Invertible generative models for inverse problems: mitigating representation error and dataset bias

Muhammad Asim [*1] **Max Daniels** [*2] **Oscar Leong** [3] **Ali Ahmed** [†1] **Paul Hand** [†2]

Abstract

Trained generative models have shown remarkable performance as priors for inverse problems in imaging – for example, Generative Adversarial Network priors permit recovery of test images from 5-10x fewer measurements than sparsity priors. Unfortunately, these models may be unable to represent any particular image because of architectural choices, mode collapse, and bias in the training dataset. In this paper, we demonstrate that invertible neural networks, which have zero representation error by design, can be effective natural signal priors at inverse problems such as denoising, compressive sensing, and inpainting. Given a trained generative model, we study the empirical risk formulation of the desired inverse problem under a regularization that promotes high likelihood images, either directly by penalization or algorithmically by initialization. For compressive sensing, invertible priors can yield higher accuracy than sparsity priors across almost all undersampling ratios, and due to their lack of representation error, invertible priors can yield better reconstructions than GAN priors for images that have rare features of variation within the biased training set, including out-of-distribution natural images. We additionally compare performance for compressive sensing to unlearned methods, such as the deep decoder, and we establish theoretical bounds on expected recovery error in the case of a linear invertible model.

Equal contributions are denoted by [*] and [†]. [1]Department of Electrical Engineering, Information Technology University, Lahore, Pakistan [2]Department of Mathematics and Khoury College of Computer Sciences, Northeastern University, Boston, MA [3]Department of Computational and Applied Mathematics, Rice University, Houston, TX. Correspondence to: Max Daniels <daniels.g@northeastern.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

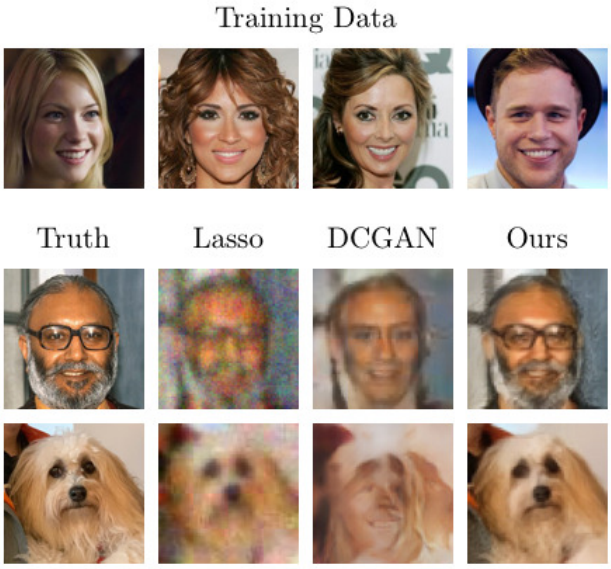

Figure 1. We train an invertible generative model with CelebA images (including those shown). When used as a prior for compressive sensing, it can yield higher quality image reconstructions than Lasso and a trained DCGAN, even on out-of-distribution images. Note that the DCGAN reflects biases of the training set by removing the man's glasses and beard, whereas our invertible prior does not.

Generative deep neural networks have shown remarkable performance as natural signal priors in imaging inverse problems, such as denoising, inpainting, compressive sensing, blind deconvolution, and phase retrieval. These generative models can be trained from datasets consisting of images of particular natural signal classes, such as faces, fingerprints, MRIs, and more (Karras et al., 2018; Minaee & Abdolrashidi, 2018; Shin et al., 2018; Chen et al., 2018). Some such models, including variational autoencoders (VAEs) and generative adversarial networks (GANs), learn an explicit low-dimensional manifold that approximates a natural signal class (Goodfellow et al., 2014; Kingma & Welling, 2014; Rezende et al., 2014). We will refer to such models as GAN priors. These priors can be used for inverse problems by attempting to find the signal in the range of

the generative model that is most consistent with provided measurements. When the GAN has a low dimensional latent space, this allows for a low dimensional optimization problem that operates directly on the natural signal class. Consequently, generative priors can obtain significant performance improvements over classical methods. For example, GAN priors have been shown to outperform sparsity priors at compressive sensing with 5-10x fewer measurements in some cases. Additionally, GAN priors have led to novel theory for signal recovery in the linear compressive sensing and nonlinear phase retrieval problems (Bora et al., 2017; Hand & Voroninski, 2018; Hand et al., 2018), and they have also shown promising results for the nonlinear blind image deblurring problem (Asim et al., 2019).

A significant drawback of GAN priors for solving inverse problems is that they can have large representation error or bias due to architecture and training. That is, a desired image may not be in or near the range of a particular trained GAN. Representation error can occur both for in-distribution and out-of-distribution images. For in-distribution, it can be caused by inappropriate latent dimensionality and mode collapse. For out-of-distribution images, representation error can be large in part because the GAN training process explicitly discourages such images due to the presence of the concurrently trained discriminator network. For many imaging inverse problems, it is important to be able to recover signals that are out-of-distribution relative to training data. For example, in scientific and medical imaging, novel objects or pathologies may be expressly sought. Additionally, desired signals may be out-of-distribution because a training dataset has bias and is unrepresentative of the true underlying distribution. As an example, the CelebA dataset (Liu et al., 2015) is biased toward people who are young, who do not have facial hair or glasses, and who have a light skin tone. As we will see, a GAN prior trained on this dataset learns these biases and exhibits image recovery failures because of them.

