Self-supervised Deep Image Restoration via Adaptive Stochastic Gradient Langevin Dynamics

Weixi Wang, Ji Li, and Hui Ji
Department of Mathematics, National University of Singapore, 119076, Singapore
wangweixi@u.nus.edu, matliji@nus.edu.sg, matjh@nus.edu.sg

Abstract

While supervised deep learning has been a prominent tool for solving many image restoration problems, there is an increasing interest on studying self-supervised or unsupervised methods to address the challenges and costs of collecting truth images. Based on the neuralization of a Bayesian estimator of the problem, this paper presents a self-supervised deep learning approach to general image restoration problems. The key ingredient of the neuralized estimator is an adaptive stochastic gradient Langevin dynamics algorithm for efficiently sampling the posterior distribution of network weights. The proposed method is applied on two image restoration problems: compressed sensing and phase retrieval. The experiments on these applications showed that the proposed method not only outperformed existing non-learning and unsupervised solutions in terms of image restoration quality, but also is more computationally efficient.

1. Introduction

Image restoration is about calculating an image $\boldsymbol{x}$ from a collection of its measurements, denoted by $\boldsymbol{y}$, whose relationship can be described as

$$\boldsymbol{y} = \Psi(\boldsymbol{x}) + \boldsymbol{n}, \tag{1}$$

where Ψ denotes the image formation process and $\boldsymbol{n}$ denotes measurement noise. Image restoration is one fundamental problem encountered in a wide range of image-related applications. For example, image restoration in digital photography, compressed sensing, computed tomography (CT), and magnetic resonance imaging (MRI) in medical imaging, phase retrieval for scientific imaging, and many others. In general, the problem (1) is an ill-posed inverse problem, whose direct inversion is either not unique or sensitive to measurement noise.

Over last decades, regularization method, or equivalently Bayesian estimator, has been the dominant tool for image restoration. These non-learning methods either impose certain prior or assume certain prior distribution on images for addressing solution ambiguity and noise sensitivity However, it remains challenging to define an accurate image prior. In recent years, deep neural network (DNN) emerges as a a prominent tool for solving inverse problems; see *e.g.* [11, 12, 16, 17, 32, 38, 49, 56, 58, 60, 61]. The majority of existing DNN-based solutions are supervised on an external training dataset with truth images. Such a prerequisite on truth images limits their wider applications in practice. For example, collecting truth images can be very challenging and costly in medical imaging and scientific imaging. Also, the generalization performance of a supervised learning method can be a concern in practice, if the network is trained over a biased dataset where the structures of test images are not present in training samples.

In recent years, it is receiving an increasing interest on developing deep learning methods for imaging, which do not require truth images for training DNNs. The so-called *plug-and-play* prior attempts to address such an issue by adopting some pre-trained denoising network in an iterative image restoration scheme; See *e.g.* [35, 48, 54, 62]. While these methods do not explicitly call truth images, the pre-trained networks are still supervised over the dataset with truth images related to the image for restoration. Another approach is using *generative adversarial network* (GAN) to synthesize training samples for training the network; see *e.g.* [43]. Similarly, the performance of GAN-based methods highly relies on the effectiveness of the pre-trained GAN model on simulating truth images. While GAN has been very effective on simulating the images in specific domain such as face and text, it is not so for other types of images, *e.g.*, medical images and scientific images.

The methods above are not completely free from the prerequisite of accessing related truth images. Recently, there has been a rapid progress on unsupervised or self-supervised learning for image denoising using un-trained DNNs; See *e.g.* [3, 14, 27, 40, 41, 50, 53]. However, the generalization of these denoising networks to ill-posed image restoration problems is not trivial. The existence of the non-

zero null space

$$\text{Null}(\Psi) = \{\boldsymbol{x} : \Psi(\boldsymbol{x}) = \boldsymbol{0}\} \neq \{\boldsymbol{0}\}$$

in image restoration make it quite different from image denoising, as the error induced by the existence of non-trivial space $\text{Null}(\Psi) \neq \{\boldsymbol{0}\}$ cannot be treated as random noise. One pioneering work is the so-called deep image prior (DIP) [53] which shows that there exists certain implicit prior induced by a convolutional neural network (CNN). The DIP states that regular image structures appear before random noise when training a CNN. Such a prior has been exploited in many image restoration tasks, *e.g*. super-resolution [53], CT reconstruction [22], image separation [21] and blind deblurring [31, 44]. In addition to DIP, there are some other approach to address the overfitting caused by the absence of truth images during training. For CS image reconstruction, Pang *et al*. [39] proposed to train a Bayesian DNN with Gaussian random weights. Heckle [23] used an under-parametrized network. Metzler *et al*. [34] and Zhussip *et al*. [66] proposed to regularize the denoising network in an iterative scheme by Stein's unbiased risk estimator (SURE).

1.1. Motivation and main idea

A self-supervised deep learning method for image restoration is very attractive to the applications where collecting truth images is challenging, *e.g*. medical and scientific imaging. This paper is about studying an efficient and effective method for training a NN to process testing data, where neither pre-trained model nor training sample with truth image is used during training.

The proposed method is derived from the so-called *minimum mean squared error* (MMSE) estimator of the problem (1) defined by

$$\widehat{\boldsymbol{x}} = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}, \tag{2}$$

where $p(\boldsymbol{x}|\boldsymbol{y})$ denotes the posterior distribution. One way to calculate (2) is using the Monte Carlo (MC) method. Instead of directly sampling $p(\boldsymbol{x}|\boldsymbol{y})$, we use a generative CNN $f(\epsilon_0; \theta)$, parameterized by θ, to re-parametrize $\boldsymbol{x}$:

$$\boldsymbol{x} = f(\epsilon_0; \theta).$$

Then, $\widehat{\boldsymbol{x}}$ in (2) can be re-expressed as

$$\widehat{\boldsymbol{x}} = \int f(\epsilon_0; \theta) p(\theta|\boldsymbol{y}, \epsilon_0) d\theta. \tag{3}$$

Such a re-parametrization allows us to utilize implicit image prior induced by CNN. Then, the remaining task is how to efficiently sample the posterior distribution

$$\pi(\theta) = p(\theta|\boldsymbol{y}, \epsilon_0),$$

to calculate the integral (3) via the MC method:

$$\widehat{\boldsymbol{x}} \approx \sum_k f(\epsilon_0; \theta_k), \quad \text{where } \theta_k \sim p(\theta_k|\boldsymbol{y}, \epsilon_0).$$

Efficient sampling restricted in feasible set. How to efficiently sample θ is critical for an accurate calculation of the integral (3). A natural treatment is, instead of sampling θ in the whole space, we only sample those parameters in a feasible set Ω, where the density function $\pi(\theta)$ concentrates. Suppose measurement noise $\boldsymbol{n}$ is i.i.d. Gaussian white noise with variance σ^2. By large number theory, we have, for image size $N \to \infty$,

$$L(\theta) = \frac{1}{N}\|\Psi(f(\epsilon_0; \theta)) - \boldsymbol{y}\|_2^2 = \frac{1}{N}\|\Psi(\boldsymbol{x}) - \boldsymbol{y}\|_2^2 \to \sigma^2.$$

In other words, with sufficiently large image size, the density of θ whose $f(\epsilon_0; \theta)$ is close to $\boldsymbol{x}$, concentrates within the set:

$$\Omega_\epsilon := \{\theta : \sigma^2 - \epsilon \leq L(\theta) \leq \sigma^2 + \epsilon\}, \tag{4}$$

where ϵ is a small threshold. A detailed analysis of (4) is provided in the supplementary file. To conclude, the samples from $\pi(\theta)$ within the feasible set Ω_ϵ defined by (4) are sufficient for accurately calculating the integral (3).

Adaptive SGLD for restricted MC sampling. SGLD is an Markov chain Monte Carlo (MCMC) sampling algorithm, which is proposed in [55] for efficiently sampling network weights. SGLD simulates dynamics of molecular systems with stochastic differential equation given by $d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2}dW_t$, where W_t is stationary Gaussian process with zero-mean and L is a loss function.

In the context of DNN, SGLD is proposed in [55] for efficiently sampling network weights. SGLD is a Markov chain Monte Carlo (MCMC) sampling technique, which simulates dynamics of molecular systems with stochastic differential equation given by $d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2}dW_t$, where W_t is stationary Gaussian process with zero-mean and L is a loss function.

In this paper, we proposed a new type of SGLD for effectively sampling from $\pi(\theta) = p(\theta|\boldsymbol{y}, \epsilon_0)$ within the feasible set Ω_ϵ defined by (4). The corresponding stochastic differential equation is defined by

$$d\theta_t = -\nabla L(\theta_t)dt + \beta \exp(c_0(\frac{\sigma^2}{L(\theta_t)} - 1))dW_t,$$

where $L(\theta)$ is defined by (1.1), in the case where $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 I)$ and θ follows an uniform distribution. Then the discretization of the equation above leads to an adaptive stochastic gradient Langevin dynamics (ASGLD):

$$\theta_{k+1} = \theta_k - \gamma_k \cdot \nabla L(\theta_k) + \beta \exp(c_0(\frac{\sigma^2}{L(\theta_k)} - 1))\sqrt{\gamma_k} \cdot \epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, I)$. It can be seen that in comparison to

classic SGLD, ASGLD adaptively adjust the magnitude of noise perturbation based on the loss function $L(\theta)$. The samples from ASGLD will concentrate within the feasible set Ω_ϵ. See Section 3 for more details.

