Robust Compressed Sensing MRI with Deep Generative Priors

Ajil Jalal*† ajiljalal@utexas.edu

Marius Arvinte*† arvinte@utexas.edu

Giannis Daras‡ giannisdaras@utexas.edu

Eric Price‡ ecprice@cs.utexas.edu

Alexandros G. Dimakis† dimakis@austin.utexas.edu

Jonathan I. Tamir† jtamir@utexas.edu

Abstract

The CSGM framework (Bora-Jalal-Price-Dimakis'17) has shown that deep generative priors can be powerful tools for solving inverse problems. However, to date this framework has been empirically successful only on certain datasets (for example, human faces and MNIST digits), and it is known to perform poorly on out-of-distribution samples. In this paper, we present the first successful application of the CSGM framework on clinical MRI data. We train a generative prior on brain scans from the fastMRI dataset, and show that posterior sampling via Langevin dynamics achieves high quality reconstructions. Furthermore, our experiments and theory show that posterior sampling is robust to changes in the ground-truth distribution and measurement process.

1 Introduction

Compressed sensing [12, 8] has enabled reductions to the number of measurements needed for successful reconstruction in a variety of imaging inverse problems. In particular, it has led to shorter scan times for magnetic resonance imaging (MRI) [25, 39], and most MRI vendors have released products leveraging this framework to accelerate clinical workflows.

More recently, deep learning techniques have been used as powerful data-driven reconstruction methods for inverse problems [20, 29]. There are two broad families of deep learning inversion techniques [29]: end-to-end supervised and distribution-learning approaches. End-to-end supervised techniques use a training set of measured images and deploy convolutional neural networks (CNNs) and other architectures to learn the inverse mapping from measurements to image see e.g. [16, 2, 26]. End-to-end methods are trained for specific imaging anatomy and measurement models and show excellent performance in these tasks. However, reconstruction quality can suffer when applied out of distribution, such as under various types of natural measurement and anatomy perturbations [3, 11].

In this paper we study deep learning inversion techniques based on distribution learning. The most common family of such techniques, known also as Compressed Sensing with Generative Modeling (CSGM) [6] uses pre-trained generative models as priors. These methods have only recently been applied to MRI and have not yet been shown to be competitive with supervised end-to-end methods. The very recent work [22] trains a StyleGAN for magnitude-only DICOM images but requires the presence of side-information and studies simulated measurements for reconstruction. The deviation from the true MRI measurement model and the use of magnitude images are known to be problematic when evaluating performance [31]. Untrained and unamortized generators [17] have also been recently explored [11], showing promising results in some cases. Further, [10] studies the harder problem of learning a generative model for a class of images using only partial observations, as first proposed in AmbientGAN [7].

[*] Ajil Jalal and Marius Arvinte contributed equally to this work.

[†] The University of Texas at Austin, Department of Electrical and Computer Engineering

[‡] The University of Texas at Austin, Department of Computer Science

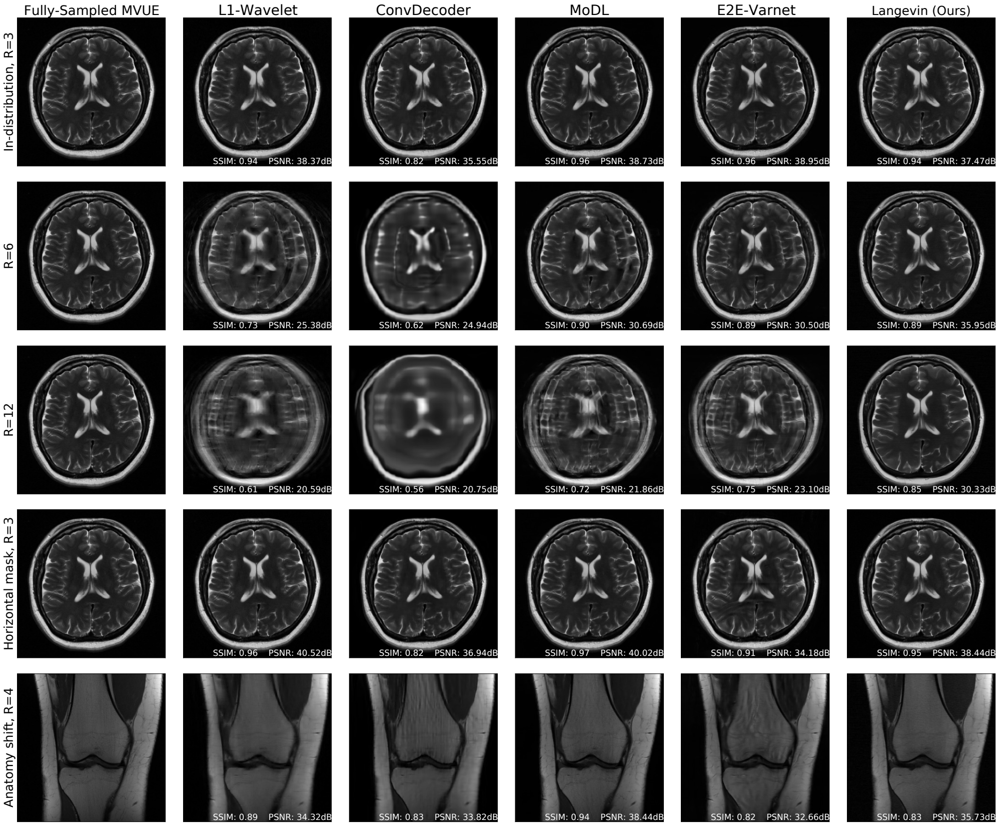


Figure 1: Comparison of reconstruction methods for in-distribution, sampling-shift, and anatomy-shift images. All methods and hyperparameters were optimized on T2-weighted *brain* scans with a vertical sampling mask, and tested at higher accelerations, horizontal masks, and on knee & abdomen scans. Our reconstructions are competitive with state-of-the-art methods, and introduce fewer artifacts out of distribution. All measurements are multicoil k-space from the NYU fastMRI dataset and the supervised baselines are trained from scratch on MVUE targets for a fair comparison.

1.1 Contributions

- We successfully train a score-based deep generative model (NCSNv2 [34]) for complex-valued, T2-weighted brain MR images without any assumptions on the measurement scheme. When applied to multi-coil MRI reconstruction under the CSGM framework, we achieve competitive performance compared to end-to-end deep learning methods when the test-time data are sampled within distribution. This is shown in the top row of Figure 1 and quantitatively in Figure 2.

- We give evidence that posterior sampling should give high-quality reconstructions. First, we show that for any measurements (including the Fourier measurements in MRI) that posterior sampling with the correct prior is within constant factors of the optimal recovery method; second, even if the prior is wrong but gives α mass to the true distribution, we show that posterior sampling for Gaussian measurements is nearly optimal with just an additive $O(\log(1/\alpha))$ loss. These results are in Appendix A.

- We empirically show that our approach is robust to test-time distribution shifts including different sampling patterns and imaging anatomy. The former is unsurprising given that our model was trained without knowledge of the measurement scheme. As a consequence, our approach provides a degree of flexibility in choosing scan parameters – a common situation in routine clinical imaging. In contrast, in Figure 1 and 6, we empirically show that end-to-end methods do not always enjoy the same robustness guarantees, in some cases leading to severe degradation in reconstruction quality when applied out-of-distribution.

- As shown in Figure 3, our method can be used to obtain multiple samples from the posterior by running

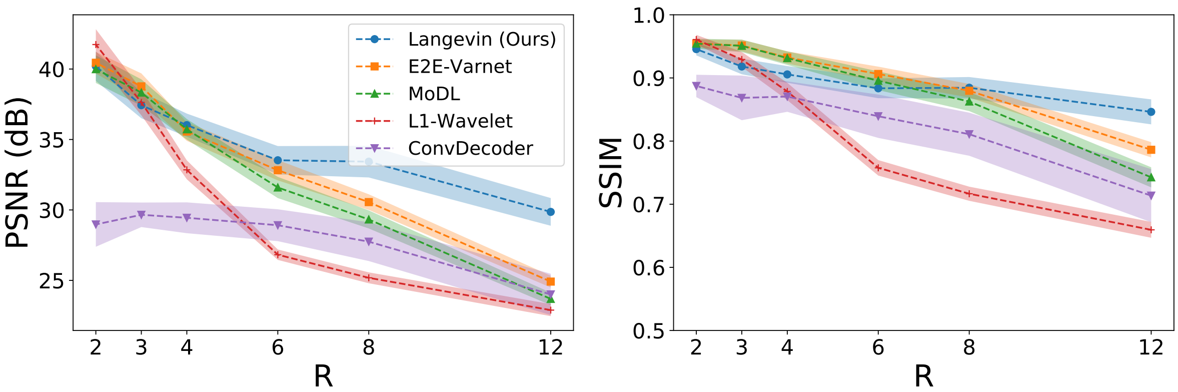


Figure 2: Quantitative results for T2-weighted MRI brains. The left and right plots show PSNR and SSIM as a function of the acceleration R. Our method is competitive with state-of-the-art baselines at low accelerations, and is better at higher accelerations. Note that the MoDL and E2E-VarNet baselines were trained to optimize SSIM at R=3 & 6, and hence achieve good SSIM scores, although one can see aliasing at R=6 in Row 2 of Figure 1 and Figure 10 in the Appendix. Figure 4 and 5 in the Appendix show quantitative results on different anatomies.

Langevin dynamics with different random initializations. This allows us to get multiple reconstructions which can be used to obtain confidence intervals for each reconstructed voxel and visualize our reconstruction uncertainty on a voxel-by-voxel resolution. Uncertainty quantification can be incorporated into end-to-end methods, e.g., using variational auto-encoders [13], but this requires changes to the architecture. Our method does not require any modification and multiple reconstruction samplers can be run in parallel.

Our main results are succinctly summarized in Figure 1: we achieve equivalent reconstruction performance using a reduced training set when evaluated in-distribution and is robust when evaluated out-of-distribution.

2 System Model and Algorithm

2.1 Multi-coil Magnetic Resonance Imaging

MRI is a medical imaging modality that makes measurements using an array of radio-frequency coils placed around the body. Each coil is spatially sensitive to a local region, and measurements are acquired directly in the spatial frequency, or *k-space*, domain. To decrease scan time, reduce operating costs, and improve patient comfort, a reduced number of k-space measurements are acquired in clinical use and reconstructed by incorporating explicit or implicit knowledge of the spatial sensitivity maps [32, 30, 14]. Formally, the vector of measurements $y_i \in \mathbb{C}^L$ acquired by the i^{th} coil can be characterized by the forward model [30]:

$$y_i = PFS_i x^* + w_i, \quad i = 1, ..., N_c, \tag{1}$$

where $x^* \in \mathbb{C}^N$ is the image containing N pixels, S_i is an operator representing the point-wise multiplication of the i^{th} coil sensitivity map, F is the spatial Fourier transform operator, P represents the k-space sampling operator, and we assume $w_i \sim \mathcal{N}_c\left(0, \sigma^2 I\right)$ for simplicity. Importantly, note that the same under-sampling operator is applied to all N_c coils. The acceleration factor R denotes the degree of under-sampling in the k-space domain, i.e., $R = N/L$. Note that we use the true acceleration factor R, and this does not match the values in fastMRI [23] [1]. Given multi-coil measurements y, sensitivity maps represented by S and the sampling operator P, the goal of MR image reconstruction is to estimate the underlying image variable x^*.

[1] See https://github.com/facebookresearch/fastMRI/blob/main/fastmri/data/subsample.py for the fastMRI acceleration definition

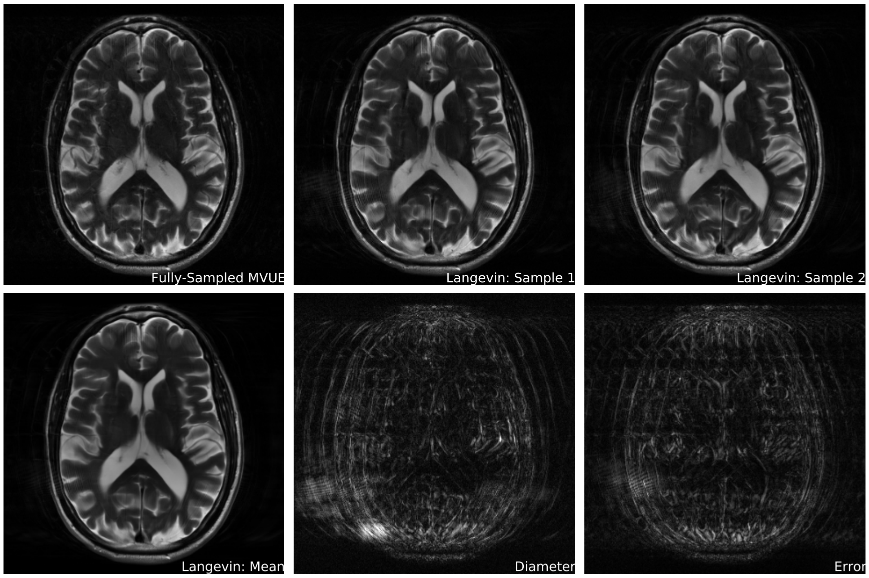


Figure 3: Our method can be run in parallel to generate different random reconstructions, and estimate uncertainty of our reconstructions. The top-left image shows the fully-sampled MVUE, top-middle top-right show different random samples from Langevin dynamics at an acceleration of $R = 12$. The bottom-left image shows the mean over 32 runs of Langevin dynamics, bottom-middle shows the difference between the two furthest reconstructions, and the bottom-right shows the error between the MVUE and mean reconstruction.

