Homotopic Gradients of Generative Density Priors for MR Image Reconstruction

Cong Quan, Jinjie Zhou, Yuanzheng Zhu, Yang Chen, Shanshan Wang, Dong Liang, Senior Member, IEEE, Qiegen Liu, Senior Member, IEEE,

Abstract—Deep learning, particularly the generative model, has demonstrated tremendous potential to significantly speed up image reconstruction with reduced measurements recently. Rather than the existing generative models that often optimize the density priors, in this work, by taking advantage of the denoising score matching, homotopic gradients of generative density priors (HGGDP) are proposed for magnetic resonance imaging (MRI) reconstruction. More precisely, to tackle the low-dimensional manifold and low data density region issues in generative density prior, we estimate the target gradients in higher-dimensional space. We train a more powerful noise conditional score network by forming high-dimensional tensor as the network input at the training phase. More artificial noise is also injected in the embedding space. At the reconstruction stage, a homotopy method is employed to pursue the density prior, such as to boost the reconstruction performance. Experiment results imply the remarkable performance of HGGDP in terms of high reconstruction accuracy; only 10% of the k-space data can still generate images of high quality as effectively as standard MRI reconstruction with the fully sampled data.

Index Terms—Magnetic Resonance Imaging (MRI), Imag reconstruction, Unsupervised learning, Generative model, Homotopic gradients.

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) noninvasively depicts structural and functional features inside a patient in rich contrasts, supporting today’s medical diagnosis and researches. However, one major drawback of MRI is the relatively slow data acquisition speed, causing non-idealized space resolution, patient discomfort and hinders applications in time-critical diagnoses. Under the premise of guaranteeing the image quality, various acceleration techniques have been developed to speed up acquisition time.

Recently, the development of compressed sensing (CS) theory has shown that it is possible to reconstruct MR images from much fewer k-space (i.e. Fourier space) measurements [1]. Specifically, CS allows fast acquisition that bypasses the Nyquist-Shannon sampling criteria with more aggressive undersampling [2]. To apply the CS theory to MRI reconstruction, it must find a proper sparsifying transformation to make the signal sparse, e.g., wavelet transformation [3], and solve a minimization with regularizers. As shown in Fig. 1, the major work of CS-MRI can be categorized primarily into two groups: model-based methods [4]-[14] and deep learning methods [14]-[29].

![Fig. 1. Classification of the existing CS-MRI algorithms.](image)

In model-based methods, CS-MRI focuses on applying predefined sparsify transforms, such as the discrete cosine transform (DCT) [4], total variation (TV) [5], [6], discrete wavelet transform (DWT) [3], [7] or contourlet transform [8], and developing efficient numerical algorithms to solve nonlinear optimization problems [9], [10]. Besides, CS-MRI has been extended to a more flexible sparse representation learned directly from data using dictionary learning, which expresses local features of reconstructed images more accurately compared to predefined universal transforms [11]-[14]. Even though dictionary learning-based methods achieve higher reconstruction quality, the reconstruction process still suffers from longer running time due to the expense of heavy computational burden for dictionary training and sparse coding.

Fueled by the success of deep learning [15], many techniques have been proposed to implement efficient and accurate MRI reconstruction. These approaches mainly can be categorized as follows: supervised and unsupervised scheme [16]. Supervised learning is a learning technique that uses pairs of data samples [2], [17]-[24]. For instance, Schlemper et al. trained a deep cascade of convolutional neural networks (DCCNN) to map undersampled data to ground-truth data [17]. Aggarwal et al. introduced a model-inspired network learning framework with a convolution neural network (CNN) based regularization prior [19]. Liu et al. unrolled the iterative feature refinement procedures in IFR-CS to a supervised model-driven network, dubbed IFR-Net [24]. However, it is revealed that the training of supervised learning requires a huge number of labeled samples and a great deal of expertise to ensure proper convergence. Generative adversarial networks (GAN) has shown great potential in solving CS-MRI [25]-[27]. Although it is originally used as an un-
supervised learning way, a number of GAN-based algorithms fall into the supervised scheme, as they only combine the popular end-to-end network with discriminator of GAN. For instance, Conditional generative adversarial network-based approach (DAGAN) [2] not only designed end-to-end U-Net network to reduce aliasing artifacts, but also coupled the adversarial loss with an innovative content loss. Both image and frequency domains are incorporated in the network.

Unsupervised learning systems are used when paired data labels are not available. Autoencoder (AE) is a type of neural network for learning efficient data encoding in unsupervised way. Until now, two variants of AE were applied to MRI reconstruction. Particularly, Tezcan et al. used variational autoencoders (VAE) as a tool for describing the data density prior [28]. Liu et al. [29] employed enhanced denoising autoencoders (DAE) as prior to MRI reconstruction (DAEPRec), which leverages DAE network as an explicit prior and maximized the likelihood using gradient descent by backpropagating the autoencoder error for MRI reconstruction. Zhang et al. [30] further exploited the gradient of the data density prior to reconstruction, termed EDMSPRec.

These two works provided promising results, unlike VAE in [28], they are not directly related to generative models, lacking intuitive interpretation in the terminology of machine learning. Besides of the variants of AE, auto-regressive generative models like the PixelCNN were verified in MRI reconstruction recently [31].

Lately, some underlying theoretic schemes with regard to DAE were reported by different research groups. Vincent et al. proposed denoising score matching (DSM) and revealed that DAE is equivalent to performing DSM [32], [33]. Lately, Song et al. introduced a new generative model named noise conditional score networks (NCSN), where samples are produced via Langevin dynamics using gradients of the data distribution estimated with SM progressively [34]. Motivated by the aforementioned relations that DAE and DSM can be used as good tools for generative model [35], [36], in this study are devoted to leveraging this tool to accelerate the MRI reconstruction. Specifically, we explicitly immerse in embedded priors on gradients of the auxiliary data distribution to attain higher-dimensional gradients, to recover MR images more precisely. Essentially, constructing and conducting noisy prior in higher-dimensional space favors to tackle the low-dimensional manifold and low data density region issues in generative models. Results indicate that our proposal stands out for its efficiency. We highlight our main contributions as follows:

- By investigating the two major issues of low-dimensional manifold hypothesis and low data density region that occurred in estimating gradients of the data distribution, an enhanced method equipped with a higher-dimensional structure is proposed.
- Rather than the classical iterative algorithms that integrating prior information in iterative reconstruction, our algorithm views the observation as a conditional variable and incorporates it into the homotopic generative procedure.

The rest of this paper is presented as follows. Section II briefly describes some previous relevant works on CS-MRI and the relationship between DAE and DSM. In Section III, we elaborate the proposed problem formulation of the training and MRI reconstruction stages based on higher-dimensional embedding and homotopy techniques, respectively. Section IV presents the MRI reconstruction performance of the proposed scheme, includes the experimental setup, reconstruction comparisons and ablation study. Section V concludes with topics and future works are given for the end.