1.2. Main contribution

This paper proposed a self-supervised method for solving ill-posed image restoration problems, without requiring any external truth image. The method trains a CNN which approximates the MMSE estimator of the problem via MC-based integration, where the key is how to efficiently sample the posterior distribution. The answer from this paper is an adaptive SGLD method, an efficient MCMC scheme that focuses on concentrated regions of the posterior distribution.

The proposed method is applied to solve image restoration problems in two imaging modalities: CS and phase retrieval. The experiments show that the proposed method not only provides state-of-the-art performance among all existing dataset-free solutions, but also is more computationally efficient. See below for the summary of the differences between ASGLD and most related unsupervised methods.

- ASGLD vs. DIP [53] and its extensions: DIP utilizes the implicit prior induced by a CNN and trains the network with early-stopping. ASGLD also training a network and its algorithm is motivated by MCMC sampling based approximation to the MMSE estimator of the problem.
- ASGLD vs. BNN [39]: Both train a network to approximate the MMSE estimator of the problem. BNN approximates it via variation approximation using a Bayesian neural network (BNN) with random weights. ASGLD approximates it using MC-sampling-based integration implemented using an efficient MCMC sampler.
- ASGLD vs. plain SGLD : In comparison to classic SGLD with constant noise variance for general MCMC sampling, ASGLD proposes a new SGLD scheme with adaptive noise variance, which enables one to efficiently calculate the MC-based integration.

See below for the summary of main contributions:

- An MC-sampling-based MMSE estimator of image restoration with untrained deep network.
- A new adaptive SGLD scheme of MCMC sampling for efficient calculation of MC-based integration.
- Noticeable performance improvement over existing unsupervised methods in two image restoration tasks.
- A general self-supervised method with potential applications to other ill-posed inverse problems in imaging.

2. Related work

There is an abundant literature on inverse problems. Due to space limitation, we only cover the most related ones.

Regularization methods with manually-defined prior. Regularization has been one prominent method for image restoration, which imposes a pre-defined prior on the latent image to address the ill-posed-ness of the problem. Most regularization methods can be viewed as an MAP estimator which minimizes $\min_{\boldsymbol{x}} -\log p(\boldsymbol{x}|\boldsymbol{y}) = \min_{\boldsymbol{x}} -\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{x})$, where $p(x)$ denotes prior distribution of x. Different priors on x are proposed in the past, including Gaussian prior on $\boldsymbol{x}$, Laplacian prior on $\nabla\boldsymbol{x}$ [6, 9, 10, 47] or related measurements such as wavelet coefficients [5, 18].

Supervised learning methods. Supervised deep learning has been used for solving a wide range of image restoration problems. One approach is to train a network on a dataset with many pairs $\{(\boldsymbol{y}, \boldsymbol{x})\}$ that maps input measurements to the image for restoration; see *e.g.* [12, 13, 51, 63, 64]. Such an approach works well for denoising, but it does not utilize the information of the follow process Ψ. To exploit the information of Ψ in the network, the so-called optimization unrolling approach unrolls some iteration schemes and replaces the prior-relating operations by a denoising CNN with learnable parameters; see *e.g.* [1, 15, 17, 42, 56, 59].

Deep learning methods with pre-trained network. In the optimization unrolling scheme, the embedded CNNs can be viewed as image denoisers encoding the prior of the image. The so-called deep learning with plug-and-play prior directly plug the pre-trained denoising network into the iterative scheme; see *e.g.* [35, 46, 48]. Instead of using pre-trained denoising network, pre-trained GAN is also used for providing the prior to regularize the prediction from the network; Ankit Raj *et al.* [43] used a pre-trained GAN to replace the hand-crafted prior for compressed sensing.

Unsupervised learning specifically for image denoising. There has been a steady progress on unsupervised learning for denoising or similar tasks such as in-painting. The Noise2Noise [29] trained the network from two noisy instances of the same scene. Noise2void [27] proposed a blind-spot technique for training a denoising network on a set of noisy images. SURE-based regularization method [50] proposed a Stein's unbiased estimator of the supervised loss function from noisy images. A dropout NN is proposed in [41] to train the NN on a single noisy image. R2R [40] introduced a data augmentation scheme to provide an unbiased estimate of the loss function supervised over truth images.

Self-supervised and unsupervised learning for image restoration. Ulyanov *et al.* [53] proposes the DIP which uses early stopping for avoiding overfitting, since the training favors regular structure over random patterns during early iterations. DIP has been exploited for many image restoration tasks, including blind deblurring [44] and image matting [57]. For further improving the effectiveness of DIP on addressing overfitting, Heckel *et al.* [24] proposed an under-parameterized deep decoder to handle the overfitting.

Metzler *et al*. [34] and Zhussip *et al*. [66] proposed to regularize the denoiser by SURE in the iterative approximate message passing (AMP) scheme. Pang *et al*. [39] proposed a variational approximation method to the MMSE estimator by training a BNN with is weights following normal distributions. Cheng *et al*. [14] proposed to use classic SGLD for image denoising and in-painting. Chen *et al*. [11] assumed equivalence in data and trained an NN in the dual space only on measurements. Li *et al*. [31] proposed using a dropout-based MC sampling method for blind image deblurring.

3. MCMC sampling with ASGLD

This section is devoted to a detailed discussion of the ASGLD method, with more details in supplementary file.

Approximate MMSE estimator via MC-sampling. Consider an inverse problem expressed as

$$\boldsymbol{y} = \Psi(\boldsymbol{x}) + \boldsymbol{n}, \tag{6}$$

where $\boldsymbol{y}$ denotes the measurement set, $\boldsymbol{x}$ denotes the image , and $\boldsymbol{n}$ denotes measurement noise. Here we consider Gaussian white noise: $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 I)$. Let $f(\epsilon_0; \theta) = \boldsymbol{x}$ denote a generative network with weights θ, which maps an initial seed ϵ_0 to the image $\boldsymbol{x}$. Then, we have an MMSE estimator of $\boldsymbol{x}$, or the so-called conditional mean, given by

$$\widehat{\boldsymbol{x}} = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} = \int f(\epsilon_0; \theta) p(\theta|\boldsymbol{y}, \epsilon_0) d\theta. \tag{7}$$

In our approach, the integral above is calculated by the MC method. Then, the remaining question is how to sample the posterior distribution defined by

$$\pi(\theta) = p(\theta|\boldsymbol{y}, \epsilon_0).$$

As the dimension of θ is very high, to efficiently approximate the integration, the sampler should focus on the samples which make sufficient contribution to the calculation of the integral. In other words, the samples should concentrate in the subset where $\pi(\theta)$ is significant. Thus, we consider to have a restricted sampler whose samples concentrated in

$$\Omega_\epsilon := \{\theta : \sigma^2 - \epsilon \leq L(\theta) \leq \sigma^2 + \epsilon\}, \tag{8}$$

for some small constant ϵ, where

$$L(\theta) = \frac{1}{N} \|\Psi(f(\epsilon_0; \theta)) - \boldsymbol{y}\|_2^2. \tag{9}$$

Adaptive Langevin dynamics. One popular MC sampler is the so-called MCMC sampler. For deep network, it is proposed in [55] that Langevin dynamics, an MCMC sampler, can be used for sampling the posterior distribution related to network weights. Langevin dynamic originally is for describing the dynamics of molecular systems with stochastic differential equation given by

$$d\theta_t = -\nabla L(\theta_t) dt + \sqrt{2} dW_t, \tag{10}$$

where L denotes a loss function. The discretization of the equation above gives the so-called SGLD:

$$\theta_{k+1} = \theta_k - \gamma_k \cdot \nabla L(\theta_k) + \sqrt{2\gamma_k} \cdot \epsilon, \tag{11}$$

where $\epsilon \sim \mathcal{N}(0, I)$. It can be seen that the iteration (11) can be viewed as the noisy stochastic gradient descent (SGD) method, corrupted by additional Gaussian white noise.

To accurately calculate (7) with sufficient computational efficiency, we need to restrict the sampler such that the samples can concentrate in the feasible set Ω_ϵ defined by (8). One natural idea is that, if one increases noise perturbation when θ is inside Ω_ϵ and decreases the noise perturbation when θ is outside of Ω_ϵ in SGLD, the resulting sampler will then be more likely to take random walk inside the feasible set Ω_ϵ. Based on such an idea, we proposed a new form of Langevin dynamics for restricted MCMC sampling.

Suppose that θ follows an uniform distribution: $\theta \sim 1_{[-T,T]}$ with sufficiently large T. Then, by Bayesian rule,

$$\pi(\theta) = p(\theta|\boldsymbol{y}, \epsilon_0) \propto p(\boldsymbol{y}|\theta, \epsilon_0) p(\theta). \tag{12}$$

Taking negative logarithm on both sides, we have then

$$-\log p(\theta|\boldsymbol{y}, \epsilon_0) = \frac{N}{2\sigma^2} L(\theta) + \text{const.}, \tag{13}$$

where L is defined by (9). We propose a new form of Langevin dynamics with adaptive stochastic term, whose underlying stochastic differential equation is defined by

$$d\theta_t = -\nabla L(\theta_t) dt + \beta \exp(c_0(\frac{\sigma^2}{L(\theta_t)} - 1)) dW_t. \tag{14}$$

Then, its discretization gives

$$\theta_{k+1} = \theta_k - \gamma_k \cdot \nabla L(\theta_k) + \beta \exp(c_0(\frac{\sigma^2}{L(\theta_k)} - 1)) \sqrt{\gamma_k} \cdot \epsilon, \tag{15}$$

where $\epsilon \sim \mathcal{N}(0, I)$. The iteration (15) is called ASGLD.

