Learning Sampling and Model-Based Signal Recovery for Compressed Sensing MRI

LEARNING SAMPLING AND MODEL-BASED SIGNAL RECOVERY FOR COMPRESSED SENSING MRI

Iris A.M. Huijben[1], Bastiaan S. Veeling[2], and Ruud J.G. van Sloun[1]

[1]Dept. of Electrical Engineering, Eindhoven University of Technology, The Netherlands
[2]Dept. of Computer Science, University of Amsterdam, The Netherlands

ABSTRACT

Compressed sensing (CS) MRI relies on adequate undersampling of the k-space to accelerate the acquisition without compromising image quality. Consequently, the design of optimal sampling patterns for these k-space coefficients has received significant attention, with many CS MRI methods exploiting variable-density probability distributions. Realizing that an optimal sampling pattern may depend on the downstream task (e.g. image reconstruction, segmentation, or classification), we here propose joint learning of both task-adaptive k-space sampling and a subsequent model-based proximal-gradient recovery network. The former is enabled through a probabilistic generative model that leverages the Gumbel-softmax relaxation to sample across trainable beliefs while maintaining differentiability. The proposed combination of a highly flexible sampling model and a model-based (sampling-adaptive) image reconstruction network facilitates exploration and efficient training, yielding improved MR image quality compared to other sampling baselines.

Index Terms— Compressed sensing, model-based deep learning, magnetic resonance imaging

1. INTRODUCTION

Magnetic resonance imaging (MRI) is an invaluable non-invasive medical imaging modality that enables reliable diagnostic imaging for a wide range of clinical applications with unmatched soft-tissue contrast and high resolution. It is however also associated with long acquisition times, thereby diminishing patient comfort, compromising high-quality dynamic imaging (e.g. cardiac), and yielding high procedural costs with limited throughput. To overcome this, accelerated MRI approaches have extensively been studied, initially resulting in parallel imaging methods that are commonplace in today's MRI devices, such as SENSE [1].

More recently, methods that exploit both redundancy in the spatial frequency components as well as structural

image priors (e.g. sparsity in some domain) have been proposed. These compressed sensing (CS) approaches [2, 3, 4] speed-up data acquisition by undersampling the full k-space through hand-tailored sampling schemes, while retaining image quality by exploiting the aforementioned sparsity during image reconstruction. Popular sampling schemes follow from variable-density probability density functions (PDF) [5], which are mostly based on the empirical observation that lower frequencies should be sampled more densely than high frequencies to enable adequate image recovery. The authors of [6] bridge the gap between these empirically found variable-density sampling designs and the mathematical justification. In practice, optimal k-space sampling may however depend on the anatomy and the imaging task at hand (e.g whole body, dynamic cardiac, or segmentation), and these approaches therefore fail to exploit the full data distribution available.

With the aim of exploiting such information, several data-driven approaches have been introduced for optimization of k-space sampling. Among these, greedy algorithms were proposed [7, 8] to handle the factorial scaling of this problem. Other approaches leverage gradient-based learning of sampling schemes through compact (gradient-distributing) interpolation kernels [9]. This was however found to inhibit effective exploration and flexibility, with learned subsampling schemes not deviating far from their initialization.

In pursuit of a more effective and flexible approach to joint learning of sampling and image recovery, we here propose to adopt a generative k-space sampling model [10], termed Deep Probabilistic Subsampling (DPS), along with a concurrent unfolded proximal gradient network [11] for recovery. DPS leverages the recently proposed Gumbel-softmax relaxation for differentiable sampling from categoricals and distributes trainable beliefs over relevant k-space coefficients. The unfolded recovery model exploits both the known measurement domain transform (Fourier) and the sampling itself, while learning an effective image proximal mapping. This allows for fast co-adaptation of the reconstruction network to the selected samples and promotes exploration of an optimal sampling pattern.

The remainder of this paper is structured as follows. We start by giving a description of the system model in section

2.1, after which we elaborate on our subsampling and reconstruction methods in sections 2.2 and 2.3, respectively. Section 3 elucidates upon the conducted experiments: sparse signal recovery from partial Fourier measurements (3.1), and MRI reconstruction from partial k-space measurements (3.2). We conclude in section 4.