II. PRELIMINARY

A. CS-MRI

In CS theory [4], assuming is an MR image to be reconstructed and is the under-sampled k-space measure, the reconstructed image can be estimated by solving the following optimization:

$$\text{Min}_x \left\{ |F_p x - y|_2^2 + \lambda R(x) \right\}$$

(1)

where $$F_p = PF$$ is the measurement matrix, $$P$$ is an under-sampling matrix and $$F$$ is a Fourier transform. Parameter $$\lambda$$ balances the importance of data-fidelity term $$\|F_p x - y\|_2^2$$ and regularization term $$R(x)$$.

Classical model-based CS-MRI approaches often adopt an iterative scheme to alternatively update the intermediate solution constrained by data-consistency and prior information. Early supervised deep learning-based CS-MRI [37], [38] learns a mapping function $$x = f(x_0)$$ between the zero-filling reconstructed image $$x_0$$ and fully-sampled reconstructed image $$x$$ by training a CNN with lots of pair data. The supervised end-to-end learning approaches essentially use a discriminative fashion to learn an implicit prior, lacking flexibility and robustness. In this work, we turn to the explicit prior construction for CS-MRI via generative models.

B. From DAE to DSM

By taking advantage of the high nonlinearity and capacity of neural networks, AE has become one of the most important unsupervised learning methods for the task of representation learning [17], [39]. In order to improve the insensitivity of AE on input data, DAE tackles a partially corrupted input and is trained to recover the original undistorted input [13], [28], [29],[39]. Essentially, denoising in DAE is acted as a regularized criterion for learning to capture useful structure from the input data.

Bigdeli et al. [40] used the magnitude of the autoencoder error as a prior (DAEP) for image recovery and the DAE $$A_o$$ is trained to minimize the expectation over all input images $$x$$:

$$L_{med}(A) = E_x \left[ \left\| A_o(x + \eta) - x \right\|_2 \right]$$

(2)

where the output $$A_o(x)$$ is trained by adding artificial Gaussian noise $$\eta$$ with standard deviation $$\sigma$$.

According to [40], the network output $$A_o(x)$$ is related to the true data density $$p(x)$$ as follows:

$$A_o(x) = x + \int_{-\infty}^{\infty} \frac{g_o(\eta)p(x-\eta)d\eta}{\int_{-\infty}^{\infty} g_o(\eta)p(x-\eta)d\eta}$$

$$= x + \nabla \log \int_{-\infty}^{\infty} g_o(\eta)p(x+\eta)d\eta$$

(3)

where $$g_o(\eta)$$ represents a local Gaussian kernel with standard deviation $$\sigma$$. Due to this observation, experiments by Liu et al. [29] shown that a simple and effective way to incorporate the DAE prior into highly undersampling MRI
reconstruction can achieve promising performance.

As seen in Eq. (3), DAE learns a mean-shift vector field from a given set of data samples, which is proportional to the gradient of the logarithm of the prior. Hence, a new prior called deep mean-shift prior (DMSP) was proposed by Bigdeli et al. [41]. They utilized it in a gradient descent approach to perform Bayes risk minimization. DMSP is formulated as follows:

$$
\nabla \text{prior}(x) = \nabla \log \int g_\eta(p(x + \eta) d\eta = \left[ (A_\sigma(x) - x) / \sigma^2 \right]
$$

By extending the naive DMSP with integration of multi-model aggregation and multi-channel network learning, Zhang et al. [30] proposed a high-dimensional embedding network derived prior, and applied the learned prior to single-channel MRI reconstruction via variable augmentation technique.

Recently, Vincent showed that there is an equivalence between DAE and DSM [33]. Furthermore, Block et al. shed new light on what the DAE learns from a distribution, showing that optimizing DAE loss is equivalent to optimizing DSM loss [42] (e.g. let $p$ be a differentiable density), the connection between DAE and DSM is as follows:

**Theorem 1.** The DAE loss

$$
L_{\text{DAE}}(A) = E_{\nu, \eta} \left[ \| A(\eta) - x \|^2 \right]
$$

and the DSM loss

$$
L_{\text{DSM}}(S) = E_{\nu} \left[ \| S(x) - \nabla \log p_\sigma(x) \|^2 \right]
$$

with

$$
S(x) = \frac{A(x) - x}{\sigma^2}
$$

are equivalent up to a term that does not depend on $A$ or $S$.

More details of Theorem 1 are provided in Supplementary material. As seen from DAE in Theorem 1, DSM optimizes the data distribution more directly. In this work, based on the connection between DAE and DSM, we will extend our preliminary idea of multi-channel and multi-view noise schemes [29], [30] on DAE to the minimization on gradients of the data distribution. More importantly, we will reveal the underlying benefits of this idea from the viewpoint of machine learning.

**III. PROPOSED HGGDP MODEL**

In the previous section, we discuss the DAE and DSM, where the estimation is done from data density to gradient of data density. In this section, we concentrate on effectively improving the performance of native DSM on MRI reconstruction, coined deep gradients of generative density prior (HGGDP). Specifically, we introduce a way of to avert the data confined to low data density regions and avoid difficulties from the manifold hypothesis, i.e., forming high-dimensional tensor and injecting artificial noise.

Fig. 2 shows a whole procedure of HGGDP and HGGDPRec, the original MR data manifold can be seen in Fig. 2(a). As long as the dimension is high enough, there exit two measure concentration properties for random vectors: The length distribution of random vectors becomes concentrated at $\sqrt{\sigma}$ and a random vector is always close to orthogonal to a fixed vector. As analyzed by the DSM procedure [43], [44], noisy samples concentrate on a set $\mathcal{X}_{\sqrt{\sigma}}$ that has a distance with $(\sqrt{\sigma} - \varepsilon, \sqrt{\sigma} + \varepsilon)$ from the data manifold $\mathcal{X}$, where $\varepsilon<\sqrt{\sigma}$. Performing DSM learning data with a fixed noise level $\sigma$ in high-dimensional space provides little information about the density outside this set. During training, derivative of loglikelihood is forced to point toward data manifold, as illustrated in Fig. 2(b). Hence how to reduce the low-data density regions is urgent.

Similar issue also motivates researchers to use multiple noise scales [34][43]. This may explain the difficulty previous models based on DSM with single noise scale encountered when sampled with noise initialization. To further inherit the spirit in [34][43], we enhance the DSM by multi-channel MR images with multi-view noise. The training and reconstruction phases can be shown in Fig. 2 (c) and Fig. 2 (d). To illustrate this point, in the right region of Fig. 2 we show some example of intermediate samples during annealed sampling process. Displayed are samples obtained by single step denoising from the noisy sample carrying different levels of noise. One can see that as the sample point approach the data manifold, more and more details are present in the denoised sample. We provide more result and explanation from this in Section III. B. It is an ideal situation that all the sample points in training and reconstructing phase approach to the original data manifold. Thus, the purified data manifold is obtained in Fig. 2 (e) finally.