Several recent priors have been developed that have lower representation error than GANs. One class of approaches are unlearned neural network priors, such as the Deep Image Prior and the Deep Decoder (Ulyanov et al., 2018; Heckel & Hand, 2019). These are neural networks that are randomly initialized, and whose weights are optimized at inversion time to best fit provided measurements. They have practically zero representation error for natural images. Because they are untrained, there is no training set or training distribution, and hence there is no notion of in- or out-of-distribution natural images. Another class of approaches include updating the weights of the trained GAN at inversion time in an image adaptive way, such as the IAGAN (Hussein et al., 2020). Such an approach could be interpreted as using the GAN as a warm start for a Deep Image Prior. A further approach is Latent Convolutional Models (Athar et al., 2019), in which a generative prior is trained using high dimensional latent representations which are structured as the parameters of a randomly initialized convolutional neural network.

In this paper, we study flow-based invertible neural networks as signal priors. These networks are mathematically invertible (one-to-one and onto) by architectural design (Dinh et al., 2017; Gomez et al., 2017; Jacobsen et al., 2018; Kingma & Dhariwal, 2018). Consequently, they have zero representation error and are capable of recovering any image, including those significantly out-of-distribution relative to a training set; see Figure 1. We call the domain of an invertible generator the latent space, and we call the range of the generator the signal space. These must have equal dimensionality. The strengths of these invertible models include: their architecture allows exact and efficient latent-variable inference, direct log-likelihood evaluation, and efficient image synthesis; they have the potential for significant memory savings in gradient computations; and they can be trained by directly optimizing the likelihood of training images. This paper emphasizes an additional strength: *because they lack representation error, invertible models can mitigate dataset bias and improve recovery performance on inverse problems, including for signals that are out-of-distribution relative to training data.*

We present a method for using pretrained generative invertible neural networks as priors for imaging inverse problems. An invertible generator, once trained, can be used for a wide variety of inverse problems, with no specific knowledge of those problems used during the training process. As an invertible net permits a likelihood estimate for all images, image recovery can be posed as seeking the highest likelihood image that is consistent with provided measurements.

As a proxy for the image log-likelihood, we pose an optimization of squared data-fit over the latent space under regularization by likelihood of latent representations. In the case of denoising, we explicitly penalize log-likelihood of latent codes, while in compressive sensing and inpainting, regularization is achieved algorithmically. This is due in part to initializion with latent code zero.

The contributions of this paper are as follows. We train a generative invertible model using the CelebA dataset. With this fixed model as a signal prior, we study its performance at multiple inverse problems.

- For image denoising, invertible neural network priors can yield sharper images with higher PSNRs than BM3D (Dabov et al., 2007).
- For compressive sensing of in-distribution images, invertible neural network priors can yield higher PSNRs than GANs with low-dimensional latent dimensionality (both DCGAN and PG-GAN), Image Adaptive

GANs, sparsity priors, and a Deep Decoder across a wide range of undersampling ratios.

- Invertible neural networks exhibit graceful performance decay for compressive sensing on out-of-distribution images. They can yield significantly higher PSNRs than GANs with low latent dimensionality across a wide range of undersampling ratios. They can additionally yield higher or comparable PSNRs than the Deep Decoder when there are sufficiently many measurements.
- We introduce a likelihood-based theoretical analysis of compressive sensing under invertible generative priors in the case that the generator is linear. Given m linear measurements of an n-dimensional signal, we prove a theorem establishing upper and lower bounds of expected squared recovery error in terms of the sum of the squares of the smallest $n - m$ singular values of the model.

2. Method

We assume that we have access to a pretrained generative Invertible Neural Network (INN), $G : \mathbb{R}^n \to \mathbb{R}^n$. We write $x = G(z)$ and $z = G^{-1}(x)$, where $x \in \mathbb{R}^n$ is an image that corresponds to the latent representation $z \in \mathbb{R}^n$. We will consider a G that has the Glow architecture introduced in (Kingma & Dhariwal, 2018). For a short introduction to the Glow model, see section 2.1.

We consider recovering an image x from possibly-noisy linear measurements of the form,

$$y = Ax + \eta,$$

where $A \in \mathbb{R}^{m \times n}$ is a measurement matrix and $\eta \in \mathbb{R}^m$ models noise. Given a pretrained invertible generator G, we have access to likelihood estimates for all images $x \in \mathbb{R}^n$. Hence, it is natural to attempt to solve the above inverse problem by a maximum likelihood formulation given by

$$\min_{x \in \mathbb{R}^n} \|Ax - y\|^2 - \gamma \log p_G(x),$$

where p_G is the likelihood function over x induced by G, $\|\cdot\|$ is the Euclidean norm, and γ is a hyperparameter. We have found this formulation to be empirically challenging to optimize, and instead we solve an optimization problem over latent space that encourages high likelihood latent representations. In the case of denoising, we solve

$$\min_{z \in \mathbb{R}^n} \|AG(z) - y\|^2 + \gamma \|z\|^2 \tag{1}$$

In the case of compressive sensing, we fix $\gamma = 0$ and solve

$$\min_{z \in \mathbb{R}^n} \|AG(z) - y\|^2 \tag{2}$$

Unless otherwise stated, we initialize both formulations at $z_0 = 0$.

