Evaluation of the robustness of learned MR image reconstruction to systematic deviations between training and test data for the models from the fastMRI challenge

Patricia M. Johnson[1], Geunu Jeong[2], Kerstin Hammernik[3,14], Jo Schlemper[4], Chen Qin[3,15], Jinming Duan[3,16], Daniel Rueckert[3,14], Jingu Lee[2], Nicola Pezzoti[5], Elwin De Weerdt[10], Sahar Yousefi[6], Mohamed S. Elmahdy[6], Jeroen Hendrikus Franciscus Van Gemert[10], Chistophe Schuelke[11], Mariya Doneva[11], Tim Nielsen[11], Sergey Kastryulin[12], Boudewijn P. F. Lelieveldt[6], Matthias J. P. Van Osch[6], Marius Staring[6], Eric Z. Chen[7], Puyang Wang[8], Xiao Chen[7], Terrence Chen[7], Vishal M. Patel[8], Shanhui Sun[7], Hyungseob Shin[13], Yohan Jun[13], Taejoon Eo[13], Sewon Kim[13], Taeseong Kim[13], Dosik Hwang[13], Patrick Putzky[17], Dimitrios Karkalousos[18], Jonas Teuwen[19] Nikita Miriakov[19] Bart Bakker[20] Matthan Caan[18] Max Welling[17] Matthew J. Muckley[9], and Florian Knoll[1]

[1] Department of Radiology, NYU Langone Health, New York, NY, USA
[2] AIRS Medical, Seoul, Korea
[3] Department of Computing, Imperial College London, London, UK
[4] Hyperfine Research Inc., Guilford, CT, USA
[5] Philips Research, Eindhoven, The Netherlands
[6] Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands
[7] United Imaging Intelligence, Cambridge USA
[8] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD,USA
[9] Facebook AI Research, New York, NY, USA
[10] Philips Healthcare, Best, The Netherlands
[11] Philips Research, 22335 Hamburg, Germany
[12] Philips Research, 121205 Moscow, Russia
[13] Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
[14] AI in Healthcare and Medicine, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany
[15] Institute for Digital Communications, School of Engineering, University of Edinburgh, Edinburgh, UK
[16] School of Computer Science, University of Birmingham, Birmingham, UK
[17] Amsterdam Machine Learning Lab, University of Amsterdam, The Netherlands
[18] Dept. of Biomedical Engineering and Physics, University of Amsterdam, The Netherlands
[19] Radboud University Medical Center, Netherlands Cancer Institute
[20] Philips Research, The Netherlands

Preprint. See published version at https://doi.org/10.1007/978-3-030-88552-6_3.

Abstract. The 2019 fastMRI challenge was an open challenge designed to advance research in the field of machine learning for MR image reconstruction. The goal for the participants was to reconstruct undersampled MRI k-space data. The original challenge left an open question as to how well the reconstruction methods will perform in the setting where there is a systematic difference between training and test data. In this work we tested the generalization performance of the submissions with respect to various perturbations, and despite differences in model architecture and training, all of the methods perform very similarly.

1 Introduction

The goal of the 2019 fastMRI challenge [1] was to evaluate the performance of current state of the art machine learning methods for MR image reconstruction [2–11] on a large-scale standardized dataset [12, 13]. The goal for the participants was to reconstruct images from undersampled MRI k-space data, and the challenge consisted of three tracks: For 2 tracks, we provided the original multi-receive channel data from MR scanners, and the tracks differed in the undersampling rate (R=4 and R=8). In the third track, we provided simulated single-channel data, which reduced the computational burden and reduced the learning curve of working with MR data. We received 33 total submissions, and during the course of the challenge, we identified the top 4 submissions per track in terms of structural similarity [14] to the fully sampled ground truth as finalists. These results of these top 4 finalists were then evaluated in a second stage of the challenge by clinical radiologists, which ultimately determined the winners of the challenge. An overview of example results from the finalists of each track is given in figure 1.

