Supplementary Materials for "Self-supervised Deep Image Restoration via Adaptive Stochastic Gradient Langevin Dynamics"

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

1. A detailed discussion of the feasible set (5)

Recall the Chebyshev's inequality for a random variable X states that $P(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}$. In our case, $X = \frac{1}{N}\|\Phi(\boldsymbol{x}) - (\Phi(\boldsymbol{x}) + \boldsymbol{n})\|_2^2 = \frac{1}{N}\sum_{i=1}^N n_i^2$, where $\boldsymbol{n} = [n_i]_{i=1}^N$ denotes the vector of i.i.d. random variables $n_i \sim \mathcal{N}(0, \sigma^2)$. Recall the second and fourth moment of n_i is σ^2 and $3\sigma^4$, thus, we have $\mathbb{E}[X] = \sigma^2$ and

$$\text{Var}(n_i^2) = \mathbb{E}[n_i^4] - \sigma^4 = 2\sigma^4 \Longrightarrow \text{Var}(X) = (2\sigma^4)/N$$

which leads to $P(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{2\sigma^4}{N\epsilon^2}$. Let $\epsilon = 0.1\sigma^2$. Then, $0.9\sigma^2 \leq X \leq 1.1\sigma^2$ with probability at least $1 - \frac{200}{N}$. As $N = O(10^4)$ in our case, Thereofre, we define the feasible set as (5) in our paper.

2. 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

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

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 [12] 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 [6] 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 self-supervised deep learning methods, as well as that from traditional and supervised methods. For example, in comparison to two self-supervised learning methods. The results from DIP [10] contains noticeable noise and the results from BNN [9] blurred out image details. In contrast, the results from the ASGLD method have the sharpest image details and have least noticeable noise.

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%.

The observation is consistent with that for phase retrieval. For two compared unsupervised learning methods, the results from DIP [10] has noticeable noise and the results from BNN [9] have image detailed smoothed out. Overall, the results from the ASGLD have most image details and least noise.

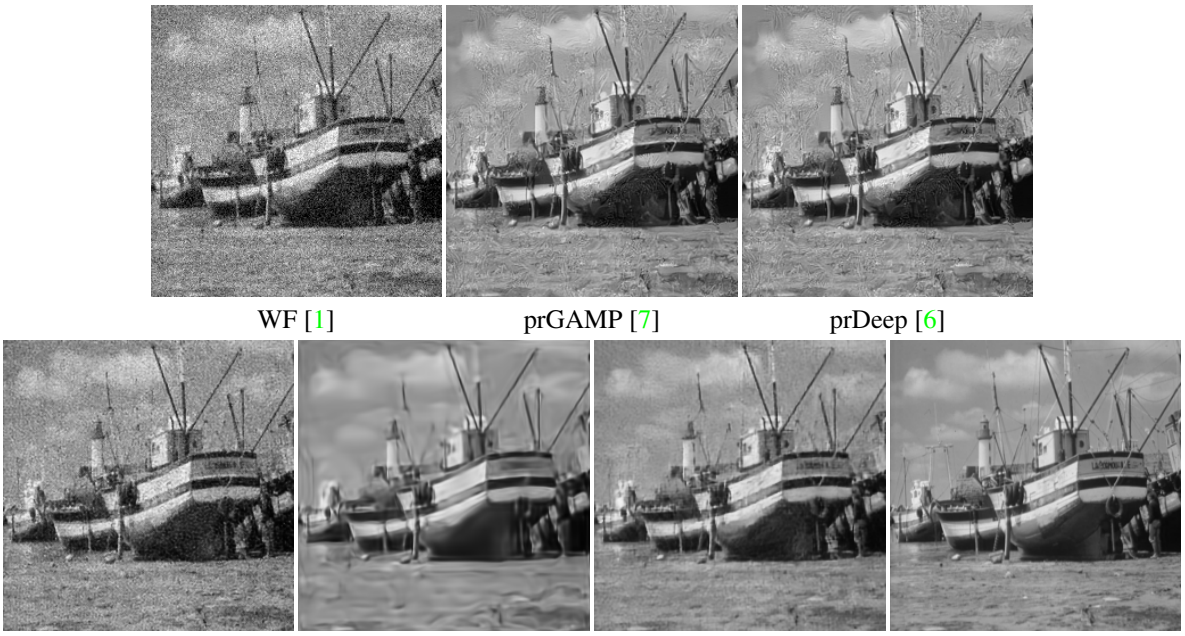


Figure 1. Phase retrieval results of "boat" with bipolar mask and Gaussian measurement data with SNR=15.