Throughout the paper, bold capital letters denote matrices. We index matrices by row-first notation, i.e. $\mathbf{a}_m$ denotes the mth row-vector of matrix $\mathbf{A}$, of which $a_{m,n}$ is the scalar in the nth column.

2. METHODS

2.1. System model

We define a partial Fourier measurement $\mathbf{y} \in \mathbb{C}^M$, acquired by subsampling the full vector of Fourier coefficients $\mathbf{x} \in \mathbb{C}^N$, using $\mathbf{A}_{(\Phi)} \in \{0, 1\}^{M \times N}$, parameterized by Φ:

$$\mathbf{y} = \mathbf{A}_{(\Phi)}\mathbf{x}. \quad (1)$$

For each row $\mathbf{a}_m$ in $\mathbf{A}_{(\Phi)}$ we require $\|\mathbf{a}_m\|_1 = 1$, making $\mathbf{A}_{(\Phi)}$ a subset selector that selects M out of N ($M \ll N$) elements from $\mathbf{x}$. We aim to achieve a subsequent task $\mathbf{s}$ (e.g. image reconstruction) from $\mathbf{y}$ through a nonlinear task function $f_\theta(\cdot)$ that is differentiable with respect to its trainable parameters θ:

$$\hat{\mathbf{s}} = f_\theta(\mathbf{y}). \quad (2)$$

2.2. DPS: Deep probabilistic subsampling

To enable training of both the parameters for the subsampling scheme, Φ, and the task network, θ, we adopt a probabilistic generative sampling model that circumvents the non-differentiable nature of subset selection [10]. Rather than attempting to train the sampling directly, we instead update beliefs across all possible Fourier (or k-space) coefficients.

To that end, we introduce M independent categorical random variables, following distributions $\text{Cat}_m(N, \boldsymbol{\pi}_m)$, with $m \in \{1, \dots, M\}$, that all express beliefs across the full set of N coefficients though the N normalized class probabilities $\pi_{m,n}$ in $\boldsymbol{\pi}_m \in \mathbb{R}^N$. We define a logit $\phi_{m,n}$ as the natural logarithm of the unnormalized class probability, such that:

$$\pi_{m,n} = \frac{\exp \phi_{m,n}}{\sum_{i=1}^N \exp \phi_{m,i}}. \quad (3)$$

To draw a sample from a categorical, we exploit the Gumbel-max trick [12]. This trick perturbs the unnormalized logits ϕ_m with Gumbel noise $\mathbf{e}_m$ in order to create randomized keys from which the highest key is selected. Subsequently a length-N one-hot vector is created, i.e. a vector that contains only one non-zero element (with value 1) at the index of the drawn sample, through a function that we denote as $\text{onehot}_N(\cdot)$. We iteratively draw samples without replacement across the M categoricals, which we implement by masking previously sampled categories (here, Fourier coefficients) through $\mathbf{w}_{m-1} \in \{-\infty, 0\}$. Each row of $\mathbf{A}_{(\Phi)}$ thus takes the following form:

$$\mathbf{a}_m = \text{onehot}_N \left\{ \underset{n \in \{1\dots N\}}{\arg\max} \left\{ w_{m-1,n} + \phi_{m,n} + e_{m,n} \right\} \right\}. \quad (4)$$

To enable error backpropagation, we adopt the Gumbel-softmax trick, thereby relaxing the non-differentiable $\arg\max(\cdot)$ function (hard sampling) by replacing it with the $\text{softmax}_\tau(\cdot)$ function (termed soft sampling) [13, 14], having temperature parameter τ. We thus define:

$$\nabla_{\boldsymbol{\phi}_m} \mathbf{a}_m := \nabla_{\boldsymbol{\phi}_m} \mathbb{E}_{\mathbf{e}_m} \left[\text{softmax}_\tau \{\mathbf{w}_{m-1} + \boldsymbol{\phi}_m + \mathbf{e}_m\} \right] =$$
$$\nabla_{\boldsymbol{\phi}_m} \mathbb{E}_{\mathbf{e}_m} \left[\frac{\exp\{(\mathbf{w}_{m-1} + \boldsymbol{\phi}_m + \mathbf{e}_m)/\tau\}}{\sum_{i=1}^N \exp\{(w_{m-1,i} + \phi_{m,i} + e_{m,i})/\tau\}} \right]. \quad (5)$$

Across all experiments presented in this paper, the unnormalized logits Φ are initialized by i.i.d zero-mean Gaussian noise ($\sigma^2 = \frac{1}{4}$) samples.