2.2 Posterior Sampling

The algorithm we consider is *posterior sampling* [19, 33]. That is, given an observation of the form $y = Ax^* + w$, where $y \in \mathbb{C}^M$, $A \in \mathbb{C}^{M \times N}$, $w \sim \mathcal{N}_c(0, \sigma^2 I)$, and $x^* \sim \mu$, the posterior sampling recovery algorithm outputs $\widehat{x}$ according to the posterior distribution $\mu(\cdot|y)$.

In order to sample from the posterior, we use *Langevin Dynamics* [5]. Assuming we have access to $\nabla_x \log \mu(x|y)$, we can sample from $\mu(x|y)$ by running noisy gradient ascent:

$$x_{t+1} \leftarrow x_t + \eta_t \nabla_{x_t} \log \mu(x_t|y) + \sqrt{2\eta_t}\, \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, 1). \tag{2}$$

Prior work [5] has shown that as $t \to \infty$ and $\eta_t \to 0$, Langevin dynamics will correctly sample from $\mu(x|y)$. In practice, vanilla Langevin Dynamics are slow to converge. Hence, the work in [33] proposes *annealed* Langevin Dynamics, where the marginal distribution of x at iteration t is modelled as $\mu_t = \mu * \mathcal{N}(0, \beta_t^2)$ and the generative model is trained to estimate the score function $f(x_t; \beta_t) := \nabla_{x_t} \log((\mu * \mathcal{N}(0, \beta_t^2)(x_t))$.

Since the distribution of $y|x^*$ is Gaussian in Eqn (1), our final algorithm with annealed Langevin dynamics is: for $x_0 \sim \mathcal{N}_c(0, I)$ and for all $t = 0, \cdots, T-1$,

$$x_{t+1} \leftarrow x_t + \eta_t \left(f(x_t; \beta_t) + \frac{A^H(y - Ax_t)}{\gamma_t^2 + \sigma^2} \right) + \sqrt{2\eta_t}\, \zeta_t, \quad \zeta_t \sim \mathcal{N}(0; I). \tag{3}$$

Note that the parameters $T, \{\beta_t\}_{t=0}^{T-1}$ were fixed during training of the generative model, and hence the only hyperparameters during inference are $\{\eta_t\}_{t=0}^{T-1}$ and $\{\gamma_t\}_{t=0}^{T-1}$. Appendix G describes hyper-parameter values used in our experiments.

3 Conclusions

This paper reports the first wide-scale, successful application of the CSGM framework for multi-coil MR image reconstruction under realistic sampling conditions, and provides theoretical evidence for the robustness of posterior sampling. Our score-based model was trained on a subset of brain MRI scans without explicit information about the sampling scheme (unlike end-to-end baselines) or the reconstruction task and shows a considerable degree of generalization to out-of-distribution samples such as abdomen and knee MRI. These scans were acquired using different MRI vendors with different pulse sequence parameters and at different institutions. We postulate that adding a small set of diverse training samples to the generative model could further improve robustness, and that these samples may not necessarily be restricted to MR images.

Though promising, our initial results are still limited to fast spin-echo imaging only and all data were retrospectively under-sampled. Further study is required to demonstrate prospective performance in a larger body of heterogeneous MRI data. Our method also currently requires a high compute cost at inference time, as well as the need for a pre-trained generative model. Clinical use requires fast reconstruction in addition to fast scanning. Future work should investigate whether score-based models can be trained without a fully-sampled training set as well as investigate approaches to reducing computation time.

4 Acknowledgements

Ajil Jalal, Giannis Daras and Alex Dimakis have been supported by NSF Grants CCF 1763702, 1934932, AF 1901281, 2008710, 2019844, the NSF IFML 2019844 award as well as research gifts by Western Digital, Interdigital, WNCG and MLL, computing resources from TACC and the Archie Straiton Fellowship.

Eric Price has been supported by NSF Award CCF-1751040 (CAREER), NSF Award CCF-2008868, and NSF IFML 2019844.

Marius Arvinte and Jon Tamir have been supported by NSF IFML 2019844 award, ONR grant N00014-19-1-2590, NIH Grant U24EB029240, and an AWS Machine Learning Research Award.

References

[1] http://mridata.org/.

[2] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.

[3] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[5] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Seminaire de probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

[6] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[7] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. *ICLR*, 2:5, 2018.

[8] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[9] Thierry Champion, Luigi De Pascale, and Petri Juutinen. The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

[10] EK Cole, JM Pauly, SS Vasanawala, and F Ong. Unsupervised mri reconstruction with generative adversarial networks. arxiv 2020. *arXiv preprint arXiv:2008.13065*.

[11] Mohammad Zalbagi Darestani, Akshay Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. *arXiv preprint arXiv:2102.06103*, 2021.

[12] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[13] Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep mri reconstruction. *IEEE Transactions on Medical Imaging*, 40(1):239–250, 2021.

[14] Mark A. Griswold, Peter M. Jakob, Robin M. Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.

[15] Matthieu Guerquin-Kern, Laurent Lejeune, Klaas Paul Pruessmann, and Michael Unser. Realistic analytical phantoms for parallel magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 31(3):626–636, 2011.

[16] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[17] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. *arXiv preprint arXiv:1810.03982*, 2018.

[18] Siddharth Iyer, Frank Ong, Kawin Setsompop, Mariya Doneva, and Michael Lustig. Sure-based automatic parameter selection for espirit calibration. *Magnetic Resonance in Medicine*, 84(6):3423–3437, 2020.

[19] Ajil Jalal, Sushrut Karmalkar, Alexandros G Dimakis, and Eric Price. Instance-optimal compressed sensing via posterior sampling. *arXiv preprint arXiv:2106.11438*, 2021.

[20] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[22] Varun A Kelkar and Mark A Anastasio. Prior image-constrained reconstruction using style-based generative models. *arXiv preprint arXiv:2102.12525*, 2021.

[23] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, 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*, 2(1):e190007, 2020.

[24] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[25] Michael Lustig, David Donoho, and John 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*, 58(6):1182–1195, 2007.

[26] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[27] Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images. *IEEE transactions on medical imaging*, 39(4):1064–1072, 2019.

[28] Matthew J. Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021.

[29] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *arXiv preprint arXiv:2005.06001*, 2020.

[30] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(5):952–962, 1999.

[31] Efrat Shimron, Jonathan Tamir, Ke Wang, and Michael Lustig. Subtle inverse crimes: Naively using publicly available images could make reconstruction results seem misleadingly better! *Proceedings of The ISMRM*, 2021.

[32] Daniel K Sodickson and Warren J Manning. Simultaneous acquisition of spatial harmonics (smash): fast imaging with radiofrequency coil arrays. *Magnetic resonance in medicine*, 38(4):591–603, 1997.

[33] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11918–11930, 2019.

[34] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

[35] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–73. Springer, 2020.

[36] Martin Uecker, Christian Holme, Moritz Blumenthal, Xiaoqing Wang, Zhengguo Tan, Nick Scholand, Siddharth Iyer, Jon Tamir, and Michael Lustig. mrirecon/bart: version 0.7.00, March 2021.

[37] Martin Uecker, Peng Lai, Mark J Murphy, Patrick Virtue, Michael Elad, John M Pauly, Shreyas S Vasanawala, and Michael Lustig. Espirit—an eigenvalue approach to autocalibrating parallel mri: where sense meets grappa. *Magnetic resonance in medicine*, 71(3):990–1001, 2014.

[38] Martin Uecker, Frank Ong, Jonathan I Tamir, Dara Bahri, Patrick Virtue, Joseph Y Cheng, Tao Zhang, and Michael Lustig. Berkeley advanced reconstruction toolbox. In *Proc. Intl. Soc. Mag. Reson. Med*, volume 23, 2015.

[39] Shreyas S. Vasanawala, Marcus T. Alley, Brian A. Hargreaves, Richard A. Barth, John M. Pauly, and Michael Lustig. Improved pediatric mr imaging with compressed sensing. *Radiology*, 256(2):607–616, 2010. PMID: 20529991.

[40] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[41] Zhou Wang and Alan C Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–156, 2006.

[42] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. 2018.

A Appendix: Theoretical Results

Background and Notation. We first introduce background and notation required for our theoretical results. $\|\cdot\|$ refers to the ℓ_2 norm. In this section alone, for simplicity of exposition, we will assume that all matrices and vectors are real valued.

For two probability distributions μ, ν on some normed space Ω, and for any $q \geq 1$, the Wasserstein-q [40, 4] and Wasserstein-∞ [9] distances are defined as:

$$\mathcal{W}_q(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \left(\mathop{\mathbb{E}}_{(u,v)\sim\gamma} \left[\|u-v\|^q\right] \right)^{1/q}, \quad \mathcal{W}_\infty(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \left(\gamma\text{-}\mathop{\mathrm{ess\,sup}}_{(u,v)\in\Omega^2} \|u-v\| \right).$$

where $\Pi(\mu,\nu)$ denotes the set of joint distributions whose marginals are μ, ν. The above definition says that if $\mathcal{W}_\infty(\mu,\nu) \leq \varepsilon$, and $(u,v)\sim\gamma$, then $\|u-v\| \leq \varepsilon$ almost surely.

The $(\varepsilon,\delta)-$*approximate covering number* [19], is defined as the smallest number of ε-radius balls required to cover $1-\delta$ mass under a distribution.

Definition A.1 ((ε,δ)-approximate covering number)**.** *Let μ be a distribution on $\mathbb{R}^N$. For some parameters $\varepsilon > 0, \delta \in [0,1]$, the* (ε,δ)-approximate covering number *of μ is defined as*

$$\mathrm{Cov}_{\varepsilon,\delta}(\mu) := \min\left\{k : \mu\left[\cup_{i=1}^k B(x_i,\varepsilon)\right] \geq 1-\delta, x_i \in \mathbb{R}^N\right\},$$

where $B(x,\varepsilon)$ is the ℓ_2 ball of radius ε centered at x.

Distributional robustness under Gaussian measurements. First, we consider mismatch between the ground-truth distribution, denoted by μ, and the generator distribution, denoted by ν. Prior work [19] has shown that if (i) $\mathcal{W}_q(\mu, \nu) \leq \varepsilon$ for some $q \geq 1$ and (ii) we are given $M \geq O(\log \mathrm{Cov}_{\varepsilon,\delta}(\mu))$ Gaussian measurements, then posterior sampling with respect to ν will recover $x^* \sim \mu$ up to an error of $\varepsilon/\delta^{1/q}$ with probability $1 - \delta$. Closeness in Wasserstein distance is a reasonable assumption in certain examples, such as when μ is the distribution of celebrity faces and ν is the distribution of a generator trained on FlickrFaces [21]. However, this assumption is unsatisfactory when we consider distributions of abdominal and brain MR scans, for example, since images of these anatomies look entirely different.

We define the following weaker notion of divergence between distributions. Informally, this new definition tells us that ν and μ are "close" if they can each be split into components which are close in $\mathcal{W}_\infty$ distance, such that the close components contain a sufficiently large fraction under ν and μ. Formally, this is defined as:

Definition A.2 ((δ, α)-$\mathcal{W}_\infty$ divergence)**.** *For two probability distributions ν and μ, and parameters $\delta, \alpha \in [0, 1]$, the (δ, α)-$\mathcal{W}_\infty$ divergence is defined as*

$$(\delta, \alpha)\text{-}\mathcal{W}_\infty(\mu, \nu) := \inf\{\varepsilon \geq 0 :$$
$$\exists \mu', \mu'', \nu', \nu'' \in \mathcal{M}(\mathbb{R}^N) \textit{ s.t. } \mu = (1 - \delta)\mu' + \delta\mu'', \nu = (1 - \alpha)\nu' + \alpha\nu'', \mathcal{W}_\infty(\mu', \nu') = \varepsilon.\}$$

Lemma A.3 highlights that this is a strict generalization of Wasserstein distances, in the sense that closeness in Wasserstein distance implies closeness in this new divergence.

Lemma A.3 ($\mathcal{W}_q$ implies (δ, α)-$\mathcal{W}_\infty$)**.** *If two distributions μ and ν satisfy $\mathcal{W}_q(\mu, \nu) \leq \varepsilon$ for some $q \geq 1$, then they satisfy (δ, δ)-$\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon/\delta^{1/q}$. Futhermore, there exist distributions that satisfy (δ, δ)-$\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$, but $\mathcal{W}_q(\mu, \nu) = \infty$ for all $q \geq 1$.*

Since the (δ, α)-$\mathcal{W}_\infty$ divergence is a generalization of Wasserstein distances, it is not clear that the main Theorem in [19] holds for distributions that are close in this new divergence. The following result shows a rather surprising fact: if (δ, α)-$\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$ then posterior sampling with $M = O\left(\log\left(\frac{1}{1-\alpha}\right) + \log \mathrm{Cov}_{\varepsilon,\delta}(\mu)\right)$ measurements will still succeed with probability $\geq 1 - O(\delta)$.

Theorem A.4. *Let $\delta, \alpha \in [0, 1]$, and $\varepsilon > 0$ be parameters. Let μ, ν be arbitrary distributions over $\mathbb{R}^N$ satisfying (δ, α)-$\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$. Let $x^* \sim \mu$ and suppose $y = Ax^* + w$, where $A \in \mathbb{R}^{M \times N}$ and $w \in \mathbb{R}^M$ are i.i.d. Gaussian normalized such that $A_{ij} \sim \mathcal{N}(0, 1/M)$ and $w_i \sim \mathcal{N}(0, \sigma^2/M)$, with $\sigma \gtrsim \varepsilon$. Given y and the fixed matrix A, let $\widehat{x}$ be the output of posterior sampling with respect to ν.*

Then for $M \geq O\left(\log\left(\frac{1}{1-\alpha}\right) + \min(\log \mathrm{Cov}_{\sigma,\delta}(\mu), \log \mathrm{Cov}_{\sigma,\delta}(\nu))\right)$, there exists a universal constant $c > 0$ such that with probability at least $1 - e^{-\Omega(M)}$ over A, w,

$$\Pr_{x^* \sim \mu, \widehat{x} \sim \nu(\cdot|y)} [\|x^* - \widehat{x}\| \geq c(\varepsilon + \sigma)] \leq \delta + e^{-\Omega(M)}.$$

For our running example of ν being a generator trained on brain scans, and μ the distribution of abdominal scans, we can set ν' to be the distribution of our generator restricted to abdominal scans, and we can let μ' be the distribution restricted to "inliers" in μ. This shows that even if our generator places an *exponentially small* probability mass(i.e., $1 - \alpha \ll 1$) on the set of abdominal scans, we can still recover abdominal scans with a *polynomial additive* increase in the number of measurements (i.e., $\log(1/(1 - \alpha))$).

