Received: 28 July 2022 | Revised: 5 February 2023 | Accepted: 8 February 2023 | Published on: 13 March 2023

DOI: 10.1002/mrm.29624

RESEARCH ARTICLE

Magnetic Resonance in Medicine

Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models

Guanxiong Luo[1] | **Moritz Blumenthal**[1,2] | **Martin Heide**[1] | **Martin Uecker**[1,2,3,4]

[1]Institute for Diagnostic and Interventional Radiology, University Medical Center Göttingen, Göttingen, Germany

[2]Institute of Biomedical Imaging, Graz University of Technology, Graz, Austria

[3]German Centre for Cardiovascular Research (DZHK) Partner Site Göttingen, Göttingen, Germany

[4]Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen, Germany

Correspondence
Guanxiong Luo, Institute for Diagnostic and Interventional Radiology, University Medical Center Göttingen,
Robert-Koch-Str. 40, 37075 Göttingen, Germany.
Email:
guanxiong.luo@med.uni-goettingen.de

Purpose: We introduce a framework that enables efficient sampling from learned probability distributions for MRI reconstruction.

Method: Samples are drawn from the posterior distribution given the measured k-space using the Markov chain Monte Carlo (MCMC) method, different from conventional deep learning-based MRI reconstruction techniques. In addition to the maximum a posteriori estimate for the image, which can be obtained by maximizing the log-likelihood indirectly or directly, the minimum mean square error estimate and uncertainty maps can also be computed from those drawn samples. The data-driven Markov chains are constructed with the score-based generative model learned from a given image database and are independent of the forward operator that is used to model the k-space measurement.

Results: We numerically investigate the framework from these perspectives: (1) the interpretation of the uncertainty of the image reconstructed from undersampled k-space; (2) the effect of the number of noise scales used to train the generative models; (3) using a burn-in phase in MCMC sampling to reduce computation; (4) the comparison to conventional ℓ_1-wavelet regularized reconstruction; (5) the transferability of learned information; and (6) the comparison to fastMRI challenge.

Conclusion: A framework is described that connects the diffusion process and advanced generative models with Markov chains. We demonstrate its flexibility in terms of contrasts and sampling patterns using advanced generative priors and the benefits of also quantifying the uncertainty for every pixel.

KEYWORDS

Bayesian inference, generative modeling, image reconstruction, inverse problems, Markov chain Monte Carlo, posterior sampling

Parts of this work were presented at the ISMRM 2022.

1 | INTRODUCTION

Modern MRI formulates reconstruction from raw data in Fourier space (k-space) as an inverse problem. Undersampling to reduce acquisition time then leads to an ill-posed reconstruction problem. To solve this problem, parallel imaging can exploit spatial information from multiple receive coils in an extended forward model.[1] Compressed sensing uses the sparsity of images in a transform domain (i.e., wavelet domain, finite differences) as prior knowledge. Combined with incoherent sampling this allows recovery of sparse images from highly undersampled data.[2,3] Learning-based techniques for compressed sensing include methods using dictionary learning[4] or a patch-based nonlocal operator.[5]

In recent years, the application of deep learning pushed these ideas forward by integrating learned prior knowledge.[6] Most of these methods can be classified into two categories: First, methods that unroll the existing iterative reconstruction algorithms into a neural network and train their parameters by maximizing the similarity to a ground truth. In Reference 7, the authors replaced the handcrafted regularization term with convolution layers, and derived a neural network from the iterative procedure of the Alternating Direction Method of Multipliers algorithm. References 8,9 investigated similar approaches. The downside of this kind of method is the need for supervised training, which requires raw k-space data with fixed known sampling patterns and corresponding ground truth images. The second category consists of methods that learn a prior from high-quality images, then plug it into existing iterative algorithms as a regularization term. In References 10-12, the image prior was constructed with a variational auto-encoder,[13] a denoising auto-encoder[14] and an autoregressive generative model,[15] respectively. These methods then compute a maximum a posterior (MAP) as the estimator of the image. These types of methods separate the learned information from the encoding matrix (sampling pattern in k-space and coil sensitivities), which permits more flexibility in practice because they allow the acquisition patterns and receive coils to change without retraining. Generative adversarial networks were also used for image reconstruction in Reference 16. There, the discriminator is used to confine the space of the output of a generator that is designed to generate images with conformity to k-space data.

Although deep learning-based approaches provide promising results, worries about the uncertainty caused by undersampling strategies and algorithms have limited their usage in clinical practice until now. Therefore, the uncertainty assessment constitutes an important step for deep learning-based approaches. The uncertainty is twofold: (1) the uncertainty of weights inside the neural network;[17,18] and (2) the uncertainty introduced by the missing k-space data points. The uncertainty from missing k-space data points can be addressed in a Bayesian imaging framework. We refer the readers to References 19,20. In Reference 11, the MAP estimator is used, but it provides only the mode of the posterior density $p(\mathbf{x}|\mathbf{y})$ and practical optimization may also even only provide a local maximum. In the setting of Bayesian inference, it is possible to investigate the full shape of posterior distribution $p(\mathbf{x}|\mathbf{y})$. In particular, it is possible to draw sample from the posterior distribution for priors based on diffusion models using the Markov chain Monte Carlo (MCMC) method as described previously by Jalal et al.[21] and others,[22-24] which are closely related to the present work. Jalel et al. use Langevin sampling to sample the posterior using score-based generative model and this is extendend in Reference 24 to also include a motion model. The method in Reference 23 uses the predictor-and-corrector framework proposed in Reference 25. These publications point out the relationship to Bayesian reconstruction and show some results related to uncertainty estimation, but a complete Bayesian formulation of this framework applied to MRI multichannel reconstruction is not provided. A general problem with this approach is the large number of iterations required during sampling, for example, Reference 23 reports the use of several thousands of iterations.

Following these ideas, a generic framework for MRI reconstruction emerges, which is based on a series of publications related to generative models,[25-29] in which the essential idea is to: (1) systematically and slowly destroy the underlying prior knowledge in a data distribution through an iterative forward diffusion process; (2) learn a reverse diffusion process that restores the patterns by a so-called score-based neural network; and (3) incorporate the forward model of the measurement into the learned reverse process. The general picture of the proposed method is illustrated in Figure 1.

In the present work, we recapitulate the framework of Bayesian reconstruction and score-based diffusion models and numerically investigate this framework from the following different perspectives: (1) the interpretation of the uncertainty of the image reconstructed from undersampled k-space; (2) the effect of the number of noise scales used the generative models on image quality on computation time; (3) using a burn-in phase in MCMC sampling to reduce computation; (4) the comparison to conventional ℓ_1-wavelet regularized reconstruction; (5) the transferability of learned information; and (6) the comparison to fastMRI challenge.[30,31]

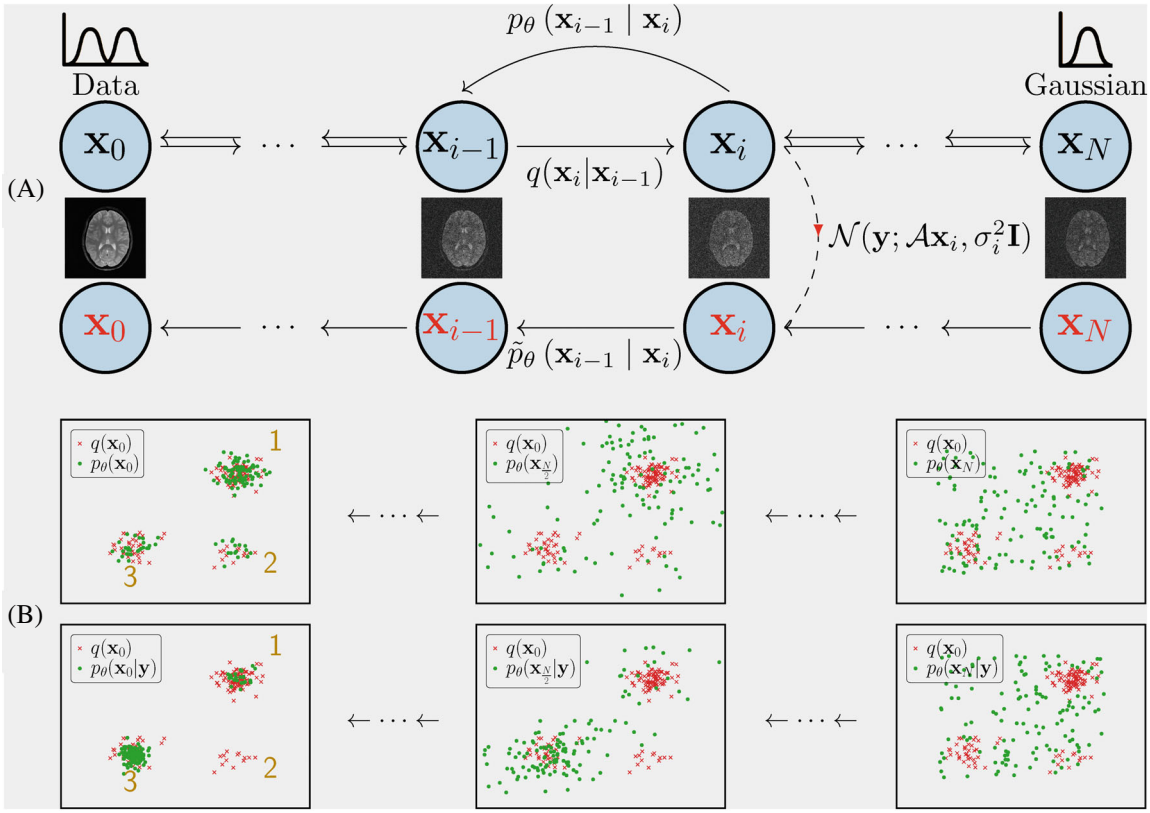

FIGURE 1 Overview of the proposed method. (A) The unknown data distribution $q(\mathbf{x}_0)$ of the training images goes through repeated Gaussian diffusion and finally reaches a known Gaussian distribution $q(\mathbf{x}_N)$, and this process is reversed by learned transition kernels $p_\theta(\mathbf{x}_{i-1}|\mathbf{x}_i)$. To compute the posterior of the image $p(\mathbf{x}|\mathbf{y})$, a new Markov chain $\tilde{p}_\theta(\mathbf{x}_{i-1}|\mathbf{x}_i)$ is constructed by incorporating the measurement model into the reverse process (red chain). (B) Training samples (red dots) from a mixture of bivariate Gaussian distribution are shown. The upper and bottom rows illustrate how samples (green dots) gradually gather around training samples in the reverse process, without and with the observation, respectively. In this example, the likelihood for the observation was a bivariate Gaussian mixture, so that cluster 2 has a lower and cluster 3 has a higher probability.