It can be seen that the stochastic term $\exp(c_0(\frac{\sigma^2}{L(\theta)} - 1))$ is adaptive to the value of $L(\theta)$. When one trains the DNN via SGD, the value of the loss function L will decrease over the iterations. In the initial iterations with large loss, the ASGLD is close to classic SGD. When $L(\theta) > \sigma^2$, noise level decreases with a smaller perturbation on SGD. In turn, the next sample θ_{t+1} is likely to have a smaller $L(\theta)$. When $L(\theta) < \sigma^2$, noise level increases with a larger perturbation on SGD. In turn, the next sample θ_{t+1} is likely to have a larger $L(\theta)$ as SGD is distorted by a large amount. In both cases, the iterative scheme (5) keeps the sequential samples being pulled back into the feasible set Ω_ϵ when the current sample is away from Ω_ϵ. The constant c_0 is chosen such that stochastic term is negligible when $L(\theta) > \frac{3}{2}\sigma^2$.

Analysis and discussion. We first showed the stationary

distribution of the stochastic different equation (15).

Theorem 3.1 (Stationary distribution)**.** *Define the density function of θ_t as $p(\theta;t)$ where θ_t is determined by (15) with random initialization. Then the stationary distribution for θ can be explicitly expressed as*

$$p_\infty(\theta) \propto \exp[-G(L(\theta)) - 2c_0\frac{\sigma^2}{L(\theta)}],$$

where $G(s) := \frac{2}{\beta^2}\int \exp(-2c_0(\frac{\sigma^2}{s}-1))ds$ is a function defined through indefinite integral.

Proof. See supplementary file for the proof. □

Indeed, with suitable β and c_0, $p_\infty(\theta)$ concentrates around $L(\theta) \approx \sigma^2$, *i.e.* the set Ω_ϵ with small ϵ. Define

$$E(s) = G(s) + 2c_0\frac{\sigma^2}{s},$$

then the iteration (15) provides the following posterior

$$p(\theta|\boldsymbol{y},\sigma,\epsilon_0) \propto \exp(-E(L(\theta))) \tag{16}$$

rather than $\exp(-\frac{N}{2\sigma^2}L(\theta))$ provided by (10). Then,

$$\frac{d}{ds}E(s) = \frac{2}{\beta^2}e^{-2c_0(\frac{\sigma^2}{s}-1)} - 2c_0\frac{\sigma^2}{s^2}.$$

Let $\beta = \sqrt{\frac{\sigma^2}{c_0}}$, we have

$$\frac{d}{ds}E(\sigma^2) = \frac{2}{\beta^2} - 2c_0\frac{1}{\sigma^2} = 0,$$

$$\frac{dd}{(ds)^2}E(s) = \frac{2}{\beta^2}e^{-2c_0(\frac{\sigma^2}{s}-1)}\frac{2c_0\sigma^2}{s^2} + \frac{4c_0\sigma^2}{s^3} > 0, \text{ for } s > 0.$$

Clearly, $E(s)$ is convex for $s > 0$ and $s = \sigma^2$ is the global minimizer of $E(s)$. Moreover, $E(s)$ has large curvature around the minimizer. For example, set $\beta = \sqrt{\frac{\sigma^2}{c_0}}$, we have $\frac{dd}{(ds)^2}E(\alpha) = \frac{4c_0}{\sigma^4}(1+c_0)$. In the case where $\sigma^2 = 0.01$, set $c_0 = 50$. Then, $\frac{dd}{(ds)^2}E(\sigma^2) > 10^8$.

Empirical illustration in CS-MRI. The illustration is conducted by applying ASGLD on CS-MRI, configured with 10% measurement noise and 1D Gaussian mask with 25% sampling ratio. See Fig. 1 for the visualization of density function $\exp(-E(s))$, and Fig. 2 for the distribution of $\{f(\epsilon_0,\theta_k)\}$ generated by ASGLD along iteration, visualized by PCA-based dimension reduction. It can be seen that the ASGLD quickly starts to sample the region close to truth after the "burn-in" iteration around 1000, and concentrate inside it afterward. In the end, the average of those samples provide an good approximation to the truth.

Training and testing. Given a generative untrained network, the network is trained via the ASGLD (15) for total K iterations. Let K_0 denotes the "burn-in" number where we assume the method starts to sample the parameters that are close to the feasible set. In other words, we generate a set of samples $\{\theta_k\}_{k=K}^{K_0}$, which can be used for calculating the integral of the MMSE estimator (7) by taking the average

$$\hat{\boldsymbol{x}} = \int f(\epsilon_0;\theta)p(\theta|\boldsymbol{y},\epsilon_0)d\theta \approx \frac{1}{K-K_0}\sum_{k=K_0}^{K} f(\epsilon_0;\theta_k).$$

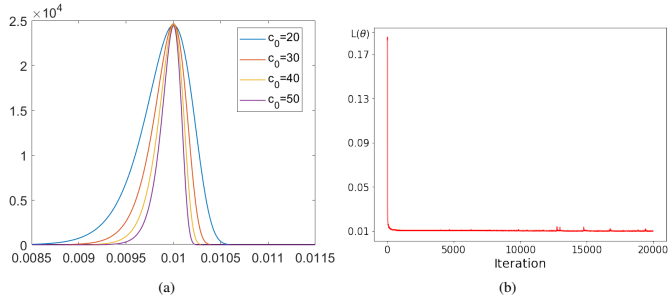


Figure 1. (a) The density function $\exp(-E(s))$ (without normalization) for different c_0 and $\sigma^2 = 0.01$ w.r.t. pixel value range $[0,1]$, the maximum is obtained at $s = 0.01$; (b) The value of $L(\theta)$ w.r.t. iteration during the training.

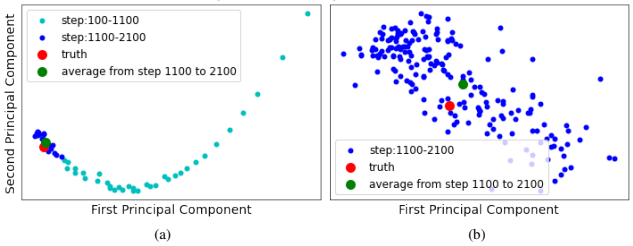


Figure 2. Visualization of the distribution of the samples generated from the ASGLD w.r.t. $\{f(\epsilon_0,\theta_k)\}$. (a) The distribution of the samples from iteration 100; (b) The distribution of the samples after iteration 1100.

4. Experiments

The proposed method is a training algorithm for training the DNN without truth images. While it is independent of network architecture, we adopt the same U-Net as DIP [53] through all experiments to exclude the effect from network architecture for fairness. The architectures are 5-layer auto-encoders with skip connections and each layer contains 128 channels. We set the fixed input ϵ_0 sampling from the uniform distribution $\mathcal{U}(0,0.1)$. All experiments are based on Pytorch on a server with NVIDIA Titan RTX GPUs. The code is publicly available at https://github.com/Wang-weixi/restricted_sampling.

In the table for quantitative comparison, the best results from all supervised methods marked as **bold**, and the best results from all traditional regularization and unsupervised learning methods are colored in blue.

4.1. Image reconstruction for CS

CS [8] is an imaging modality for faster sampling and lower energy consumption. It has a wide range of applica-

DAMP [36] TVAL3 [30] ISTA [60] DIP [53] BNN [39] ASGLD Truth

Figure 3. Results of CS image reconstruction using noisy input with ratio 25% for "Barbara".

Table 1. Average PSNR(dB) results of different methods for natural image reconstruction on Set11.

		Regularized methods		Supervised methods						Unsupervised methods			
Noise σ	CS -ratio	TVAL3 [30]	D-AMP [36]	ReconNet [28]	ISTA [60]	ISTA+ [60]	DPANet [51]	MACNet [12]	FISTANet [56]	SURE-AMP [66]	DIP [53]	BNN [39]	ASGLD
0	40%	30.52	33.49	-	35.97	36.02	35.04	35.31	**36.24**	34.12	33.28	35.71	35.87
	25%	26.44	28.21	25.54	32.59	32.44	31.74	**32.91**	32.60	29.76	31.33	32.30	33.06
	10%	21.35	21.16	22.68	26.64	26.49	26.99	**27.68**	26.94	22.65	27.40	27.49	28.15
10	40%	26.66	29.25	-	27.98	**31.09**	30.01	30.34	31.07	31.14	28.87	30.39	31.11
	25%	24.75	26.35	24.36	27.26	29.20	29.25	29.31	**30.32**	28.50	27.36	28.67	29.35
	10%	21.02	20.84	22.00	24.55	24.55	25.21	**25.56**	25.11	22.11	24.19	25.23	26.02

tions including medical imaging [20, 33] and computational photography [2, 19]. Image reconstruction for CS can be formulated as solving an ill-posed linear system:

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

where $A \in \mathbb{R}^{M\times N}$ ($M \ll N$) denotes the sensing matrix.