The winners presented their methods during the 2019 NeurIPS conference, and a detailed description of the structure of the challenge, the evaluation criteria and the results, are presented in [1]. The challenge prompted several follow up questions that were out of scope of the original evaluation. In particular, the challenge design did not include any type of domain shift between the training data and the challenge test data. Therefore, the generalization performance of the submissions with respect to perturbations was not tested. The presentations at the 2019 NeurIPS conference also raised the interesting point that in some test examples, despite providing visually and quantitatively impressive results, the reconstructions removed important image features.

In this work we systematically investigate the robustness of the approaches of the challenge finalists with respect to representative domain shifts in the data, which could occur in real-world clinical use. To this extent, we added certain perturbations to the challenge test set, and the finalists re-ran the models from their submissions without re-training. We also provide a more detailed follow-up analysis of the examples where pathology was missed in the original submissions.

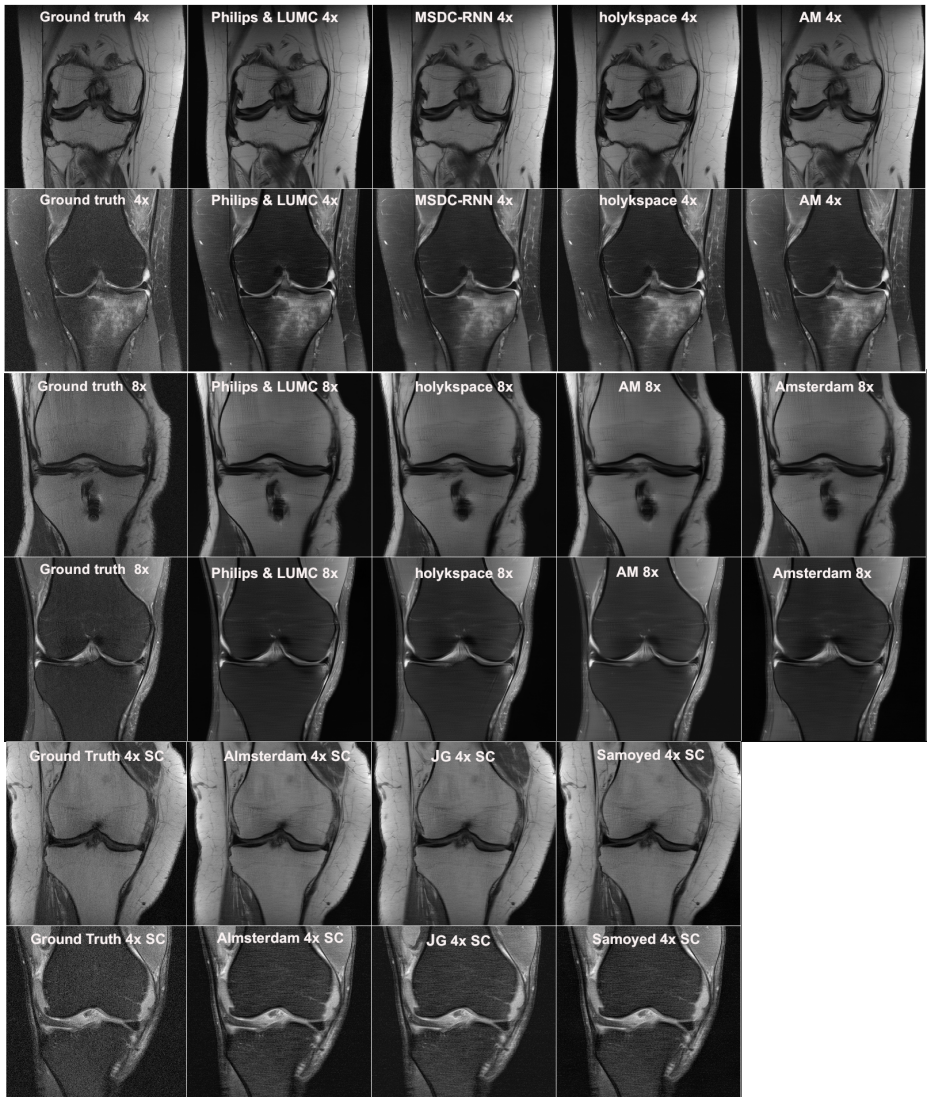

Fig. 1. Image results for the 2019 fastMRI knee challenge. Example results for the multicoil track R=4, are shown in the top two rows, results for multicoil track R =8 are shown in the middle two rows, and single coil results are shown in the bottom rows.

