

Stable Deep MRI Reconstruction Using Generative Priors

Martin Zach^{1b}, Florian Knoll^{1b}, *Member, IEEE*, and Thomas Pock^{1b}

Abstract—Data-driven approaches recently achieved remarkable success in magnetic resonance imaging (MRI) reconstruction, but integration into clinical routine remains challenging due to a lack of generalizability and interpretability. In this paper, we address these challenges in a unified framework based on generative image priors. We propose a novel deep neural network based regularizer which is trained in a generative setting on reference magnitude images only. After training, the regularizer encodes higher-level domain statistics which we demonstrate by synthesizing images without data. Embedding the trained model in a classical variational approach yields high-quality reconstructions irrespective of the sub-sampling pattern. In addition, the model shows stable behavior when confronted with out-of-distribution data in the form of contrast variation. Furthermore, a probabilistic interpretation provides a distribution of reconstructions and hence allows uncertainty quantification. To reconstruct parallel MRI, we propose a fast algorithm to jointly estimate the image and the sensitivity maps. The results demonstrate competitive performance, on par with state-of-the-art end-to-end deep learning methods, while preserving the flexibility with respect to sub-sampling patterns and allowing for uncertainty quantification.

Index Terms—MRI, generative priors, stable inverse problems.

I. INTRODUCTION

A VAST literature is dedicated to reducing examination time in MRI while retaining the diagnostic value of the resulting images. On the hardware side, parallel imaging (PI) [1] exploits spatially varying sensitivity maps of coil arrays. Such PI hardware has become standard in clinical systems, but noise amplification limits the potential speed up with classical reconstruction techniques [2]. On the algorithmic side, compressed sensing (CS) theory [3] and variational approaches enable greater acceleration under the assumption that the reconstruction has a sparse representation in some basis. Often, this basis is hand-crafted and only reflects rudimentary aspects of the prior information in the underlying distribution. As a prominent example, the total variation (TV)

assumes sparsity in the spatial image gradients (which translates to piecewise constant image intensities), and has been successfully applied to parallel MRI [4].

Hand-crafted prior information, in general, fails to capture the complexity of the underlying distribution [5]. On the other hand, data-driven approaches have shown promising results MRI reconstruction, often leading to superior results compared to classical CS techniques [6], [7], [8], [9], [10], [11], [12]. Here, information is encoded in a deep neural network, which is trained on reference data in an off-line step. Data-driven approaches have successfully been applied as pre-processing steps in k -space [7] as well as post-processing steps in image-space [8]. Variational networks (VNs) [6], [13], [14], [15] imitate the structure of an iterative reconstruction scheme by unrolling an optimization algorithm. In a dynamical setting, this approach has been proven useful for quantifying flow in four-dimensional MRI [16]. At the pinnacle, purely data-driven methods disregard any physical measurement model and instead learn direct maps from k -space to image-space (AUTOMAP, [17]).

Data-driven methods typically require large training datasets. In particular, state-of-the-art networks mimicking iterative schemes like the end-to-end VN [19] as well as AUTOMAP [17] (and related methods such as [10]) require image-data pairs for training. Such image-data pairs are scarce in medical applications, whereas reconstructed images are much more abundantly available in the form of DICOM data [8]. Data-driven methods that only assume access to reference images (but still allow for some kind of data-fidelity) include generative adversarial networks (GANs) [11] and score-based diffusion models [12]. Both methods share the problem that it is not obvious how the encoded prior information is best used: GANs suffer from the range-dilemma [20] and authors have proposed to optimize the parameters of the GAN at inference time [11], effectively turning it into a deep image prior [21]. Diffusion models have shown remarkable results in MRI reconstruction [12], [22], [23], but it is still an open question how to optimally incorporate data-fidelity in the reverse diffusion process (see [24] and the discussion in Section I-C for an overview of proposed methods). In addition, (with the exception of score-based diffusion models) *all* of the mentioned models effectively act as a point estimator, mapping a k -space datum to an image. However, having access to a distribution of reconstructions and, consequently, being able to quantify uncertainty in the reconstruction is of utmost importance in medical applications. Further, these point

Manuscript received 12 July 2023; accepted 30 August 2023. Date of publication 1 September 2023; date of current version 30 November 2023. (Corresponding author: Martin Zach.)

Martin Zach and Thomas Pock are with the Institute of Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria (e-mail: martin.zach@icg.tugraz.at; pock@icg.tugraz.at).

Florian Knoll is with the Department of Artificial Intelligence in Biomedical Engineering, Friedrich Alexander University of Erlangen–Nuernberg, 91052 Erlangen, Germany (e-mail: florian.knoll@fau.de).

Digital Object Identifier 10.1109/TMI.2023.3311345

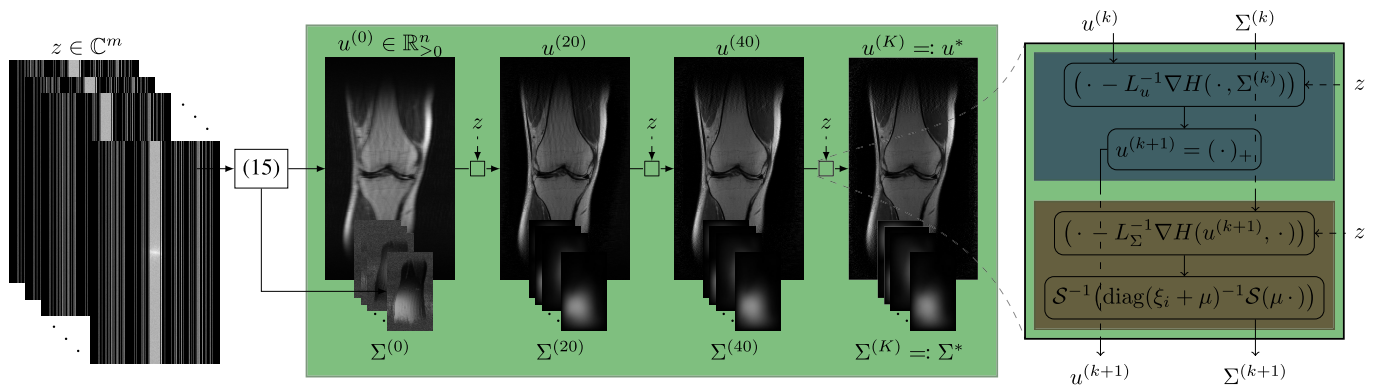


Fig. 1. Sketch of the reconstruction algorithm: To jointly reconstruct the spin density u and the sensitivity maps Σ , we impose data-fidelity, image-regularity, and coil-regularity in the iterations of iPALM [18]. The function H incorporates our learned regularizer acting on u . Details are discussed in Section II-B.

estimators typically assume a particular acquisition modality and do not generalize with respect to acquisition masks [12]. In fact, it has been shown that reconstruction quality of some data-driven approaches deteriorates when more data becomes available [25]. This severely hampers adoption of these methods in clinical practice. In any case, a method that marries the great representation power of data-driven deep neural networks with the interpretability of variational approaches is still sought after.

In this work we pursue a principled approach to parallel MRI that combines the merits of modern data-driven approaches with the benefits of classical variational approaches. In particular, we learn a highly expressive image-space prior in the well-known maximum-likelihood framework. In contrast to most other data-driven approaches, training our model does not require image-data pairs, but only assumes access to a database of reference reconstructions. To solve different PI reconstruction problems, we propose a novel algorithm to jointly estimate the image as well as the sensitivity maps in a sensitivity encoding (SENSE)-type [1] forward model. A sketch of our proposed approach is shown in Fig. 1. We demonstrate that combining this prior with suitable data-likelihood terms yields competitive performance for a multitude of PI problems. Specifically, the variational formulation makes our method agnostic to sub-sampling patterns. Additionally, the accompanying probabilistic interpretation naturally enables experts to explore the full posterior distribution of any reconstruction problem. In particular, we analyze the posterior expectation as well as the pixel-wise marginal posterior variance.

A. On the Hunt for the Optimal Regularizer

The classical variational approach for solving inverse problems amounts to minimizing an energy functional

$$\min_x E(x, z) := D(x, z) + R(x). \quad (1)$$

Here, x is the reconstruction given the datum z . The data-fidelity term D models the physical acquisition process, including stochasticity. For example, if A models the physical acquisition process, $D(x, z) = (1/2) \|Ax - z\|_2^2$ is optimal under the assumption of Gaussian measurement noise. On the other hand, the regularization term R encodes prior knowledge about the solution. Classical regularizers R include

magnitude penalization (e.g. the Tikhonov functional $x \mapsto \|x\|_2^2$) or encode smoothness assumptions (e.g. the TV $x \mapsto \sum_{i=1}^n \sqrt{(Dx)_i^2 + (Dx)_{i+n}^2}$).

Often, Eq. 1 is used without fully acknowledging its probabilistic interpretation: It computes the maximum a-posteriori (MAP) estimate of the posterior $p(x | z) \propto p(z | x)p_R(x)$ composed of a likelihood $-\log p(z | x) \propto D(x, z)$ and a prior $-\log p_R(x) \propto R(x)$. Hand-crafted regularizers, in general, can not faithfully represent the intricate statistics of the underlying distribution. Since these regularizers are overly simplistic, methods like early stopping the optimization of Eq. 1 have emerged [26]. Clearly, the need for such heuristics arises only because hand-crafted regularizers are a bad model for natural or medical images.

In this paper we pose the following question: Can we learn a regularizer such that p_R is indistinguishable from the underlying reference distribution? We answer this by encoding the regularizer as a neural network endowed with learnable parameters, which we train on reference data using maximum likelihood (see Section II). This method of learning an unnormalized probability density is known as learning an energy-based model (EBM) [27], [28], and has recently been used in the context of MRI reconstruction by [29], [30], and [31]. After training, the network can be used in arbitrary reconstruction tasks by embedding it in Eq. 1, thereby combining the versatility of variational approaches with the great representation power of deep neural networks.

On one hand, concerning versatility, the strict separation of data likelihood and prior makes this approach agnostic to the number of receiver coils and sub-sampling pattern. Further, the probabilistic interpretation enables access to a distribution of reconstructions $p(x | z)$, and we can compute different Bayesian estimators. In addition to the MAP estimate (δ_x is the Dirac measure centered at x)

$$\max_x \int p(x' | z) d\delta_x(x'), \quad (2)$$

we exploit the distribution to compute the minimum mean-squared-error (MMSE) estimate

$$\min_x \int p(x' | z) \|x - x'\|^2 dx' =: \mathbb{E}[x | z], \quad (3)$$

as well as the pixel-wise marginal variance (at the i^{th} pixel)

$$\int (x'_i - \mathbb{E}[x | z])^2 p(x'_i | z) dx'_i =: (\text{Var}[x | z])_i. \quad (4)$$

On the other hand, the representation power is empirically demonstrated by two experiments: First, our model is capable of synthesizing realistic images *without* any data. Second, we show that we can faithfully reconstruct images from severely ill-posed problems, e.g. when using random acquisition masks, where profound knowledge of the underlying anatomy is required.

B. Parallel MRI

In modern parallel MRI systems, the view of Eq. 1 is overly simplistic: These systems typically utilize coil arrays with spatially varying sensitivity maps [1] which are, in general, entirely unknown. In other words, the physical acquisition model is not well specified. Thus, for any real practical reconstruction problem, a great challenge lies in estimating these sensitivity maps accurately.