![Fig. 2. Schematic diagram of manifold learning using HGGDP. (a) Ground-truth data manifold; (b) DSM with single noise scale; (c,d) Training and reconstruction stage via projecting on multi-channel space with multiple noise scale; (e) Purified data. The legend is located at the bottom.](image)

A. **Baseline: Analysis of NCSN**

Generative models are well studied and have shown outstanding performance in modeling data distributions to capitalize on known input invariances that might form good prior knowledge for informing the parameter learning [45], [46]. However, most existing generative models, either optimizing log-likelihood or adversarial training, are devoted to describing the data distribution faithfully. For instance, VAE [47] provides a formulation and a sampling approach to generate more details of images, but the samples often suffer from being blurry.
In [34], Song et al. presented NCSN that trains a score network $S_p(x, \sigma)$ directly to estimate the gradient of data prior $\nabla \log p_\sigma(x)$ instead of data prior $p_\sigma(x)$ and employed annealed Langevin dynamics for image generation (i.e., $p_\sigma(x) \approx p_{\sigma_\text{max}}(x)$, $\sigma \rightarrow 0$). More precisely, NCSN estimates the gradients of the data distribution with DSM. Empirically, DSM is adopted as it is slightly fast and naturally fits the task of estimating scores of noise-perturbed data distributions. i.e., Assuming the noise is distribution is chosen to be $p_\sigma(x) = N(\mu(x), \sigma^2)$, subsequently it leads to

$$\nabla \log p_\sigma(x) = (\tilde{\mu} - \mu) / \sigma^2.$$  

For a given $\sigma$, the objective of DSM is

$$L(\theta; \sigma) \triangleq \frac{1}{2} E_{p_{\sigma}(x)}\left[|S_p(x, \sigma) - \nabla \log p_\sigma(x)|^2\right]$$

$$= \frac{1}{2} E_{p_{\sigma}(x)}\left[|\nabla S_p(x, \sigma) - \nabla \log p_\sigma(x)|^2\right] + C$$

(8)

where $C$ is a constant that does not depend on $\theta$.

Beside of the statistic derivation to Eq. (8) [33], [34], we provide a new intuitive derivation for it. In fact, as stated in Theorem 1, we get $S_p(x, \sigma) = \frac{1}{2} \left( A(x + \eta) - (\tilde{\mu} + \eta) \right) / \sigma^2$. On the other hand, as described in Fig. 3, if we denote $D_p(x, \sigma) = (x + \eta) - A(x + \eta)$ and $A(x + \eta) \rightarrow x$, then we get $S_p(x, \sigma) = -D_p(x, \sigma) / \sigma^2 \rightarrow -\eta / \sigma^2 = -(\tilde{\mu} - \mu) / \sigma^2$.

![Fig. 3. Visual exhibition of the network A and D.](image)

After Eq. (8) is derived, the DSM loss is combined for all $\sigma \in \{\sigma_i\}_{i=1}^I$ to get one unified objective

$$L(\theta; \{\sigma_i\}_{i=1}^I) = \frac{1}{I} \sum_{i=1}^I \lambda(\sigma_i) L(\theta; \sigma_i)$$

(9)

where $\lambda(\sigma_i) > 0$ is a coefficient function depending on $\sigma_i$.

Since Eq. (9) is a conical combination of $I$ DSM objectives, the optimal score $S_p(x, \sigma)$ minimizes Eq. (9) if and only if $S_p(x, \sigma) = \nabla \log p_\sigma(x)$ is satisfied for all $i \in \{1, 2, \ldots, I\}$.

As the training procedure of the NCSN $S_p(x, \sigma)$ is determined, annealed Langevin dynamics as a sampling approach is introduced. The annealed Langevin dynamics method recursively computes the following:

$$\tilde{x}_i = \tilde{x}_{i-1} + \frac{\alpha_i}{2} \nabla \log p_\sigma(\tilde{x}_{i-1}) + \sqrt{\alpha_i} z_i$$

$$= \tilde{x}_{i-1} + \frac{\alpha_i}{2} S_p(\tilde{x}_{i-1}, \sigma_i) + \sqrt{\alpha_i} z_i$$

(10)

where $\alpha_i$ is the step size by tuning down it gradually. $T$ is the number of iterations for each noise level, and $z_i \sim \mathcal{N}(0, 1)$.

To sum up, two key contributions of NCSN can be summarized as follows: One is perturbing the data with random Gaussian noise of various magnitudes; The other is that an annealed version of Langevin dynamics is initialized using scores corresponding to the highest noise level, and gradually annealing down the noise level until it is small enough to be indistinguishable from the original data distribution. Essentially, the core innovation behind the above contributions is the flexible usage of artificial noise.

Although NCSN has achieved good results, its application in MRI reconstruction is still leaving huge room for improvement, particularly on two major deficiencies: low data density regions and the manifold hypothesis [48], [49]. For instance, it is often assumed that the data distribution is supported on a low dimensional manifold. Then, the score will be undefined in the ambient space such that matching method will fail to provide a consistent score estimator. Additionally, the scarcity of training data in low data density regions, e.g., far from the manifold, hinders the matching accuracy and slows down the mixing of Langevin dynamics sampling.

To circumvent the above limitations, we propose HGGDP to enrich the native NCSN by taking advantage of the usage of artificial noise in higher-dimensional space. The underlying idea is to span the whole space with more samples to override the deficiency from the manifold hypothesis and enable the data not to be confined to low data density regions.

### B. HGGDP: Prior in High-dimensional Space

We provide a detailed analysis of the naive NCSN. As a result, we enrich it by introducing a high-dimensional space embedding strategy to boost the representation, and thereafter reconstruction performance.

**Theorem 2** [42]. Let $F$ be a class of $\mathbb{R}^d$-valued functions, all of which are $M/2$-Lipschitz, bounded coordinate wise by $R > 0$, containing arbitrarily good approximations of $\nabla \log p_\sigma$ on the ball of radius $R$. Let $\sigma < \sigma_{\text{max}}$ and support we have $n$ i.i.d samples from $p_\sigma$, $x_1, \ldots, x_n$. Let

$$\hat{S} \in \arg\min_{\hat{S}} \frac{1}{n} \sum_{i=1}^n \left| S(x_i) - \nabla \log p_\sigma(x_i) \right|$$

(11)

Then with probability at least $1 - 4dC - 4n e^{-C_1 n^{\alpha}}$ on the randomness due to the sample,

$$E_{p_\sigma}\left[\left| \hat{S}(x) - \nabla \log p_\sigma(x) \right| \right] \leq C(M + B)^2 (\log^3 n - 3 \log^2 (F) + \beta n d)$$

(12)

where $m = m - \frac{\sigma_M^2}{2}$ and $\beta_n = \frac{\log \log n - \log \delta}{n}$.

$C$ is a universal constant.

From **Theorem 2** [42], it is concluded that the representation boundary is related to the sample number $n$, the space dimension $d$. Particularly, the larger the sample number is, the less the representation error will be. This observation motivates us to boost the performance via sampling in higher-dimensional embedding space. The whole strategy consists of two consecutive operations: forming high-dimensional tensor and injecting artificial noise.

