Instabilities in Conventional Multi-Coil MRI Reconstruction with Small Adversarial Perturbations

Chi Zhang[1,2], Jinghan Jia[3], Burhaneddin Yaman[1,2], Steen Moeller[2], Sijia Liu[4], Mingyi Hong[1] and Mehmet Akçakaya[1,2]
[1]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN
[2]Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN
[3]University of Florida, Gainesville, FL
[4]MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA
Emails: {zhan4906, yaman013, moell018, mhong, akcakaya}@umn.edu, jinghan.jia@ufl.edu, sijia.liu@ibm.com

***Abstract*—Although deep learning (DL) has recently received significant attention in accelerated MRI, recent studies suggest that small perturbations may lead to large instabilities in DL-based reconstructions. This has also highlighted concerns for their utility in clinical settings. However, these works focus on single-coil acquisitions, which are not practically relevant. In this work, we investigate how small adversarial perturbations affect multi-coil MRI reconstruction, particularly using conventional non-DL methods. Our results indicate that for multi-coil MRI reconstruction, conventional parallel imaging and multi-coil compressed sensing (CS) methods also exhibit considerable instabilities against small adversarial perturbations. Moreover, for physics-guided DL reconstructions that utilize the forward encoding operator explicitly, such small perturbations predominantly target the linear data-consistency units. These results suggest that at high acceleration rates, adversarial attacks exploit the ill-conditioning of the forward encoding operator.**

I. Introduction

Deep learning (DL) has recently emerged as a powerful means for accelerated MRI reconstruction with improved image quality especially at higher acceleration rates [1–6]. While DL methods have been similarly transformative in multiple image processing and computer vision tasks, it is well-understood that they may be susceptible to instabilities arising from small adversarial perturbations, which lead to no noticeable differences in input but significantly impact the output, due to their non-linear nature [7–10]. These adversarial perturbations were also recently shown [11] to affect several DL methods for MRI reconstruction [1, 12–14]. Furthermore, it was suggested that both researchers and FDA should be cognizant of such issues for DL reconstructions used in MRI [11]. Following this work, several studies explored adversarial training strategies to improve the robustness of DL methods for MRI reconstruction [15, 16].

However, all of these aforementioned works regarding stability of DL reconstruction for MRI focused on single-coil datasets, while in practice, almost exclusively multi-coil MRI is used for acquisitions. Clinically, such datasets are reconstructed using linear parallel imaging techniques [17–19], and more recently using regularized methods, such as compressed sensing [20]. However, the impact of adversarial attacks on conventional multi-coil MRI reconstruction methods, especially the clinically used ones, remains not investigated. Furthermore, most regularized reconstruction strategies, including compressed sensing, as well as physics-guided DL reconstruction, involve a linear data consistency operation, which in itself may be susceptible to instabilities at high acceleration rates in the multi-coil setting.

In this work, we investigated the effect of small adversarial attacks on conventional multi-coil MRI reconstruction strategies, focusing on linear parallel imaging, and compressed sensing. Additionally, we also explored the behavior of such attacks on physics-guided DL reconstructions based on algorithm unrolling [21] to determine whether the non-linear neural network or the linear data consistency operation was more susceptible to perturbations. Our results indicate that for highly-accelerated multi-coil MRI reconstruction, parallel imaging and multi-coil compressed sensing are also susceptible to large instabilities from small adversarial perturbations.

II. Methods

A. Background on Multi-Coil MRI Reconstruction

In most clinical MRI systems, multiple receiver coils are used for data acquisition. Let $\mathbf{E}_\Omega$ be the multi-coil encoding operator that includes the coil sensitivity information as well as the partial Fourier transform with sub-sampling pattern Ω. The multi-coil acquisition model is given as:

$$\mathbf{y}_\Omega = \mathbf{E}_\Omega \mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{y}_\Omega$ denotes the acquired multi-coil data, $\mathbf{x}$ denotes the image of interest, and $\mathbf{n}$ is measurement noise. For i.i.d. Gaussian noise, the maximum likelihood estimation leads to:

$$\arg\min_{\mathbf{x}} ||\mathbf{y}_\Omega - \mathbf{E}_\Omega \mathbf{x}||_2^2 = (\mathbf{E}_\Omega^H \mathbf{E}_\Omega)^{-1} \mathbf{E}_\Omega^H \mathbf{y}_\Omega, \tag{2}$$

which forms the basis of the CG-SENSE formulation for parallel imaging [18]. Alternatively, one can perform linear interpolation in k-space using linear shift-invariant convolutional kernels, which forms the basis of the GRAPPA formulation [19].