Many strategies for estimating the sensitivity maps in an off-line step have been proposed (e.g. ESPIRiT [32], [33]). Typically, these require the acquisition of fully sampled auto-calibration lines (ACL) in the k -space center such that, in essence, a low-resolution reconstruction along with the corresponding sensitivity maps can be estimated. The disadvantage of such methods is two-fold: First, assuming a non-Cartesian sub-sampling pattern, the acquisition of the ACL essentially requires an additional (albeit low-resolution) scan. This increases examination time and opens the possibility of patient movement and subsequent misalignment artifacts. Second, high-frequency regions of the k -space are not taken into account for estimating the sensitivity maps. Generally, the sensitivity maps can be assumed to be smooth (indeed we also make this assumption). However, since errors in the sensitivity estimation adversely affect downstream tasks significantly, it is vital to exploit all available data to the best extent.

In this work, similar to [4], [34], and [35], we propose to estimate the image and sensitivity maps jointly. This joint estimation hinges on the observation that the sensitivity maps are much smoother than the imaged anatomy. To enforce this smoothness and simultaneously resolve ambiguities, we impose a simple quadratic penalization on the spatial gradient of the sensitivity maps. The joint estimation allows us to utilize all available k -space data to estimate the image as well as the sensitivity maps. The resulting optimization problem can be efficiently solved by accelerated non-convex optimization algorithms. Reconstructing a parallel MRI image takes about 5 s on consumer hardware.

C. Related Work

In this section we review some related work on learned MRI reconstruction, but restrict our attention to methods that share similarities with our proposed approach. For a more general overview of data-driven MRI reconstruction we refer to [36]. The authors of [25] give a very broad overview of potential risks of using modern techniques for medical image reconstruction in general.

1) Energy-Based Models: After submission of this paper, we were made aware of concurrent works that also utilize EBMs for MRI reconstruction. In [30], the authors propose to learn a regularizer and use proximal gradient descent for inference. The main difference to our approach is that for parallel MRI, their algorithm assumes access to sensitivity maps, which have to be precomputed using, e.g., ESPIRiT [32]. Additionally, their algorithm requires hand-tuning of step-sizes and is relatively slow. On the contrary, we propose an algorithm for joint reconstruction of image and coil sensitivities which does not require hand-tuning of step-sizes, and is fast due to being accelerated.

The authors extended their work in [31], where they learn an EBM for both image- and k -space. While this obviates the need for sensitivity estimation, it requires to train two independent networks that need to be balanced at inference time. Additionally, the k -space EBM requires fully-sampled reference k -space data, which is scarcely available. In contrast, we only require reference DICOM (magnitude) images to train one network. In addition, we perform a data-independent analysis of the learned regularizer as well as uncertainty analysis.

2) Diffusion Models: Diffusion models [22], [23], [37], [38] aim to model the gradient of the log-prior while undergoing a diffusion process: Let p_t denote the data distribution at diffusion time t . The aim is to learn a time-conditional score network s_θ such that $s_\theta(\cdot, t) \approx \nabla \log p_t$ for all $t > 0$. This approach shares many similarities with EBMs, where indeed it should hold that $\nabla R = s_\theta(\cdot, 0)$.

Diffusion models can be used to generate unconditional samples from the data distribution [37]. This is computationally demanding as it requires solving a stochastic differential equation (SDE) with high accuracy, which typically needs thousand of gradient evaluations [12]. For inverse problems, it is not clear how to optimally incorporate data-fidelity into the SDE. Proposed approaches include data projection [22], annealed Langevin dynamics [38] and diffusion posterior sampling [39]. These approaches require hand-tuning of parameters, sometimes at each step of the reverse diffusion [38], [39]. The recent work of [24] showed that none of these methods generate samples from the true posterior distribution and propose to augment the score models with normalizing flows. While their inference algorithm is parameter-free, the approach is opaque due to the introduction of the normalizing flow and is still computationally demanding.

In contrast, EBMs enjoy a natural probabilistic interpretation with access to an analytic expression of the posterior. In addition, MAP inference does not require solving an SDE, but can be done by efficient optimization algorithms. Having access to the function value (as opposed to only the gradient) also has practical applications, such as being able to inspect the regularization landscape (see Fig. 2) and utilizing backtracking in optimization algorithms [18].

In the works of [23] and [38], parallel imaging is tackled by off-line sensitivity estimation, which comes with the drawbacks outlined in Section I-B. The authors of [12] propose to reconstruct individual coil images using a model trained solely on root-sum-of-squares (RSS) reconstructions. While

the results are impressive, the computational cost consequently depends on the number of coils utilized in the physical scanner, and the authors report reconstruction times of up to 10 min. In our joint reconstruction algorithm, the gradient of the network is only evaluated once per iteration, irrespective of the number of coils. Imposing spatial regularity on the coils is extremely fast, where at each iteration we only have to compute very few fast Fourier transforms (see Section II-B).

3) Joint Reconstruction: The common approach to parallel imaging is based on a two-step approach: Coil sensitivities are estimated (typically from ACL regions) in an off-line step, after which the image is reconstructed. A joint reconstruction was first proposed by [35] based on alternating minimization, where they explicitly parametrize the sensitivities with low-order polynomials. In [34], the authors propose an iteratively regularized Gauss-Newton algorithm that enforces spatial smoothness of the sensitivity maps during the iterations. Their algorithm is not guaranteed to converge for arbitrary initializations, requires hand-tuning of parameters at each update step and leads to noisy solutions when run for too long. This algorithm was extended to incorporate classical variational penalties, such as the TV [40] or the Total Generalized Variation (TGV) [41], by the authors of [4], in order to suppress noisy solutions.

Indeed, their approach shares a lot of similarities with our proposed approach. The differences can be summarized as follows: Instead of hand-crafted regularizers, we employ modern generative learning techniques to learn an expressive regularizer from data, which leads to state-of-the-art reconstructions. In addition, instead of the iteratively regularized Gauss-Newton algorithm with additional non-trivial sub-problems for the variational penalties, we employ the inertial proximal alternating linearized minimization (iPALM) algorithm [18] for optimization. Thus, we can guarantee convergence and only require hand-tuning of two parameters (see Section II-B). A sketch of our reconstruction algorithm is shown in Fig. 1.

For completeness, we mention that the end-to-end variational network of [19] also estimates the sensitivities jointly with the image. However, their approach differs significantly to ours in that they learn a mapping from k -space to image-space *discriminatively*, where they utilize the estimated sensitivity maps to enforce data fidelity. Thus, the network only works well with a particular sub-sampling pattern and coil configuration. In contrast, we learn image features *generatively* and impose hand-crafted spatial regularity onto the sensitivity maps, therefore being agnostic to the sub-sampling pattern as well as coil configuration.

II. METHODS

A. Maximum-Likelihood Training

To learn a regularizer such that its induced distribution is indistinguishable from the underlying reference distribution we proceed as follows: We equip the regularizer with parameters and denote with $\{R_\theta : \theta \in \Theta\}$ the family of θ -parametrized functions $R_\theta : \mathcal{X} \rightarrow \mathbb{R}$. Θ is a suitably selected set of parameters, \mathcal{X} is the space of the underlying distribution.

We discuss our particular choice of the θ -parametrized family in Section II-D. R_θ induces a Gibbs distribution on \mathcal{X} with density

$$p_\theta(x) := \exp(-R_\theta(x))Z_\theta^{-1}, \quad (5)$$

where $Z_\theta = \int_{\mathcal{X}} \exp(-R_\theta(\xi)) d\xi$ is the partition function. We find the optimal parameters by maximizing the likelihood of reference data — drawn from a reference distribution p_{data} — under our model:

$$\min_{\theta \in \Theta} \{\Gamma(\theta) := \mathbb{E}_{x \sim p_{\text{data}}}[-\log p_\theta(x)]\}. \quad (6)$$

The loss function Eq. 6 admits the gradient [42] $\nabla \Gamma(\theta) = \mathbb{E}_{x \sim p_{\text{data}}}[\nabla_\theta R_\theta(x)] - \mathbb{E}_{x \sim p_\theta}[\nabla_\theta R_\theta(x)]$.

For any interesting regularizer, computing the partition function Z_θ is intractable. Thus, we resort to Markov chain Monte Carlo (MCMC) techniques to approximate $\mathbb{E}_{x \sim p_\theta}[\nabla_\theta R_\theta(x)]$. In particular, following [27] we use the unadjusted Langevin algorithm (ULA) [43], which iterates

$$x^{(j)} \sim \mathcal{N}\left(x^{(j-1)} + \frac{\zeta}{2} \nabla \log p_\theta(x^{(j-1)}), \zeta \text{Id}\right), \quad j \in \llbracket 1, J \rrbracket. \quad (7)$$

Here, $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution on \mathcal{X} with mean μ and covariance Σ , and $\llbracket \cdot, \cdot \rrbracket$ denotes an integer interval. Further, $\zeta > 0$ is the discretization time step of the associated continuous-time Langevin diffusion stochastic differential equation, which is known to be p_θ -stationary [43], [44]. The time discretization biases ULA asymptotically, which could be corrected via metropolization. However, the poor non-asymptotic performance of metropolized ULA [45], [46] makes this less useful in practice, and we do not use metropolization. Note that by Eq. 5 $\nabla \log p_\theta = -\nabla R_\theta$.

In general, maximum-likelihood based learning of EBMs is known to be unstable [28], [47] due to the slow convergence of ULA (7). To minimize the burn-in time and aid convergence of our sampler, we utilize persistent initialization [48], [49]: Assuming p_θ changes only slightly during each update to θ , samples from previous learning states are good initial guesses for the current iteration. We implement this idea using a replay buffer that holds samples from previous iterates. After iterating (7) starting from $x^{(0)}$ drawn from the replay buffer, we write $x^{(j)}$ back into the buffer with $1 - \pi_{\text{reinit}}$ chance. Otherwise, we draw a random sample from p_{data} to write it to the buffer. Upon reinitialization, to help mode coverage, we randomly permute the pixels of (on average) every second sample. We refer to [27], [42], and [28] for more information of maximum-likelihood training of EBMs.

B. Reconstruction Algorithm

We utilize an image-space SENSE-type [1] forward model for parallel MRI. In detail, we relate the real-valued spin density $x \in \mathbb{R}_{\geq 0}^n$ to the noisy measurement data $z = (z_1, \dots, z_C)^\top \in \mathbb{C}^m$. Here, $m = CK$ combines the K acquired k -space data points (not necessarily on a Cartesian grid) of

$C \in \mathbb{N}$ receiver coils. The forward operator

$$A: \mathbb{R}^n \rightarrow \mathbb{C}^m, \quad x \mapsto \begin{pmatrix} \mathcal{F}_M(\sigma_1 \odot x) \\ \mathcal{F}_M(\sigma_2 \odot x) \\ \vdots \\ \mathcal{F}_M(\sigma_C \odot x) \end{pmatrix} \quad (8)$$

utilizes the sensitivity maps $(\sigma_c)_{c=1}^C \in (\mathbb{C}^n)^C$ of the C receiver coils, as well as the sampling operator $\mathcal{F}_M: \mathbb{C}^n \rightarrow \mathbb{C}^K$ defined by a k -space trajectory M . In our model, we assume a real-valued spin-density and empirically demonstrate good performance on the fastMRI dataset [8] in Section III-D. In phase-sensitive imaging however, a complex spin-density needs to be assumed. Our approach can be generalized to this setting by, e.g. splitting real and imaginary channels as in [11] and [19], and training the regularizer on this data.