2 | THEORY

2.1 | MRI reconstruction as Bayesian inference

We consider image reconstruction as a Bayesian problem where the posterior of image $p(\mathbf{x}|\mathbf{y})$ given with the measured data $\mathbf{y}$ and a prior $p(\mathbf{x})$ learned from a database of images.[11,19,20] Here, the image is denoted as $\mathbf{x} \in \mathbb{C}^{n\times n}$, where $n \times n$ is the size of image, and $\mathbf{y} \in \mathbb{C}^{m\times m_C}$ is the vector of m complex-valued k-space samples from m_C receive coils. Assuming the noise η circularly symmetric normal with zero mean and covariance matrix $\sigma_\eta^2\mathbf{I}$, the likelihood $p(\mathbf{y}|\mathbf{x})$ for observing the $\mathbf{y}$ determined by $\mathbf{y} = \mathcal{A}\mathbf{x} + \eta$ and given the image $\mathbf{x}$ is given by a complex normal distributions

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= C\mathcal{N}\left(\mathbf{y}; \mathcal{A}\mathbf{x}, \sigma_\eta^2\mathbf{I}\right) \\ &= \left(\sigma_\eta^2\pi\right)^{-N_p} e^{-\|\sigma_\eta^{-1}\cdot(\mathbf{y}-\mathcal{A}\mathbf{x})\|_2^2}, \end{aligned} \tag{1}$$

where $\mathbf{I}$ is the identity matrix, σ_η the SD of the noise, $\mathcal{A}\mathbf{x}$ is the mean and N_p is the length of the k-space data vector. $\mathcal{A} : \mathbb{C}^{n\times n} \to \mathbb{C}^{m\times m_C}$ is the forward operator and given by $\mathcal{A} = \mathcal{P}\mathcal{F}\mathcal{S}$, where $\mathcal{S}$ are the coil sensitivity maps, $\mathcal{F}$ the two-dimensional Fourier transform, and $\mathcal{P}$ the k-space sampling operator. According to Bayes' theorem the posterior density function $p(\mathbf{x}|\mathbf{y})$ is then

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x})}{p(\mathbf{y})}. \tag{2}$$

In this work, the reconstruction is based on the sampling of this posterior distribution. We utilize an efficient technique based on the MCMC method with the application of a diffusion probabilistic generative model. This consists of two processes: (1) a forward diffusion process which converts a complicated distribution used as prior for the image into a simple Gaussian distribution; and (2) a learned finite-time reversal of this diffusion process with which a Gaussian distribution is gradually transformed back to the posterior (cf. Figure 1).

2.2 | The forward diffusion process

In probabilistic diffusion models, the data distribution characterized by density $q(\mathbf{x}_0)$ is gradually converted into an analytically tractable distribution (Gaussian noise).[28] The image $\mathbf{x}_0$ is perturbed with a sequence of noise scales $0 = \sigma_0 < \sigma_1 < \cdots < \sigma_N$. When the number of steps used for discretization $N \to \infty$, the diffusion process becomes a continuous process. Here, we consider the discrete Markov chain

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \mathbf{z}_{i-1}, \quad i = 1, \cdots, N, \tag{3}$$

where $\mathbf{z}_{i-1} \sim \mathcal{CN}(\mathbf{0}, (\sigma_i^2 - \sigma_{i-1}^2)\mathbf{I})$, that is, the ith transition kernel is then given by

$$q(\mathbf{x}_i|\mathbf{x}_{i-1}) = \mathcal{CN}\left(\mathbf{x}_i; \mathbf{x}_{i-1}, \left(\sigma_i^2 - \sigma_{i-1}^2\right)\mathbf{I}\right). \tag{4}$$

Instead of doing transitions step by step[25,32] a single perturbation kernel

$$q(\mathbf{x}_i|\mathbf{x}_0) = \mathcal{CN}\left(\mathbf{x}_i; \mathbf{x}_0, \sigma_i^2\mathbf{I}\right), \tag{5}$$

can be computed as a convolution of Gaussians. With Bayes' theorem we can write:

$$q\left(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0\right) = q\left(\mathbf{x}_i|\mathbf{x}_{i-1}\right)\frac{q\left(\mathbf{x}_{i-1}|\mathbf{x}_0\right)}{q\left(\mathbf{x}_i|\mathbf{x}_0\right)}. \tag{6}$$

Given the initial image $\mathbf{x}_0$, the posterior of a single step of the forward process is then given by (see Appendix A)

$$q(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0) = \mathcal{CN}\left(\mathbf{x}_{i-1}; \frac{\sigma_{i-1}^2}{\sigma_i^2}\mathbf{x}_i + \left(1 - \frac{\sigma_{i-1}^2}{\sigma_i^2}\right)\mathbf{x}_0, \tau_i^2\mathbf{I}\right), \tag{7}$$

with variance $\tau_i^2 := \left(\sigma_i^2 - \sigma_{i-1}^2\right)\left(\sigma_{i-1}^2/\sigma_i^2\right)$.

2.3 | Learning the reverse process

The joint distribution of the reversal diffusion process is characterized by the probability density

$$p(\mathbf{x}_N, \mathbf{x}_{N-1}, \ldots, \mathbf{x}_0) = p(\mathbf{x}_N)\prod_{i=1}^{N} p(\mathbf{x}_{i-1}|\mathbf{x}_i), \tag{8}$$

where $p(\mathbf{x}_N)$ is the initial Gaussian distribution. The reverse is given by Kolmogorov's backward equation which has the same form as the forward process.[28,32] Thus, the transitions $p(\mathbf{x}_{i-1}|\mathbf{x}_i)$ of the reverse process can be parameterized with the Gaussian transition kernel

$$p\left(\mathbf{x}_{i-1}|\mathbf{x}_i\right) = \mathcal{CN}\left(\mathbf{x}_{i-1}; \boldsymbol{\mu}\left(\mathbf{x}_i, i\right), \tau_i^2\mathbf{I}\right), \tag{9}$$

where $\boldsymbol{\mu}(\mathbf{x}_i, i)$ and $\tau_i^2\mathbf{I}$ are the mean and variance of the reverse transitions, respectively. Here, we learn the mean $\mu_{\boldsymbol{\theta}}$ of the reverse transitions using a neural network parameterized by training parameters θ. Since the learned reverse transitions $p_\theta(\mathbf{x}_{i-1}|\mathbf{x}_i)$ lead to a new density $p_{\boldsymbol{\theta}}(\mathbf{x}_0)$, which should match $q(\mathbf{x}_0)$, they can be learned by minimizing the cross entropy

$$H(p_{\boldsymbol{\theta}}, q) = -\mathbb{E}_{q(\mathbf{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}_0)\right]. \tag{10}$$

Following Reference 28 a lower bound ℓ can be written in terms of Kullback-Leibler (KL) divergence between the transition kernel Equation (9) and the posterior of forward process Equation (7)

$$\begin{aligned}\ell &= \sum_{i=2}^{N}\mathbb{E}_{q(\mathbf{x}_0)}\mathbb{E}_{q(\mathbf{x}_i|\mathbf{x}_0)}\left[D_{\mathrm{KL}}(q(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)\|\, p_\theta(\mathbf{x}_{i-1}|\mathbf{x}_i))\right] \\ &= \sum_{i=2}^{N}\mathbb{E}_{q(\mathbf{x}_0)}\mathbb{E}_{q(\mathbf{x}_i|\mathbf{x}_0)} \\ &\quad \times\left[\frac{1}{\tau_i^2}\left\|\frac{\sigma_{i-1}^2}{\sigma_i^2}\mathbf{x}_i + \left(1 - \frac{\sigma_{i-1}^2}{\sigma_i^2}\right)\mathbf{x}_0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_i, i)\right\|_2^2\right] + C,\end{aligned} \tag{11}$$

where C is a constant. The derivation of KL divergence between two Gaussian distributions is detailed in Appendix B. Using Equation (5) we can express $\mathbf{x}_i = \mathbf{x}_0 + \mathbf{z}$ with $\mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_i^2\mathbf{I}\right)$, and obtain

$$\ell = \sum_{i=2}^{N}\mathbb{E}_{\mathbf{x}_0,\mathbf{z}}\left[\frac{1}{\tau_i^2}\left\|\frac{\sigma_{i-1}^2}{\sigma_i^2}\mathbf{z} + \mathbf{x}_0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\mathbf{x}_i, i\right)\right\|_2^2\right] + C. \tag{12}$$

Thus, we can learn the mean of the reverse transitions by learning to denoise the training data disturbed by noise. In References 26,27, the generative model is estimated by minimizing the expected squared distance between the gradient of the log-probability given by the score network and the gradient of the log-probability of the observed data. This technique was extended and generalized in References 25,29. In the following, we quickly point out the connection to score matching networks. Let:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\mathbf{x}_i, i\right) - \mathbf{x}_0 = \sigma_{i-1}^2\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_i, i\right), \tag{13}$$

where $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, i)$ denotes the denoising score matching network that is conditional on the index of noise scales i. Then, we have

$$\ell = \sum_{i=2}^{N}\mathbb{E}_{\mathbf{x}_0,\mathbf{z}}\left[\frac{\sigma_{i-1}^2}{\tau_i^2}\left\|\frac{\mathbf{z}}{\sigma_i^2} - \boldsymbol{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_i, i\right)\right\|_2^2\right] + C. \tag{14}$$

Expressing the noise again as $\mathbf{z} = \mathbf{x}_i - \mathbf{x}_0$, we can rewrite

$$\mathbb{E}_{\mathbf{x}_0,\mathbf{z}}\left[\left\|\frac{\mathbf{x}_i - \mathbf{x}_0}{\sigma_i^2} - \mathbf{s}_\theta(\mathbf{x}_i, i)\right\|_2^2\right] = \mathbb{E}_{q(\mathbf{x}_0)}\mathbb{E}_{q(\mathbf{x}_i|\mathbf{x}_0)}\left[\left\|\nabla_{\mathbf{x}_i}\log q(\mathbf{x}_i|\mathbf{x}_0) - \mathbf{s}_\theta(\mathbf{x}_i, i)\right\|_2^2\right], \quad (15)$$

which shows that Equation (14) is equivalent to score matching. For the later use of the transition kernel, Equation (13) is equivalent to

$$\boldsymbol{\mu}_\theta(\mathbf{x}_i, i) - \mathbf{x}_i = \left(\sigma_i^2 - \sigma_{i-1}^2\right)\mathbf{s}_\theta(\mathbf{x}_i, i). \quad (16)$$

In summary, the score network is trained via Equation (15) to output the gradient fields that are used to construct the Markov transitions (Equation 9) which nudges coarse samples $\mathbf{x}_i$ toward finer ones $\mathbf{x}_{i-1}$, namely the reverse process. In later sections, we will discuss how we construct and train the score networks.

2.4 | Computing the posterior for MRI reconstruction

In order to compute the posterior probability $p(\mathbf{x}|\mathbf{y})$ for the image $\mathbf{x}$ given the data $\mathbf{y}$, we need to modify the learned reverse process. We achieve this by multiplying each of the intermediate distributions $p(\mathbf{x}_i)$ with the likelihood term $p(\mathbf{y}|\mathbf{x}_i)$ according to Bayes' theorem. We use $\tilde{p}(\mathbf{x}_i) = p(\mathbf{x}_i|\mathbf{y})$ to denote the resulting sequence of intermediate distributions

$$\tilde{p}(\mathbf{x}_i) \propto p(\mathbf{x}_i)\, p(\mathbf{y}|\mathbf{x}_i), \quad (17)$$

up to the unknown normalization constant. Following Reference 28, the transition from $\mathbf{x}_{i+1}$ to $\mathbf{x}_i$ of the modified reverse process is

$$\tilde{p}(\mathbf{x}_i|\mathbf{x}_{i+1}) \propto p(\mathbf{x}_i|\mathbf{x}_{i+1})\, p(\mathbf{y}|\mathbf{x}_i). \quad (18)$$

