.g., acquisition time), rather than auxiliary cost measures (e.g., number of samples);
- We can directly incorporate practical sampling constraints, such as the requirement of sampling entire rows and/or columns rather than arbitrary patterns;
- Parameter tuning is not required;
- Our learning algorithm is highly parallelizable, rendering it feasible even when the dimension and/or the number of training images is large;
- We demonstrate the effectiveness of our approach on several real-world data sets, performing favorably against existing methods.

III. Learning-based Framework

A. Overview

Our learning-based framework is outlined as follows:

- We have access to a set of fully-sampled training signals $\mathbf{x}_1, \dots, \mathbf{x}_m$ that are assumed to be representative of the unknown signal of interest $\mathbf{x}$.
- We assume that the decoder (2) is given. This decoder is allowed to be *arbitrary*, meaning that our framework can be used alongside general existing reconstruction methods, and potentially also future methods.
- For any subsampling pattern Ω, we can consider its empirical average performance on the training signals:

$$\frac{1}{m}\sum_{j=1}^{m}\eta_{\Omega}(\mathbf{x}_j), \tag{5}$$

  where $\eta_{\Omega}(\mathbf{x})$ is a performance measure (e.g., PSNR) associated with the signal $\mathbf{x}$ and its reconstruction when the sampling pattern is Ω. If $\mathbf{x}_1, \dots, \mathbf{x}_m$ are similar to $\mathbf{x}$, we should expect that any Ω such that (5) is high will also perform well on $\mathbf{x}$.
- While maximizing (5) is computationally challenging in general, we can use any preferred method to seek an *approximate* maximizer. We will pay particular attention to a *greedy algorithm*, which is parameter-free and satisfies a useful nestedness property.

We proceed by describing these points in more detail. For convenience, we initially consider the noiseless setting,

$$\mathbf{b} = \mathbf{P}_{\Omega}\mathbf{\Psi}\mathbf{x}, \tag{6}$$

and then turn to the noisy setting in Section III-F.

B. Preliminaries

Our broad goal is to determine a good subsampling pattern $\Omega \subseteq \{1, \dots, p\}$ for compressive MRI. To perform this task, we assume that we have access to a set of *training signals* $\mathbf{x}_1, \dots, \mathbf{x}_m$, with $\mathbf{x}_j \in \mathbb{C}^p$. The idea is that if an unseen signal has similar properties to the training signals, then we should expect the learned subsampling patterns to generalize well.

In addition to the reconstruction rule g of the form (2), the learning procedure has knowledge of a *performance measure*, which we would like to make as high as possible on the unseen signal. We focus primarily on PSNR in our experimental section, while also considering the SSIM index.

For implementation reasons, one may wish to restrict the sampling patterns in some way, e.g., to contain only horizontal and/or vertical lines. To account for such constraints, we assume that there exists a set $\mathcal{S}$ of subsets of $\{1,\dots,p\}$ such that the final sampling pattern must take the form

$$\Omega = \bigcup_{j=1}^{\ell} S_j, \quad S_j \in \mathcal{S} \tag{7}$$

for some $\ell > 0$. If $\mathcal{S} = \{\{1\},\dots,\{p\}\}$, then we recover the setting of [45] where the subsampling pattern may be arbitrary. However, arbitrary sampling patterns are not always feasible; for instance, masks consisting of only horizontal and/or vertical lines are often considered much more suited to practical implementation, and hence, it may be of interest to restrict $\mathcal{S}$ accordingly.

Finally, we assume there exists a *cost function* $c(\Omega) \geq 0$ associated with each subsampling pattern, and that the final cost must satisfy

$$c(\Omega) \leq \Gamma \tag{8}$$

for some $\Gamma > 0$. We will focus primarily on the case that the cost is the total number of indices in Ω (i.e., we are placing a constraint on the sampling rate), but in practical scenarios one may wish to consider the ultimate underlying cost, such as the scan time. We assume that $c(\cdot)$ is monotone with respect to inclusion, i.e., if $\Omega_1 \subseteq \Omega_2$ then $c(\Omega_1) \leq c(\Omega_2)$.

C. Theoretical motivation via statistical learning theory

Before describing our main algorithm, we present a theoretical motivation for our learning-based framework. To do so, we think of the underlying signal of interest $\mathbf{x}$ as coming from a probability distribution P. Under any such distribution, we can write down the indices with the best average performance:

$$\Omega^* = \arg\max_{\Omega\in\mathcal{A}} \mathbb{E}_P\big[\eta_\Omega(\mathbf{x})\big], \tag{9}$$

where $\mathcal{A}$ is the set of feasible Ω according to $c(\cdot)$, Γ, and $\mathcal{S}$, and we define

$$\eta_\Omega(\mathbf{x}) = \eta(\mathbf{x},\hat{\mathbf{x}}) \tag{10}$$

with $\hat{\mathbf{x}} = g(\Omega,\mathbf{b})$ and $\mathbf{b} = \mathbf{P}_\Omega\mathbf{\Psi}\mathbf{x}$.

Unfortunately, the rule in (9) is not feasible in practice, since one cannot expect to know P (e.g., one cannot reasonably form an accurate probability distribution that describes a brain image). However, if the training signals $\mathbf{x}_1,\dots,\mathbf{x}_m$ are also independently drawn from P, then there is hope that the *empirical average* is a good approximation of the true average. This leads to the following selection rule:

$$\hat{\Omega} = \arg\max_{\Omega\in\mathcal{A}} \frac{1}{m}\sum_{j=1}^{m} \eta_\Omega(\mathbf{x}_j). \tag{11}$$

The rule (11) is an instance of *empirical risk minimization* in statistical learning theory. While finding the exact maximum can still be computationally hard, this viewpoint will nevertheless dictate that we should seek indices $\Omega \in \mathcal{A}$ such that $\frac{1}{m}\sum_{j=1}^{m}\eta_\Omega(\mathbf{x}_j)$ is high.

To see this more formally, we consider the following question: If we find a set of indices $\Omega \in \mathcal{A}$ with a good empirical performance $\frac{1}{m}\sum_{j=1}^{m}\eta_\Omega(\mathbf{x}_j)$, does it also provide good performance $\mathbb{E}[\eta_\Omega(\mathbf{x})]$ on an unseen signal $\mathbf{x}$? The following proposition answers this question in the affirmative using statistical learning theory.

Proposition 1. *Consider the above setup with a performance measure normalized so that $\eta(\mathbf{x},\hat{\mathbf{x}}) \in [0,1]$.[2] For any $\delta \in (0,1)$, with probability at least $1-\delta$ (with respect to the randomness of $\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_m$), it holds that*

$$\left|\frac{1}{m}\sum_{j=1}^{m}\eta_\Omega(\mathbf{x}_j) - \mathbb{E}_P\left[\eta_\Omega(\mathbf{x})\right]\right| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2\,|\mathcal{A}|}{\delta}\right)},$$

simultaneously for all $\Omega \in \mathcal{A}$.

The proof is given in the appendix. We see that as long as m is sufficiently large compared to $|\mathcal{A}|$, the average performance attained by any given $\Omega \in \mathcal{A}$ on the training data is an accurate estimate of the true performance. This guarantee is with respect to the *worst case*, regarding all possible probability distributions P; the actual performance could exceed this guarantee in practice.

D. Greedy algorithm

While finding the exact maximizer in (11) is challenging in general, we can seek to efficiently find an *approximate* solution. There are several possible ways to do this, and Proposition 1 reveals that regardless of how we come across a mask with better empirical performance, we should favor it. In this subsection, we present a simple *greedy* approach, which is parameter-free in the sense that no parameter tuning is needed for the mask selection process once the decoder is given (though the decoder itself may still have tunable parameters). The greedy approach also exhibits a useful nestedness property (described below).

At each iteration, the greedy procedure runs the decoder g with each element of $\mathcal{S}$ that is not yet included in the mask, and adds the subset $S \in \mathcal{S}$ that increases the performance function most on average over the training images, normalized by the cost. The algorithm stops when it is no longer possible to add new subsets from $\mathcal{S}$ without violating the cost constraint. The details are given in Algorithm 1.