To ease notation, we define $\Sigma := (\sigma_c)_{c=1}^C$, $\mathcal{C} := (\mathbb{C}^n)^C$, and $|\cdot|_{\mathcal{C}}: \mathcal{C} \rightarrow \mathbb{R}_{\geq 0}^n$, $(\sigma_c)_{c=1}^C \mapsto \sqrt{\sum_{c=1}^C |\sigma_c|^2}$. ($|\cdot|$ is the complex modulus acting element-wise on its argument.) In the case of sub-sampling on a Cartesian grid, \mathcal{F}_M encodes the Fourier transform followed by the multiplication with a binary mask. In such SENSE-type setups, to quantitatively compare to fully-sampled RSS reconstructions, the spin density x has to be re-weighted by $|\Sigma|_{\mathcal{C}}$ [34]: Denoting the reconstructed image $u \in \mathbb{R}_{\geq 0}^n$, it is given by $u = x \odot |\Sigma|_{\mathcal{C}}$, which motivates an immediate change of variables

$$x = u \oslash |\Sigma|_{\mathcal{C}}. \quad (9)$$

Instead of estimating the spin density (and the sensitivity maps), we propose to directly estimate the image u (and the sensitivity maps). Formally, given the noisy measurement data $z = (z_1, \dots, z_C)^{\top}$ we propose to find

$$\arg \min_{(u, \Sigma) \in \mathbb{R}^n \times \mathcal{C}} H(u, \Sigma) + \delta_{\mathbb{R}_{\geq 0}^n}(u) + \mu F(\Sigma) \quad (10)$$

where

$$H: \mathbb{R}^n \times \mathcal{C} \rightarrow \mathbb{R},$$

$$(u, (\sigma_c)_{c=1}^C) \mapsto \frac{1}{2} \left\| \begin{pmatrix} \mathcal{F}_M(\sigma_1 \odot u \oslash |\Sigma|_{\mathcal{C}}) \\ \vdots \\ \mathcal{F}_M(\sigma_C \odot u \oslash |\Sigma|_{\mathcal{C}}) \end{pmatrix} - \begin{pmatrix} z_1 \\ \vdots \\ z_C \end{pmatrix} \right\|_2^2 + \lambda R(u) \quad (11)$$

combines data fidelity and image-regularization and

$$F: \mathcal{C} \rightarrow \mathbb{R}_{\geq 0},$$

$$(\sigma_c)_{c=1}^C \mapsto \frac{1}{2} \sum_{c=1}^C (\|D\Re(\sigma_c)\|_2^2 + \|D\Im(\sigma_c)\|_2^2) \quad (12)$$

encodes the smoothness prior on the sensitivity maps. In the above, $\delta_{\mathbb{R}_{\geq 0}^n}$ enforces non-negativity on the spin density using the indicator function

$$\delta_{\mathcal{A}}: \mathbb{R}^n \rightarrow \{0, \infty\}, \quad x \mapsto \begin{cases} 0 & \text{if } x \in \mathcal{A}, \\ \infty & \text{else.} \end{cases} \quad (13)$$

$D: \mathbb{R}^n \rightarrow \mathbb{R}^{2n}$ is the discrete gradient operator such that (Dx) contains vertically stacked horizontal and vertical first-order finite differences (see, e.g., [50]). It implements physically

motivated Dirichlet boundary conditions ($\sigma_c = 0$ outside of the image domain $\forall c$). $\lambda \in \mathbb{R}_+$ and $\mu \in \mathbb{R}_+$ are scalars trading off the strength of the regularization on the image and the sensitivity maps respectively.

Observe that the division by $|\Sigma|_{\mathcal{C}}$ in Eq. 11 follows from the identification Eq. 9. This formulation has the advantage that the sensitivity maps $(\sigma_c)_{c=1}^C$ are implicitly normalized, and we do not have to impose a normalization constraint explicitly during optimization. Further, the existence of a minimizer is ensured by coercivity of Eq. 10, which is a direct result of the enforced Dirichlet boundary conditions.

We solve Eq. 10 using the iPALM algorithm [18] with Lipschitz backtracking, summarized in Algorithm 1. The algorithm utilizes the combined fidelity-regularity functional H defined in (11). For our learned regularizer, we would have $R = R_{\theta}$ as in (16), whereas for a hand crafted regularizer R may be TV (18). All gradients are understood in the $\mathbb{C}\mathbb{R}$ -sense [51]. Recall that the proximal operator $\text{prox}_{\alpha G}: \mathcal{H} \rightarrow \mathcal{H}$ of a proper extended real-valued function $G: \mathcal{H} \rightarrow]-\infty, \infty]$ (\mathcal{H} is a Hilbert space) is the map $\bar{x} \mapsto \arg \min_x (1/2) \|\bar{x} - x\|^2 + \alpha G(x)$. Observe that $\text{prox}_{\delta_{\mathbb{R}_+^n}}$ retrieves the positive part of its argument, that is $\text{prox}_{\delta_{\mathbb{R}_+^n}}(x) = (x)_+$. $\text{prox}_{\mu F}$ can be solved in closed form by utilizing the discrete sine transform (see [52, Chapter 19.4] for a more rigorous discussion) as (denoting $\hat{i} := \sqrt{-1}$)

$$(\sigma_c)_{c=1}^C \mapsto (Q_{\mu}(\Re(\sigma_c)) + \hat{i} Q_{\mu}(\Im(\sigma_c)))_{i=1}^C. \quad (14)$$

Here, $Q_{\mu}: y \mapsto \mathcal{S}^{-1}(\text{diag}(\xi_i + \mu)^{-1} \mathcal{S}(\mu y))$ uses the discrete sine transform \mathcal{S} and the eigenvalues ξ_i of the discrete Laplace operator, which are of the form $\xi_i = 2 - 2 \cos \phi_i$ for equally spaced angles ϕ_i . We initialize the algorithm with the zero-filled (ZF) RSS reconstruction and the corresponding sensitivity maps (for all $c \in \llbracket 1, C \rrbracket$)

$$u^{(0)} = \sqrt{\sum_{c=1}^C |\mathcal{F}_M^*(z_c)|^2} \quad \text{and} \quad \sigma_c^{(0)} = \mathcal{F}_M^*(z_c) \oslash u^{(0)}. \quad (15)$$

A sketch of the reconstruction algorithm is shown in Fig. 1.

Algorithm 1 iPALM [18] Instantiation to Solve (10).
bt Is Algorithm 2.

Input : $(u^{(0)}, \Sigma^{(0)}) \in \mathbb{R}^n \times \mathcal{C}$, $K \in \mathbb{N}$, $\gamma_1 \in (0, 1)$,
 $\gamma_2 \in (0, 1)$, **initial** $(L_x, L_{\Sigma}) \in \mathbb{R}_+ \times \mathbb{R}_+$
Output: $(u^{(K)}, \Sigma^{(K)})$ solving Eq. 10

- 1 $(u^{(1)}, \Sigma^{(1)}) = (u^{(0)}, \Sigma^{(0)})$
- 2 **for** $k \in \llbracket 1, K-1 \rrbracket$ **do**
- 3 $\bar{u} = u^{(k)} + \frac{k}{k+3}(u^{(k)} - u^{(k-1)})$
- 4 $(u^{(k+1)}, L_x) = \text{bt}(H(\cdot, \Sigma^{(k)}), \delta_{\mathbb{R}_+^n}, \bar{u}, L_x, \gamma_1, \gamma_2)$
- 5 $\bar{\Sigma} = \Sigma^{(k)} + \frac{k}{k+3}(\Sigma^{(k)} - \Sigma^{(k-1)})$
- 6 $(\Sigma^{(k+1)}, L_{\Sigma}) = \text{bt}(H(u^{(k+1)}, \cdot), \mu F, \bar{\Sigma}, L_{\Sigma}, \gamma_1, \gamma_2)$

C. Experimental Data

For all experiments we utilize the fastMRI knee dataset [53]. Specifically, the training data are the RSS reconstructions of

Algorithm 2 Backtracking Procedure to Find the Local Lipschitz Constants in Algorithm 1.

Input : $E, P, x_0, L_0, \gamma_1, \gamma_2$

Output: (x, L)

```

1  $L \leftarrow L_0$ 
2 for ever do
3    $x = \text{prox}_{L^{-1}P}(x_0 - L^{-1}\nabla E(x_0))$ 
4    $d = x - x_0$ 
5   if  $E(x) \leq E(x_0) + \langle \nabla E(x_0), d \rangle + \frac{L}{2} \|d\|_2^2$  then
6      $L \leftarrow \gamma_1 L$ 
7     break
8   else  $L \leftarrow L/\gamma_2$ 

```

size $\tilde{n} = 320 \times 320$ of the multi-coil coronal proton-density weighted (CORPD) training split. We used the central 11 slices to ensure reasonable training data, resulting in a total of 5324 training slices. To have consistent intensity ranges during training, we normalized each slice by $x \mapsto ((x - \min_i x_i) / (\|x\|_\infty - \min_i x_i))$ individually (this normalization was not performed for any of the reference methods). For validation and testing, we used the multi-coil CORPD validation split, discarding samples with width $w \notin \{368, 372\}$, leaving 91 scans. The scans were split into 30 validation samples and 61 test samples by lexicographic ordering of the filenames. To be consistent with training, we again restrict our interest to the central 11 slices, resulting in 330 validation slices and 671 test slices. For the out-of-distribution experiments, we used the central 11 slices of the coronal proton-density weighted fat-suppressed (CORPD-FS) scans (again excluding width $\notin \{368, 372\}$) in the fastMRI knee validation dataset.

D. Network Architecture and Implementation Details

The family of θ -parametrized functions we consider in this work follows the simple structure

$$R_\theta = U \circ FC \circ S_L \circ S_{L-1} \circ \dots \circ S_2 \circ S_1 \circ \mathcal{C}, \quad (16)$$

where $\mathcal{C}: \mathbb{R}^n \rightarrow \mathbb{R}^{\tilde{n}}$ center-crops the validation images of size $n = 640 \times w$ to the size of the training images. The layer $S_l: x \mapsto \mathcal{L}(\tilde{W}_l \mathcal{L}(W_l x + b_l) + \tilde{b}_l)$, $l \in \llbracket 1, L \rrbracket$ utilizes the leaky ReLU $\mathcal{L}: x \mapsto \max\{\gamma x, x\}$ with leak coefficient $\gamma = 0.05$. The weights and biases $\{W_l, b_l, \tilde{W}_l, \tilde{b}_l\}$ in the l^{th} layer encode the convolution kernels of size 3×3 where \tilde{W}_l has stride 2. The number of features in each of the $L = 6$ layers follows a geometric progression with common ratio 1.75, starting at 48 features in the first layer. Finally, $FC: x \mapsto W_{FC}x$ is a fully connected layer mapping to a scalar and U is the absolute value. The entirety of the learnable parameters can be summarized as $\theta = \{W_l, b_l, \tilde{W}_l, \tilde{b}_l\}_{l=1}^L \cup \{W_{FC}\}$. We do not impose any constraints on any of the parameters, thus $\Theta \cong \mathbb{R}^{n_p}$, where $n_p = 21\,350\,640$ is the total number of parameters. Although our network is quite sizable, it has significantly less learnable parameters than, e.g., the discriminative end-to-end VN of [19] (3×10^7) or the score-based diffusion models of [12] (6.7×10^7).

We optimize Eq. 6 with AdaBelief [54] ($\beta_1 = 0.9, \beta_2 = 0.999$). In contrast to most previous works [27], [30], [31]

we did not find it necessary to regularize our model by means of, e.g., Lipschitz regularization, magnitude penalization or similar techniques. We use a learning rate of 5×10^{-4} , exponentially decreasing with rate 0.5 at update steps {500, 2000, 3000, 5000, 7000}, using a batch size of 50 for 27 000 parameter updates. To stabilize training, we smooth the data distribution by convolving it with a normal distribution of standard deviation 1.5×10^{-2} . For approximating \mathbb{E}_{p_θ} , we run ULA for $J_{\max} = 500$ steps. To accelerate training in the early stages, we use an exponential schedule, detailed by $J_h = \lceil J_{\max}(1 - \exp(-h/1000)) \rceil$, at the h^{th} parameter update. For persistent initialization, we use a replay buffer holding 8000 images with reinitialization chance $\pi_{\text{reinit}} = 1\%$. Training took approximately one month on a machine equipped with one NVIDIA Quadro RTX 8000.

