On Instabilities of Unsupervised Denoising Diffusion Models in Magnetic Resonance Imaging Reconstruction

Tianyu Han[1], Sven Nebelung[1], Firas Khader[1], Jakob Nikolas Kather[2,3,4], and Daniel Truhn[1]

[1] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany
[2] Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[3] Department of Medicine I, University Hospital Dresden, Dresden, Germany
[4] Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Abstract. Denoising diffusion models offer a promising approach to accelerating magnetic resonance imaging (MRI) and producing diagnostic-level images in an unsupervised manner. However, our study demonstrates that even tiny worst-case potential perturbations transferred from a surrogate model can cause these models to generate fake tissue structures that may mislead clinicians. The transferability of such worst-case perturbations indicates that the robustness of image reconstruction may be compromised due to MR system imperfections or other sources of noise. Moreover, at larger perturbation strengths, diffusion models exhibit Gaussian noise-like artifacts that are distinct from those observed in supervised models and are more challenging to detect. Our results highlight the vulnerability of current state-of-the-art diffusion-based reconstruction models to possible worst-case perturbations and underscore the need for further research to improve their robustness and reliability in clinical settings.

Keywords: Magnetic Resonance Imaging · Image Reconstruction · Denoising Diffusion Models.

1 Introduction

Magnetic Resonance Imaging (MRI) is essential for medical diagnostics, especially for brain diseases, due to its detailed, non-invasive imaging capabilities. However, MRI faces challenges like long acquisition times and high sensitivity to motion. Recent advancements, particularly denoising diffusion models, promise to accelerate MRI by reconstructing high-quality images from undersampled data. Unlike traditional methods, these models can operate without paired training data. However, our study reveals a critical vulnerability: susceptibility to minimal worst-case perturbations, leading to significant inaccuracies

in reconstructed images. Our research explores the robustness of diffusion models in MRI reconstruction, investigating adversarial perturbations and proposing strategies to enhance resilience. We aim to advance reliable diffusion models in clinical settings.

2 Related Works

2.1 DL-based end-to-end solution

Model-based image reconstruction methods consider imaging systems as a linear operator $\mathbf{A}$ that maps anatomical ground truth to the signal domain. Specifically, a noisy observation $\mathbf{y}$ given by m sparse measurements can be defined as

$$\mathbf{y} \in \mathbb{R}^m = \mathbf{A}\mathbf{x} + \epsilon, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the unknown, $\mathbf{A} \in \mathbb{R}^{m \times n}$ $(n > m)$ denotes an operator randomly samples k-space data, and $\epsilon \in \mathbb{R}^m$ is the measurement noise. When using standard Cartesian acquisition, we can factorize operator $\mathbf{A}$ into an operator $\mathcal{P}(\Lambda)$ and an invertible matrix $T \in \mathbb{R}^{n \times n}$ that corresponds to Fourier transform: $\mathbf{A} = \mathcal{P}(\Lambda)T$. K-space lines are selected during acquisition if $\Lambda_{ii} = 1$ in $\mathcal{P}(\Lambda)$, where $\Lambda \in \{0, 1\}^{n \times n}$ is a diagonal matrix with $tr(\Lambda) = m$.

Given the rise of DL, it's a natural flow to approximate the inverse model using convolutional neural networks (CNNs). In the CNN formulation, one force $\mathbf{x}$ to be well-approximated by the CNN reconstruction by using the following objective:

$$\min_{\theta} . \, \lambda \left\| \mathbf{x} - f_{\text{cnn}}(\mathbf{A}^H \mathbf{y} \,|\, \theta) \right\|_2^2 + \left\| \mathbf{A}\mathbf{x} - \mathbf{y} \right\|_2^2 . \tag{2}$$

Here, we denote $f_{\text{cnn}}(\cdot \,|\, \theta)$ as a CNN parameterized by θ. The CNN reconstruction can be considered as resolving a de-aliasing issue in the spatial domain because $\mathbf{A}^H \mathbf{y}$ is severely affected by aliasing from sub-Nyquist sampling. However, the performance of directly optimizing Equation 2 is inadequate since the CNN reconstruction and the data fidelity are optimized independently. The CNN is specifically trained to reconstruct the sequence without knowing the prior details of the obtained data in k-space because it works purely in the image domain.

