Solving 3D Inverse Problems using Pre-trained 2D Diffusion Models

Hyungjin Chung[1,2]*, Dohoon Ryu[1]*, Michael T. Mccann[2], Marc L. Klasky[2], Jong Chul Ye[1]
[1]Korea Advanced Institute of Science & Technology, [2]Los Alamos National Laboratory,
{hj.chung, dh.ryu, jong.ye}@kaist.ac.kr, {mccann, mklasky}@lanl.gov

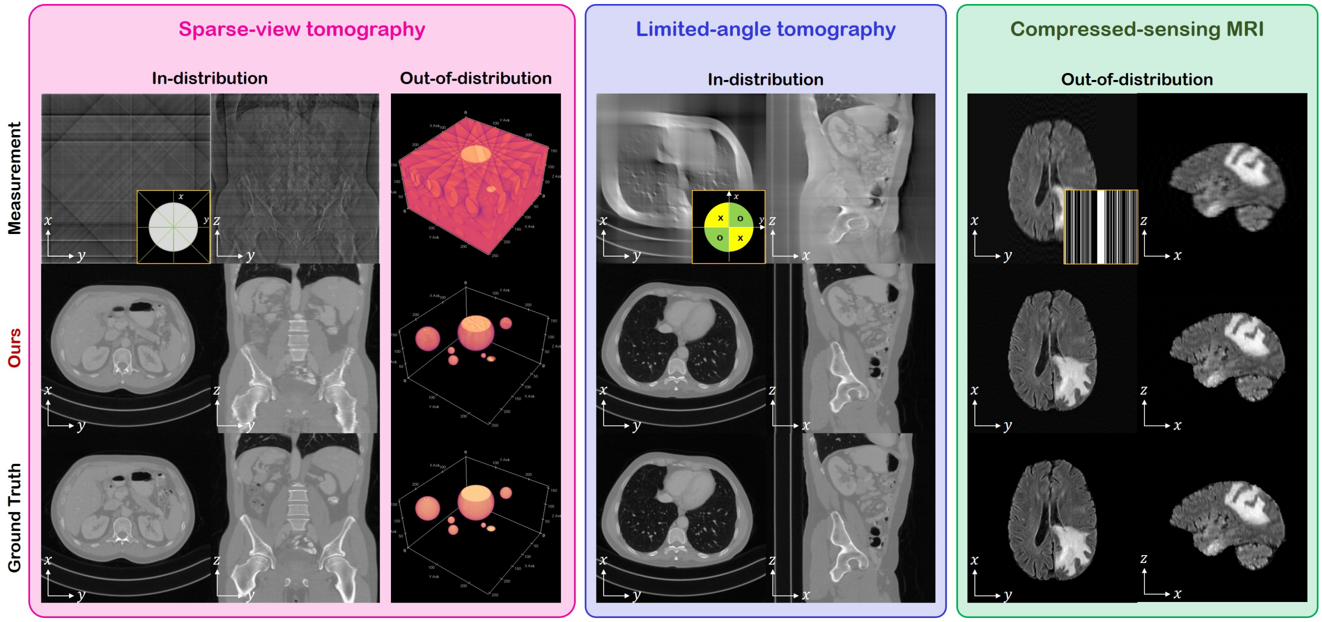

Figure 1. 3D reconstruction results with DiffusionMBIR. First row: measurement, second row: our method, third row: ground truth. Yellow inset: measurement process. Sparse-view tomography: 8-view measurement, Limited-angle tomography: [0 90]° out of [0 180]° angle measurement, Compressed-sensing MRI: 1D uniform sub-sampling of ×2 acceleration. (In-distribution): test data aligned with training data, (Out-of-distribution): test data vastly different from training data.

Abstract

Diffusion models have emerged as the new state-of-the-art generative model with high quality samples, with intriguing properties such as mode coverage and high flexibility. They have also been shown to be effective inverse problem solvers, acting as the prior of the distribution, while the information of the forward model can be granted at the sampling stage. Nonetheless, as the generative process remains in the same high dimensional (i.e. identical to data dimension) space, the models have not been extended to 3D inverse problems due to the extremely high memory and computational cost. In this paper, we combine the ideas from the conventional model-based iterative reconstruction with the modern diffusion models, which leads to a highly effective method for solving 3D medical image reconstruction tasks such as sparse-view tomography, limited angle tomography, compressed sensing MRI from pre-trained 2D diffusion models. In essence, we propose to augment the 2D diffusion prior with a model-based prior in the remaining direction at test time, such that one can achieve coherent reconstructions across all dimensions. Our method can be run in a single commodity GPU, and establishes the new state-of-the-art, showing that the proposed method can perform reconstructions

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711137899, KMDF_PR_20200901_0015), by the National Research Foundation of Korea under Grant NRF-2020R1A2B5B03001980, by the KAIST Key Research Institute (Interdisciplinary Research Group) Project, and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2020-0-01461) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

of high fidelity and accuracy even in the most extreme cases (e.g. 2-view 3D tomography). We further reveal that the generalization capacity of the proposed method is surprisingly high, and can be used to reconstruct volumes that are entirely different from the training dataset. Code available: https://github.com/HJ-harry/DiffusionMBIR

1. Introduction

Diffusion models learn the data distribution implicitly by learning the gradient of the log density (i.e. $\nabla_{\boldsymbol{x}} \log p_{\text{data}}(\boldsymbol{x})$; score function) [9, 32], which is used at inference to create generative samples. These models are known to generate high-quality samples, cover the modes well, and be highly robust to train, as it amounts to merely minimizing a mean squared error loss on a denoising problem. Particularly, diffusion models are known to be much more robust than other popular generative models [8], for example, generative adversarial networks (GANs). Furthermore, one can use pre-trained diffusion models to solve inverse problems in an unsupervised fashion [5–7, 15, 32]. Such strategies has shown to be highly effective in many cases, often establishing the new state-of-the-art on each task. Specifically, applications to sparse view computed tomography (SV-CT) [5, 31], compressed sensing MRI (CS-MRI) [6, 7, 31], super-resolution [4, 6, 15], inpainting [5, 15] among many others, have been proposed.