The motivation for formulations (1) and (2) is as follows. As a proxy for the likelihood of an image $x \in \mathbb{R}^n$, we will use the likelihood of its latent representation $z = G^{-1}(x)$. Because the invertible network G was trained to map a standard normal in $\mathbb{R}^n$ to a distribution over images, the log-likelihood of a latent representation z is proportional to $\|z\|^2$. The model induces a probability distribution over the affine space of images consistent with some given measurements, and so our proxy turns the likelihood maximization task over an affine space in x into the geometric task of finding the point on a manifold in z-space that is closest to the origin with respect to the Euclidean norm. In order to approximate that point, we run a gradient descent in z down the data misfit term starting at $z_0 = 0$.

In principle, this proxy is imperfect in that some high likelihood latent codes may correspond to low likelihood images. We find that the set of such images has low total probability and they are inconsistent with enough provided measurements. For further discussion of our choice of proxy and initialization at $z_0 = 0$, see the Supplemental Materials.

In all experiments that follow, we use an invertible Glow model, as in (Kingma & Dhariwal, 2018). Due to computational considerations, we run most of our experiments on 64×64px color images with the pixel values scaled between $[0, 1]$. For some compressive sensing experiments, we additionally trained a 128×128px Glow model in order to replicate results at this larger size. Once trained, the Glow prior is fixed for use in each of the inverse problems below.

For comparison to GAN architectures, we train a 64×64 DCGAN architecture (Radford et al., 2016) and a 128×128 PGGAN architecture (Karras et al., 2018). To use these priors in inverse problems, we use the formulation from (Bora et al., 2017), which is the formulation above in the case where the optimization is performed over $\mathbb{R}^k$, $\gamma = 0$, and initialization is selected randomly. For comparison to an unlearned neural image prior, we implement an overparameterized variant of a Deep Decoder prior at both resolutions as in (Heckel & Hand, 2019). In all experiments, we solve (2) using L-BFGS for Glow, and Adam (Kingma & Ba, 2015) for the Deep Decoder, DCGAN, and PGGAN architectures. DCGAN and PGGAN results are reported for an average of 3 runs because we observed some variance due to random initialization. To measure the quality of recovered images, we use Peak Signal-to-Noise Ratio (PSNR). For more information on the training algorithms, hyperparameters, and parameter counts for each of the tested models, see the Supplemental Materials.

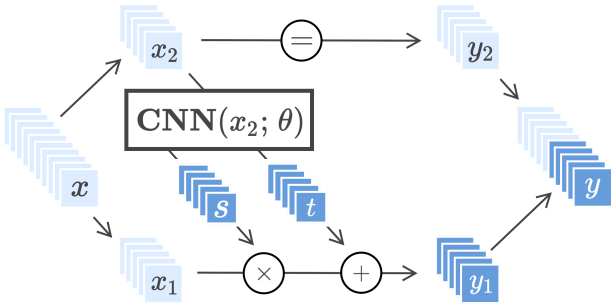

Figure 2. An Affine Coupling layer applies an affine transformation to half of the input data, here x_1. The parameters of the affine transformations, s and t, can depend in a complex, learned way on the other half of the input data. The model can be inverted, even though s and t themselves are not invertible.

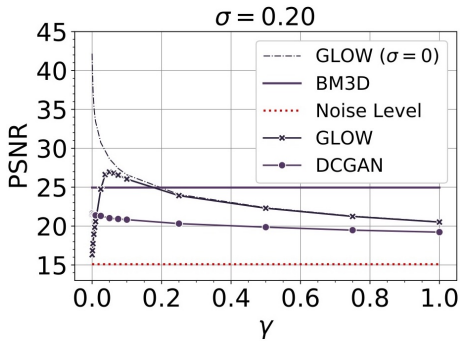

Figure 3. Recovered PSNR values as a function of γ for denoising by the Glow and DCGAN priors. Denoising results are averaged over $N = 50$ in-distribution test set images. For reference, we show the average PSNRs of the original noisy images, after applying BM3D, and under the Glow prior in the noiseless case ($\sigma = 0$).

2.1. Details of the Glow Architecture

The Glow architecture (Kingma & Dhariwal, 2018) belongs to the class of *normalizing flow models*. A normalizing flow models output signals using a composition of many flow steps which are each individually invertible. In the Glow model, flow steps use an Affine Coupling layer, in which half of the input data is used to determine the scale and translation parameters of an affine transformation applied to the other half of the input data. This operation is shown schematically in Figure 2. Each affine transformation is invertible and has an upper-triangular Jacobian, making it computationally tractable to compute the Jacobian determinant of the entire normalizing flow. In turn, given a simple prior over the latent space, the model can be efficiently trained to sample structured, high-dimensional data by directly maximizing likelihood of the training data.

To ensure that each input component can affect each output component, the Glow models incorporate a pixelwise reshuffling. In other normalizing flows, such as RealNVP (Dinh et al., 2017), this is achieved by a fixed permutation, whereas in Glow it is achieved by a learned 1x1 convolution. Both models consist of multiscale achitectures based on affine coupling and pixelwise reshuffling. We refer the reader to (Dinh et al., 2017) and (Kingma & Dhariwal, 2018) for more details.