Another line of work considers the regularized version of (2), given via

$$\arg\min_{\mathbf{x}} ||\mathbf{y}_\Omega - \mathbf{E}_\Omega \mathbf{x}||_2^2 + \mathcal{R}(\mathbf{x}), \tag{3}$$

where the first quadratic term enforces data consistency with acquired measurements, while $\mathcal{R}(\cdot)$ is a regularizer. In conventional strategies, $\mathcal{R}(\cdot)$ is chosen as a Tikhonov-type term [18] or the l_1 norm of transform domain coefficients [20]. In physics-guided DL methods relying on algorithm unrolling, the non-linear representation associated with such a regularizer is learned implicitly through neural networks.

B. Adversarial Attacks on Multi-Coil Reconstruction

As described previously, most multi-coil reconstruction algorithms utilize a form of data consistency with the acquired data $\mathbf{y}_\Omega$. For the aforementioned methods, including CG-SENSE, compressed sensing and physics-guided DL reconstruction, the consistency is enforced through $\mathbf{E}_\Omega^H \mathbf{y}_\Omega$. This is referred to as the zero-filled image, which will be denoted as

$$\mathbf{z}_\Omega = \mathbf{E}_\Omega^H \mathbf{y}_\Omega. \tag{4}$$

Therefore, we will consider multi-coil reconstruction algorithms $f(\cdot)$ as taking $\mathbf{z}_\Omega$ as input. We will utilize an l_∞ attack on the zero-filled image. The adversarial attack is performed on the zero-filled image instead of the fully-sampled image as done in [11], because the latter is not practical. First, one does not have access to fully-sampled images to generate a practical attack. Furthermore, in multi-coil MRI, the encoding operator is not known exactly, but estimated via the estimation of coil sensitivities [18]. Finally, the attack is performed in image space instead of the acquisition k-space, since it is difficult to define an l_∞-perturbation in k-space due to the varying signal magnitudes between central and outer k-space.

The l_∞ attack is a small additive perturbation $\mathbf{r}$ that satisfies $||\mathbf{r}||_\infty < \epsilon$, where ϵ is a small scalar. The impact of this an attack is evaluated by how much $f(\mathbf{z}_\Omega + \mathbf{r})$ deviates from $f(\mathbf{z}_\Omega)$. In our study, the fast gradient sign method (FGSM) was used to generate the perturbation via [7] :

$$\mathbf{r} = \epsilon \cdot \text{sign}(\nabla_{\mathbf{z}_\Omega} l(f(\mathbf{z}_\Omega), \mathbf{x})) \tag{5}$$

where $\text{sign}(\cdot)$ takes sign of inputs, $\nabla_{\mathbf{z}_\Omega}$ denotes the gradient respect to $\mathbf{z}_\Omega$. $l(\cdot)$ denotes the loss function. For the multi-coil reconstructions discussed in this study, $l(\cdot)$ was chosen as the MSE loss. The scalar ϵ is set to $||\mathbf{z}_\Omega||_\infty/255$ to ensure a sufficiently small perturbation level that causes no noticeable visual differences between $\mathbf{z}_\Omega$ and $\mathbf{z}_\Omega + \mathbf{r}$.

C. Experimental Setup

Several experiments were performed to assess the stability of multi-coil MRI reconstruction methods to small adversarial perturbations, using both uniform and random undersampling:

Image-domain based linear parallel imaging. CG-SENSE was used as the clinically relevant image-domain linear parallel imaging approach [18]. In order to implement CG-SENSE for $f(\cdot)$, the CG algorithm was unrolled for 10 conjugate gradient steps, in order to obtain $\nabla_{\mathbf{z}_\Omega} l(f(\mathbf{z}_\Omega), \mathbf{x})$ via backpropagation (Fig.1). Separate adversarial attacks were generated for uniform and random undersampling patterns. Additionally,