CS for natural image acquisition. Following the setting of [39, 60], we consider CS-based natural image acquisition via block-wise random Gaussian sensing matrix. The images from Set11 [60] and BSD68 [27] are cropped into non-overlapped blocks of size 33×33 to generate the measurements. The sensing matrix A of size $M \times N$ ($N = 1089$) is sampled from $\mathcal{N}(0, I)$ with row-wise orthogonalization. Gaussian white noise with s.t.d. 10 is added to the measurements. Three different down-sampling ratios are tested. The methods for comparison includes regularization methods: TVAL3 [30] and D-AMP [36]; six supervised learning methods: ReconNet [28], ISTA and ISTA+ [60], DPANet [51], MACNet [12] and FISTANet [56]; three unsupervised methods: SURE-AMP [66], DIP [53], and BNN [39]. We set the learning rate as 0.001. For noisy data, $K = 10000, K_0 = 2000$. For noise-free data, $K = 30000, K_0 = 5000$.

See Table 1–2 for quantitative comparison of different methods in different configurations. It showed that ASGLD not only outperformed all non-learning and unsupervised deep learning methods by a noticeable margin, but also remained very competitive against supervised learning methods. See Figure 3 for visual comparison of different methods of one example and supplementary file for more examples.

CS-MRI. The second CS application is CS-based MRI, an important technology for rapid MRI imaging. The experiment follows the procedure of [32, 39]. The measurement matrix A is implemented using random Fourier downsampling matrices and the dataset contains 21 MRI images from ADNI (Alzheimer's Disease Neuroimaging Initiative). Three down-sampling masks with 25% down-sampling rate are tested: 1D Gaussian mask, 2D Gaussian masks, and radial mask; see [32, 39] for more details. ASGLD is compared to different methods, including direct zero-filling (ZF) method [4] and TV regularization method [33]; supervised ADMM-Net [59]; plug-and-play method [32] with three different networks: SCAE, SNLAE, GAN; tuning free plug-and-play deep learning TFPnP [54]; three unsupervised learning methods: DIP [53], BNN [39] and EI [11]. Both noiseless measurement and noisy measurement with noise level $\sigma = 10\%$ are tested. The learning rate is 0.002. For noisy data, $K = 7000, K_0 = 1000$. For noise-free data, $K = 15000, K_0 = 5000$.

See Table 3 for quantitative comparison of different methods in different configurations. See Figure 4 for the visualization of the results from different methods. It can be seen that ASGLD is overall the top performer among the methods for comparison including supervised learning methods. See supplementary file for more experiments and visualizations.

4.2. Phase retrieval

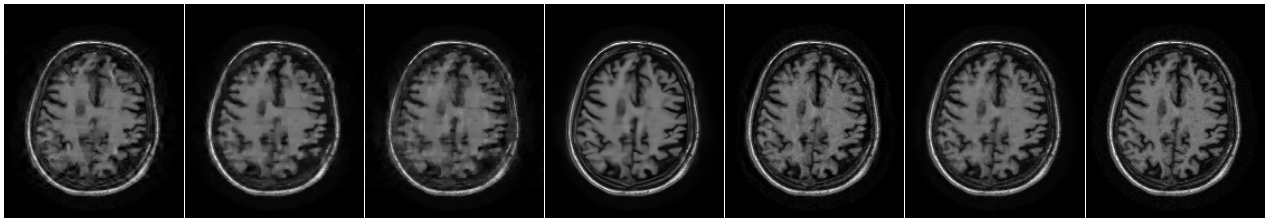
Phase retrieval is an imaging technology used in many areas of engineering and science, *e.g.* diffraction imaging [25, 26], microscopy imaging [65], and holographic

Table 2. Average PSNR(dB) results of different methods for natural image reconstruction on BSD68.

		Regularized methods		Supervised methods						Unsupervised methods			
Noise σ	CS -ratio	TVAL3 [30]	D-AMP [36]	ReconNet [28]	ISTA [60]	ISTA+ [60]	DPANet [51]	MACNet [12]	FISTANet [56]	SURE-AMP [66]	DIP [53]	BNN [39]	ASGLD
0	40%	29.39	28.03	-	32.17	32.17	31.33	31.39	**32.25**	30.26	30.10	31.28	31.36
	25%	26.48	25.57	25.31	29.36	29.29	29.00	**29.42**	29.18	27.02	27.78	28.63	29.51
	10%	22.49	21.92	23.16	25.32	25.29	25.57	**25.80**	25.09	22.53	24.82	25.24	25.51
10	40%	26.15	26.55	-	26.68	28.98	28.78	28.92	**29.01**	28.16	25.24	28.13	28.75
	25%	24.80	24.87	24.12	25.84	27.26	27.24	27.57	**28.12**	26.14	24.07	26.47	27.16
	10%	22.03	21.70	22.36	23.86	23.86	24.34	**24.63**	24.34	21.93	22.46	23.79	24.56

Table 3. Average PSNR (dB) results of different MRI reconstruction methods.

		Regularized methods		Supervised methods					Unsupervised methods			
Methods	Noise	ZF [4]	TV prior [33]	ADMM -Net [59]	SCAE [32]	SNLAE [32]	GAN [32]	TFPnP [54]	EI [11]	DIP [53]	BNN [39]	ASGLD
1D Gaussian	0	23.06	25.77	28.99	29.37	29.06	27.47	**29.94**	28.98	31.80	31.38	32.18
	10%	20.37	22.25	22.98	22.72	**24.37**	23.32	24.24	22.98	23.38	25.65	26.03
2D Gaussian	0	25.3	32.79	34.97	**35.61**	32.85	32.94	33.88	31.76	35.63	36.10	36.07
	10%	22.38	24.92	25.84	26.06	26.06	26.15	**26.89**	24.61	24.41	27.12	27.40
Radial	0	25.45	32.32	33.67	**33.94**	32.53	32.26	33.24	31.14	33.81	34.08	34.38
	10%	22.38	25.16	25.96	26.13	26.38	25.33	**27.12**	24.73	24.54	27.07	27.37



ADMM-Net [59] SNLAE [32] GAN [32] TFPnP [54] BNN [39] ASGLD Truth

Figure 4. Reconstruction results of CS-MRI with 1D Gaussian mask of sampling ratio 25%.

Table 4. Average PSNR(dB) results of different phase retrieval methods.

Dataset		Unnatural-6 [35]							Natural-6 [35]						
		Traditional		Plug-and-play		Unsupervised			Traditional		Plug-and-play		Unsupervised		
Noise type		WF [7]	DOLPHIn [52]	prGAMP [37]	prDeep [35]	DIP [53]	BNN [39]	ASGLD	WF [7]	DOLPHIn [52]	prGAMP [37]	prDeep [35]	DIP [53]	BNN [39]	ASGLD
AWGN	10	20.37	24.71	30.00	30.20	28.71	29.99	31.53	15.33	23.68	25.28	26.11	24.54	24.65	25.30
	15	26.18	26.70	32.81	32.13	32.49	31.89	34.53	21.12	26.78	28.19	28.79	29.59	29.52	30.30
	20	31.47	29.80	35.81	35.44	32.08	32.22	37.31	26.42	30.09	30.87	31.33	32.17	30.19	32.32
Poisson	9	38.67	30.12	39.41	39.01	34.16	32.96	43.67	38.80	31.21	38.12	37.87	36.21	36.64	40.41
	27	28.61	26.81	32.53	33.03	33.36	31.88	37.09	29.02	27.12	31.29	31.71	30.33	30.98	34.61
	81	17.95	22.11	25.32	27.17	26.25	29.81	30.01	18.61	19.28	23.97	25.23	24.40	24.29	28.29

imaging [45]. It needs to solve a non-linear problem:

$$\boldsymbol{b} = \sqrt{|\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{n}},$$

where $|\cdot|$ denotes absolute value and $\boldsymbol{A}$ is a sensing matrix composed by Discrete Fourier transform and bipolar (or uniform) random masks; See more details in [35]. The experiments are conducted on 2 datasets: Unnatural-6 and Natural-6 sets with 6 images each [35]. For a fair comparison, measurement data are obtained from three bipolar masks. We consider two types of noise: AWGN and Poisson. The noise

WF [7] prDeep [35] prGAMP [37] DIP [53] BNN [39] ASGLD Truth

Figure 5. Results of phase retrieval with bipolar masks and sample data with SNR=15.

Table 5. Comparison of ASGLD with classic SGLD on noisy measurements in terms of PSNR(dB).

Tasks	CS ($\sigma = 10$)						MRI ($\sigma = 25.5$)			Phase Retrieval					
	Set11			BSD68			Gaussian		Radial	AWGN (in SNR)			Poisson (α)		
	10%	25%	40%	10%	25%	40%	1D	2D		10	15	20	9	27	81
SGLD	25.20	28.49	30.31	24.13	26.28	28.02	25.11	26.25	26.30	27.64	31.66	33.25	40.74	33.09	26.73
ASGLD	26.02	29.33	31.11	24.56	27.16	28.70	26.06	27.45	27.40	28.42	32.42	34.82	42.04	35.85	29.15

level for AWGN is measured by SNR, while Poisson noise is measured by α in noise $\boldsymbol{n} \sim \mathcal{N}\left(0, \alpha^2|\boldsymbol{A}\boldsymbol{x}|^2\right)$ (a large α indicates low SNR). The compared methods include traditional non-learning Wirtinger flow (WF) method [7] and dictionary learning DOLPHIn [52]; two plug-and-play methods: prGAMP with BM3D denoiser [37] and supervised learned denoiser prDeep [35]; two unsupervised learning methods: DIP [53] and BNN [39]. We set the learning rate as 0.01. For both noiseless and noisy data, $K = 10000, K_0 = 2000$.