The sampling at each intermediate distribution of Markov transitions Equation (18) is performed with the unadjusted Langevin algorithm[33]

$$\mathbf{x}_i^{k+1} \leftarrow \mathbf{x}_i^k + \frac{\gamma}{2}\nabla_{\mathbf{x}_i}\log\tilde{p}(\mathbf{x}_i^k|\mathbf{x}_{i+1}) + \sqrt{\gamma}\mathbf{z}, \quad (19)$$

where $\mathbf{z}$ is standard complex Gaussian noise $\mathcal{CN}(0, \mathbf{I})$. We now go over to the modified learned process $\tilde{p}_\theta(\mathbf{x}_i|\mathbf{x}_{i+1})$ parameterized by θ and obtain the log-derivative with respect to $\mathbf{x}_i$ using the learned reverse transitions $p_\theta(\mathbf{x}_i|\mathbf{x}_{i+1})$ as

$$\nabla_{\mathbf{x}_i}\log\tilde{p}_\theta(\mathbf{x}_i|\mathbf{x}_{i+1}) = \nabla_{\mathbf{x}_i}\log p_\theta(\mathbf{x}_i|\mathbf{x}_{i+1}) + \nabla_{\mathbf{x}_i}\log p(\mathbf{y}|\mathbf{x}_i). \quad (20)$$

From Equation (9) and Equation (16), we have

$$\nabla_{\mathbf{x}_i}\log p_\theta(\mathbf{x}_i|\mathbf{x}_{i+1}) = \frac{1}{\tau_{i+1}^2}\left(\sigma_{i+1}^2 - \sigma_i^2\right)\mathbf{s}_\theta(\mathbf{x}_{i+1}, i), \quad (21)$$

and from Equation (1) we have

$$\nabla_{\mathbf{x}_i}\log p(\mathbf{y}|\mathbf{x}_i) = -\frac{1}{\sigma_\eta^2}\left(\mathcal{A}^H\mathcal{A}\mathbf{x}_i - \mathcal{A}^H\mathbf{y}\right). \quad (22)$$

After inserting these expressions into Equation (19) we obtain

$$\mathbf{x}_i^{k+1} \leftarrow \mathbf{x}_i^k + \frac{\gamma}{2\tau_{i+1}^2}\left(\sigma_{i+1}^2 - \sigma_i^2\right)\mathbf{s}_\theta\left(\mathbf{x}_i^k, i\right) - \frac{\gamma}{2\sigma_\eta^2}\left(\mathcal{A}^H\mathcal{A}\mathbf{x}_i^k - \mathcal{A}^H\mathbf{y}\right) + \sqrt{\gamma}\mathbf{z}. \quad (23)$$

The starting point for each chain $\mathbf{x}_i^0 = \mathbf{x}_{i+1}^{\mathrm{K}}$ is the last sample from the previous distribution $\tilde{p}(\mathbf{x}_{i+1}|\mathbf{x}_{i+2})$ after K Langevin steps. We found it advantageous to modify the likelihood term in each step according $\sigma_\eta^2 = \tau_{i+1}/\lambda$, which should approach the variance of the data noise in the last step. Since the noise variance was unknown for the data set we used, we empirically selected a λ that determines how strong the k-space data consistency is relative to the prior. We set γ to $2\tau_{i+1}^2$. At last, the algorithm used to sampling the posterior is presented in Algorithm 1.

Algorithm 1. SAMPLING THE POSTERIOR WITH A MCMC METHOD

1: Give the acquired k-space $\mathbf{y}$.
2: Construct the forward operator $\mathcal{A}$ with sampling pattern $\mathcal{P}$ and coil sensitivities $\mathcal{S}$.
3: Set the Langevin steps K, the factor λ, the start noise level index N, and γ.
4: Generate $\mathbf{x}_{\mathrm{N}}^0$ from a suitable Gaussian distribution (e.g., $\mathcal{CN} \sim (0, \mathbf{I})$).
5: **for** i in $\{\mathrm{N}-1, \cdots, 1\}$ **do**
6: Draw samples from $\tilde{p}(\mathbf{x}_i|\mathbf{x}_{i+1})$ by running K Langevin steps with Equation (23).
7: **end for**

To characterize the shape of a posterior, we run multiple chains to draw samples in parallel. To reduce the amount of computation, the burn-in phase is introduced as shown in Figure 2. That means only one chain proceeds through the several beginning noise levels, and after that we split it up into multiple Markov chains using the sample from the burn-in phase as initial point indicated by the

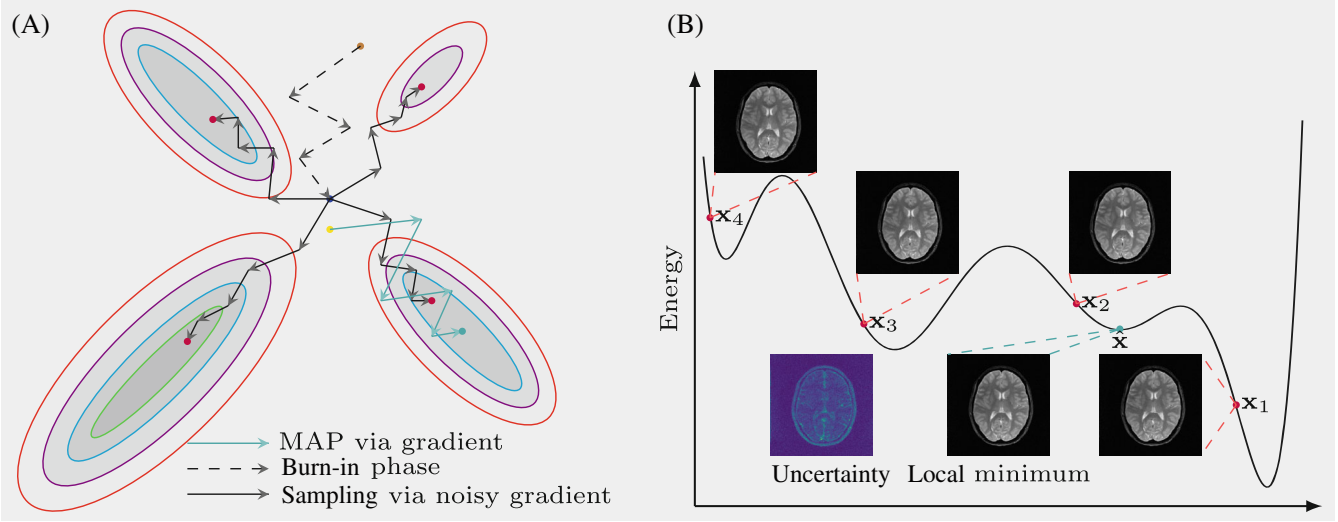

FIGURE 2 Illustration for the sampling of the posterior $p(\mathbf{x}|\mathbf{y})$. (A) The four possible sampling trajectories are indicated the solid lines, sharing the same burn-in phase (dashed line). The MAP approach via gradient descent reaches a locally optimal solution. (B) Possible reconstructions are showed over the energy curve and the uncertainty map is the pixelwise variance over samples.

blue dot. To further reduce computation, we introduce the continuously decreasing noise scales, which reduces the number of iterations when performing Langevin dynamics at each intermediate distribution.

2.5 | The analysis of samples

Given a posterior probability distribution $p(\mathbf{x}|\mathbf{y})$ the minimum mean square error (MMSE) estimator minimizes the mean square error:

$$\mathbf{x}_{\text{MMSE}} = \arg\min_{\tilde{\mathbf{x}}} \int \|\tilde{\mathbf{x}} - \mathbf{x}\|^2 p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \mathbb{E}[x|y]. \quad (24)$$

The MMSE estimator cannot be computed in a closed form, and numerical approximations are typically required. Since we demonstrated how to generate samples from the posterior in previous sections, let us consider the samples $\mathbf{x}_0^{\text{K}}$ at the last stage, and a consistent estimate of $\mathbf{x}_{\text{MMSE}}$ can be computed by averaging those samples, i.e. the empirical mean of samples converges in probability to $\mathbf{x}_{\text{MMSE}}$ due to weak law of large numbers. The variance of those samples is a solution to the error assessment for the reconstruction if we trust the model parameterized by Equation (9) that is learned from a image database. The 95% confidence interval is computed for each pixel with its mean and variance. Since a wider confidence interval means a larger margin of error, the mean is overlaid with it to indicate the variability of each pixel, and up to a certain point, the variability can cause a visual change on the image (cf. Section 3.3.8).

3 | METHODS

3.1 | Score networks' architecture

The denoising score network is designed to predict the noise given an image degraded by Gaussian noise of a particular scale σ_i. To improve the quality of the predictions for different noise scales, we consider networks conditional on discrete and pseudo-continuous noise scales. The discrete one has a much larger gap between σ_i and σ_{i-1} than the pseudo-continuous one and usually has a smaller number of noise scales N, while the pseudo-continuous network is adaptive to a certain trained range of noise scales. The sequence of noise scales $\{\sigma_i\}_{i=1}^{\text{N}}$ is geometrically generated following the scheme in Reference 25, that is, $\sigma_i = \sigma\left(\frac{i}{N}\right) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{i-1}{N-1}}$.

For a discrete model, we add modified instance normalization layers that are conditional on the index of the noise scales following each convolution layer. The conditional instance normalization[34] is

$$\hat{f}_k = \Phi[i,k]\frac{f_k - \mu_k}{s_k} + \Omega[i,k], \quad (25)$$

where $\Phi \in \mathbb{R}^{N\times C}$ and $\Omega \in \mathbb{R}^{N\times C}$ are learnable parameters, k denotes the index of a feature map f_k, μ_k and s_k are the means and SD over its spatial locations of the kth feature map computed in each pass through the network, and i denotes the index of σ in $\{\sigma_i\}_{i=1}^{\text{N}}$.

For a continuous model, we let networks be conditional on the index of noise scales by inserting random

Fourier features.[35] Three steps used to encode a noise index into random features are as follows:

- Draw a random vector which has i.i.d. Gaussian m entries with the specified standard deviation,
- Scale the random vector with the index i, then multiply it with 2π,
- Apply sines and cosines to the scaled random vector, then concatenate them into $m \times 2$ matrix,

where m is embedding size. The encoded index is added to all the blocks listed in Table S1.

With either one of the two modifications above, a network $\mathbf{s}_\theta(\mathbf{x}, i)$ has two inputs, that is, noise corrupted image $\mathbf{x}$ and noise index i. Real and imaginary parts of the images are interpreted as separate channels when input into the neural network. RefineNet[36] is the backbone of all the score networks used in this work (cf. Figure S2). Three variants from that are trained for different reconstruction experiments. The architectures of three networks are presented in detail in Table S1. We refer the readers to the codes available online for more information about them. We labeled the three networks with $\texttt{NET}_1$, $\texttt{NET}_2$, and $\texttt{NET}_3$, respectively, for ease of reference in the following. $\texttt{NET}_1$ is conditional on discrete noise scales, $\texttt{NET}_2$ and $\texttt{NET}_3$ are conditional on continuous noise scales. We introduce self-attention modules into $\texttt{NET}_3$ to capture long-range dependencies by adding nonlocal blocks as described previously[37] so that the network has the capability to model the dataset of high-resolution images.

3.2 | Dataset, training, and inference

We trained $\texttt{NET}_1$ and $\texttt{NET}_2$ on a dataset acquired by us already used and described in Reference 11. $\texttt{NET}_3$ was trained on a subset of the fastMRI dataset.[30] Our dataset has 1300 images containing T1-weighted, T2-weighted, T2-weighted fluid-attenuated inversion recovery (FLAIR), and T2*-weighted brain images from 13 healthy volunteers examined with clinical standard-of-care protocols. The brain images from fastMRI dataset[30] were used for benchmark that contains T1-weighted (some with post contrast), T2-weighted and FLAIR images. For the detailed information of both dataset, we refer readers to corresponding publication. Regarding the data partitioning, we first separated all multislice volumes into training and testing groups. Then we split the volume into two-dimensional slices (i.e., images). Reference images—denoted $\mathbf{x}_0$ in the theory—were reconstructed from fully sampled multichannel k-space. Then, these complex image datasets after coil combination were normalized to a maximum magnitude of 1. The coil sensitivity maps were computed with BART toolbox using ESPIRiT.[38,39] 1300 images of size 256×256 from the dataset used in Reference 11 were used to train $\texttt{NET}_1$ and $\texttt{NET}_2$. 1000 images were used for training, and 300 images were used for testing. All networks are trained for 1000 epochs, that is, iterations over all training images. For the training of $\texttt{NET}_3$, we used the T2-weighted FLAIR contrast images of size 320×320 that are reconstructed from fastMRI raw k-space data. 2937 images are for training, 326 images are for testing.