Near-optimality under arbitrary measurement processes. The previous result required Gaussian matrices to handle the distribution shift. Our next result shows that for an *arbitrary* measurement process, and assuming that there is no distribution shift between the generator and the ground truth distribution, posterior sampling is almost the best algorithm for this *fixed* measurement process.

Theorem A.5. *Let $x^* \sim \mu$ and let $y = \mathcal{A}(x^*)$ be measurements generated from x^* for some arbitrary forward operator $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^M$. Then if there exists an algorithm that uses y as inputs and outputs x' such that*

$$\|x^* - x'\| \leq \varepsilon \textit{ with probability } 1 - \delta,$$

then posterior sampling $\widehat{x} \sim \mu(\cdot|y)$ will satisfy

$$\|x^* - \widehat{x}\| \leq 2\varepsilon \textit{ with probability } \geq 1 - 2\delta.$$

Remark on combining these results. Our theoretical results above show that posterior sampling is (1) highly robust to distribution shift under Gaussian measurements, and (2) accurate with arbitrary measurements without distribution shift. A natural hope would be to combine these two results and show that it is robust to distribution shift under Fourier measurements. Unfortunately, this is *not* true for general distributions: for example, if μ and ν are both random distributions over Fourier-sparse signals, then Fourier measurements will usually give zero information about the signal, so cannot convince the sampler to sample near μ rather than ν.

A.1 Proof of Lemma A.3

Proof. Let Γ be a coupling between μ, ν such that $\mathbb{E}_{(u,v)\sim\Gamma}\left[\|u-v\|^q\right] \leq \varepsilon^q$. Then an application of Markov's inequality gives

$$\Pr[\|u-v\| \geq \varepsilon/\delta^{1/q}] \leq \delta. \tag{4}$$

Now, we can split the distribution Γ into two unnormalized components Γ', Γ'' defined as

$$\begin{aligned}\Gamma'(u,v) &= \Gamma(u,v)\mathbf{1}\{\|u-v\| < \varepsilon/\delta^{1/q}\},\\ \Gamma''(u,v) &= \Gamma(u,v)\mathbf{1}\{\|u-v\| \geq \varepsilon/\delta^{1/q}\}.\end{aligned}$$

Using Γ', Γ'', we can define measures μ', μ'', ν', ν'', via

$$\begin{aligned}\mu'(B) &:= \Gamma'(B,\Omega),\\ \mu''(B) &:= \Gamma''(B,\Omega),\\ \nu'(B) &:= \Gamma'(\Omega,B),\\ \nu''(B) &:= \Gamma''(\Omega,B),\end{aligned}$$

where B is any measurable set and Ω is the state-space.

Since Γ is a valid coupling between μ, ν, and Γ', Γ'' are disjoint distributions, for any measurable $B \subseteq \Omega$, we have:

$$\begin{aligned}\mu(B) &= \Gamma(B,\Omega),\\ &= \Gamma'(B,\Omega) + \Gamma''(B,\Omega),\\ &= \mu'(B) + \mu(B''),\\ &= \mu'(\Omega)\frac{\mu'(B)}{\mu'(\Omega)} + \mu''(\Omega)\frac{\mu''(B)}{\mu''(\Omega)}.\end{aligned}$$

Using Eqn (4), we can conclude that $\mu'(\Omega) \geq 1-\delta, \mu''(\Omega) \leq \delta$. Setting $\mu' \leftarrow \mu'/\mu'(\Omega)$ and $\mu'' \leftarrow \mu''/\mu''(\Omega)$, we can now rewrite μ as $\mu = (1-\delta)\mu' + \delta\mu''$. A similar argument for ν gives $\nu = (1-\delta)\nu' + \delta\nu''$.

By construction, μ', ν' can be $\mathcal{W}_\infty$ coupled via Γ' to within a distance of $\varepsilon/\delta^{1/q}$. This shows that (δ,δ)-$\mathcal{W}_\infty(\mu,\nu) \leq \varepsilon/\delta^{1/q}$.

Now we need to show that two distributions can be close in (δ,δ)-$\mathcal{W}_\infty$, but $\mathcal{W}_q = \infty$ for all q. Consider two scalar distributions μ, ν defined as

$$\mu = \begin{cases} 0 & \text{with probability } 1-\delta,\\ r & \text{with probability } \delta, \end{cases},$$

$$\nu = \begin{cases} \varepsilon & \text{with probability } 1-\delta,\\ -r & \text{with probability } \delta. \end{cases}$$

Clearly, these distributions satisfy (δ,δ)-$\mathcal{W}_\infty(\mu,\nu) \leq \varepsilon$, but $\mathcal{W}_q(\mu,\nu) \approx r$ for all q. As $r \to \infty$, we get $\mathcal{W}_q(\mu,\nu) \to \infty$ for all $q \geq 1$.

□

A.2 Proof of Theorem A.4

In order to prove the Theorem, we make use of the following three Lemmas from [19].

Lemma A.6. *For $c \in [0, 1]$, let $H := (1 - c)H_0 + cH_1$ be a mixture of two absolutely continuous distributions H_0, H_1 admitting densities h_0, h_1. Let y be a sample from the distribution H, such that $y|z^* \sim H_{z^*}$ where $z^* \sim$ Bernoulli(c).*

Define $\widehat{c}_y = \frac{ch_1(y)}{(1-c)h_0(y)+ch_1(y)}$, and let $\widehat{z}|y \sim$ Bernoulli$(\widehat{c}_y)$ be the posterior sampling of z^* given y. Then we have

$$\Pr_{z^*,y,\widehat{z}}[z^* = 0, \widehat{z} = 1] \leq 1 - TV(H_0, H_1).$$

Lemma A.7. *Let y be generated from x^* by a Gaussian measurement process with noise rate σ. For a fixed $\tilde{x} \in \mathbb{R}^n$, and parameters $\eta > 0, c \geq 4e^2$, let P_{out} be a distribution supported on the set*

$$S_{\tilde{x},out} := \{x \in \mathbb{R}^n : \|x - \tilde{x}\| \geq c(\eta + \sigma)\}.$$

Let $P_{\tilde{x}}$ be a distribution which is supported within an $\eta-$radius ball centered at $\tilde{x}$.

For a fixed A, let $H_{\tilde{x}}$ denote the distribution of y when $x^* \sim P_{\tilde{x}}$. Let H_{out} denote the corresponding distribution of y when $x^* \sim P_{out}$. Then we have:

$$\mathop{\mathbb{E}}_{A}[TV(H_{\tilde{x}}, H_{out})] \geq 1 - 4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

Lemma A.8. *Let R, P, denote arbitrary distributions over $\mathbb{R}^n$ such that $\mathcal{W}_\infty(R, P) \leq \varepsilon$.*

Let $x^* \sim R$ and $z^* \sim P$ and let y and u be generated from x^* and z^* via a Gaussian measurement process with m measurements and noise rate σ. Let $\widehat{x} \sim P(\cdot|y, A)$ and $\widehat{z} \sim P(\cdot|u, A)$. For any $d > 0$, we have

$$\Pr_{x^*,A,w,\widehat{x}}[\|x^* - \widehat{x}\| \geq d + \varepsilon] \leq e^{-\Omega(m)} + e^{\left(\frac{4\varepsilon(\varepsilon+2\sigma)m}{2\sigma^2}\right)} \Pr_{z^*,A,w,\widehat{z}}[\|z^* - \widehat{z}\| \geq d].$$

Theorem A.4. *Let $\delta, \alpha \in [0, 1]$, and $\varepsilon > 0$ be parameters. Let μ, ν be arbitrary distributions over $\mathbb{R}^N$ satisfying (δ, α)-$\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$. Let $x^* \sim \mu$ and suppose $y = Ax^* + w$, where $A \in \mathbb{R}^{M \times N}$ and $w \in \mathbb{R}^M$ are i.i.d. Gaussian normalized such that $A_{ij} \sim \mathcal{N}(0, 1/M)$ and $w_i \sim \mathcal{N}(0, \sigma^2/M)$, with $\sigma \gtrsim \varepsilon$. Given y and the fixed matrix A, let $\widehat{x}$ be the output of posterior sampling with respect to ν.*

Then for $M \geq O\left(\log\left(\frac{1}{1-\alpha}\right) + \min(\log \mathrm{Cov}_{\sigma,\delta}(\mu), \log \mathrm{Cov}_{\sigma,\delta}(\nu))\right)$, there exists a universal constant $c > 0$ such that with probability at least $1 - e^{-\Omega(M)}$ over A, w,

$$\Pr_{x^* \sim \mu, \widehat{x} \sim \nu(\cdot|y)}[\|x^* - \widehat{x}\| \geq c(\varepsilon + \sigma)] \leq \delta + e^{-\Omega(M)}.$$

Proof. We know from (δ, α)-$\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$ that there exist μ', ν', μ'', ν'' and a finite distribution Q supported on a set S such that

1. $\mathcal{W}_\infty(\mu', \nu') \leq \varepsilon$,
2. $\min\{\mathcal{W}_\infty(\nu', Q), \mathcal{W}_\infty(\mu', Q)\} \leq \sigma$,
3. $\mu = (1 - \delta)\mu' + \delta\mu''$ and $\nu = (1 - \alpha)\nu' + \alpha\nu''$.

Suppose $\mathcal{W}_\infty(\nu', Q) \leq \sigma$. If not, then $\mathcal{W}_\infty(\mu', Q) \leq \sigma$, and by (1), we see that $\mathcal{W}_\infty(\nu', Q) \leq \sigma + \varepsilon$, and we will use this in the proof instead. By decomposing $\mu = (1 - \delta)\mu' + \delta\mu''$, we have

$$\Pr_{x^* \sim \mu, \widehat{x} \sim \nu(\cdot|y)}[\|x^* - \widehat{x}\| \geq (2c + 1)\sigma + \varepsilon] \leq \delta + (1 - \delta) \Pr_{x^* \sim \mu', \widehat{x} \sim \nu(\cdot|y)}[\|x^* - \widehat{x}\| \geq (2c + 1)\sigma + \varepsilon]. \tag{5}$$

We now bound the second term on the right hand side of the above equation. For this term, consider the joint distribution over $x^*, A, w, \widehat{x}$. By Lemma A.8, we can replace $x^* \sim \mu'$ with $z^* \sim \nu'$, replace $y = Ax^* + w$ with $u = Az^* + w$, and replace $\widehat{x} \sim \nu(\cdot|A, y)$ with $\widehat{z} \sim \nu(\cdot|A, u)$ to get the following bound

$$\Pr_{x^* \sim \mu', A, w, \widehat{x} \sim \nu(\cdot|A,y)}[\|x^* - \widehat{x}\| \geq (2c + 1)\,\sigma + \varepsilon] \leq e^{-\Omega(m)} + e^{\left(\frac{2\varepsilon(\varepsilon+2\sigma)m}{\sigma^2}\right)} \Pr_{z^* \sim \nu', A, w, \widehat{z} \sim \nu(\cdot|u,A)}[\|z^* - \widehat{z}\| \geq (2c + 1)\sigma]. \tag{6}$$

We now bound the second term in the right hand side of the above inequality. Let Γ denote an optimal $\mathcal{W}_\infty-$coupling between ν' and Q.

For each $\tilde{z} \in S$, the conditional coupling can be defined as

$$\Gamma(\cdot|\tilde{z}) = \frac{\Gamma(\cdot, \tilde{z})}{Q(\tilde{z})}.$$

By the $\mathcal{W}_\infty$ condition, each $\Gamma(\cdot|\tilde{z})$ is supported on a ball of radius σ around $\tilde{z}$.

Let $E = \{z^*, \widehat{z} \in \mathbb{R}^n : \|z^* - \widehat{z}\| \geq (2c+1)\,\sigma\}$ denote the event that $z^*, \widehat{z}$ are far apart. By the coupling, we can express ν' as

$$\nu' = \sum_{\tilde{z} \in S} Q(\tilde{z})\Gamma(\cdot|\tilde{z}).$$

This gives

$$\Pr_{z^* \sim \nu', A, w, \widehat{z} \sim \nu(\cdot|A,u)} [E] = \sum_{\tilde{z}^* \in S} Q(\tilde{z}^*) \mathop{\mathbb{E}}_{z^* \sim \Gamma(\cdot|\tilde{z}^*), A, w, \widehat{z} \sim \nu(\cdot|A,u)} [1_E]\,.$$

For each $\tilde{z}^* \in S$, we now bound $Q(\tilde{z}^*)\, \mathbb{E}_{z^* \sim \Gamma(\cdot|\tilde{z}^*), A, w, \widehat{z} \sim \nu(\cdot|A,u)} [1_E]\,.$

For each $\tilde{z}^* \in S$, we can write ν as $\nu = (1-\alpha)\, Q_{\tilde{z}^*} \nu_{\tilde{z}^*,0} + c_{\tilde{z}^*,1}\nu_{\tilde{z}^*,1} + c_{\tilde{z}^*,2}\nu_{\tilde{z}^*,2}$, where the components of the mixture are defined in the following way. The first component $\nu_{\tilde{z}^*,0}$ is $\Gamma(\cdot|\tilde{z}^*)$, the second component is supported within a $2c\sigma$ radius of $\tilde{z}^*$, and the third component is supported outside a $2c\sigma$ radius of $\tilde{z}^*$.