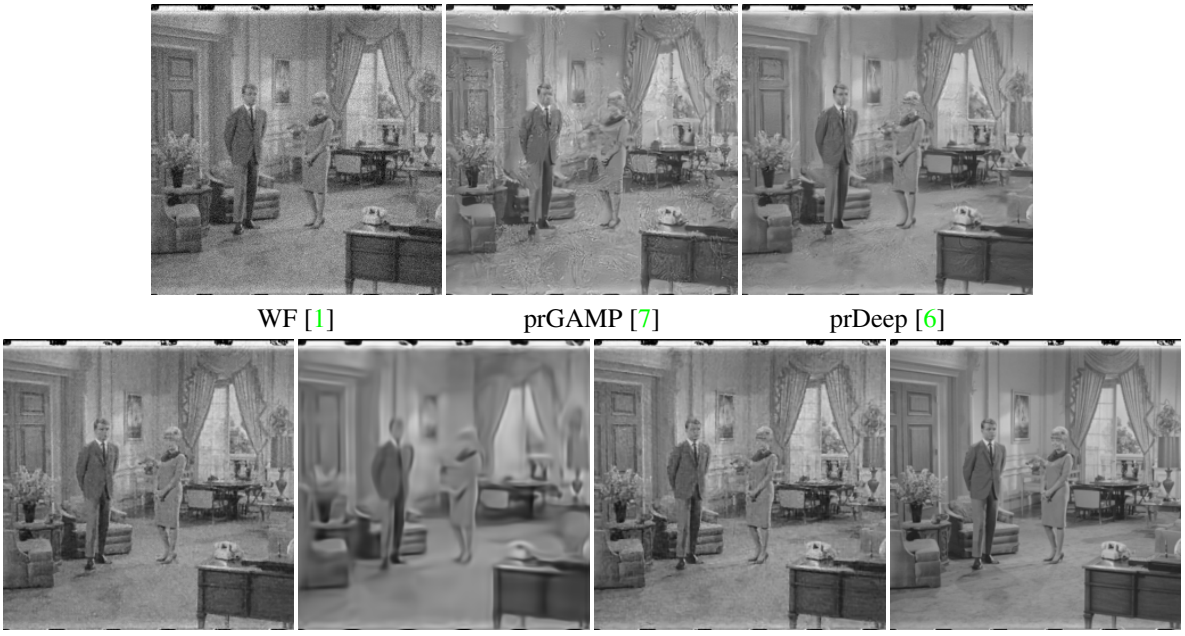
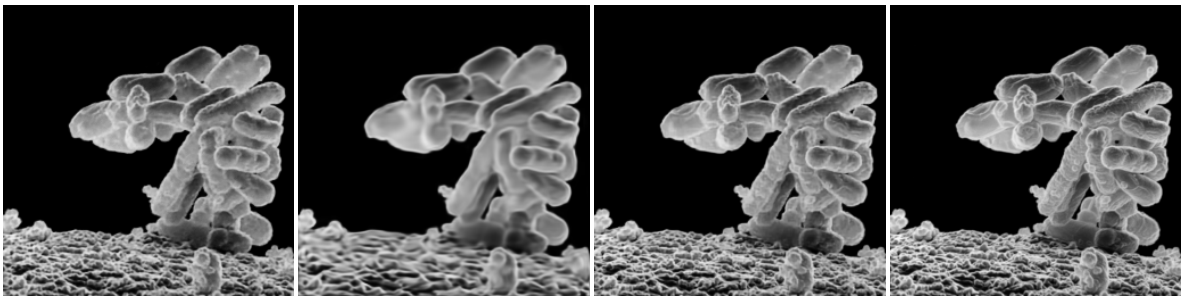


Figure 2. Phase retrieval results of "couple" with bipolar mask and Poisson measurement data with $\alpha = 27$.