Nevertheless, to the best of our knowledge, all the methods considered so far focused on 2D imaging situations. This is mostly due to the high-dimensional nature of the generative constraint. Specifically, diffusion models generate samples by starting from pure noise, and iteratively denoising the data until reaching the clean image. Consequently, the generative process involves staying in the *same* dimension as the data, which is prohibitive when one tries to scale the data dimension to 3D. One should also note that training a 3D diffusion model amounts to learning the 3D prior of the data density. This is undesirable in two aspects. First, the model is data hungry, and hence training a 3D model would typically require thousands of *volumes*, compared to 2D models that could be trained with less than 10 volumes. Second, the prior would be needlessly complicated: when it comes to dynamic imaging or 3D imaging, exploiting the spatial/temporal correlation [12, 33] is standard practice. Naively modeling the problem as 3D would miss the chance of leveraging such information.

Another much more well-established method for solving 3D inverse problems is model-based iterative reconstruction (MBIR) [14, 17], where the problem is formulated as an optimization problem of weighted least squares (WLS), constructed with the data consistency term, and the regularization term. One of the most widely acknowledged regularization in the field is the total variation (TV) penalty [18, 28], known for its intriguing properties: edge-preserving, while imposing smoothness. While the TV prior has been widely explored, it is known to fall behind the data-driven prior of the modern machine learning practice, as the function is too simplistic to fully model how the image “looks like”.

In this work, we propose DiffusionMBIR, a method to combine the best of both worlds: we incorporate the MBIR optimization strategy into the diffusion sampling steps in order to *augment* the data-driven prior with the conventional TV prior, imposed to the z-direction only. Particularly, the standard reverse diffusion (i.e. denoising) steps are run independently with respect to the z-axis, and hence standard 2D diffusion models can be used. Subsequently, the data consistency step is imposed by aggregating the slices, then taking a single update step of the alternating direction method of multipliers (ADMM) [3]. This step effectively coerces the cross-talk between the slices with the measurement information, and the TV prior. For efficient optimization, we further propose a strategy in which we call *variable sharing*, which enables us to only use a *single* sweep of ADMM and conjugate gradient (CG) per denoising iteration. Note that our method is fully general in that we are not restricted to the given forward operator at test time. Hence, we verify the efficacy of the method by performing extensive experiments on sparse-view CT (SV-CT), limited angle CT (LA-CT), and compressed sensing MRI (CS-MRI): out method shows consistent improvements over the current diffusion model-based inverse problem solvers, and shows strong performance on *all* tasks (For representative results, see Fig. 1. For conceptual illustration of the inverse problems, see Fig. 2).

In short, the main contributions of this paper is to devise a diffusion model-based reconstruction method that 1) operate with the voxel representation, 2) is memory-efficient such that we can scale our solver to much higher dimensions (i.e. $> 256^3$), and 3) is not data hungry, such that it can be trained with less than ten 3D volumes.

2. Background

Model-based iterative reconstruction (MBIR). Consider a linear forward model for an imaging system (e.g. CT, MRI)

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ is the measurement (i.e. sinogram, k-space), $\boldsymbol{x} \in \mathbb{R}^n$ is the image that we wish to reconstruct, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the discrete transform matrix (i.e. Radon, Fourier[1]), and $\boldsymbol{n}$ is the measurement noise in the system. As the problem is ill-posed, a standard approach for the inverse problem that estimates the unknown image $\boldsymbol{x}$ from

[1] While we denote real-valued transforms and measurements for the simplicity of exposition, the discrete Fourier transform (DFT) matrix, and the corresponding measurement are complex-valued.

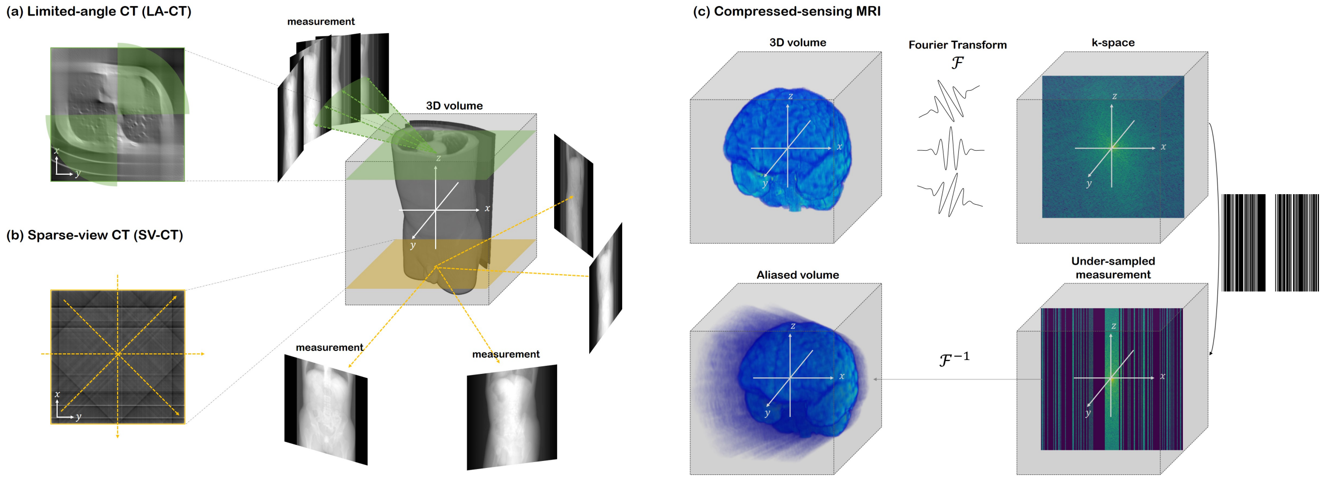

Figure 2. Visualization of the measurement process for the three tasks we tackle in this work: (a) Limited angle CT (LA-CT)—measurement model of Fig. 4, (b) sparse view CT (SV-CT)—measurement model of Fig. 3,6,7, (c) compressed sensing MRI (CS-MRI)—measurement model of Fig. 8.

the measurement $\boldsymbol{y}$ is to perform the following regularized reconstruction:

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + R(\boldsymbol{x}), \tag{2}$$

where R is the suitable regularization for $\boldsymbol{x}$, for instance, sparsity in some transformed domain. One widely used function is the TV penalty, $R(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_{2,1}$, where $\boldsymbol{D} := [\boldsymbol{D}_x, \boldsymbol{D}_y, \boldsymbol{D}_z]^T$ computes the finite difference in each axis. Minimization of (2) can be performed with robust optimization algorithms, such as fast iterative soft thresholding algorithm (FISTA) [2] or ADMM.