An important feature of this method is the *nestedness* property that allows one to immediately adapt for different costs Γ (e.g., different sampling rates). Specifically, one can record the order in which the elements are included in Ω during the mask optimization for a high cost, and use this to infer the mask corresponding to lower costs, or to use as a starting point for higher costs. Note that this is not possible for most parametric methods, where changing the sampling rate requires one to redo the parameter tuning.

We briefly note that alternative greedy methods could easily be used. For instance:

[2] As a concrete example, suppose we are interested in the squared error $\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2$. If the input is normalized to $\|\mathbf{x}\|_2^2 = 1$, then it can be shown that any estimate $\hat{\mathbf{x}}$ only improves if it is scaled down such that $\|\hat{\mathbf{x}}\|_2^2 \leq 1$. It then follows easily that $\eta(\mathbf{x},\hat{\mathbf{x}}) = 1 - \frac{1}{4}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2$ always lies in $[0,1]$.

Algorithm 1 Greedy mask optimization

Input: Training data $\mathbf{x}_1, \dots, \mathbf{x}_m$, reconstruction rule g, sampling subsets $\mathcal{S}$, cost function c, maximum cost Γ
Output: Sampling pattern Ω

$\Omega \leftarrow \emptyset$;
while $c(\Omega) \leq \Gamma$ **do**
  for $S \in \mathcal{S}$ such that $c(\Omega \cup S) \leq \Gamma$ **do**
    $\Omega' = \Omega \cup S$
    For each j, set $\mathbf{b}_j \leftarrow \mathbf{P}_{\Omega'} \mathbf{\Psi} \mathbf{x}_j$, $\hat{\mathbf{x}}_j \leftarrow g(\Omega', \mathbf{b}_j)$
    $\eta(\Omega') \leftarrow \frac{1}{m} \sum_{j=1}^{m} \eta(\mathbf{x}_j, \hat{\mathbf{x}}_j)$
  $\Omega \leftarrow \Omega \cup S^*$, where

$$S^* = \underset{S \,:\, c(\Omega \cup S) \leq \Gamma}{\arg\max} \frac{\eta(\Omega \cup S) - \eta(\Omega)}{c(\Omega \cup S) - c(\Omega)}$$

return Ω

- One could start with $\Omega = \{1, \dots, p\}$ (i.e., sampling the entire Fourier space) and then *remove* samples until a feasible pattern is attained;
- One could adopt a hybrid approach in which samples are both added and removed iteratively until some convergence condition is met.

In our experiments, however, we focus in the procedure in Algorithm 1, which we found to work well.

In another related work [43], an iterative approach is taken in which only a single nonlinear reconstruction is implemented in each iteration of mask selection, starting with an initial mask whereas we run separate reconstructions for each candidate to be added to the sampling pattern, starting with the empty set. Moreover, [43] makes use of several parameters, such as the number of the regions with higher errors to which the samples are moved iteratively, the size of these regions, the power of the polynomial used for a weighting function, etc., which need to be tuned for each experiment. Our greedy algorithm has the advantage of avoiding such heuristics and additional parameters. While the proposed algorithm requires a larger number of computations for mask selection, these computations can be easily parallelized and performed efficiently.

E. Parametric approach with learning

An alternative approach is to generate a number of candidate masks $\Omega_1, \dots, \Omega_L$ using one or more parametric variable-density methods (possibly with a variety of different choices of parameters), and then to apply the learning-based idea to these candidate masks: Choose the one with the best empirical performance on the training set. While similar ideas have already been used when performing parameter sweeps in existing works (e.g., see [39]), our framework provides a more formal justification to why the empirical performance is the correct quantity to optimize. The details are given in Algorithm 2, where we assume that all candidate masks are feasible according to the sampling subsets $\mathcal{S}$ and cost function c.

F. Noisy setting

So far, we have considered the case that both the acquired signal $\mathbf{b}$ and the training signals $\mathbf{x}_1, \dots, \mathbf{x}_m$ are noiseless. In this subsection, we consider a noisy variant of our setting: The acquired signal is given by

$$\mathbf{b} = \mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x} + \mathbf{w} \tag{12}$$

for some noise term $\mathbf{w} \in \mathbb{C}^p$, and the learning algorithm does not have access to the exact training signals $\mathbf{x}_1, \dots, \mathbf{x}_m$, but instead to noisy versions $\mathbf{z}_1, \dots, \mathbf{z}_m$, where

$$\mathbf{z}_j = \mathbf{x}_j + \mathbf{v}_j, \quad j = 1, \dots, m \tag{13}$$

with $\mathbf{v}_j$ representing the noise.

We observe that the selection rule in (11) can no longer be used, since the learning algorithm does not have direct access to $\mathbf{x}_1, \dots, \mathbf{x}_m$. The simplest alternative is to substitute the noisy versions of the signals and use (11) with $\mathbf{z}_j$ in place of $\mathbf{x}_j$. It turns out, however, that we can do better if we have access to a *denoiser* $\xi(\mathbf{z})$ that reduces the noise level. Specifically, suppose that

$$\xi(\mathbf{z}) = \mathbf{x}_j + \widetilde{\mathbf{v}}_j \tag{14}$$

for some reduced noise $\widetilde{\mathbf{v}}_j$ such that $\mathbb{E}[\|\widetilde{\mathbf{v}}_j\|] \leq \mathbb{E}[\|\mathbf{v}_j\|]$. We then propose the selection rule

$$\hat{\Omega} = \underset{\Omega \in \mathcal{A}}{\arg\max} \frac{1}{m} \sum_{j=1}^{m} \eta(\mathbf{x}_j + \widetilde{\mathbf{v}}_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega \mathbf{\Psi}(\mathbf{x}_j + \mathbf{v}_j)), \tag{15}$$

where $\hat{\mathbf{x}}(\mathbf{b})$ denotes the decoder applied to $\mathbf{b}$. Note that we still use the *noisy* training signal in the choice of $\mathbf{b}$; by doing so, we are *learning how to denoise*, which is necessary because the unseen test signal is itself noisy as per (12).

To understand how well the above rule generalized to unseen signals, we would like to compare the empirical performance on the right-hand side of (15) to the true average performance on an unseen signal, defined as

$$\overline{\eta}_{\text{noisy}}(\Omega) = \mathbb{E}\big[\eta(\mathbf{x}, \hat{\mathbf{x}})\big] \tag{16}$$

with $\hat{\mathbf{x}} = g(\Omega, \mathbf{b})$ and $\mathbf{b} = \mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x} + \mathbf{w}$. The following proposition quantifies this comparison.

Proposition 2. *Consider the above noisy setup with* $\mathbf{w}$ *and* $\{\mathbf{v}_j\}_{j=1}^m$ *having independent Gaussian entries of the same variance, and a performance measure* $\eta(\mathbf{x}, \hat{\mathbf{x}}) \in [0, 1]$ *that satisfies the continuity assumption* $|\eta(\mathbf{x}, \hat{\mathbf{x}}) - \eta(\mathbf{x}', \hat{\mathbf{x}})| \leq L\|\mathbf{x} - \mathbf{x}'\|_2$ *for all* $\mathbf{x}, \mathbf{x}'$ *and some* $L > 0$. *For any* $\delta \in (0, 1)$,

Algorithm 2 Choosing from a set of candidate masks

Input: Training data $\mathbf{x}_1, \dots, \mathbf{x}_m$, reconstruction rule g, candidate masks $\Omega_1, \dots, \Omega_L$
Output: Sampling pattern Ω

for $\ell = 1, \dots, L$ **do**
  For each j, set $\mathbf{b}_j \leftarrow \mathbf{P}_{\Omega_\ell} \mathbf{\Psi} \mathbf{x}_j$, $\hat{\mathbf{x}}_j \leftarrow g(\Omega_\ell, \mathbf{b}_j)$
  $\eta_\ell \leftarrow \frac{1}{m} \sum_{j=1}^{m} \eta(\mathbf{x}_j, \hat{\mathbf{x}}_j)$
$\Omega \leftarrow \Omega_{\ell^*}$, where $\ell^* = \arg\max_{\ell=1,\dots,L} \eta_\ell$
return Ω

with probability at least $1-\delta$ (with respect to the randomness of the noisy training signals), it holds that

$$\left| \frac{1}{m} \sum_{j=1}^{m} \eta(\mathbf{x}_j + \widetilde{\mathbf{v}}_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega \mathbf{\Psi}(\mathbf{x}_j + \mathbf{v}_j)) - \overline{\eta}_{\text{noisy}}(\Omega) \right| \leq L\mathbb{E}[\|\widetilde{\mathbf{v}}\|_2] + \sqrt{\frac{1}{2m} \log\left(\frac{2\,|\mathcal{A}|}{\delta}\right)}, \quad (17)$$

simultaneously for all $\Omega \in \mathcal{A}$, where $\widetilde{\mathbf{v}}_j = \xi(\mathbf{v}_j)$ is the effective noise remaining in the j-th denoised training signal, and $\widetilde{\mathbf{v}}$ has the same distribution as any given $\widetilde{\mathbf{v}}_j$.