### Forming higher-dimensional tensor: In regions of low data density, the original DSM may not have enough evidence to estimate score functions accurately, due to the lack of data samples. EDAAEPReC proposed by Liu et al. handled complex data by concatenating the real and imaginary parts as channels, that is, the real and imaginary components are set as the network input simultaneously via replication technology. They elaborated that prior information learned from high-dimensional tensor is more effective for MRI reconstruction than the low-dimensional counterpart [29].

Inspired by [29], we use replication technology for the
generative model NCSN. In detail, we utilize a general transformation \( X = H(x) \) to establish a \( N \)-channel higher-dimensional tensor e.g., the vector variable with \( N \) channels as \( X = [x, x, \ldots, x] \). The goal of stacking to be \( X \) is to obtain much more data information at high-dimensional manifold and high-density regions, thus avoiding some difficulties for both accuracy in score estimation and sampling with Langevin dynamics. Subsequently, the HGGDP is trained with \( X \) in high-dimensional space as network input and the parameterized \( S_{\theta}(X, \sigma) \) is obtained.

To show the superior performance of higher-dimensional formulation strategy, we visualize the data scores \( \nabla_{\theta} \log p_{data}(x) \) and the estimated scores \( S_{\theta}(X, \sigma) \) of HGGDP in Fig. 4. As can be seen, after expanding the space dimension, the diversity of score matching increases. Additionally, the trouble of low data density regions is alleviated as the data samples in 3D space increases.

In our case of handling complex-valued MR image, the transformation \( H(x) \) is adopted to form high dimensional tensor \( X \), as shown in Fig. 5. A \( N \)-channel high-dimensional tensor \( X = \{[x_{\text{real}}, x_{\text{imag}}], \ldots, [x_{\text{real}}, x_{\text{imag}}]\} \) is the channel concatenation of real and imaginary components. In this work \( X = \{[x_{\text{real}}, x_{\text{imag}}], [x_{\text{real}}, x_{\text{imag}}], [x_{\text{real}}, x_{\text{imag}}]\} \) with the setting of \( N = 3 \).

**Injecting artificial noise:** The manifold hypothesis states that data in the real world tend to concentrate on low-dimensional manifolds embedded in an ambient space. There are two vital issues faced by score-based generative models. On one hand, when \( X \) is restricted to a low dimensional manifold, the score \( \nabla_{\theta} \log p_{data}(x) \) is undefined, which is a gradient taken in the ambient space. On the other hand, a consistent score estimator is afforded only when the support of the data distribution is the whole space; If the data reside on a low-dimensional manifold, it will be inconsistent. Nevertheless, Song et al. discovered that perturbing the data with random Gaussian noise of various magnitudes can ensure the resulting distribution does not collapse to a low dimensional manifold [34].

After forming \( X \) in higher-dimensional space, artificial noise injection is needed to fill in the whole space. Similarly as in Eq. (9), the unified objective is executed on all samples \( X \) for all \( \sigma \in \{\sigma_i\}_{i=1}^{L} \) to pursue the optimal score,

\[
\hat{L}(\theta; \{\sigma_i\}_{i=1}^{L}) = \frac{1}{L} \sum_{i=1}^{L} \hat{L}(\sigma_i; \theta; \sigma_i)
\]  

Furthermore, after learning the parameterized score of probability density \( S_{\theta}(X, \sigma) \) from \( X \), artificial noise injection is also done at the reconstruction phase. Specifically, a simulated phase map is produced with varying levels of Gaussian noise added to \( X \). More critically, injecting noise in higher-dimensional space will produce more diverse samples in the perturbed data distribution than that in the original data distribution, thus improving score estimation accuracy. Taking advantage of it, the reconstruction result is obtained by an operation that makes simulated phase map as the input of \( S_{\theta}(X, \sigma) \) via gradually annealed noise to iterate repeatedly. i.e.,

\[
X' = X^{-1} + \frac{\sigma}{2} S_{\theta}(X^{-1}, \sigma) + \sqrt{\sigma} \epsilon
\]  

**Fig. 5. Visualization of the three-channel high-dimensional tensor \( X \).**

**Fig. 6. Performance comparison of “injecting noise in high-dimensional tensor”: (a) Training loss in DSM of the naive NCSN and the advanced HGGDP at each epoch. (b) Image quality comparison on the brain data at 15% radial sampling: Reconstruction images, error maps (Red) and zoom-in results (Green).**

The benefit of injecting noise in high-dimensional space is visualized in Fig. 6. Fig. 6(a) shows the training losses of the objects that occurred in noisy single-channel and three-channel space, respectively. The network is trained to estimate the data score on SIAT dataset that perturbed with small Gaussian noise. Notice that the imposed Gaussian noise \( N(0, 0.0001) \) is very small for images with pixel values in the range \([0, 1]\), and is almost indistinguishable to human eyes. As can be seen, both loss curves first decrease and then fluctuate as the number of epochs increases. Two distinct
phenomena can be observed. First, the loss with single-channel fluctuations is much lower than the result in the three-channel circumstance, indicating better convergence property. Second, the loss value trained on three-channel MR images is much lower than the single-channel counterpart, which implies better score accuracy. In addition, Fig. 6(b) visualizes the reconstruction with enlarged and difference views by NCSN and HGGDP. In comparison, HGGDP achieves higher reconstruction accuracy and gives more faithful result.

As illustrated in Fig. 7(a), we use annealed Langevin dynamics to further sample from the high-dimensional noisy data distribution with multi-view noise. In order to intuit the procedure of annealed Langevin dynamics, we provide the intermediate samples in Fig. 7(b), where each row shows how samples evolve from pure random noise to high-resolution MR images.

![Annealed Langevin Dynamics](image)

**Fig. 7.** Pipeline of sampling from the high-dimensional noisy data distribution with multi-view noise and intermediate samples. (a) Conceptual diagram of the sampling on high-dimensional noisy data distribution with multi-view noise. (b) Intermediate samples of annealed Langevin dynamics.

C. HGGDPRrec: A Homotopic Method

As proved in [33] and shown in Eq. (8), the DSM is optimally accurate only when \( p_r(x) \rightarrow p_{\text{data}}(x) \) (i.e., \( \nabla \log p_r(x) \rightarrow \nabla \log p_{\text{data}}(x) \)), \( \sigma \rightarrow 0 \). Therefore, a homotopic process is needed to attain accurate density prior. Rather than the classical iterative algorithms that integrating prior information in iterative reconstruction, the proposed HGGDPRrec views the observation as a conditional variable and incorporates it into the iterative generative procedure.