See Table 4 for the results. It can be seen that the ASGLD is the top performer among all compared methods. It also is the most robust to measurement noise. See Figure 5 for visual comparison of different methods of one sample.

Table 6. Comparison of model size, no. of iterations and computing time of different methods to achieve reported results for MRI reconstruction with measurement noise ($\sigma = 25.5$).

Methods	# of params	#of iterations	elapsed time	PSNR
BNN	2M	1×10^4	$\approx$ 520s	25.70
SGLD	2.2M	1×10^4	$\approx$ 640s	26.14
ASGLD	2.2M	0.3×10^4	$\approx$190s	26.42

4.3. Ablation study

In this study, we would like to see how much advantage on performance and efficiency of the proposed ASGLD over plain SGLD. The experiments are conducted on both applications. See Table 5 for the ablation study on the performance, and Table 6 for the study on the efficiency.

It can be seen from Table 5 that, thanks for our proposed effective restricted sampling strategy, in all three experiments, the proposed ASGLD outperformed the same implementation but with SGLD by a large margin. For computational efficiency, ASGLD is about 3 times faster than the one with SGLD with better result, the same for BNN. In addition, it can be seen from Figure 6 that the proposed ASGLD is also very stable during the training.

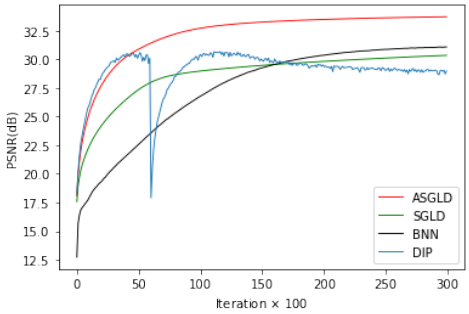


Figure 6. PSNR curves of different methods over iterations on phase retrieval (Poisson noise, $\alpha = 27$).

Limitation. The proposed method aims at training an NN for solving inverse problem without accessing any truth images. As an unsupervised solution, it cannot have a pre-trained model to process new coming data, as supervised methods can. Such an issue limits its applicability to certain applications which require real-time processing of image data.

5. Conclusion

This paper presents a self-supervised deep learning method for general image restoration and reconstruction problems, which is built on an adaptive stochastic gradient Langevin dynamics for effective MCMC sampling used for integral calculation. The proposed method is universal and has its advantages over existing unsupervised learning methods in terms of both reconstruction quality and computational efficiency. In future, we would like to investigate its applications to other problems, as well as further improve its computational efficiency.

Acknowledgement

This work was funded by Singapore MOE AcRF Tier 1 Research Grant R146000229114.

References

[1] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Trans. Med. Imag.*, 37(6):1322–1332, 2018. 3

[2] G. Arce, D. Brady, L. Carin, H. Arguello, and D. Kittle. Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Process. Mag.*, 31(1):105–115, 2013. 5

[3] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *ICML*, pages 524–533. PMLR, 2019. 1

[4] M. A. Bernstein, S. B. Fain, and S. J. Riederer. Effect of windowing and zero-filled reconstruction of MRI data on spatial resolution and acquisition strategy. *Journal of Magnetic Resonance Imaging*, 14(3):270–280, 2001. 6, 7

[5] J. Cai, H. Ji, C. Liu, and Z. Shen. Blind motion deblurring from a single image using sparse approximation. In *CVPR*, pages 104–111, 2009. 3

[6] Jian-Feng Cai, Bin Dong, Stanley Osher, and Zuowei Shen. Image restoration: total variation, wavelet frames, and beyond. *Journal of the American Mathematical Society*, 25(4):1033–1089, 2012. 3

[7] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inf. Theory*, 61(4):1985–2007, 2015. 7, 8

[8] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006. 5

[9] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997. 3

[10] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE Trans. Image Process.*, 7(3):370–375, 1998. 3

[11] Dongdong Chen, Julián Tachella, and Mike E Davies. Equivariant imaging: Learning beyond the range space. In *ICCV*, pages 4379–4388, 2021. 1, 4, 6, 7

[12] Jiwei Chen, Yubao Sun, Qingshan Liu, and Rui Huang. Learning memory augmented cascading network for compressed sensing of images. In *ECCV*, pages 513–529. Springer, 2020. 1, 3, 6, 7

[13] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1256–1272, 2016. 3

[14] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A bayesian perspective on the deep image prior. In *CVPR*, pages 5443–5451, 2019. 1, 4

[15] Q. Ding, G. Chen, X. Zhang, Q. Huang, H. Ji, and H. Gao. Low-dose CT with deep learning regularization via proximal forward backward splitting. *Physics in Medicine & Biology*, 2020. 3

[16] Qiaoqiao Ding, Hui Ji, Hao Gao, and Xiaoqun Zhang. Learnable multi-scale fourier interpolation for sparse view ct image reconstruction. In *MICCAI*, pages 286–295, 2021. 1

[17] Qiaoqiao Ding, Yuesong Nan, Hao Gao, and Hui Ji. Deep learning with adaptive hyper-parameters for low-dose ct image reconstruction. *IEEE Trans. Comput. Imag.*, 7:648–660, 2021. 1, 3

[18] Bin Dong, Hui Ji, Jia Li, Zuowei Shen, and Yuhong Xu. Wavelet frame based blind image inpainting. *Applied and Computational Harmonic Analysis*, 32(2):268–279, 2012. 3

[19] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, 2008. 5

[20] U. Gamper, P. Boesiger, and S. Kozerke. Compressed sensing in dynamic MRI. *Magnetic Resonance in Medicine*, 59(2):365–373, 2008. 5

[21] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, pages 11026–11035, 2019. 2

[22] Kuang Gong, Ciprian Catana, Jinyi Qi, and Quanzheng Li. PET image reconstruction using deep image prior. *IEEE Trans. Med. Imag.*, 38(7):1655–1665, 2018. 2

[23] R Heckel. Regularizing linear inverse problems with convolutional neural networks. In *NeurIPS 2019 Medical Imaging meets NeurIPS workshop*, 2019. 2

[24] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. In *ICLR*, 2018. 3

[25] J. Holloway, M. S. Asif, M. K. Sharma, N. Matsuda, R. Horstmeyer, O. Cossairt, and A. Veeraraghavan. Toward long-distance subdiffraction imaging using coherent camera arrays. *IEEE Trans. Comput. Imaging*, 2(3):251–265, 2016. 6

[26] M. R. Kellman, E. Bostan, N. A. Repina, and L. Waller. Physics-based learned design: Optimized coded-illumination for quantitative phase imaging. *IEEE Trans. Comput. Imaging*, 5(3):344–353, 2019. 6

[27] A. Krull, T. Buchholz, and F. Jug. Noise2void-learning denoising from single noisy images. In *CVPR*, pages 2129–2137, 2019. 1, 3, 6

[28] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *CVPR*, pages 449–458, 2016. 6, 7

[29] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2Noise: Learning image restoration without clean data. In *ICML*, pages 2965–2974, 2018. 3

[30] Chengbo Li, Wotao Yin, Hong Jiang, and Yin Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56(3):507–530, 2013. 6, 7

[31] Ji Li, Yuesong Nan, and Hui Ji. Un-supervised learning for blind image deconvolution via monte-carlo sampling. *Inverse Problems*, 2022. 2, 4

[32] Jiulong Liu, Tao Kuang, and Xiaoqun Zhang. Image reconstruction by splitting deep learning regularization from iterative inversion. In *MICCAI*, pages 224–231. Springer, 2018. 1, 6, 7

[33] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid mr imaging. *Mag-*

netic Resonance in Medicine, 58(6):1182–1195, 2007. 5, 6, 7

[34] C. Metzler, A. Mousavi, R. Heckel, and R. Baraniuk. Unsupervised learning with stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018. 2, 4

[35] Christopher Metzler, Phillip Schniter, Ashok Veeraraghavan, et al. prdeep: robust phase retrieval with a flexible deep network. In *ICML*, pages 3501–3510. PMLR, 2018. 1, 3, 7, 8

[36] C. A. Metzler, A. Maleki, and R. Baraniuk. From denoising to compressed sensing. *IEEE Trans. Inf. Theory*, 62(9):5117–5144, 2016. 6, 7

[37] Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. Bm3d-prgamp: Compressive phase retrieval based on bm3d denoising. In *ICIP*, pages 2504–2508. IEEE, 2016. 7, 8

[38] Y. Nan, Y. Quan, and H. Ji. Variational-EM-based deep learning for noise-blind image deblurring. In *CVPR*, pages 3626–3635, June 2020. 1

[39] Tongyao Pang, Yuhui Quan, and Hui Ji. Self-supervised bayesian deep learning for image recovery with applications to compressive sensing. In *ECCV*, pages 475–491, 2020. 2, 3, 4, 6, 7, 8

[40] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *CVPR*, pages 2043–2052, 2021. 1, 3

[41] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *CVPR*, pages 1890–1898, 2020. 1, 3

[42] Yuhui Quan, Peikang Lin, Yong Xu, Yuesong Nan, and Hui Ji. Nonblind image deblurring via deep learning in complex field. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021. 3