3. Applications

3.1. Denoising

We consider the denoising problem with $A = I_n$ and $\eta \sim \mathcal{N}(0, \sigma^2 I_n)$, as given by formulation equation 1. We evaluate the performance of a Glow prior, a DCGAN prior, and BM3D for two different noise levels on 64px in-distribution images ($n = 64 \times 64 \times 3 = 12288$).

Figure 3 shows the recovered PSNR values as a function of γ for denoising by the Glow and DCGAN priors, along with the PSNR by BM3D. The figure shows that the performance of the regularized Glow prior increases with γ, and then decreases. If γ is too low, then the network fits to the noise in the image. If γ is too high, then data fit is not enforced strongly enough. We study this effect for an extensive range of γ and noise levels, which may be found in the Supplemental Materials. We see in Figure 3 that an appropriately regularized Glow prior can outperform BM3D by almost 2 dB. The experiments also reveal that appropriately regularized Glow priors outperform the DCGAN prior, which suffers from representation error and is not aided by the regularization. A visual comparison of the recoveries at the noise level $\sigma = 0.1$ using Glow, DCGAN priors, and BM3D can be seen in Figure 4. Note that the recoveries with Glow are sharper than BM3D. See the Supplemental Materials for more quantitative and qualitative results.

3.2. Compressive Sensing

In compressive sensing, one is given undersampled linear measurements of an image, and the goal is to recover the image from those measurements. In our notation, $A \in \mathbb{R}^{m \times n}$ with $m < n$. As the image x is undersampled, there is an affine space of images consistent with the measurements, and an algorithm must select which is most 'natural.' A common proxy for naturalness in the literature has been sparsity with respect to the DCT or wavelet bases. With a GAN prior, an image is considered natural if it lies in or near the range of the GAN. For an invertible prior, image likelihood is a proxy for naturalness, and under our proxy for likelihood, we consider an image to be natural if it has a latent representation of small norm.

We study compressive sensing in the case that A is an $m \times n$ matrix of i.i.d. $\mathcal{N}(0, 1/m)$ entries and where η is standard

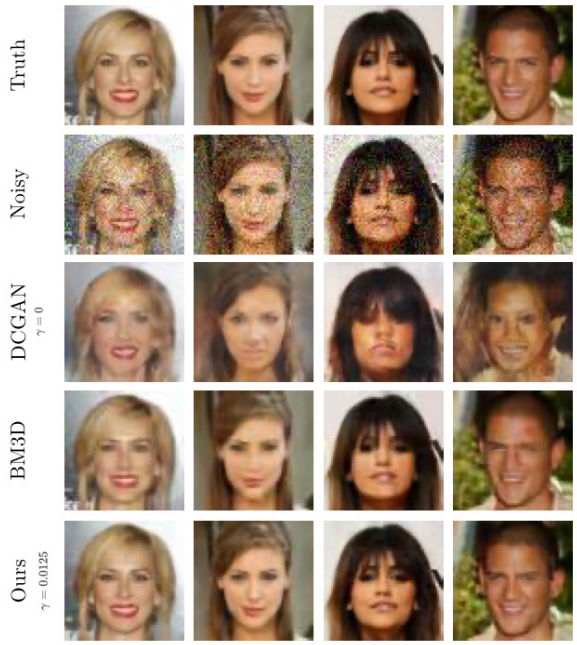

Figure 4. Denoising results using the Glow prior, the DCGAN prior, and BM3D at noise level $\sigma = 0.1$. Note that the Glow prior gives a sharper image than BM3D in these cases.

i.i.d. Gaussian noise normalized such that $\sqrt{\mathbb{E}\|\eta\|^2} = 0.1$. We present our results under formulation (2), the $\gamma = 0$ simplification of (1).

We compare the Glow prior to various other learned and unlearned image priors. This includes GANs, as used in (Bora et al., 2017), and IAGAN, as used in (Hussein et al., 2020). In the 64px case ($n = 64 \times 64 \times 3 = 12288$), we compare to a DCGAN, and in the 128px case ($n = 128 \times 128 \times 3 = 49152$) we compare to a PGGAN, both trained on the CelebA-HQ dataset. We also compare to unlearned image priors, including an overparameterized Deep Decoder and a sparsity prior in the DCT basis[1]. To assess the performance of each image prior, we report the mean PSNR of recovered test set images from both the training distribution of the learned models ("in-distribution" images) and other datasets ("out-of-distribution" images). Our in-distribution images are sampled from a test set of CelebA-HQ images which were withheld from all learned models during their training procedures. Out-of-distribution images are sampled randomly from the Flickr Faces High Quality dataset, which provides images with features of variation that are rare among CelebA images (eg. skin tone, age, beards, and glasses) (Karras et al., 2019). The 64px and 128px recovery experiments have test sets with $N = 1000$ and $N = 100$ images respectively.

[1]The inverse problems with Lasso were solved by $\min_z \|A\Phi z - y\|_2^2 + 0.01\|z\|_1$ using coordinate descent. We observe similar performance between the DCT basis and a Wavelet basis.

Surprisingly, the Glow prior exhibits superior performance on compressive sensing tasks with no likelihood penalization in the objective (2). We find that $z_0 = 0$ is a particularly good choice of initialization, for which one does not benefit from likelihood penalization, while for other choices of initialization direct penalization of likelihood may improve performance. We provide additional experiments exploring this phenomena in the Supplemental Materials.