Formally, let $B_{\tilde{z}^*}$ denote the ball of radius $c\sigma$ centered at $\tilde{z}^*$, and let $B^c_{\tilde{z}^*}$ be its complement. The constants are defined via the following Lebesque integrals, and the mixture components for any Borel measurable B are defined as

$$c_{\tilde{z}^*,1} := \int_{B_{\tilde{z}^*}} d\nu - (1-\alpha)\, Q_{\tilde{z}^*} \int_{B_{\tilde{z}^*}} d\Gamma(\cdot|\tilde{z}^*),$$

$$c_{\tilde{z}^*,2} := \int_{B^c_{\tilde{z}^*}} d\nu - (1-\alpha)\, Q_{\tilde{z}^*} \int_{B^c_{\tilde{z}^*}} d\Gamma(\cdot|\tilde{z}^*),$$

$$\nu_{\tilde{z}^*,0}(B) := \Gamma(B \cap B_{\tilde{z}^*}|\tilde{z}^*) = \Gamma(B|\tilde{z}^*) \text{ since } \operatorname{supp}(\Gamma(\cdot|\tilde{z}^*)) \subset B_{\tilde{z}^*},$$

$$\nu_{\tilde{z}^*,1}(B) := \begin{cases} \frac{1}{c_{\tilde{z}^*,1}}\nu(B \cap B_{\tilde{z}^*}) - \frac{1-\alpha}{c_{\tilde{z}^*,1}} Q_{\tilde{z}^*}\Gamma(B \cap B_{\tilde{z}^*}|\tilde{z}^*) & \text{if } c_{\tilde{z}^*,1} > 0, \\ \text{do not care} & \text{otherwise.} \end{cases},$$

$$\nu_{\tilde{z}^*,2}(B) := \begin{cases} \frac{1}{c_{\tilde{z}^*,2}}\nu(B \cap B^c_{\tilde{z}^*}) - \frac{1-\alpha}{c_{\tilde{z}^*,2}} Q_{\tilde{z}^*}\Gamma(B \cap B^c_{\tilde{z}^*}|\tilde{z}^*) & \text{if } c_{\tilde{z}^*,2} > 0, \\ \text{do not care} & \text{otherwise.} \end{cases}.$$

Notice that if z^* is sampled from $\Gamma(\cdot|\tilde{z}^*)$, then by the W_∞ condition, we have $\|z^* - \tilde{z}^*\| \leq \sigma$. Furthermore, if $\widehat{z}$ is $(2c+1)\,\sigma$ far from z^*, an application of the triangle inequality implies that it must be distributed according to $\nu_{\tilde{z}^*,2}$. That is,

$$\begin{aligned} Q(\tilde{z}^*) \mathop{\mathbb{E}}_{z^* \sim \Gamma(\cdot|\tilde{z}^*), A, w, \widehat{z} \sim \nu(\cdot|A,u)} [1_E] &\leq \mathop{\mathbb{E}}_{A,w,z^*} \Pr\left[z^* \sim \nu_{\tilde{z}^*,0}, \widehat{z} \sim \nu_{\tilde{z}^*,2}(\cdot|u)\right] \\ &\leq \frac{1}{1-\alpha} \mathop{\mathbb{E}}_{A} \left[1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})\right], \end{aligned}$$

where $H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2}$ are the push-forwards of $\nu_{\tilde{z}^*,0}, \nu_{\tilde{z}^*,2}$ for A fixed and the last inequality follows from Lemma A.6.

Notice that if we sum over all $\tilde{z}^* \in S$, then the LHS of the above inequality is an expectation over $z^* \sim \nu'$. This gives:

$$\Pr_{z^* \sim \nu', A, w, \widehat{z} \sim \nu(\cdot|u,A)} [E] \leq \frac{1}{1-\alpha} \sum_{\tilde{z}^* \in S} \mathop{\mathbb{E}}_{A} \left[1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})\right].$$

Notice that $\nu_{\tilde{z}^*,0}$ is supported within an $\sigma-$ball around $\tilde{z}^*$, and $\nu_{\tilde{z}^*,2}$ is supported outside a $2c\sigma-$ball of $\tilde{z}^*$. By Lemma A.7 we have

$$\mathop{\mathbb{E}}_{A}[TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})] \geq 1 - 4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

This implies

$$\begin{aligned}\Pr_{z^* \sim \nu', A, w, \widehat{z} \sim \nu(\cdot|u,A)} [\|z^* - \widehat{z}\| \geq (2c+1)\sigma] &\leq \frac{1}{1-\alpha} \sum_{\tilde{z}^* \in S} \mathop{\mathbb{E}}_{A} \left[(1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2}))\right], \\ &\leq \frac{1}{1-\alpha} 4|S| e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}, \\ &\leq 4e^{-\frac{m}{4}\log\left(\frac{c}{4e^2}\right)},\end{aligned}$$

where the last inequality is satisfied if $m \geq 4\log\left(\frac{1}{1-\alpha}\right) + 4\log\left(|S|\right)$.

Substituting in Eqn (6), if $c > 4\exp\left(2 + \frac{8\varepsilon(\varepsilon+2\sigma)}{\sigma^2}\right)$, we have

$$\Pr_{x^* \sim \mu', A, w, \widehat{x} \sim \nu(\cdot|A,y)} [\|x^* - \widehat{x}\| \geq (2c+1)\,\sigma + \varepsilon] \leq e^{-\Omega(m)}.$$

This implies that there exists a set $S_{A,w}$ over A, w satisfying $\Pr_{A,w}[S_{A,w}] \geq 1 - e^{-\Omega(m)}$, such that for all $A, w \in S_{A,w}$, we have

$$\Pr_{x^* \sim \mu', \widehat{x} \sim \nu(\cdot|y)} [\|x^* - \widehat{x}\| \geq (2c+1)\sigma + \varepsilon] \leq e^{-\Omega(m)}.$$

Substituting in Eqn (5), we have

$$\Pr_{x^* \sim \mu, \widehat{x} \sim \nu(\cdot|y)} [\|x^* - \widehat{x}\| \geq (2c+1)\sigma + \varepsilon] \leq \delta + e^{-\Omega(m)}.$$

Rescaling c gives us our result.

At the beginning of the proof, we had assumed that $\mathcal{W}_\infty(\nu', Q) \leq \sigma$. If instead $\mathcal{W}_\infty(\mu', Q) \leq \sigma$, then we need to replace σ in the above bound by $\sigma + \varepsilon$. Rescaling c in the above bound gives us the Theorem statement.

□

A.3 Proof of Theorem A.5

Theorem A.5. *Let $x^* \sim \mu$ and let $y = \mathcal{A}(x^*)$ be measurements generated from x^* for some arbitrary forward operator $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^M$. Then if there exists an algorithm that uses y as inputs and outputs x' such that*

$$\|x^* - x'\| \leq \varepsilon \textit{ with probability } 1 - \delta,$$

then posterior sampling $\widehat{x} \sim \mu(\cdot|y)$ will satisfy

$$\|x^* - \widehat{x}\| \leq 2\varepsilon \textit{ with probability } \geq 1 - 2\delta.$$

Proof. Without loss of generality, let the optimal algorithm output x' that is a deterministic function of y. Yao's lemma tells us that this is as good as a randomized function of y.

Then by the statement of the Lemma, we have

$$1-\delta = \Pr[\|x^* - x'\| \leq \varepsilon] = \mathop{\mathbb{E}}_{y}\left(\Pr[\|x^* - x'\| \leq \varepsilon | y]\right).$$

Now, the probability that $\|x^* - \widehat{x}\| \leq 2\varepsilon$ can be expressed as

$$\begin{aligned}
\Pr[\|x^* - \widehat{x}\| \leq 2\varepsilon] &= \mathop{\mathbb{E}}_{y}\left(\Pr[\|x^* - \widehat{x}\| \leq 2\varepsilon | y]\right), \\
&\geq \mathop{\mathbb{E}}_{y}\left(\Pr[\|x^* - x'\| \leq \varepsilon \wedge \|x' - \widehat{x}\| \leq \varepsilon | y]\right), \\
&= \mathop{\mathbb{E}}_{y}\left(\Pr[\|x^* - x'\| \leq \varepsilon | y] \cdot \Pr[\|x' - \widehat{x}\| \leq \varepsilon | y]\right), \\
&= \mathop{\mathbb{E}}_{y}\left(\Pr[\|x^* - x'\| \leq \varepsilon | y]^2\right), \\
&\geq \left(\mathop{\mathbb{E}}_{y}\left(\Pr[\|x^* - x'\| \leq \varepsilon | y]\right)\right)^2, \\
&= (1-\delta)^2 \geq 1 - 2\delta,
\end{aligned}$$

where the second line follows from a triangle inequality, the third line follows since $x^*, \widehat{x}$ are independent conditioned on y, the fourth line follows since $\widehat{x}|y$ is distributed according to $x^*|y$, and the fifth line follows from Jensen's inequality.

□

B Appendix: Experimental Results

We perform retrospective under-sampling in all experiments, i.e., given fully-sampled k-space measurements from the NYU fastMRI [23, 42] and Stanford MRI [1] datasets, we apply sampling masks and evaluate the performance of all considered algorithms on the reconstructed data. Depending on scan parameters (e.g., 3D scans for the Stanford knee data in Appendix F), we appropriately slice and sample the data in the proper dimension so as to not commit any inverse crime [15, 31].

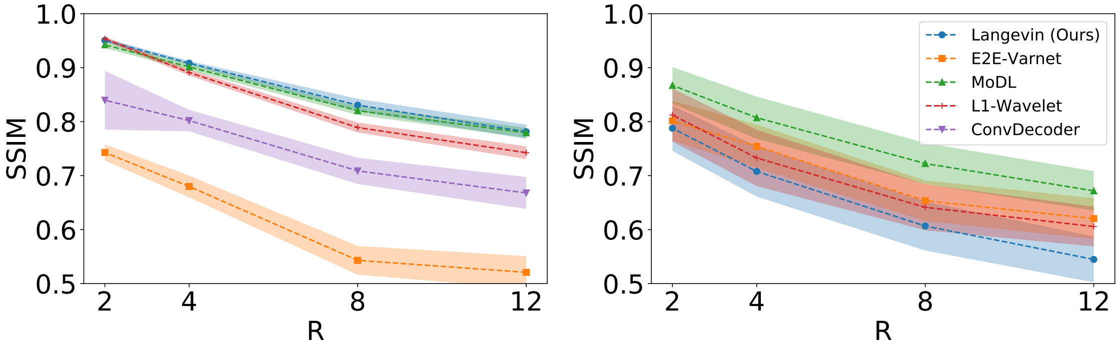
We first highlight that an advantage of the proposed approach is the invariance to the sampling scheme during training. In contrast, this is a design choice that must be made for supervised end-to-end methods, which here were trained on equispaced, vertical sampling masks, following the fastMRI 2020 challenge guidelines [42, 28]. As our results show, this affords us a significant degree of robustness across a wide distribution of sampling masks during inference.

We train a score-based model, NCSNv2 [34], on a small subset of scans from the NYU fastMRI brain dataset. Specifically, we train using T2-weighted images at a field strength of 3 Tesla for a total of 14,539 2D training slices. We calculate the MVUE from the fully sampled data and use the ESPIRiT algorithm [37, 18] applied to the fully-sampled central portion of k-space to estimate the sensitivity maps. The backbone network for our model is a RefineNet [24]. Since the generator's output is expected to be complex-valued, we treat the real and imaginary parts as separate image channels. Details about the architectures are given in Appendix G.

We use an ℓ_1-Wavelet regularized reconstruction algorithm [25] as a parallel imaging and compressed sensing baseline. We use the publicly available implementation from the BART toolbox [38, 36] and optimize the regularization hyper-parameter using the same subset of samples from the brain dataset that was used to train our method. We find that $\lambda = 0.01$ performs the best on the training data and use this value for all experiments.

We consider three different deep learning baselines: MoDL [2], E2E-VarNet [35], and the ConvDecoder architecture [11]. The first two methods, MoDL and E2E-VarNet, use supervised learning and belong to the general class of unrolled optimization methods, while the ConvDecoder is a self-supervised method that optimizes an individual model per scan.

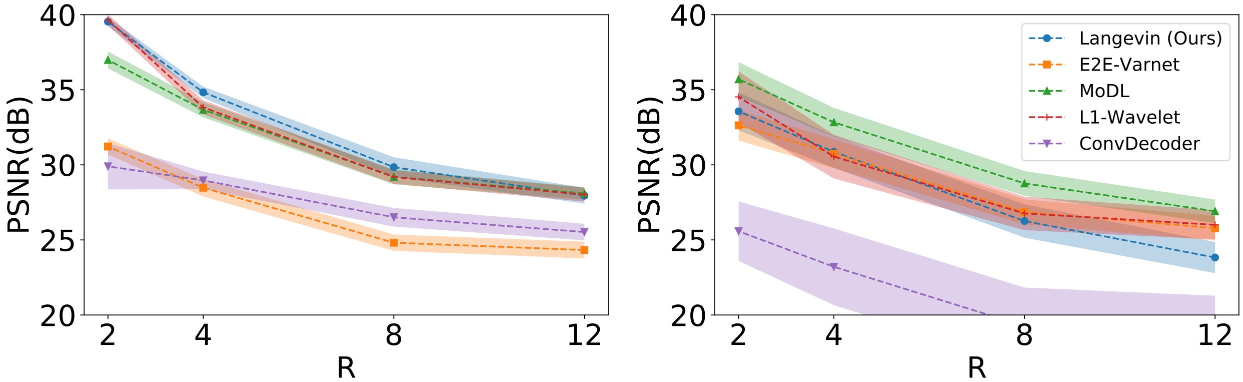
We train the MoDL baseline using a residual network backbone on the same training dataset as our method, at an acceleration factor $R = \{3, 6\}$ and equispaced under-sampling, with a supervised SSIM loss on the MVUE image.