The proof is given in the appendix. We observe that the second term coincides with that of the noiseless case in Proposition 1, whereas the first term represents the additional error due to the residual noise after denoising. It is straightforward to show that such a term is unavoidable in general.[3]

Hence, along with the fact that more training signals leads to better generalization, Proposition 2 reveals the intuitive fact that *the ability to better denoise the training signals leads to better generalization*. In particular, if we can do *perfect denoising* (i.e., $\|\widetilde{\mathbf{v}}_j\| = 0$) then we get the same generalization error as the noiseless case.

In Algorithm 3, we provide the learning-based procedure with an arbitrary denoising function ξ. Note that if we choose the identity function $\xi(\mathbf{z}) = \mathbf{z}$, then we reduce to the case where no denoising is done.

Algorithm 3 Learning-based mask selection with denoising

Input: Noisy training data $\mathbf{z}_1, \ldots, \mathbf{z}_m$, reconstruction rule g, denoising algorithm $\xi(\mathbf{z})$, and either the triplet $(\mathcal{S}, c, \Gamma)$ or candidate masks $\Omega_1, \ldots, \Omega_L$

Output: Sampling pattern Ω

1: $\mathbf{x}'_j \leftarrow \xi(\mathbf{z}_j)$ for $j = 1, \ldots, m$
2: Select Ω using Algorithm 1 or 2 with $\eta(\mathbf{x}'_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{z}_j))$ replacing $\eta(\mathbf{x}_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}_j))$ throughout.
3: **return** Ω

IV. Numerical Experiments

In this section, we provide numerical experiments demonstrating that our learning-based framework provides high-performing sampling patterns for a diverse range of reconstruction algorithms. Our simulation code and data are publicly available online. [4]

A. Implementation details

Reconstruction rules. We consider the decoders described in Section II-A, which we refer to as BP, TV, BM3D, and NN (i.e., neural network). For BP in (3), we let the sparsifying operator $\mathbf{\Phi}$ be the shearlet transform [56], and for both BP and TV, we implement the minimization using NESTA [57] , for which we set the maximum number iterations to 20000, the denoising parameter to $\epsilon = 0$, the tolerance value and the smoothing parameter to $\mu = 10^{-5}$, and the number of continuation steps to $T = 1$.

For BM3D, we use the code available in [58]. We take the observation fidelity parameter $\alpha = 0$, the number of outer iterations $\mathcal{J} = 20$ and the regularization parameters as $\lambda_{\max} = 200$ and $\lambda_{\min} = 0.01$. We also use a varying number of inner iterations between 1 and 10 as described in [3].

For the NN decoder, we use the network structure from [5], only slightly modifying certain parameters. We choose depth of the architecture as $n_d = 3$ and depth of the cascade as $n_c = 5$. We set the mini-batch size for training to 20. We use the same training signals for learning indices and tuning the network weights. Since it is difficult to optimize these jointly, we perform alternating optimization: Initialize the weights, and then alternate between learning indices with fixed weights, and learning weights with fixed indices. We perform up to three iterations of this procedure, which we found to be sufficient for convergence.

As was done in [5], we initialize the network weights using the initialization of He *et al.* [59], and perform network weight optimization using the Adam algorithm [60] with step size $\alpha = 10^{-2}$ and decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Moreover, we apply an additional ℓ_2 weight regularization penalty of 10^{-6}. Each time we train the network, we run the training for 7000 epochs (i.e., passes over the training data). We use the Python implementation available in [5].

In principle, it may sometimes be preferable to change the reconstruction parameters as the greedy algorithm adds indices and increases the current sampling rate. However, we did not find such an approach to provide further benefit in the present setting, so here we stick to the above approach where the reconstruction parameters remain fixed.

Mask selection methods. In addition to the greedy method in Algorithm 1, we consider parametric randomized variable-density methods with learning-based optimization according to Algorithm 2; the details are provided in the relevant subsections below. Moreover, we consider the following two baselines from the existing literature:

- *(Coherence-based)* We consider the parametric approach of [31] with parameters specifying (i) the size of a fully-sampled region at low frequencies; and (ii) the polynomial rate of decay of sampling at higher frequencies. As suggested in [31], we choose the parameters to optimize an *incoherence function*, meaning that no training data is used. The minimization is done using Monte Carlo methods, and we do this using the code used in [31] available online.
- *(Single-image)* We consider the approach of [41] in which only a single training image is used. Specifically, this image determines a probability density function where the probability is proportional to energy, and then the samples are randomly selected by drawing from this distribution.

Data sets. The MRI data used in the following subsections was acquired on a 3T MRI system (Magnetom Trio Scanner, Erlangen, Germany). The protocols were approved by the

[3] For instance, to give an example where the generalization error must contain the $\mathbb{E}[\|\mathbf{v}\|_2]$ term, it suffices to consider the ℓ_2-error in the trivial case that $\mathbf{x} = \mathbf{x}_1 = \ldots = \mathbf{x}_m$ with probability one, and $\hat{\mathbf{x}}$ also outputs the same deterministic signal.

[4] https://lions.epfl.ch/lb-csmri

local ethics committee, and all subjects gave written informed consent.

The data set used in the first three experiments (subsections) below consists of 2D T1-weighted brain scans of seven healthy subjects, which were scanned with a FLASH pulse sequence and a 12-channel receive-only head coil. In our experiments, we use 20 slices of sizes 256×256 from five such subjects (two for training, three for testing). Data from individual coils was processed via a complex linear combination, where coil sensitivities were estimated from an 8×8 central calibration region of k-space [61]. The acquisition used a field of view (FOV) of 220 × 220 mm^2 and a resolution of 0.9 × 0.7 mm^2. The slice thickness was 4.0 mm. The imaging protocol comprised a flip angle of 70°, a TR/TE of 250.0/2.46 ms, with a scan time of 2 minutes and 10 seconds.

The data set used in subsection E below consists of angiographic brain scans of five healthy subjects acquired with 12-channel receive-only head coil and 20 slices from each are used in our experiments (two subjects for training, three subjects for testing). The size of the slices is 256×256. A 3D TOF sequence was used with FOV of 204×204×51 mm^3, 0.8×0.8×0.8 mm^3 resolution, flip angle of 18°, magnetization-transfer contrast, a TR/TE of 47/4.6 ms, and a scan time of 16 min 25 sec.

B. Comparison to baselines

We first compare to the above-mentioned baselines for a single specific decoder, namely, BP. We use a conventional method of sampling in which readouts are performed as lines at different *phase encodes*, corresponding to a horizontal line in Fourier space. Hence, our subsampling masks consist of only full horizontal lines, and we let $\mathcal{S}$ in Section III-B be the set of all horizontal lines accordingly.

We use our greedy algorithm to find a subset of such lines at a given budget on the total number of samples (or equivalently, the sampling rate). From the data of the five subjects with 20 slices each, we take the first 2 subjects (40 slices total) as training data. Once the masks are obtained, we implement the reconstructions on the remaining 3 subjects (60 slices total). As seen in Figure 1, the learning-based approach outperforms the baselines across all sampling rates shown.

C. Cross-performances of decoders

Next continuing in the same setting as the previous subsection, we compare all four decoders (TV, BP, BM3D, and NN), and evaluate how a mask optimized for one decoder performs when applied to a different decoder. We refer to these as *cross-performances*, and the results are shown in Table I. Here we report both the PSNR (top) and SSIM values (bottom), but the training optimizes only the PSNR; see Section IV-E for training with respect to the SSIM.