As shown in Figure 5, we find that on its training distribution, the Glow prior outperforms both the learned and unlearned alternatives for a wide range of undersampling ratios. Surprisingly, in the case of extreme undersampling, Glow substantially outperforms these methods even though it does not maintain a direct low-dimensional parameterization of the signal manifold. In both the 64px and 128px cases, the GAN architectures quickly saturate due to their representation error. For out-of-distribution images, the Glow prior exhibits graceful performance decay, and is still highly performant in a large measurement regime. See figures 6 and 7 for a visual comparison of recovered images from the CelebA and FFHQ test sets for the 128px case. The PGGAN's performance reveals biases of the underlying dataset and limitations of low-dimensional modeling, as the PGGAN fails completely to represent features like darker skin tones or accessories, which are uncommon in CelebA. In contrast, the Glow prior mitigates this bias, demonstrating image recovery for natural images that are not representative of the CelebA training set.

3.3. Inpainting

In inpainting, one is given a masked image of the form $y = M \quad x$, where M is a masking matrix with binary entries and $x \in \mathbb{R}^n$ is an n-pixel image. As in compressive sensing, we solve (2) to try to recover x, among the affine space of images consistent with the measurements. As before, using the minimizer $\hat{z}$, the estimated image is given by $G(\hat{z})$. We show qualitative inpainting results in Fig. 8. Our experiments reveal the same story as forcompressive sensing. If initialized at $z_0 = 0$, then Glow model under the empirical risk formulation with $\gamma = 0$ exhibits high PSNRs on test images while the DCGAN is limited by its representation error.

4. Theory

We now introduce a likelihood-based theory for compressive sensing under invertible priors in the case of a linear invertible model. We will provide an estimate on the expected error of the recovered signal in terms of the singular values of the model. Specifically, we consider a fixed invertible linear $G : \mathbb{R}^n \to \mathbb{R}^n$. Assume signals are generated by a distribution $p_G(x)$, given by $x = Gz$, where $z \sim \mathcal{N}(0, I_n)$. Equivalently, $p_G = \mathcal{N}(0, GG^{\mathrm{T}})$.

