Solving Inverse Problems with a Flow-based Noise Model

Jay Whang [1] **Qi Lei** [2] **Alexandros G. Dimakis** [3]

Abstract

We study image inverse problems with a normalizing flow prior. Our formulation views the solution as the maximum a posteriori estimate of the image conditioned on the measurements. This formulation allows us to use noise models with arbitrary dependencies as well as non-linear forward operators. We empirically validate the efficacy of our method on various inverse problems, including compressed sensing with quantized measurements and denoising with highly structured noise patterns. We also present initial theoretical recovery guarantees for solving inverse problems with a flow prior.

1. Introduction

Inverse problems seek to reconstruct an unknown signal from observations (or *measurements*), which are produced by some process that transforms the original signal. Because such processes are often lossy and noisy, inverse problems are typically formulated as reconstructing $\boldsymbol{x}$ from its measurements

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\delta} \tag{1}$$

where f is a known deterministic *forward operator* and $\boldsymbol{\delta}$ is an additive noise which may have a complex structure itself. An impressively wide range of applications can be posed under this formulation with an appropriate choice of f and $\boldsymbol{\delta}$, such as compressed sensing (Candes et al., 2006; Donoho, 2006), computed tomography (Chen et al., 2008), magnetic resonance imaging (MRI) (Lustig et al., 2007), and phase retrieval (Candes et al., 2015a;b).

In general for a non-invertible forward operator f, there can be potentially infinitely many signals that match given observations. Thus the recovery algorithm must critically rely on *a priori* knowledge about the original signal to find the most plausible solution among them. Sparsity has classically been a very influential structural prior for various inverse problems (Candes et al., 2006; Donoho, 2006; Baraniuk et al., 2010). Alternatively, recent approaches introduced deep generative models as a powerful signal prior, showing significant gains in reconstruction quality compared to sparsity priors (Bora et al., 2017; Asim et al., 2019; Van Veen et al., 2018; Menon et al., 2020).

However, most existing methods assume the setting of Gaussian noise model and are unsuitable for structured and correlated noise, making them less applicable in many real-world scenarios. For instance, when an image is ruined by hand-drawn scribbles or a piece of audio track is overlayed with human whispering, the noise to be removed follows a very complex distribution. These settings deviate significantly from Gaussian noise settings, yet they are much more realistic and deserve more attention. In this paper, we propose to use a normalizing flow model to represent the noise distribution and derive principled methods for solving inference problems under this general noise model.

Contributions.

- We present a general formulation for obtaining maximum a posteriori (MAP) reconstructions for dependent noise and general forward operators. Notably, our method can leverage deep generative models for both

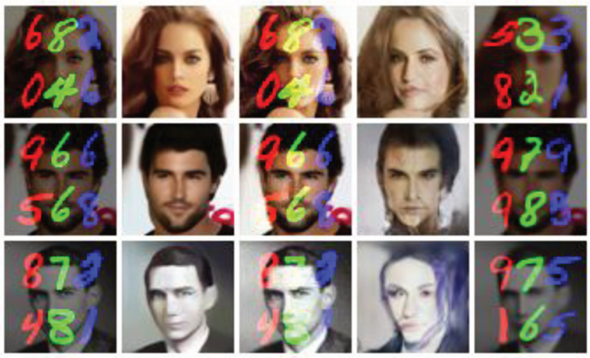


Figure 1. Result of denoising MNIST digits. The first column contains noisy observations, and subsequent columns contain reconstructions.

[1]Dept. of Computer Science, UT Austin, TX, USA [2]Dept. of Electrical and Computer Engineering, Princeton University, NJ, USA [3]Dept. of Electrical and Computer Engineering, UT Austin, TX, USA. Correspondence to: Jay Whang <jaywhang@utexas.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

the original image and the noise.

- We extend our framework to the general setting where the signal prior is given as a latent-variable model, for which likelihood evaluation is intractable. The resulting formulation presents a unified view on existing approaches based on GAN, VAE and flow priors.

- We empirically show that our method achieves excellent reconstruction in the presence of noise with various complex and dependent structures. Specifically, we demonstrate the efficacy of our method on various inverse problems with structured noise and non-linear forward operators.

- We provide the initial theoretical characterization of likelihood-based priors for image denoising. Specifically, we show a reconstruction error bound that depends on the local concavity of the log-likelihood function.

2. Background

2.1. Normalizing Flow Models

Normalizing flow models are a class of likelihood-based generative models that represent complex distributions by transforming a simple distribution (such as standard Gaussian) through an invertible mapping (Tabak & Turner, 2013). Compared to other types of generative models, flow models are computationally flexible in that they provide efficient sampling, inversion and likelihood estimation (Papamakarios et al., 2019, and references therein).

Concretely, given a differentiable invertible mapping $G : \mathbb{R}^n \to \mathbb{R}^n$, the samples $\boldsymbol{x}$ from this model are generated via $\boldsymbol{z} \sim p_G(\boldsymbol{z}), \boldsymbol{x} = G(\boldsymbol{z})$. Since G is invertible, change of variables formula allows us to compute the log-density of $\boldsymbol{x}$:

$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}) + \log \left|\det J_{G^{-1}}(\boldsymbol{x})\right|, \qquad (2)$$

where $J_{G^{-1}}(\boldsymbol{x})$ is the Jacobian of G^{-1} evaluated at $\boldsymbol{x}$. Since $\log p(\boldsymbol{z})$ is a simple distribution, computing the likelihood at any point $\boldsymbol{x}$ is straightforward as long as G^{-1} and the log-determinant term can be efficiently evaluated.

Notably, when a flow model is used as the prior for an inverse problem, the invertibility of G guarantees that it has an unrestricted range. Thus the recovered signal can represent images that are out-of-distribution, albeit at lower probability. This is a key distinction from a GAN-based prior, whose generator has a restricted range and can only generate samples from the distribution it was trained on. As pointed out by Asim et al. (2019) and also shown below in our experiments, this leads to performance benefits on out-of-distribution examples.

2.2. Inverse Problems with a Generative Prior

We briefly review the existing literature on the application of deep generative models to inverse problems. While vast literature exists on compressed sensing and other inverse problems, the idea of replacing the classical sparsity-based prior (Candes et al., 2006; Donoho, 2006) with a neural network was introduced relatively recently. In their pioneering work, Bora et al. (2017) proposed to use the generator from a pre-trained GAN or a VAE (Goodfellow et al., 2014; Kingma & Welling, 2013) as the prior for compressed sensing. This led to a substantial gain in reconstruction quality compared to classical methods, particularly at small number of measurements.

Following this work, numerous studies have investigated different ways to utilize various neural network architectures for inverse problems (Mardani et al., 2018; Heckel & Hand, 2019; Mixon & Villar, 2018; Pandit et al., 2019; Lucas et al., 2018; Shah & Hegde, 2018; Liu & Scarlett, 2020; Kabkab et al., 2018; Mousavi et al., 2018; Raj et al., 2019; Sun et al., 2019). One straightforward extension of (Bora et al., 2017) proposes to expand the range of the pre-trained generator by allowing sparse deviations (Dhar et al., 2018). Similarly, Shah & Hegde (2018) proposed another algorithm based on projected gradient descent with convergence guarantees. Van Veen et al. (2018) showed that an *untrained* convolutional neural network can be used as a prior for imaging tasks based on Deep Image Prior by Ulyanov et al. (2018).

More recently, Wu et al. (2019) applied techniques from meta-learning to improve the reconstruction speed, and Ardizzone et al. (2018) showed that by modelling the forward process with a flow model, one can *implicitly* learn the inverse process through the invertibility of the model. Asim et al. (2019) proposed to replace the GAN prior of (Bora et al., 2017) with a normalizing flow model and reported excellent reconstruction performance, especially on out-of-distribution images.