Once again, the learning-based approach outperforms the baselines by approximately 2.5-3.5 dB for all decoders considered. We observe that the greedy method always finds the best performing mask for the given reconstruction algorithms that we use. The performance drop is typically small when the masks optimized for other decoders are used. In Figure 2, we further illustrate these observations with a single slice from the test data, showing the masks and reconstructions along with their PSNR and SSIM values. In this figure, we also compare to a pure low-pass mask given in the bottom row. It can be seen that the greedy masks outperform the the low pass mask as well, in terms of PSNR, SSIM and also visual quality, as they offer sharper images with less aliasing artefacts by balancing between low and high frequency components. On the other hand, as can be seen from the zoomed-in regions, the pure low-pass mask introduces strong blurring, whereas the other baseline masks (coherence-based and single image) cause highly visible aliasing due to suboptimal sampling across low to intermediate frequencies.

Computation times for the greedy mask optimization on a parallel computing cluster depend strongly on the reconstruction algorithm in use, and are as follows for a mask of 25% sampling rate using MATLAB's Parallel Computing Toolbox

TABLE I
PSNR AND SSIM PERFORMANCES AVERAGED ON 60 TEST SLICES AT 25% SUBSAMPLING RATE. THE ENTRIES WHERE THE LEARNING IS MATCHED TO THE DECODER AND PERFORMANCE MEASURE ARE SHOWN IN BOLD.

Mask \ Decoder	TV	BP	BM3D	NN
Coherence-based	30.76	31.48	30.04	32.02
Single-image	32.79	33.32	32.42	33.67
TV-greedy	**34.84**	36.08	35.95	36.04
BP-greedy	34.76	**36.16**	36.11	36.17
BM3D-greedy	34.77	36.04	**36.19**	35.92
NN-greedy	34.81	36.05	36.16	**36.36**
Low Pass	31.96	32.41	32.59	32.59

Mask \ Decoder	TV	BP	BM3D	NN
Coherence-based	0.832	0.85	0.822	0.798
Single-image	0.876	0.889	0.879	0.854
TV-greedy	0.907	0.922	0.921	0.869
BP-greedy	0.906	0.923	0.921	0.859
BM3D-greedy	0.906	0.922	0.922	0.909
NN-greedy	0.907	0.923	0.923	0.925
Low Pass	0.876	0.888	0.893	0.893

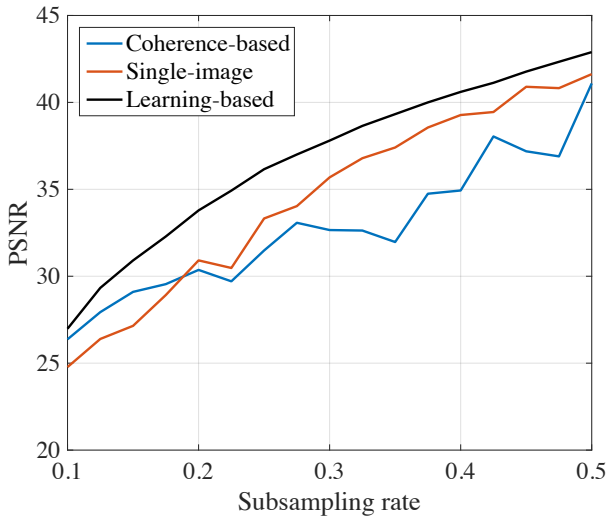

Fig. 1. PSNR as a function of subsampling rates with BP reconstruction.

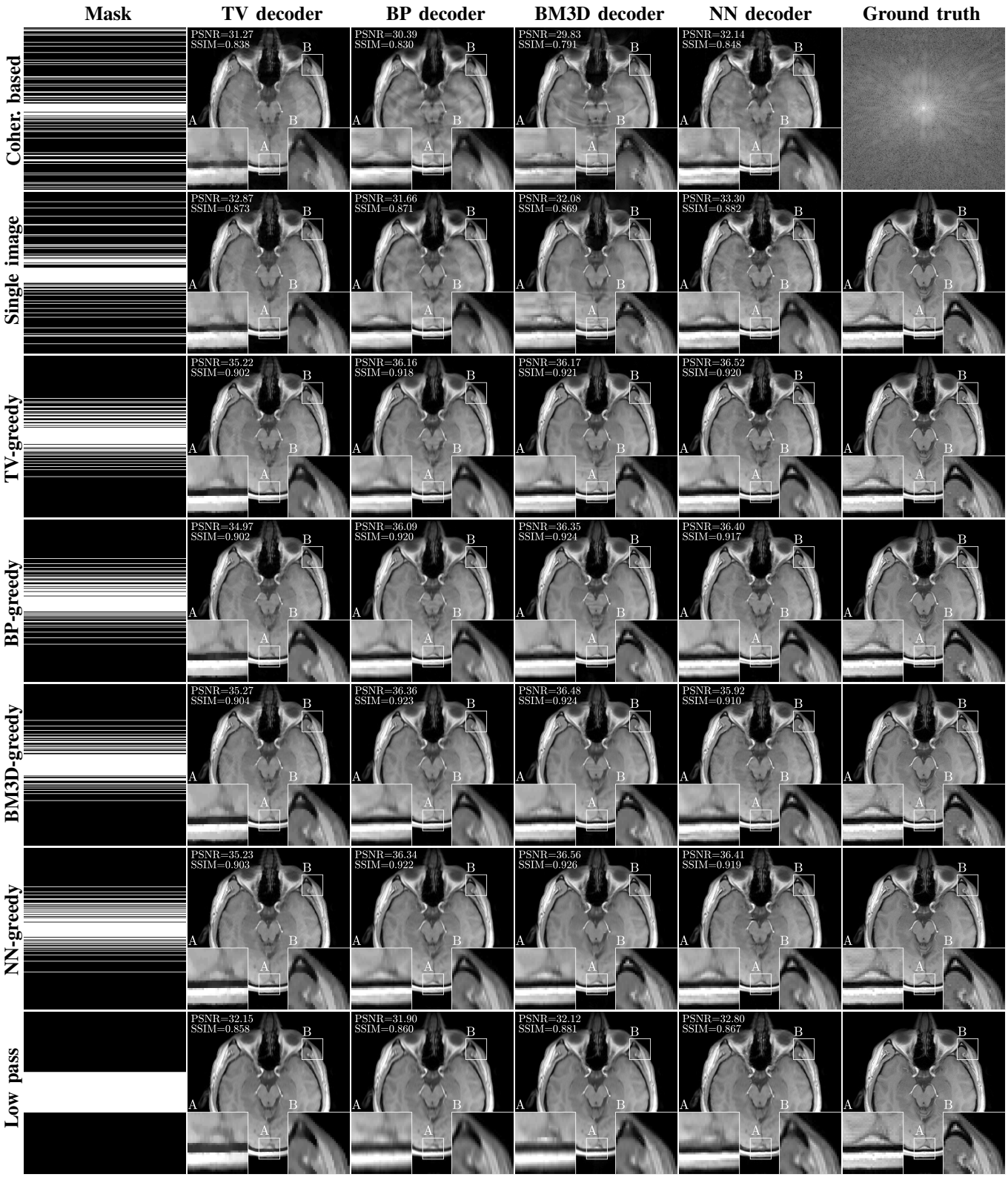

Fig. 2. MRI reconstruction example for the test subject 2, slice 4 at 25% subsampling rate. Sampling masks consist of horizontal lines (phase encodes) and are obtained by the baseline methods [31], [41] and the greedy method proposed in Algorithm 1 where PSNR is used as the performance measure. PSNR (in dB) and SSIM values are shown on the images. The last row shows the performance of purely low-pass mask with different decoders. We put the ground truth into the each row of the last column for the ease of visual comparison, except for the first row, where we present the k-space of the ground truth image in log-scale.

with 256 CPU nodes: (TV) 2 hours and 41 minutes; (BP with shearlets) 3 hours and 23 minutes; (BM3D) 5 hours and 24 minutes. The coherence-based algorithm takes 10 seconds on 256 nodes. The single-image based adaptive algorithm is

TABLE II
PSNR AND SSIM PERFORMANCES AT 25% SUBSAMPLING RATE AVERAGED ON 60 TEST SLICES.