2.2 Bayesian image reconstruction

In the Bayesian picture, the MRI measurement $\mathbf{y}$ and tissue signal $\mathbf{x}$ are coupled by a measurement distribution in this probabilistic formulation: $p(\mathbf{y} \,|\, \mathbf{x}) = q_\epsilon(\mathbf{y} - \mathbf{A}\mathbf{x})$, where $q_\epsilon(\cdot)$ stands for the noise distribution. The conditional distribution $p(\mathbf{y} \,|\, \mathbf{x})$ represents a forward process of measuring $\mathbf{y}$ from $\mathbf{x}$, which is also described by the linear forward model (Equation 1). The reconstruction problem is then viewed as drawing samples from the posterior distribution $p(\mathbf{x} \,|\, \mathbf{y})$. In general, we can obtain such a posterior using Bayes' theorem: $p(\mathbf{x} \,|\, \mathbf{y}) = p(\mathbf{y} \,|\, \mathbf{x})p(\mathbf{x})/p(\mathbf{y})$. The following Bayes' rule for score functions results from taking gradients concerning x on both sides of this expression: $\nabla_{\mathbf{x}} \log p(\mathbf{x} \,|\, \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{y} \,|\, \mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{x})$. Note, the data prior term

$\nabla_{\mathbf{x}} \log p(\mathbf{x})$ can be efficiently estimated by a denoising diffusion model [4, 13]. Incorporating measured observation $\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ into our system is an essential step that transforms an unguided diffusion into a conditional one. In general, one can add such a correction term to unconditional diffusion steps either via approximating the likelihood gradient or directly performing closed-form data consistency.

2.3 Reconstructing MRI using diffusion prior and posterior sampling

Given a time-dependent diffusion-based model $s_{\theta^*}(\mathbf{x}_t, t)$ that has been trained to approximate the data score function via a diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ that was produced by perturbing $\mathbf{x}$ with an SDE. The procedure of unconditional sampling chooses a series of time steps and iterates by $\hat{x}_{t_{i-1}} = h(\hat{x}_{t_i}, z_i, s_{\theta^*}(\hat{\mathbf{x}}_t, t_i))$, where function h represents an SDE solver and $z_i \sim \mathcal{N}(0, I)$. In MRI reconstruction, function h needs an additional step k prepended to itself to impose the constraint implied by measurements, resulting in

$$\begin{aligned} \hat{x}'_{t_i} &= k(\hat{x}_{t_i}, \hat{y}_{t_i}, \lambda) \\ \hat{x}_{t_{i-1}} &= h(\hat{x}'_{t_i}, z_i, s_{\theta^*}(\hat{\mathbf{x}}_t, t_i)). \end{aligned} \tag{3}$$

The above k function interacts with measured k-space entries and thus should follow an image fidelity objective and a k-space consistency objective:

$$\hat{x}'_t = T^{-1} \left[\lambda \Lambda P^{-1}(\Lambda) \hat{y}_t + (1 - \lambda) \Lambda T \hat{x}_t + (I - \Lambda) T \hat{x}_t \right]. \tag{4}$$

3 Worst-Case Instabilities in MRI reconstruction

As demonstrated in Fig. 1, our worst-case noise is designed to be small in k-space but can induce a significant mismatch in the slice recovered by the undersampled version of the perturbed ground truth. Let f: $\mathbb{R}^m \to \mathbb{R}^n$ be a trained neural network f mapping an undersampled measurement to an image. Finding an adversarial direction $\delta \in \mathbb{R}^m$ in the measurement domain can be viewed as solving an optimization with the following objective:

$$\max_{\delta : \|\delta\| \leq \epsilon} \mathbb{E}_{\mathbf{y}} \left[\|f(\mathbf{y}; \theta) - f(\mathbf{y} + \delta; \theta)\|_2^2 \right]. \tag{5}$$

Here, we confine allowed perturbation sets to be a hypersphere l_2 ball around any $\mathbf{y}$ with a norm ϵ, i.e., $\|\delta\|_2 \leq \epsilon \|\mathbf{y}\|_2$. Following [1], a projected gradient descent (PGD) method [7] was used to maximize the objective in Equation 5.