E. Simulation Study and Posterior Sampling

To construct a real-valued simulation study, we use the RSS validation and test data detailed in Section II-C, and retrospectively sub-sample the corresponding k -space data. To approximately map the reconstructions to the same intensities seen during training, we normalize the data by $z \mapsto (z / \|u^{(0)}\|_\infty)$, and normalize the final reconstruction by $u^* \mapsto u^* \|u^{(0)}\|_\infty$. To run inference on this data, we utilize Algorithm 1 with one sensitivity map that is fixed at the identity ($C = 1, \sigma_1 = \mathbf{1}$ is fixed as the one-vector). We found the optimal regularization parameter λ by grid search on the validation dataset.

In addition to computing MAP estimators in the sense of Eq. 10, we also examine the posterior distribution. In particular, we approximate the MMSE (Eq. 3) and variance (Eq. 4) integrals by MCMC: We run ULA on the Gibbs density of H , where again $C = 1$ and $\sigma_1 = \mathbf{1}$ is fixed as the one-vector, i.e. $p(u | z) \propto \exp(-H(u, \mathbf{1}))$. The algorithm is initialized with uniform noise. We discard the first 10 000 samples to reach a steady state, and then save every 15th iteration, for a total of 160 000 iterations (resulting in 10 000 saved samples). We found that the regularization parameter λ barely influenced the results of the posterior sampling and consequently set it to $\lambda = 1$ for all sub-sampling patterns.

F. Parallel Imaging

For the parallel imaging experiments, we run Algorithm 1 with $K = 100$. As in the real-valued simulation study, we normalize the data by $z \mapsto (z / \|u^{(0)}\|_\infty)$, and normalize the final reconstruction by $u^* \mapsto u^* \|u^{(0)}\|_\infty$. To find the optimal regularization parameters, we fix $\mu = 10$ and obtain λ by linear least-squares regression of the initial residuum $\sum_{c=1}^C \left\| \mathcal{F}_M(\sigma_c^{(0)}(z^{\text{val},i}) \odot u^{(0)}(z^{\text{val},i})) - (z^{\text{val},i})_c \right\|_2^2$ against $\min_\lambda \|u^*(z^{\text{val},i}, \lambda) - u^{\text{val},i}\|_2^2$ (found by grid search) for all image-data pairs $(u^{\text{val},i}, z^{\text{val},i})$ in the validation set. Here, we view the initial image $u^{(0)}$, the sensitivity maps $(\sigma_c)_{c=1}^C$, and the optimal reconstruction u^* as maps to make dependencies explicit. This regression is performed *only* for 4-fold Cartesian sub-sampling with 8% ACL and other sub-sampling patterns use the same fit. Generalization experiments marked with † use the linear λ -fit calculated on CORPD data.

For experiments marked with *, we re-ran the regression on CORPD-FS data on a 4-fold Cartesian sub-sampling with 8% ACL.

A particular characteristic of our reconstruction approach is that its intensities are not quantitatively comparable to the reference. In detail, although we normalize the reconstruction by the RSS of the sensitivity maps, we observed that especially in low-intensity regions (e.g. air) the reconstruction did not match the reference. To remedy this and allow for fair quantitative evaluation, we utilize the validation data to fit a spline curve (cubic splines, 5 equally spaced knots) against the scatter of reconstructed and reference intensities. For the generalization experiments, we fit the spline curve again on an independent CORPD-FS validation dataset. The spline curves for both CORPD and CORPD-FS are shown in Fig. 8 in the appendix. The insets show that our reconstructions prefer zero-intensity in background regions, whereas the reference images have non-zero background intensity.

For parallel imaging, we define the MMSE as follows: Let $(u^*, (\sigma_c^*)_{c=1}^C)$ be a solution to Eq. 10. Then, we fix $(\sigma_c^*)_{c=1}^C$ and sample the conditional probability $p(u \mid z, (\sigma_c^*)_{c=1}^C)$. This amounts to performing Langevin sampling on the Gibbs density of $H(\cdot, (\sigma_c^*)_{c=1}^C): \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}$. The sensitivities may also be included in the Langevin procedure, but we empirically found no noticeable difference to freezing them. We believe that this is due to the strong imposed spatial regularity, which corresponds to a “narrow” prior.

We evaluate the quality of the estimated sensitivity maps by computing the null-space residual [32]: Let $u_c = \mathcal{F}^*(z_c)$, $c = 1, \dots, C$, denote the fully-sampled coil images. The null-space residual

$$\pi_c = \frac{\sigma_c}{|\Sigma|_C} \sum_{i=1}^C \bar{\sigma}_i u_i - u_c, \quad (17)$$

where $\bar{\cdot}$ denotes complex conjugation, should only contain noise since $u_i = \sigma_i u$ when σ_i is exact. Thus, any residual signal points to sub-optimal sensitivity estimates.

G. Comparison and Evaluation

We compare our approach to the following methods: For a hand-crafted prior, we chose the Charbonnier smoothed total variation

$$\text{TV} : \mathbb{R}_{\geq 0}^n \ni x \mapsto \sum_{i=1}^n \sqrt{(\text{D}x)_i^2 + (\text{D}x)_{i+n}^2 + \epsilon^2} \quad (18)$$

with $\epsilon = 10^{-3}$. In the real-valued simulation study, we compare against the fastMRI baseline method [8] as well as the diffusion-based approach of [12]. However, due to time and computational constraints, we limit our comparison to one arbitrarily picked image per sub-sampling pattern. The implementation as well as the trained model are taken from their github repository, thus the training database differs to ours as it includes the CORPD-FS data. We use 2000 steps in the reverse diffusion. As a state-of-the-art discriminative approach for parallel MRI, we compare against the end-to-end VN approach from [19]. The implementation was taken from the fastMRI github repository with default parameters. The

fastMRI baseline method as well as the VN were trained on the subset of the fastMRI dataset detailed Section II-C, with masks generated using random 4-fold Cartesian sub-sampling with 8% ACL.

We compare the reconstructions quantitatively using the peak signal-to-noise ratio (PSNR), normalized mean-squared error (NMSE) and structural similarity (SSIM) [55]. SSIM uses a 7×7 uniform filter and parameters $K_1 = 0.01$, $K_2 = 0.03$. We define the acceleration factor (denoted Acc. in tables showing quantitative results) as the ratio of the image size and the acquired k -space points on the Cartesian grid.

Our training and evaluation code, along with pre-trained models is available at <https://github.com/VLOGroup/stable-deep-mri>.

III. RESULTS

A. Data-Independent Analysis

The local regularization landscape is a proxy for many interesting properties, relating to (local) convexity, generalization capabilities and adversarial robustness [56]. To effectively visualize the $\tilde{n} = (320 \times 320)$ -dimensional landscape, we follow [57]: Let $(x^{(k)})_{k=1}^K, x^{(k)} \in \mathbb{R}^{\tilde{n}}$ be samples drawn from the Langevin process Eq. 7, starting from uniform noise. Denote with (v_1, v_2) the first two principal components of the matrix $[x^{(0)} - x^{(K)}; x^{(1)} - x^{(K)}; \dots; x^{(K-1)} - x^{(K)}]$ and denote $\bar{x} := x^{(K)} + 1/(K-1) \sum_{k=1}^{K-1} (x^{(k)} - x^{(K)})$. For the experiments, we set $K = 10000$.

In Fig. 2 we show $\mathbb{R}^2 \ni (\xi_1, \xi_2) \mapsto \exp(-R(\bar{x} + \xi_1 v_1 + \xi_2 v_2)/\tilde{T})$, where $\tilde{T} = 7$ was chosen to yield visually pleasing results, along with the Langevin trajectory. The figure shows two interesting properties: First, the samples from the Langevin process are almost indistinguishable from the reference data, which demonstrates that our prior is extremely strong and faithfully represents the distribution of the training data. In particular, this implies that the learned regularizer — by construction — should only be applied to reconstruction tasks where the underlying data is also drawn from the same distribution. The experiments in Section III-E show that our regularizer can reasonably reconstruct knee images of different contrast by adapting regularization strength, but prior work [58] demonstrates that performance quickly degrades when confronted with, e.g., rotated images. Second, in two dimensions, the landscape appears smooth on the considered domain and almost log-concave around modes of the induced distribution. Although not applicable directly, these findings should be taken as evidence that the high-dimensional landscape is also reasonably well-behaved. This is corroborated empirically by the ease of optimization: For all reconstruction tasks, we only need in the order of 10 iterations of iPALM.

B. Simulation Study

We show qualitative results of the retrospectively sub-sampled real-valued data in Fig. 3 (left), where the results are shown in increasing acceleration. In the second row we show the results from 4-fold Cartesian sub-sampling with 8% ACL. Note that this is the setting on which the U-Net was trained, and indeed it yields satisfactory reconstructions. The TV reconstruction for this task is not able to fully remove

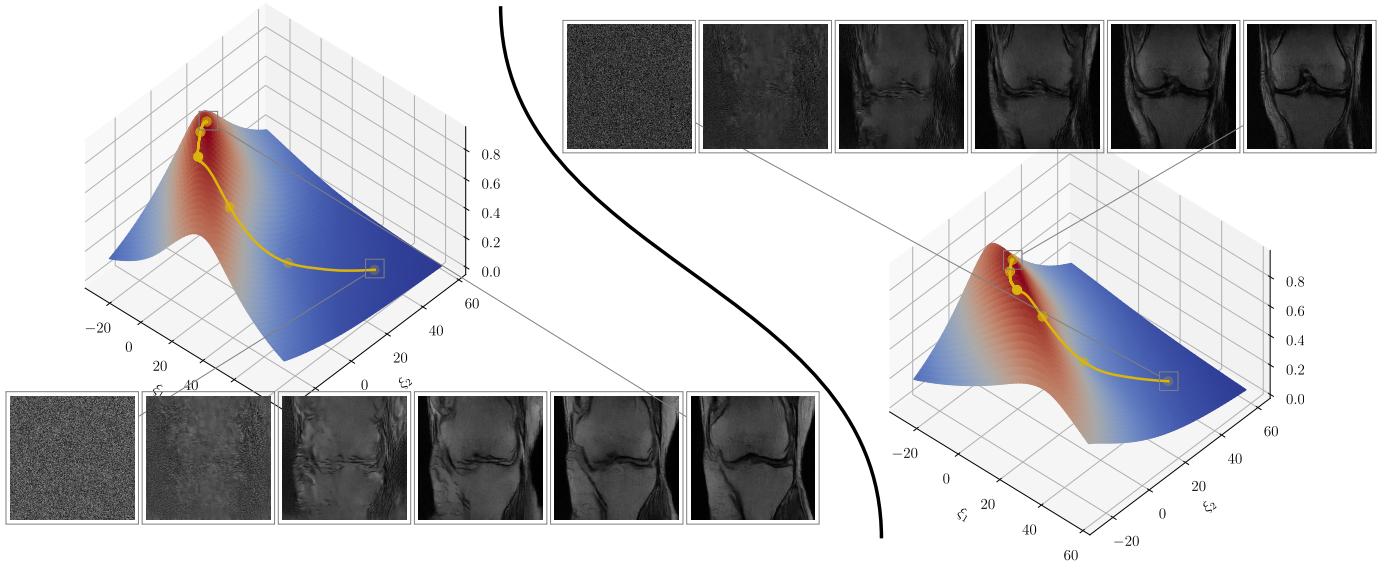


Fig. 2. Surface plots of the local regularization landscape $\mathbb{R}^2 \ni (\xi_1, \xi_2) \mapsto \mathbf{exp}(-R(\bar{x} + \xi_1 v_1 + \xi_2 v_2)/\tilde{T})$. The Langevin trajectory (Eq. 7) is shown in gold along with some representative states.

the sub-sampling artifacts, but increasing regularization would lead to significant loss of detail. Our approach is at least on par qualitatively with the discriminative U-Net, and the quantitative analysis in Table I shows superiority over all reference methods.