2 Methods

2.1 Image perturbations

In the first set of experiments, we evaluate the response of the different methods to small structural changes in the images by in-painting objects to a proton density weighted image. We added squares with varying intensities (figure 2) and a resolution grid (figure 3), and then regenerate a simulated k-space from the perturbed image, which was reconstructed with each model.

To evaluate the effect of a mismatched SNR between the training and test data on network performance, we added noise to the input k-space of a fat-saturated data set. The noise level was estimated by calculating the standard deviation (σ) of voxel intensities from a slice with no anatomy. Two different amounts of noise were then added to the k-space. Gaussian noise with a standard deviation of 0.5σ and σ were added to the model input, simulating an SNR of 2/3 and 1/2 of the original SNR of the image.

The next experiment evaluates the robustness of each method when the input data has a different number of coils. We performed coil compression on the original 15 channel data, simulating 10 and 5 channel data. Coil compression was performed using the scipy SVD algorithm.

The final experiments explored two cases in which pathology was removed in the challenge reconstructions. For each case, the data were retrospectively under-sampled with two realizations of random under-sampling, including the original sampling pattern used in the challenge. The sampling patterns were consistent with R=4 or R=8 random under sampling commonly used in compressed sensing. In addition, we sampled 16 lines at the center of k-space.

2.2 Description of 2019 fastMRI approaches

The network from Philips & LUMC, referred to as Adaptive-CS-Network [15], is a deep cascade approach that builds on the ISTA-Net model using a multiscale regularizer. The model consists of 25 unrolled iterations, with different design in each iteration. It also includes the MR physics priors, such as the phase behaviour and background identification, which are provided to the model with a "nudge" approach. The model is trained using a Multiscale-SSIM combined with L1 loss, and sequential refinement on different data populations.

The model submitted by AIRS medical (labelled as AM and JG) is an Auto-Calibrating Deep Learning Network. It consists of a neural network block and an auto-calibration block, which were iteratively applied. The neural network block, based on the U-net, was trained in combined complex image space with l1 loss. The auto-calibration block enforced the null space constraint $\boldsymbol{N}x = 0$, where $\boldsymbol{N}$ is a convolution operator corresponding to the null space, via the conjugate gradient method [16]. After the multiple cascade of the blocks, a refinement U-net processed the complex data and generated magnitude images.

The model originally referred to as MSDC-RNN is a Pyramid Convolutional RNN (PC-RNN) model [17], which includes three convolutional RNN modules

to iteratively reconstruct images at different scales. The spatial sizes of feature maps in the three convolutional RNN modules are downsampled by 4x, 2x, 1x, respectively. Each convolutional RNN module has five iterations. The reconstructed images in coarse to fine scales are combined by a final CNN module. The model takes the multiple coils as multi-channel inputs and was trained with the NMSE loss [12] and the SSIM loss [18] on the coil combined images for 60 epochs.

Σ-Net [19] by holykspace (Imperial College London) ensembles multiple learned unrolled reconstruction networks. First, sensitivity networks involving explicit coil sensitivity maps were trained for a gradient descent, proximal gradient, and variable splitting scheme, followed by style transfer for further fine-tuning to the reference. Second, parallel coil networks were deployed to learn the coil combination implicitly. All networks, with a Down-Up Network as backbone [20], were unrolled for 10 steps. Training was conducted with a combined L1+SSIM loss, followed by individual fine-tuning for contrasts and acceleration factors. Additional networks trained with a GAN loss and a self-supervised approach complete the final ensemble of Σ-Net.