WF [1] prGAMP [7] prDeep [6]



DIP [10] BNN [9] ASGLD [4] Truth

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

DAMP [8] TVAL3 [5] ISTA [12] MACNet [3]

DIP [10] BNN [9] ASGLD Truth

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

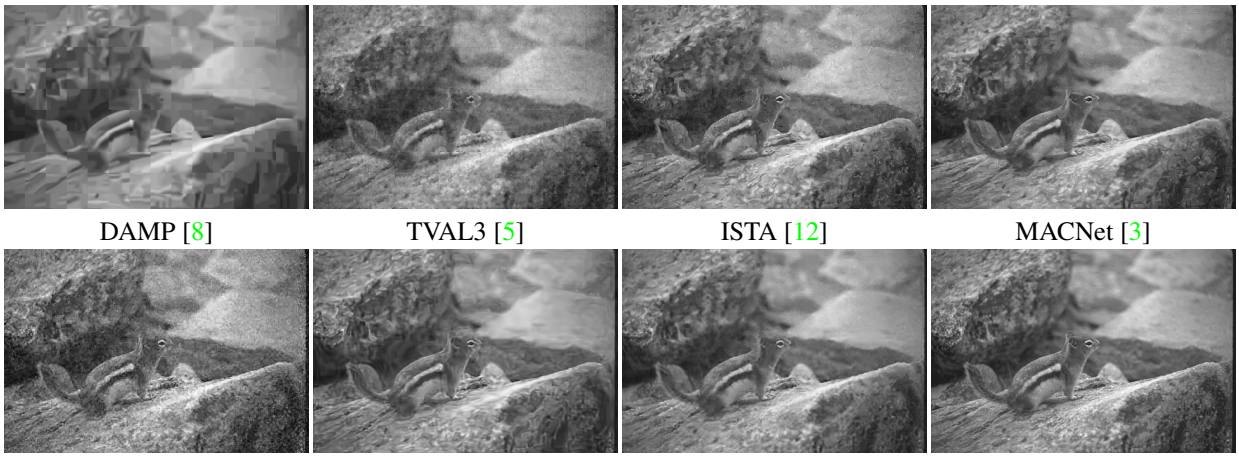


Figure 5. Visualization of different results for CS-based natural image acquisition using noisy input $\sigma = 10$ with ratio 25%.

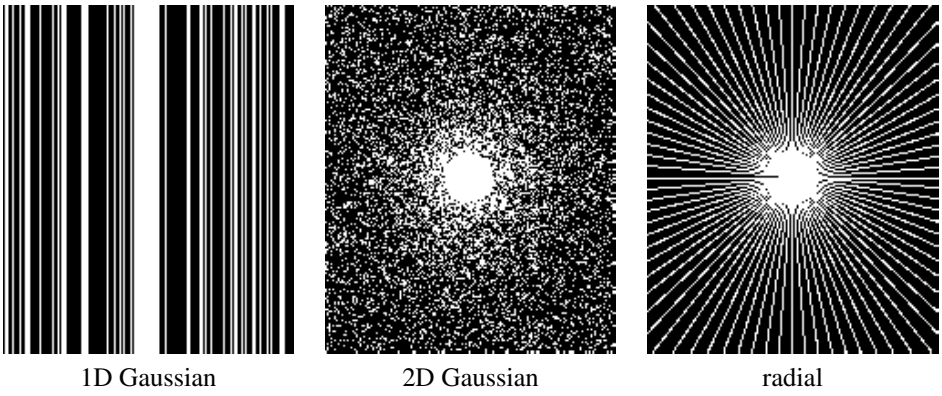
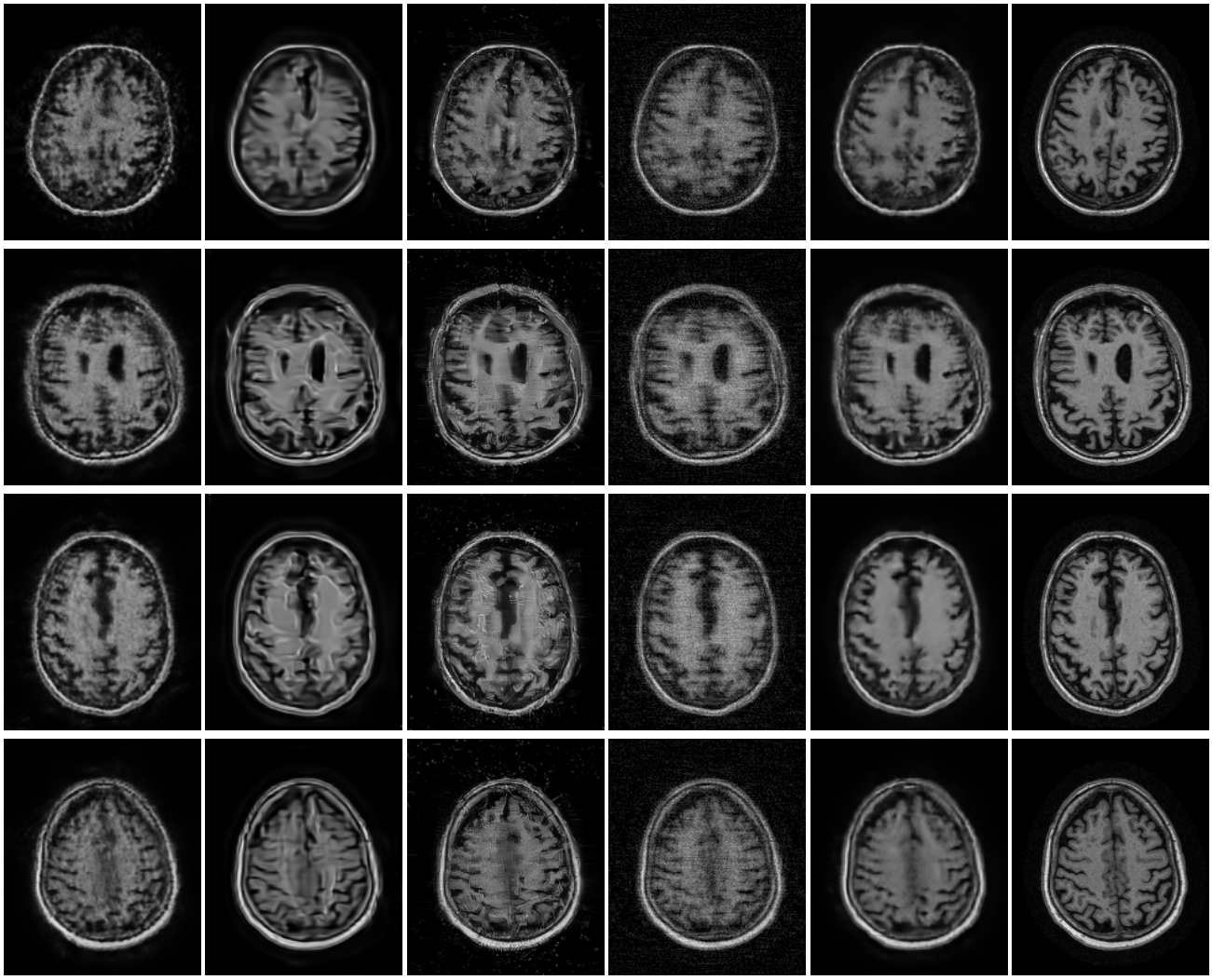