Three score networks are implemented with Tensorflow.[40] The hyperparameters used to train the three score networks are listed in Table S2. With the trained networks, we implemented MCMC sampling Algorithm 1 with Tensorflow and Numpy,[41] and then explored the posterior $p(\mathbf{x}|\mathbf{y})$ in different experimental settings. We trained three score networks once separately for all the experiments we did in this work. These three models can support all experiments performed in this study with variable undersampling patterns, coil sensitivity maps, channel numbers. It took around 43 and 67 s, respectively, to train $\texttt{NET}_1$ and $\texttt{NET}_2$ for one epoch on one NVIDIA A100 GPU with 80GB. For $\texttt{NET}_3$, it took around 500 s per epoch on two NVIDIA A100 GPUs using the multi-GPU support from Tensorflow. In the spirit of reproducible research, codes and data to reproduce all experiments are made available*.

3.3 | Experiments

3.3.1 | Single coil unfolding

To investigate how the Markov chain explores the solution space of the inverse problem $\mathbf{y} = \mathcal{A}\mathbf{x} + \eta$, we designed the single coil unfolding experiment. The single channel k-space is simulated out of multichannel k-space data. The odd lines in k-space are retained. Ten samples were drawn from the posterior $p(\mathbf{x}|\mathbf{y})$. $\texttt{NET}_1$ was used to construct transition kernels and the parameters in Algorithm 1 are $K = 50, N = 10, \lambda = 6$. We redo the experiment with the object shifted to bottom. This experiment has an inherent ambiguity which cannot be resolved using the data alone and where the reconstruction is strongly determined by the prior. Thus, it mimics in a synthetic setting a situation with high undersampling where hallucinations were observed in the reconstruction of some deep-learning methods.[42]

3.3.2 | Multicoil reconstruction

Multichannel data points from Cartesian k-space are randomly picked with variable-density poisson-disc sampling and the central 20×20 region is fully acquired. The acquisition mask covers 11.8% k-space and the

corresponding zero-filled reconstruction is shown Figure 4B. We initialized 10 chains and the $\mathbf{x}_{\text{MMSE}}$ was computed using different numbers of samples. $\texttt{NET}_1$ was used to construct transition kernels and the parameters in Algorithm 1 are $K = 30, N = 15, \lambda = 13$. To visualize the process of sampling, we use peak-signal-noise-ratio (PSNR in dB) and similarity index (SSIM) as metrics to track intermediate samples. The comparisons are made between the magnitude of $\mathbf{x}_{\text{MMSE}}$ and the ground truth $\tilde{\mathbf{x}}$ after normalized with ℓ_2-norm.

3.3.3 | More noise scales

To investigate how the number of noise scales influences the proposed method, we reconstructed the image from the undersampled k-space that was used in the multicoil experiment. $\texttt{NET}_2$ was used to construct transition kernels and the parameters in Algorithm 1 are $K = 5, N = 70, \lambda = 25$.

3.3.4 | Investigation of the Burn-in Phase

To investigate the burn-in phase illustrated in Figure 2, we split up into multiple chains at a certain noise scale when drawing samples from the posterior $p(\mathbf{x}|\mathbf{y})$. For instance, we denote by $(\mathbf{x}_{\text{MMSE}}, 60)$ the $\mathbf{x}_{\text{MMSE}}$ that is computed with 10 samples drawn from $p(\mathbf{x}|\mathbf{y})$ by splitting up into 10 chains at the 60th noise scale. By changing the splitting point, we got different sets of samples that are from chains of different length and computed the final $\mathbf{x}_{\text{MMSE}}$, respectively. We have two sets of $\mathbf{x}_{\text{MMSE}}$ that are reconstructed from the undersampled k-space using two sampling patterns separately. The central 20×20 region is obtained and the k-space, outside the center, is randomly picked up retrospectively (10%, 20%). $\texttt{NET}_2$ was used to construct Markov transition kernels and the parameters in Algorithm 1 are $K = 5, N = 70, \lambda = 25$.

3.3.5 | Investigation into MAP

To verify the samples are located around the local modality of the posterior, we disabled the disturbance with noise after stochastic inference with the last distribution $\tilde{p}(\mathbf{x}_0|\mathbf{x}_1)$ and ran 200 iterations more to get extended samples. What is more, we repeated this procedure with determinate inference, in which the disturbance was disabled during sampling iterations to get one deterministic sample, that is, MAP estimation. A Poisson-disc sampling pattern is generated without variable density and with twofold undersampling along phase and frequency encoding directions. $\texttt{NET}_2$ was used to construct transition kernels and the parameters in Algorithm 1 are $K = 5, N = 70, \lambda = 25$.

3.3.6 | Comparison to ℓ_1-regularized Reconstruction

A comparison using the fastMRI dataset was used to evaluate the performance of the proposed method. We noticed that the raw k-space data is padded with zeros to make them have the same dimension. The effect caused by zero paddings is investigated in Reference 43. Since we only used the images that were reconstructed from the zero padded k-space for training, the issue caused by the synthesized k-space does not exist in our work. The undersampling pattern for each slice is randomly generated in all retrospective experiments. $\texttt{NET}_3$ was used to construct transition kernels. The parameters in Algorithm 1 are $K = 3, N = 90, \lambda = 20$ and 10 samples were drawn to compute $\mathbf{x}_{\text{MMSE}}$. The data range for computing PSNR and SSIM is determined by the maximum over each slice.

3.3.7 | Transferability

To investigate the transferability of learned prior information from T2 FLAIR images to other contrasts, we acquired T1-weighted (pulse repetition time = 2000 ms, inversion time = 900 ms, echo time = 9 ms) and T2-weighted (pulse repetition time = 9000 ms, inversion time = 2500 ms, echo time = 81 ms) FLAIR k-space data using a two-dimensional multislice turbo spin-echo sequence with a 16-channel head coil at 3T (Siemens, 3T Skyra). $\texttt{NET}_3$ (trained with T2 FLAIR images) was used to construct transition kernels. The parameters in Algorithm 1 are $K = 5, N = 70, \lambda = 20$.

3.3.8 | Comparison to fastMRI challenge

As a comparison to the unrolled neural network, the XPDNet[31] is selected as the reference which ranked 2nd in the fastMRI challenge. Two networks were trained for acceleration factors 4 and 8, using retrospectively undersampled data from the fastMRI dataset[10] using equidistant Cartesian masks and the trained models that are publicly available[†]. For the proposed method, $\texttt{NET}_3$ was used to construct transition kernels. The parameters in Algorithm 1 are $K = 4, N = 90, \lambda = 20$. The confidence interval after thresholding is used as the color map to indicate that a region has high uncertainty. Be consistent with the evaluation the XPDNet provided, 30 FLAIR volumes are used for validation to compute metrics.

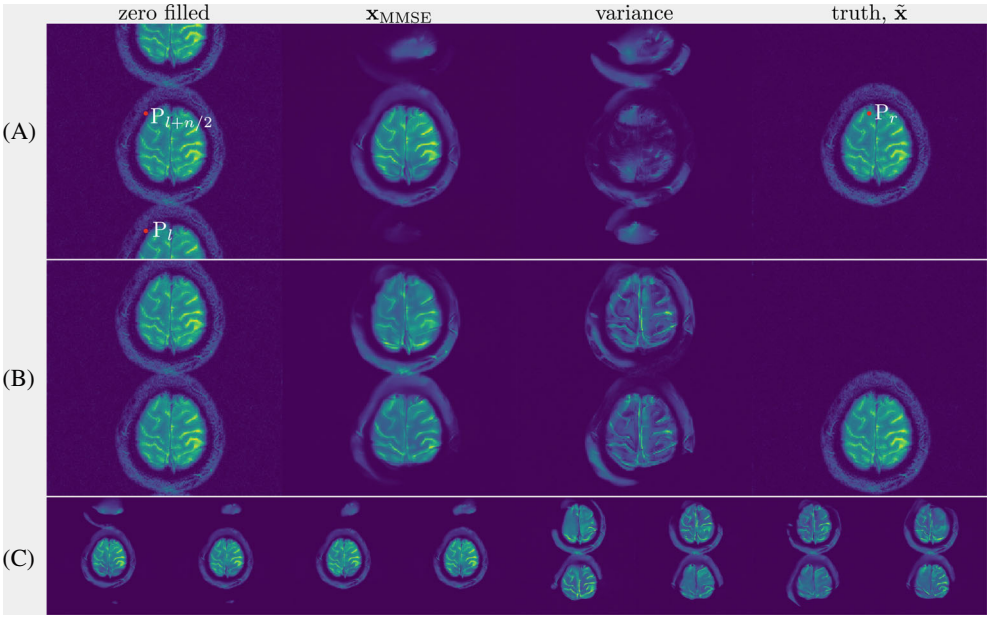

FIGURE 3 Single-coil unfolding with NET_1. The k-space is undersampled by skipping every second line. Aliased images, $\mathbf{x}_{\mathrm{MMSE}}$, variance maps and ground truth are shown. (A) The object is centered. (B) The object is shifted. (C) Selected solutions are presented. The left four are centered and the right four are shifted.

4 | RESULTS

4.1 | Single coil unfolding

As expected, the lack of spatial information from coil sensitivities without parallel imaging leads to huge errors and folding artifacts still exist in $\mathbf{x}_{\mathrm{MMSE}}$ as shown in Figure 3. Since only odd lines are acquired, all images in which the superposition of points P_l and $\mathrm{P}_{l+2/n}$ equals to the points P_r in ground truth are solutions to $\mathbf{y} = \mathcal{A}\mathbf{x} + \epsilon$ with the same error (the residual norm $\|\mathbf{y} - \mathcal{A}\mathbf{x}\|^2$). Selected solutions are presented in Figure 3C. The variance map indicates the uncertainty of the solutions, which in this experiment is similar to the hallucinations observed in for some deep-learning methods for high undersampling.[42] The errors of the estimation $\mathbf{x}_{\mathrm{MMSE}}$ are largely reduced compared to the zero-filled reconstruction because of prior knowledge from the learned reverse process (cf. Figure 3A). The shift of the object increases the symmetry and then leads to even bigger errors as learned reverse process know less about images that were shifted (cf. Figure 3B).

4.2 | Multicoil reconstruction

Figure 4 shows the results for the multicoil experiment. Figure 4A shows the evolution of the samples' PSNR and SSIM over the transitions of the data-driven Markov chain. Intermediate samples are presented in Figure S1. The convergence of samples at each noise level was reached as indicated by the PSNR and SSIM curves. When there are more samples, the $\mathbf{x}_{\mathrm{MMSE}}$ converges to higher PSNR and SSIM. In Figure 4B, 10 converged samples were used to compute $\mathbf{x}_{\mathrm{MMSE}}$ and the variance map. Comparing with the ground truth, the variance map mainly reflects the edge information, which can be interpreted by the uncertainty that is introduced by the undersampling pattern used in k-space where many high-frequency data points are missing but the low frequency data points are fully acquired. In contrast to the single-coil unfolding, the local spatial information from coil sensitivities reduces the uncertainties of missing k-space data. Moreover, error maps qualitatively correspond to the variance map, with larger errors in higher variance regions as shown in Figure 4C. Lastly, the average over more samples leads to smaller error.