[43] Ankit Raj, Yuqi Li, and Yoram Bresler. GAN-based projector for faster recovery with convergence guarantees in linear inverse problems. In *ICCV*, pages 5602–5611, 2019. 1, 3

[44] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *CVPR*, pages 3341–3350, 2020. 2, 3

[45] Yair Rivenson, Yibo Zhang, Harun Gunaydin, Da Teng, and Aydogan Ozcan. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light, Science & Applications*, 7, 2017. 6

[46] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. 3

[47] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 3

[48] Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *ICML*, pages 5546–5557. PMLR, 2019. 1, 3

[49] W. Shi, F. Jiang, S. Liu, and D. Zhao. Scalable convolutional neural network for image compressed sensing. In *CVPR*, pages 12290–12299, 2019. 1

[50] Shakarim Soltanayev and Se Young Chun. Training deep learning based denoisers without ground truth data. In *NeurIPS*, pages 3257–3267, 2018. 1, 3

[51] Yubao Sun, Jiwei Chen, Qingshan Liu, Bo Liu, and Guodong Guo. Dual-path attention network for compressed sensing image reconstruction. *IEEE Trans. Image Process.*, 29:9482–9495, 2020. 3, 6, 7

[52] Andreas M Tillmann, Yonina C Eldar, and Julien Mairal. Dolphin—dictionary learning for phase retrieval. *IEEE Trans. Signal Process.*, 64(24):6485–6500, 2016. 7

[53] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018. 1, 2, 3, 5, 6, 7, 8

[54] Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Carola-Bibiane Schönlieb, and Hua Huang. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *ICML*, pages 10158–10169. PMLR, 2020. 1, 6, 7

[55] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, pages 681–688. Citeseer, 2011. 2, 4

[56] Jinxi Xiang, Yonggui Dong, and Yunjie Yang. FISTA-Net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging. *IEEE Trans. Med. Imag.*, 40(5):1329–1339, 2021. 1, 3, 6, 7

[57] Yong Xu, Baoling Liu, Yuhui Quan, and Hui Ji. Unsupervised deep background matting using deep matte prior. *IEEE Trans. Circuits Syst. Video Technol.*, 2021. 3

[58] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep admm-net for compressive sensing MRI. In *NeurIPS*, pages 10–18, 2016. 1

[59] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing MRI. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 10–18, 2016. 3, 6, 7

[60] Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *CVPR*, pages 1828–1837, 2018. 1, 6, 7

[61] Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson WH Lau, and Ming-Hsuan Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *CVPR*, pages 3817–3825, 2017. 1

[62] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 1

[63] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017. 3

[64] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Trans. Image Process.*, 27(9):4608–4622, 2018. 3

[65] Guoan Zheng, Roarke Horstmeyer, and Changhuei Yang. Wide-field, high-resolution fourier ptychographic microscopy. *Nature Photonics*, 7:739–745, 2013. 6

[66] M. Zhussip, S. Soltanayev, and S. Chun. Training deep learning based image denoisers from undersampled measurements without ground truth and without image prior. In *CVPR*, pages 10255–10264, 2019. 2, 4, 6, 7

Supplementary Materials for "Restricted MCMC Sampling for Solving Inverse Imaging Problems with Untrained Deep Network"

Anonymous CVPR submission

Paper ID 7221

1. Proof of Theorem 3.1

In this section, we present the detailed proof of the main result stated in the main manuscript. Recall that the proposed method is built on the stochastic differential equation (SDE) defined by

$$d\theta_t = -\nabla L(\theta_t)dt + \beta \exp(c_0(\frac{\sigma^2}{L(\theta_t)} - 1))dW_t. \tag{1}$$

In this section, we derive the stationary distribution derived from the above dynamics.

Theorem 3.1 (Stationary distribution)**.** *Define the density function of θ_t as $p(\theta; t)$ where θ_t is determined by (1) with random initialization. Then the stationary distribution for θ can be explicitly expressed as*

$$p_\infty(\theta) \propto \exp[-G(L(\theta)) - 2c_0\frac{\sigma^2}{L(\theta)}],$$

where $G(s) := \frac{2}{\beta^2}\int \exp(-2c_0(\frac{\sigma^2}{s} - 1))ds$ is a function defined through indefinite integral.

Proof. Denote

$$A(\theta_t) = -\nabla L(\theta_t) \in \mathbb{R}^n \quad \text{and } B(\theta_t) = \beta \exp(c_0(\frac{\sigma^2}{L(\theta_t)} - 1))I_n \in \mathbb{R}^{n\times n}, \tag{2}$$

where I_n is an identity matrix, we have then

$$d\theta_t = A(\theta_t)dt + B(\theta_t) \cdot dW_t.$$

The dynamics of probability current $p(\theta, t)$ is governed by the Fokker-Plank equation

$$\frac{d}{dt}p(\theta, t) + \sum_{i=1}^{n} \partial_{\theta_i} J_i(\theta, t) = 0, \tag{3}$$

where

$$J_i(\theta, t) = A_i(\theta)p(\theta, t) - \frac{1}{2}\sum_{j=1}^{n}\frac{\partial}{\partial\theta_j}[B^2(\theta)_{i,j}p(\theta, t)], \quad i = 1, \ldots, n$$

and $B^2(\theta) = B(\theta) \cdot B(\theta)^\top$.

Consider the stationary distribution $p_\infty(\theta) := p(\theta, t)$ with $t \to \infty$ and replace $p(\theta, t)$ by $p_\infty(\theta)$, we have

$$\sum_{i=1}^{n} \partial_{\theta_i} J_i(\theta, t) = 0.$$

We further require

$$J_i(\theta, t) = 0, \quad \text{for all } i.$$

Then, we have

$$p_\infty(\theta)A(\theta) = \frac{1}{2}\sum_j \partial_j(p_\infty(B(\theta)B(\theta)^\top)_j) = \frac{1}{2}p_\infty[B(\theta)B(\theta)^\top \cdot \nabla] + \frac{1}{2}B(\theta)B(\theta)^\top \cdot \nabla p_\infty.$$

By direct calculation, we have

$$\nabla \log p_\infty(\theta) = (B(\theta)B(\theta)^\top)^{-1}[2A(\theta) - (B(\theta)B(\theta)^\top \cdot \nabla)].$$

Substituting the explicit form of $A(\theta)$ and $B(\theta)$ (2) into above equality, we have

$$\begin{aligned}\nabla \log p_\infty(\theta) &= \frac{1}{\beta^2}\exp(-2c_0(\frac{\sigma^2}{L(\theta)} - 1))[2\nabla L(\theta) - \nabla(\beta^2 \exp(2c_0(\frac{\sigma^2}{L(\theta)} - 1)))] \\ &= \frac{2}{\beta^2}\exp(-2c_0(\frac{\sigma^2}{L(\theta)} - 1))\nabla L(\theta) - \nabla 2c_0(\frac{\sigma^2}{L(\theta)}) := Z(\theta).\end{aligned}$$

Noted that the vector function $Z(\theta)$ satisfies

$$\partial_i Z_j(\theta) = \partial_j Z_i(\theta),$$

thus $Z(\theta)$ is integrable. Hence the stationary distribution exists and it satisfies

$$p_\infty(\theta) \propto \exp(G(L(\theta)) - \frac{2c_0\sigma^2}{L(\theta)})$$

with $G(s) := \frac{2}{\beta^2}\int \exp(-2c_0(\frac{\sigma^2}{s} - 1))ds$.

For the uniqueness, consider the KL divergence between $p_t(\theta)$ and $p_\infty(\theta)$:

$$F(t) := \mathrm{KL}(p_t(\theta), p_\infty(\theta)) = \int p_t(\theta)\ln(\frac{p_t(\theta)}{p_\infty(\theta)})d\theta.$$

Then

$$\partial_t F(t) = \int \partial_t p_t \ln(\frac{p_t}{p_\infty})d\theta + \int p_t \partial_t \ln\frac{p_t}{p_\infty}d\theta = \int \partial_t p_t \ln(\frac{p_t}{p_\infty})d\theta$$

The Fokker-Plank equation is

$$\partial_t p_t(\theta) = -\sum_i \partial_i(p_t(\theta)A_i(\theta)) + \frac{1}{2}\sum_{i,j}\partial_{i,j}(p_t B(\theta)^2_{i,j}).$$

Substituting this equation into to $\partial_t F(t)$, we have

$$\partial_t F(t) = \sum_i \int p_t(\theta)[A_i(\theta)\partial_i \ln(\frac{p_t(\theta)}{p_\infty(\theta)})d\theta + \sum_{i,j}\int B^2_{i,j}(\theta)\partial_{i,j}\ln(\frac{p_t(\theta)}{p_\infty(\theta)})]d\theta.$$

Noted that

$$\begin{aligned}\partial_i \ln\frac{p_t}{p_\infty} &= \frac{p_\infty}{p_t}\partial_i(\frac{p_t}{p_\infty}) \\ \partial_{i,j}\ln(\frac{p_t}{p_\infty}) &= (\frac{p_\infty}{p_t})^2[\partial_{i,j}(\frac{p_t}{p_\infty})(\frac{p_t}{p_\infty}) - \partial_i(\frac{p_t}{p_\infty})\partial_j(\frac{p_t}{p_\infty}),]\end{aligned}$$

we have

$$\partial_t F(t) = -\mathbb{E}_{p_t}[\|\nabla_\theta \ln(\frac{p_t}{p_\infty})\|^2_{B^2}],$$

where for any column vector $x \in R^n$, $\|x\|^2_{B^2} = x^\top B^2 x$. Thus as long as $\nabla_\theta \ln(\frac{p_t}{p_\infty}) \neq 0$, we have $\partial_t F(t) < 0$. Suppose $p_t \to p'_\infty$, because $F(t)$ is lower bounded, when $t \to \infty$, $\nabla_\theta \ln(\frac{p'_\infty}{p_\infty}) = 0$. So

$$p'_\infty(\theta) = p_\infty(\theta).$$

The result provides the uniqueness of the stationary distribution. □

2. Robustness of ASGLD to possible estimation error of noise level of measurement

The propose method requires the prior of noise level of the measurement, the standard deviation (s.t.d.). As in practice, noise level usually is estimated either by empirical data or some estimator, the estimation might not be exact. In the experiment, we show how robust of the proposed method to possible estimation error of measurement. The experiment is conducted on CS acquisition for natural image. Different noise levels are used by ASGLD, where the truth noise level is $\sigma = 10$. See Table 1 for the results on the dataset Set11 [**?**] under 3 different sampling rates, in the presence of AGWN with $\sigma = 10$. It can be seen that the proposed ASGLD is robust to such estimation error, the performance impact is negligible with 10% error ratio, and remains small even with 20% error ratio.