Specifically, the generative procedure of HGGDPRrec for MRI reconstruction is a homotopic process: at the iterative preparation stage, we generate a list of noise levels \( \{\sigma_i\}_{i=1}^l \) that are reduced proportionally for each step of the outer loop. At the iteration stage, we add noise according to the noise level \( \sigma_i \), from large to small at the iterative reconstruction result of the outer loop, which helps smoothly transfer the benefits of large noise levels to low noise levels where the perturbed data are almost indistinguishable from the original ones. In the meantime, we incorporate the data consistency into the iterative reconstruction gradient update. Finally, at \( \sigma_l \), its impact can no longer be recognized by the naked eye, HGGDPRrec has reached perfect quality.

It should be emphasized that the annealed/homotopic strategy used in HGGDPRrec has a long history in severely non-convex optimization for reconstruction. For instance, in [50], Trzasko et al. aimed to attain highly undersampled MRI reconstruction via homotopic L0 minimization. In [51], Wong et al. extended the homotopic L0 minimization scheme for regional sparsified domain. In those works, they continually reduce the parameter in the regularization term to pursue the global solution in non-convex minimization. Here, we adopt a similar strategy in the case of statistical distribution learning to minimize the non-smooth and non-convex functional \( \nabla_x \log p_{\text{data}}(X) \).

Specifically, at each iteration of the annealed Langevin dynamics, we update the solution via data consistency constraint after Eq. (14), i.e., let \( \hat{x} = \text{Mean}(X) \), it yields

\[
\begin{equation}
\arg\min_x \|y - F_x \hat{x}\|^2 + \lambda \|x - \hat{x}\|^2
\end{equation}
\]

The least-square (LS) minimization in Eq. (15) can be solved as follows:

\[
(F_x^T F_x + \lambda I) x^* = F_x^T y + \lambda \hat{x}
\]

Let \( F \in \mathbb{C}^{M \times H} \) denote the full Fourier encoding matrix which is normalized as \( F^T F = I_M \). \( F(x) \) stands for the updated value at under-sampled K-space location \( k \), and \( \Omega \) represents the sampled subset of data, it yields

\[
F(x) = \begin{cases} 
F(k), & k \notin \Omega \\
FF^T y(k) + \lambda Fx(k), & k \in \Omega 
\end{cases}
\]

In summary, as explained in Algorithm 1, the whole MRI reconstruction procedure consists of two-level loops: The outer loop handles \( \nabla_x \log p_{x}(X) \) to approximate \( \nabla_x \log p_{\text{data}}(X) \), while the inner loop decouples to be an alternating process of the estimated gradient of data prior \( \nabla_x \log p_{x}(X) \) and the LS scheme. Detailed flowchart of HGGDPRrec is shown in Fig. 7.

| Algorithm 1 HGGDPRrec |
|-----------------------|
| **Training stage** |
| Dataset: Multi-channel dataset: \( X = \{x_1, x_2, x_3, \ldots x_N\} \) |
| Outputs: Trained HGGDP \( S_h(X, \sigma) \) |
| **Reconstruction stage** |
| Setting: \( \sigma \in \{\sigma_1, \sigma_2, \ldots, \sigma_l\} \) and \( \lambda \) |
| 1: for \( \ell \leftarrow 1 \) to \( L \) do (Outer loop) |
| 2: \( \sigma_\ell = \sigma / \sigma_{\ell - 1} \) |
| 3: for \( t \leftarrow 1 \) to \( T \) do (Inner loop) |
| 4: Draw \( z \sim N(0,1) \) and \( X' = [x_1', x_2', \ldots x_N'] \) |
| 5: \( X' = X' + \sigma / 2 S_h(X'_{\ell, \sigma}) + \sqrt{1 / \sigma} z \) |
| 6: Update \( x^* = \text{Mean}(X') \) and Eq. (17) |
| 7: end for |
| 8: \( \hat{x} \leftarrow x^* \) |
| 9: end for |
| **Return** \( \hat{x} \) |

IV. EXPERIMENTS

A. Experiment Setup

1) Datasets: The experiments were performed with brain images from SIAT to demonstrate the generalization ability and robustness of HGGDPRrec. SIAT dataset was provided
by Shenzhen Institutes of Advanced Technology, the Chinese Academy of Science. Informed consents were obtained from the imaging subject in compliance with the institutional review board policy. The raw data were acquired from 3D turbo spin-echo (TSE) sequence with T2 weighting by a 3.0T whole-body MR system (SIEMENS MAGNETOM Trio Tim), which has 192 slices per slab, and the thickness of each slice was 0.86 mm. Typically, the field of view and voxel size were set to be 220×220 mm$^2$ and 0.9×0.9×0.9 mm$^3$, respectively. Moreover, the number of coils is 12 and the collected dataset includes 500 2D complex-valued MR images. Affine transformation was adopted for data augmentation, and 8000 patches are obtained for training by slicing the enhanced image into 64×64. At the reconstruction stage, we use a variety of sampling schemes on another 31 complex-valued MR images.

2) **Model Training**: During the learning phase, we use fully sampled MR images as the network input and disturb it simultaneously via random Gaussian noise of various amplitudes. Notice that the input and output are all complex-valued images with the same size and each includes real and imaginary components. Additionally, RefineNet [39] with instance normalization, dilated convolutions and U-Net-type architectures[52] was selected as the network structure. Adam was selected as an optimizer with the learning rate 0.005, which was halved every 5000 iterations. Subsequently, the HGGDP model was trained after 1e5 iterations with the batch size of 32 that took around 20 hours, which is performed with Pytorch interface on 2 NVIDIA Titan XP GPUs, 12GB RAM.

3) **Sampling Masks**: The sampling in the Fourier domain is realized using three different undersampling strategies, namely 2D random, pseudo radial and 1D cartesian sampling. Specifically, the accelerated factors are varied over five values, $R = 3.3, 4, 5, 6.7, 10$. Fig. 9 displays 3 sampling masks and some fully-sampled representative MR images.

4) **Compared Methods**: Several state-of-the-art CS-MRI methods were implemented to compete against our proposed method. For example, patch-based DLMRI [53], reference-derived sparse representation method PANO [54], dictionary learning method FDLCP [55], low-rank-based NLR-CS [56], end-to-end DC-CNN [17], enhanced denoising autoencoder based EDADPRec [29] and NCSNRRec [34]. In the meanwhile, we quantified the reconstruction performance by peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and high-frequency error norm (HFEN). For convenient reproducibility, source code is available at: https://github.com/yqx7150/HGGDP.

![Fig. 8. Pipeline of HGGDP training process for prior learning and HGGDPRRec procedure for MRI reconstruction.](image)

![Fig. 9. Representative sampling masks and test images.](image)

**B. Reconstruction Comparisons**

Quantitative PSNR/SSIM/HFEN results of the above experiments with various sampling rates and patterns are summarized in Table I. Intuitively, the average results of 31 test images obtained by HGGDPRec are more than 2 dB over the naive NCSNRRec. Furthermore, HGGDPRec yields the highest values in the most majority of the sampling rates. For example, under the accelerated factors $R = 4, 5$, the highest PSNR values achieved by all the compared methods are 34.49 dB and 33.49 dB, which are lower than the values of 35.55 dB and 34.40 dB obtained by HGGDPRec.