3. Our Method

3.1. Notations and Setup

We use bold lower-case variables to denote vectors, $\|\cdot\|$ to denote ℓ_2 norm, and $\odot$ to denote element-wise multiplication. We also assume that we are given a pre-trained latent-variable generative model $p_G(\boldsymbol{x})$ that we can efficiently sample from. Importantly, we assume the access to a noise distribution p_Δ parametrized as a normalizing flow, which itself can be an arbitrarily complex, pre-trained distribution. We let f denote the deterministic and differentiable forward operator for our measurement process. Thus an observation is generated via $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\delta}$ where $\boldsymbol{x} \sim p_G(\boldsymbol{x})$ and $\boldsymbol{\delta} \sim p_\Delta(\boldsymbol{\delta})$.

Note that while f and p_Δ are known, they cannot be treated as fixed across different examples, e.g. in compressed sensing the measurement matrix is random and thus only known at the time of observation. This precludes the use of end-to-end training methods that require having a fixed forward operator.

3.2. MAP Formulation

When the likelihood under the prior $p_G(\boldsymbol{x})$ can be computed efficiently (e.g. when it is a flow model), we can pose the inverse problem as a MAP estimation task. Concretely, for a given observation $\boldsymbol{y}$, we wish to recover $\boldsymbol{x}$ as the MAP estimate of the conditional distribution $p_G(\boldsymbol{x}|\boldsymbol{y})$:

$$\begin{aligned}
&\arg\max_{\boldsymbol{x}} \log p(\boldsymbol{x}|\boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{x}} \left[\log p(\boldsymbol{y}|\boldsymbol{x}) + \log p_G(\boldsymbol{x}) - \log p(\boldsymbol{y})\right] \\
&\overset{(1)}{=} \arg\max_{\boldsymbol{x}} \left[\log p_\Delta(\boldsymbol{y} - f(\boldsymbol{x})) + \log p_G(\boldsymbol{x})\right] \\
&\triangleq \arg\min_{\boldsymbol{x}} L_{\text{MAP}}(\boldsymbol{x}; \boldsymbol{y}),
\end{aligned}$$

where

$$L_{\text{MAP}}(\boldsymbol{x}; \boldsymbol{y}) = -\log p_\Delta(\boldsymbol{y} - f(\boldsymbol{x})) - \log p_G(\boldsymbol{x}). \quad (3)$$

Note that in (1) we drop the marginal density $\log p(\boldsymbol{y})$ as it is constant and rewrite $p(\boldsymbol{y}|\boldsymbol{x})$ as $p_\Delta(\boldsymbol{y} - f(\boldsymbol{x}))$.

Recalling that the generative procedure for the flow model is $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I), \boldsymbol{x} = G(\boldsymbol{z})$, we arrive at the following loss:

$$L_{\text{MAP}}(\boldsymbol{z}; \boldsymbol{y}) \triangleq -\log p_\Delta(\boldsymbol{y} - f(G(\boldsymbol{z}))) - \log p_G(G(\boldsymbol{z})) \quad (4)$$

The invertibility of G allows us to minimize the above loss with respect to either $\boldsymbol{z}$ or $\boldsymbol{x}$:

$$\begin{aligned}
&\arg\min_{\boldsymbol{z}} \; L_{\text{MAP}}(\boldsymbol{z}; \boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{z}} \left[-\log p_\Delta\left(\boldsymbol{y} - f(G(\boldsymbol{z}))\right) - \log p_G(G(\boldsymbol{z}))\right] \\
&= \arg\min_{\boldsymbol{x}} \left[-\log p_\Delta\left(\boldsymbol{y} - f(\boldsymbol{x})\right) - \log p_G(\boldsymbol{x}))\right] \\
&= \arg\min_{\boldsymbol{x}} \; L_{\text{MAP}}(\boldsymbol{x}; \boldsymbol{y})
\end{aligned}$$

We have experimented with optimizing the loss both in image space $\boldsymbol{x}$ and latent space $\boldsymbol{z}$, and found that the latter achieved better performance across almost all experiments. Since the above optimization objective is differentiable, any gradient-based optimizer can be used to approximately find the minimizer. In practice, even with an imperfect model and approximate optimization, we observe that our approach performs well across a wide range of tasks as shown in the experimental results below.

3.3. MLE Formulation

When the signal prior does not provide tractable likelihood evaluation (e.g. for the case of GAN and VAE), we view the problem as a maximum-likelihood estimation under the *noise model*. Thus we attempt to find the signal that maximizes noise likelihood within the support of $p_G(\boldsymbol{x})$ and arrive at a similar, but different loss:

$$\begin{aligned}
&\arg\max_{\boldsymbol{x} \in \text{supp}\; p(\boldsymbol{x})} \; \log p_\Delta(\boldsymbol{y} - f(\boldsymbol{x})) \\
&= \arg\max_{\boldsymbol{z}} \log p_\Delta(\boldsymbol{y} - f(G(\boldsymbol{z}))) \\
&\triangleq \arg\min_{\boldsymbol{z}} L_{\text{MLE}}(\boldsymbol{z}; \boldsymbol{y}),
\end{aligned}$$

where

$$L_{\text{MLE}}(\boldsymbol{z}; \boldsymbol{y}) \triangleq -\log p_\Delta(\boldsymbol{y} - f(G(\boldsymbol{z}))). \quad (5)$$

3.4. Prior Work

In (Bora et al., 2017), the authors proposed to use a deep generative prior for inverse problem, but the choice of models was restricted to GANs and VAEs with explicit low-dimensional prior. Subsequently Asim et al. (2019) generalized this paradigm using Flow-based models. We describe here the methods proposed in those papers in detail. Importantly, we show that their approaches are special cases of our MAP/MLE formulations, under Gaussian noise assumptions. Furthermore, note that both papers considered linear inverse problems, so they correspond to the case where $f(\boldsymbol{x}) = A\boldsymbol{x}$ under our notation.

GAN Prior: (Bora et al., 2017) considers the following loss:

$$L_{\text{Bora}}(\boldsymbol{z}; \boldsymbol{y}) = \|\boldsymbol{y} - AG(\boldsymbol{z})\|^2 + \lambda \|z\|^2, \quad (6)$$

which tries to project the input $\boldsymbol{y}$ onto the range of the generator G with ℓ_2 regularization on the latent variable. Aside from the regularization term, this corresponds exactly to our MLE loss for a Gaussian p_Δ. While (Bora et al., 2017) motivated this objective as a projection on the range of G, our approach reveals a probabilistic interpretation based on the MLE objective for the noise.

Flow Prior: (Asim et al., 2019) replaces the GAN prior of (Bora et al., 2017) with a flow model. In that paper the authors consider the objective below that tries to simultaneously match the observation and maximize the likelihood of the reconstruction under the model:

$$L(\boldsymbol{z}; \boldsymbol{y}) = \|\boldsymbol{y} - AG(\boldsymbol{z})\|^2 - \gamma \log p_G(\boldsymbol{x}), \quad (7)$$

for some hyperparameter $\gamma > 0$. This loss is a special case of our MAP loss for isotropic Gaussian noise $\boldsymbol{\delta} \sim$

$\mathcal{N}(\mathbf{0}, \gamma I)$, since the log-density of $\boldsymbol{\delta}$ becomes $\log p_\Delta(\boldsymbol{\delta}) = -\frac{1}{2\gamma}\|\boldsymbol{\delta}\|^2 - C$ for a constant C. However, Asim et al. (2019) report that due to optimization difficulty, they found the following proxy loss to perform better in experiments:

$$L_{\text{Asim}}(\boldsymbol{z}; \boldsymbol{y}) = \|\boldsymbol{y} - AG(\boldsymbol{z})\|^2 + \gamma \|\boldsymbol{z}\|. \tag{8}$$

This is again related to a specific instance of our loss when the flow model is volume-preserving (i.e. the log-determinant term is constant). Continuing from eq. (2): This allows us to recover the ℓ_2-regularized version of L_{Asim}:

We reiterate that both of the aforementioned objectives are special cases of our formulation for the case of zero-mean isotropic Gaussian noise. Thus we expect that our method better handles non-Gaussian noise and experimentally confirm that our approach indeed leads to better reconstruction performance for noises with nonzero mean or conditional dependence across different pixel locations.