(a) Abdomens

(b) Knees

Figure 4: Average test SSIM in various scenarios, across a range of accelerations factors R. Higher R indicates a smaller number of acquired measurements. Our approach shows the best performance and lowest reconstruction variance both in- and out-of-distribution at test-time. Shaded regions indicate 95% confidence intervals.



(a) Abdomens

(b) Knees

Figure 5: Average test PSNR in various scenarios, across a range of accelerations factors R. Higher R indicates a smaller number of acquired measurements. Our approach shows the best performance and lowest reconstruction variance both in- and out-of-distribution at test-time. Shaded regions indicate 95% confidence intervals. Note that hyperparameters for all methods were chosen for the in-distribution setting and reused under data shifts.

For the E2E-VarNet baseline, we train a U-Net with a depth of four stages, with 18 hidden channels in the first stage, for a total of 29 million learnable parameters. This model also include a smaller deep neural network that is used to estimate the sensitivity maps. This is also a U-Net, with four stages, but only eight hidden channels after the first stage, for an additional 0.7 million parameters. The model is trained for a number of 12 unrolls, and separate image networks are used at each unroll. We train this model for 50 epochs using an Adam optimizer with default PyTorch parameters and a learning rate of 2e−4, as well as gradient clipping to a maximum magnitude of 1, on the fully-sampled MVUE reconstructions from the brain T2 contrast used to train all methods. We use a batch size of 1 for training and use a supervised SSIM loss between the absolute values of ground truth and reconstructed MVUE images at accelerations $R = \{3, 6\}$, using a vertical, equispaced sampling pattern, same as all other baselines.

For the ConvDecoder baseline, we use the architecture for brain data in [11] and optimize the number of fitting iterations on a subset of samples from the training data. We find that 10000 iterations are sufficient to reach a stable average performance at $R = 3$.

We evaluate reconstruction performance using the MVUE of the fully sampled data as a reference image and measure the structural similarity index (SSIM) [41] between the absolute values of the reconstruction and ground-truth MVUE images.

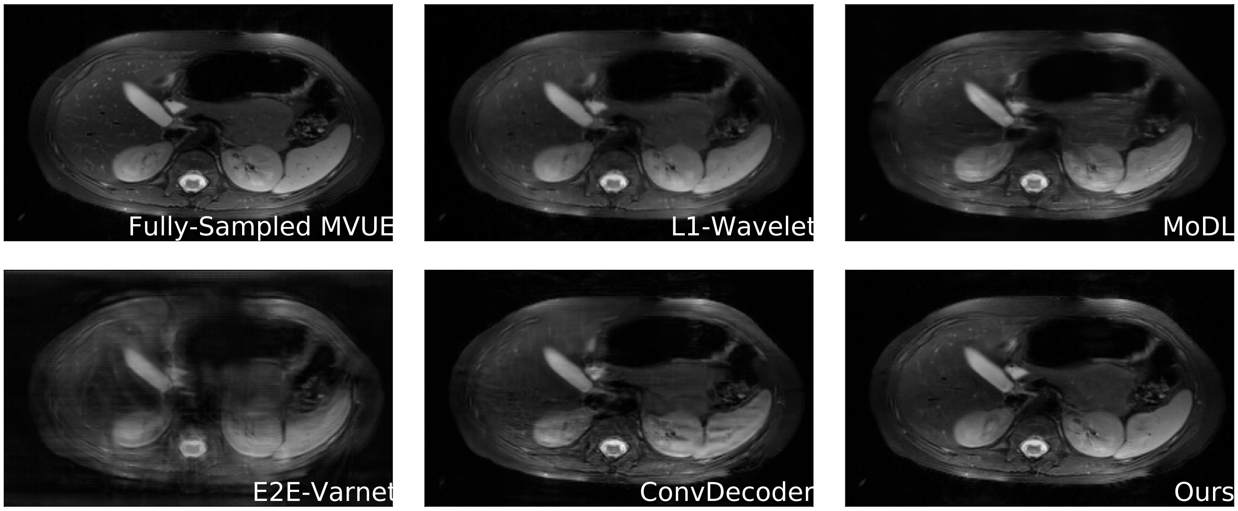


Figure 6: Comparative reconstructions of a 2D abdominal scan with uniform random under-sampling in the horizontal direction at $R = 4$. None of the methods were trained to reconstruct abdomen MRI. Our method uses a score-based generative model trained on brain images (as explained), and obtains good qualitative abdomen reconstructions.

B.1 In-Distribution Performance

In this experiment, we test all models using the same forward model that matches the training conditions for the baselines: vertical, equispaced sampling patterns. Examples of various sampling patterns are shown in Appendix C.

Figure 1 (top row) and Figure 2 show qualitative and quantitative results, respectively, for the case when there is no distributional mismatch between the training and inference sampling patterns. From Figure 2, we notice that our method surpasses all baselines in the PSNR and SSIM metric at higher acceleration factors and is competitive at lower acceleration factors. Note that at accelerations of $R \geq 6$, MoDL and E2E-Varnet attain good SSIM and PSNR scores, but Figures 1, 10, 11 show that these methods suffer from heavy aliasing. This highlights the importance of qualitative evaluations in medical image reconstruction and the limitations of existing image quality metrics [27].

We find that ℓ_1-Wavelet suffers both qualitatively and quantitatively at high acceleration factors, while the ConvDecoder is also a competitive architecture, but incurs a large computational cost. When benchmarked on an NVIDIA RTX 2080Ti GPU, our method takes 16 minutes and 0.95 GB of memory to reconstruct a high-resolution brain scan, whereas the ConvDecoder takes longer than 80 minutes and 6.6 GB of memory. While our method is limited by the inference time and is not in the range of end-to-end models (where reconstruction takes at most on the order of seconds and 3.5 GB of memory), multiple scans can be reconstructed in parallel due to the reduced memory footprint.

B.2 Out-of-Distribution Performance

Test-time anatomy shifts. We now consider the much more difficult problem of reconstructing different anatomies than the ones seen during training. This was previously investigated in [11], which concluded that all methods suffer a drastic shift due to the various changes in scan parameters between body parts. In contrast to prior work, our main finding is that the proposed score-based model retains a significant degree of robustness under these shifts, and outputs excellent qualitative reconstructions. In some cases, end-to-end methods retain robustness as well.

Figure 4c shows SSIM scores obtained on reconstructed abdominal scans obtained from [1] at an acceleration factor of $R = 4$. This represents both an anatomy and sampling pattern shift, and it can be seen that our method, MoDL and the ℓ_1-Wavelet algorithm retain their competitive advantage, while the ConvDecoder and E2E-VarNet suffer severe performance losses. Figure 6 further shows a qualitative comparison of a

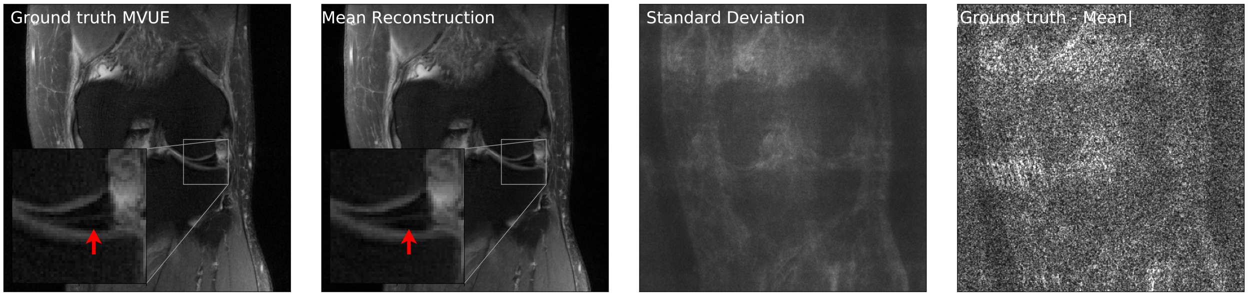


Figure 7: Our method successfully recovers fine details and can provide an estimate of the reconstruction error. The left column shows a knee from the fastMRI dataset, along with an annotated meniscus tear (indicated by red arrow in zoomed inset). Given measurements at an acceleration factor of $R = 4$, we obtain 48 independent reconstructions via posterior sampling. The second column shows the pixel-wise average of reconstructions, the third column shows the pixel-wise standard deviation, and the fourth column shows the magnitude of the error between the ground truth and the mean reconstruction. Note that our generative prior has never seen such pathology, as it was trained on T2-weighted brain scans.

reconstructed abdominal scan, with highlighted artifacts. Appendix E shows another abdomen scan.

Finally, Figure 4d shows SSIM scores obtained on fastMRI knee reconstructions, while Figure 1 (bottom row) shows the accompanying qualitative plots. This anatomy is challenging especially because of the poor signal-to-noise ratio conditions, which can be seen even in the ground-truth image. It can be noticed that this is the most severe shift for all methods, but our approach still shows the best performance at $R = 2, 4$ and a significantly lower variance. Appendix D shows more examples of knee reconstructions with and without fat suppression.

B.3 Uncertainty Estimation

Our method can also provide uncertainty estimates for each reconstructed pixel by running multiple reconstruction samplers. For a given observation y, we can obtain independent samples $\widehat{x}_1, \cdots, \widehat{x}_K \sim \mu(\cdot|y)$, for K sufficiently large. Now, using the conditional mean estimate $\bar{x} = \sum_{i=1}^{K} \widehat{x}_i / K$, we can compute the pixel-wise standard deviation $\sqrt{\sum_{i=1}^{K} |\widehat{x}_i - \bar{x}|^2 / K}$, and this gives an estimate of the error in each pixel. As shown in Fig 7, the pixel-wise standard deviation is a good estimate of the ground truth error $|x^* - \bar{x}|$. Additionally, notice that the reconstructions are able to recover fine details such as the annotated meniscus tear[2] in Fig 7 and predict low uncertainty for these features. We again emphasize that our model was not trained on knee scans.

Figure 15 in Appendix D shows another example of an annotated meniscus tear.

C Appendix: fastMRI Brain

C.1 Examples of Sampling Masks

Figure 8 shows example of some of the masks used throughout the experiments in the paper and their corresponding reconstructions. Note that the type of mask used is coupled with the scan parameters (e.g., two-dimensional slices from a three-dimensional scan will use a 2D grid of points).

We also highlight that, in all cases, a central region of the k-space is kept fully sampled and is used to estimate the coil sensitivity maps for all methods. The bottom row of Figure 8 shows naive reconstructions of a single coil image using the zero-filled k-space. This shows that different types of masks lead to different types of aliasing patterns in the image domain, motivating the need for robust image reconstruction algorithms.

[2] https://discuss.fastmri.org/t/219

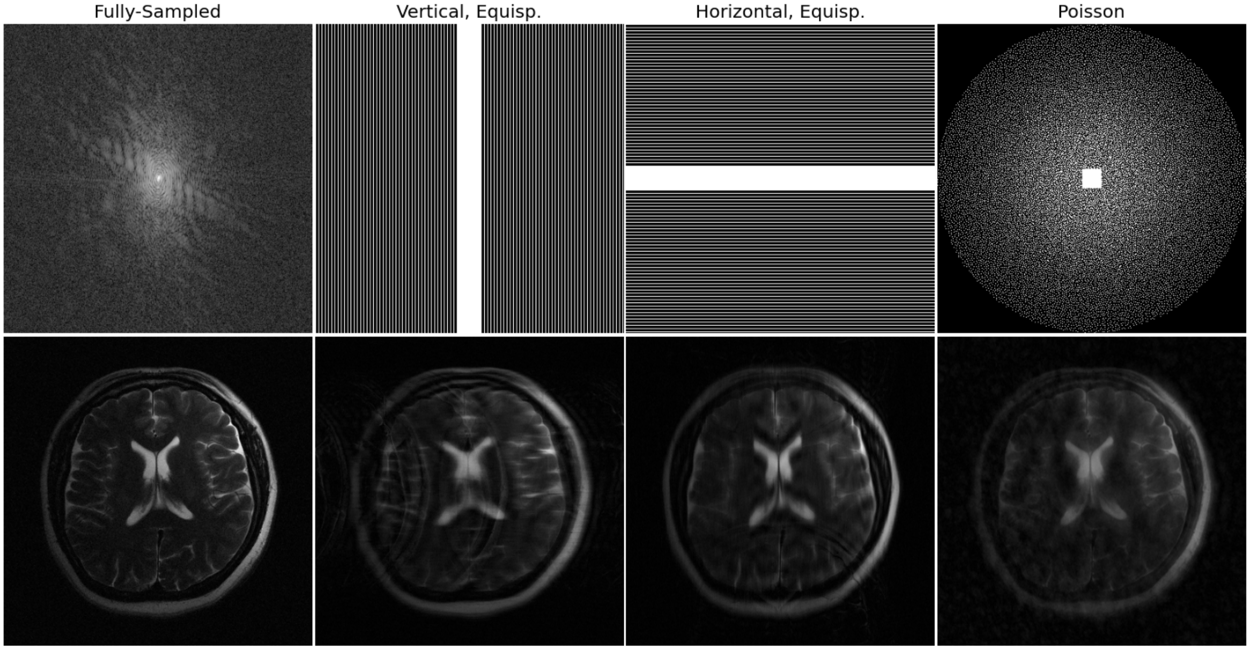


Figure 8: Examples of sampling patterns used throughout the experiments (top) and naive reconstructions (bottom). Top: The leftmost image shows the log-magnitude of the fully sampled k-space measurements corresponding to a single coil. The remaining images show three possible sampling masks, all with acceleration factor $R = 4$ but drastically different patterns. Bottom: Each image shows the magnitude of the reconstruction obtained by a two-dimensional IFFT applied to the sampled k-space.