Score-based diffusion models. A score-based diffusion model is a generative model that defines the generative process as the *reverse* of the data *noising* process. Specifically, consider the stochastic process $\{\boldsymbol{x}(t) \triangleq \boldsymbol{x}_t\}, t \in [0,1]$, where we introduce the *time* variable t to represent the evolution of the random variable. Particularly, we define $p(\boldsymbol{x}_0) \triangleq p_{\text{data}}(\boldsymbol{x})$, i.e. the data distribution, and $p(\boldsymbol{x}_T)$ to approximately a Gaussian distribution. The evolution can be formalized with the following stochastic differential equation

$$d\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)\,dt + g(t)\,d\boldsymbol{w}, \tag{3}$$

where $\boldsymbol{f}(\boldsymbol{x}, t) : \mathbb{R}^{n+1} \mapsto \mathbb{R}^n$ is the drift function, $g(t) : \mathbb{R} \mapsto \mathbb{R}$ is the scalar diffusion function, and $\boldsymbol{w}$ is the $n-$dimensional standard Brownian motion [27]. Let $\boldsymbol{f}(\boldsymbol{x}, t) = 0, g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$. Then, the SDE simplifies to the following Brownian motion

$$d\boldsymbol{x} = \sqrt{\frac{d[\sigma^2(t)]}{dt}}\,d\boldsymbol{w}, \tag{4}$$

in which the mean remains the same across the evolution, while Gaussian noise will be continuously added to $\boldsymbol{x}$, eventually approaching pure Gaussian noise as the noise term dominates. This is the so called variance-exploding SDE (VE-SDE) in the literature [32], and as we construct all our methods on the VE-SDE, we derive what follows from (4). Directly applying Anderson's theorem [1, 32] leads to the following reverse SDE

$$d\boldsymbol{x} = -\frac{d[\sigma^2(t)]}{dt}\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)\,dt + \sqrt{\frac{d[\sigma^2(t)]}{dt}}\,d\bar{\boldsymbol{w}}, \tag{5}$$

where $dt, d\bar{\boldsymbol{w}}$ are the reverse time differential, and the reverse standard $n-$dimensional Brownian motion. (5) defines the *generative* process of the diffusion model, where the equation can be solved by numerical integration. Notably, the key workhorse in the integration step is the score function $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$, that can be trained with denoising score matching (DSM) [34]

$$\min_{\theta} \mathbb{E}_{t,\boldsymbol{x}(t)}\left[\lambda(t)\|\boldsymbol{s}_\theta(\boldsymbol{x}(t), t) - \nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}(t)|\boldsymbol{x}(0))\|_2^2\right], \tag{6}$$

where $\boldsymbol{s}_\theta(\boldsymbol{x}(t), t) : \mathbb{R}^{n\times 1} \mapsto \mathbb{R}^n$ is a time-dependent neural network, and $\lambda(t)$ is the weighting scheme. Since $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}(t)|\boldsymbol{x}(0))$ is simply the *residual noise* added to $\boldsymbol{x}(t)$ from $\boldsymbol{x}(0)$ scaled with noise variance, optimizing for (6) amounts to training a residual denoiser across multiple noise scales - a fairly robust training scheme. While the training is robust, the equivalence between DSM and explicit score matching (ESM) can be established [34] in the optimization sense, and hence $s_{\theta^*}(\boldsymbol{x}(t), t) \simeq \nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$ can be used as a plug-in approximate in practice, i.e.

$$d\boldsymbol{x} \simeq -\frac{d[\sigma^2(t)]}{dt}\boldsymbol{s}_{\theta^*}(\boldsymbol{x}(t), t)\,dt + \sqrt{\frac{d[\sigma^2(t)]}{dt}}\,d\bar{\boldsymbol{w}}. \tag{7}$$

One can solve (7) with, e.g. the predictor-corrector (PC) sampler [32] by discretization of the time interval $[0, 1]$ to N bins.

3D diffusion. The generative process (i.e. reverse diffusion), explicitly represented by (7), runs in the full data dimension $\mathbb{R}^n$. It is widely known that the voxel representation of 3D data is heavy, and scaling the data size over 64^3 requires excessive GPU memory [19]. For example, a recent work that utilizes diffusion models for 3D shape reconstruction [35] uses the dataset of size 64^3. Other works that use diffusion models for 3D generative modeling typically focus on the more efficient point cloud representation [20, 21, 40], where the number of point clouds remain less than a few thousand (e.g. $2048 \ll 64^3$ in [20, 40]). Naturally, point cloud representations are efficient but extremely sparse, certainly not suitable for the problem of tomographic reconstruction, where we require accurate estimation of the interior.

There is one concurrent workshop paper that aims for designing a diffusion model that can model the 3D voxel representation [25]. In [25], the authors train a score function that can model $160 \times 224 \times 160$ volumes, by training a latent diffusion model [26], where the latent dimension is relatively small (i.e. $20 \times 28 \times 20$). However, the model requires 1000 synthetic 3D volumes for the training dataset. More importantly, using *latent* diffusion models for solving inverse problems is not straightforward, and has never been reported in literature.