Table 1. The results from the ASGLD with different inputs of the estimated noise level $\tilde{\sigma}^2$.

$\tilde{\sigma}^2$	$0.8\sigma^2$	$0.9\sigma^2$	σ^2	$1.1\sigma^2$	$1.2\sigma^2$
40%	30.72	30.98	31.11	31.13	31.09
25%	29.15	29.29	29.35	29.37	29.36
10%	25.87	25.94	26.02	26.07	26.02

3. Visual inspection of more results from the experiment on phase retrieval

In this section, we visually show more results from the experiment on phase retrieval. For Gaussian measurement data, see Figure 1 for the results of different methods on one natural image. For Poisson measurement data with $\alpha = 27$ (see main manuscript and [**?**] for more details), see Figure 2 and 3 for visual inspection of the results from different methods on one sample natural image and one sample unnatural image.

It can be seen that overall, the results from the proposed ASGLD contain more details and have less noise, in comparison to that from other unsupervised deep learning methods, as well as that from traditional and supervised methods. For example, in comparison to two unsupervised learning methods. The results from DIP [**?**] contains noticeable noise and the results from BNN [**?**] blurred out image details. In contrast, the results from the ASGLD method have the sharpest image details and have least noticeable noise.

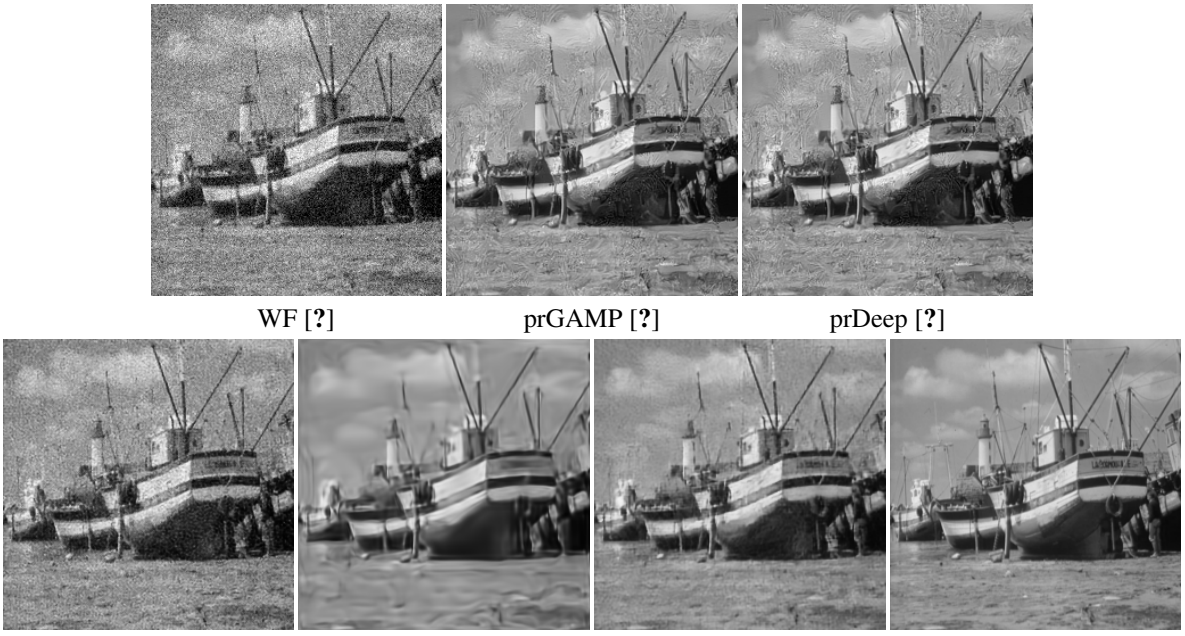


Figure 1. Phase retrieval results of ”boat” with bipolar mask and Gaussian measurement data with SNR=15.

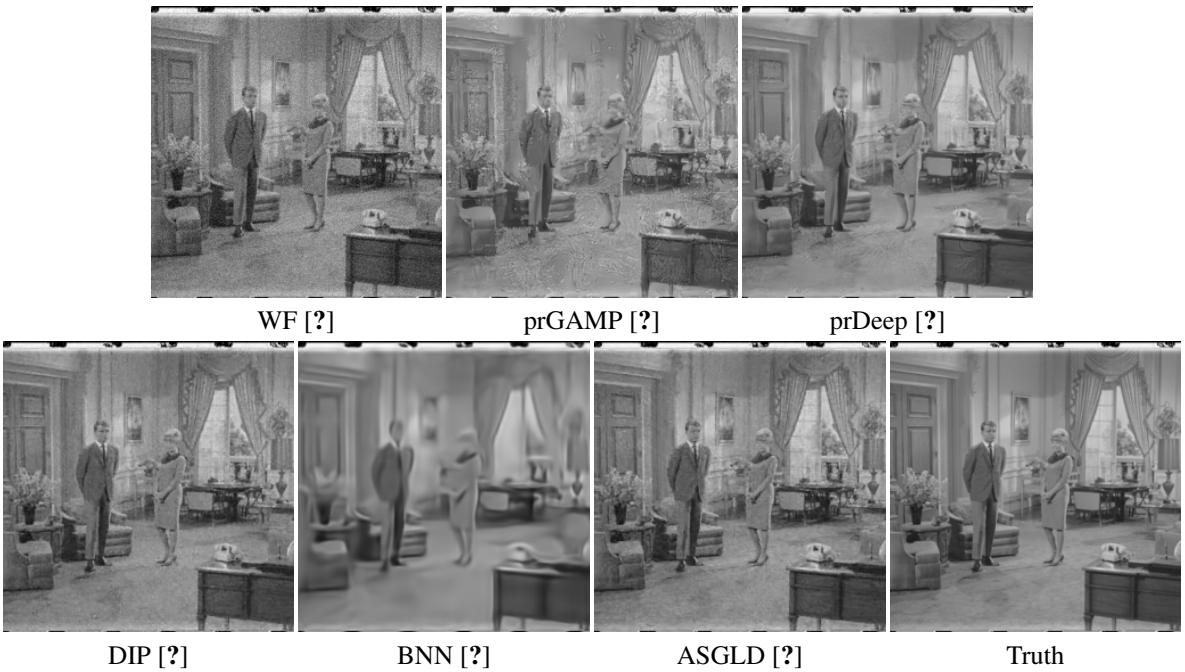


Figure 2. Phase retrieval results of "couple" with bipolar mask and Poisson measurement data with $\alpha = 27$.

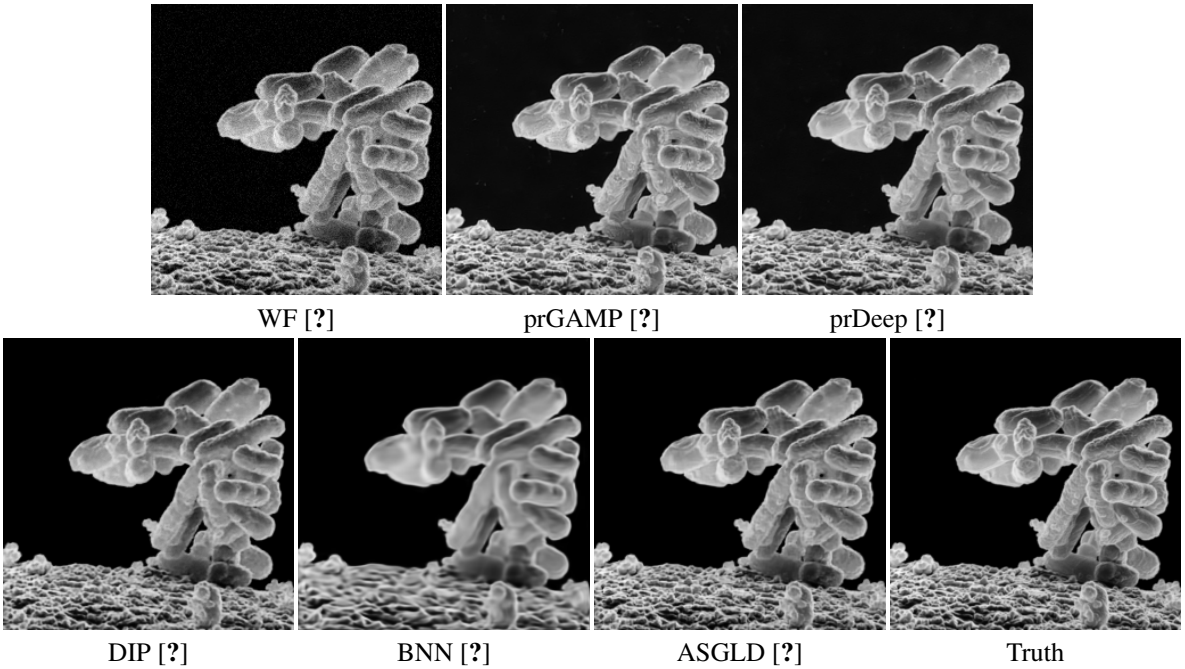


Figure 3. Phase retrieval results of "Ecoli" with bipolar mask and Poisson measurement data with $\alpha = 27$.