Connections to Blind Source Separation: Here we focus on the connection between our formulation and blind source separation. For the denoising case where we have identity forward operator, we see that our observation is simply the sum of two random variables $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\delta}$. Given two flow-based priors (one for each of $\boldsymbol{x}$ and $\boldsymbol{\delta}$), the task of extracting $\boldsymbol{x}$ from $\boldsymbol{y}$ thus becomes a blind source separation problem with two sources. While rich literature exists for various source separation problems (Hu et al., 2017; Subakan & Smaragdis, 2018; Wang & Chen, 2018; Hoshen, 2019), two recent studies are particularly relevant to our setting as they make use of a neural network prior.

In Double-DIP, Gandelsman et al. (2019) utilize Deep Image Prior (Ulyanov et al., 2018) as a signal prior to perform blind source separation from multiple mixtures. This work differs from ours in that we focus on single-mixture setting with pre-trained signal priors. Our use of pre-trained priors is a key distinction, since DIP is untrained and may not apply to other modalities and datasets. In contrast, our method is applicable as long as we are able to train a deep generative prior for the signal and the noise.

In (Jayaram & Thickstun, 2020), the authors use a flow-based prior (Kingma & Dhariwal, 2018) for blind source separation. Unlike our approach, however, they sample from the posterior using Langevin dynamics (Welling & Teh, 2011; Neal et al., 2011). The authors use simulated annealing to speed up mixing, and this approach would in theory be able to asymptotically sample from the correct posterior. The advantage of our approach is that it is generally faster (as it avoids costly MCMC procedure), and it can be applied to non-likelihood based priors for the signal $\boldsymbol{x}$.

4. Theoretical Analysis

In this section we provide some theoretical analysis of our approach in denoising problems with a flow-based prior. Unlike most prior work, we take a probabilistic approach and avoid making specific structural assumptions on the signal we want to recover, such as sparsity or being generated from a low-dimensional Gaussian prior.

For denoising, we show that better likelihood estimates lead to lower reconstruction error. Note that while our experiments employed flow models, our results apply to any likelihood-based generative model. Detailed proof is included in the appendix.

4.1. Recovery Guarantee for Denoising

Suppose we observe $\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{\delta}$ with Gaussian noise $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ with $\|\boldsymbol{\delta}\| = r$. We perform MAP inference by minimizing the following loss with gradient descent:

$$\begin{aligned} L_{\text{MAP}}(\boldsymbol{x}) &= -\log p_\Delta(\boldsymbol{y} - \boldsymbol{x}) - \log p(\boldsymbol{x}) \\ &= \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + q(\boldsymbol{x}), \end{aligned} \tag{9}$$

where we write $q(\boldsymbol{x}) \triangleq -\log p(\boldsymbol{x})$ for notational convenience. Notice that the image we wish to recover is a natural image with high probability rather than an arbitrary one, and reconstruction is not expected to succeed for the latter case. Thus we consider the case where the ground truth image $\boldsymbol{x}^*$ is a local maximum of p.

Theorem 4.1. *Let $\boldsymbol{x}^*$ be a local optimum of the model $p(\boldsymbol{x})$ and $\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{\delta}$ be the noisy observation. Assume that q satisfies local strong convexity within the ball around $\boldsymbol{x}^*$ defined as $B_r^d(\boldsymbol{x}^*) \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}^*\| \le r\}$, i.e. the Hessian of q satisfies $H_q(\boldsymbol{x}) \succeq \mu I \; \forall \boldsymbol{x} \in B_r^d(\boldsymbol{x}^*)$ for for some $\mu > 0$. Then gradient descent starting from $\boldsymbol{y}$ on the loss function (9) converges to $\bar{\boldsymbol{x}}$, a local minimizer of $L_{MAP}(\boldsymbol{x})$, that satisfies:*

$$\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\| \le \frac{1}{\mu\sigma^2 + 1}\|\boldsymbol{\delta}\|$$

Even though the theorem is quite straightforward, it still serves as some initial understanding on the denoising task under a likelihood-based prior. It sheds light on how the reconstruction is affected by the structure of the probabilistic model and the likelihood of the natural signal one wants to recover. This theorem shows that a well-conditioned model with large μ leads to better denoising and confirms that our MAP formulation encourages reconstructions with high density. Thus, the maximum-likelihood training objective is directly aligned with better denoising performance.

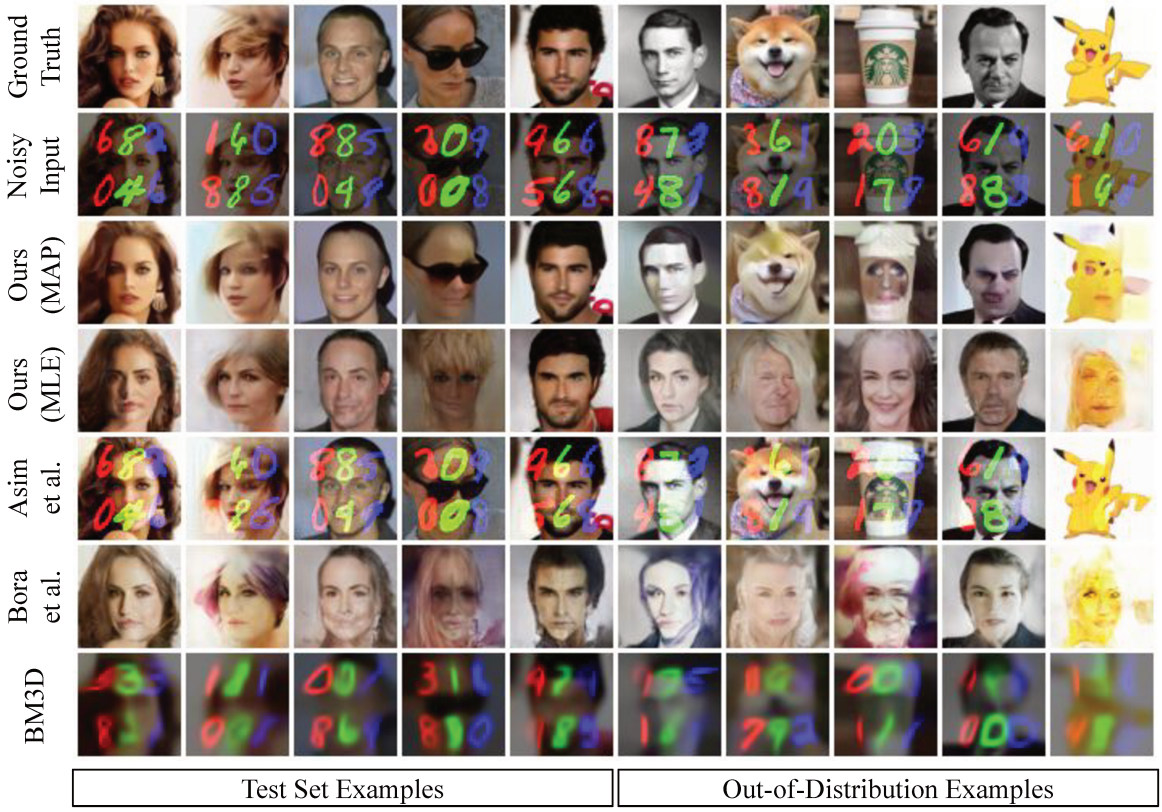


Figure 2. Result of removing MNIST noise from CelebA-HQ faces. Notice that without any understanding of the complex noise structure, baseline methods fail to produce good reconstructions.