The third row details the reconstructions for spiral sub-sampling. Here, the acceleration factor is approximately 5, with a more densely sampled k -space center, which is typically advantageous for traditional reconstruction techniques. The TV reconstruction removes most of the sub-sampling artifacts, although some are still visible, especially in the background. The discriminative U-Net approach in this task (and indeed all tasks apart from Cartesian sub-sampling, on which it was trained) struggles to discriminate between details in the anatomy and sub-sampling artifacts. Thus, it hallucinates details into the reconstruction that are not reflected in the data. In contrast, our approach is able to faithfully reconstruct the knee with no visible artifacts. The results are similar for the pseudo-radial sub-sampling pattern with 45 spokes shown in the fourth row.

The 3-fold random sub-sampling shown in the first row is particularly interesting. Here, the k -space center is not more densely sampled than any other region, which manifests in the zero-filling reconstruction by large-scale intensity shifts. None of the reference methods are able to correct this, since they do not have knowledge of the anatomy of the human knee. On the other hand, our approach, due to its generative nature, can restore the general shape of the knee well, and due to the variational approach also faithfully keeps details that are present in the data. The generative approach alleviates the requirement for sampling schemes to densely sample the k -space center, which is in stark contrast to the general theory.

We show a quantitative comparison against the diffusion-based approach of [12] in Table II that includes reconstruction time and number of trainable parameters. The accompanying qualitative results are shown in Fig. 9 in the

TABLE I

QUANTITATIVE RESULTS FOR THE REAL-VALUED SIMULATION STUDY USING DIFFERENT k -SPACE TRAJECTORIES. NUMBERS IN PARENTHESIS INDICATE THE ACCELERATION FACTOR, BOLD TYPEFACE INDICATES THE BEST METHOD

		ZF	TV	U-Net	Ours	
					MAP	MMSE
Random (3)	¶	12.93	20.81	19.52	30.87	32.44
	‡	64.24	10.22	12.83	1.34	1.02
	§	0.48	0.73	0.57	0.86	0.90
Cartesian (4)	¶	24.16	31.47	34.16	35.53	36.17
	‡	5.96	0.88	0.44	0.32	0.28
	§	0.70	0.85	0.89	0.89	0.91
Spiral (≈ 5)	¶	21.21	31.25	27.76	35.35	36.21
	‡	10.24	0.92	1.90	0.34	0.28
	§	0.62	0.85	0.78	0.88	0.90
Radial (≈ 6)	¶	27.02	32.86	31.76	35.04	35.47
	‡	3.29	0.64	0.76	0.36	0.33
	§	0.48	0.71	0.65	0.86	0.89

¶: PSNR (dB) \uparrow , ‡: NMSE ($\times 10^2$) \downarrow , §: SSIM \uparrow

appendix. The results are separated since we only evaluated on one image per sub-sampling pattern due to time and computational constraints, and thus it does not constitute a comprehensive comparison. Our MMSE estimate consistently beats the diffusion-based approach, and our MAP estimate is only inferior for the random sub-sampling pattern. To generate the 10 000 samples for the MMSE estimate takes only about 30% more time compared to the one sample generated by the reverse SDE. On the other hand, computing our MAP estimate is about 80 times faster.

C. Uncertainty Quantification Through Posterior Sampling

The natural probabilistic interpretation of our variational approach gives rise to a distribution of reconstructions for any particular reconstruction problem. We exploit this distribution

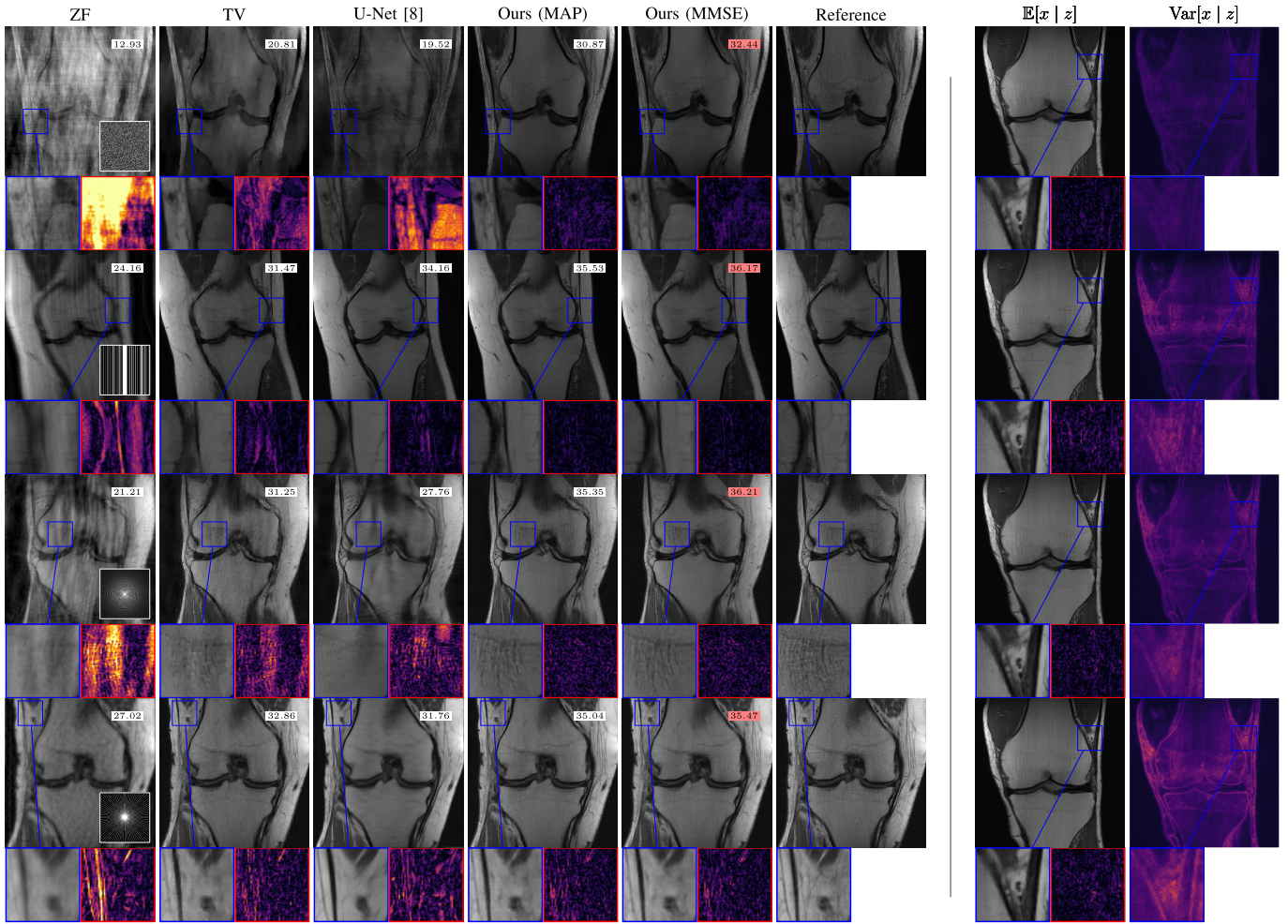


Fig. 3. Real-valued simulation study results: 1th row: Random sub-sampling with acceleration factor 3, 2th row: 4-fold Cartesian sub-sampling with 8% ACL, 3th row: Spiral sub-sampling (5-fold acceleration), 4th row: Radial sub-sampling using 45 spokes. The inlays show the sub-sampling pattern (white), a detail zoom (blue), and the magnitude of the difference to the reference (red, 0 \rightarrow 0.2). Left: Reconstruction results. The numbers show PSNR (over the entire test dataset), with the best method emphasized in red. Right: Uncertainty quantification through marginal posterior variance (0 \rightarrow 0.0025).

TABLE II

QUANTITATIVE COMPARISON (PSNR) AGAINST THE DIFFUSION-BASED METHOD OF [12]. BOLD TYPEFACE INDICATES THE BEST METHOD

		Random	Cartesian	Spiral	Radial	Time (s)	Parameters
Ours	MAP	31.78	34.97	36.07	36.02	3.13	2.1×10^7
	MMSE	34.76	35.55	37.00	36.56	333.3	
[12]		33.31	34.65	35.66	34.98	251.6	6.8×10^7

to compute pixel-wise variance maps as well as the MMSE estimate. The results in Fig. 3 (right) show large variance around small structures, and more variance when less data is available: The three-fold random sub-sampling shows the least variance, while the approximately six-fold radial sub-sampling shows the most. The MMSE estimator also yields visually pleasing reconstructions. In fact, Table I and Table II establish quantitative superiority over the MAP estimate (see Section IV for a possible explanation).

D. Parallel Imaging

The first row of Fig. 4 shows a classical Cartesian sub-sampling pattern, with an acceleration factor of 4 and using

8% ACL. This coincides with the training setup of the discriminative end-to-end variational network approach of [19]. Consequently, it shows the best performance quantitatively as well as qualitatively. Our method also yields competitive results, achieving a PSNR of 35.23dB. This is in line with our expectations, as in general we can not expect a generative approach to beat a discriminative counterpart. However, the strength of our approach becomes apparent when we slightly change the sub-sampling pattern: The performance of the end-to-end VN deteriorates significantly when the phase-encoding direction is swapped, or when less ACL are acquired. We emphasize that between these tasks the number of acquired data points is fixed. With these minimal changes in the sub-sampling pattern, the method of [19] is no longer able to fully remove back-folding artifacts or introduces severe hallucinations, while our method yields comparable results for all three tasks. This is also reflected in the quantitative evaluation in Table III, where our method beats the VN approach by a significant margin.

Shifting towards radial and 2D Gaussian sub-sampling patterns, we observe that the VN approach introduces severe artifacts. We believe that this is a combination of the

TABLE III

QUANTITATIVE RESULTS FOR PARALLEL IMAGING WITH DIFFERENT SUB-SAMPLING PATTERNS ON IN- AND OUT-OF-DISTRIBUTION DATA. THE † COLUMN SHOWS RESULTS USING THE CORPD λ -FIT, WHILE THE * COLUMN HAS CORPD-FS-ADAPTED PARAMETERS (SEE SECTION II-F)

Acc.	ACL		In-distribution (CORPD)				Out-of-distribution (CORPD-FS)					
			ZF	TV	VN	Ours		ZF	TV	VN	Ours	
						MAP	MMSE				†	*
Cartesian	8 %	¶	27.19	31.87	36.92	35.23	35.28	26.09	31.30	30.00	30.60	31.71
		‡	2.24	0.79	0.24	0.36	0.36	5.35	1.48	2.38	1.95	1.35
		§	0.74	0.81	0.92	0.89	0.89	0.68	0.73	0.77	0.73	0.73
	8 % (horiz.)	¶	31.13	33.03	24.72	36.23	36.01	26.58	31.56	28.57	30.89	31.65
		‡	0.93	0.59	4.01	0.28	0.30	5.15	1.40	2.90	2.24	1.37
		§	0.81	0.83	0.67	0.90	0.90	0.71	0.73	0.71	0.75	0.73
	4 %	¶	24.14	25.81	32.16	35.33	35.22	24.98	29.91	29.12	29.93	31.26
		‡	4.51	3.48	0.70	0.35	0.36	6.65	2.09	2.79	2.92	1.53
		§	0.69	0.70	0.89	0.89	0.89	0.65	0.71	0.76	0.73	0.72
Radial	≈ 11	¶	28.76	32.47	20.56	34.46	34.16	25.06	31.15	26.26	31.43	31.36
		‡	1.57	0.67	10.13	0.42	0.45	7.37	1.53	4.97	1.45	1.57
		§	0.75	0.81	0.69	0.86	0.86	0.62	0.71	0.69	0.73	0.70
2D Gaussian	8	¶	32.10	34.14	20.74	35.35	35.41	26.75	31.52	23.46	31.87	32.09
		‡	0.74	0.45	9.95	0.34	0.34	5.46	1.42	9.93	1.43	1.25
		§	0.84	0.85	0.68	0.88	0.89	0.71	0.75	0.68	0.76	0.75

¶: PSNR (dB) †, ‡: NMSE ($\times 10^2$) ‡, §: SSIM †

sensitivity estimation sub-network failing as well as the image sub-network being confronted with previously unseen sub-sampling artifacts. Indeed, for this task the more general TV approach is superior to the VN, although reconstructions appear over-smoothed. In line with expectations, our approach reconstructs the image satisfactorily, yielding the best performance quantitatively as well as qualitatively.