4.3 | More noise scales

We also plotted the curve of PSNRs and SSIMs over iterations in Figure 5A for NET_2 which uses continuous noise scales. The PSNR and SSIM of $\mathbf{x}_{\mathrm{MMSE}}$, which is computed with 10 samples, are 37.21 dB and 0.9360, respectively. Two $\mathbf{x}_{\mathrm{MMSE}}$ reconstructed separately with the application of NET_1 and NET_2 are presented in Figure 5B and variance maps are presented as well. The variance of the samples that are drawn with NET_2 is less than those drawn with NET_1, which means that we are more confident about the reconstruction using NET_2. When we zoom into the region that has more complicated structures, the boundaries between white matter and gray matter are more distinct in the image recovered with NET_2 and the details are more obvious, as shown in Figure 5C. Hence, increasing the number of noise scales in NET_2 relative to NET_1 reduces the number of iterations and improves the quality of

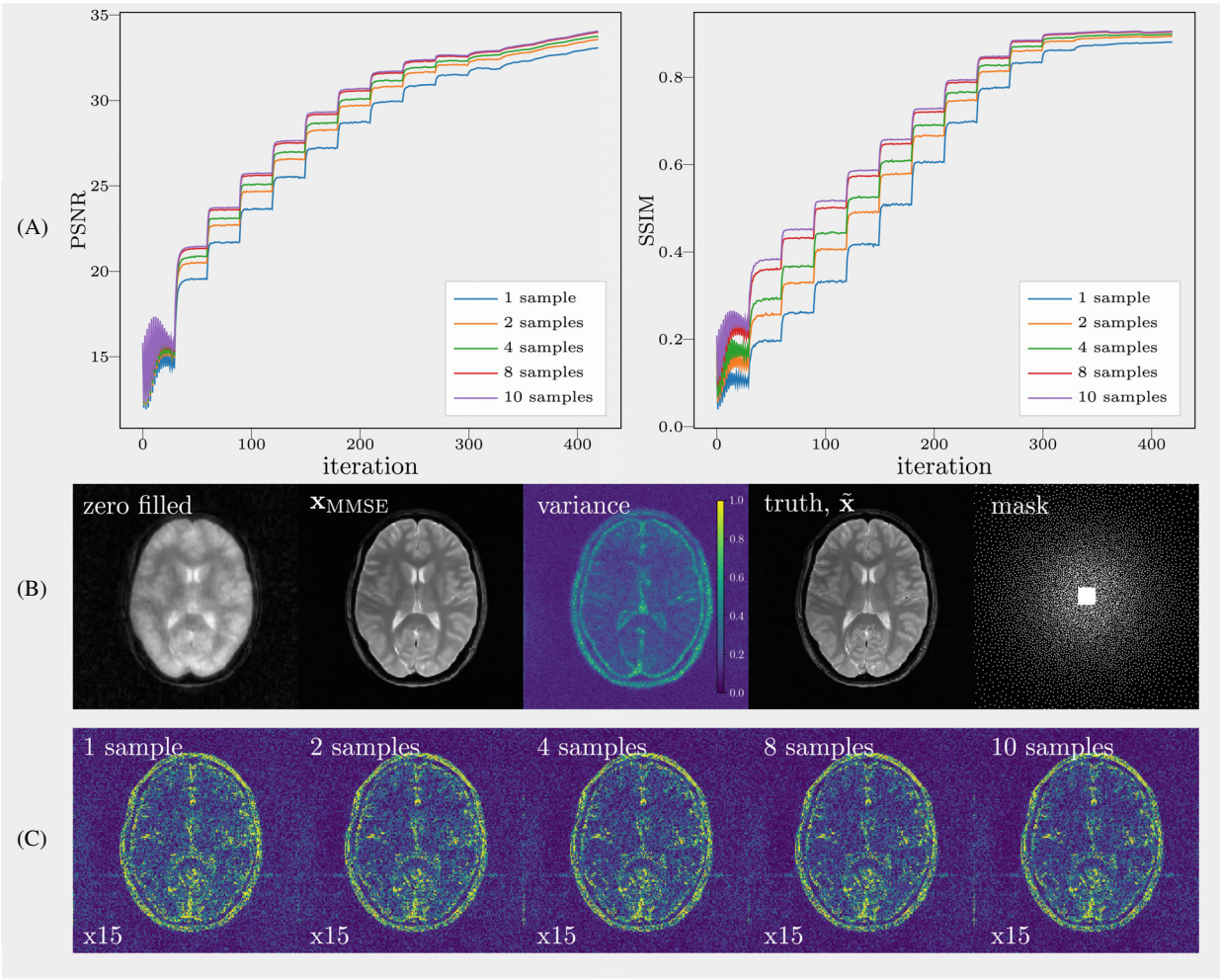

FIGURE 4 Multicoil reconstruction with NET_1. Results: (A) The curves of peak-signal-noise-ratio (PSNR) and similarity index (SSIM) over iterations for $\mathbf{x}_{\mathrm{MMSE}}$s estimated by averaging a different number of samples. (B) Zero-filled, $\mathbf{x}_{\mathrm{MMSE}}$, variance maps, truth and mask are presented. The final PSNR and the SSIM of $\mathbf{x}_{\mathrm{MMSE}}$ are 34.05dB and 0.9050, respectively. (C) The error maps between different $\mathbf{x}_{\mathrm{MMSE}}$s and the ground truth are presented.

reconstruction using score networks of comparable size. More noise scales make chains constructed with NET_2 exploit the prior knowledge from training image dataset more effectively than chains constructed with NET_1 which has fewer noise scales.

4.4 | Investigation of the burn-in phase

The two sets of $\mathbf{x}_{\mathrm{MMSE}}$ are presented in Figure 6. In Figure 6A, the earlier we split chains, the closer the $\mathbf{x}_{\mathrm{MMSE}}$ gets to the truth. Especially, when we zoom into the region that has complicated structures (indicated by the red rectangle), the longer chains make fewer mistakes. The slightly distorted structure is seen in $(\mathbf{x}_{\mathrm{MMSE}}, 60)$ highlighted with blue circles. The distortion has disappeared in $(\mathbf{x}_{\mathrm{MMSE}}, 0)$ but some details are still missing. However, given more k-space data points, the longer chains do not cause a huge visual difference in the $\mathbf{x}_{\mathrm{MMSE}}$ as shown in Figure 6B, even though there is a slight increase in PSNR and SSIM. Although fewer data points mean more uncertainties, longer chains permit better exploration of the solution space, as shown by this experiment. Here, the image $(\mathbf{x}_{\mathrm{MMSE}}, 60)$ took about one fourth of the time (4 min and 30 s) to compute than the image $(\mathbf{x}_{\mathrm{MMSE}}, 0)$. For moderate undersampling rates, a burn-in phase is recommended for reducing computation time.

4.5 | Investigation of the MAP

In Figure 7, we plotted the curves of PSNR and SSIM over extended iterations for NET_2 and presented reconstructions that are from the MMSE and MAP estimator. As indicated by zoom-in images and curves in Figure 7A,C, the extended samples converge to a consistent estimate

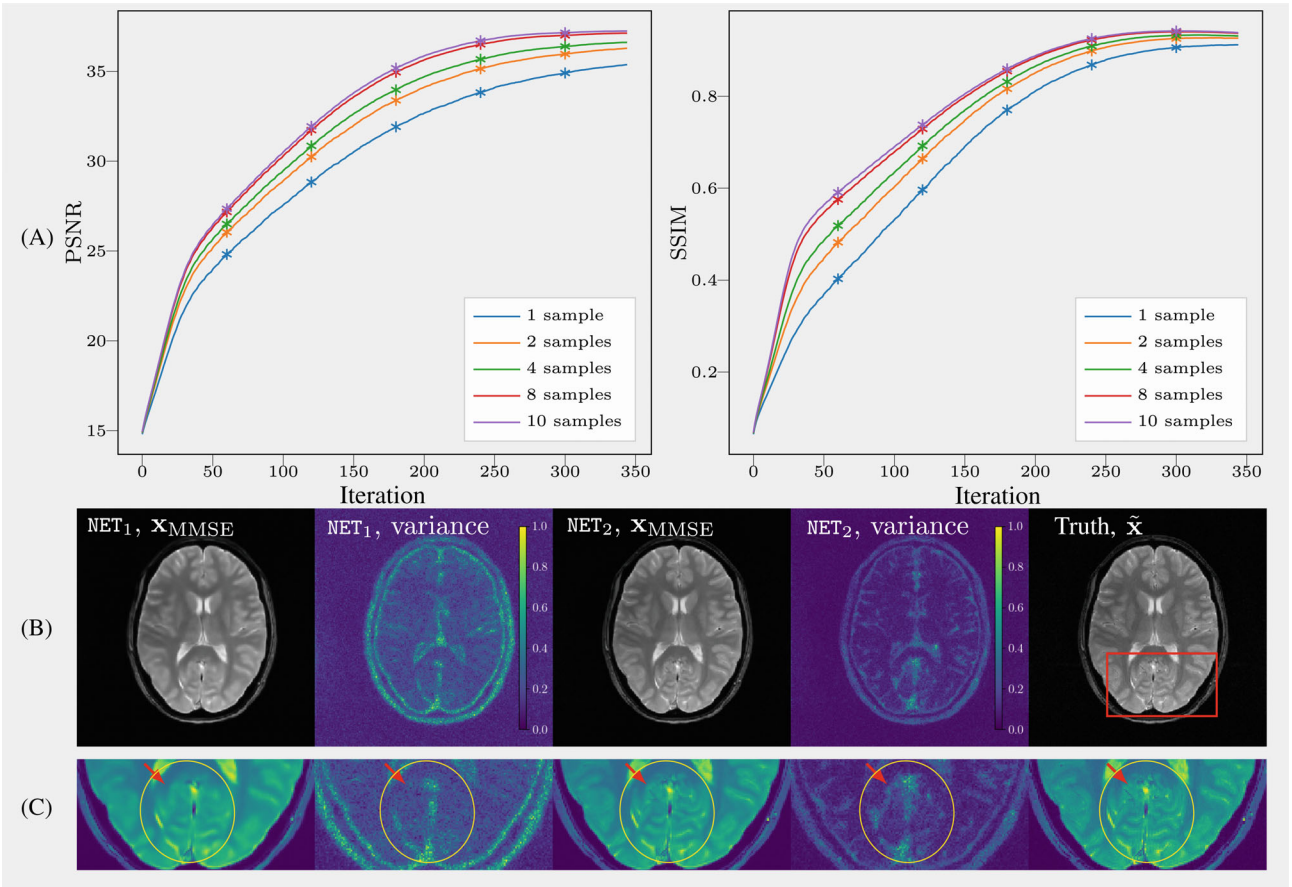

FIGURE 5 Effect of using continuous noise scales in $\texttt{NET}_2$. (A) The convergence curves of peak-signal-noise-ratio (PSNR) and similarity index (SSIM) over iterations for $\texttt{NET}_2$. (B) Reconstructed minimum mean square error and variance maps for $\texttt{NET}_2$ and $\texttt{NET}_2$. (C) Zoomed view of selected structures (yellow circle, red arrow).

of the MAP. Measured by PSNR and SSIM, the MAP has better quality than individual samples. As expected, the MMSE obtained from averaging 10 (nonextended) samples has better PSNR and SSIM than the MAP.