Solving inverse problems with diffusion. Solving the reverse SDE with the approximated score function (7) amounts to sampling from the prior distribution $p(\boldsymbol{x})$. For the case of solving inverse problem, we desire to sample from the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$, where the relationship between the two can be formulated by the Bayes' rule $p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})/p(\boldsymbol{y})$, leading to

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|\boldsymbol{y}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}). \tag{8}$$

Here, the likelihood term enforces the data consistency, and thereby inducing samples that satisfy $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. Two approaches to incorporating (8) exist in the literature. First, one can split the update step into the 1) prior update (i.e. denoising), and then 2) projection in to the measurement subspace [6, 32]. Formally, in the discrete setting,

$$\boldsymbol{x}'_{i-1} \leftarrow \mathrm{Solve}(\boldsymbol{x}_{i-1}, \boldsymbol{s}_{\theta^*}), \tag{9}$$

$$\boldsymbol{x}_i \leftarrow \mathcal{P}_{\{\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}\}}(\boldsymbol{x}'_{i-1}), \tag{10}$$

where Solve denotes a general numerical solver that can solve the reverse-SDE in (7), and $\mathcal{P}_C$ denotes the projection operator to the set C. Specifically, when using the Euler-Maruyama discretization, the equation reads[2]

$$\boldsymbol{x}'_{i-1} \leftarrow (\sigma_i^2 - \sigma_{i-1}^2)\boldsymbol{s}_{\theta^*}(\boldsymbol{x}_{i-1}, i-1) + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\boldsymbol{\epsilon}, \tag{11}$$

$$\boldsymbol{x}_i \leftarrow \mathcal{P}_{\{\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}\}}(\boldsymbol{x}'_{i-1}). \tag{12}$$

Note that the stochasticity of $\mathrm{Solve}(\cdot)$ is implicitly defined. It was shown in [32] that using the PC solver, which alternates between the numerical SDE solver and monte carlo markov chain (MCMC) steps leads to superior performance. Throughout the manuscript, we refer to a single step of PC sampler as $\mathrm{Solve}(\cdot)$ unless specified otherwise. Alternatively, one can try to explicitly approximate the gradient of the log likelihood and take the update in a single step [5].

3. DiffusionMBIR

3.1. Main idea

To efficiently utilize the diffusion models for 3-D reconstruction, one possible solution would be to apply 2-D diffusion models slice by slice. Specifically, (9),(10) could be applied *parallel* with respect to the $z-$axis. However, this approach has one fundamental limitation. When the steps are run without considering the inter-dependency between the slices, the slices that are reconstructed will not be coherent with each other (especially when we have sparser view angles). Consequently, when viewed from the coronal/sagittal slice, the images contain severe artifacts.(see Fig. 3,4 (d) row 2-3).

In order to address this issue, we are interested in combining the advantages from the MBIR and the diffusion model to oppress unwanted artifacts. Specifically, our proposal is to adopt the alternating minimization approach in (9),(10), but rather than applying them in 2-D domain, the diffusion-based denoising step in (9) is applied slice-by-slice, whereas the 2-D projection step in (10) is replaced with the ADMM update step in 3-D volume. Specifically, we consider the following sub-problem

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\boldsymbol{D}_z\boldsymbol{x}\|_1, \tag{13}$$

where unlike the conventional TV algorithms that take $\|\boldsymbol{D}\boldsymbol{x}\|_1$, we only take the ℓ_1 norm of the finite difference in the $z-$axis. This choice stems from the fact that the prior with respect to the xy plane is already taken care of with the neural network $\boldsymbol{s}_{\theta^*}$, and all we need to imply is the spatial correlation with respect to the remaining direction. In other words, we are augmenting the generative prior with the model-based sparsity prior. From our experiments, we

[2] For all equations and algorithms that are presented, we refer to the sampled random Gaussian noise as $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$, unless specified otherwise.

observe that our prior augmentation strategy is highly effective in producing coherent 3D reconstructions throughout all the three axes.

3.2. Algorithmic steps

We arrive at the update steps[3]

$$\boldsymbol{x}^+ = (\boldsymbol{A}^T\boldsymbol{A} + \rho\boldsymbol{D}_z^T\boldsymbol{D}_z)^{-1}(\boldsymbol{A}^T\boldsymbol{y} + \rho\boldsymbol{D}^T(\boldsymbol{z} - \boldsymbol{w})) \tag{14}$$

$$\boldsymbol{z}^+ = \mathcal{S}_{\lambda/\rho}(\boldsymbol{D}_z\boldsymbol{x}^+ + \boldsymbol{w}) \tag{15}$$

$$\boldsymbol{w}^+ = \boldsymbol{w} + \boldsymbol{D}_z\boldsymbol{x}^+ - \boldsymbol{z}^+, \tag{16}$$

where ρ is the hyper-paremeter for the method of multipliers, and $\mathcal{S}$ is the soft thresholding operator. Moreover, (14) can be solved with conjugate gradient (CG), which efficiently finds a solution for $\boldsymbol{x}$ that satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$: we denote running K iterations of CG with initial point $\boldsymbol{x}$ as $\texttt{CG}(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{x}, K)$. Full derivation for the ADMM steps is provided in Supplementary section A. For simplicity, we denote one sweep of (14),(15),(16) as $\boldsymbol{x}^+, \boldsymbol{z}^+, \boldsymbol{w}^+ = \texttt{ADMM}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$. Iterative application of $\texttt{ADMM}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ would robustly solve the minimization problem in (13). Hence, the naive implementation of the proposed algorithm would be

$$\boldsymbol{x}'_{i-1} \leftarrow \text{Solve}(\boldsymbol{x}_i, \boldsymbol{s}_{\theta^*}), \tag{17}$$

$$\boldsymbol{x}_{i-1} \leftarrow \operatorname*{argmin}_{\boldsymbol{x}'_{i-1}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}'_{i-1}\|_2^2 + \|\boldsymbol{D}_z\boldsymbol{x}'_{i-1}\|_1. \tag{18}$$

Specifically, (17) would amount to parallel denoising for each slice, whereas (18) augments the z-directional TV prior and impose consistency. See the detailed (slow version) solving steps in Algorithm 2 of the supplementary section. Here, note that there are three sources of iteration in the algorithm: 1) Numerical integration of SDE, indexed with i, 2) ADMM iteration, and 3) the inner CG iteration, used to solve (14).

Since diffusion models are slow in itself, multiplicative additional cost of factors 2),3) will be prohibitive, and should be refrained from. In the following, we devise a simple method to reduce this cost dramatically.

Fast and efficient implementation (variable sharing) In Algorithm 2, we re-initialize the primal variable $\boldsymbol{z}$, and the dual variable $\boldsymbol{w}$, everytime before the ADMM iteration runs for the i^{th} iteration of the SDE. In turn, this would lead to slow convergence of the ADMM algorithm, as burn-in period for the variables $\boldsymbol{z}, \boldsymbol{w}$ would be required for the first few iterations. Moreover, since solving for diffusion models would have large number of discretization steps N, the difference between the two adjacent iterations $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$ is minimal. When dropping the values of $\boldsymbol{z}, \boldsymbol{w}$ from the $i+1^{\text{th}}$ iteration and re-initializing at the i^{th} iteration, one would be dropping valuable information, and wasting compute. Hence, we propose to initialize both $\boldsymbol{z}_N, \boldsymbol{w}_N$ as a *global* variable before the start of the SDE iteration, and keep the updated values throughout. Interesting enough, we find that choosing $M = 1, K = 1$, i.e. *single* iteration for both ADMM and CG are necessary for high fidelity reconstruction. Our fast version of DiffusionMBIR is presented in Algorithm 1.