Besides of the quantitative comparison, the visual quality is also highlighted. As can be seen in Figs. 10-11, visual quality of reconstructions for different methods varies.
DLMRI and NLR-CS reconstructions have limitations in recovering the structure and texture. FDLCP utilizes the similarity and geometric orientation of the patch, which contains more details than DLMRI. In order to further compare visual quality, the zoom-in images and error maps are illustrated through the green and red boxes on the screen. The visual quality indicates that EDAEPRec is better than FDLCP, PANO and DC-CNN methods in terms of more abundant recovered edge details. However, it still suffers from some undesirable artifacts and loses details such as ringing, jaggy and staircase artifacts. To sum up, HGDDPRec can achieve more satisfactory results with clearer contours, sharper edges, and finer image details under various sampling masks.

| Table I |
| --- |
| RECONSTRUCTION PSNR, SSIM AND HFVEN VALUES OF THREE TEST IMAGES AT VARIOUS SAMPLING TRAJECTORIES AND UNDER SAMPLING PERCENTAGES. |
| (a) | DLMRI | PANO | FDLCP | NLR-CS | DC-CNN | RefineGAN | MEDSMPRec | EDAEPRec | NSCSRec | HGDDPRec |
| R=3, 4, 5 | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Radial | 0.61 | 0.87 | 0.77 | 0.92 | 0.83 | 0.94 | 0.86 | 0.95 | 0.88 | 0.96 | 0.90 | 0.98 |
| (b) | DLMRI | PANO | FDLCP | NLR-CS | DC-CNN | RefineGAN | MEDSMPRec | EDAEPRec | NSCSRec | HGDDPRec |
| R=6, 7 | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Radial | 0.61 | 0.87 | 0.77 | 0.92 | 0.83 | 0.94 | 0.86 | 0.95 | 0.88 | 0.96 | 0.90 | 0.98 |

C. Ablation Study

We discuss the impact of several factors on HGDDPRec reconstruction: Size of $I$ in $\{\sigma_I\}_I$, number of channels, size of training patch, and initial values.

First, we use disturbance data with $\sigma$ level random Gaussian noise as input in the HGDDPRec to train the network. Here, we set 4 groups of noises with $I = 8, 10, 12, 15$ as comparative experiments to study (e.g. the same proportional sequence with 1-0.01 but the different ratio).

From Table II, it can be seen that the PSNR, SSIM, and HFVEN under $I = 10, 12, 15$ are much higher than $I = 8$. It is unsurprising that the more prior information and the higher quality of the image reconstruction may be obtained as the size of $I$ increases. Especially, the PSNR result obtained by HGDDPRec becomes stable after $I = 10$.

Second, the influence with the number of channels at network input on the reconstruction results is investigated. The number of channels to be 2, 3 and 4 is compared in the experiment. It can be seen from Table III that setting to be 3 is the best choice.

Third, HGDDP is trained on image patches with different size, which are 32, 64, and 96, respectively. When generating patch, we randomly flip the patch to obtain richer and more gradient of prior information. As can be seen from Table IV, when patch size is set to be 64, the best results are obtained.

Finally, we investigate how different initial values would affect the efficacy of the proposed method to reconstruct MR images. To this end, two different initializations are used respectively, namely initializing with the uniform noise $x \sim N(-1,1)$ and zero-filled data. Table V presents the results produced by our method with two initializations. The results gained by HGDDPRec are almost the same regardless of initializations, which empirically indicates that our method is insensitive to initialization.

| Table II |
| --- |
| PERFORMANCE OF RECONSTRUCTING 31 TEST IMAGES BY DIFFERENT SIZE OF $I$ AT RADIAL SAMPLING TRAJECTORIES. |
| $I$ | 8 | 10 | 12 | 15 |
| --- | --- | --- | --- | --- |
| R=4 | PSNR | 38.47 | 35.55 | 35.47 | 35.48 |
| SSIM | 0.9240 | 0.9287 | 0.9243 | 0.9233 |
| HFVEN | 0.51 | 0.51 | 0.51 | 0.50 |
| R=10 | PSNR | 30.70 | 30.70 | 30.70 | 30.70 |
| SSIM | 0.8289 | 0.8488 | 0.8379 | 0.8385 |
| HFVEN | 1.56 | 1.29 | 1.32 | 1.28 |

| Table III |
| --- |
| PERFORMANCE OF RECONSTRUCTING 31 TEST IMAGES BY DIFFERENT NUMBER OF CHANNELS AT RADIAL SAMPLING TRAJECTORIES. |
| Channel number | 2 | 3 | 4 |
| --- | --- | --- | --- |
| R=4 | PSNR | 35.22 | 35.55 | 35.14 |
| SSIM | 0.9277 | 0.9289 | 0.9265 |
| HFVEN | 0.53 | 0.53 | 0.52 |
| R=10 | PSNR | 30.69 | 30.88 | 30.37 |
| SSIM | 0.8466 | 0.8488 | 0.8403 |
| HFVEN | 1.31 | 1.29 | 1.37 |

| Table IV |
| --- |
| PERFORMANCE OF RECONSTRUCTING 31 TEST IMAGES BY DIFFERENT SIZE OF TRAINING PATCHES AT RADIAL SAMPLING TRAJECTORIES. |
| Patch size | 32 | 64 | 96 |
| --- | --- | --- | --- |
| R=4 | PSNR | 34.93 | 35.55 | 35.48 |
| SSIM | 0.9253 | 0.9288 | 0.9282 |
| HFVEN | 0.57 | 0.57 | 0.52 |
| R=10 | PSNR | 29.91 | 30.88 | 30.89 |
| SSIM | 0.8345 | 0.8488 | 0.8485 |
| HFVEN | 1.49 | 1.29 | 1.3 |
As the cost time is an important factor in the reconstruction process, the iteration number $T$ of the inner loop is investigated. In Table VI, it can be found that the reconstruction quality had no obvious change after $T = 60$. As the acceleration factor $R$ increase, reconstruction with larger $T$ is needed to attain the best result.

**Fig. 10.** Reconstruction comparison on pseudo radial sampling at acceleration factor $R = 5$. Top: Reference, reconstruction by DLMRI, PANO, FDLCP. Bottom: Reconstruction by NLR-CS, DC-CNN, EDAEPRec, HGGDPRRec. Green and red boxes illustrate the zoom in results and error maps, respectively.