C.2 More Exemplar Reconstructions

Figures 9 throughout 13 show detailed qualitative reconstructions on different brain scans from the fastMRI dataset. We highlight Figures 12 and 13, which represent a contrast shift from the in-distribution data (T1 and FLAIR vs. T2, respectively). Our method still produces excellent qualitative reconstructions.

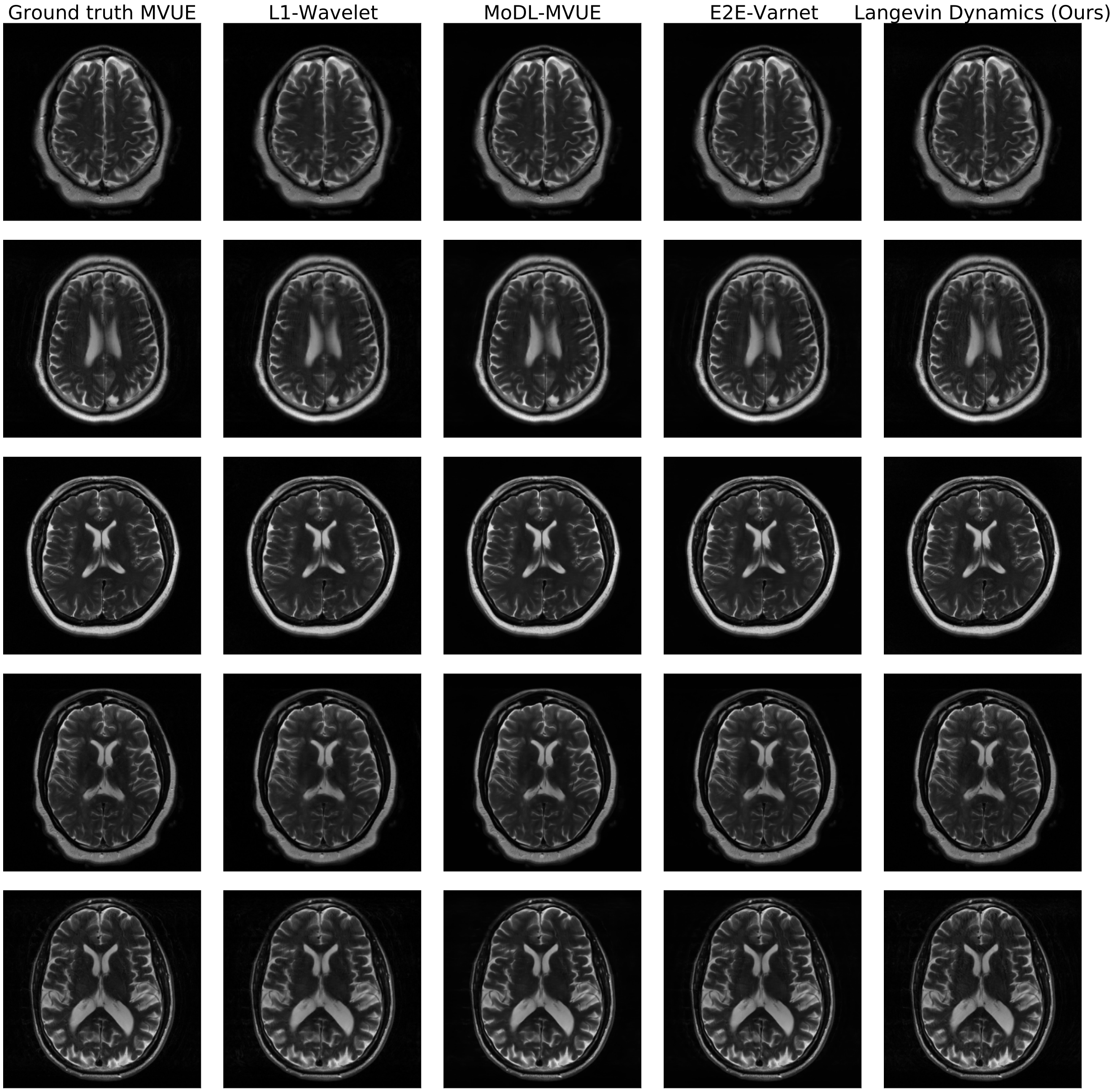


Figure 9: In-distribution brain reconstructions, at an acceleration factor of $R = 3$ and an equispaced vertical mask in k-space. Our model was trained on T2-weighted brain images from the fastMRI dataset. These results show that our method is competitive with state-of-the-art methods such as E2E-VarNet.

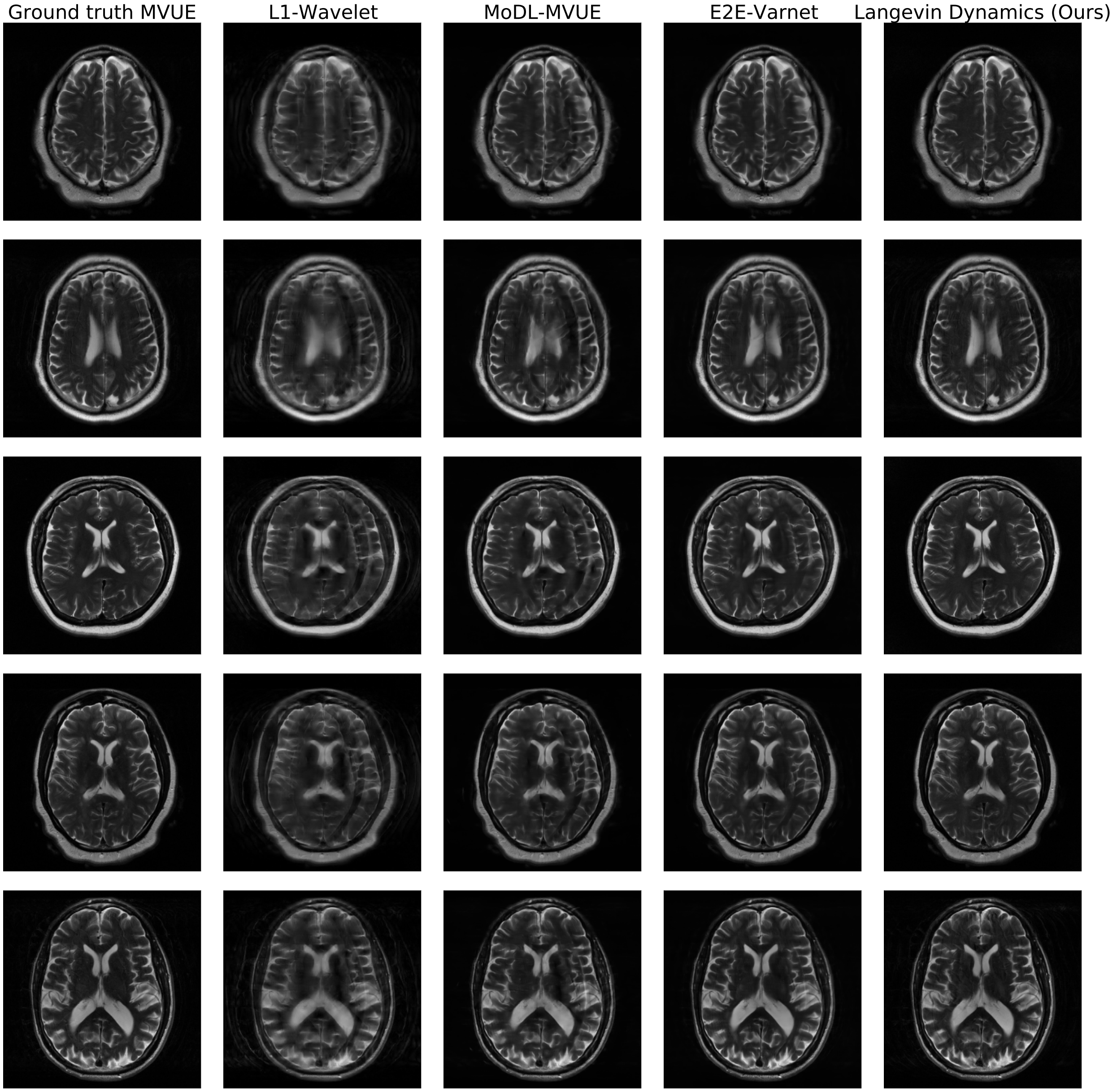


Figure 10: In-distribution brain reconstructions, at an acceleration factor of $R = 6$ and an equispaced vertical mask in k-space. Our model was trained on T2-weighted brain images from the fastMRI dataset. These results show that our method is competitive with state-of-the-art methods such as E2E-VarNet, and retains its performance at higher acceleration factors, unlike L1-Wavelet and MoDL.

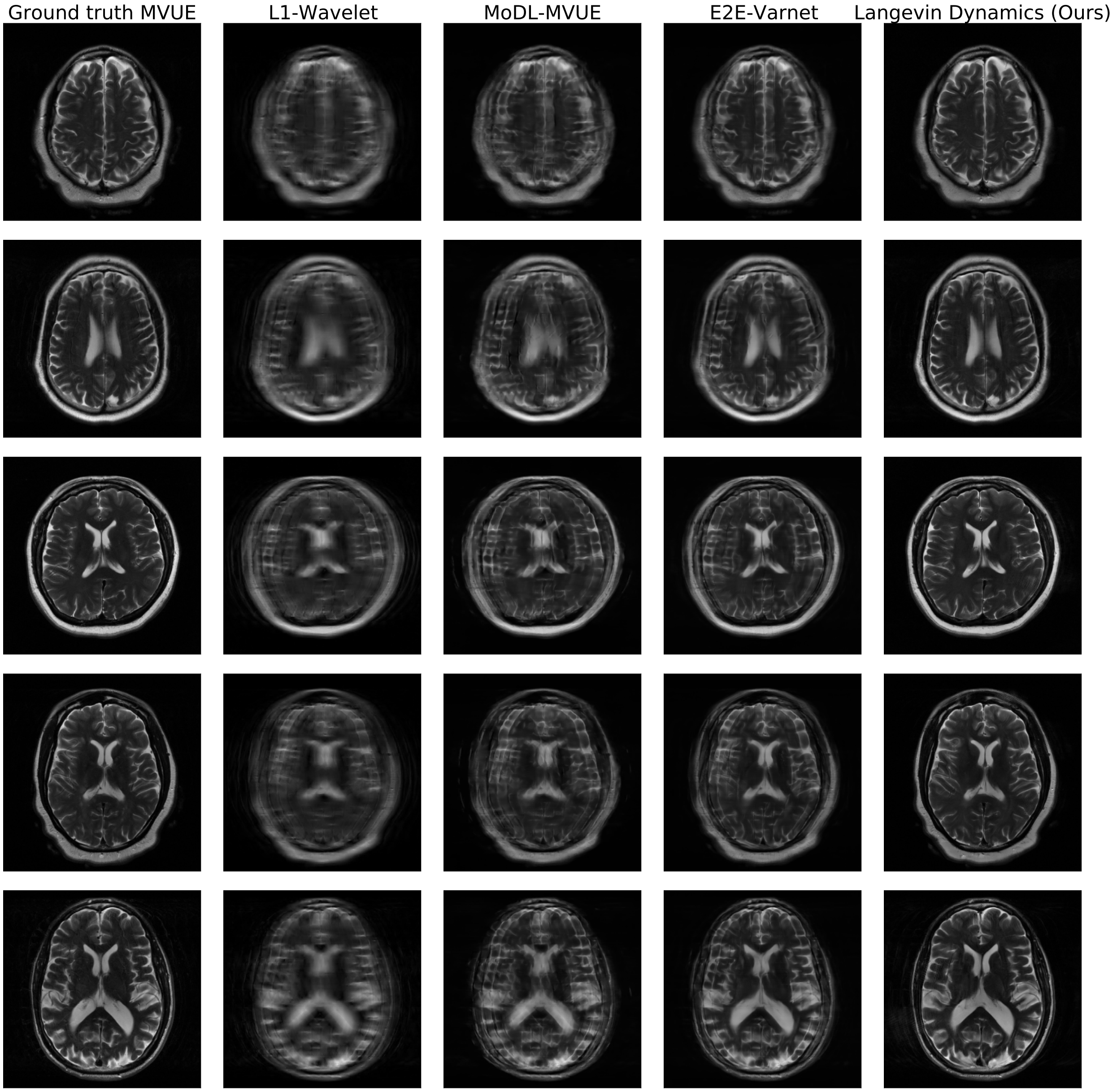


Figure 11: In-distribution brain reconstructions, at an acceleration factor of $R = 12$ and an equispaced vertical mask in k-space. Our model was trained on T2-weighted brain images from the fastMRI dataset. These results show that our method retains its performance at higher acceleration factors.