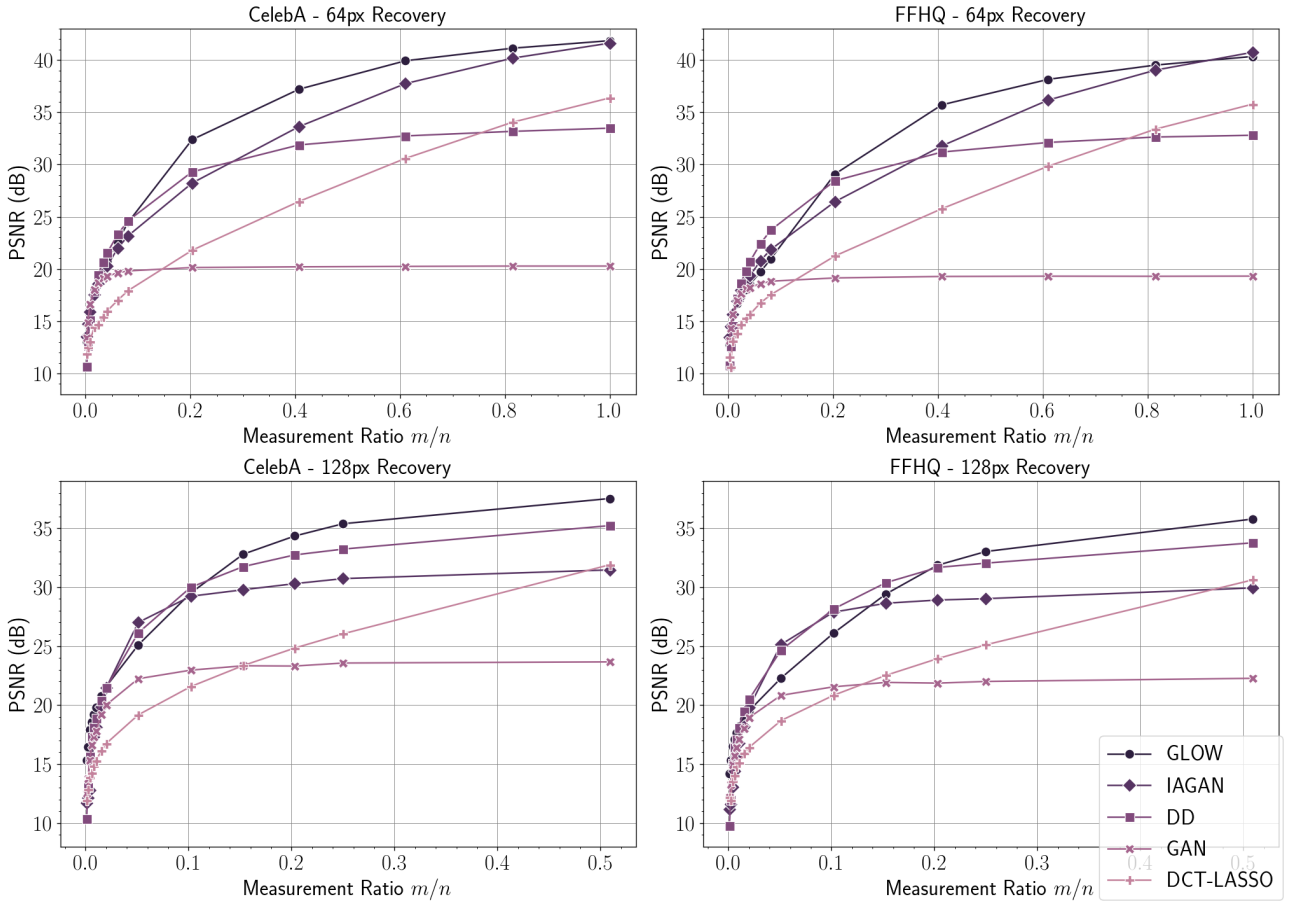

Figure 5. Performance of Glow, GAN, and IAGAN priors (learned) and the Deep Decoder and Lasso-DCT priors (unlearned) across various undersampling ratios in the 64px and 128px case. The 64px and 128px experiments use $N = 1000$ and $N = 100$ test set images respectively.

For an unknown sample $x_0 \sim p_G$, suppose we are provided noiseless measurements Ax_0, where $A \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries. We consider the maximum likelihood estimate of x_0:

$$\hat{x} := \arg\max_{x\in\mathbb{R}^n} p_G(x) \text{ s.t. } Ax = Ax_0. \tag{3}$$

The following theorem provides both upper and lower bounds on the absolute expected squared error in terms of the singular values of G.

Theorem 1. *Suppose $x_0 \sim p_G$ where $p_G = \mathcal{N}(0, GG^{\mathrm{T}})$ and $G \in \mathbb{R}^{n\times n}$ has singular values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n > 0$. Let $A \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0,1)$ entries where $4 \leqslant m < n$. Then the maximum likelihood estimator $\hat{x}$ obeys*

$$\sum_{i>m} \sigma_i^2 \leqslant \mathbb{E}_A \mathbb{E}_{x_0\sim p_G} \|\hat{x} - x_0\|^2 \leqslant m \sum_{i>m-2} \sigma_i^2. \tag{4}$$

The relative expected squared error could be computed by $\frac{\mathbb{E}_A \mathbb{E}_{x_0\sim p_G} \|\hat{x}-x_0\|^2}{\mathbb{E}_{x_0\sim p_G}\|x_0\|^2}$, noting that $\mathbb{E}_{x_0\sim p_G}\|x_0\|^2 = \sum_{k\geq 1}\sigma_k^2$.

The lower bound of this theorem establishes that under a linear invertible generative model, m Gaussian measurements give rise to at least as much error as would be given by the best m-dimensional signal model, i.e. the model corresponding to the only the top m directions of highest variance. The upper bound on this theorem establishes that up to a factor of m, expected square error is bounded above by the error given by the best $m-2$ dimensional model.

Note that if the singular values decay quickly enough, then the expected recovery error under the linear invertible model decreases to a small value as $m \to n$. Specifically, this is achieved if $\sigma_i = o(i^{-1/2})$. This conclusion is in contrast with the theory for GANs with low latent dimensionality. In that literature, recovery error does not decrease to 0 as $m \to n$; instead it saturates at the representation error of the GAN model. In the present case of a linear model, the best k-dimensional model would have expected square recovery error at least $\sum_{i>k}\sigma_i^2$ regardless of m. Similar work in (Yu & Sapiro, 2011) also showed that m Gaussian measurements of a Gaussian signal give rise to an error proportional to

Figure 6. Compressive sensing on CelebA images with $m = 7,500$ ($\approx 20\%$) of measurements. Visual comparisons: CS under the Glow prior, PGGAN prior, and the overparameterized Deep Decoder prior. For images in-distribution, we observe qualitatively sharper recoveries from the Glow Prior than from the Deep Decoder.

Figure 7. Compressive sensing on FFHQ images with $m = 7,500$ ($\approx 20\%$) of measurements. Visual comparisons: CS under the Glow prior, PGGAN prior, and the overparameterized Deep Decoder prior. For images out-of-distribution, the images recovered by the Deep Decoder and the Glow priors are both qualitatively and quantitatively (by PSNR) comparable.

the best m-term approximation. The difference between that work and our results is that we have an explicit upfront constant in the upper bound whereas in (Yu & Sapiro, 2011) it is estimated via Monte Carlo simulations.

The theorem is proved in the Supplemental Materials.

5. Discussion

We have demonstrated that pretrained generative invertible models can be used as natural signal priors in imaging inverse problems. Their strength is that *every desired image is in the range of an invertible model*, and the challenge that they overcome is that *every undesired image is also in the range of the model and no explicit low-dimensional representation is kept*. We demonstrate that this formulation can quantitatively and qualitatively outperform BM3D at denoising. For compressive sensing on in-distribution images, invertible priors can have lower recovery errors than Deep Decoder, GANs with low dimensional latent representations, and Lasso, across a wide range of undersampling ratios. We show that the performance of our invertible prior behaves gracefully with slight performance drops for out-of-distribution images. We additionally prove a theoretical upper and lower bound for expected squared recover error in the case of a linear invertible generative model.

The idea of analyzing inverse problems with invertible neural networks has appeared in Ardizzone et al. (2019). The authors study estimation of the complete posterior parameter distribution under a forward process, conditioned on observed measurements. Specifically, the authors approximate a particular forward process by training an invertible neural network. The inverse map is then directly available. In order to cope with information loss, the authors augment the measurements with additional variables. This work differs from ours because it involves training a separate model for every particular inverse problem.