Both white- and black-box perturbations were evaluated in this study. In a white-box scenario, we have full access to the target reconstruction model θ and can perform gradient-based PGD attacks on its reconstruction. Second, in the black-box setting, we tested the success rate of transferring the above perturbations to the remaining models, e.g., diffusion-based reconstructions.

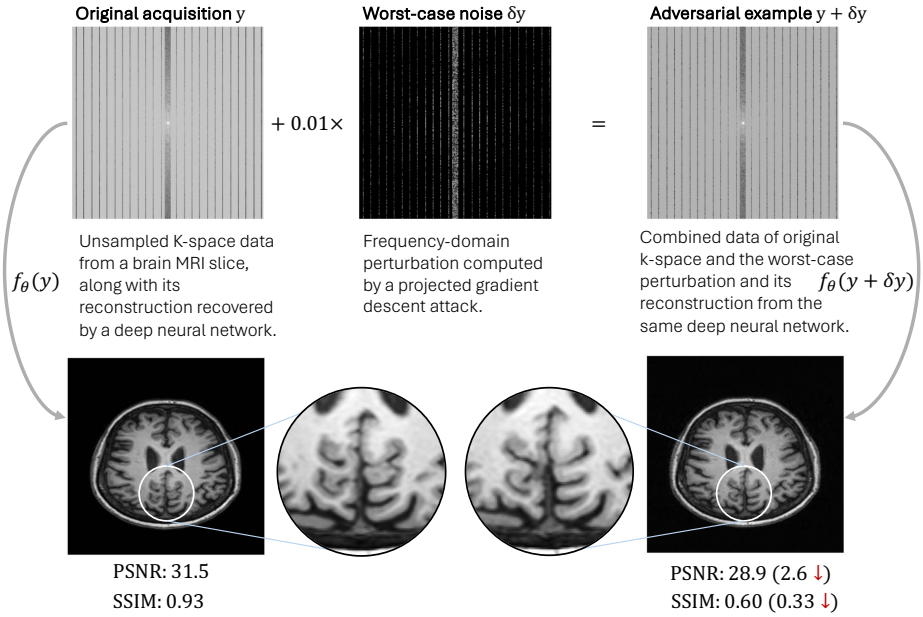

Fig. 1: MRI reconstruction can be vulnerable to worst-case perturbations, which add noise to the original k-space signal and manipulate the reconstruction process of undersampled data. The resulting reconstructions can show false gray matter structures that are difficult for humans to detect (see zoomed-in plots).

Algorithm 1 Worst-case perturbation in k-space

Require: A model with its current parameter θ: f_θ; Loss function: $L = -L_2$; Partially measured k-space: ksp; Acquisition mask: M

Require: Adam Optimizer Opt; Perturbation strength: ϵ; Number of iterations: T; Learning rate: α; Constant $c = 1e4$

Initialize perturbation: $\delta \leftarrow \delta_r + i \cdot \delta_i$, where $\delta_r, \delta_i \sim \mathcal{N}(0, I)$

$\delta \leftarrow \delta \times \frac{\|ksp\|}{\|\delta\| \times c}$

Initialize Optimizer: $Opt \leftarrow$ `Adam(param` $= [\delta]$, `lr` $= \alpha)$

Get standard reconstruction: $x \leftarrow f_\theta(ksp \odot M, M)$

for `i in range(T)` **do**

  $\hat{x} \leftarrow f_\theta((ksp + \delta) \odot M, M)$

  Opt.`zero_grad()`

  $loss \leftarrow L(x, \hat{x})$

  $loss$.`backward()`

  if $\|\delta\| > \epsilon$ **then**

    $\delta \leftarrow \delta \times \frac{\|ksp\|}{\|\delta\|} \cdot \epsilon$

  end if

end for

return δ

4 Experiments and Results

4.1 Data

All neuroimaging data were employed from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a multicenter, longitudinal study of 2,463 participants between the ages of 55 and 90 who had or were at high risk of developing dementia and Alzheimer's disease. Each participant has gone through T_1 weighted structure MRI measured by Magnetization Prepared RApid Gradient Echo (MP-RAGE) [8]. For our study, we utilized 80% of the dataset, corresponding to 1,970 participants and 13,651 scans, for training, and randomly selected 108 scans from the remaining 20% of ADNI participants (493 individuals) for testing. Using 80% of the dataset for training and validation follows a standard 80-20 train-validation split, ensuring robust model evaluation.