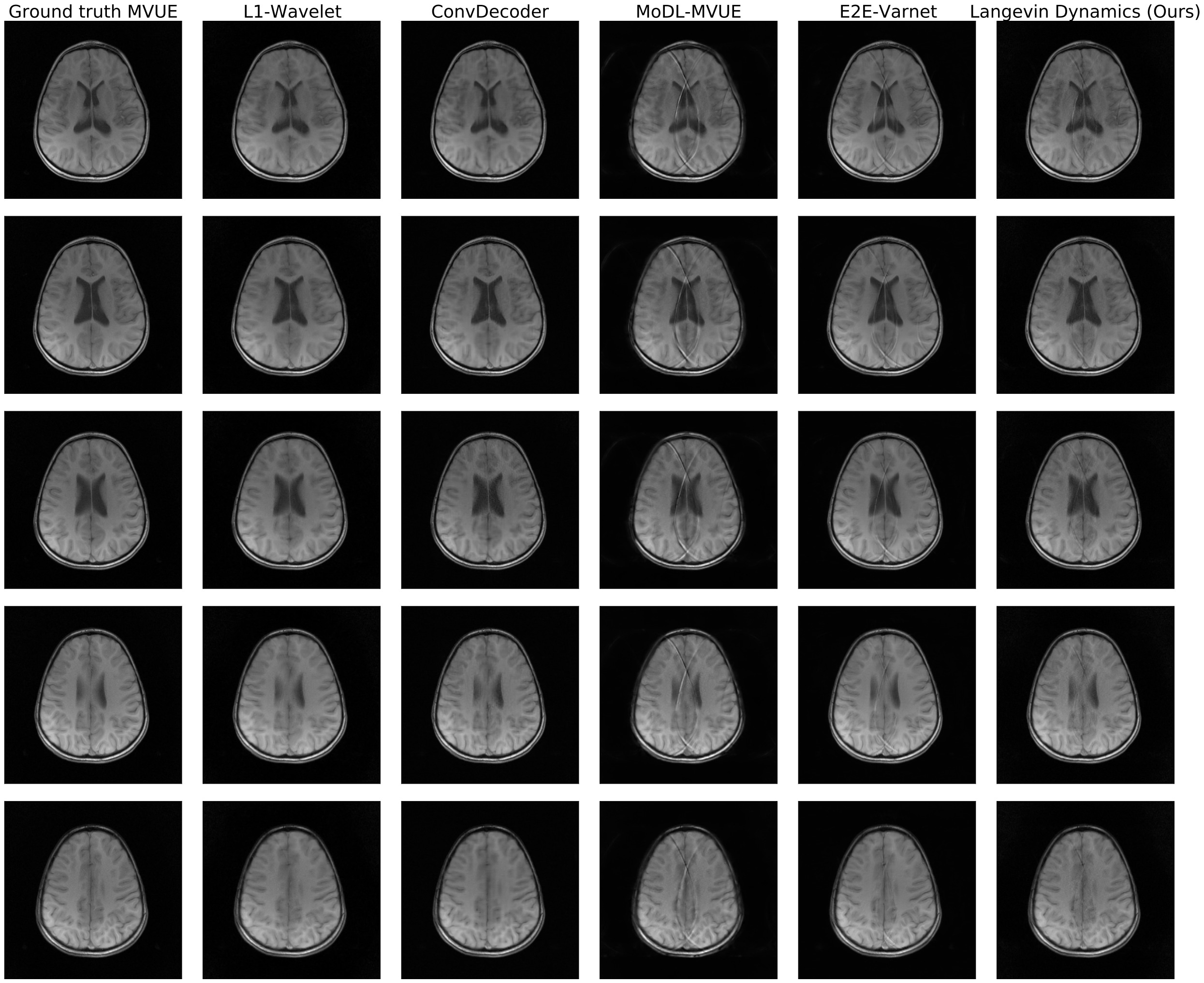


Figure 12: Brain reconstructions under a contrast shift, at an acceleration of $R = 3$. Our method was trained on T2-weighted brains, while these are T1-weighted brains, and our method is clearly robust to this contrast shift.

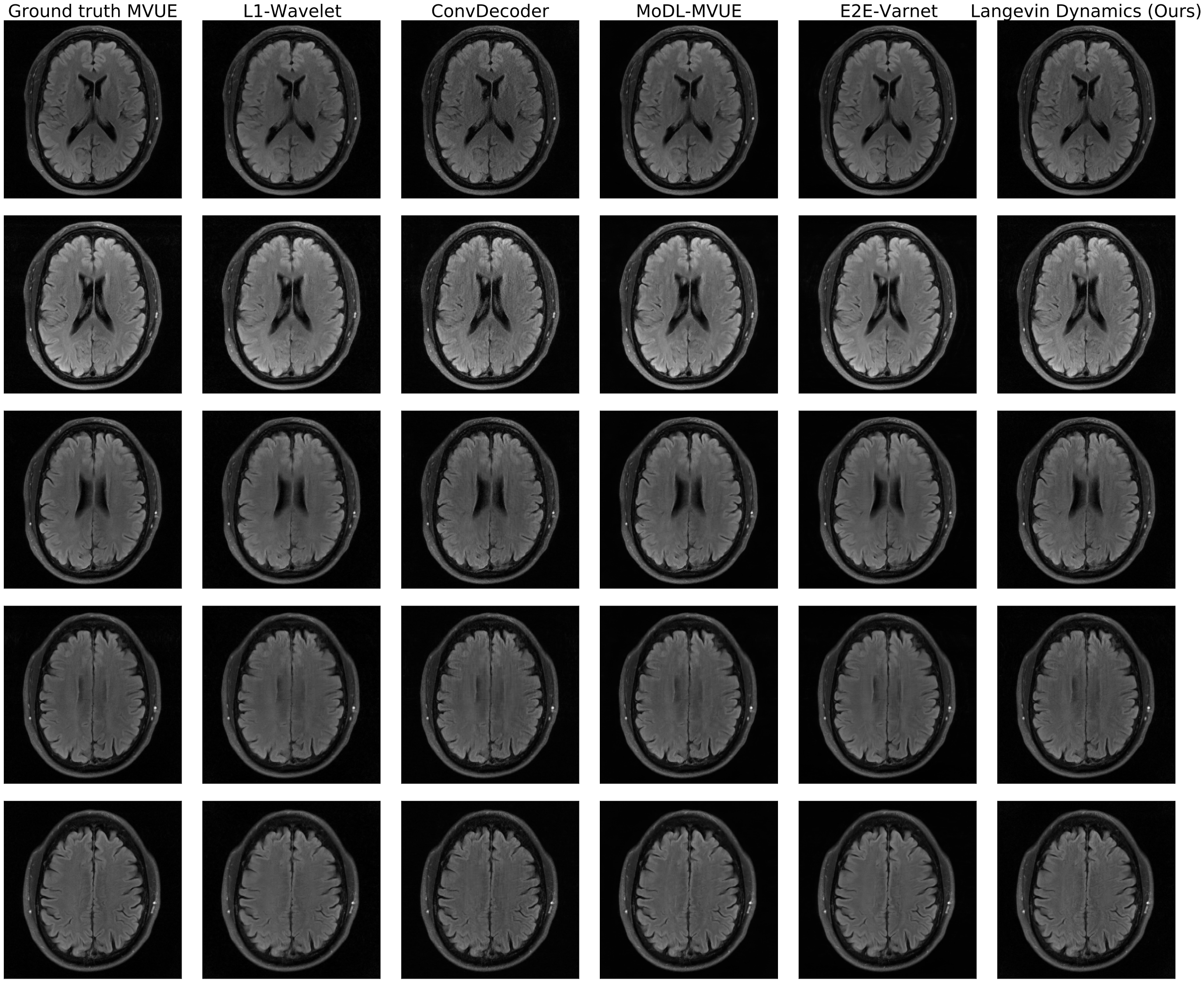


Figure 13: Brain reconstructions under a contrast shift, at an acceleration of $R = 3$. Our method was trained on T2-weighted brains, while these are FLAIR brains, and our method is clearly robust to this contrast shift.

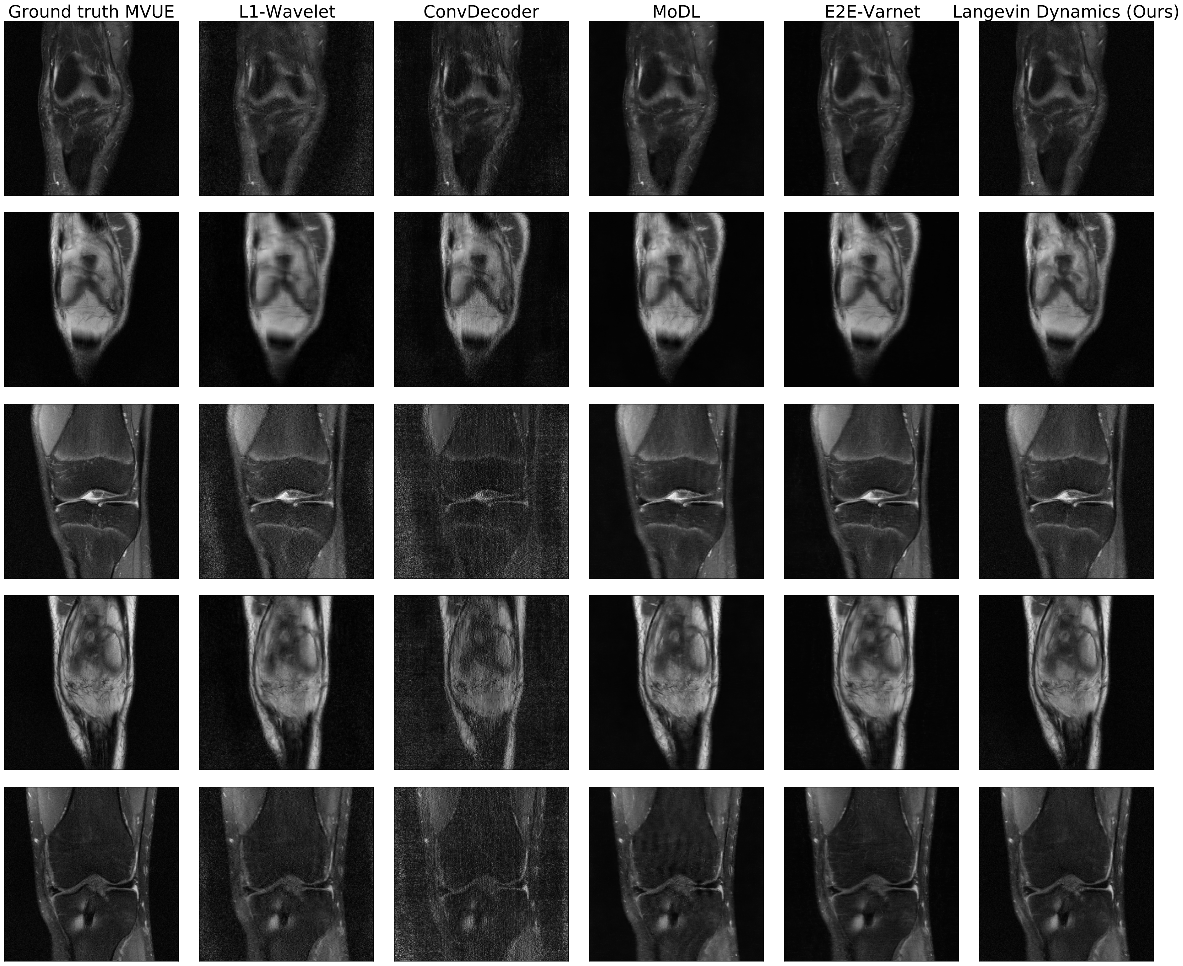


Figure 14: fastMRI knee reconstructions at an acceleration factor of $R = 4$ and a random vertical mask in k-space. All methods were trained on fastMRI brains, and this shows that our method is more robust than other methods with respect to anatomy shift. We highlight that the rows alternate between a fat suppressed (FS) and non-fat suppressed (NFS) contrast scan. The FS scans are difficult to reconstruct for the baselines, since they also exhibit a much lower signal-to-noise ratio than the brain training data, whereas our approach is robust to this type of shift as well.

D Appendix: fastMRI Knee

Figure 14 shows further examples of knee reconstructions for the fastMRI dataset. We highlight that the rows alternate between a fat suppressed (FS) and non-fat suppressed (NFS) contrast scan. The FS scans are difficult to reconstruct for the baselines, since they also exhibit a much lower signal-to-noise ratio than the brain training data, whereas our approach is robust to this type of shift as well.

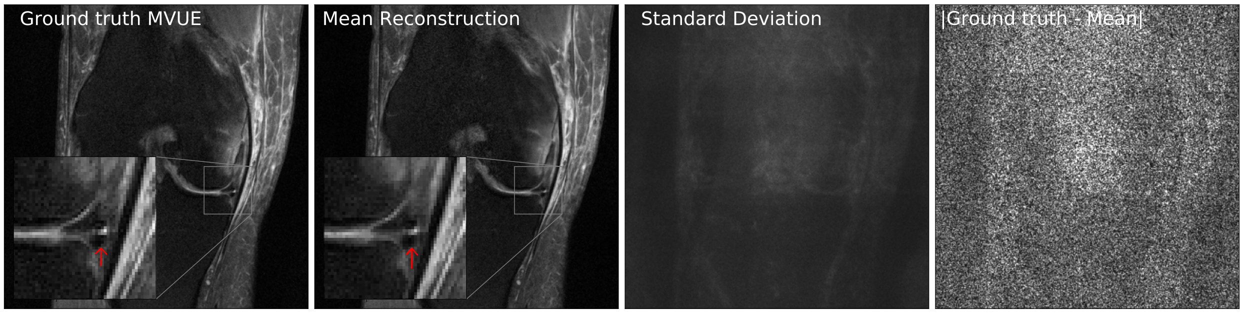


Figure 15: Our method successfully recovers fine details and can provide an estimate of the reconstruction error. The left column shows a knee from the fastMRI dataset, along with an annotated meniscus tear (indicated by red arrow in zoomed inset). Given measurements at an acceleration factor of $R = 4$, we obtain 48 independent reconstructions via posterior sampling. The second column shows the pixel-wise average of reconstructions, the third column shows the pixel-wise standard deviation, and the fourth column shows the magnitude of the error between the ground truth and the mean reconstruction. Note that our generative prior has never seen such a pathology, as it was trained on T2-weighted brain scans.

E Appendix: Abdomen

Figure 16 shows an additional example of a reconstructed abdominal scan. This is obtained from the same volume as the figure in the main text, and has a resolution of 164×320 voxels, but a much larger field of view, leading to a resolution shift for all models.