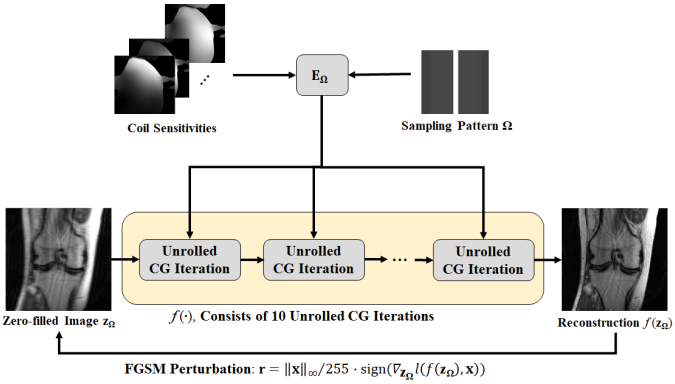

Fig. 1. Generation of FGSM perturbations for multi-coil MRI data using CG-SENSE reconstruction. CG-SENSE is unrolled for 10 steps. The gradient with respect to the input for FGSM is calculated using backpropagation and the multi-coil encoding operator, which includes coil sensitivity and undersampling pattern information.

CG-SENSE reconstructions were performed with different levels of Tikhonov regularization to study the effect of changing the condition number of the inverted linear system.

k-space based linear parallel imaging. For uniform undersampling, the performance of GRAPPA against adversarial attacks was also investigated. In order to do so, the l_∞ attack generated for CG-SENSE in the image domain was transformed into multi-coil k-space. As there are infinitely many k-space perturbations on Ω that lead to the same $\mathbf{r}$, the minimum ℓ_2 solution was picked:

$$\mathbf{v} = (\mathbf{E}_\Omega \mathbf{E}_\Omega^H)^{-1} \mathbf{E}_\Omega \mathbf{r}. \tag{6}$$

5×4 linear convolutional kernels, calibrated on auto-calibration signal (ACS) without perturbation, were used for GRAPPA reconstruction [19].

Multi-coil compressed sensing. The stability of multi-coil compressed sensing reconstruction was investigated using the adversarial attack generated on the CG-SENSE algorithm for random undersampling patterns. Weighted ℓ_1 norm of the Daubechies-4 wavelet transform was used as the regularizer [20]. The objective function in (3) was solved using variable splitting with quadratic penalty. The quadratic penalty and soft-thresholding parameters were optimized heuristically for optimal visible performance in the unperturbed setting.

Physics-guided DL reconstruction. Algorithm unrolling for variable splitting with quadratic penalty algorithm was used for physics-guided DL reconstruction [1–3] leading to two sub-problems

$$\mathbf{u}^{(i)} = \arg\min_{\mathbf{u}} \mu ||\mathbf{x}^{(i-1)} - \mathbf{u}||_2^2 + \mathcal{R}(\mathbf{u}) \tag{7}$$

$$\mathbf{x}^{(i)} = (\mathbf{E}_\Omega^H \mathbf{E}_\Omega + \mu \mathbf{I})^{-1} (\mathbf{z}_\Omega + \mu \mathbf{u}^{(i)}), \tag{8}$$

where $\mathbf{u}^{(i)}$ and $\mathbf{x}^{(i)}$ are an auxiliary variable and an image estimate at the i^{th} iteration respectively, and μ is a learnable quadratic relaxation parameter. In this setup, the proximal operation for the regularizer in (7) was implicitly implemented using a convolutional neural network, which was based on a ResNet [3]. The data consistency step in (8) was implemented

using the CG algorithm [2]. This unrolled network was trained in a supervised manner using a normalized ℓ_1-ℓ_2 loss in k-space [3] over 300 imaging slices, using Adam optimizer, learning rate = 0.001, 100 epochs. FGSM perturbation was generated accordingly with the same normalized ℓ_1-ℓ_2 using the full unrolled network.

D. Imaging Data

Multi-coil raw k-space data of knee and brain MRI scans from the fastMRI database [22] were employed in this study. Coronal-PD knee and FLAIR brain imaging were performed on 3T systems (Magnetom Skyra; Siemens, Erlangen, Germany) with a 15-channel knee coil and a 16-channel head coil respectively. Our primary investigations for uniform and random undersampling patterns used an acceleration rate (R) of 4 with 24 lines of ACS data in the center of k-space. Additionally, to test the effect of instabilities at more clinically relevant acceleration rates, adversarial perturbations were studied for uniform undersampling at $R = 2$ with 24 ACS lines. For uniform undersampling, CG-SENSE, GRAPPA and physics-guided DL were investigated, while for random undersampling, CG-SENSE and multi-coil compressed sensing were studied.