In contrast, our work studies how to use a pretrained invertible generator for a variety of inverse problems not known at training time. Training invertible networks is challenging and computationally expensive; hence, it is desirable to separate the training of off-the-shelf invertible models from potential applications in a variety of scientific domains. In additional work by Putzky & Welling (2019), the authors also exploit the efficient gradient calculations of invertible nets for improved MR reconstruction.

Why do invertible neural networks perform well for both in-distribution and out-of-distribution images?
One reason that the invertible prior performs so well is because it has no representation error. Any image is potentially recoverable, even if the image is significantly outside of the training distribution. In contrast, methods based on projecting onto an explicit low-dimensional representation of a natural signal manifold (such as GAN priors) will have

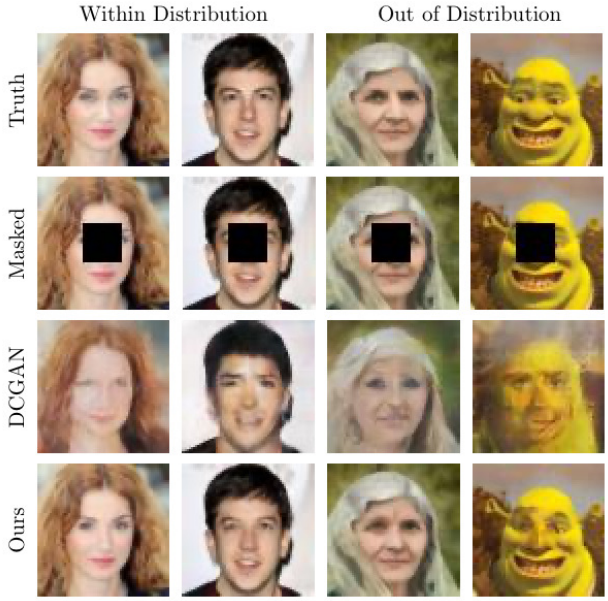

Figure 8. Inpainting on random images with the Glow and DCGAN priors. In the left columns, images are taken from the training distribution of the learned priors, and the right column includes two images from the wild.

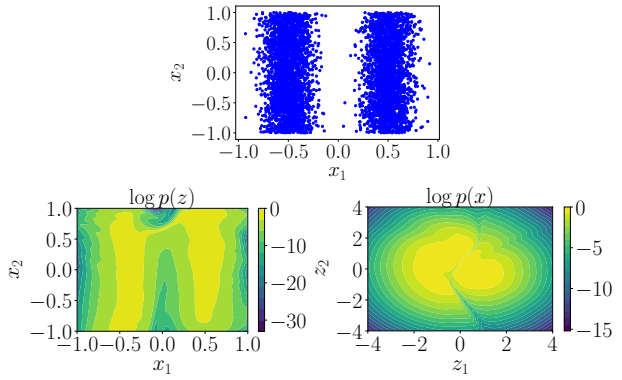

Figure 9. An invertible net was trained on the data points in x-space (top), resulting in the given plots of latent z-likelihood versus x (bottom left), and x-likelihood versus latent representation z (bottom right).

representation error, perhaps due to modeling assumptions, mode collapse, or bias in a training set. Such methods will see performance prematurely saturate as the number of measurements increases. In contrast, an invertible prior would not see performance saturate. In the extreme case of having a full set of exact measurements, an invertible prior could in principle recover any image exactly.

How do invertible priors respect the low dimensionality of natural signals? And how does our theory inform this?
A surprising feature of invertible priors is that they perform well even though they do not maintain explicit low-dimensional representations of natural signals. Instead they have a fully dimensional representation of the natural signal class. Naturally, any signal class will truly be fully dimensional, for example due to sensor noise, but with different importances to different directions. Those directions in latent space corresponding to noise perturbations will have much smaller of an effect than corresponding perturbations in semantically meaningful directions. As an illustration, we observe with trained Glow models that the singular values of the Jacobian of G at a natural image exhibit significant decay, as we show in the Supplemental Materials. In principle, signal models of any given dimensionality could be extracted from G, though it is not obvious how to compute these. *The power of invertible priors is that each additional measurement acts to roughly increment the dimensionality of the modeled natural signal manifold. That is, more measurements permit exploiting a higher dimensional and, hence, lower-error signal model. In contrast, GAN-based recovery theory exploits a model of fixed dimensionality regardless of the number of measurements*. Our theoretical analysis provides justification for this explanation based on linear invertible generators G. Each singular value σ_i of G quantifies the signal variation due to a particular direction in latent space, and the expected squared error given m random measurements is upper bounded by $m \sum_{i>m-2} \sigma_i^2$. Note that the best m-dimensional manifold (given by the top m singular values and vectors) would yield at best an expected squared error of $\sum_{i>m} \sigma_i^2$. The multiplicative m term and the sum's starting index may not be optimal, and improvement of this bound is left for future research.