4.2 Supervised baselines

CNNs have established a new state-of-the-art MRI reconstruction, vastly beyond the traditional baselines. One typical approach utilizes auto-encoder architectures, such as U-Net [12, 15], which solves the medical inverse problem in an end-to-end fashion. We selected a Unet-based baseline (ResUnet++) as it is the most widely used CNN backbone in MRI image reconstruction. In experiments, we trained a ResUnet++ [5] model with a batch size of 16 and a learning rate of 0.001, using 50 epochs on ADNI training set. Another branch of models such as ADMM [14] and i-RIM [9, 10] generalizes the idea of iterative compressed sensing reconstruction that unrolls the data-flow graph via a cascade of neural networks. We selected i-RIM model due to its superior performance in vairous MRI reconstruction challenges, especially, the FastMRI challenge [9, 6]. In our experiments, the i-RIM model was trained using a batch size of 4 and a learning rate of 0.001 with 50 epochs. We trained both supervised models with an acceleration factor of 8.

4.3 Denoising diffusion reconstruction

Our training method is similar to Song et al [13]. Since the Predictor-Corrector (PC) sampler has generally higher performance for VE-SDEs, we employ it here in place of the numerical SDE solver to generate samples. The corrector in this PC sampler follows Langevin dynamics that solely rely on the scores, whereas the predictor refers to a numerical solver for the reverse-time SDE. We perform an additional data consistency [11, 13] step to adapt the PC sampler for solving inverse problems. 1,000 noise scales and 1 step of the Langevin correction for each noise scale were selected, resulting in a total of 2,000 steps of score model evaluation in the PC sampler. Besides, the signal-to-noise ratio (SNR) η controls the step size ϵ in Langevin dynamics. In our setting, η was set to 0.517 and the λ in the data consistency operation was set to 1.0.

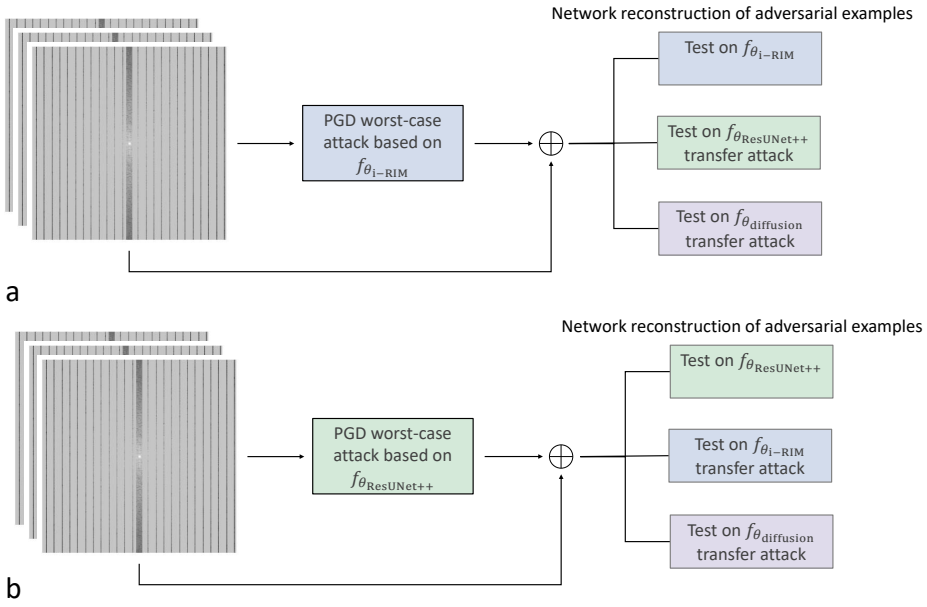

Fig. 2: We designed experiments to evaluate the susceptibility of trained i-RIM and ResUnet++ models to white- and black-box attacks (**a** and **b**).