Mask \ Rule	TV-PSNR	TV-SSIM	BP-PSNR	BP-SSIM
parametric	35.02	0.913	36.00	0.926
greedy	35.89	0.927	37.37	0.942

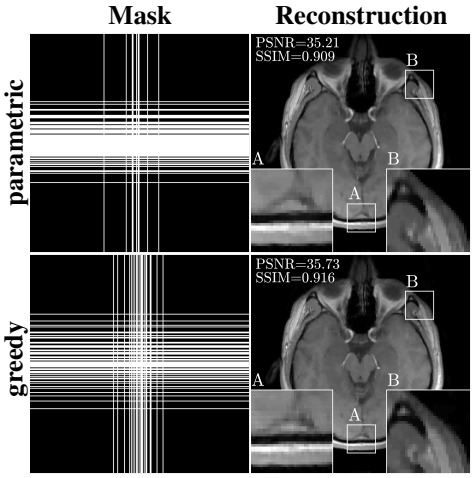

Fig. 3. Masks obtained and example reconstructions under TV decoding at 25% sampling rate, for the parametric method of [35] combined with Algorithm 2, and the greedy method given in Algorithm 1. Both horizontal and vertical lines are permitted. The reconstruction shown is for subject 2, slice 4.

quite fast, running in 2 seconds on a single node. For the NN decoder, the greedy algorithm takes 2 hours and 19 minutes on 40 GPU nodes using multiprocessing package of Python. Note that these computations for mask selection are carried out *offline*, and therefore, we contend that the longer computation time for the greedy mask selection should not be considered a critical issue.

D. Comparison of greedy and parametric methods

We now perform an experiment comparing the greedy approach (Algorithm 1) and the parametric approach with learning (Algorithm 2). In contrast with the previous experiments, we consider measurements in the 2D Fourier space along both horizontal and vertical lines. As described in [35], this is done via a pulse sequence program that switches between *phase encoding* and *frequency encoding*, and can provide improvements over the approach of using only horizontal lines.

We tune the parameters of [35] on the training data using Algorithm 2. The first two of the three parameters are d_x and d_y, which are the sizes of the fully sampled central regions in horizontal and vertical directions. We sweep this for $d_x, d_y \in \{2, 4, \ldots, d_{\max}\}$ where $d_{\max}$ is maximum feasible fully sampled region size for a given subsampling rate. The last parameter D is the degree of the polynomial that defines the probability distribution function from which random masks are drawn. We sweep over $D \in \{1, 3, 5, \ldots, 13\}$. We then randomly draw 5 masks for each choice of parameters, and we use the mask that gives the best average PSNR on the training data, as per Algorithm 2.

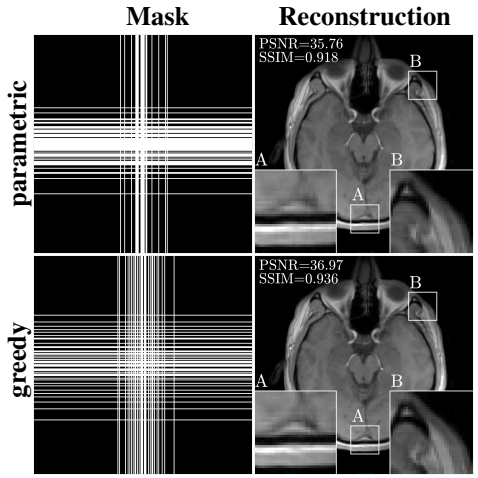

Fig. 4. Masks obtained and example reconstructions under BP decoding at 25% sampling rate, for the parametric method of [35] combined with Algorithm 2, and the greedy method given in Algorithm 1. Both horizontal and vertical lines are permitted. The reconstruction shown is for subject 2, slice 4.

As seen in Table II, the greedy approach outperforms the parametric approach for both the TV and BP reconstruction algorithms. Interestingly, the masks obtained are also visually rather different (*cf.*, Figures 3 and 4, which also show the reconstructions for a single slice), with the greedy masks being more "spread" rather than taking a continuum of rows at low frequencies. It can be also noticed that both methods choose more horizontal lines than vertical lines, due to the fact the the energy in k-space is distributed relatively more broadly across the horizontal direction, as can be seen on the top-right corner of Figure 2.

E. Cross-performances of performance measures

In the previous experiments, we focused on the PSNR performance measure. Here we show that considering different measures can lead to different optimized masks, and that it is important to learn a pattern targeted to the correct performance measure. Specifically, we consider both the PSNR and the structural similarity index (SSIM) [1]. Also different from the previous experiments, we use the data set of angiographic brain scans instead of T1-weighted scans (see Section IV-A for details). We return to the method of taking horizontal lines only in the sampling pattern.

Table III gives the PSNR and SSIM performances for the TV and BP decoders, under the masks obtained via the greedy algorithm (*cf.*, Algorithm 1) with the two different performance measures and the decoders at 30% sampling rate. These results highlight the fact that certain decoders are often better suited to certain performance measures. Here, TV is suited to the PSNR measure, as both tend to prefer concentrating the sampling pattern at low frequencies, whereas BP is better suited to SSIM, with both preferring a relatively higher proportion of high frequencies. Note also that in some columns, the performance is not highest on the rows where the training is matched to the decoder and the performance measure (shown in bold), but slightly lower than the highest

TABLE III
RECONSTRUCTION PERFORMANCES AT 30% SUBSAMPLING RATE AVERAGED OVER 60 ANGIO TEST SLICES. THE CASES THAT THE TRAINING IS MATCHED TO THE PERFORMANCE MEASURE AND DECODER ARE HIGHLIGHTED IN BOLD.

Decoder	TV		BP	
Metric / Mask	SSIM	PSNR	SSIM	PSNR
Coherence-based	0.738	29.82	0.698	27.85
Single-Image	0.766	30.72	0.744	28.85
SSIM-greedy-TV	**0.795**	31.90	0.779	30.50
PSNR-greedy-TV	0.799	**32.74**	0.806	32.97
SSIM-greedy-BP	0.797	32.72	**0.806**	33.08
PSNR-greedy-BP	0.795	32.54	0.804	**32.85**
Low Pass	0.771	31.70	0.777	31.93

values, which is most likely either due to limited training data or the suboptimality of the greedy algorithm.

These observations are further illustrated in Figure 5 (only for TV decoder and its masks due to space constraints), where we show the optimized masks, the reconstructions on a single slice and as the maximum intensity projection (MIP) of the volume this slice belongs to [62]. We see in particular that the two masks are somewhat different, with that for the PSNR containing more gaps at higher frequencies and fewer gaps at lower frequencies. We also observe that compared to the data used in the previous subsections, the angiographic data used in this experiment is more concentrated at the center of the $k-$space. The greedy algorithm is able to adapt to this change and obtain masks that have more lower frequencies.