Top-M sampling for large N Defining M categorical random variables with N categories offers high expressiveness, but comes at the cost of having a large amount of trainable parameters and poses memory challenges. This particularly holds for large N, e.g. when sampling MRI k-space coefficients. As such, we also consider sampling M elements without replacement from only a single categorical distribution with trained logits, termed top-M sampling.

As shown by Kool *et al.* [15], top-M hard sampling is equivalent to sampling M times without replacement from the same distribution. As such, we can define

$$\mathbf{A}_{(\Phi)} = \text{Mhot}_N \{ \underset{n \in \{1\dots N\}}{\text{topM}} (\phi_n + e_n) \}, \quad (6)$$

where $\text{topM}(\cdot)$ returns the indices of the unique M highest values, and $\text{Mhot}_N(\cdot)$ creates M one-hot encoded rows from the returned indices by $\text{topM}(\cdot)$, yielding the rows of $\mathbf{A}_{(\Phi)}$.

The authors of [16] demonstrated that iterative sampling without replacement from a single Gumbel-softmax distribution is a valid top-M relaxation. As such we can directly leverage (5) for backpropagation.

2.3. Model-based deep networks for signal recovery

In CS, the ill-posed problem of signal/image reconstruction from compressive measurements is typically solved using proximal gradient methods [11]. The iterative nature makes them tediously slow however, and the adopted regularizers are typically selected empirically. Recently, the use of deep neural networks for signal/image reconstruction has gained popularity, being a fast, fully data-driven alternative. Common network designs are closely related to the U-Net [17, 18]. While popular, such a structure is inherently not sampling-adaptive and does not exploit available knowledge

on the acquisition. We therefore propose to leverage a deep unfolded proximal gradient method [11], in which iterations of a known iterative algorithm are unfolded as a feedforward neural network. The update rule of a this proximal gradient scheme takes the following form:

$$\hat{\mathbf{s}}^{(k+1)} = \mathcal{P}_{(\zeta)}^{(k)} \left\{ \hat{\mathbf{s}}^{(k)} - \alpha_{(\psi)}^{(k)} \mathbf{F}^H \mathbf{A}_{(\Phi)}^T \left(\mathbf{A}_{(\Phi)} \mathbf{F} \hat{\mathbf{s}}^{(k)} - \mathbf{A}_{(\Phi)} \mathbf{x} \right) \right\} \tag{7}$$

$$= \mathcal{P}_{(\zeta)}^{(k)} \left\{ B^{(k)} \hat{\mathbf{s}}^{(k)} + C^{(k)} [\mathbf{A}_{(\Phi)}^T \mathbf{A}_{(\Phi)} \mathbf{x}] \right\}, \tag{8}$$

where $\mathbf{F} \in \mathbb{C}^{N \times N}$ is the Fourier transform matrix, $\mathbf{F}^H$ is its Hermitian, $\alpha^{(k)}$ and $\mathcal{P}_{(\zeta)}(\cdot)$ are the stepsize and the proximal operator (parameterized by ζ), respectively, $B^{(k)} = \mathbf{I} - \alpha_{(\psi)}^{(k)} \mathbf{F}^H \mathbf{A}_{(\Phi)}^T \mathbf{A}_{(\Phi)} \mathbf{F}$, and $C^{(k)} = \alpha_{(\psi)} \mathbf{F}^H$. Each iteration in (7) is a model-based (MB) network layer, with trainable parameters ζ. Depending on the application, we will either fully learn a suitable proximal mapping $\mathcal{P}_{(\zeta)}(\cdot)$ using a neural network (MRI reconstruction), or only learn a layer-wise shrinkage parameter for a known analytic proximal operator (sparse recovery).

This MB-approach not only facilitates the incorporation of domain transforms and sampling patterns in the network architecture (allowing it to co-adapt when learning the sampling scheme), but also greatly reduces the amount of trainable parameters compared to conventional convolutional networks.