To analyze our reconstruction algorithm also with respect to the estimated sensitivity maps, we show example estimations along with ESPIRiT [32] in Fig. 6. The figure shows estimations for the 4-fold Cartesian sub-sampling with 8% ACL reconstruction problem shown in the first row of Fig. 4. Due to the re-parametrization Eq. 9, our reconstruction algorithm does not require pixel-wise normalized sensitivity maps. Hence, they look very physically plausible and match the reference well.

We additionally analyze the estimation qualitatively by visualizing the RSS null-space residual [32] $|(\pi_c)_{c=1}^C|_c$. In Fig. 7, we compare the sensitivity maps from our joint estimation with ESPIRiT for Cartesian sub-sampling patterns with 8% ACL and 4% ACL (first and third row in Fig. 4 respectively). For both tasks, the sensitivity maps from our joint estimation algorithm can reproduce the data very well. In particular, while the ESPIRiT estimation leads to slightly better results when 8% ACL are available, the estimation deteriorates for 4%. In contrast, our estimation remains stable, as it can exploit data that is not in the k -space center.

E. Generalization

By construction, the learned model encodes the distribution of the training data. In the previous sections we have demonstrated thoroughly that the prior is agnostic to shifts in the sub-sampling pattern, which is an expected consequence of the generative approach. However, a natural question is whether the approach is also robust with respect to shifts in the

underlying distribution. To study this, we apply the regularizer to an underlying distribution of fat-suppressed images.

The quantitative and qualitative results in Table III and Fig. 5 indicate that our method generalizes better to unseen data than the end-to-end VN approach of [19], although the performance degrades significantly in both cases. To highlight the advantage of the tuneable regularization parameters in our approach, we show results using the λ -fit (see Section II-F) calculated on non-fat-suppressed data, as well as using parameters adapted to the task: The ability to tune the influence of the regularizer leads to improved performance when confronted with previously unseen data.

IV. DISCUSSION

Fig. 2 demonstrates that our learned regularizer encodes the training distribution. As a consequence, we can reconstruct images from severely ill-posed problems (such as random sub-sampling pattern) satisfactorily. However, this also means that the performance of our learned regularizer is expected to significantly decline when applied to, e.g., different anatomy (rotation is explored in [58]). Diffusion models [12] are another instance of models that are capable of generating realistic-looking images from the data-distribution, but for inverse problems it is not clear if this is needed. Since our network is purely convolutional, it can easily be “localized” by removing the deeper layers, effectively recovering the fields-of-experts model [29]. This would remove the majority of the trainable parameters (and consequently the training cost), and we hypothesize that it would lead to better generalization as local features are largely shared between different anatomical structures.

The results in Section III-D demonstrate that the regularizer can reconstruct images from multi-coil data satisfactorily, irrespective of the acquisition mask. This is a very significant advantage over the reference methods, since in practice a

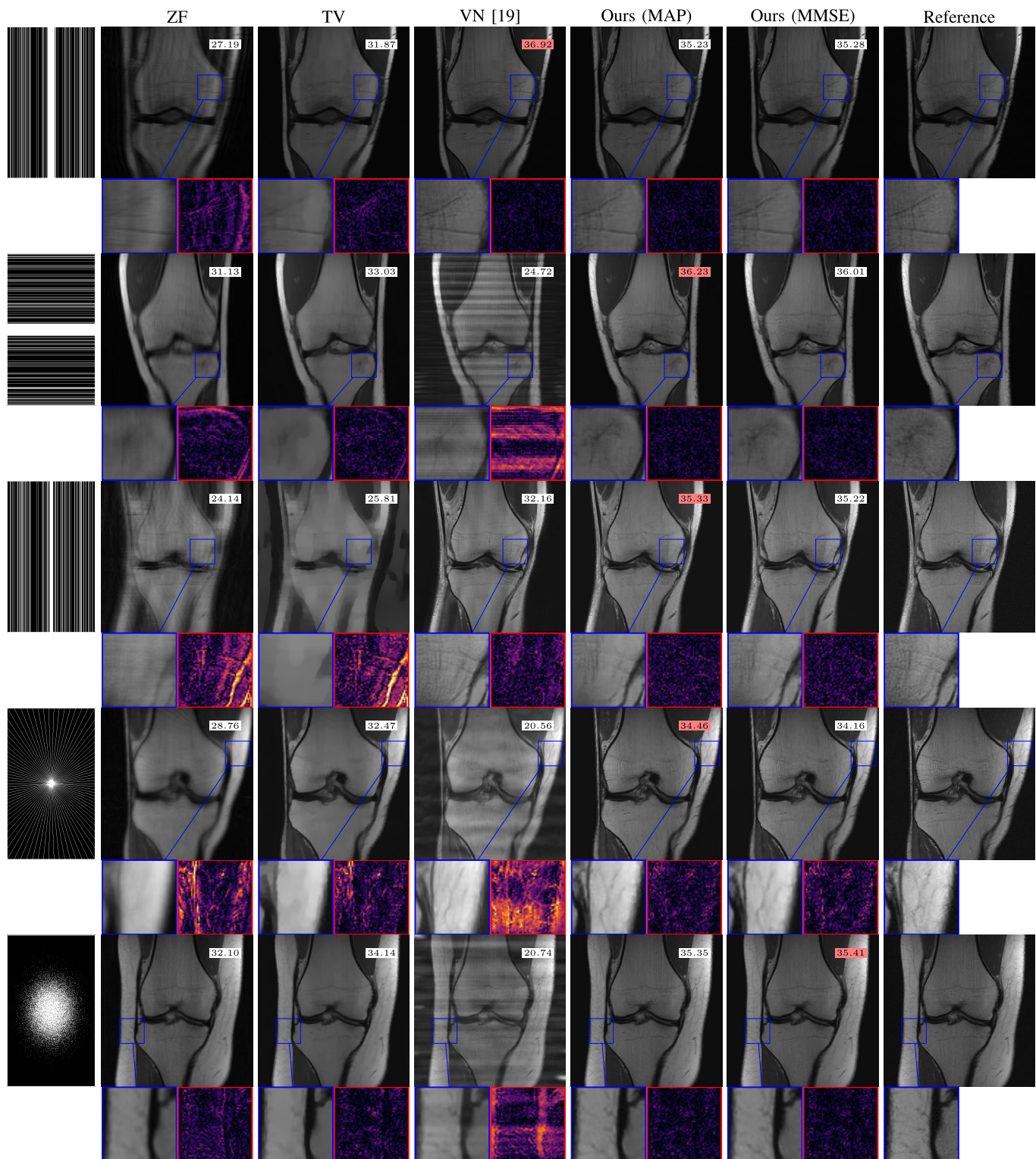


Fig. 4. Parallel imaging on in-distribution data: 1st row: 4-fold Cartesian sub-sampling with 8% ACL, 2nd row: as previous with swapped phase encoding direction, 3rd row: 4-fold Cartesian sub-sampling with 4% ACL, 4th row: Pseudo-radial sampling with 45 equidistant spokes (\approx 11-fold sub-sampling), 5th row: 2D Gaussian sampling (8-fold sub-sampling). The inlays show the sub-sampling mask, a detail zoom (blue), and the magnitude of the difference to the reference (red, 0 \rightarrow 0.2).

particular situation might necessitate a different sub-sampling pattern. As an example, phase encoding direction is often swapped when a blood vessel is located in a particularly disadvantageous position. In addition, our regularizer does

not need re-training when a new sub-sampling pattern is discovered to be advantageous.

The results on multi-coil data also empirically demonstrate that a real-valued regularizer suffices to reconstruct the under-

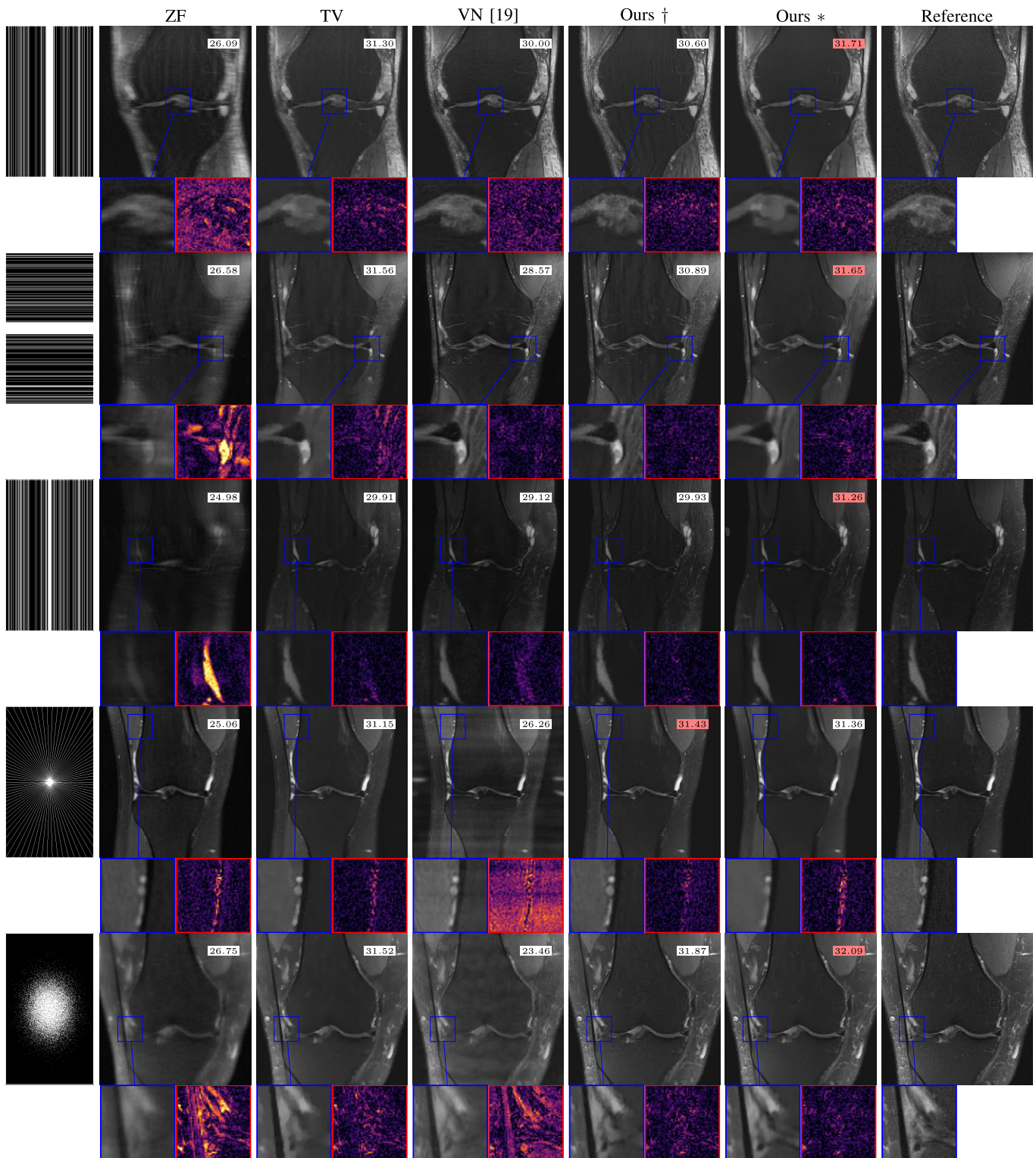


Fig. 5. Parallel imaging on out-of-distribution data (†: regularization parameters from non-fat-suppressed data, *: adapted parameters): 1st row: 4-fold Cartesian sub-sampling with 8% ACL, 2nd row: as previous with swapped phase encoding direction, 3rd row: 4-fold Cartesian sub-sampling with 4% ACL, 4th row: Pseudo-radial sampling with 45 equidistant spokes (\approx 11-fold sub-sampling). 5th row: 2D Gaussian sampling (8-fold sub-sampling). The inlays show the sub-sampling mask, a detail zoom (blue), and the magnitude of the difference to the reference (red, 0 \rightarrow 0.2).