Algorithm 1 DiffusionMBIR (fast; variable sharing)

```
Require: $s_{\theta^*}, N, \lambda, \rho, \{\sigma_i\}$
$\boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{0}, \sigma_T^2\boldsymbol{I})$
$\boldsymbol{z}_N \leftarrow$ torch.zeros_like($\boldsymbol{x}_N$)
$\boldsymbol{w}_N \leftarrow$ torch.zeros_like($\boldsymbol{x}_N$)
for $i = N - 1 : 0$ do                                  ▷ SDE iteration
    $\boldsymbol{x}'_i \leftarrow \text{Solve}(\boldsymbol{x}_{i+1}, s_{\theta^*})$
    $\boldsymbol{A}_{\texttt{CG}} \leftarrow \boldsymbol{A}^T\boldsymbol{A} + \rho\boldsymbol{D}_z^T\boldsymbol{D}_z$
    $\boldsymbol{b}_{\texttt{CG}} \leftarrow \boldsymbol{A}^T\boldsymbol{y} + \rho\boldsymbol{D}_z^T(\boldsymbol{z}_{i+1} - \boldsymbol{w}_{i+1})$
    $\boldsymbol{x}_i \leftarrow \texttt{CG}(\boldsymbol{A}_{\texttt{CG}}, \boldsymbol{b}_{\texttt{CG}}, \boldsymbol{x}'_i, 1)$
    $\boldsymbol{z}_i \leftarrow \mathcal{S}_{\lambda/\rho}(\boldsymbol{D}_z\boldsymbol{x}_i + \boldsymbol{w}_{i+1})$
    $\boldsymbol{w}_i \leftarrow \boldsymbol{w}_{i+1} + \boldsymbol{D}_z\boldsymbol{x}_i - \boldsymbol{z}_i$
end for
return $\boldsymbol{x}_0$
```

Another caveat is when running the neural network forward pass through the entire volume is not feasible memory-wise, for example, when fitting the solver into a single commodity GPU. One can circumvent this by dividing the batch dimension[4] into sub-batches, running the denoising step for the sub-patches separately, and then aggregating them into the full volume again. The $\texttt{ADMM}$ step can be applied to the full volume after the aggregation, which would yield the same solution with Algorithm 1. For both the slow and the fast version of the algorithm, one can also apply a projection to the measurement subspace at the end when one wishes to exactly match the measurement constraint.

4. Experimental setup

We conduct experiments on three most widely studied tasks in medical image reconstruction: 1) sparse view CT (SV-CT), 2) limited angle CT (LA-CT), and 3) compressed sensing MRI (CS-MRI). Specific details can be found in the supplementary section B.

Dataset. For both CT reconstruction tasks (i.e. SV-CT, LA-CT) we use the data from the AAPM 2016 CT low-dose grand challenge. All volumes except for one are used for training the 2D score function, and one volume is held-out

[3] Detailed derivation of the following optimization steps can be found in Supplementary section A.

[4] In our implementation, the batch dimension corresponds to the $z-$axis, as 2D slices are stacked.

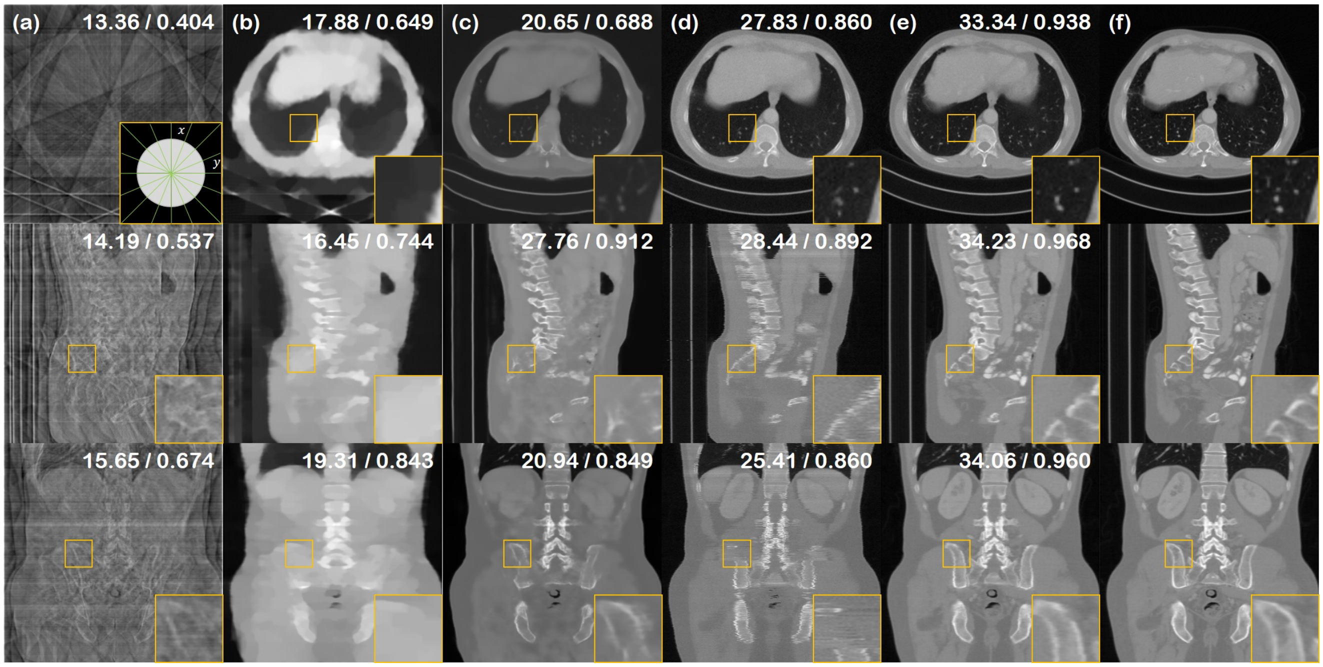

Figure 3. 8-view SV-CT reconstruction results of the test data (First row: axial slice, second row: sagittal slice, third row: coronal slice). (a) FBP, (b) ADMM-TV, (c) Lahiri *et al.* [16], (d) Chung *et al.* [5], (e) proposed method, (f) ground truth. PSNR/SSIM values presented in the upper right corner. Green lines in the inset of first row (a): measured angles.