3. EXPERIMENTS

3.1. Partial Fourier sampling of sparse signals

Experiment setup   We first demonstrate DPS for sparse signal reconstruction from partial Fourier measurements [5, 19, 20]. To that end, we synthetically generate random sparse signal vectors $\mathbf{s} \in \mathbb{R}^{128}$, with 5 non-zero coefficients, which we subsequently Fourier-transform to yield $\mathbf{x} \in \mathbb{C}^{128}$. We compare subsampling using a fixed random pattern (classic CS) to DPS, using either top-M or top-1 sampling.

Reconstruction model   The MB recovery network is based on an unfolded version of the iterative shrinkage and thresholding algorithm [21]. It comprises 3 layers that follow the iterations in (8), with $B^{(k)}$ and $C^{(k)}$ here replaced by trainable fully-connected layers, and $\mathcal{P}_{(\zeta)}(\cdot)^{(k)}$ being a smoothed sigmoid-based soft-thresholding operator [22], with learned thresholding parameters for each k. We compare this MB network to a standard fully-connected (FC) network with 5 hidden layers containing $256, 512, 256, 128,$ and 128 nodes, respectively, each followed by leaky ReLU activations.

Training details   The model is trained by minimizing the mean-squared error (MSE) between the model predictions and the ground truth signals $\mathbf{s}$. For DPS, we promote training towards one-hot distributions $\text{Cat}_m(N, \boldsymbol{\pi}_m)$ through an additional mild entropy penalty (multiplier $\mu = 1e-8$) on the
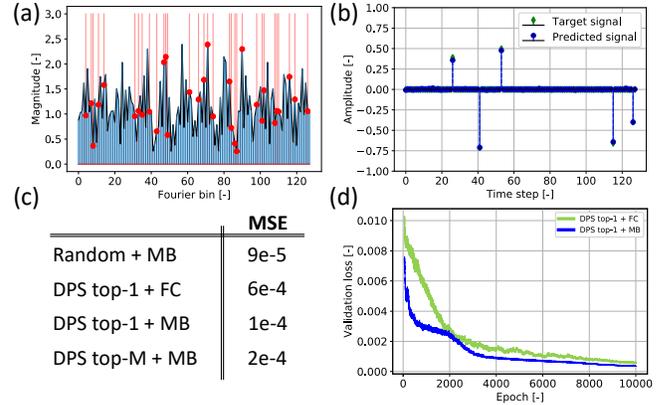

Fig. 1. (a) Partial Fourier measurements by Deep Probabilistic Subsampling (DPS) with $N/M = 4$ (red) and all Fourier coefficients (blue). (b) Reconstructed sparse signal using the proposed model-based (MB) deep network. (c) MSE values for DPS vs random partial Fourier measurements, and MB vs fully-connected (FC) reconstruction networks. (d) Training graphs for DPS top-1 with MB (blue) and FC (green) networks for the first 10,000 epochs.

	MSE
Random + MB	9e-5
DPS top-1 + FC	6e-4
DPS top-1 + MB	1e-4
DPS top-M + MB	2e-4

trainable logits. Note that this penalty is not applied in case of top-M sampling. The resulting total loss is minimized using stochastic gradient descend across mini-batches of 16 randomly-generated Fourier-transformed data vectors. To this end we adopt the ADAM solver ($\beta_1 = 0.9, \beta_2 = 0.999,$ and $\epsilon = 1e-7$) [23], with separate learning rates for the generative sampling model and the reconstruction network ($8e-2$ and $1e-3$, respectively). We train until the validation loss converged, with the temperature parameter τ gradually lowered from 5.0 to 0.5.

Results   Figure 1b visualizes sparse signal reconstruction using the proposed MB deep network with learned subsampling by DPS (top-1 sampling), selecting 25% of the Fourier coefficients (see fig. 1a, samples are displayed in red). The sparse signal components are well-recovered. Figure 1c provides the MSE values for all tested models, and the training graphs (first 10,000 epochs) for DPS top-1 with MB and fully-connected recovery networks are displayed in fig. 1d. We clearly see the positive effect of a MB recovery network compared to a FC network in terms of performance and convergence, which we attribute to the MB-network's ability to efficiently co-adapt with the changing sampling pattern.