5. Experiments

Our experiments are designed to test how well our MAP/MLE formulation performs in practice as we deviate from the commonly studied setting of linear inverse problems with Gaussian noise. Specifically, we focus on two aspects: (1) complex noise with dependencies and (2) non-linear forward operator. For all our experiments, we quantitatively evaluate each method by reporting peak signal-to-noise ratio (PSNR). We also visually inspect sample reconstructions for qualitative assessment.

Models: We trained multi-scale RealNVP models on two image datasets MNIST and CelebA-HQ (LeCun et al., 1998; Liu et al., 2015). Due to computational constraints, all experiments were done on 100 randomly-selected images (1000 for MNIST) from the test set, as well as out-of-distribution images. We additionally train a DCGAN on the CelebA-HQ dataset for MLE experiments as well as (Bora et al., 2017). A detailed description of the datasets, models and hyperparameters are provided in the appendix.

Baseline Methods: We compare our approach to the methods of (Bora et al., 2017) and (Asim et al., 2019), as they are two recently proposed approaches that use deep generative prior on inverse problems. Depending on the task, we also compare against BM3D (Dabov et al., 2006), a popular off-the-shelf image denoising algorithm, and LASSO (Tibshirani, 1996) with Discrete Cosine Transform (DCT) basis as appropriate. Note that for the 1-bit compressed sensing experiment, most existing techniques do not apply, since our task involves quantization as well as non-Gaussian noise.

We point out that the baselines methods are not designed to make use of the noise distribution, whereas our method does utilize it. Thus, the experiments are not meant to be taken as direct comparisons, but rather as an empirical evidence that the MAP formulation indeed benefits from the knowledge of noise structure.

Smoothing Parameter: Since our objective Equation (4) depends on the density $p_G(\boldsymbol{x})$ given by the flow model, our recovery of $\boldsymbol{x}$ depends heavily on the quality of density estimates from the model. Unfortunately likelihood-based models exhibit counter-intuitive properties, such as assigning higher density on out-of-distribution examples or random noise over in-distribution examples. (Nalisnick et al., 2018; Choi et al., 2018; Hendrycks & Dietterich, 2018; Nalisnick et al., 2019). We empirically observe such behavior from our models as well. To remedy this, we use a smoothed

version of the model density $p_G(x)^\beta$ where $\beta \geq 0$ is the *smoothing parameter*. Since the two extremes $\beta = 0$ and $\beta = \infty$ correspond to only using the noise density and the model density, respectively, β controls the degree to which we rely on p_G. Thus the loss we minimize becomes

$$L_{\text{MAP}}(\boldsymbol{z}; \boldsymbol{y}, \beta) = -\log p_\Delta(\boldsymbol{y} - f(G(\boldsymbol{z}))) - \beta \log p_G(G(\boldsymbol{z})) \tag{10}$$

5.1. Results

We test our methods on denoising and compressed sensing tasks that involve various structured, non-Gaussian noises as well as a non-linear forward operator. Note that many existing methods cannot be applied in this setting as they are designed for linear inverse problems and a specific noise model. While specialized algorithms do exist (e.g. for quantized compressed sensing), we note that our method is general and can be applied in a wide range of settings without modification.

5.1.1. Denoising MNIST Digits

The measurement process is $\boldsymbol{y} = 0.5 \cdot \boldsymbol{x} + \boldsymbol{\delta}_{\text{MNIST}}$, where $\boldsymbol{\delta}_{\text{MNIST}}$ represents MNIST digits added at different locations and color channels. Each digit itself comes from a flow model trained on the MNIST dataset. As shown in Figure 2, our method successfully removes MNIST noise in both MAP and MLE settings. Recall that we use the same GAN-based prior for "Ours (MLE)" and (Bora et al., 2017). The difference in the reconstruction quality between these two methods in Figure 3 confirms that even for non-likelihood based priors (e.g. GAN and VAE), an accurate noise model is critical to accurate signal recovery. While the method of Bora et al. (2017) manages to remove MNIST digits, we note that this is because its outputs are forced to be in the range of the DCGAN.

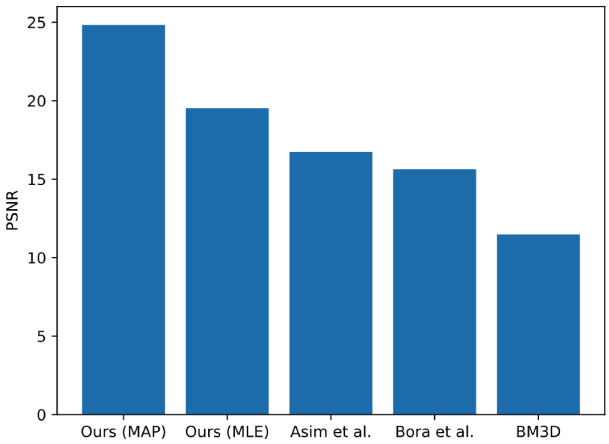


Figure 3. Reconstruction PSNR on the MNIST denoising task.

For the rest of the experiments, we focus on the MAP formulation as it generally outperforms the MLE formulation. We posit that this is due to the restricted range of our DCGAN used for experiments.

5.1.2. Noisy Compressed Sensing

Now we consider the measurement process $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\delta}_{\text{sine}}$ where $A \in \mathbb{R}^{m \times d}$ is a random Gaussian measurement matrix and $\boldsymbol{\delta}$ has positive mean with variance that follows a sinusoidal pattern. Specifically, the k-th pixel of the noise has standard deviation $\sigma_k \propto \exp\left(\sin(\frac{2\pi k}{16})\right)$ normalized to have the maximum variance of 1.

Figure 4a shows that our method is able to make a better use of additional measurements. Interestingly in Figure 4b, all three methods with deep generative prior produced plausible human faces. However, the reconstructions from Asim et al. (2019) and Bora et al. (2017) significantly differ from the ground truth images. We posit that this is due to the implicit Gaussian noise assumption made by the two methods, again showing the benefits of explicitly incorporating the knowledge of noise distribution.

5.1.3. Removing Sinusoidal Noise

We consider another denoising task with observation $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\delta}_{\text{sine}}$. This is the 2-dimensional version of periodic noise, where the noise variance for all pixels within the k-th row follows σ_k from above.

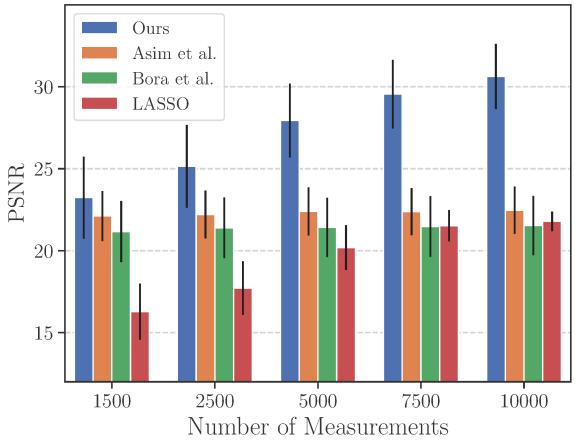
From Figure 5 and Figure 6a, we see that baseline methods do not perform well even though the noise is simply Gaussian at each pixel. This reemphasizes an important point: without an understanding of the *structure* of the noise, algorithms designed to handle Gaussian noise have difficulty removing them when we introduce variability across different pixel locations.