The Amsterdam submission is the i-RIM model, [21] an invertible variant of the RIM [22] for reconstructing accelerated MR-images [23]. The 480-layers model consists of 12 down-sampling blocks. Except the zero-filled reconstruction, a 1-hot vector was also given as input for encoding field-strength and fat-suppression meta-data. For singlecoil data we chose 64 feature layers and for multicoil 96. For multicoil data, k-space measurements and individual coil images were stacked, without sensitivity modeling in this context. We cropped the images to 368x368 pixels, and for smaller sizes we applied zero-padding. Finally, we used the Adam optimizer with learning-rate 10-4, and the SSIM as loss function.

The Samoyed model utilizes consecutive CNN blocks in the image domain for de-aliasing [24] [25], with interleaved data consistency layers that adopt trainable regularization parameters [4] [6]. Each CNN block comprises 5 dilated convolutional layers with 64 feature maps followed by Leaky Rectified Linear Units. Every feature map of the 4th convolutional layer in each block is concatenated to the feature map of the 2nd convolutional layer in the next block (i.e., dense-connection) to prevent information wash-out. The L1 loss and the SSIM loss were applied only to the foreground area, so that the learning could be focused on the anatomical area rather than the background area. The model was separately trained on each acquisition protocol due to SNR mismatch.

All the finalists methods make use of some sort of data-consistency, demonstrating the importance of leveraging the data early in the reconstruction chain compared to techniques that rely solely on the reconstructed images.

3 Results

The performance of the submissions with respect to small structural changes are shown in figures 2 and 3. The first perturbation, which consists of 6 in-painted

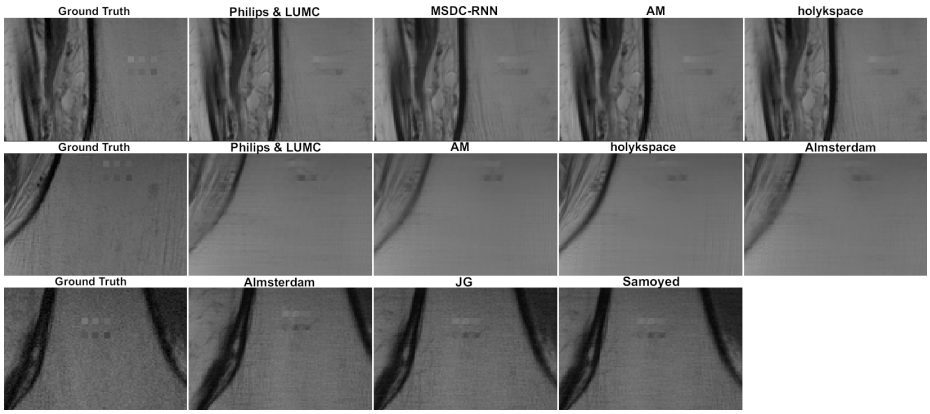

Fig. 2. Image results for contrast experiments. The perturbation is a set of in-painted squares with varying intensity. The R=4, R=8 and single coil results are shown in the top, middle and bottom rows respectively.

squares of varying intensities, appears blurred in all reconstructed images. This is consistent between tracks and for all submissions. In the multi-coil R=4 reconstructions (figure 2, top row) the highest and lowest signal intensity squares remain distinguishable. We observe blurring and ghosting of the in-painted resolution grids in the R=4 reconstructions shown in figure 3. This effect is more significant for the lower grid with vertical lines. In the R=8 reconstructions, the resolution grids are almost completely removed.

The performance of the submissions with respect to increased noise is shown in figure 4. A systematic mismatch in SNR between training and test data results in reconstructed images that appear over-smoothed. The Philips & LUMC reconstructions have lower SSIM but appear less smooth than reconstructions from other methods.

Most of the methods appear to be robust to a mismatch in the number of coils between the training and test data. The image quality is very similar for the reconstructions of 5 , 10 and 15 channel data. The exception is the AImsterdam model, for which sensitivity maps were not included during training, but rather each coil image was treated as a separate channel. Results of the coil compression experiments for R=8 are shown in figure 5.