III. Results

Figure 2 depicts the results of uniform undersampling reconstructions at $R = 4$ using CG-SENSE and GRAPPA. Without adversarial perturbations, both methods provided stable reconstructions albeit with residual aliasing artifacts due to the higher acceleration rate, with GRAPPA visibly outperforming CG-SENSE. As expected, the additive small adversarial perturbation led no visually noticeable differences compared to the original fully-sampled or zero-filled images. However, both CG-SENSE and GRAPPA failed under the attack, with visible artifacts in both reconstruction results.

Figure 3 shows the reconstruction results of $R = 4$ random undersampling using CG-SENSE and multi-coil compressed sensing with similar observations. Without perturbation, both methods successfully reconstruct the image, with compressed sensing displaying lower reconstruction noise as expected.

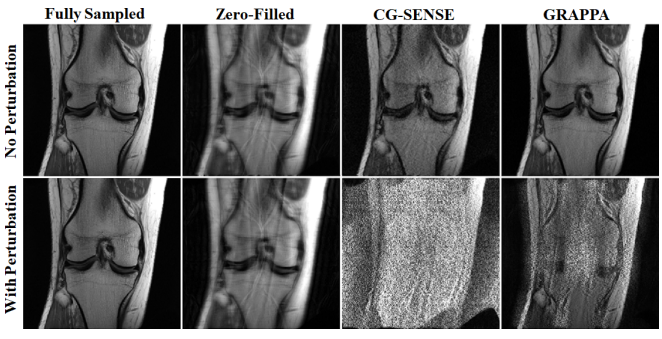

Fig. 2. Reconstruction results for $R = 4$ uniform undersampling with CG-SENSE and GRAPPA. Without perturbation, both methods work as expected, albeit with some artifacts due to the higher acceleration rate. The adversarial perturbation leads to no visual differences compared at the input, while causing both methods to fail at the output.

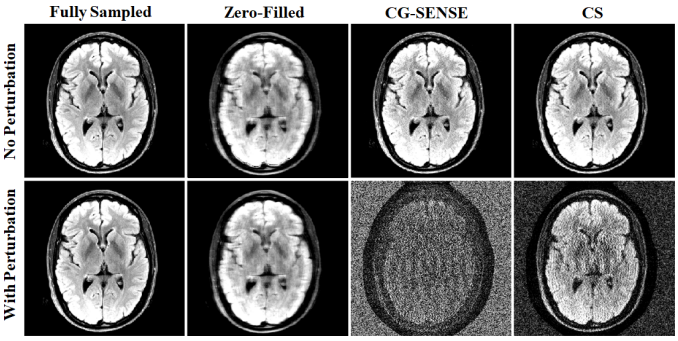

Fig. 3. CG-SENSE and compressed sensing (CS) results for $R = 4$ random undersampling. Without perturbations, both methods lead to good quality reconstructions. With the small adversarial perturbation, both CG-SENSE and CS fail with visible artifacts.

Adversarial perturbations caused no visual differences compared to the fully-sampled reference and zero-filled images. However, following this small adversarial perturbation, both method failed with visible artifacts.

Figure 4 depicts the Tikhonov-regularized CG-SENSE results for $R = 4$ uniform undersampling with and without the adversarial attack. With increasing regularization factor, the Tikhonov-regularized CG-SENSE demonstrated improved robustness against the attack. However, as expected, this led to a failure to unalias the image with increasing regularization, resulting in blurrier images.

Figure 5 shows reconstructions using physics-guided DL reconstruction with $R = 4$ uniform undersampling. Physics-guided DL successfully reconstructed the unperturbed image, offering higher quality than the linear methods in Figure 2. With the adversarial attacks, the DL reconstruction failed, similar to the reports in [11] for single-coil data. To further probe which part of the unrolled network is targeted by the adversarial perturbation, its effect was explored using a single-pass through the CNN regularizer and the data consistency units. There is no major change when the attack is run through the CNN regularizer, but the output shows major artifacts when the attack is passed through the data-consistency units, suggesting the end-to-end attack may target the linear data consistency unit even if generated end-to-end.