5.1.4. Noisy 1-bit Compressed Sensing

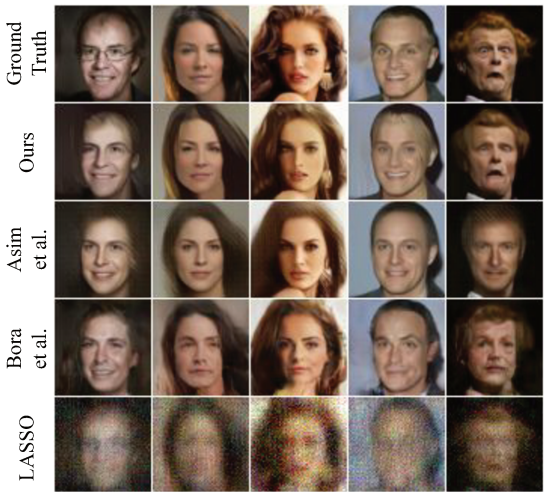
This task considers a combination of a non-linear forward operator as well as a non-Gaussian noise. The measurement process is $\boldsymbol{y} = \text{sign}(A\boldsymbol{x}) + \boldsymbol{\delta}_{\text{sine}}$, identical to noisy compressed sensing except with the sign function. This is the most extreme case of quantized compressed sensing, since $\boldsymbol{y}$ only contains the (noisy) sign $\{+1, -1\}$ of the measurements. Because the gradient of sign function is zero everywhere, we use Straight-Through Estimator (Bengio et al., 2013) for backpropagation. See Figure 7a and Figure 7b for a comparison of our method to the baselines at varying numbers of measurements.

5.1.5. Sensitivity to Hyperparameters

We also observe that our method is more robust to hyperparameter choices than (Asim et al., 2019), as shown



(a) PSNR at different measurement counts (best viewed in color). The approaches by Asim et al. (2019) and Bora et al. (2017) show little improvements from having more measurements due to their inability to utilize the noise model.



(b) Reconstructions at 2500 measurements. Notice that even though existing approaches produce reconstructions that resemble human faces, they do not match the ground truth as well as our method.

Figure 4. Experiment results for noisy compressed sensing on CelebA-HQ images.

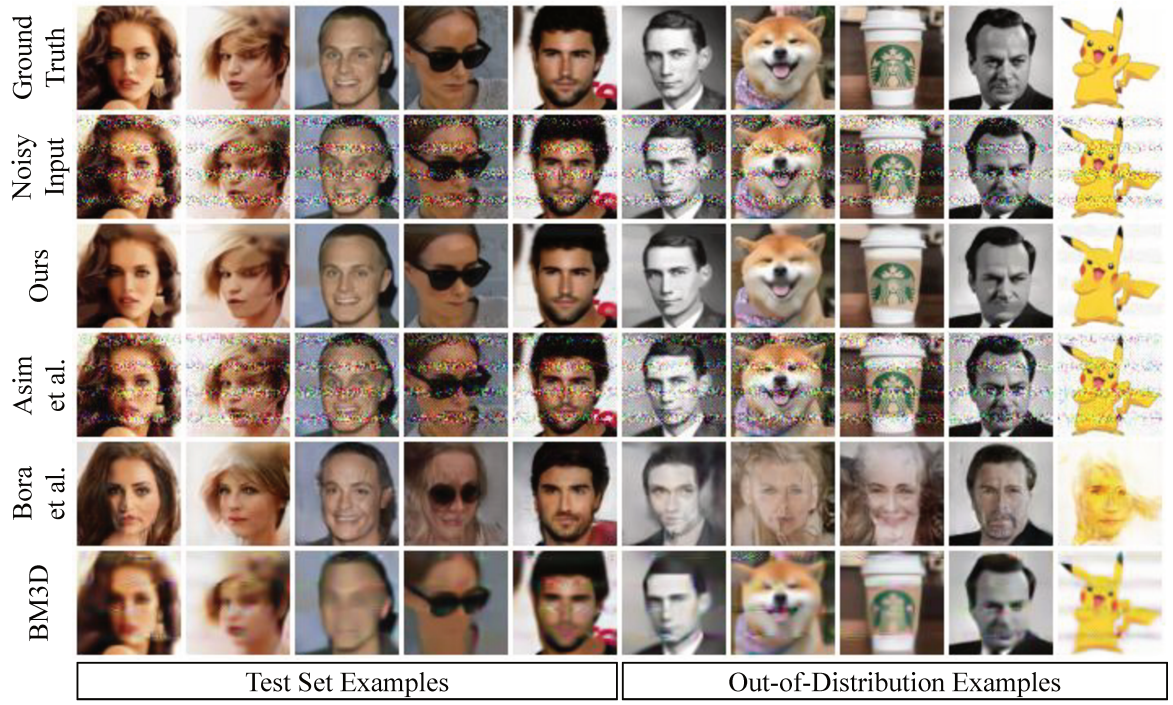
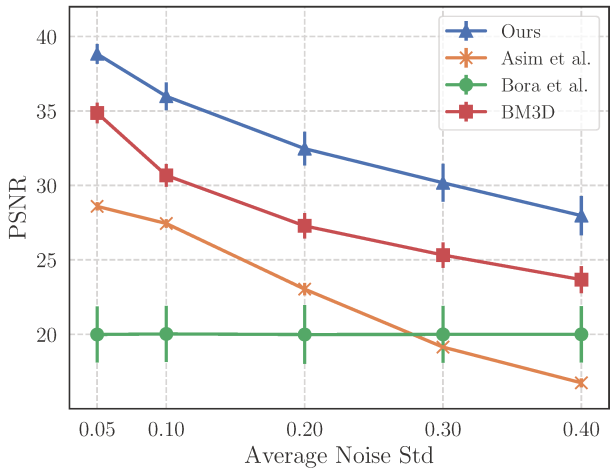


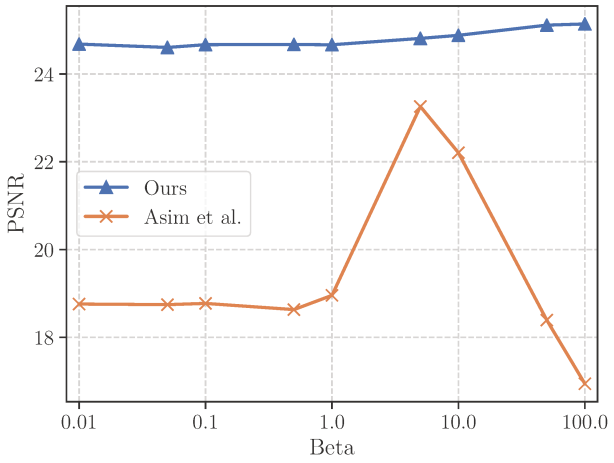
Figure 5. Result of denoising SINUSOIDAL noise on CelebA-HQ faces and out-of-distribution images.

in Figure 6b. This may come as a surprise given that our objective has an additional log-determinant term. We speculate that this is due to the regularization term in $L_{\text{Asim}}(\boldsymbol{z}; \boldsymbol{y}) = \|\boldsymbol{y} - f(G(\boldsymbol{z}))\|^2 + \gamma \|\boldsymbol{z}\|$.