3.2. MR images from partial k-space measurements

Experiment setup   To show the merit of DPS for MRI k-space sampling, we leverage the NYU fastMRI database that contains a vast amount of knee MRI volumes [20]. We selected only the single-coil measurements, and the outer slices of all volumes (which mostly depicted background) were removed. We randomly selected 8000 MRI slices for training, and 2000 and 3000 for validation and testing, respectively.
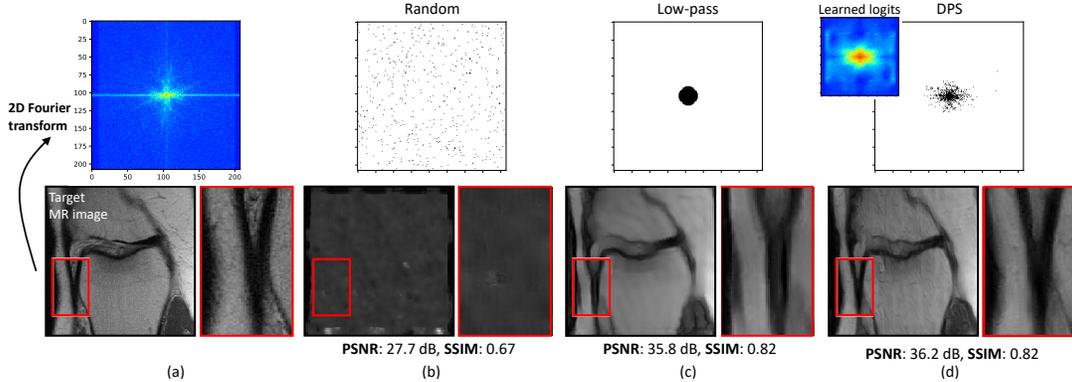
Fig. 2. MR image reconstruction from partial k-space measurements ($N/M = 80$) using (b) uniform random sampling, (c) low-pass sampling, and (d) learned sampling by DPS. (d, top inset) learned distribution logits in DPS, expressing beliefs over k-space coefficients. (a) The target image and full k-space, of which only 1.2% of the coefficients are sampled in (b-d).

The images were cropped to the central 208×208 pixels, and normalized between 0 and 1, after which, by means of the 2D fast-Fourier transform, corresponding k-space images $\mathbf{x} \in \mathbb{C}^{208 \times 208}$ were generated.

We compare learned subsampling by DPS to a fixed uniform random or low-pass pattern. For DPS, we here adopt top-M sampling from a single trained distribution to alleviate memory challenges due to the large number of potential k-space coefficients. Results are evaluated through the PSNR and structural similarity index (SSIM) [24].

Reconstruction model   Each layer in the MB reconstruction network follows the update rule of (7) (3 unfoldings), with trainable stepsize $\alpha_{(\psi)}^{(k)}$ implemented as a small 3×3 convolutional kernel. The proximal mapping $\mathcal{P}_{(\zeta)}^{(k)}$ was also learned [25], by implementing it as 3 ReLU-activated 2D-convolutional layers with 16 features (3×3 kernels), followed by a linear 2D-convolutional layer (3×3 kernels), mapping the output to a single image.

Training details   We again adopt the MSE loss, but in addition leverage an adversarial loss to promote visually plausible reconstructions. To that end, we adopt a discriminator network D_ξ that is trained to distinguish real from generated MR images [26], while the sampling and image reconstruction parameters are jointly optimized to maximize discriminator loss. The discriminator consists of 3 2D-convolutional layers with 64 features, 3×3 kernels and a stride of 2, each followed by a Leaky ReLU activation. The resulting feature maps are then spatially aggregated into a feature vector through global average pooling, followed by dropout (40%). A final fully-connected layer with sigmoid activation then maps the features to a single output probability.

In addition to the image MSE and adversarial loss, we also penalize the MSE between the discriminator features for generated images and true images. To yield the total training loss, these 3 losses are weighted by multipliers 1, $5e-6$ and

$1e-7$, respectively. We again use a distinct learning rate for the parameters of the generative sampling model and the other parameters, being 0.01 and $2e-4$, respectively. The temperature parameter τ was fixed at 5.0. Training was performed for 500 epochs with mini-batches containing 8 k-space spectra, using the ADAM solver [23] with settings as in section 3.1.