Two cases from the original challenge where readers identified missing or less visible pathology in the reconstructed images were under-sampled with a different random sampling pattern with equivalent acceleration. The first case was a proton density weighted image with an acceleration of R=4 , the reconstructions using the original sampling pattern and new sampling pattern are shown in the first and second rows of figure 6, respectively. The pathology, indicated by the arrow in the ground truth image is more visible on the new reconstructions. The second case is a fat saturated image with acceleration R=8. The reconstructions

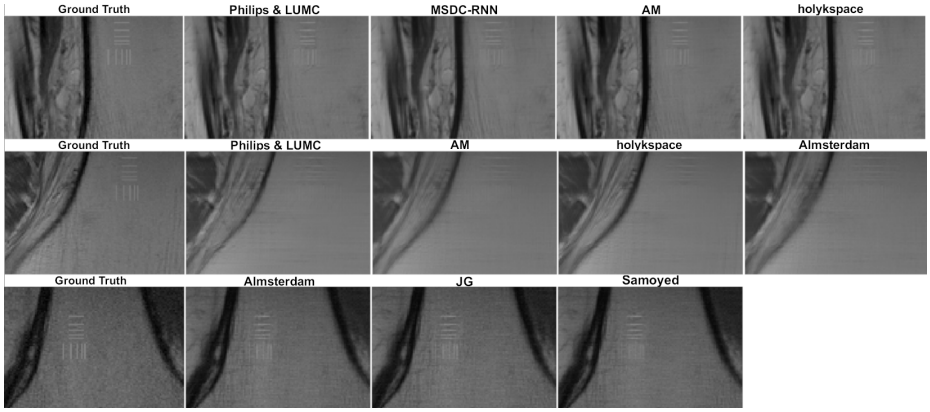

Fig. 3. Image results for resolution experiments. The perturbation is an in-painted resolution grid. The R=4, R=8 and single coil results are shown in the top, middle and bottom rows respectively

using the original and new sampling patterns are shown in the third and fourth row of 6 respectively. In this example the meniscal tear (indicated by the arrow in the ground truth image) is much more clear with the second sampling pattern than the original.

4 Discussion & Conclusion

In this work we evaluate the robustness of deep learning reconstruction methods submitted to the 2019 fastMRI challenge. All of these methods achieved high image quality but their robustness to possible domain shifts between training and test data remained an open question. In real world clinical use, the images may have unique features not seen in the training data, and they may vary in terms of SNR and coil configuration. Our results show that all of the methods remove small structures not seen in the training data and generate over smoothed images when the model input has lower SNR. They all appear to be robust to data with different coil configurations. We performed a follow-up analysis of the examples where pathology was missed in the original challenge reconstructions. Using a new realization of a random undersampling pattern with matched acceleration factors, the pathology was preserved. This suggests that the choice of sampling pattern can make a substantial difference in the clinical value of the image.

All DL reconstruction methods discussed in this work provide impressive results in the absence of a domain shift between training and test data. Our results show that all methods perform remarkably similarly in the presence of several perturbations despite differences in network architecture and training as well as showing the importance of the right sampling pattern for the reconstruction quality.