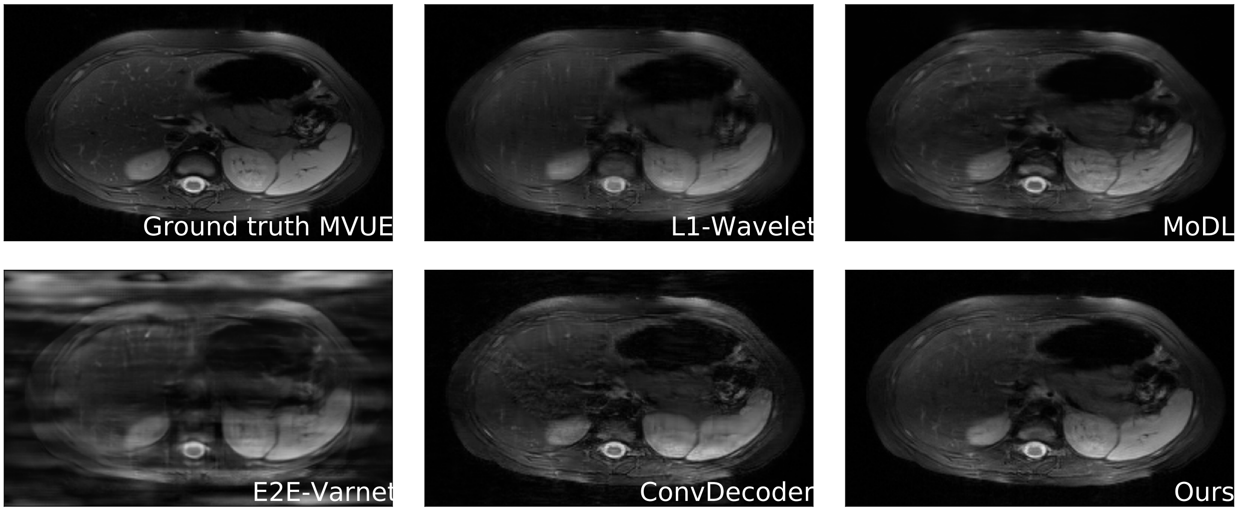


Figure 16: Comparative reconstructions of a 2D abdominal scan with uniform random under-sampling in the horizontal direction at an acceleration factor of $R = 4$. None of the methods were trained to reconstruct abdomen MRI. Our method uses a score-based generative model trained on brain images (as explained), and obtains excellent qualitative abdomen reconstructions.

F Appendix: Stanford Knee

Figures 17 and 18 show quantitative and qualitative reconstruction under an anatomy shift induced by testing axial knee scans. In this case, we first obtain a complete three-dimensional fast spin echo (3D-FSE) knee scan from the publicly available repository at `mridata.org`. To obtain two-dimensional slices, we apply an IFFT operator on the readout axis and select 24 equally spaced slices for evaluation. Each slice has a resolution of 320×256 pixels.

G Appendix: Implementation

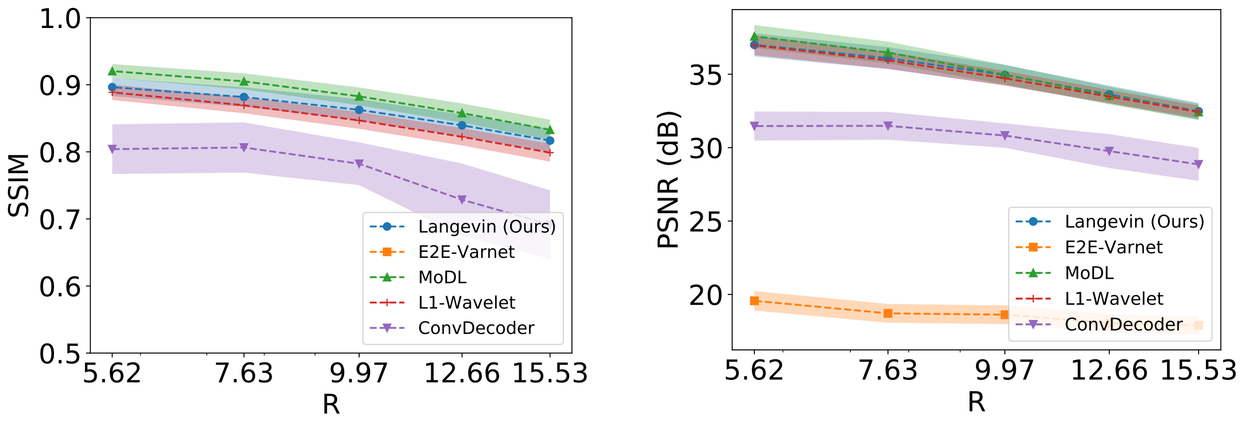
G.1 Score-Based Generative Model

Training the model We use the implementation from `https://github.com/ermongroup/ncsnv2`. As MRI data are complex valued, we changed the generator such that the output and input have two channels, one each for the real and imaginary components. We did not change the architecture otherwise.

We used the FlickrFaces (FFHQ) configs file from the NCSNv2 repo, except we set `sigma_begin` = 232, and `sigma_end` = 0.0066. This is because of the smaller number of channels in MRI when compared to FFHQ.

Dynamic range of the data. MRI data exhibits a lot of variation in the dynamic range. For example, the fastMRI dataset has max pixel value on the order of 10^{-4}, while the abdomen and Stanford knee data has max pixels on the order of 10^5. In order to deal with this variation, during *training*, we normalize each image by the 99 percentile pixel value. During inference time, when we do not have access to the ground-truth image, we normalize the reconstruction using the 99 percentile pixel value of the *pseudo-inverse.* We observe that this heuristic is sufficient to get good results.

Invariance to image shapes. Due to the convolutional nature of NCSNv2, although we trained on 384×384 images, we can still apply them to knees, T1-weighted & FLAIR brains, and abdomens, although all of these have different dimension shapes.



(a) SSIM

(b) PSNR

Figure 17: Reconstruction SSIM and PSNR on Stanford Knees as a function of the acceleration R. This dataset is considerably different from the others, as they are 3D scans. We sample k-space measurements according to Poisson masks, which gives improved incoherence, and hence we find no statistical difference between L1-Wavelet, MoDL, and our method. Note that all hyper-parameter selection and model training was done on brains from the fastMRI dataset.

Hyperparameters for inference We run annealed Langevin dynamics given in Eqn (3) for $T = 6933$ steps.

The γ_0 and γ_T hyperparameters in Eqn (3) were set to $\gamma_0 = 232, \gamma_T = 0.00066$. The γ_t values vary geometrically after every three steps, i.e.,

$$\gamma_t = \gamma_0 r^{\lfloor t/3 \rfloor},\ t = 0, \cdots, 6932,$$

where $r := \left(\frac{\gamma_T}{\gamma_0}\right)^{1/(T/3-1)} = \left(\frac{0.00066}{232}\right)^{1/2310} = 0.994487.$

The learning rate η_t in Eqn (3) also varies geometrically, as

$$\eta_t = 9 \cdot 10^{-6} \cdot \left(\frac{\gamma_t}{\gamma_T}\right)^2.$$

We use these hyperparameters for *all datasets, all accelerations, and all sampling patterns*.

G.2 E2E-VarNet Baseline

The backbone for the image reconstruction network is a U-Net with a depth of four stages, and 18 hidden channels in the first stage, for a total of 29 million learnable parameters. This model also include a smaller deep neural network that is used to estimate the sensitivity maps. This is also a U-Net, with four stages, but only eight hidden channels after the first stage, for an additional 0.7 million parameters. The model is trained for a number of 12 unrolls, and separate image networks are used at each unroll.

We train this model for 50 epochs using an Adam optimizer with default PyTorch parameters and a learning rate of 2e−4, as well as gradient clipping to a maximum magnitude of 1, on the fully-sampled MVUE reconstructions from the brain T2 contrast used to train all methods. We use a batch size of 1 for training and use a supervised SSIM loss between the absolute values of ground truth and reconstructed MVUE images at accelerations $R = \{4, 8\}$, using a vertical, equispaced sampling pattern, same as all other baselines.

Finally, it is worth mentioning that the network used to estimate the sensitivity maps explicitly uses the fully-sampled, vertical ACS region, as shown in Figure 8, both during training and inference. This makes testing with other mask patterns non-trivial for this baseline. To alleviate this, we always feed the image obtained from the vertical ACS region (for example, in the case of horizontal masks, we intentionally zero out other sampled lines that would fall in this region), to not introduce incoherent aliasing in this image.

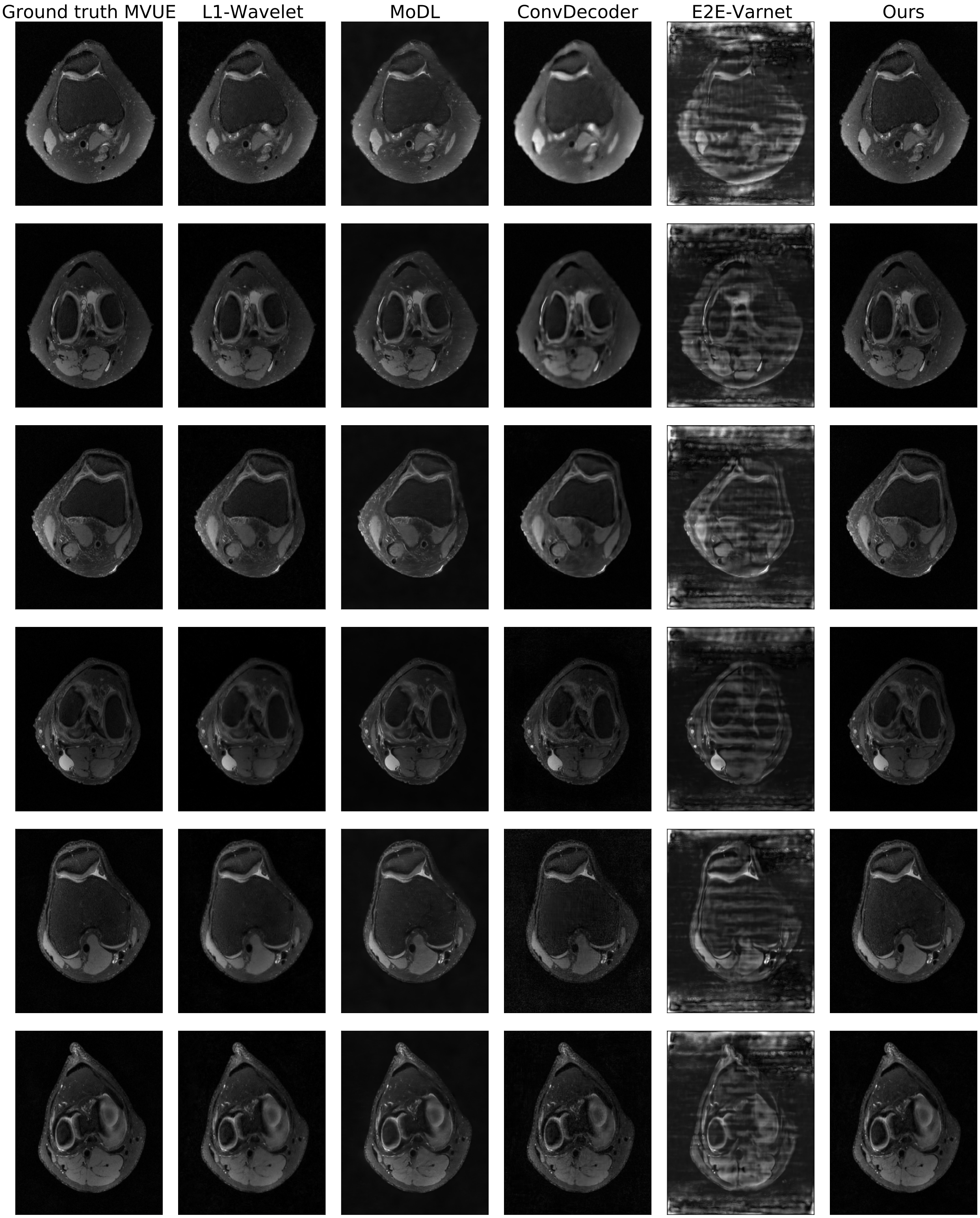


Figure 18: Qualitative reconstructions obtained by all methods on the Stanford Knees dataset at an acceleration of $R = 5.62$. This dataset is considerably different from the others, as they are 3D scans. We sample k-space measurements according to Poisson masks, which gives improved incoherence, and hence we find no statistical difference between L1-Wavelet, MoDL, and our method. Note that all hyper-parameter selection and model training was done on brains from the fastMRI dataset.

G.3 MoDL Baseline

We train a MoDL model that uses a backbone residual network with a depth of six layers, three equispaced residual connections (that feed from the first three layers to the last three layers) and 64 hidden channels, with a total of 220000 trainable parameters. Unlike E2E-VarNet, the same backbone network is used across all unrolls, and the data consistency term is given by a Conjugate Gradient (CG) operator, truncated to six steps.

We train MoDL for a number of six unrolls, leading to a total of 36 CG steps and six network applications in the unroll. We use the Adam optimizer with default PyTorch parameters and learning rate 2e−4, as well as gradient clipping to a maximum magnitude of 1. We train for five epochs, using a batch size of 1 on exactly the same T2 brain scans as all methods and a supervised SSIM loss at accelerations $R = \{4, 8\}$ using a vertical, equispaced sampling pattern and find that this is sufficient to achieve excellent in-distribution reconstruction, due to the large dataset and small network size.

Since MoDL and all other methods, except E2E-VarNet, require external sensitivity map estimates to be provided to them, we use the ESPIRiT algorithm without any eigenvalue cropping to estimate a single set of sensitivity maps, one for each coil.

H Additional Results

Figure 5 shows the test PSNR evaluated in the same conditions as Figure 4 in the main text. This highlights that our model is also robust in this metric, but also the shortcomings of PSNR in evaluating reconstruction performance, since for some methods the metric does not improve when more samples are available (R is lower).