	8-view						**4-view**						**2-view**					
	Axial*		**Coronal**		**Sagittal**		**Axial***		**Coronal**		**Sagittal**		**Axial***		**Coronal**		**Sagittal**	
Method	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DiffusionMBIR (ours)	**33.49**	**0.942**	**35.18**	**0.967**	**32.18**	**0.910**	**30.52**	**0.914**	**30.09**	**0.938**	**27.89**	**0.871**	24.11	0.810	23.15	**0.841**	**21.72**	**0.766**
Chung *et al.* [5]	28.61	0.873	28.05	0.884	24.45	0.765	27.33	0.855	26.52	0.863	23.04	0.745	**24.69**	**0.821**	**23.52**	0.806	20.71	0.685
Lahiri *et al.* [16]	21.38	0.711	23.89	0.769	20.81	0.716	20.37	0.652	21.41	0.721	18.40	0.665	19.74	0.631	19.92	0.720	17.34	0.650
FBPConvNet [11]	16.57	0.553	19.12	0.774	18.11	0.714	16.45	0.529	19.47	0.713	15.48	0.610	16.31	0.521	17.05	0.521	11.07	0.483
ADMM-TV	16.79	0.645	18.95	0.772	17.27	0.716	13.59	0.618	15.23	0.682	14.60	0.638	10.28	0.409	13.77	0.616	11.49	0.553

Table 1. Quantitative evaluation of SV-CT (8, 4, 2-view) (PSNR, SSIM) on the AAPM 256×256 test set. **Bold**: Best, under: second best.*: the plane where the diffusion model prior takes place. Holds the same for Table. 3,4.*: the plane where the diffusion model prior takes place. Holds the same for Table. 3,4

for testing. For the task of CS-MRI, we take the data from the multimodal brain tumor image segmentation benchmark (BRATS) [22] 2018 FLAIR volume for testing. Note that we use a pre-trained score function that was trained on fastMRI knee [36] images only, and hence we need not split the train/test data here.

Network training, inference. For CT tasks, we train the `ncsnpp` model [32] on the AAPM dataset which consists of about 3000 2D slices of training data. For the CS-MRI task, we take the pre-trained model checkpoint from[5] [7]. For inference (i.e. generation; inverse problem solving), we base our sampler on the predictor-corrector (PC) sampling scheme of [32]. We set $N = 2000$, which amounts to 4000 iterations of neural function evaluation with $\boldsymbol{s}_{\theta^*}$.

Comparison methods and evaluation. For CT tasks, we first compare our method with Chung *et al.* [5], which is another diffusion model approach for CT reconstruction, outperforming [30]. As using the manifold constrained gradient (MCG) of [5] requires 10GB of VRAM for a *single* 2D slice (256×256), it is infeasible for us to leverage such gradient step for our 3D reconstruction. Thus, we employ the projection onto convex sets (POCS) strategy of [5], which amounts to taking algebraic reconstruction technique (ART) in each data-consistency imposing step. We also compare against some of the best-in-class fully supervised methods. Namely, we include Lahiri *et al.* [16], and FBPConvNet [11] (SV-CT) / Zhang *et al.* [37] (LA-CT) as baselines. For the implementation of [16], we use 2 stages, but implement the De-streaking CNN as U-Nets rather than simpler CNNs. Finally, we include isotropic ADMM-TV, which uses the regularization function $R(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_{2,1}$.

In the CS-MRI experiments, we compare against score-MRI [7] as a representative diffusion model-based solver. Moreover, we include comparisons with DuDoRNet [39],