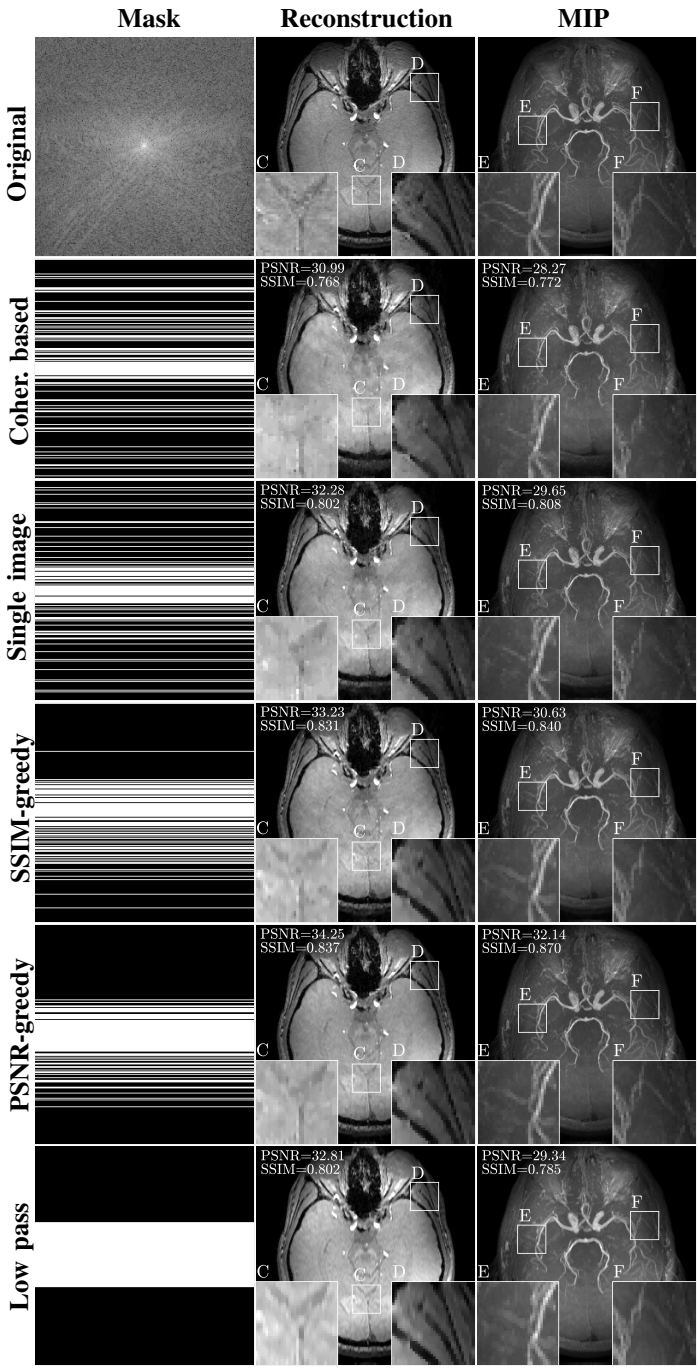

Fig. 5. Masks obtained, example reconstructions, and MIP views of a volume under TV decoding at 30% sampling rate. The mask in the fourth row is obtained using the SSIM as the performance measure in Algorithm 1, and the following mask is obtained using the PSNR. We also present the performances of the coherence-based [31] and single-image based [41] masks. The last row shows the low-pass mask performance. The reconstruction shown is for subject 1, slice 15 in the middle column, and for the MIP of the whole brain in the last column. In the first row, we present the ground truth as a single slice and as MIP; these are used as references when computing the errors.

F. Experiments with additive noise

The data we used in the previous subsections has very low levels of noise. In order to test the validity of our claims in the noisy setting, we add bivariate circularly symmetric complex random Gaussian noise to our normalized complex images, with a noise standard deviation of $\sigma = 3 \times 10^{-4}$ for both the real and imaginary components. Since the ground truth images are normalized, this noise level gives an average signal-to-noise ratio (SNR) of 25.68 dB. We set the denoising parameter of NESTA to $\epsilon = 10$ for TV minimization and to $\epsilon = 1.1$ for BP with shearlets which work well with the various masks and images used in this section. In Algorithm 1, we measure the error at each iteration with respect to denoised image that is obtained using BM3D denoising algorithm [4]. Note that the ground truth should not be used in the learning algorithm, since it is unknown in practice. On the other hand, in the testing part, we compute the errors with respect to the ground truth images.

As can be seen from Figure 6 and Table IV, the greedy algorithm is still capable of finding a better mask compared to the other baseline masks. Therefore, in this example, our approach is robust with respect to noise. Note that we train with respect to the PSNR, but also report the SSIM values. Note also that compared to the case where the noise levels were very low, the mask obtained in noisy setting is slightly closer to a low-pass mask. The reason for this is that the noise hides the relatively weaker signal present at high frequencies, while only having a minimal effect on the stronger signal present at lower frequencies.

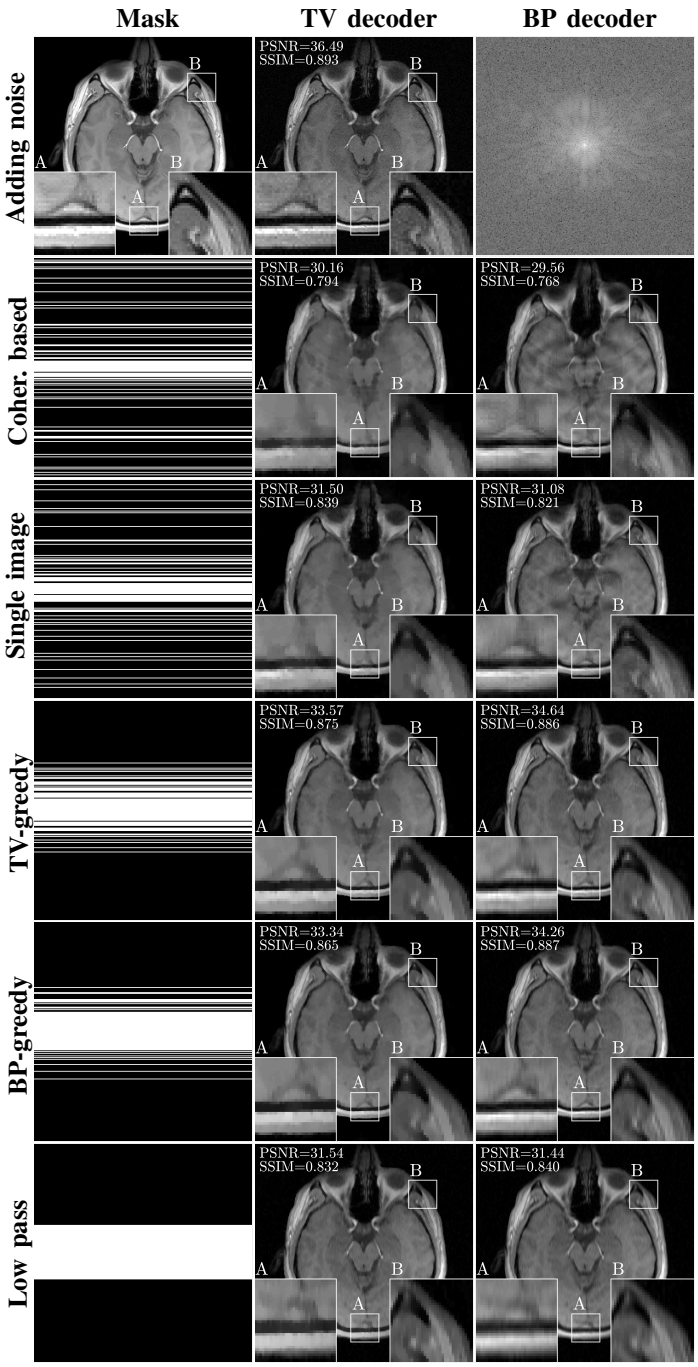

Fig. 6. Masks obtained and example reconstructions under TV and BP decoding at 25% sampling rate. PSNR is used as the performance measure in the greedy method given in Algorithm 1. The reconstruction shown is for subject 2, slice 4. We also present the performances of coherence-based [31] and single-image based [41] masks. In the first row, we present the ground truth, noisy ground truth, and its k-space. The last row shows the low-pass mask performance.

TABLE IV
PSNR AND SSIM PERFORMANCES AT 25% SUBSAMPLING RATE WITH ADDITIVE NOISE, AVERAGED OVER 60 TEST SLICES AND 10 RANDOM NOISE DRAWS. THE CASES THAT THE TRAINING IS MATCHED TO THE PERFORMANCE MEASURE AND DECODER ARE HIGHLIGHTED IN BOLD.