4.4 Experimental design

First, we trained supervised and unsupervised models on reconstruction tasks in the training cohort and evaluated the performance in the test cohort with three acceleration schemes. We demonstrated diffusion models deliver comparable performance to state-of-the-art supervised models while showing significantly better generalization to unknown acquisition processes. Then, as demonstrated in Fig. 2, we assessed the robustness of the trained models against white- and black-box adversarial perturbations. Both scenarios are important as a white-box attack corresponds to malfunctions of an internal MRI system, while black-box perturbations reveal vulnerabilities of current deep-learning reconstruction towards possible external adversarial interferences.

4.5 Quantitative and visual evaluation

Worst-case instabilities of supervised models. The results of our experiments are presented in Fig. 3, which shows the SSIM loss as a function of perturbation strength. Subplots c and d of Fig. 3 show the white-box perturbations [3] obtained by attacking i-RIM and ResUnet++ models, respectively. We found that both i-RIM (the blue line in subplot (a)) and ResUnet++ (the green line in subplot (b)) are unstable and can be easily biased by tiny adversarial perturbations.

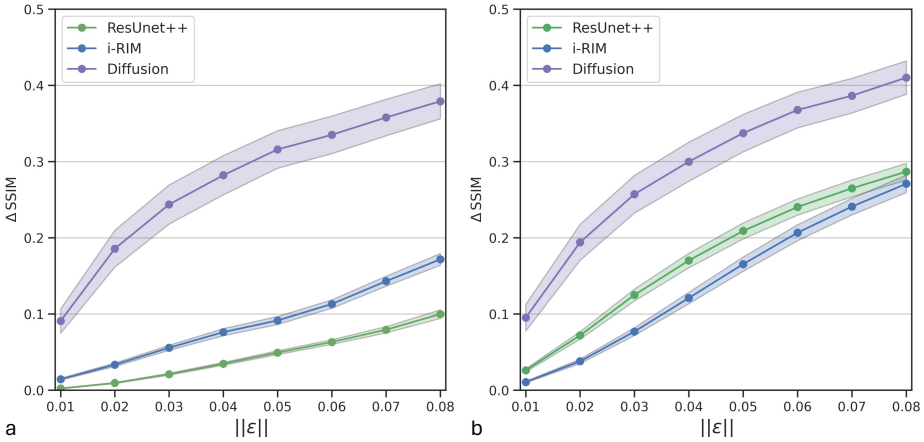

Fig. 3: We visualized the impact of perturbation amplitude on model performance, measured by the ΔSSIM metric. Subplot (a) shows that all models experienced a drastic drop in SSIM as the perturbation amplitude increased using worst-case perturbations generated by i-RIM. Similar findings were observed with adversarial perturbations via the ResUnet model, in (b).

Worst-case transferability to diffusion models. Empirical evidence for the transferability of adversarial examples has been investigated in classification applications, but rarely demonstrated in regression tasks such as image reconstruction. In Fig. 2a, we explore a black-box scenario where adversarial perturbations are generated against a surrogate model, such as an i-RIM, rather than the actual models used for reconstruction, such as diffusion. Similar to supervised models, as shown in Figures 3, the unsupervised diffusion model is also susceptible to worst-case distribution shifts in the form of adversarial perturbations. Even with a minimal adversarial perturbation of $\epsilon = 0.01$ crafted from ResUnet++ parameters, the diffusion model distorts the gray matter structure in its reconstruction (Fig. 4). This result is significant because diffusion models are trained similarly to denoisers, as evidenced by the appearance of a Gaussian-noise-like artifact when we adversarially perturb them (Fig. 4).