[5] `https://github.com/HJ-harry/score-MRI`

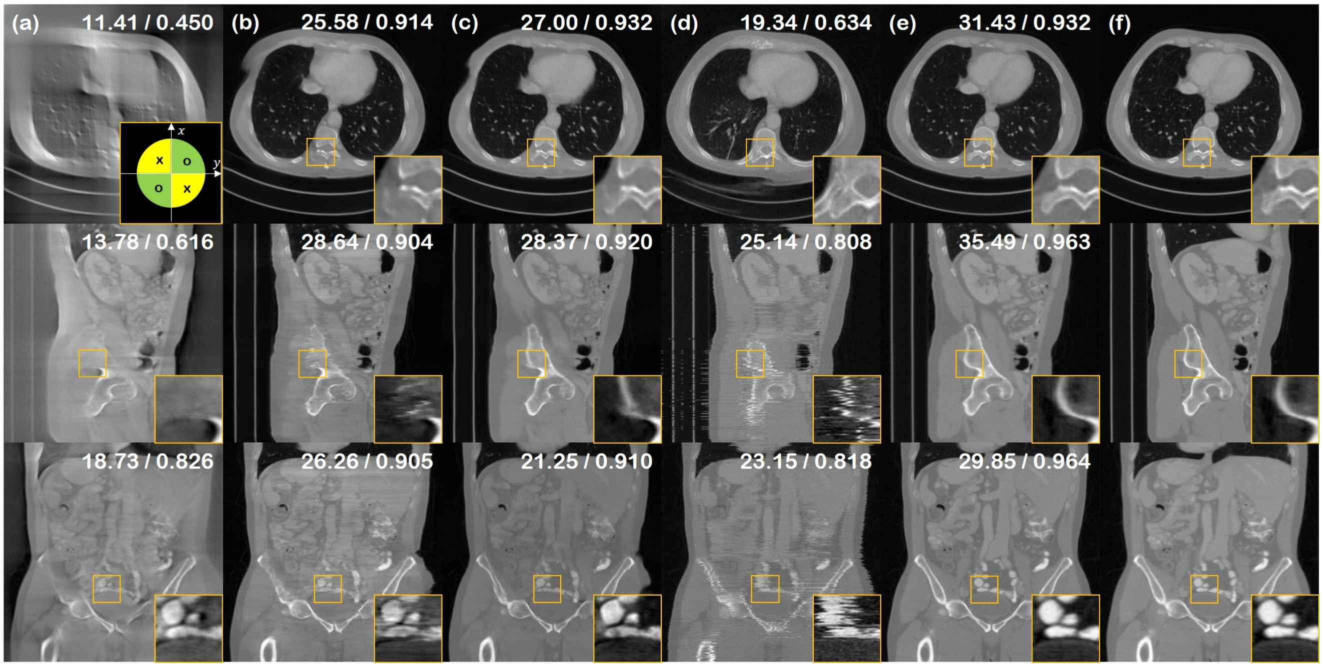

Figure 4. 90° LA-CT reconstruction results of the test data (First row: axial slice, second row: sagittal slice, third row: coronal slice). (a) FBP, (b) Zhang *et al.* [37], (c) Lahiri *et al.* [16], (d) Chung *et al.* [5], (e) proposed method, (f) ground truth. PSNR/SSIM values presented in the upper right corner. Green area in the inset of first row (a): measured, Yellow area in the inset of first row (a): not measured.

and U-Net [36]. Note that all networks including ours, the training dataset (fastMRI knee) was set deliberately different from the testing data (BRATS Flair), as it was shown that diffusion model-based inverse problem solvers are fairly robust to out of distribution (OOD) data in CS-MRI settings [7, 10]. Quantitative evaluation was performed with two standard metrics: peak-signal-to-noise-ratio (PSNR), and structural similarity index (SSIM). We report on metrics that are averaged over each planar direction, as we expect different performance for xy-slices as compared to xz- and yz-slices.

5. Results

In this section, we present the results of the proposed method. For further experiments and ablation studies, see supplementary section C.

Sparse-view CT. We present the quantitative metrics of the SV-CT reconstruction results in Table 1. The table shows that the proposed method outscores the baselines by large margins in most of the settings. Fig. 3 and Supplementary Fig. 6 show the 8,4-view SV-CT reconstruction result. As shown at the first row of each figure, axial slices of the proposed method have restored much finer details compared to the baselines. Furthermore, the results of sagittal and coronal slices in the second and third rows imply that DiffusionMBIR could maintain the structural connectivity of the original structures in all directions. In contrast, Chung et al. [5] performs well on reconstructing the axial slices, but do not have spatial integrity across the z direction, leading to shaggy artifacts that can be clearly seen in coronal/sagittal slices. Lahiri et al. [16] often omits important details, and is not capable of reconstruction, especially when we only have 4 number of views. ADMM-TV hardly produces satisfactory results due to the extremely limited setting.

Limited angle CT. The results of the limited angle tomography is presented in Table 3 and Fig. 4. We test on the case where we have measurements in the $[0, 90]°$ regime, and no measurements in the $[90, 180]°$ regime. Hence, the task is to *infill* the missing views. Consistent with what was observed from SV-CT experiments, we see that DiffusionMBIR improves over the conventional diffusion model-based method [5], and also outperforms other fully supervised methods, where we see even larger gaps in performance between the proposed method and all the other methods. Notably, Chung et al. [5] leverages no information from the adjacent slices, and hence has high degree of freedom on how to infill the missing angle. As the reconstruction is stochastic, we cannot impose consistency across the different slices. Often, this results in the structure of the torso being completely distorted, as can be seen in the first row of Fig. 4 (d). In contrast, our augmented prior imposes smoothness across frames, and also naturally robustly preserves the structure.

Compressed Sensing MRI. We test our method on the reconstruction of 1D uniform random sub-sampled images,

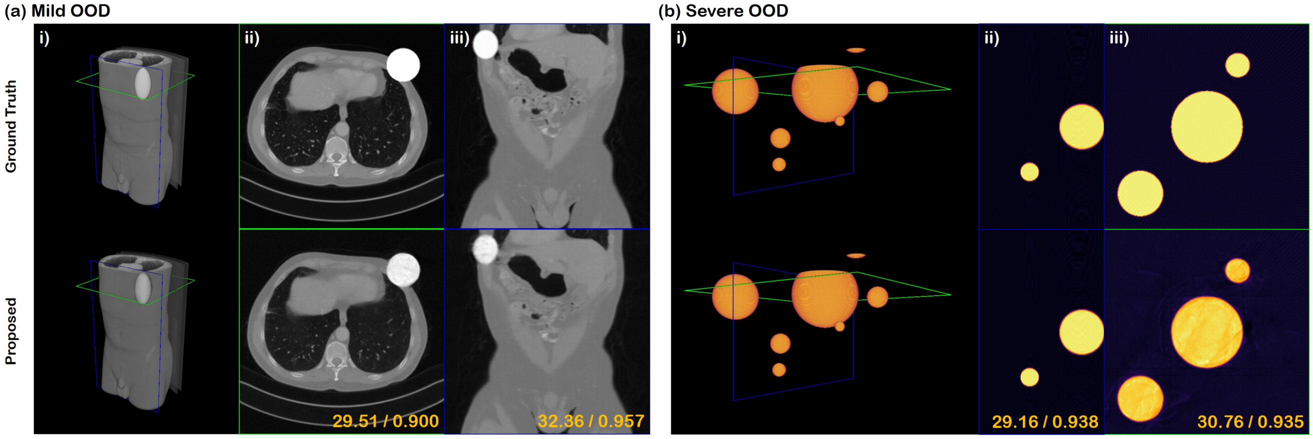

Figure 5. 8-view SV-CT reconstruction results of the OOD data (Same geometry as in Fig. 3). (a) Ellipsis laid on top of the test data volume, (b) Phantom that consists of spheres located randomly.

as was used in [36]. Specifically, we keep 15% of the autocalibrating signal (ACS) region in the center, and retain only the half of the k-space sampling lines, corresponding to approximately 2× acceleration factor for the acquisition scheme. The results are presented in Table 4 and supplementary Fig. 8. Consistent with what was observed in the experiments with SV-CT and LA-CT, we observe large improvements over the prior arts.