lying spin-density, at least for the acquisition sequences used in our dataset [8]. Similarly to [22] and [23], our approach could easily be extended to account for complex images if needed. In particular, this is the case for contrasts that inherently

rely on complex-valued information, such as phase-contrast MRI.

Generalization is not limited to the acquisition mask: The regularizer also performs well in out-of-distribution

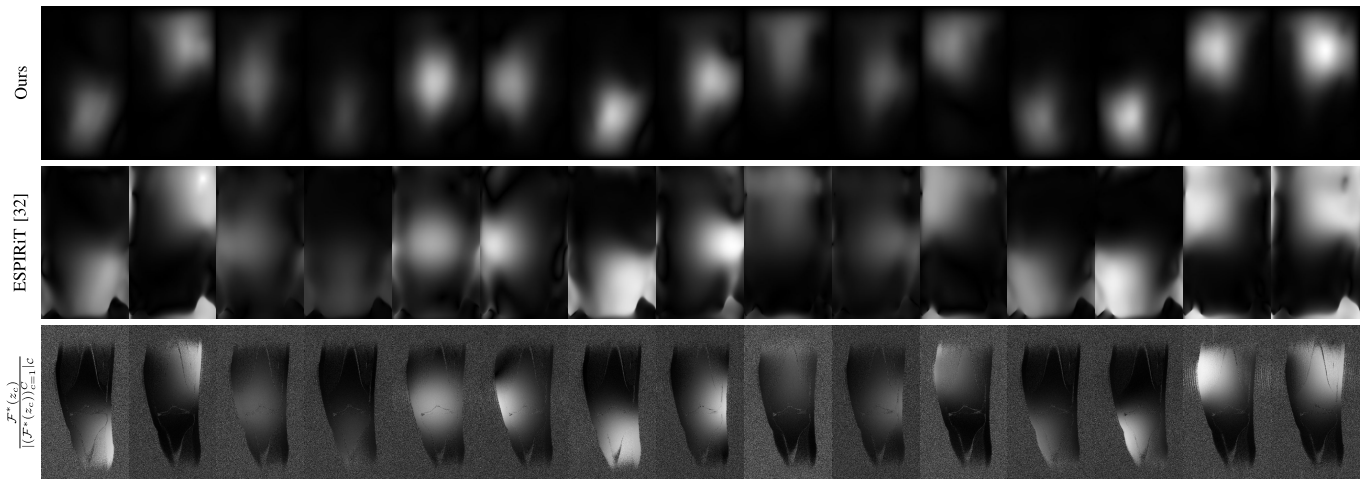


Fig. 6. Magnitude of the estimated sensitivities using our joint estimation algorithm (top) versus the ESPIRiT [32] estimation (middle) for the reconstruction problem shown in Fig. 4 (1st row). The bottom row shows the reference coil sensitivities computed with the fully-sampled data.

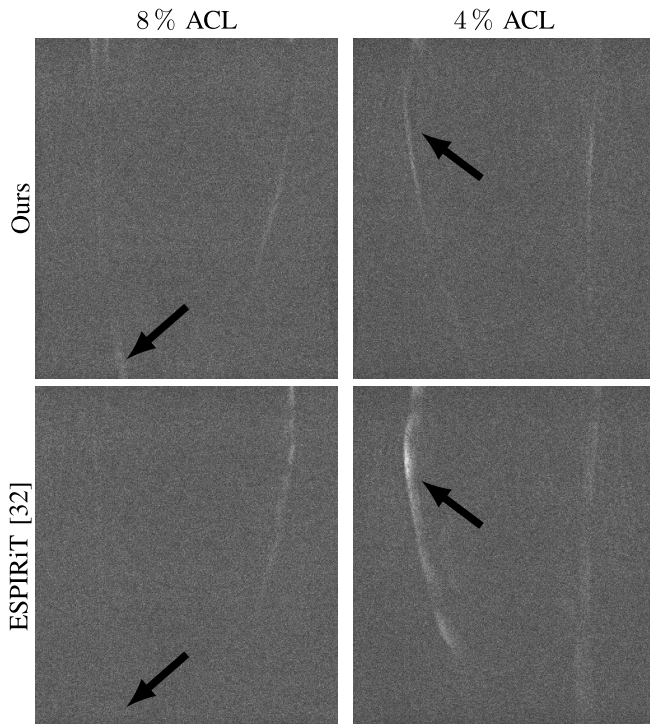


Fig. 7. RSS null-space residual $|\langle \pi_c \rangle_{c=1}^C|_{\mathcal{C}}$ using our joint estimation algorithm (top) and ESPIRiT [34] (bottom). We consider the two Cartesian sub-sampling reconstruction problems in the first and third row of Fig. 4: 8% ACL (left) and 4% ACL (right).

experiments, which highlight the importance of the ability to control its influence. Table III shows that adapting the strength of the regularization to the underlying data strongly improves performance. We note that the regularization strength was always tuned to a 4-fold Cartesian sub-sampling task with 8% ACL (see Section II-F). The results for radial sub-sampling indicate that this is a sub-optimal fit for the out-of-distribution data: The fit calculated on the CORPD data († column) performs better than the fit calculated on CORPD-FS data. Fig. 5 suggests that the adapted parameters lead to an over-smoothed reconstruction. Clearly, the performance

could be improved by adapting the regularization strength to the CORPD-FS data *and* the radial sub-sampling. In general, adapting the regularization strength to acquisition mask as well as the data is greatly beneficial. However, this is typically not possible with other data-driven approaches.

The accompanying probabilistic interpretation is significant in two ways: First, the regularizer can be inspected by data-independent analysis as in Fig. 2, where experts can easily visualize preferred structures. This aids in interpreting the encoded information and thus improves the confidence with which clinicians may view the reconstruction. Second, the variance maps provide the clinician with additional information about the inherent uncertainty in the reconstruction. Fig. 3 (right) clearly shows high variance around small anatomic structures. If the clinician’s decision making is based on such regions, they might deem it necessary to acquire additional information.

In the single-coil simulation study, quantitative and qualitative analysis shows superiority of the MMSE estimate over the MAP estimate. We believe that this is related to the training procedure: During training, the regularizer is only ever confronted with slightly noisy images in the Langevin process (see Section II-A), and injecting the same noise in the reconstruction algorithm improves overall performance. We are unsure why this performance improvement does not translate to the parallel imaging experiments, but hypothesize that the our joint estimation biases the reconstruction such that this effect is no longer observable.

Finally, we emphasize that our algorithm for reconstructing parallel MRI is fast: Algorithm 1 converges in around 100 iterations, which takes around 5s on an NVIDIA Titan RTX using approximately 2GB of memory. This is in stark contrast to score-based diffusion models, where the reverse diffusion process typically has to be solved with high accuracy and consequently reconstruction time is in the order of 10min [12]. Speed is advantageous in practice, as images can be viewed while the patient is still in the scanner. This allows fast changes to the sub-sampling pattern, should they be needed.

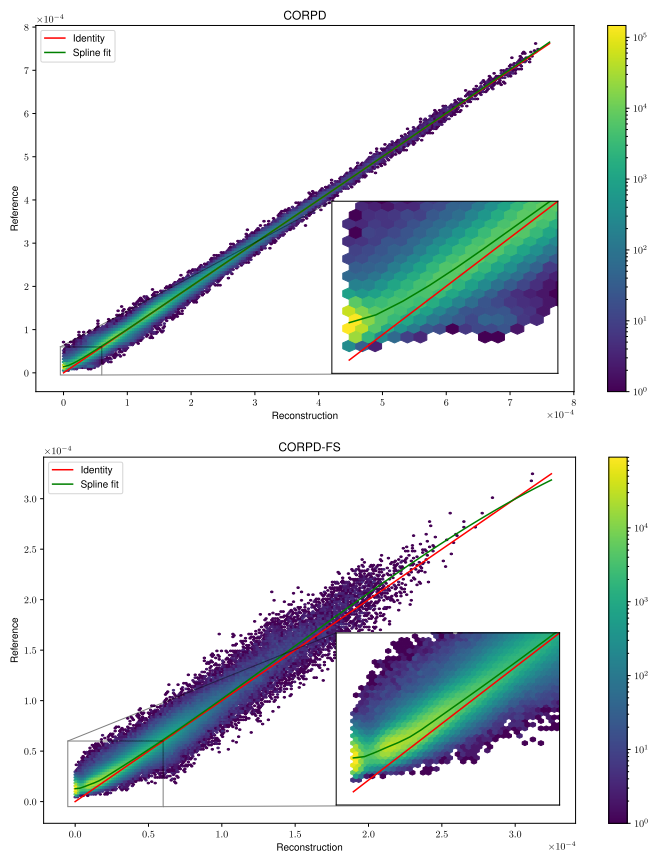


Fig. 8. Spline fit computed on a validation set for CORPD (top) and CORPD-FS data (bottom). The figures show the log-histogram of reconstructed versus reference intensities with the insets show the region around zero, where they deviate the strongest.

V. CONCLUSION

We utilize modern generative learning techniques to train a regularizer such that it encodes the underlying distribution of the training data faithfully. By embedding the regularizer in a variational reconstruction framework, we can satisfactorily reconstruct single-coil and parallel MRI by adapting the data likelihood term. Quantitative and qualitative analysis indicate competitive reconstruction performance, on-par with (or sometimes superior to) fully supervised methods. In addition, our approach is agnostic to changes in the acquisition mask, while the supervised reference methods introduce severe hallucinations when confronted with previously unseen data. The knowledge of the human anatomy encoded in the regularizer even allows us to reconstruct high quality images from random sampling masks. Finally, on out-of-distribution data, our method is still on-par or better than hand-crafted regularizers such as the TV or other supervised methods. This indicates that our method also generalizes better with respect to shifts in the underlying distribution.