Finally, Figure 6 depicts the reconstruction results of the more clinically relevant acceleration of $R = 2$ uniform

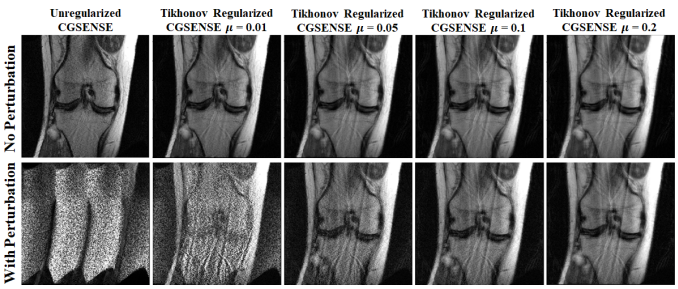

Fig. 4. Tikhonov-regularized CG-SENSE results for $R = 4$ uniform undersampling. Tikhonov-regularized CG-SENSE demonstrate improved robustness against the attack with increasing regularization strength, albeit at the cost of failing to dealias the image for higher μ even without perturbations.

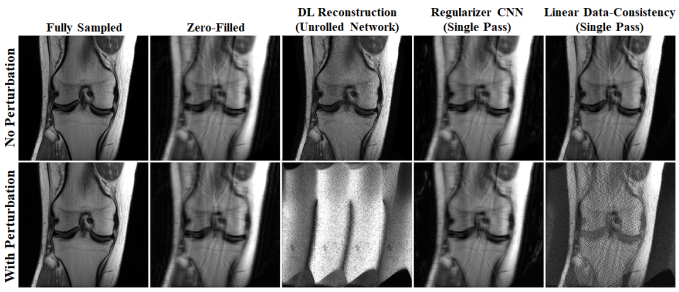

Fig. 5. Reconstruction results for $R = 4$ uniform undersampling with physics-guided DL reconstruction. The DL method successfully reconstructs the multi-coil MRI data without perturbation, but fails under the attack. The single-pass results show that the CNN regularizer has minor changes under the attach, while the linear data consistency unit has been more substantially impacted by the attack.

undersampling, using CG-SENSE and GRAPPA. Unlike the results at $R = 4$, both CG-SENSE and GRAPPA are robust to adversarial attacks at this rate, with only minor differences in the results with and without perturbation. This suggests that the adversarial attacks for conventional algorithms primarily affect the reconstruction when the forward encoding operator is highly ill-conditioned, which depends both on the acceleration rate and coil configuratoin.

IV. Discussion and Conclusions

In this work, we investigated the effect of small adversarial perturbation on multi-coil MRI reconstruction strategies, including clinically used parallel imaging methods. Uniform and random undersampling patterns were investigated on multi-coil knee and brain MRI data. The generated additive perturbations lead no visual differences compared to the original image, while causing significant reconstruction failures for all tested methods at high acceleration rates. Our results demonstrate that for multi-coil MRI datasets, conventional reconstruction strategies, such as parallel imaging and multi-coil compressed sensing are also susceptible to large instabilities from small additive adversarial perturbations. For physics-guided DL reconstruction that utilize the forward encoding operator explicitly, our results suggest that adversarial attacks predominantly affect the linear data consistency units.

Ill-conditioning of the multi-coil encoding operator is a well-discussed topic for CG-SENSE in non-Cartesian acquisitions, often leading to an early stopping criterion in practice [23]. For multi-coil MRI encoding operators, the condition number depends on coil configuration and R, and is hard to compute in general. However, our results suggest that adversarial attacks enable a method to exploit this ill-conditioning. These observations are further supported by the results at $R = 2$ for parallel imaging compared to $R = 4$. At the more clinically used rate of 2, both CG-SENSE and GRAPPA are robust to the small adversarial perturbations, suggesting $\mathbf{E}_\Omega$ is well-conditioned at this rate. At the higher rate of 4, the multi-coil encoding operator is more ill-conditioned as expected, which is exploited by the adversarial attack.

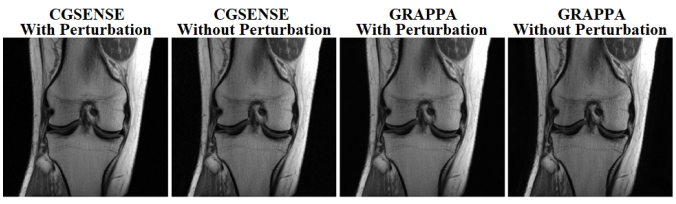

Fig. 6. CG-SENSE and GRAPPA reconstructions for the more clinically relevant $R = 2$ uniform undersampling. Both methods provide high-quality reconstruction with or without perturbations, with the adversarial perturbations causing only minor differences, due to the improved conditioning of the multi-coil encoding operator at this lower acceleration rate.