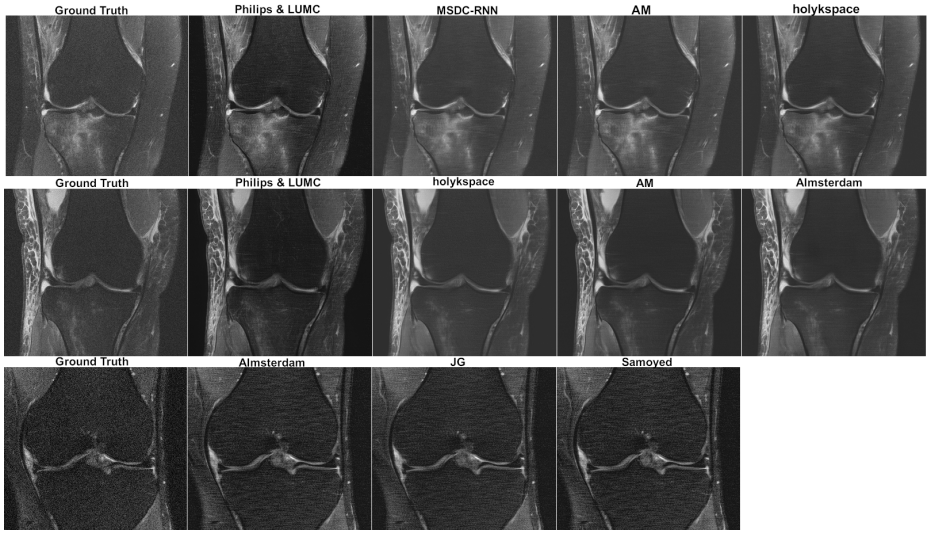

Fig. 4. Image results for noise experiments. Gaussian noise was added to the input k-space. The R=4, R=8 and single coil results are shown in the top, middle and bottom rows respectively

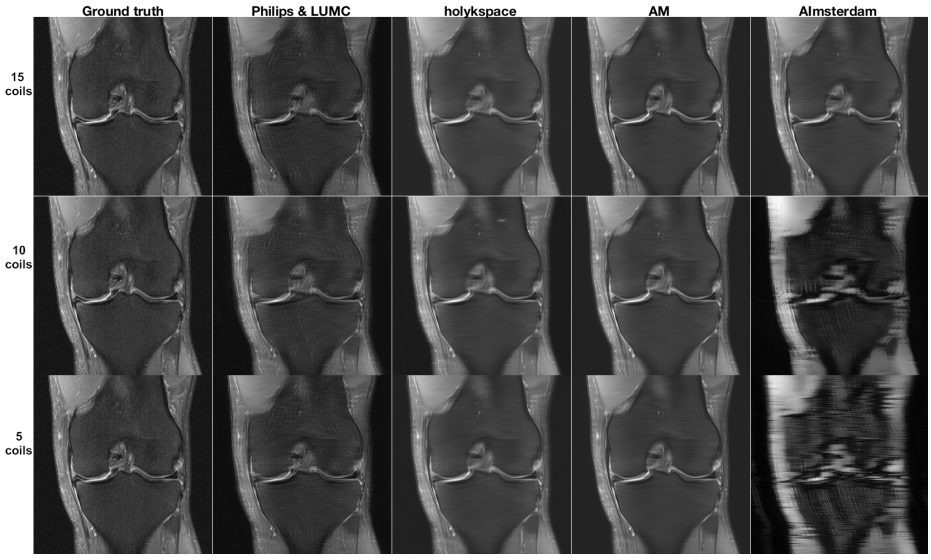

Fig. 5. Image results for R = 8 coil compression experiments. The network input was 5, 10 and 15 coils

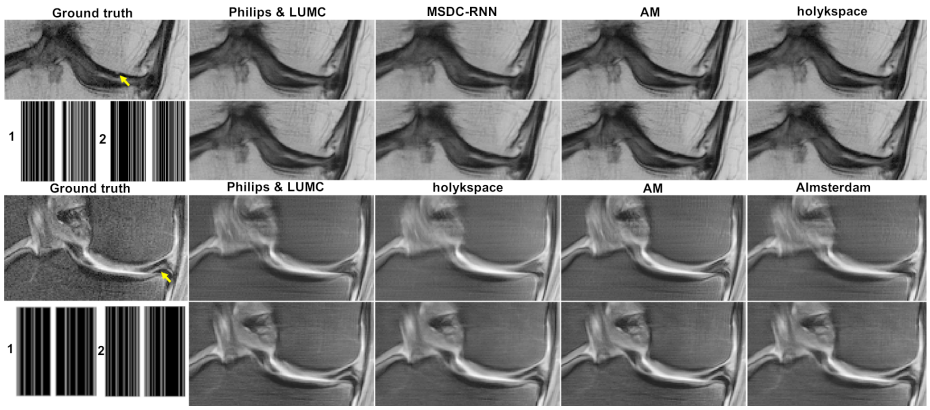

Fig. 6. Image results for two different sampling patterns. The original sampling patterns (rows 1 and 3) resulted in removed pathology. The pathology is more visible with the modified sampling pattern (rows 2 and 4).The original (1) and new (2) sampling patterns for each case are shown on the left of rows 2 and 4.