It is known that samples from d-dimensional isotropic Gaussian concentrate around a thin "shell" around the sphere of radius $\sqrt{d}$. This suggests that the range of $\|\boldsymbol{z}\|$ corresponding to natural images may be small. Thus, forcing the latent variable $\boldsymbol{z}$ to have small norm without taking the log-determinant term into account could lead to a sudden drop in the reconstruction quality.

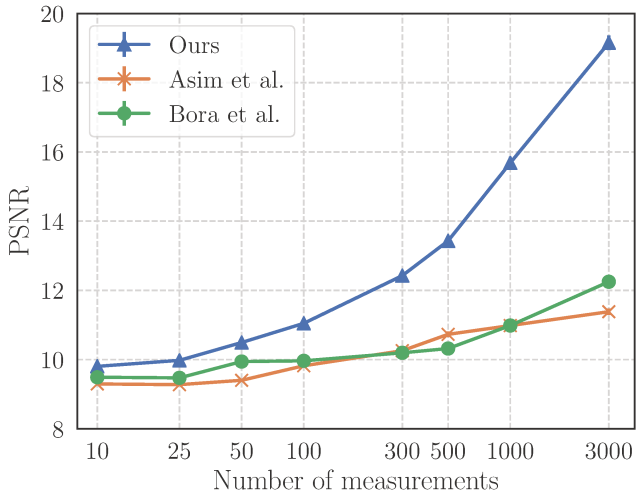


(a) Result for SINUSOIDAL denoising on CelebA-HQ images at various noise rates. Our method achieves the same reconstruction performance even when the noise has up to 3× higher average standard deviation compared to the best baseline method (BM3D).

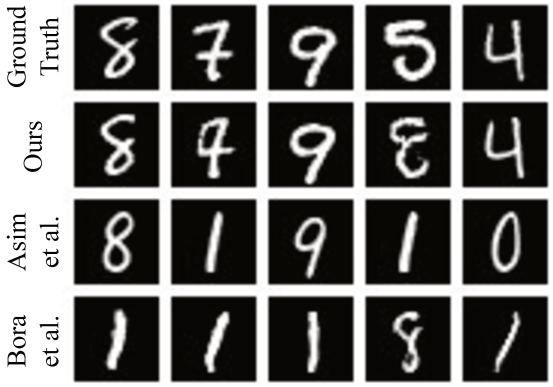


(b) Compressed sensing performance of our method and (Asim et al., 2019) at different hyperparameter values. For our method, we vary the smoothing parameter β in L_{MAP}. For Asim et al. (2019), we vary the regularization coefficient γ in L_{Asim}.

Figure 6. SINUSOIDAL denoising results (left) and hyperparameter sensitivity plot (right).



(a) Result of 1-bit noisy compressed sensing at different measurement counts. Our method achieves the same reconstruction performance using up to 2× fewer measurements compared to the best baseline method (Asim et al., 2019).



(b) Reconstructions from noisy 1-bit compressed sensing with 3000 binary measurements. Notice that our method fails more gracefully compared to other methods, i.e. even when the reconstructions differ from the ground truth, substantial parts of the reconstructions are still correct. On the other hand, other methods predict a completely different digit.

Figure 7. Results of 1-bit compressed sensing experiments.

6. Conclusion

We propose a novel method to solve inverse problems for general differentiable forward operators and structured noise. Our method generalizes that of (Asim et al., 2019) to arbitrary differentiable forward operators and non-Gaussian noise distributions. The power of our approach stems from the flexiblity of invertible generative models which can be combined in a modular way to solve MAP inverse problems in very general settings, as we demonstrate.

For future work, it would be interesting to consider extending our method to blind source separation problems (Amari et al., 1996), as the noise and signal considered in our setting are actually exchangeable for the MAP case. Furthermore, one may investigate the applicability of the MLE formulation with even more general family of generative models such as energy-based models (Gao et al., 2020) and score

networks (Song & Ermon, 2019). On the theoretical side, one central question that remains open is to analyze the optimization problem we formulated. In this paper we empirically minimize this loss using gradient descent, but some theoretical guarantees would be desirable, possibly under assumptions, e.g. random weights following the framework of (Hand & Voroninski, 2020).

Acknowledgements

This research has been supported by NSF Grants CCF 1934932, AF 1901292, 2008710, 2019844 the NSF IFML 2019844 award as well as research gifts by Western Digital, WNCG and MLL, computing resources from TACC and the Archie Straiton Fellowship.

References

Amari, S.-i., Cichocki, A., Yang, H. H., et al. A new learning algorithm for blind signal separation. In *Advances in neural information processing systems*, pp. 757–763, 1996.

Ardizzone, L., Kruse, J., Wirkert, S. J., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U. Analyzing inverse problems with invertible neural networks. *CoRR*, abs/1808.04730, 2018. URL http://arxiv.org/abs/1808.04730.

Asim, M., Ahmed, A., and Hand, P. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *CoRR*, abs/1905.11672, 2019. URL http://arxiv.org/abs/1905.11672.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. *IEEE Transactions on information theory*, 56(4):1982–2001, 2010.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 537–546. JMLR. org, 2017.

Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59 (8):1207–1223, 2006.

Candes, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. Phase retrieval via matrix completion. *SIAM review*, 57 (2):225–251, 2015a.

Candes, E. J., Li, X., and Soltanolkotabi, M. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015b.

Chen, G.-H., Tang, J., and Leng, S. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Medical physics*, 35 2:660–3, 2008.

Choi, H., Jang, E., and Alemi, A. A. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Image denoising with block-matching and 3d filtering. In *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, volume 6064, pp. 606414. International Society for Optics and Photonics, 2006.

Dhar, M., Grover, A., and Ermon, S. Modeling sparse deviations for compressed sensing using generative models, 2018.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Gandelsman, Y., Shocher, A., and Irani, M. " double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11026–11035, 2019.

Gao, R., Nijkamp, E., Kingma, D. P., Xu, Z., Dai, A. M., and Wu, Y. N. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Transactions on Information Theory*, 2020.

Heckel, R. and Hand, P. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations*, 2019.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

Hoshen, Y. Towards unsupervised single-channel blind source separation using adversarial pair unmix-and-remix. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3272–3276. IEEE, 2019.

Hu, W., Liu, R., Lin, X., Li, Y., Zhou, X., and He, X. A deep learning method to estimate independent source number. In *2017 4th International Conference on Systems and Informatics (ICSAI)*, pp. 1055–1059, 2017. doi: 10.1109/ICSAI.2017.8248441.

Jayaram, V. and Thickstun, J. Source separation with deep generative priors. In *International Conference on Machine Learning*, pp. 4724–4735. PMLR, 2020.

Kabkab, M., Samangouei, P., and Chellappa, R. Task-aware compressed sensing with generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 292–303, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lucas, A., Iliadis, M., Molina, R., and Katsaggelos, A. K. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Mardani, M., Sun, Q., Donoho, D., Papyan, V., Monajemi, H., Vasanawala, S., and Pauly, J. Neural proximal gradient descent for compressive imaging. In *Neural Information Processing Systems*, pp. 9573–9583, 2018.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2437–2445, 2020.

Mixon, D. G. and Villar, S. Sunlayer: Stable denoising with generative networks. *arXiv preprint arXiv:1803.09319*, 2018.

Mousavi, A., Dasarathy, G., and Baraniuk, R. G. A data-driven and distributed approach to sparse signal representation and recovery. In *International Conference on Learning Representations*, 2018.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Nalisnick, E., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Pandit, P., Sahraee, M., Rangan, S., and Fletcher, A. K. Asymptotics of map inference in deep networks. *arXiv preprint arXiv:1903.01293*, 2019.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

Raj, A., Li, Y., and Bresler, Y. Gan-based projector for faster recovery with convergence guarantees in linear inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5602–5611, 2019.