Why is the likelihood of an image's latent representation a reasonable proxy for the image's likelihood?
The training process for an invertible generative model attempts to learn a target distribution in image space by directly maximizing the likelihood of provided samples from that distribution, given a standard Gaussian prior in latent space. High probability regions in latent space map to regions in image space of equal probability. Hence, broadly speaking, regions of small values of $\|z\|$ are expected to map to regions of large likelihoods in image space. There will be exceptions to this property. For example, natural image distributions have a multimodal character. The preimage of high probability modes in image space will correspond to high likelihood regions in latent space. Because the generator G is invertible and continuous, interpolation in latent space of these modes will provide images of high likelihood in z but low likelihood in the target distribution. To demonstrate this, we trained a Real-NVP (Dinh et al., 2017) invertible neural network on the two dimensional set of points depicted in Figure 9 (top panel). The lower left and right panels show that high likelihood regions in latent space

generally correspond to higher likelihood regions in image space, but that there are some regions of high likelihood in latent space that map to points of low likelihood in image space and in the target distribution. We see that the spurious regions are of low total probability and would be unlikely to be the desired outcomes of an inverse problem arising from the target distribution.

How can solving compressive inverse problems be successful without direct penalization of the image likelihood or its proxy?
If there are fewer linear measurements than the dimensionality of the desired signal, an affine space of images is consistent with the measurements. In our formulation, regularization does not occur by direct penalization by our proxy for image likelihood; instead, it occurs implicitly by performing the optimization in z-space with an initialization of $z_0 = 0$. The set of latent representations z that are consistent with the compressive measurements define a m-dimensional nonlinear manifold. As per the likelihood proxy mentioned above, the spirit of our formulation is to find the point on this manifold that is closest to the origin with respect to the Euclidean norm. Our specific way of estimating this point is to perform a gradient descent down a data misfit term in z-space, starting at the origin. While a gradient flow typically will not find the closest point on the manifold, it empirically finds a reasonable approximation of that point. This type of algorithmic regularization is akin to the linear invertible model setting where it is well known that running gradient descent to optimize an underdetermined least squares problem with initialization $z_0 = 0$ will converge to the minimum norm solution.

The results of this paper provide further evidence that reducing representational error of generators can significantly enhance the performance of generative models for inverse problems in imaging. This idea was also recently explored in (Athar et al., 2019), where the authors trained a GAN-like prior with a high-dimensional latent space. The high dimensionality of this space lowers representational error, though it is not zero. In their work, the high-dimensional latent space had a structure that was difficult to directly optimize, so the authors successfully modeled latent representations as the output of an untrained convolutional neural network whose parameters are estimated at test time.

Their paper and ours raises an important question: which generator architectures provide a good balance between low representation error, ease of training, and ease of inversion? This question is important, as solving (2) in our 128×128px color images experiments took on the order of 11 minutes using an NVIDIA 2080 Ti GPU.

New developments are needed on architectures and frameworks in between low-dimensional generative priors and fully invertible generative priors. Such methods could leverage the strengths of invertible models while being much cheaper to train and use.

Acknowledgements

PH is partially supported by NSF CAREER Grant DMS-1848087. OL acknowledges support by the NSF Graduate Research Fellowship under Grant No. DGE-1450681.

References

Ardizzone, L., Kruse, J., Rother, C., and Köthe, U. Analyzing inverse problems with invertible neural networks. In *International Conference on Learning Representations*, 2019.

Asim, M., Shamshad, F., and Ahmed, A. Blind image deconvolution using deep generative priors. *arXiv:1802.04073 [cs]*, Feb 2019. arXiv: 1802.04073.

Athar, S., Burnaev, E., and Lempitsky, V. Latent convolutional models. In *International Conference on Learning Representations*, 2019.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 537–546. PMLR, Aug 2017.

Chen, Y., Shi, F., Christodoulou, A. G., Xie, Y., Zhou, Z., and Li, D. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G. (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, volume 11070 of *Lecture Notes in Computer Science*, pp. 91–99. Springer, 2018. doi: 10.1007/978-3-030-00928-1_11.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8): 2080–2095, 2007.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. 2017.

Gomez, A. N., Ren, M., Urtasun, R., and Grosse, R. B. The reversible residual network: Backpropagation without storing activations. In Guyon, I., Luxburg, U. v., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information*

Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 2214–2224, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. *Generative Adversarial Nets*, pp. 2672–2680. Curran Associates, Inc., 2014.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 970–978. PMLR, 2018.

Hand, P., Leong, O., and Voroninski, V. *Phase Retrieval Under a Generative Prior*, pp. 9136–9146. Curran Associates, Inc., 2018.

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

Hussein, S. A., Tirer, T., and Giryes, R. Image-adaptive gan based reconstruction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 3121–3129. AAAI Press, 2020.

Jacobsen, J.-H., Smeulders, A. W. M., and Oyallon, E. i-revnet: Deep invertible networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 10236–10245, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425.

Minaee, S. and Abdolrashidi, A. Finger-gan: Generating realistic fingerprint images using connectivity imposed gan. *CoRR*, abs/1812.10482, 2018.

Putzky, P. and Welling, M. Invert to learn to invert. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 444–454, 2019.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 1278–1286. JMLR.org, 2014.

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In Gooya, A., Goksel, O., Oguz, I., and Burgos, N. (eds.), *Simulation and Synthesis in Medical*

Imaging - Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, volume 11037 of *Lecture Notes in Computer Science*, pp. 1–11. Springer, 2018. doi: 10.1007/978-3-030-00536-8_1.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. Deep image prior. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 9446–9454. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00984.

Yu, G. and Sapiro, G. Statistical compressed sensing of gaussian mixture models. *IEEE Trans. Signal Process.*, 59 (12):5842–5858, 2011. doi: 10.1109/TSP.2011.2168521.