**Fig. 11.** Reconstruction comparison on 2D Random sampling at acceleration factor $R = 6.7$. Top: Reference, reconstruction by DLMRI, PANO, FDLCP. Bottom: Reconstruction by NLR-CS, DC-CNN, EDAEPRec, HGGDPRRec. Green and red boxes illustrate the zoom in results and error maps, respectively.

| Initial value | $R=3.3$ | $R=4$ | $R=5$ | $R=10$ |
|---------------|---------|-------|-------|--------|
| Uniform Noise | PSNR: 36.18, SSIM: 0.9364, HFEN: 0.42 | 35.55, 0.9288, 0.50 | 34.40, 0.9130, 0.65 | 30.88, 0.8488, 1.29 |
| $\chi \sim N(-1,1)$ | PSNR: 36.09, SSIM: 0.9307, HFEN: 0.44 | 35.44, 0.9215, 0.52 | 34.29, 0.9032, 0.68 | 30.68, 0.8319, 1.35 |

**TABLE VI**

| Initial value | $T=20$ | $T=40$ | $T=60$ | $T=80$ | $T=100$ |
|---------------|--------|--------|--------|--------|--------|
| $R=4$ | PSNR: 34.79, SSIM: 0.9248, HFEN: 0.61 | 35.37, 0.9276, 0.53 | 35.51, 0.9283, 0.50 | 35.54, 0.9286, 0.50 | 35.55, 0.9287, 0.50 |
| $R=10$ | PSNR: 28.49, SSIM: 0.8011, HFEN: 1.94 | 30.35, 0.8416, 1.43 | 30.70, 0.8459, 1.34 | 30.83, 0.8478, 1.30 | 30.85, 0.8488, 1.29 |

**V. CONCLUSIONS**

In this paper, a homotopic gradient of generative density prior was proposed. The generative modeling scheme first estimates the DSM and then employs Langevin diffusion for sampling. The proposed generative framework leverages the gradient of data as prior and improves the native NCSN for high diagnostic-quality image reconstruction. The present method involves two main characteristics: higher-dimensional and homotopic iteration. These factors jointly avert the data confined to low data density regions and avoid difficulties from the manifold hypothesis. Comprehensive experiment results demonstrate that HGGDPRRec achieved superior performance. In the forthcoming future, more ways to form high-dimensional space will be exploited and imaging modalities will be applied to validate its efficiency, such as CT and PET reconstruction.

**REFERENCES**

[1] D.L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.
[2] G. Yang, S. Yu, H. Dong, G. Slabaugh, P.L. Dragotti, X. Ye et al., “DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction,” *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310-1321, Dec. 2017.
[3] Z. Lai, X. Qu, Y. Liu, D. Guo, J. Ye, Z. Zhan et al., “Image reconstruction of compressed sensing MRI using graph-based redundant wavelet transform,” *Med. Image Anal.*, vol. 27, pp. 93–104, 2016.
[4] M. Lustig, D. Donoho, and J.M. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182-1195, Dec. 2007.
[5] J. Yang, Y. Zhang, and W. Yin, “A fast alternating direction method for TVL1-L2 signal reconstruction from partial fourier data,” *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 288–297, Apr. 2010.
Abbye works for high-resolution reconstruction in biomedi,
generative adversarial networks for denoising autoencoders,
"Extracting and composing robust features with denoising autoencoders," Proc. Int. Conf. Learn. Represent., pp. 1096-1103, 2008.

P. Vincent, "A connection between score matching and denoising autoencoders," Neural Comput., vol. 23, no. 7, pp. 1661-1674, May 2011.

Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," Proc. Adv. Neural Inf. Process. Syst., pp. 11895-11907, 2019.

M.Y. Azar, V. Varadarajan, L. Hamey, and U. Tupakula, "Autoencoder-based feature learning for cyber security applications," Proc. IEEE Int. Joint Conf. Neural Netw., pp. 3854-3861, Jul. 2017.

S. Wang, Z. Ding, and Y. Fu, “Feature selection guided auto-encoder,” Proc. AAAI Conf. Artif. Intell., 2017.

S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu et al., “Accelerating magnetic resonance imaging via deep learning,” IEEE Trans. Int. Symp. Biomed. Imag., pp. 514-517, 2016.

J. Schlemper, J.V. Hajnal, A. Price, and D. Rueckert, “A deep cascade of convolutional neural networks for dynamic MR image reconstruction,” IEEE Trans. Med. Imag., vol. 37, no. 2, pp. 491-503, Feb. 2018.

G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1925-1934, 2017.

S.A. Bigdeli, and M. Zwick, “Image restoration using autoencoding priors,” arXiv preprint :arXiv:1703.09964, 2017.

S.A. Bigdeli, M. Zwick, P. Favaro, and M. Jin, “Deep mean-shift priors for image restoration,” Proc. Adv. Neural Inf. Process. Syst., pp. 763-772, 2017.

A. Block, Y. Mroueh, and A. Rakhlin, “Generative modeling with denoising auto-encoders and Langevin sampling,” arXiv preprint :arXiv:2002.00107, 2020.

Z. Li, Y. Chen, and F.T. Sommer, “Annealed denoising score matching: Learning energy-based models in high-dimensional spaces,” arXiv preprint :arXiv:1910.07762, 2019.

S. Saremi and A. Hyvarinen, “Neural empirical bays,” arXiv preprint :arXiv:1903.02334, 2019.

Y. Bengio, A. Courville, and P. Vincent, “Reprentation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

D.P. Kingma, P. Dhariwal, “Glow: Generative flow with invertible 1×1 convolution,” Proc. Adv. Neural Inf. Process. Syst., pp. 10215-10224, 2018.

D.P. Kingma, and M. Welling, “Auto-encoding variational bayes,” arXiv preprint :arXiv:1312.6114, 2013.

H. Narayanan and S. Mitter, “Sample complexity of testing the manifold hypothesis,” Proc. Adv. Neural Inf. Process. Syst., 2010.

S. Riffai, Y. Dauphin, P. Vincent, Y. Bengio, and X. Muller, “The manifold tangent classifier,” Proc. Adv. Neural Inf. Process. Syst., 2011.

J. Trzasko and A. Manduca, “Highly undersampled magnetic resonance image reconstruction via homotopic L0 minimization,” IEEE Trans. Med. Imag., vol. 28, no. 1, pp. 106-121, Jul. 2009.

A. Wong, A. Mishra, P. Fieguth, and D.A. Clausi, “Sparse reconstruction of Breast MRI using homotopic L0 minimization in a regional sparsified domain,” IEEE Trans. Biomed. Eng., vol. 60, no. 3, pp. 743-752, Oct. 2013.

O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” Proc. Int. Conf. Med. Image Comput. Assist. Intervent., pp. 234-241, Oct. 2015.

S. Ravivshankar and Y. Bresler, “MR image reconstruction from highly undersampled k-space data by dictionary learning,” IEEE Trans. Med. Imag., vol. 30, no. 5, pp. 1028-1041, Nov. 2011.

X. Qu, Y. Hou, F. Lam, D. Guo, J. Zhong, and Z. Chen, “Magnetic resonance image reconstruction from undersampled measurements using a patch-based nonlocal operator,” Med. Image Anal., vol. 18, no. 6, pp. 843-856, Aug. 2014.

B. Wohlberg, “Efficient algorithms for convolutional sparse representations,” IEEE Trans. Image Process., vol. 25, no. 1, pp. 301-315, Oct. 2016.