4. Visual inspection of more results for CS: Natural image acquisition and MRI

In this section, we show more examples for visual inspection of the results from different methods for CS. See Figure 4 and 5 for visual inspection of CS for natural image acquisition. For CS-MRI, see Figure 6 for the visualization of three masks for sampling the Fourier measurement used in the experiments. See Figure 7 visual inspection of the reconstructed images from different methods, in the case of noisy measurements ($\sigma = 10\%$) with 1D Gaussian mask and sampling ratio 25%.

Figure 4. Visualization of different results for CS-based natural image acquisition of "Lena256", using noisy data $\sigma = 10$ with ratio 25%.

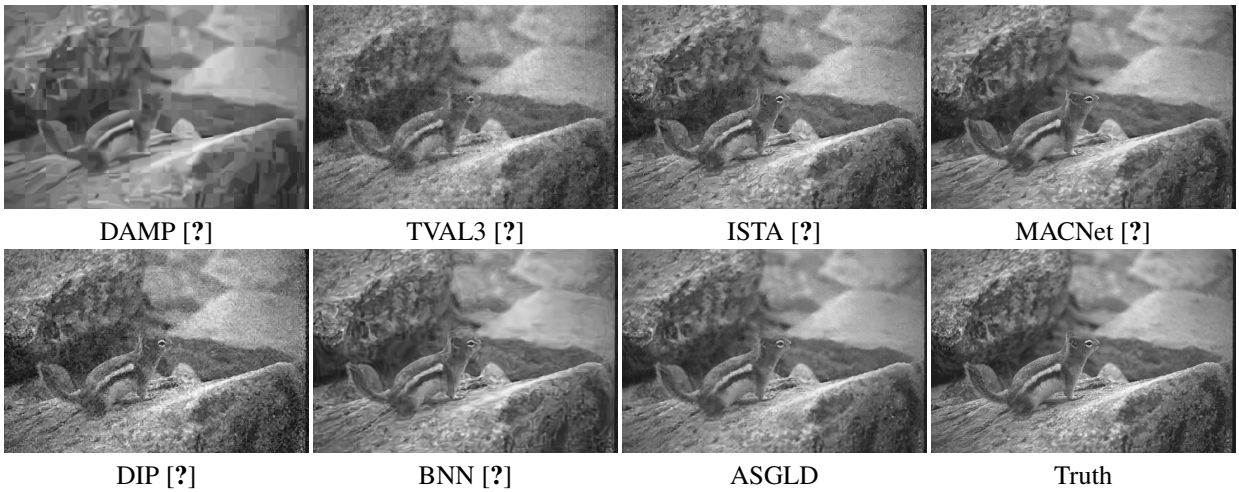
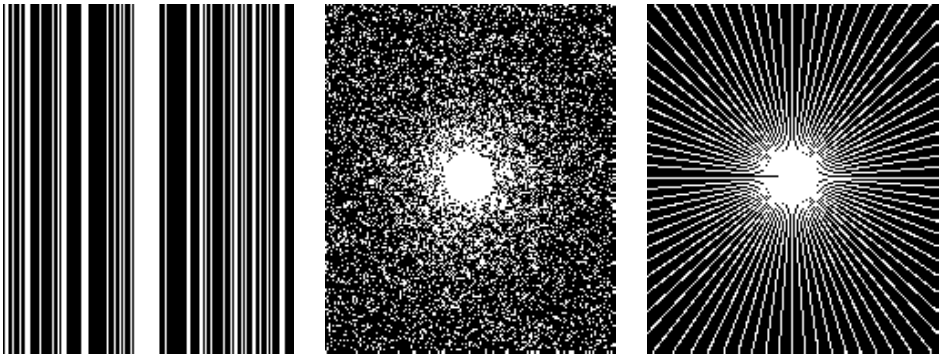


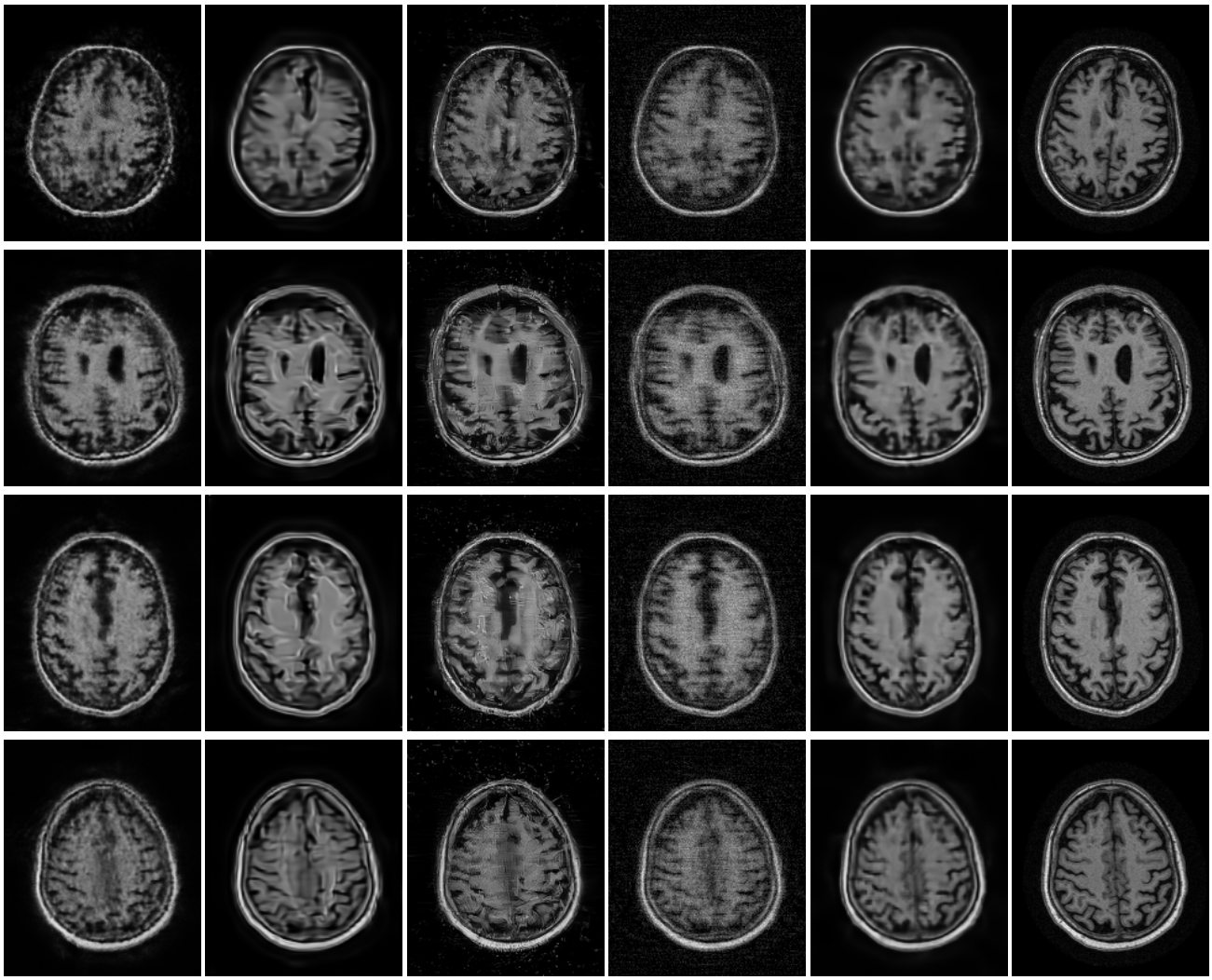
Figure 5. Visualization of different results for CS-based natural image acquisition using noisy input $\sigma = 10$ with ratio 25%.

The observation is consistent with that for phase retrieval. For two compared unsupervised learning methods, the results from DIP [?] has noticeable noise and the results from BNN [?] have image detailed smoothed out. Overall, the results from the ASGLD have most image details and least noise.



1D Gaussian 2D Gaussian radial

Figure 6. Three different types of sampling masks of sample ratio 25%



DIP [?] BNN [?] TFPnP [?] EI [?] ASGLD Truth

Figure 7. MRI reconstruction results with 1D Gaussian mask of sampling ratio 25% and 10% noise.