Shah, V. and Hegde, C. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4609–4613. IEEE, 2018.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2019.

Subakan, Y. C. and Smaragdis, P. Generative adversarial source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 26–30. IEEE, 2018.

Sun, Y., Liu, J., and Kamilov, U. S. Block coordinate regularization by denoising. *NeurIPS*, 2019.

Tabak, E. G. and Turner, C. V. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Van Veen, D., Jalal, A., Soltanolkotabi, M., Price, E., Vishwanath, S., and Dimakis, A. G. Compressed sensing with deep image prior and learned regularization. *arXiv preprint arXiv:1806.06438*, 2018.

Wang, D. and Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10): 1702–1726, 2018.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Wu, Y., Rosca, M., and Lillicrap, T. P. Deep compressed sensing. *CoRR*, abs/1905.06723, 2019. URL http://arxiv.org/abs/1905.06723.

A. Omitted Proof

A.1. Proof for Denoising

Proof of Theorem 4.1. We first show that gradient descent with sufficiently small learning rate will converge to $\bar{\boldsymbol{x}}$, the locally-optimal solution of Equation (9). Recall the loss function $L(\boldsymbol{x}) := q(\boldsymbol{x}) + \frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{y}\|^2$ (we subsume the scaling $\frac{1}{2}$ into $\frac{1}{\sigma^2}$ without loss of generality). Notice in the ball $B_r^d(\boldsymbol{x}^*) := \left\{\boldsymbol{x}\in\mathbb{R}^d \,|\, \|\boldsymbol{x}-\boldsymbol{x}^*\| \le r\right\}$, L is $\left(\mu+\frac{1}{\sigma^2}\right)$ strongly-convex. We next show there is a stationary point $\bar{\boldsymbol{x}} \in B_r^d(\boldsymbol{x}^*)$ of $L(\boldsymbol{x})$.

$$\begin{aligned}\nabla L(\bar{\boldsymbol{x}}) = 0 &\implies \nabla q(\bar{\boldsymbol{x}}) + \frac{1}{\sigma^2}(\bar{\boldsymbol{x}}-\boldsymbol{y}) = 0\\ &\implies \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*) = \frac{1}{\sigma^2}(\boldsymbol{y}-\bar{\boldsymbol{x}})\\ &\implies \langle \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*), \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle\\ &= \frac{1}{\sigma^2}\langle \boldsymbol{y}-\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle\end{aligned}$$

From strong convexity of q,

$$\langle \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*), \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle \ge \mu\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\|^2.$$

Thus,

$$\begin{aligned}&\frac{1}{\sigma^2}\langle \boldsymbol{y}-\boldsymbol{x}^*, \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle\\ &= \frac{1}{\sigma^2}\langle (\boldsymbol{y}-\bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}}-\boldsymbol{x}^*), \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle\\ &= \frac{1}{\sigma^2}\langle \boldsymbol{y}-\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle + \frac{1}{\sigma^2}\langle \bar{\boldsymbol{x}}-\boldsymbol{x}^*, \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle\\ &= \langle \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*), \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle + \frac{1}{\sigma^2}\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\|^2\\ &\ge \mu\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\|^2 + \frac{1}{\sigma^2}\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\|^2\\ &= \left(\mu+\frac{1}{\sigma^2}\right)\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\|^2\end{aligned}$$

Finally, by Cauchy-Schwartz inequality,

$$\langle \boldsymbol{y}-\boldsymbol{x}^*, \bar{\boldsymbol{x}}-\boldsymbol{x}^*\rangle \le \|\boldsymbol{y}-\boldsymbol{x}^*\|\cdot\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\|.$$

So we get $\|\bar{\boldsymbol{x}}-\boldsymbol{x}^*\| \le \frac{1}{1+\mu\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}^*\| \le \|\boldsymbol{\delta}\| \le r$, in other words, $\bar{\boldsymbol{x}} \in B_r^d(\boldsymbol{x}^*)$.

Notice L is $\left(\mu+\frac{1}{\sigma^2}\right)$ strongly-convex in $B_r^d(\boldsymbol{x}^*)$, which contains the stationary point $\bar{\boldsymbol{x}}$. Therefore $\bar{\boldsymbol{x}}$ is a local minimizer of $L(\boldsymbol{x})$. Also note that we implicitly require q to be twice differentiable, meaning in a compact set $B_r^d(\boldsymbol{x}^*)$ its smoothness is upper bounded by a constant M. Thus gradient descent starting from $\boldsymbol{y} \in B_r^d(\boldsymbol{x}^*)$ with learning rate smaller than $\frac{1}{M}$ will converge to $\bar{\boldsymbol{x}}$ without leaving the (convex) set $B_r^d(\boldsymbol{x}^*)$. □

B. Additional Experimental Results

Here we include experimental results and details not included in the main text. Across all the experiments, we individually tuned the hyperparameters for each method.

B.1. Experimental Details

Dataset. For MNIST, we used the default split of 60,000 training images and 10,000 test images of (LeCun et al., 1998). For CelebA-HQ, we used the split of 27,000 training images and 3,000 test images as provided by (Kingma & Dhariwal, 2018).

During evaluation, the following Python script was used to select 1000 MNIST images and 100 CelebA-HQ images from their respective test sets:

```
np.random.seed(0)
indices_mnist = np.random.choice(
    10000, 1000, False)
np.random.seed(0)
indices_celeba = np.random.choice(
    3000, 100, False)
```

Note that CelebA-HQ images were further resized to 64×64 resolution.

Noise Distributions. For the sinusoidal noise used in the experiments, the standard deviation of the k-th pixel/row is calculated as:

$$\sigma_k = 0.1\cdot\left(\exp\left(\sin(2\pi\cdot\frac{k}{16})\right)-1\right)/(e-1),$$

clamped to be in range $[0.001, 1]$. For Figure 9b, we used vary the coefficient 0.1 to values in $\{0.05, 0.1, 0.2, 0.3, 0.4\}$.

For the radial noise used in the additional experiment below, the standard deviation of each pixel with ℓ_2 distance is d from the center pixel (31, 31) is computed as: $\sigma_k = 0.1\cdot\exp(-0.005\cdot d^2)$, clamped to be in range $[0.001, 1000]$.

B.2. Additional Result: Removing RADIAL Noise

Consider the measurement process $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\delta}_{\text{radial}}$, where each pixel follows a Gaussian distribution, but with variance that decays exponentially in distance to the center point. For a pixel whose ℓ_2 distance to the center pixel is d, the standard deviation is computed as $\sigma(d) = \exp\left(-0.005\cdot d^2\right)$. See Figure 8 and Figure 9a for reconstructions as well as PSNR plot comparing the methods considered.