W. Dong, G. Shi, X. Li, Y. Ma, and F. Huang, “Compressive sensing via nonlocal low-rank regularization,” IEEE Trans. Image Process., vol. 23, no. 8, pp. 3618-3632, Jun. 2014.
Supplementary Material

Supp-1: The equivalence between DAE and DSM.

Theorem S-1. The DAE loss and the DSM loss are equivalent up to a term that does not depend on \( A \) or \( S \).

Proof. Let \( y = x + \eta \), \( \eta = g_{x_{\sigma}} \), \( x \sim p \), then we have,
\[
L_{\text{DAE}}(A) = E_{\nu_{y_{\eta_{\sigma}}}}[d(A(x+\eta)-x)]^2
= \int [d(A(y)-y+\eta)p(y-\eta)g(\eta)d\eta]
= \int [d(A(y)-y)p(y)g(y)d\eta]+\int [2\eta.A(y)-y]p(y-\eta)g(\eta)d\eta
+\int d[\eta]p(y-\eta)g(\eta)d\eta
\]
The last term above does not depend on \( A \) and so we may ignore it. We focus now on the second term. Let \( \xi = g_{x_{\sigma}} \) be a standard Gaussian and let \( s'(x) = A(x)-x \), we have
\[
\int \langle s',s'(x+\eta)\rangle g(\eta)d\eta = \sigma^2 \sigma^2 [\int \langle \xi',s'(x+\sigma\xi)\rangle g_{1}(\xi)d\xi]
= \sigma^2 \sigma^2 [\int \langle \xi',s'(x+\sigma\xi)\rangle g_{1}(\xi)d\xi] = \sigma^2 \sigma^2 \text{div} \int s'(x+\eta)g(\eta)d\eta
\]
where the Gaussian derivative definition is used above. Now, note that as we know that \( p_{\sigma} \) is a density, it must tend to zero as \( \|x\| \to \infty \). Thus we may apply the divergence theorem to get
\[
\int [2\eta.A(y)-y]p(y-\eta)g(\eta)d\eta = 2\sigma^2 \sigma^2 \int p(x)\text{div} (E_{x_{\sigma_{\sigma}}} [s'(x+\eta)])dx
= 2\sigma^2 \sigma^2 \int p(x)\text{div} (s'(x))p_{\sigma}(x)dy
= -2\sigma^2 \sigma^2 \int p_{\sigma}(x)dy
\]
This thus we have that
\[
L_{\text{DAE}}(A) = E_{\nu_{y_{\eta_{\sigma}}}}[A(x)-x]^2 - 2\sigma^2 \sigma^2 \int s'(x)\text{V} \log p_{\sigma}(x)] + C(p,\sigma)
= E_{\nu_{y_{\eta_{\sigma}}}}[s'(x) - \sigma^2 \text{V} \log p_{\sigma}(x)] + C(p,\sigma)
= E_{\nu_{y_{\eta_{\sigma}}}}[s'(x)\sigma^2 \text{V} \log p_{\sigma}(x)] + C(p,\sigma)
\]
where \( C(p,\sigma) \) does not depend on \( A \). Dividing by \( \sigma^2 \) and setting \( S(x)=s'(x)/\sigma^2 \) shows that
\[
L_{\text{DSM}}(S) = E_{\nu_{y_{\eta_{\sigma}}}}[S(x) - \text{V} \log p_{\sigma}(x)]^2 = \frac{1}{\sigma^2} E_{\nu_{y_{\eta_{\sigma}}}}[S(x) - \text{V} \log p_{\sigma}(x)]^2 + C(p,\sigma)
\]
Thus, the two losses are equivalent to minimize with respect to \( A \) or \( S \).

Supp-2: The proof of Eq. (8).

Theorem S-2. The following two optimization objectives
\[
J_{\text{EMDO}}(\theta) = E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} S_{\eta_{\sigma}}(\hat{x}) - \frac{\partial \log p_{\sigma}(\hat{x})}{\partial \hat{x}}]^2
\]
and
\[
J_{\text{DSMO}}(\theta) = E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} S_{\eta_{\sigma}}(\hat{x}) - \frac{\partial \log p_{\sigma}(\hat{x})}{\partial \hat{x}}]^2
\]
are equivalent up to a term that does not depend on \( \theta \).

Proof. Eq. (S6) can be rewritten as
\[
J_{\text{EMDO}}(\theta) = E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} S_{\eta_{\sigma}}(\hat{x})]^2 - G(\theta) + C_2
\]
where \( C_2 = E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} \partial \log p_{\sigma}(\hat{x})]^2 \) is a constant that does not depend on \( \theta \) and
\[
G(\theta) = E_{\nu_{y_{\eta_{\sigma}}}}[\int S_{\eta_{\sigma}}(\hat{x}) \frac{\partial \log p_{\sigma}(\hat{x})}{\partial \hat{x}} d\hat{x}]
\]
Substituting this expression for \( G(\theta) \) in Eq. (S8) yields
\[
J_{\text{EMDO}}(\theta) = E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} S_{\eta_{\sigma}}(\hat{x})]^2 - E_{\nu_{y_{\eta_{\sigma}}}}[\int S_{\eta_{\sigma}}(\hat{x}) \frac{\partial \log p_{\sigma}(\hat{x})}{\partial \hat{x}}] + C_2
\]
\[
= E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} S_{\eta_{\sigma}}(\hat{x})]^2 - \frac{\partial \log p_{\sigma}(\hat{x})}{\partial \hat{x}}] + C_3
\]
\[
= J_{\text{DSMO}}(\theta) + C_3
\]
where \( C_3 = C_2 - E_{\nu_{y_{\eta_{\sigma}}}}[\frac{1}{2} \partial \log p_{\sigma}(\hat{x})]^2 \) is a constant that does not depend on \( \theta \). Thus, it is shown that the two optimization objectives are equivalent.

Supp-3: The definition of \( R_n(G) \)
Let \( G \) be a class of real valued functions on \( \mathbb{R}^d \) and let \( S=(x_1,\cdots,x_n) \) be \( n \) samples from \( \mathbb{R}^d \). We define the Rademacher average with respect to the sample as
\[
\hat{R}_n(G,S) = E_{\nu_{\eta_{\sigma}}}[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i)]
\]
where \( \epsilon_i \) are i.i.d random variables taking values \( \{\pm 1\} \) with probability \( \frac{1}{2} \) each. We define the Rademacher com-
plexity of the function class $\mathcal{G}$ as $$\mathcal{R}_n(\mathcal{G}) = \sup_{\mathcal{S} \subset \{\mathbb{R}^d\}} \hat{\mathcal{R}}_n(\mathcal{G}, \mathcal{S})$$ (S11)

For a class of $\mathbb{R}^k$-valued functions $\mathcal{G}$, we denote by $$\mathcal{R}_n(\mathcal{G}) = \sum_{i=1}^{k} \mathcal{R}_n(\mathcal{G}_i)$$ where $\mathcal{G}_i$ is the restriction of $\mathcal{G}$ to the $i$-th coordinate.