4.6 | Comparison to ℓ_1-regularized Reconstruction

The reconstructions with different methods are presented in Figure 8. ℓ_1-ESPIRiT denotes the reconstruction with the `pics` command of BART toolbox using ℓ_1-wavelet regularization (0.01), which mostly recovers general structures while smoothing out some details. In $\mathbf{x}_{\text{MMSE}}$, the majority of details are recovered, and the texture is almost identical to the ground truth, although some microscopic structures are still missing. Each subject has 16 slices and the metrics of three subjects presented in Table S3 are the average over slices of each subject. It is worth mentioning that PSNR and SSIM are influenced by the value-range of a slice in the evaluation of MR images.

4.7 | Transferability

Figure 9 shows a $\texttt{NET}_3$ trained with T2 FLAIR contrast used to reconstruct a T1 FLAIR image (red box) in comparison to a T2 FLAIR image. No loss of quality can be observed.

4.8 | Comparison to fastMRI challenge

As discussed in Reference 44, the ground truth matters when computing comparison metrics. We plotted the metrics of 30 volumes against a root sum of squares and a coil combined image (CoilComb) in Figure S3, which shows XPDNet favors root sum of squares that was used as labels for training it while $\mathbf{x}_{\text{MMSE}}$ favors the other. Besides, the data range can be determined slice by slice or volume by volume, and the influences of that are not ignorable.

Both methods provide nearly aliasing-free reconstruction at four- or eightfold acceleration. However, the hallucinations appear when using eightfold acceleration, highlighted with the green color (cf. Figure 10).

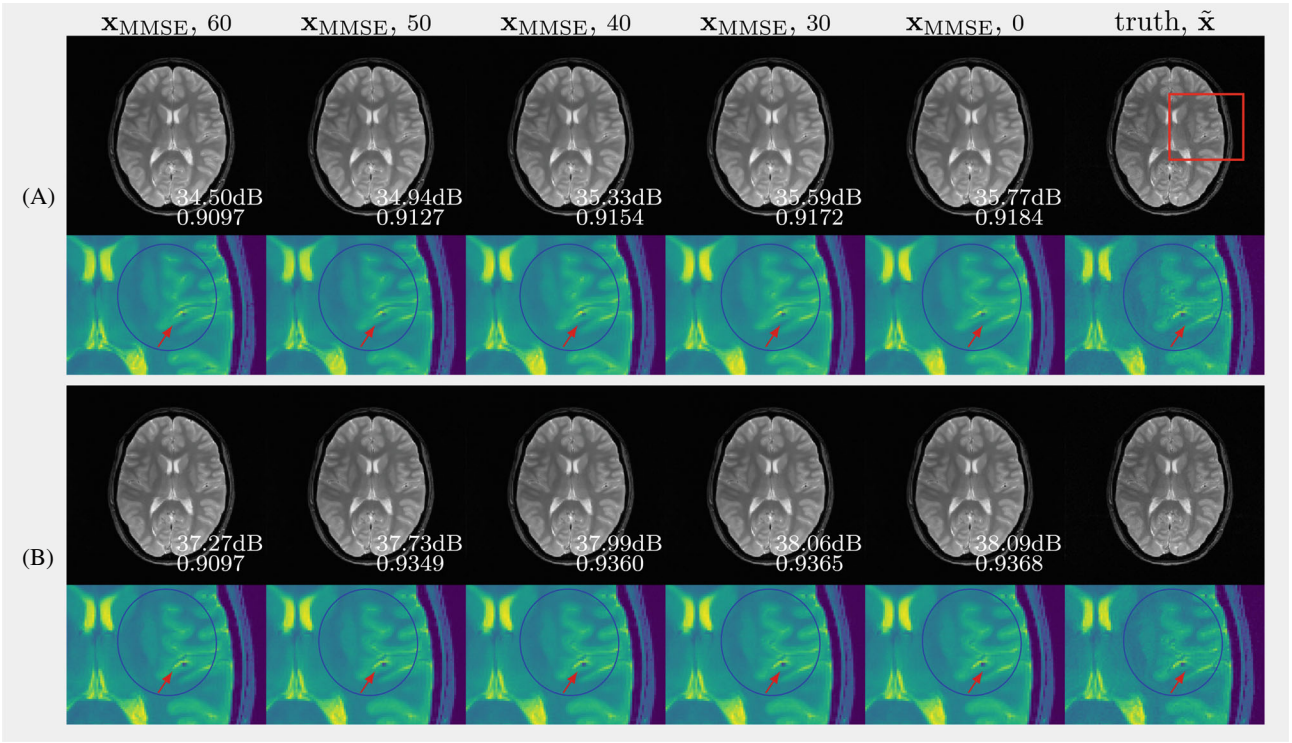

FIGURE 6 To investigate the burn-in phase the effect of splitting chains at different time points is shown for NET_2 for reconstruction with (A) 10% k-space data points and (B) 20% k-space data points.

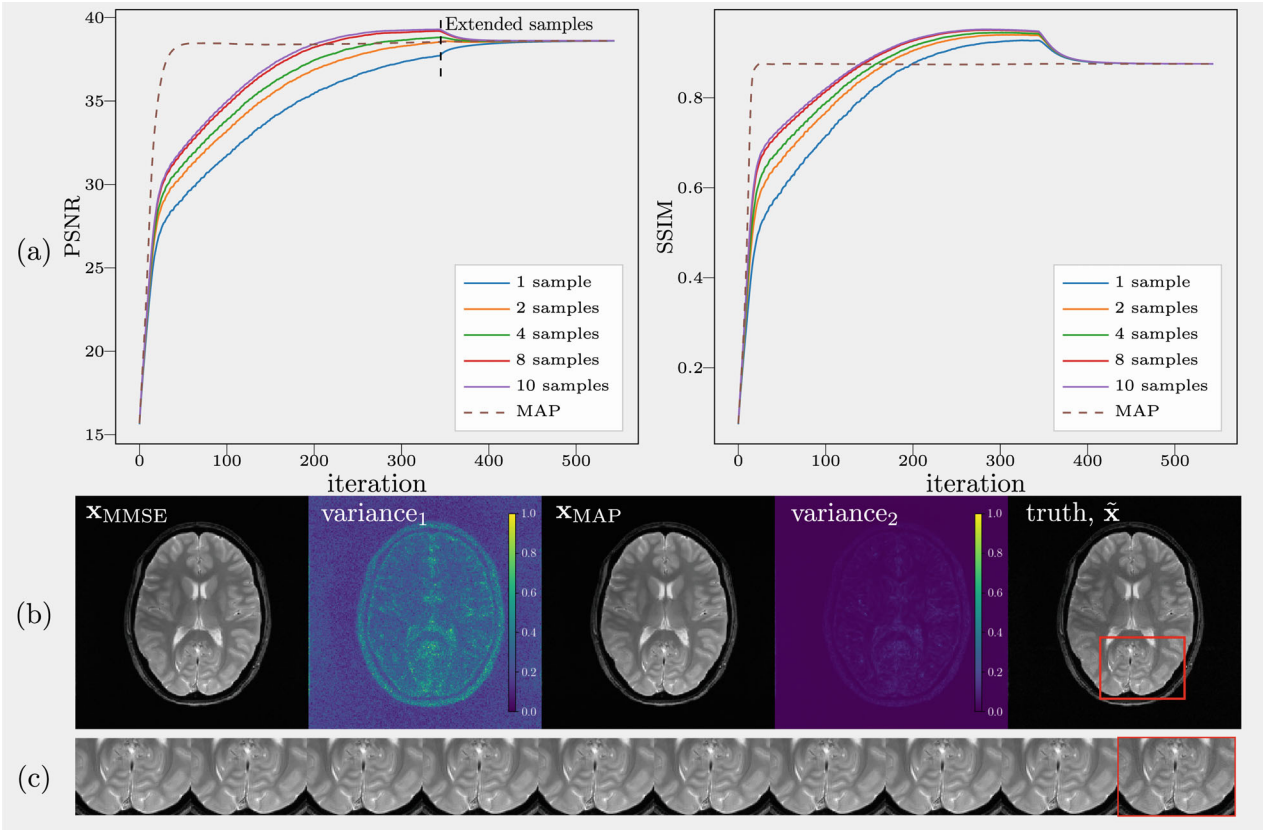

FIGURE 7 Investigation of the maximum a posterior (MAP) reconstructed with NET_2. 200 extended iterations after random exploration versa a deterministic estimate of MAP that are indicated by solid and dashed lines, respectively. (A) The curves of peak-signal-noise-ratio (PSNR) and similarity index (SSIM) over iterations. (B) The subfigure variance$_1$ and variance$_2$ were computed from unextended samples and extended samples respectively. $\mathbf{x}_{\text{MAP}}$ is an extended sample. (C) The zoom-in region of nine extended samples and the ground truth.

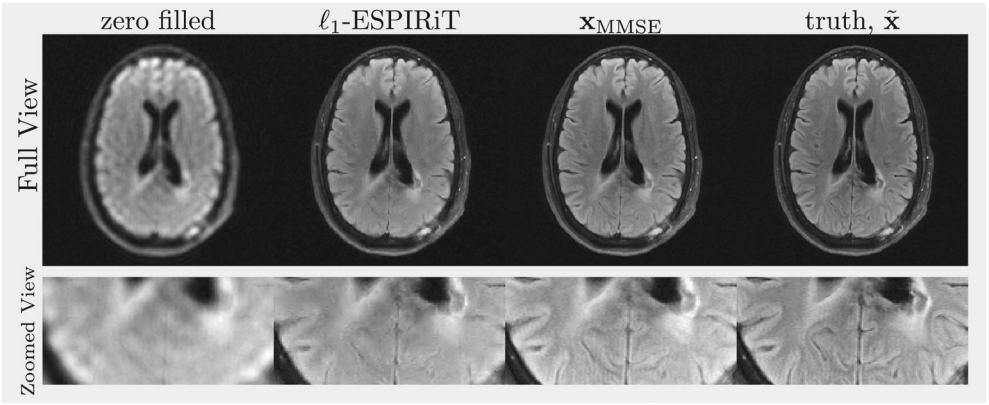

FIGURE 8 Comparison of the minimum mean square error computed with NET_3 to the ℓ_1-wavelet regularized and zero-filled reconstruction. The high resolution image (320 × 320) was reconstructed from k-space data using 10-fold undersampling. The regularization parameter was set to 0.01.

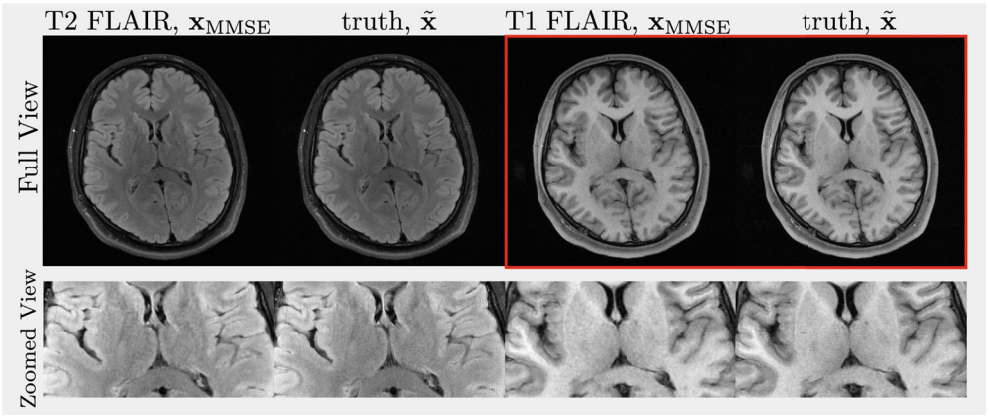

FIGURE 9 Transferability: Reconstruction of T2 and T1 fluid-attenuated inversion recovery images (FLAIR) (red box) using a Poisson-disc pattern with 8× undersampling in k-space using NET_3 trained on T2 FLAIR images.