5 Conclusion

In summary, our study highlights the vulnerability of both state-of-the-art supervised models and unsupervised diffusion models to adversarial perturbations from the MRI signal domain. We found that worst-case perturbations can effectively transfer between independently trained regression models, similar to the transferability observed in classification tasks. While diffusion models are generally robust against anatomical and test-time distribution shifts, our findings indicate that even tiny adversarial perturbations can cause these models to generate fake tissue structures that may mislead clinicians. Furthermore, at

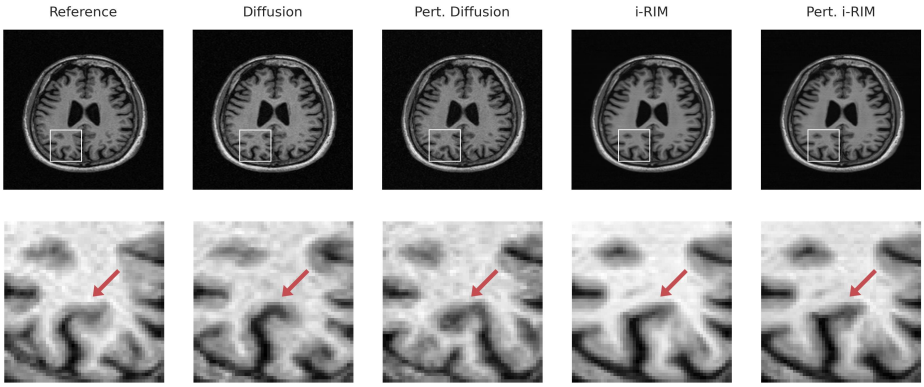

Fig. 4: To demonstrate, we crafted worst-case black-box perturbations [2] using an independent ResUnet++ model and applied them to unsupervised diffusion reconstruction and supervised i-RIM. The application of worst-case inference to unsupervised reconstruction can create misleading artifacts in brain tissue, which can be seen as red arrows in the subplot below.

larger perturbation amplitudes, diffusion models exhibit noise-like artifacts that are distinct from those observed in supervised models and may be more difficult for clinicians to detect.

We hypothesize that the main reason for this vulnerability is due to the perturbed K-space misleads the reverse iterative diffusion process, creating non-physical artifacts. Classical regularization techniques like total variance regularization might offer better robustness in such scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Darestani, M.Z., Chaudhari, A.S., Heckel, R.: Measuring robustness in deep learning based compressive sensing. In: International Conference on Machine Learning. pp. 2433–2444. PMLR (2021)
2. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., et al.: Adversarial attacks and adversarial robustness in computational pathology. Nature Communications **13**(1), 5711 (2022)
3. Han, T., Nebelung, S., Pedersoli, F., Zimmermann, M., Schulze-Hagen, M., Ho, M., Haarburger, C., Kiessling, F., Kuhl, C., Schulz, V., et al.: Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. Nature communications **12**(1), 4315 (2021)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
5. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). pp. 225–2255. IEEE (2019)
6. Knoll, F., Murrell, T., Sriram, A., Yakubova, N., Zbontar, J., Rabbat, M., Defazio, A., Muckley, M.J., Sodickson, D.K., Zitnick, C.L., et al.: Advancing machine learning for mr image reconstruction with an open competition: Overview of the 2019 fastmri challenge. Magnetic resonance in medicine **84**(6), 3054–3070 (2020)
7. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
8. Mugler III, J.P., Brookeman, J.R.: Three-dimensional magnetization-prepared rapid gradient-echo imaging (3d mp rage). Magnetic resonance in medicine **15**(1), 152–157 (1990)
9. Putzky, P., Karkalousos, D., Teuwen, J., Miriakov, N., Bakker, B., Caan, M., Welling, M.: i-rim applied to the fastmri challenge. arXiv preprint arXiv:1910.08952 (2019)
10. Putzky, P., Welling, M.: Invert to learn to invert. Advances in neural information processing systems **32** (2019)
11. Qin, C., Schlemper, J., Caballero, J., Price, A.N., Hajnal, J.V., Rueckert, D.: Convolutional recurrent neural networks for dynamic mr image reconstruction. IEEE transactions on medical imaging **38**(1), 280–290 (2018)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
13. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
14. Sun, J., Li, H., Xu, Z., et al.: Deep admm-net for compressive sensing mri. Advances in neural information processing systems **29** (2016)
15. Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al.: fastmri: An open dataset and benchmarks for accelerated mri. arXiv preprint arXiv:1811.08839 (2018)