In a similar manner, higher degree of Tikhonov regularization, which helps control the ill-conditioning of the inverted linear system, reduces the effect of the adversarial attacks, albeit at the cost of worse reconstruction performance. However, for regularized reconstructions that also invert a penalized linear system as in (8), this suggests a higher μ value is desirable. Especially given that several local minima exist for a typical DL training optimization landscape [24, 25], it may be desirable to pick a solution with higher μ value to help stabilize the reconstruction further. Further work is warranted to test the feasibility of this approach in practice.

Finally, while the instability of DL reconstruction methods for MRI has generated significant interest, it is worthwhile to interpret these in the broader context, especially for multi-coil MRI datasets, where even conventional linear strategies exhibit instabilities at higher acceleration rates.

ACKNOWLEDGMENT

This work was partially supported by NIH P41EB015894, NIH U01EB025144, NIH P41EB027061, NSF CAREER CCF-1651825.

References

[1] K. Hammernik, T. Klatzer, et al., "Learning a variational network for reconstruction of accelerated MRI data," *Magn Reson Med*, vol. 79, pp. 3055–3071, 2018.

[2] H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Trans Med Imag*, vol. 38, no. 2, pp. 394–405, 02 2019.

[3] B. Yaman, S. A. H. Hosseini, et al., "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magn Reson Med*, vol. 84, no. 6, pp. 3172–3191, 12 2020.

[4] S. A. H. Hosseini, B. Yaman, et al., "Dense recurrent neural networks for accelerated MRI: history-cognizant unrolling of optimization algorithms," *IEEE J Sel Top Sig Proc*, vol. 14, no. 6, pp. 1280–1291, Oct 2020.

[5] M. J. Muckley, B. Riemenschneider, et al., "Results of the 2020 fastmri challenge for machine learning mr image reconstruction," *IEEE Trans Med Imaging*, vol. 40, no. 9, pp. 2306–2317, 2021.

[6] F. Knoll, K. Hammernik, et al., "Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues,"

IEEE Signal Process Mag, vol. 37, no. 1, pp. 128–140, 2020.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[8] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, and Deepfool, "A simple and accurate method to fool deep neural networks," in *Proceedings of the CVPR*, pp. 2574–2582.

[9] A. Madry, A. Makelov, et al., "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[10] C. Szegedy, W. Zaremba, et al., "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[11] V. Antun, F. Renna, et al., "On instabilities of deep learning in image reconstruction and the potential costs of ai," *PNAS*, vol. 117, no. 48, pp. 30088–30095, 2020.

[12] B. Zhu, J. Z. Liu, et al., "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[13] G. Yang, S. Yu, et al., "Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction," *IEEE Trans Med Imaging*, vol. 37, no. 6, pp. 1310–1321, 2017.

[14] J. Schlemper, J. Caballero, et al., "A deep cascade of convolutional neural networks for mr image reconstruction," in *Proceedings of IPMI*. Springer, 2017, pp. 647–658.

[15] K. Cheng, F. Calivá, et al., "Addressing the false negative problem of deep learning mri reconstruction models by adversarial attacks and robust training," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 121–135.

[16] F. Calivá, K. Cheng, R. Shah, and V. Pedoia, "Adversarial robust training of deep learning mri reconstruction models," *arXiv preprint arXiv:2011.00070*, 2020.

[17] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "Sense: sensitivity encoding for fast mri," *Magn Reson Med*, vol. 42, no. 5, pp. 952–962, 1999.

[18] K. P. Pruessmann, M. Weiger, P. Börnert, and P. Boesiger, "Advances in sensitivity encoding with arbitrary k-space trajectories," *Magn Reson Med*, vol. 46, no. 4, pp. 638–651, Oct 2001.

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

[20] 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, Dec 2007.

[21] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process Mag*, vol. 38, no. 2, pp. 18–44, 2021.

[22] F. Knoll, J. Zbontar, et al., "fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning," *Radiology: Artificial Intelligence*, vol. 2, no. 1, pp. e190007, 2020.

[23] O. Maier, S. H. Baete, et al., "Cg-sense revisited: Results from the first ismrm reproducibility challenge," *Magn Reson Med*, vol. 85, no. 4, pp. 1821–1839, 2021.

[24] K. Kawaguchi, "Deep learning without poor local minima," in *Proceedings of NIPS*, 2016, vol. 29.

[25] A. Choromanska, M. Henaff, et al., "The loss surfaces of multilayer networks," in *Artificial intelligence and statistics*. PMLR, 2015, pp. 192–204.