All in all, a deep learning-based method has enough capability to generate a realistic-looking image even when the problem is highly underdetermined as a result of undersampling, but the uncertainties inside it cannot be ignored.

5 | DISCUSSION

Generally, the Bayesian statistical approach provides a foundation for sampling the posterior $p(\mathbf{x}|\mathbf{y})$ and a natural mechanism for incorporating the prior knowledge that is learned from images. The generative model is used to construct Markov chains to sample the posterior. The utilization of probabilistic generative models allows: (1) flexibility for changing the forward model of measurement; (2) exact sampling from the posterior term $p(\mathbf{x}|\mathbf{y})$; and (3) the estimation of uncertainty due to limited k-space data points.

5.1 | Uncertainties of reconstruction

One advantage of the proposed approach over classical deterministic regularization methods is that it allows the quantification of uncertainties of the reconstruction with the variance map. That requires MCMC sampling technique. The loss of spatial information of coils leads to the failure of unfolding, as demonstrated in Section 4.1. High undersampling implies a high uncertainty about the solution, which may lead to hallucinations as observed in Reference 42 and Figure 10. The regions with aliasing correspond to the high variance areas of the uncertainty map. With multiple coils, the reduction of high frequency data points in k-space leads to the loss of fine details, as demonstrated in Section 4.2. The $\mathbf{x}_{\mathrm{MMSE}}$ represents the reconstruction with minimum mean square error and the variance map evaluates the confidence interval of $\mathbf{x}_{\mathrm{MMSE}}$. Furthermore, it is possible to derive error bounds from the variance of the posterior as reported Reference 45.

5.2 | Overfitting and distortion

The proposed algorithm is an iterative refining procedure that starts from generating coarse samples with rich variations under large noise, before converging to fine samples with less variations under small noise. For early iterations of the algorithm, each parameter update mimics stochastic

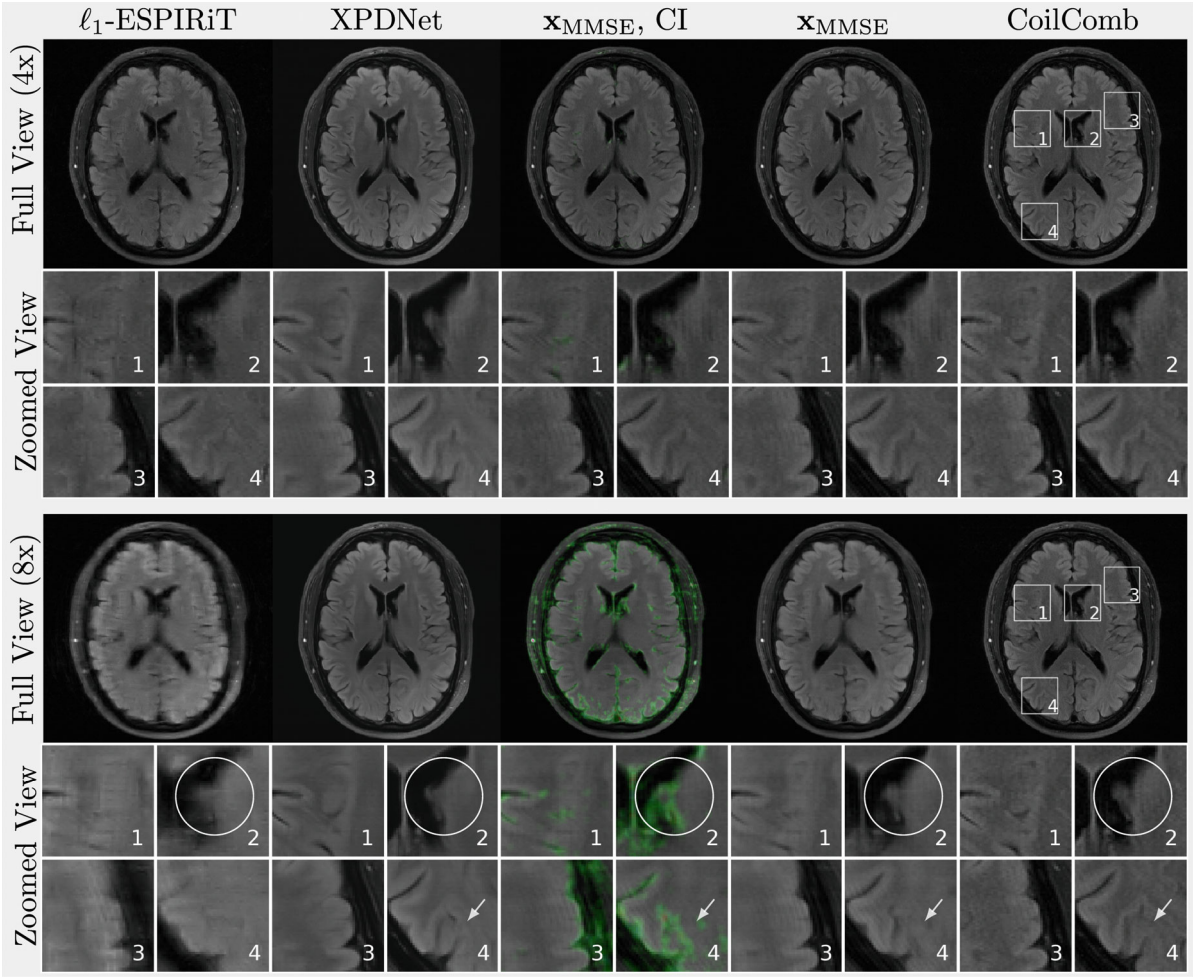

FIGURE 10 Comparison to fastMRI challenge. From the leftmost to rightmost column, reconstructions are ℓ_1-ESPIRiT, XPDNet, $\mathbf{x}_{\text{MMSE}}$ highlighted with confidence interval, $\mathbf{x}_{\text{MMSE}}$ and a fully sampled coil-combined image (CoilComb). Hallucinations appear when using eightfold acceleration along the phase-encoding direction (horizontal) and are highlighted with the confidence interval after thresholding. Selected regions of interests are presented in a zoomed view.

gradient descent; however, as the algorithm approaches a local minimum, the gradient shrinks and the chain produces the samples from the posterior. Lastly, we noticed that the balance between the learned transition and the data consistency plays an important role generally in the generation of realistic samples; here we refer readers to Figure S4. The larger λ, the stronger the consistency of data. Besides, we found that a large value of K is required for using the discrete noise conditional score network in Algorithm 1 while a smaller value is sufficient for the continuous noise scales. While the N in Algorithm 1 is larger for the continuous case, the total number of iterations in both cases is comparable.

5.3 | Computational burden

The promising performance of this method comes at the price of demanding computation. It takes around 10 min to reproduce the results in Figure 5 while ℓ_1-ESPIRiT takes about 5 s with BART for a single slice. The possible solutions to the computation burden are to: (1) accelerate the inferencing of neural networks; (2) parallelize the sampling process when multiple chains are used; and (3) reduce the number of iterations using more efficient MCMC sampling techniques. Furthermore, reducing the scale of networks is also viable. The introduction of burn-in experiment in Section 4.4 is a direct way to overcome this shortcoming when the undersampling factor is moderate.

5.4 | Relationship to generative models

To our knowledge, the construction of image models to exploit prior knowledge was first introduced in Reference 46 in which the handcrafted model which extracts edge information was used for image restoration. Following that

framework, the learned generic image priors from generative perspective are investigated in References 25,47,48, which permits more expressive modeling. In the medical imaging field, image priors learned with variational autoencoder[10,13] and PixelCNN[11,15] were applied to MRI image reconstruction. As a comparison to the method in Reference 11, the result is presented in Figure S5. Compared with some unrolled network based deep learning image reconstruction methods, the application of image priors is independent of k-space data and coil sensitivities, which permits a more versatile use of the method using different k-space acquisition strategies.

5.5 | Limitations

PSNR and SSIM only give a partial and distorted view of image quality. The influence of the ground truth and noise properties of the background have a severe influence, as does the selected data range used for computing the metrics. Thus, rating of image quality by human readers would be an important next step in the evaluation of the technique. Also the clinical usefulness of the uncertainty maps requires further investigations. To facilitate the use in clinical studies, we implemented the sampling in the BART toolbox.[49,50]

6 | CONCLUSION

The proposed reconstruction method combines concepts from machine learning, Bayesian inference and image reconstruction. In the setting of Bayesian inference, the image reconstruction is realized by drawing samples from the posterior term $p(\mathbf{x}|\mathbf{y})$ using data-driven Markov chains, providing a minimum mean square reconstruction and uncertainty estimation. The prior information can be learned from an existing image database, where the generic generative priors based on the diffusion process allow for flexibility regarding contrast, coil sensitivities, and sampling pattern.

ACKNOWLEDGMENTS

We acknowledge funding by the "Niedersächsisches Vorab" funding line of the Volkswagen Foundation. We would like to thank Xiaoqing Wang for his help in preparing this manuscript as well as Christian Holme for help with our computer systems. Open Access funding enabled and organized by Projekt DEAL.

ENDNOTES

*https://github.com/mrirecon/spreco
†https://huggingface.co/zaccharieramzi

ORCID

Guanxiong Luo https://orcid.org/0000-0001-8005-4639
Moritz Blumenthal https://orcid.org/0000-0002-2127-8365
Martin Heide https://orcid.org/0000-0002-4129-7395
Martin Uecker https://orcid.org/0000-0002-8850-809X

REFERENCES

1. Pruessmann KP, Weiger M, Boernert P, Boesiger P. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn Reson Med*. 2001;46:638-651.
2. Lustig M, Donoho D, Pauly JM. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn Reson Med*. 2007;58:1182-1195.
3. Block KT, Uecker M, Frahm J. Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint. *Magn Reson Med*. 2007;57:1086-1098.
4. Ravishankar S, Bresler Y. MR image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Trans Med Imaging*. 2011;30:1028-1041.
5. Qu X, Hou Y, Lam F, Guo D, Zhong J, Chen Z. Magnetic resonance image reconstruction from undersampled measurements using a patch-based nonlocal operator. *Med Image Anal*. 2014;18:843-856.
6. Wang S, Su Z, Ying L, et al. Accelerating magnetic resonance imaging via deep learning. Paper presented at: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016:514-517.
7. Yang Y, Sun J, Li H, Xu Z. Deep ADMM-Net for compressive sensing MRI. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2016.
8. Aggarwal HK, Mani MP, Jacob M. MoDL: model-based deep learning architecture for inverse problems. *IEEE Trans Med Imaging*. 2019;38:394-405.
9. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn Reson Med*. 2017;79:3055-3071.
10. Tezcan Kerem C, Baumgartner Christian F, Luechinger R, Pruessmann KP, Konukoglu E. MR image reconstruction using deep density priors. *IEEE Trans Med Imaging*. 2019;38:1633-1642.
11. Luo G, Zhao N, Jiang W, Hui ES, Cao P. MRI reconstruction using deep Bayesian estimation. *Magn Reson Med*. 2020;84:2246-2261.
12. Liu Q, Yang Q, Cheng H, Wang S, Zhang M, Liang D. Highly undersampled magnetic resonance imaging reconstruction using autoencoding priors. *Magn Reson Med*. 2020;83:322-336.
13. Kingma DP, Welling M. Auto-encoding variational bayes. Paper presented at: 2nd International Conference on Learning Representations, ICLR 2014. Conference Track Proceedings; April 14-16, 2014; Banff, Canada.
14. Alain G, Bengio Y. What regularized auto-encoders learn from the data-generating distribution. *J Mach Learn Res*. 2014;15:3563-3593.
15. Salimans T, Karpathy A, Chen X, Kingma DP. PixelCNN++: improving the PixelCNN with discretized logistic mixture likelihood and other modifications. Paper presented at: 5th International Conference on Learning Representations, ICLR