References

1. F. Knoll, T. Murrell, A. Sriram, N. Yakubova, J. Zbontar, M. Rabbat, A. Defazio, M. J. Muckley, D. K. Sodickson, C. L. Zitnick, and M. P. Recht, "Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge," *Magnetic Resonance in Medicine*, p. mrm.28338, jun 2020.
2. Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-Net for compressive sensing MRI," in *Advances in Neural Information Processing Systems*, 2016, pp. 10–18.
3. K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated MRI data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.
4. J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, 2017.
5. G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, and D. Firmin, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1310–1321, 2017.
6. T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang, "KIKI-net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *Magnetic Resonance in Medicine*, vol. 80, no. 5, pp. 2188–2201, 2018.
7. H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 394–405, 2018.
8. B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.
9. F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson, "Assessment of the generalization of learned image reconstruction and the potential for transfer learning," *Magnetic Resonance in Medicine*, vol. 81, no. 1, pp. 116–128, 2019.
10. F. Knoll, K. Hammernik, C. Zhang, S. Moeller, T. Pock, D. K. Sodickson, and M. Akçakaya, "Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues," *IEEE Signal Processing Magazine*, vol. 37, no. 1, pp. 128–140, 2020.
11. B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magnetic Resonance in Medicine*, vol. 84, no. 6, pp. 3172–3191, 2020.
12. J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui, "fastMRI: An open dataset and benchmarks for accelerated MRI," *arXiv preprint arXiv:1811.08839*, 2018.
13. F. Knoll, J. Zbontar, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, J. Pinkerton, D. Wang, N. Yakubova, E. Owens, C. L.

Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui, "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, p. e190007, 2020.

14. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
15. N. Pezzotti, S. Yousefi, M. S. Elmahdy, J. H. F. Van Gemert, C. Schuelke, M. Doneva, T. Nielsen, S. Kastryulin, B. P. F. Lelieveldt, M. J. P. Van Osch, E. De Weerdt, and M. Staring, "An adaptive intelligence algorithm for undersampled knee mri reconstruction," *IEEE Access*, vol. 8, pp. 204 825–204 838, 2020.
16. 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, mar 2014.
17. P. Wang, E. Z. Chen, T. Chen, V. M. Patel, and S. Sun, "Pyramid convolutional RNN for MRI reconstruction," *arXiv:1912.00543*, 2019.
18. H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," *arXiv*, 2015.
19. K. Hammernik, J. Schlemper, C. Qin, J. Duan, R. M. Summers, and D. Rueckert, "Σ-Net: Ensembled Iterative Deep Neural Networks for Accelerated Parallel MR Image Reconstruction," in *Proceedings of the International Society of Magnetic Resonance in Medicine (ISMRM)*, 2020, p. 0987.
20. S. Yu, B. Park, and J. Jeong, "Deep Iterative Down-Up CNN for Image Denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
21. P. Putzky and M. Welling, "Invert to learn to invert," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf
22. ——, "Recurrent inference machines for solving inverse problems," 2017.
23. K. Lønning, P. Putzky, J.-J. Sonke, L. Reneman, M. W. Caan, and M. Welling, "Recurrent inference machines for reconstructing heterogeneous mri data," *Medical Image Analysis*, vol. 53, pp. 64–78, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518306078
24. "Accelerating cartesian mri by domain-transform manifold learning in phase-encoding direction," *Medical Image Analysis*, vol. 63, p. 101689, 2020.
25. Y. Jun, T. Eo, H. Shin, T. Kim, H.-J. Lee, and D. Hwang, "Parallel imaging in time-of-flight magnetic resonance angiography using deep multistream convolutional neural networks," *Magnetic Resonance in Medicine*, vol. 81, no. 6, pp. 3840–3853, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27656