Out-of-distribution performance. It was shown in the context of CS-MRI that diffusion models are surprisingly robust to the out-of-distribution (OOD) data [7, 10]. For example, the score function trained with proton-density weighted, coronal knee scans only was able to generalize to nearly all the different anatomy and contrasts that were never seen at the test time. Would such generalization capacity also hold in the context of CT? Here, we answer with a positive, and show that with the proposed method, one can use the same score function even when the targeted ground truth is vastly different from those in the training dataset.

In Fig. 5, we show two different cases (i.e. mild, severe) of such OOD reconstruction. For both cases, we see that 8-view is enough to produce high fidelity reconstructions that closely estimate the ground truth. Note that our prior is constructed from the anatomy of the human body, which is vastly different from what is given in Fig. 5 (b). Intuitively, in the Bayesian perspective, this means that our diffusion prior would have placed very little mass on images such as Fig. 5 (b). Nonetheless, we can interpret that diffusion models tend to place *some* mass even to these heavily OOD regions. Consequently, when incorporated with enough likelihood information, it is sufficient to guide proper posterior sampling, as seen in this experiment. Such property is particularly useful in medical imaging, where training data mostly consists of normal patient data, and the distribution is hence biased towards images without pathology. When at test time, we are given a patient scan that contains lesions, we desire a method that can fully generalize in such cases.

Choice of augmented prior. In this work, we proposed to augment the diffusion generative prior with the model-based TV prior, in which we chose to impose the TV constraint only in the redundant $z-$direction, while leaving the $xy-$plane intact. This design choice stems from our assumption that diffusion prior *better* matches the actual prior distribution than the TV prior, and mixing with the TV prior might compromise the ability of diffusion prior.

To verify that this is indeed the case, we conducted an ablation study on TV(xyz) and TV(z) prior. Supplementary Fig. 10 shows the visual and numerical results on the priors. Both priors have scored high on PSNR and SSIM, but we can figure out that the images with TV prior on xy-plane are blurry compared to the samples from the proposed one. The analysis implies that the usage of TV(xyz) prior shifted the result a bit far apart from the well-trained diffusion prior.

6. Conclusion

In this work, we propose DiffusionMBIR, a diffusion model-based reconstruction strategy for performing 3D medical image reconstruction. We show that all we need is a 2D diffusion model that can be trained with little data ($<$ 10 volumes), augmented with a classic TV prior that operates on the redundant z direction. We devise a way to seamlessly integrate the usual diffusion model sampling steps with the ADMM iterations in an efficient way. The results demonstrate that the proposed method is capable of achieving state-of-the-art reconstructions on Sparse-view CT, Limited-angle CT, and Compressed-sensing MRI. Specifically for Sparse-view CT, we show that our method is capable of providing accurate reconstructions even with as few as two views. Finally, we show that DiffusionMBIR is capable of reconstructing OOD data that is vastly different from what is presented in the training data.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 3

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. 3

[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 2, 11

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[5] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 6, 7, 12, 13

[6] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4

[7] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical Image Analysis*, page 102479, 2022. 2, 6, 7, 8, 13, 14

[8] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. 2

[10] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. In *Advances in Neural Information Processing Systems*, volume 34, pages 14938–14954, 2021. 7, 8

[11] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. 6, 13

[12] Hong Jung, Kyunghyun Sung, Krishna S Nayak, Eung Yeop Kim, and Jong Chul Ye. k-t focuss: a general compressed sensing framework for high resolution dynamic mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 61(1):103–116, 2009. 2

[13] Eunhee Kang, Junhong Min, and Jong Chul Ye. A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. *Medical physics*, 44(10):e360–e375, 2017. 11

[14] Masaki Katsura, Izuru Matsuda, Masaaki Akahane, Jiro Sato, Hiroyuki Akai, Koichiro Yasaka, Akira Kunimatsu, and Kuni Ohtomo. Model-based iterative reconstruction technique for radiation dose reduction in chest ct: comparison with the adaptive statistical iterative reconstruction technique. *European radiology*, 22(8):1613–1623, 2012. 2

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. 2

[16] Anish Lahiri, Marc Klasky, Jeffrey A Fessler, and Saiprasad Ravishankar. Sparse-view cone beam ct reconstruction using data-consistent supervised and adversarial learning from scarce training data. *arXiv preprint arXiv:2201.09318*, 2022. 6, 7, 12, 13

[17] Lu Liu. Model-based iterative reconstruction: a promising algorithm for today's computed tomography imaging. *Journal of Medical imaging and Radiation sciences*, 45(2):131–136, 2014. 2

[18] Yan Liu, Jianhua Ma, Yi Fan, and Zhengrong Liang. Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction. *Physics in Medicine & Biology*, 57(23):7923, 2012. 2

[19] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[20] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 4

[21] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. 4

[22] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 6, 12

[23] Frédéric Noo, Michel Defrise, and Rolf Clackdoyle. Single-slice rebinning method for helical cone-beam ct. *Physics in Medicine & Biology*, 44(2):561, 1999. 11

[24] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014. 11

[25] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. *arXiv preprint arXiv:2209.07162*, 2022. 4

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 4

[27] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019. 3

[28] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777, 2008. 2

[29] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021. 12

[30] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021. 6

[31] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022. 2

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*, 2021. 2, 3, 4, 6, 12

[33] Jiaqi Sun, Alireza Entezari, and Baba C Vemuri. Exploiting structural redundancy in q-space for improved eap reconstruction from highly undersampled (k, q)-space in dmri. *Medical image analysis*, 54:122–137, 2019. 2

[34] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 3

[35] Dominik JE Waibel, Ernst Röoell, Bastian Rieck, Raja Giryes, and Carsten Marr. A diffusion model predicts 3d shapes from 2d microscopy images. *arXiv preprint arXiv:2208.14125*, 2022. 4

[36] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018. 6, 7, 8, 12, 13, 14

[37] Hanming Zhang, Liang Li, Kai Qiao, Linyuan Wang, Bin Yan, Lei Li, and Guoen Hu. Image prediction for limited-angle tomography via deep learning with convolutional neural network. *arXiv preprint arXiv:1607.08707*, 2016. 6, 7, 13

[38] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 13

[39] Bo Zhou and S Kevin Zhou. Dudornet: learning a dual-domain recurrent network for fast mri reconstruction with deep t1 prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4273–4282, 2020. 6, 13, 14

[40] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 4