Decoder	TV		BP	
Metric / Mask	SSIM	PSNR	SSIM	PSNR
Coherence-based	0.799	29.73	0.808	30.29
Single-Image	0.828	30.86	0.843	31.58
TV-greedy	0.878	**33.07**	0.896	34.21
BP-greedy	0.875	32.89	0.893	**34.07**
Low Pass	0.859	31.37	0.872	31.80

V. CONCLUSION

We have presented a versatile learning-based framework for selecting masks for compressive MRI, using training signals to optimize for a given decoder and anatomy. As well as having a rigorous justification via statistical learning theory, our approach is seen to provide improved performance on real-world data sets for a variety of reconstruction methods. Since our framework is suited to general decoders, it can potentially be used to optimize the indices for new reconstruction methods that are yet to be discovered. In this work, we focused on 1D subsampling for 2D MRI, 2D subsampling (via horizontal and vertical lines) for 2D MRI, and 1D subsampling for 3D MRI, but our greedy approach can potentially provide an automatic way to optimize the sampling in the settings of 2D subsampling for 3D MRI and non-Cartesian sampling, as opposed to constructing a randomized pattern on a case-by-case basis. For the setting of 3D MRI, there is an additional computational challenge to our greedy algorithm, since the candidate set is large.

In future studies, we will also seek to validate the performance under the important practical variation of *multi-coil* measurements, as well as applications beyond MRI such as computer tomography, phase retrieval, and ultrasound. We finally note that in this paper, the number of subjects and training images used was relatively small, and we anticipate that larger data sets would be of additional benefit in realizing the full power of our theory.

APPENDIX

A. Proof of Proposition 1

Using the fact that η lies in $[0,1]$ and applying Hoeffding's inequality [63], we obtain for any $\Omega \in \mathcal{A}$ and $t > 0$ that

$$\left| \frac{1}{m} \sum_{j=1}^{m} \eta_\Omega(\mathbf{x}_j) - \mathbb{E}_P \left[\eta_\Omega(\mathbf{x}) \right] \right| \leq t,$$

with probability at least $1 - 2\exp(-2mt^2)$. Since the probability of a union of events is upper bounded by the sum of the individual probabilities (i.e., the union bound), we find that the same inequality holds for *all* $\Omega \in \mathcal{A}$ with probability at least $1 - 2\left|\mathcal{A}\right|\exp(-2mt^2)$. The proposition follows by setting $\delta = 2\left|\mathcal{A}\right|\exp(-2mt^2)$ and solving for t.

B. Proof of Proposition 2

By the fact that the Fourier transform is a unitary operation and i.i.d. Gaussian vectors are invariant under unitary transforms, we have

$$\begin{aligned}\overline{\eta}_{\text{noisy}}(\Omega) &= \mathbb{E}\left[\eta(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{P}_\Omega \mathbf{\Psi}\mathbf{x} + \mathbf{w}))\right] \\ &= \mathbb{E}\left[\eta(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{P}_\Omega \mathbf{\Psi}(\mathbf{x} + \mathbf{v})))\right], \end{aligned} \tag{18}$$

where $\hat{\mathbf{x}}(\mathbf{b})$ denotes the estimator applied to the noisy output $\mathbf{b}$, and $\mathbf{v}$ has the same distribution as any given $\mathbf{v}_j$.

Let $\widetilde{\mathbf{v}} = \xi(\mathbf{v})$ be the denoised version of $\mathbf{v}$. Using the triangle inequality, we write

$$\begin{aligned}&\left|\frac{1}{m}\sum_{j=1}^{m}\eta(\mathbf{x}_j + \widetilde{\mathbf{v}}_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}(\mathbf{x}_j + \mathbf{v}_j))) - \overline{\eta}_{\text{noisy}}(\Omega)\right| \\ &= \Bigg|\frac{1}{m}\sum_{j=1}^{m}\eta(\mathbf{x}_j + \widetilde{\mathbf{v}}_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}(\mathbf{x}_j + \mathbf{v}_j))) \\ &\qquad - \mathbb{E}\left[\eta(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}(\mathbf{x} + \mathbf{v})))\right]\Bigg| \\ &\leq \Bigg|\frac{1}{m}\sum_{j=1}^{m}\eta(\mathbf{x}_j + \widetilde{\mathbf{v}}_j, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}(\mathbf{x}_j + \mathbf{v}_j))) \\ &\qquad - \mathbb{E}\left[\eta(\mathbf{x} + \widetilde{\mathbf{v}}, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}(\mathbf{x} + \mathbf{v})))\right]\Bigg| \\ &\quad + \left|\mathbb{E}\left[\eta(\mathbf{x} + \widetilde{\mathbf{v}}, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}\mathbf{x} + \mathbf{w}))\right] - \mathbb{E}\left[\eta(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{P}_\Omega\mathbf{\Psi}\mathbf{x} + \mathbf{w}))\right]\right|.\end{aligned}$$

Using (18) and following the proof of Proposition 1, the first term is upper bounded by $\sqrt{\frac{1}{2m}\log\left(\frac{2|\mathcal{A}|}{\delta}\right)}$ with probability at least $1-\delta$. Moreover, by the continuity condition assumed in the proposition statement, the second term above is upper bounded by $L\,\mathbb{E}\left[\|\widetilde{\mathbf{v}}\|_2\right]$, thus completing the proof.

ACKNOWLEDGMENT

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 725594 - time-data), and from Hasler Foundation Program: Cyber Human Systems (project number 16066). It was also sponsored by the Department of the Navy, Office of Naval Research (ONR) under a grant number N62909-17-1-2111.

REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] E. M. Eksioglu, "Decoupled algorithm for MRI reconstruction using nonlocal block matching model: BM3D-MRI," *J. Math. Imag. Vision*, vol. 56, no. 3, pp. 430–440, 2016.

[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "BM3D image denoising with shape-adaptive principal component analysis," in *Sig. Proc. Adaptive Sparse Struc. Repr. (SPARS)*, 2009. [Online]. Available: https://hal.inria.fr/inria-00369582

[5] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for MR image reconstruction," in *Int. Conf. Inf. Proc. Medical Imag.*, 2017, pp. 647–658.

[6] M. Murphy, M. Alley, J. Demmel, K. Keutzer, S. Vasanawala, and M. Lustig, "Fast ℓ_1-spirit compressed sensing parallel imaging mri: Scalable parallel implementation and clinically feasible runtime," *IEEE transactions on medical imaging*, vol. 31, no. 6, pp. 1250–1262, 2012.

[7] M. Uecker, P. Lai, M. J. Murphy, P. Virtue, M. Elad, J. M. Pauly, S. S. Vasanawala, and M. Lustig, "Espirit?an eigenvalue approach to autocalibrating parallel mri: where sense meets grappa," *Magnetic resonance in medicine*, vol. 71, no. 3, pp. 990–1001, 2014.

[8] K. H. Jin, D. Lee, and J. C. Ye, "A general framework for compressed sensing and parallel mri using annihilating filter based low-rank hankel matrix," *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 480–495, 2016.

[9] 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, 2011.

[10] ——, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2013.

[11] ——, "Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging," *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2519–2557, 2015.

[12] ——, "Data-driven learning of a union of sparsifying transforms model for blind compressed sensing," *IEEE Transactions on Computational Imaging*, vol. 2, no. 3, pp. 294–309, 2016.

[13] S. G. Lingala, Y. Hu, E. DiBella, and M. Jacob, "Accelerated dynamic mri exploiting sparsity and low-rank structure: kt slr," *IEEE transactions on medical imaging*, vol. 30, no. 5, pp. 1042–1054, 2011.

[14] H. Yoon, K. S. Kim, D. Kim, Y. Bresler, and J. C. Ye, "Motion adaptive patch-based low-rank approach for compressed sensing cardiac cine mri," *IEEE transactions on medical imaging*, vol. 33, no. 11, pp. 2069–2085, 2014.

[15] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components," *Magnetic Resonance in Medicine*, vol. 73, no. 3, pp. 1125–1136, 2015.

[16] S. Ravishankar, B. E. Moore, R. R. Nadakuditi, and J. A. Fessler, "Low-rank and adaptive sparse signal (lassi) models for highly accelerated dynamic imaging," *IEEE transactions on medical imaging*, vol. 36, no. 5, pp. 1116–1128, 2017.

[17] S. G. Lingala and M. Jacob, "Blind compressive sensing dynamic mri," *IEEE transactions on medical imaging*, vol. 32, no. 6, pp. 1132–1145, 2013.

[18] Y. Wang and L. Ying, "Compressed sensing dynamic cardiac cine mri using learned spatiotemporal dictionary," *IEEE transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1109–1120, 2014.