Results   Figure 2 shows the obtained MRI reconstructions for an example from the hold-out test set. The k-space was subsampled by a factor 80 (i.e. only 1.2% of the coefficients are sampled), and the results of DPS were compared to non-trained sampling by two baseline methods (random uniform and low-pass). Random uniform sampling (fig. 2b) did not yield plausible reconstructions, lacking sufficient low-frequency coefficients. While low-pass sampling (fig. 2c) resulted in improved reconstructions, it was outperformed by DPS (fig. 2d). The latter achieves an apparent sharpness and detail that is particularly evident from the zoomed areas. Across the full test set, its PSNR value was moreover higher. When inspecting the learned logits we observe that DPS generally favours low frequencies while also distributing samples towards the higher frequencies.

4. CONCLUSIONS

In this paper we proposed an end-to-end deep learning solution for CS MRI that enables joint optimization of a data-driven task-adaptive subsampling model (deep probabilistic subsampling, DPS) and the parameters of a dedicated model-based signal/image recovery network. We found that learned subsampling yielded improved MR image reconstruction compared to the evaluated non-trained patterns, and showed that the use of a model-based deep reconstruction network facilitates efficient training. Future work includes evaluation of DPS for other MR imaging tasks such as segmentation and classification, as well as its use across a broader range of task-adaptive sampling applications in different fields.

5. REFERENCES

[1] K.P. Pruessmann and Scheidegger M.B. Boesiger P. Weiger, M., "SENSE: sensitivity encoding for fast MRI," *Magnetic resonance in medicine*, vol. 42, no. 5, pp. 952–962, 1999.

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

[3] Y.C. Eldar and G. Kutyniok, *Compressed sensing: Theory and Applications*, Cambridge University Press, 2012.

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

[5] 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.

[6] Ben Adcock, Anders C Hansen, et al., "Breaking the coherence barrier: A new theory for compressed sensing," in *Forum of Mathematics, Sigma*. Cambridge University Press, 2017, vol. 5.

[7] K.H. Jin, M. Unser, and K.M. Yi, "Self-supervised deep active accelerated MRI," *arXiv preprint arXiv:1901.04547*, 2019.

[8] Thomas Sanchez, Baran Gözcü, et al., "Scalable learning-based sampling optimization for compressive dynamic MRI," *arXiv preprint arXiv:1902.00386*, 2019.

[9] T. Weiss, O. Senouf, et al., "PILOT: Physics-informed learned optimal trajectories for accelerated MRI," *arXiv preprint arXiv:1909.05773*, 2019.

[10] I.A.M Huijben, B.S. Veeling, et al., "Learning sub-sampling and signal recovery with applications in ultrasound imaging," *arXiv preprint arXiv:1908.05764*, 2019.

[11] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[12] E. J. Gumbel, "Statistical theory of extreme values and some practical applications," *NBS Applied Mathematics Series*, vol. 33, 1954.

[13] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with gumbel-softmax," *stat*, vol. 1050, pp. 17, 2017.

[14] C Maddison, A Mnih, and Y Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *International Conference on Machine Learning*, 2017.

[15] "Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement, author=Kool, W. and Van Hoof, H. and Welling, M., journal=arXiv preprint arXiv:1903.06059, year=2019," .

[16] S.M. Xie and S. Ermon, "Reparameterizable subset sampling via continuous relaxations," in *International Joint Conference on Artificial Intelligence*, 2019.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[18] C.D. Bahadir, A.V. Dalca, and M.R. Sabuncu, "Learning-based optimization of the under-sampling pattern in MRI," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 780–792.

[19] M. Lustig, D.L. Donoho, et al., "Compressed sensing MRI," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 72, 2008.

[20] J. Zbontar, F. Knoll, et al., "fastMRI: An open dataset and benchmarks for accelerated MRI," *arXiv preprint arXiv:1811.08839*, 2018.

[21] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *International Conference on Machine Learning*. Omnipress, 2010, pp. 399–406.

[22] A.M. Atto, D. Pastor, and G. Mercier, "Smooth sigmoid wavelet shrinkage for non-parametric estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3265–3268.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *2014 International Conference on Learning Representations*, 2014.

[24] Z. Wang, A.C. Bovik, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[25] M. Mardani, Q. Sun, et al., "Neural proximal gradient descent for compressive imaging," in *Advances in Neural Information Processing Systems*, 2018, pp. 9573–9583.

[26] C. Ledig, L. Theis, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.