In addition to a competitive reconstruction performance, our method has a natural accompanying probabilistic interpretation through statistical modeling and Bayesian inference. This allows experts to explore a distribution of reconstructions, while other data-driven approaches only act as point estimators. Our experiments indicate that this distribution encodes

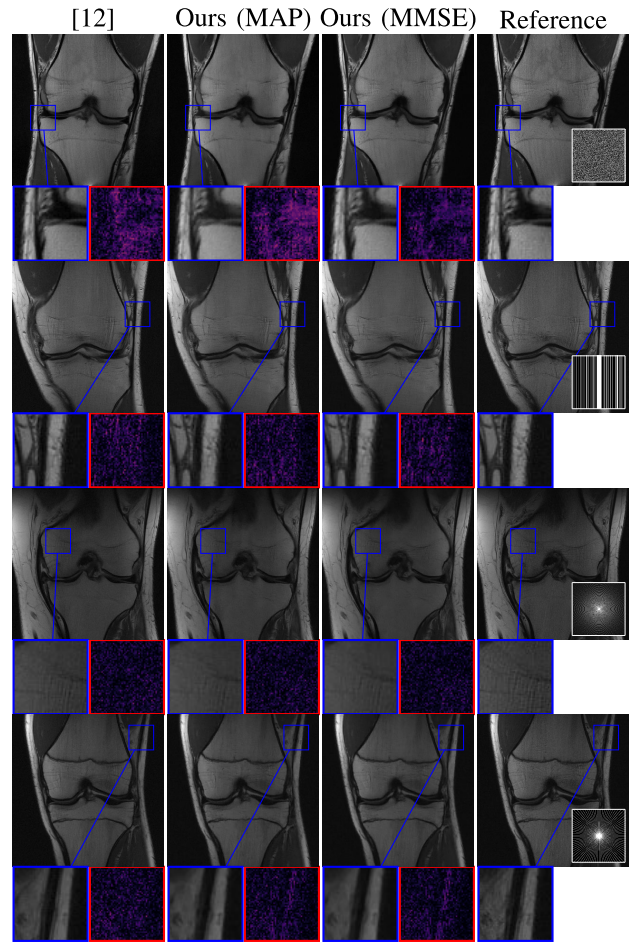


Fig. 9. Qualitative comparison against [12] on in-distribution data: 1st row: Random sub-sampling with acceleration factor 3, 2nd row: 4-fold Cartesian sub-sampling with 8% ACL, 3rd row: Spiral sub-sampling (5-fold acceleration), 4th row: Radial sub-sampling using 45 spokes. The inlays show the sub-sampling pattern (white), a detail zoom (blue), and the magnitude of the difference to the reference (red, 0 to 0.2).

important diagnostic information, such as high uncertainty around small anatomical structures. We believe that this additional information can aid the practitioner’s decision making. Further, we can perform a data-independent analysis of the model by visualizing the induced distribution. This allows to interpret the information encoded in the regularizer, whereas other data-driven methods solely act as black-boxes in this respect.

To reconstruct parallel MRI, we propose a fast algorithm to jointly estimate the image and sensitivity maps. In contrast to off-line sensitivity estimation, our approach does not require auto-calibration lines, can easily be applied to non-Cartesian sampling trajectories, and uses all available data. The resulting optimization problem can be solved efficiently using non-convex optimization algorithms. High-fidelity reconstructions are available in approximately 5s on consumer hardware.

In general, we believe that reconstruction approaches based on generative priors have huge potential. A natural extension of our work would be to combine the image-prior with a learned sensitivity-prior. In addition, future research could focus on extending the simple architecture used in this paper

with more modern building blocks, such as attention layers. A related direction is to research if local convolutional models can replicate the performance of our generative prior.

APPENDIX

A. Spline Fit

Fig. 8 shows the scatter plot of reconstructed versus reference intensities along with the spline fit discussed in Section III-D.

B. Additional Results

A qualitative comparison against the diffusion-based method of [12] is presented in Fig. 9. The accompanying quantitative evaluation is shown in Table II in the main body.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their time and effort spent on reviewing the manuscript. Their comments and suggestions helped them to greatly improve the quality of the manuscript.

REFERENCES

- [1] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, no. 5, pp. 952–962, Nov. 1999.
- [2] P. M. Robson, A. K. Grant, A. J. Madhuranthakam, R. Lattanzi, D. K. Sodickson, and C. A. McKenzie, "Comprehensive quantification of signal-to-noise ratio and g-factor for image-based and k-space-based parallel imaging reconstructions," *Magn. Reson. Med.*, vol. 60, no. 4, pp. 895–907, Oct. 2008.
- [3] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [4] F. Knoll, C. Clason, K. Bredies, M. Uecker, and R. Stollberger, "Parallel imaging with nonlinear reconstruction using variational penalties," *Magn. Reson. Med.*, vol. 67, no. 1, pp. 34–41, Jun. 2011.
- [5] J. Huang and D. Mumford, "Statistics of natural images and models," in *Proc. CVPR*, vol. 1, 1999, pp. 541–547.
- [6] K. Hammernik et al., "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, Nov. 2017.
- [7] M. Akçakaya, S. Moeller, S. Weingärtner, and K. Uğurbil, "Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction: Database-free deep learning for fast imaging," *Magn. Reson. Med.*, vol. 81, no. 1, pp. 439–453, Sep. 2018.
- [8] J. Zbontar et al., "fastMRI: An open dataset and benchmarks for accelerated MRI," 2018, *arXiv:1811.08839*.
- [9] P. Putzky and M. Welling, "Invert to learn to invert," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–11.
- [10] B. Zhou and S. K. Zhou, "DuDoRNet: Learning a dual-domain recurrent network for fast MRI reconstruction with deep T1 prior," in *Proc. CVPR*, 2020, pp. 4273–4282.
- [11] D. Narnhofer, K. Hammernik, F. Knoll, and T. Pock, "Inverse GANs for accelerated MRI reconstruction," *Proc. SPIE*, vol. 11138, pp. 381–392, Sep. 2019.
- [12] H. Chung and J. C. Ye, "Score-based diffusion models for accelerated MRI," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102479.
- [13] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, Jun. 2017.
- [14] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock, "Variational networks: Connecting variational methods and deep learning," in *Pattern Recognition (Lecture Notes in Computer Science)*. Berlin, Germany: Springer, 2017, pp. 281–293.
- [15] J. Cheng, H. Wang, L. Ying, and D. Liang, "Model learning: Primal dual networks for fast MR imaging," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019, pp. 21–29.
- [16] V. Vishnevskiy, J. Walheim, and S. Kozerke, "Deep variational network for rapid 4D flow MRI reconstruction," *Nature Mach. Intell.*, vol. 2, no. 4, pp. 228–235, Apr. 2020.
- [17] 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, Mar. 2018.
- [18] T. Pock and S. Sabach, "Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems," *SIAM J. Imag. Sci.*, vol. 9, no. 4, pp. 1756–1787, Jan. 2016.
- [19] A. Sriram et al., "End-to-end variational networks for accelerated MRI reconstruction," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2020.
- [20] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. ICML*, vol. 70, 2017, pp. 537–546.
- [21] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.
- [22] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–18.
- [23] G. Luo, M. Blumenthal, M. Heide, and M. Uecker, "Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models," *Magn. Reson. Med.*, vol. 90, no. 1, pp. 295–311, Mar. 2023.
- [24] B. T. Feng, J. Smith, M. Rubinstein, H. Chang, K. L. Bouman, and W. T. Freeman, "Score-based diffusion models as principled priors for inverse imaging," 2023, *arXiv:2304.11751*.
- [25] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of AI," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30088–30095, May 2020.
- [26] E. Kobler, A. Effland, K. Kunisch, and T. Pock, "Total deep variation for linear inverse problems," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7546–7555.
- [27] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–11.
- [28] E. Nijkamp, M. Hill, T. Han, S.-C. Zhu, and Y. N. Wu, "On the anatomy of MCMC-based maximum likelihood learning of energy-based models," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 4, pp. 5272–5280.
- [29] S. Roth and M. J. Black, "Fields of experts: A framework for learning image priors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Oct. 2005, pp. 860–867.
- [30] Y. Guan, Z. Tu, S. Wang, Y. Wang, Q. Liu, and D. Liang, "Magnetic resonance imaging reconstruction using a deep energy-based model," *NMR Biomed.*, vol. 36, no. 3, p. e4848, Nov. 2022.
- [31] Z. Tu, C. Jiang, Y. Guan, J. Liu, and Q. Liu, "K-space and image domain collaborative energy-based model for parallel MRI reconstruction," *Magn. Reson. Imag.*, vol. 99, pp. 110–122, Jun. 2023.
- [32] M. Uecker et al., "ESPIRiT—An eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA," *Magn. Reson. Med.*, vol. 71, no. 3, pp. 990–1001, May 2013.
- [33] C. A. McKenzie, E. N. Yeh, M. A. Ohliger, M. D. Price, and D. K. Sodickson, "Self-calibrating parallel imaging with automatic coil sensitivity extraction," *Magn. Reson. Med.*, vol. 47, no. 3, pp. 529–538, Feb. 2002.
- [34] M. Uecker, T. Hohage, K. T. Block, and J. Frahm, "Image reconstruction by regularized nonlinear inversion-joint estimation of coil sensitivities and image content," *Magn. Reson. Med.*, vol. 60, no. 3, pp. 674–682, Sep. 2008.
- [35] L. Ying and J. Sheng, "Joint image reconstruction and sensitivity estimation in SENSE (JSENSE)," *Magn. Reson. Med.*, vol. 57, no. 6, pp. 1196–1202, 2007.
- [36] G. Zeng et al., "A review on deep learning MRI reconstruction without fully sampled k-space," *BMC Med. Imag.*, vol. 21, no. 1, p. 195, Dec. 2021.
- [37] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021, pp. 1–36.
- [38] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, "Robust compressed sensing MRI with deep generative priors," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 14938–14954.
- [39] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *Proc. ICLR*, 2023, pp. 1–30.

- [40] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.
- [41] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM J. Imag. Sci.*, vol. 3, no. 3, pp. 492–526, 2010.
- [42] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.
- [43] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, pp. 341–363, Dec. 1996.
- [44] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin diffusions," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 60, no. 1, pp. 255–268, Jan. 1998.
- [45] A. Durmus, É. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau," *SIAM J. Imag. Sci.*, vol. 11, no. 1, pp. 473–506, Jan. 2018.
- [46] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal, "Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation," *Statist. Comput.*, vol. 31, no. 3, pp. 1–18, May 2021.
- [47] Y. Du, S. Li, J. B. Tenenbaum, and I. Mordatch, "Improved contrastive divergence training of energy based models," in *Proc. ICML*, 2021, pp. 1–16.
- [48] S. C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," *Int. J. Comput. Vis.*, vol. 27, no. 2, pp. 107–126, 1998.
- [49] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1064–1071.
- [50] A. Chambolle and T. Pock, "An introduction to continuous optimization for imaging," *Acta Numerica*, vol. 25, pp. 161–319, May 2016.
- [51] K. Kreutz-Delgado, "The complex gradient operator and the CR-calculus," 2009, *arXiv:0906.4835*.
- [52] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.
- [53] F. Knoll et al., "FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190007.
- [54] J. Zhuang et al., "Adabelief optimizer: Adapting stepsizes by the belief in observed gradients," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18795–18806.
- [55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [56] D. Stutz, M. Hein, and B. Schiele, "Relating adversarially robust generalization to flat minima," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7787–7797.
- [57] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Red Hook, NY, USA: Curran Associates, 2018, pp. 1–11.
- [58] M. Zach, E. Kobler, and T. Pock, "Computed tomography reconstruction using generative energy-based priors," in *Proc. OAGM Workshop*, Graz, Austria: Verlag der Technischen Universität Graz, 2021, pp. 52–58.