[19] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t focuss: A general compressed sensing framework for high resolution dynamic mri," *Magnetic resonance in medicine*, vol. 61, no. 1, pp. 103–116, 2009.

[20] J. Tsao, P. Boesiger, and K. P. Pruessmann, "k-t blast and k-t sense: Dynamic mri with high frame rate exploiting spatiotemporal correlations," *Magnetic resonance in medicine*, vol. 50, no. 5, pp. 1031–1042, 2003.

[21] N. Aggarwal and Y. Bresler, "Patient-adapted reconstruction and acquisition dynamic imaging method (paradigm) for mri," *Inverse Problems*, vol. 24, no. 4, p. 045015, 2008.

[22] B. Sharif, J. A. Derbyshire, A. Z. Faranesh, and Y. Bresler, "Patient-adaptive reconstruction and acquisition in dynamic imaging with sensitivity encoding (paradise)," *Magnetic Resonance in Medicine*, vol. 64, no. 2, pp. 501–513, 2010.

[23] B. Bilgic, V. K. Goyal, and E. Adalsteinsson, "Multi-contrast reconstruction with bayesian compressed sensing," *Magnetic resonance in medicine*, vol. 66, no. 6, pp. 1601–1615, 2011.

[24] J. M. Duarte-Carvajalino, C. Lenglet, K. Ugurbil, S. Moeller, L. Carin, and G. Sapiro, "A framework for multi-task bayesian compressive sensing of dw-mri," in *Proceedings of the CDMRI MICCAI workshop*, 2012, pp. 1–13.

[25] Y. Huang, J. Paisley, Q. Lin, X. Ding, X. Fu, and X.-P. Zhang, "Bayesian nonparametric dictionary learning for compressed sensing mri," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5007–5019, 2014.

[26] D. Lee, J. Yoo, and J. C. Ye, "Deep residual learning for compressed sensing mri," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 15–18.

[27] Y. S. Han, J. Yoo, and J. C. Ye, "Deep learning with domain adaptation for accelerated projection reconstruction mr," *arXiv preprint arXiv:1703.01135*, 2017.

[28] 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.

[29] M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly *et al.*, "Deep generative adversarial networks for compressed sensing automates mri," *arXiv preprint arXiv:1706.00051*, 2017.

[30] C. M. Sandino, N. Dixit, J. Y. Cheng, and S. S. Vasanawala, "Deep convolutional neural networks for accelerated dynamic magnetic resonance imaging," *preprint*.

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

[32] S. Vasanawala, M. Murphy, M. T. Alley, P. Lai, K. Keutzer, J. M. Pauly, and M. Lustig, "Practical parallel imaging compressed sensing MRI: Summary of two years of experience in accelerating body MRI of pediatric patients," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011, pp. 1039–1043.

[33] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," *arXiv preprint arXiv:1302.0561v3*, 2013.

[34] B. Adcock, A. C. Hansen, and B. Roman, "The quest for optimal sampling: Computationally efficient, structure-exploiting measurements for compressed sensing," in *Compressed Sensing and Its Applications*. Springer, 2015, pp. 143–167.

[35] H. Wang, D. Liang, and L. Ying, "Pseudo 2D random sampling for compressed sensing MRI," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 2672–2675.

[36] N. Chauffert, P. Ciuciu, J. Kahn, and P. Weiss, "Variable density sampling with continuous trajectories," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1962–1992, 2014.

[37] C. Boyer, P. Weiss, and J. Bigot, "An algorithm for variable density sampling with block-constrained acquisition," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1080–1107, 2014.

[38] J. Bigot, C. Boyer, and P. Weiss, "An analysis of block sampling strategies in compressed sensing," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2125–2139, 2016.

[39] F. Knoll, C. Clason, C. Diwoky, and R. Stollberger, "Adapted random sampling patterns for accelerated MRI," *Magnetic resonance materials in physics, biology and medicine*, vol. 24, no. 1, pp. 43–50, 2011.

[40] Y. Zhang, B. S. Peterson, G. Ji, and Z. Dong, "Energy preserved sampling for compressed sensing MRI," *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[41] J. Vellagoundar and R. R. Machireddy, "A robust adaptive sampling method for faster acquisition of MR images," *Magnetic resonance imaging*, vol. 33, no. 5, pp. 635–643, 2015.

[42] D.-d. Liu, D. Liang, X. Liu, and Y.-t. Zhang, "Under-sampling trajectory design for compressed sensing MRI," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 73–76.

[43] S. Ravishankar and Y. Bresler, "Adaptive sampling design for compressed sensing MRI," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 3751–3755.

[44] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, "Optimization of k-space trajectories for compressed sensing by Bayesian experimental design," *Magnetic resonance in medicine*, vol. 63, no. 1, pp. 116–126, 2010.

[45] L. Baldassarre, Y.-H. Li, J. Scarlett, B. Gözcü, I. Bogunovic, and V. Cevher, "Learning-based compressive subsampling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 809–822, 2016.

[46] L. Weizman, Y. C. Eldar, and D. Ben Bashat, "Compressed sensing for longitudinal MRI: An adaptive-weighted approach," *Medical Physics*, vol. 42, no. 9, pp. 5195–5208, 2015.

[47] M. Mardani, G. B. Giannakis, and K. Ugurbil, "Tracking tensor subspaces with informative random sampling for real-time mr imaging," *arXiv preprint arXiv:1609.04104*, 2016.

[48] J. Choi and H. Kim, "Implementation of time-efficient adaptive sampling function design for improved undersampled mri reconstruction," *Journal of Magnetic Resonance*, vol. 273, pp. 47–55, 2016.

[49] F. Zijlstra, M. A. Viergever, and P. R. Seevinck, "Evaluation of variable density and data-driven k-space undersampling for compressed sensing magnetic resonance imaging," *Investigative radiology*, vol. 51, no. 6, pp. 410–419, 2016.

[50] Y. Li, R. Yang, C. Zhang, J. Zhang, S. Jia, and Z. Zhou, "Analysis of generalized rosette trajectory for compressed sensing mri," *Medical physics*, vol. 42, no. 9, pp. 5530–5544, 2015.

[51] H. Wang, X. Wang, Y. Zhou, Y. Chang, and Y. Wang, "Smoothed random-like trajectory for compressed sensing mri," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 404–407.

[52] L. Feng, R. Grimm, K. T. Block, H. Chandarana, S. Kim, J. Xu, L. Axel, D. K. Sodickson, and R. Otazo, "Golden-angle radial sparse parallel mri: Combination of compressed sensing, parallel imaging, and golden-angle radial sampling for fast and flexible dynamic volumetric mri," *Magnetic resonance in medicine*, vol. 72, no. 3, pp. 707–717, 2014.

[53] F. Hilbert, T. Wech, D. Hahn, and H. Köstler, "Accelerated radial fourier-velocity encoding using compressed sensing," *Zeitschrift für Medizinische Physik*, vol. 24, no. 3, pp. 190–200, 2014.

[54] C. Lazarus, P. Weiss, N. Chauffert, F. Mauconduit, M. Bottlaender, A. Vignaud, and P. Ciuciu, "Sparkling: Novel non-cartesian sampling schemes for accelerated 2d anatomical imaging at 7t using compressed sensing," in *25th annua meeting of the International Society for Magnetic Resonance Imaging*, 2017.

[55] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[56] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer, "Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets," *ACM Transactions on Mathematical Software (TOMS)*, vol. 42, no. 1, p. 5, 2016.

[57] S. Becker, J. Bobin, and E. J. Candès, "Nesta: A fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.

[58] E. Eksioglu, "BM3D-MRI project page," http://web.itu.edu.tr/eksioglue/pubs/BM3D_MRI.htm.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Conf. Comp. Vision*, 2015, pp. 1026–1034.

[60] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[62] J. W. Wallis, T. R. Miller, C. A. Lerner, and E. C. Kleerup, "Three-dimensional display in nuclear medicine," *IEEE Transactions on Medical Imaging*, vol. 8, no. 4, pp. 297–230, 1989.

[63] P. Massart, *Concentration Inequalities and Model Selection*. Berlin: Springer-Verl., 2007.