2017, Conference Track Proceedings. OpenReview.net; Toulon, France, April 24-26, 2017.
16. Mardani M, Gong E, Cheng JY, et al. Deep generative adversarial neural networks for compressive sensing MRI. *IEEE Trans Med Imaging*. 2018;38:167-179.
17. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural network. *International Conference on Machine Learning*. PMLR; 2015:1613-1622.
18. Narnhofer D, Effland A, Kobler E, Hammernik K, Knoll F, Pock T. Bayesian uncertainty estimation of learned variational MRI reconstruction. *IEEE Trans Med Imaging*. 2022;41:279-291.
19. Daniela C, Erkki S. Hypermodels in the Bayesian imaging framework. *Inverse Probl*. 2008;24:034013.
20. Stuart AM. Inverse problems: a Bayesian perspective. *Acta Num*. 2010;19:451-559.
21. Jalal A, Arvinte M, Daras G, Price E, Dimakis AG, Tamir J. Robust compressed sensing MRI with deep generative priors. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman VJ, eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc; 2021:14938-14954.
22. Luo G, Heide M, Uecker M. Using data-driven Markov chains for MRI reconstruction with joint uncertainty estimation. Paper presented at: Proceedings of International Society of Magnetic Resonance in Medicine; 2022; London, UK:0298.
23. Chung H, Ye JC. Score-based diffusion models for accelerated MRI. *Med Image Anal*. 2022;80:102479.
24. Levac B, Jalal A, Tamir JI. Accelerated motion correction for MRI using score-based generative models. *arXiv:2211.00199*. 2022.
25. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. Paper presented at: International Conference on Learning Representations; 2021.
26. Hyvärinen A. Estimation of non-normalized statistical models by score matching. *J Mach Learn Res*. 2005;6:695-709.
27. Pascal V. A connection between score matching and denoising autoencoders. *Neural Comput*. 2011;23:1661-1674.
28. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: Francis B, David B, eds. *Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research*. Vol 37. PMLR; 2015:2256-2265.
29. Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. In: Wallach HM., Larochelle H, Beygelzimer A, D'Alché-Buc F, Fox EB, Garnett R, eds. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*; *December 8-14, 2019*; Vancouver, BC:11895-11907.
30. Zbontar J, Knoll F, Sriram A, et al. fastMRI: an open dataset and benchmarks for accelerated MRI. *arXiv*. 2019.
31. Ramzi Z, Ciuciu P, Starck J-L. XPDNet for MRI reconstruction: an application to the 2020 fastMRI challenge. *arXiv*. 2020.
32. Särkkä S, Solin A. *Applied Stochastic Differential Equations*. Cambridge University Press; 2019.
33. Douc R, Moulines E, Priouret P, Soulier P. *Markov Chains*. Springer; 2018 ch.2:38-41.
34. Huang X, Belongie SJ. Arbitrary style transfer in real-time with adaptive instance normalization. Paper presented at: IEEE International Conference on Computer Vision, ICCV 2017; IEEE Computer Society; October 22-29, 2017; Venice, Italy:1510-1519.
35. Rahimi A, Recht B. Random features for large-scale kernel machines. In: Platt J, Koller D, Singer Y, Roweis S, eds. *Advances in Neural Information Processing Systems NIPS'07*. Vol 20. Curran Associates, Inc; 2008:1177-1184.
36. Lin G, Milan A, Shen C, Reid ID. RefineNet: multi-path refinement networks for high-resolution semantic segmentation. Paper presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017; IEEE Computer Society; July 21-26, 2017; Honolulu, HI:5168-5177.
37. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. Paper presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT; 2018:7794-7803.
38. Uecker M, Lai P, Murphy MJ, et al. ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. *Magn Reson Med*. 2014;71:990-1001.
39. Uecker M, Rosenzweig S, Holme H. Christian M., et al. *mrirecon/bart: version 0.6.00*. 2020.
40. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation OSDI'16*. USENIX Association; 2016:265-283.
41. Harris CR, Millman K, Jarrod WSJ, et al. Array programming with NumPy. *Nature*. 2020;585:357-362.
42. Muckley MJ, Riemenschneider B, Radmanesh A, et al. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. *IEEE Trans Med Imag*. 2021;40:1.
43. Shimron E, Tamir JI, Wang K, Lustig M. Implicit data crimes: machine learning bias arising from misuse of public data. *Proc Natl Acad Sci*. 2022;119:e2117203119.
44. Arvinte M, Tamir J. The truth matters: a brief discussion on MVUE vs. RSS in MRI reconstruction. Paper presented at: European Society for Magnetic Resonance in Medicine and Biology; Virtual Conference, Barcelona, Spain; 2021.
45. Narnhofer D, Habring A, Holler M, Pock T. Posterior-variance-based error quantification for inverse problems in imaging. *arXiv:2212.12499*. 2022.
46. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*. 1984;PAMI-6:721-741.
47. Roth S, Black MJ. Fields of experts: a framework for learning image priors. Paper presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA; 2005:860-867.
48. Schmidt U, Gao Q, Roth S. A generative perspective on MRFs in low-level vision. Paper presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA; 2010:1751-1758.
49. Luo G, Blumenthal M, Uecker M. Using data-driven image priors for image reconstruction with BART. Paper presented at: Proceedings of International Society of Magnetic Resonance in Medicine; Virtual Conference; 2021:3768.
50. Blumenthal M, Luo G, Schilling M, Holme H, Christian M, Uecker M. Deep, deep learning with BART. *Magn Reson Med*. 2023;89:678-693.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

FIGURE S1: Samples and x_{MMSE} from intermediate distributions are presented here. Each x_{MMSE} is the average over 10 samples.
FIGURE S2: Overview of RefineNet and refine blocks
FIGURE S3: PSNR and SSIM metrics for different ground truths and data ranges.
FIGURE S4: Samples reconstructed with different λ. Two selected samples with a particular λ are presented in (A) and (B). The variance maps over 10 samples reconstructed with each λ are shown in (C).
FIGURE S5: Reconstruction using different prior-based methods with Poisson-disc sampling with 10x undersampling in k-space.
TABLE S1: Architectures of the score networks.
TABLE S2: Hyperparameters for training
TABLE S3: Average PSNR (dB) and SSIM (%) for test subjects

How to cite this article: Luo G, Blumenthal M, Heide M, Uecker M. Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models. *Magn Reson Med*. 2023;90:295-311. doi: 10.1002/mrm.29624

APPENDIX A. REWRITE IN TERMS OF POSTERIOR

Because the forward diffusion is a Markov process and start at $\mathbf{x}_0$, with Bayes' rule we have

$$q(\mathbf{x}_i|\mathbf{x}_{i-1},\mathbf{x}_0) = q(\mathbf{x}_{i-1}|\mathbf{x}_i)\frac{q(\mathbf{x}_i|\mathbf{x}_0)}{q(\mathbf{x}_{i-1}|\mathbf{x}_0)}. \tag{A1}$$

Substituting density function into Equation (A1) yields

$$q(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0) = q(\mathbf{x}_i|\mathbf{x}_{i-1})\cdot\frac{q(\mathbf{x}_{i-1}|\mathbf{x}_0)}{q(\mathbf{x}_i|\mathbf{x}_0)}. \tag{A2}$$

$$= \frac{1}{\sqrt{(2\pi\beta_i^2)^{N_p}}}\cdot\frac{b_i^{2N_p}}{b_{i-1}^{2N_p}}\exp\left[-\left(\frac{\|\mathbf{x}_i-\mathbf{x}_{i-1}\|^2}{\beta_i^2} + \frac{\|\mathbf{x}_{i-1}-\mathbf{x}_0\|^2}{b_{i-1}^2} - \frac{\|\mathbf{x}_i-\mathbf{x}_0\|^2}{b_i^2}\right)\right]. \tag{A3}$$

Let $\frac{b_{i-1}^2+\beta_i^2}{b_i^2} = 1$, which is satisfied with Equation (5), we have

$$q(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0) = \frac{1}{\sqrt{(2\pi\beta_i^2)^{N_p}}}\cdot\frac{b_i^{2N_p}}{b_{i-1}^{2N_p}}\exp\left[-\left(\frac{\|\mathbf{x}_i-\mathbf{x}_{i-1}\|^2}{\beta_i^2} + \frac{\|\mathbf{x}_{i-1}-\mathbf{x}_0\|^2}{b_{i-1}^2} - \frac{\|\mathbf{x}_i-\mathbf{x}_0\|^2}{b_i^2}\right)\right], \tag{A4}$$

$$= \frac{1}{\sqrt{(2\pi\beta_i^2)^{N_p}}}\cdot\frac{b_i^{2N_p}}{b_{i-1}^{2N_p}}\exp\left[-\left(\frac{\|\mathbf{x}_{i-1}-\boldsymbol{\mu}\|^2}{\beta_i^2\cdot\frac{b_{i-1}^2}{b_i^2}}\right)\right], \tag{A5}$$

where

$$\boldsymbol{\mu} = \frac{b_{i-1}^2}{b_i^2}\cdot\mathbf{x}_i + \frac{\beta_i^2}{b_i^2}\cdot\mathbf{x}_0. \tag{A6}$$

APPENDIX B. KL DIVERGENCE OF TWO GAUSSIAN DISTRIBUTIONS

Let $p(\mathbf{x}) = \mathcal{CN}(\boldsymbol{\mu}_1,\sigma_1^2\mathbf{I})$ and $q(\mathbf{x}) = \mathcal{CN}(\boldsymbol{\mu}_2,\sigma_2^2\mathbf{I})$ and the KL divergence is defined by

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(\mathbf{x})\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)d\mathbf{x}.$$

Therefore,

$$\begin{aligned} &D_{KL}(P\|Q)\\ &= \int[\log p(\mathbf{x}) - \log q(\mathbf{x})]p(\mathbf{x})d\mathbf{x}\\ &= \mathbb{E}_{p(\mathbf{x})}\left[N_p\log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{\sigma_2^2}\|\mathbf{x}-\boldsymbol{\mu}_2\|^2 - \frac{1}{\sigma_1^2}\|\mathbf{x}-\boldsymbol{\mu}_1\|^2\right]\\ &= N_p\cdot\log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{\sigma_2^2}\mathbb{E}_{p(\mathbf{x})}\left[\|\mathbf{x}-\boldsymbol{\mu}_2\|^2\right] - 1. \end{aligned}$$

where N_p is the dimensionality $n\times n\times 2$. Noting that

$$\|\mathbf{x}-\boldsymbol{\mu}_2\|^2 = \|\mathbf{x}-\boldsymbol{\mu}_1\|^2 + 2\mathbb{R}(\mathbf{x}-\boldsymbol{\mu}_1)^H(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2$$

we arrive at

$$\begin{aligned} D_{KL}(P\|Q) &= N_p\cdot\log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{\sigma_2^2}\left(\mathbb{E}_{p(\mathbf{x})}\left[\|\mathbf{x}-\boldsymbol{\mu}_1\|^2\right]\right.\\ &\quad + 2\mathbb{R}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^H\mathbb{E}_{p(\mathbf{x})}\left[\mathbf{x}-\boldsymbol{\mu}_1\right]\\ &\quad \left.+\mathbb{E}_{p(\mathbf{x})}\left[\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2\right]\right) - 1\\ &= N_p\cdot\log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2+\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{\sigma_2^2} - 1. \end{aligned}$$