Figure 6. Three different types of sampling masks of sample ratio 25%



DIP [10] BNN [9] TFPnP [11] EI [2] ASGLD Truth

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

References

[1] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inf. Theory*, 61(4):1985–2007, 2015. 4, 5

[2] D. Chen, J. Tachella, and M. E. Davies. Equivariant imaging: Learning beyond the range space. In *ICCV*, pages 4379–4388, 2021. 7

[3] J. Chen, Y. Sun, Q. Liu, and R. Huang. Learning memory augmented cascading network for compressed sensing of images. In *ECCV*, pages 513–529. Springer, 2020. 5, 6

[4] Z. Cheng, M. Gadelha, S. Maji, and D. Sheldon. A bayesian perspective on the deep image prior. In *CVPR*, pages 5443–5451, 2019. 4, 5

[5] C. Li, W. Yin, H. Jiang, and Y. Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56(3):507–530, 2013. 5, 6

[6] C. Metzler, P. Schniter, A. Veeraraghavan, et al. prdeep: robust phase retrieval with a flexible deep network. In *ICML*, pages 3501–3510. PMLR, 2018. 3, 4, 5

[7] C. A. Metzler, A. Maleki, and R. G. Baraniuk. Bm3d-prgamp: Compressive phase retrieval based on bm3d denoising. In *ICIP*, pages 2504–2508. IEEE, 2016. 4, 5

[8] C. A. Metzler, A. Maleki, and R. G. Baraniuk. From denoising to compressed sensing. *IEEE Trans. Inf. Theory*, 62(9):5117–5144, 2016. 5, 6

[9] T. Pang, Y. Quan, and H. Ji. Self-supervised bayesian deep learning for image recovery with applications to compressive sensing. In *ECCV*, pages 475–491, 2020. 3, 4, 5, 6, 7

[10] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018. 3, 4, 5, 6, 7

[11] K. Wei, A. Aviles-Rivero, J. Liang, Y. Fu, C.-B. Schönlieb, and H. Huang. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *ICML*, pages 10158–10169. PMLR, 2020. 7

[12] J. Zhang and B. Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *CVPR*, pages 1828–1837, 2018. 3, 5, 6