B.3. Additional Result: 1-bit Compressed Sensing

Figure 9b shows the performance of each method at different noise scales for a fixed number of measurements. We

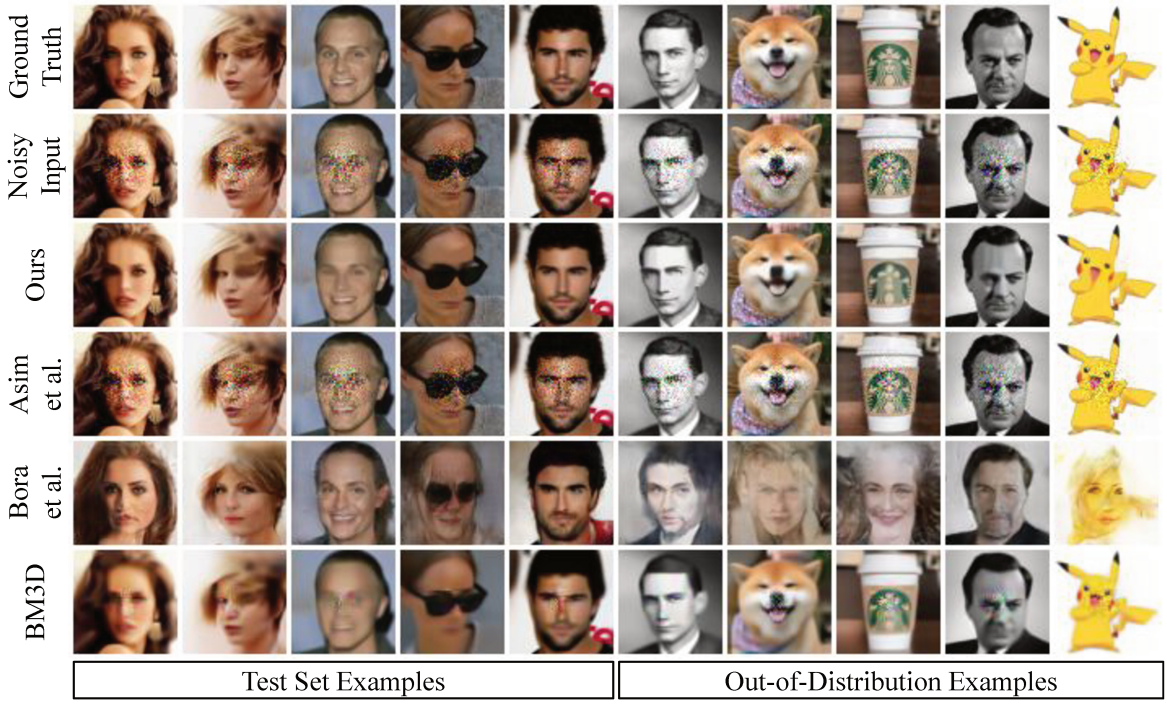
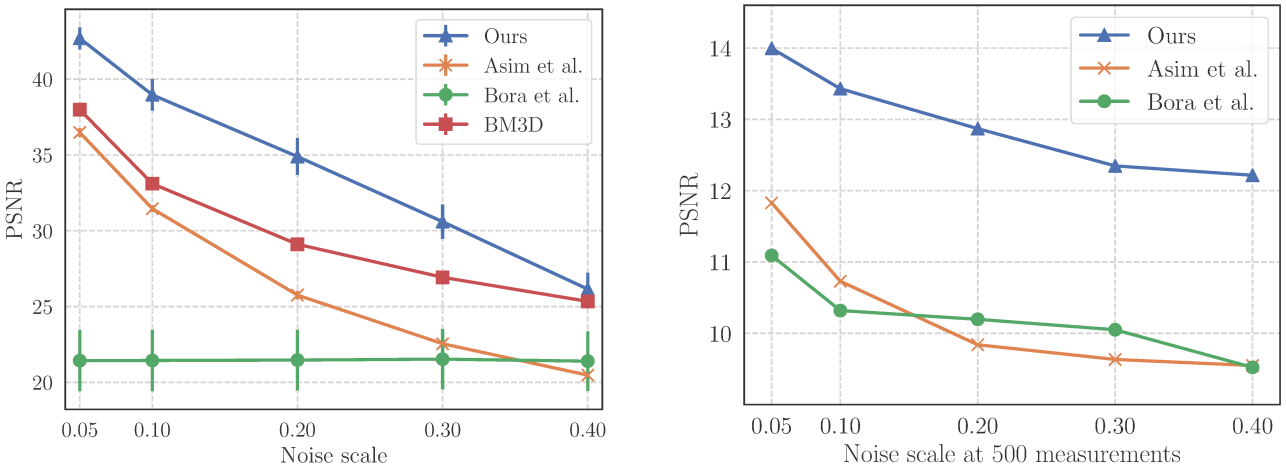


Figure 8. Result of denoising RADIAL noise on CelebA-HQ faces and out-of-distribution images.



(a) Result on denoising RADIAL noise at varying noise rates. Our method achieves the same reconstruction performance even when the noise has approximately $1.5\times$ higher noise scale compared to the best baseline method which is BM3D for this setting.

(b) Result of 1-bit compressed sensing at different noise scale. Our method obtains the best reconstructions, achieving similar PSNR as (Asim et al., 2019) when the noise scale is $8\times$ higher.

Figure 9. RADIAL denoising results (left) and 1-bit compressed sensing results at different noise levels (right).

observe that our method performs consistently better at all noise levels.

C. Model Architecture and Hyperparameters

For the RealNVP models we trained, we used multiscale architecture as was done in (Dinh et al., 2016), with residual networks and regularized weight normalization on convolutional layers. Following (Kingma & Dhariwal, 2018), we used 5-bit color depth for the CelebA-HQ model. Hyperparameters and samples from the models can be found in Table 1 and Figure 10.

Hyperparameter	CelebA-HQ	MNIST
Learning rate	5e−4	1e−3
Batch size	16	128
Image size	$64 \times 64 \times 3$	$28 \times 28 \times 1$
Pixel depth	5 bits	8 bits
Number of epochs	300	200
Number of scales	6	3
Residual blocks per scale	10	6
Learning rate halved every	60 epochs	40 epochs
Max gradient norm	500	100
Weightnorm regularization	1e−5	5e−5

Table 1. Hyperparameters used for RealNVP models.

Figure 10. Samples from the RealNVP models used in our experiments.

Figure 11. Out-of-distribution images used in our experiments. We included different types of out-of-distribution instances including grayscale images and cartoons with flat image areas.

D. Experiment Hyperparameters

Here we list the hyperparameters used for each experiment. We used the Adam optimizer (Kingma & Ba, 2014) for all appropriate methods below.

Denoising MNIST Digits.

- Learning rate: 0.02
- Optimization steps for Ours (MAP) and (Asim et al., 2019): 400
- Optimization steps for Ours (MLE) and (Bora et al., 2017): 1000
- Smoothing parameter for Ours (MAP & MLE): $\beta = 1.0$
- Regularization for (Asim et al., 2019): $\gamma = 0.0$
- Regularization for (Bora et al., 2017): $\lambda = 0.01$

Noisy Compressed Sensing.

- Learning rate: 0.02
- Optimization steps for Ours (MAP) and (Asim et al., 2019): 300
- Optimization steps for (Bora et al., 2017): 1000
- Smoothing parameter for Ours (MAP): $\beta = 100$
- Regularization for (Asim et al., 2019): $\gamma = 10$
- Regularization for (Bora et al., 2017): $\lambda = 0.001$
- Regularization for LASSO: $\lambda = 0.01$

Denoising Sinusoidal Noise.

- Learning rate: 0.02
- Optimization steps for Ours (MAP) and (Asim et al., 2019): 150
- Optimization steps for (Bora et al., 2017): 1000
- Smoothing parameter for Ours (MAP): $\beta = 0.5$
- Regularization for (Asim et al., 2019): $\gamma = 2.0$
- Regularization for (Bora et al., 2017): $\lambda = 0.01$

Noisy 1-bit Compressed Sensing.

- Learning rate: 0.02
- Optimization steps for Ours (MAP) and (Asim et al., 2019): 200
- Optimization steps for (Bora et al., 2017): 1000
- Smoothing parameter for Ours (MAP): $\beta = 1.0$
- Regularization for (Asim et al., 2019): $\gamma = 1.0$
- Regularization for (Bora et al., 2017): $\lambda = 0.01$