Bayesian Image Reconstruction using Deep Generative Models

Razvan V. Marinescu  
MIT CSAIL  
razvan@csail.mit.edu

Daniel Moyer  
MIT CSAIL  
dmoyer@csail.mit.edu

Polina Golland  
MIT CSAIL  
polina@csail.mit.edu

Abstract

Machine learning models are commonly trained end-to-end and in a supervised setting, using paired (input, output) data. Examples include recent super-resolution methods that train on pairs of (low-resolution, high-resolution) images. However, these end-to-end approaches require re-training every time there is a distribution shift in the inputs (e.g., night images vs daylight) or relevant latent variables (e.g., camera blur or hand motion). In this work, we leverage state-of-the-art (SOTA) generative models (here StyleGAN2) for building powerful image priors, which enable application of Bayes’ theorem for many downstream reconstruction tasks. Our method, Bayesian Reconstruction through Generative Models (BRGM), uses a single pre-trained generator model to solve different image restoration tasks, i.e., super-resolution and in-painting, by combining it with different forward corruption models. We keep the weights of the generator model fixed, and reconstruct the image by estimating the Bayesian maximum a-posteriori (MAP) estimate over the input latent vector that generated the reconstructed image. We further use Variational Inference to approximate the posterior distribution over the latent vectors, from which we sample multiple solutions. We demonstrate BRGM on three large and diverse datasets: (i) 60,000 images from the Flick Faces High Quality dataset [1] (ii) 240,000 chest X-rays from MIMIC III [2] and (iii) a combined collection of 5 brain MRI datasets with 7,329 scans [3]. Across all three datasets and without any dataset-specific hyperparameter tuning, our simple approach yields performance competitive with current task-specific state-of-the-art methods on super-resolution and in-painting, while being more generalisable and without requiring any training. Our source code and pre-trained models are available online: https://razvanmarinescu.github.io/brgm/

1 Introduction

While end-to-end supervised learning is currently the most popular paradigm in the research community, it suffers from several problems. First, distribution shifts in the inputs often require re-training, as well as the effort of collecting an updated dataset. In some settings, such shifts can occur often (hospital scanners are upgraded) and even continuously (population is slowly aging due to improved healthcare). Secondly, current state-of-the-art machine learning (ML) models often require prohibitive computational resources, which are only available in a select number of companies and research
High Res. with adversarial losses, it is not clear which solution image is retrieved, and how to sample multiple solutions from the posterior distribution. This can be repeated for other corruption processes ($f_2$, $f_3$) such as masking, motion, to-grayscale, as well as for other parametrisations of the process (e.g., super-resolution with different kernels or factors).

Deep generative models have recently obtained state-of-the-art results in simulating high-quality images from a variety of computer vision datasets. Generative Adversarial Networks (GANs) such as StyleGAN2 [4] and StyleGAN-ADA [5] have been demonstrated for unconditional image generation, while BigGAN has shown impressive performance in class-conditional image generation [6]. Similarly, Variational Autoencoder-based methods such as VQ-VAE [7] and β-VAE [8] have also been competitive in several image generation tasks. Other lines of research in deep generative models are auto-regressive models such as PixelCNN [9] and PixelRNN [10], as well as invertible flow models such as NeuralODE [11], Glow [12] and RealNVP [13]. While these models generate high-quality images that are similar to the training distribution, they are not directly applicable for solving more complex tasks such as image reconstruction.

A particularly important application domain for generative models is inverse problems, which aim to reconstruct an image that has undergone a corruption process such as blurring. Previous work has focused on regularizing the inversion process using smoothness [14] or sparsity [15–17] priors. However, such priors often result in blurry images, and do not enable hallucination of features, which is essential for such ill-posed problems. More recent deep-learning approaches [18–21] address this challenge using training data made of pairs of (low-resolution, high-resolution) images. However, one fundamental limitation is that they compute the pixelwise or perceptual loss in the high-resolution/in-painted space, which leads to the so-called averaging effect [22]: since multiple high-resolution images map to the same low-resolution image $I$, the loss function minimizes the average of all such solutions, resulting in a blurry image. Some methods [22, 23] address this through adversarial losses, which force the model to output a solution that lies on the image manifold. However, even with adversarial losses, it is not clear which solution image is retrieved, and how to sample multiple solutions from the posterior distribution.

To overcome the averaging of all possible solutions to ill-posed problems, one can build methods that estimate by design all potential solutions, or a distribution of solutions, for which a Bayesian framework is a natural choice. Bayesian solutions for image reconstruction problems include: Markov Random Field (MRF) priors for denoising and in-painting [24], generative models of photosensor responses from surfaces and illuminants [25], MRF models that leverage global statistics for in-painting [26], Bayesian quantification of the distribution of scene parameters and light direction for inference of shape, surface properties and motion [27], and sparse derivative priors for super-resolution and image demosaicing [28]. The seminal work of [29] formulated the image reconstruction task as the Bayesian optimization over an energy function in a lattice-like physical system.
Starting with [30], several works proposed solving inverse problems using deep generative priors [31–34], which constrain the solution to belong to the learned image manifold of a deep generative model. While all these methods focus on deriving a single point estimate, some recent methods further derive a distribution of potential solutions [36, 37] through Langevin dynamics. However, Langevin dynamics has slow mixing, and additionally, their loss function is only loosely based on a Bayesian MAP estimate. Without a fully-Bayesian model, it is difficult to derive more complex measures of uncertainty, such as α-level confidence intervals. While [35, 38] solve the inverse problem in a Bayesian setting, they only demonstrate it for invertible flow models. It is thus still unclear how to use models other than flows, such as GANs or VAEs, to perform this reconstruction in a Bayesian framework, and how to use GANs to capture a distribution of solutions for a given corrupted image.

In this work, we propose a Bayesian method to perform reconstruction through deep generative priors built using state-of-the-art GAN models. Given a pre-trained generator $G$ (here StyleGAN2), a known corruption model $f$, and a corrupted image $I$ to be restored, we minimize at inference time the Bayesian MAP estimate $w^* = \arg \max_w p(w)p(I \circ G(w))$, where $w$ is latent vector given as input to $G$. We further adapt previous Variational Inference (VI) methods [39, 40] to approximate the posterior $p(w|I)$ with a Gaussian distribution, thus enabling sampling of multiple solutions. Our key theoretical contributions are (i) the formulation of image reconstruction in a principled Bayesian framework using deep generative priors and (ii) the adaptation of previous deep-learning based VI methods to sample multiple reconstructions, while on the application side, we demonstrate our method on three datasets, including two challenging medical datasets, and show its competitive performance against four state-of-the-art methods.

1.1 Related work

Deep Generative Prior (DGP) methods: A related work to ours is PULSE [41], which uses pre-trained StyleGAN models for face super-resolution. Other DGP methods [30–34] are close to our work, especially those that derive a distribution of potential solutions either through (i) direct sampling from conditioned flow models [35], (ii) Langevin dynamics [36, 37] or (iii) Variational Inference (VI) [38], each having several advantages and disadvantages: (i) conditioned flow models generate independent samples from the exact posterior, but use more restricted transformations due to the need to have computable Jacobian determinants, (ii) Langevin dynamics also sample the exact posterior, but generate highly correlated samples and can get stuck in distributional modes while (iii) VI methods can generate a large number of independent samples from an approximate posterior, but require fitting the variational parameters. Our sampling method that uses VI is closest to [38], but we demonstrate it for GAN models such as StyleGAN2 instead of normalizing flows as in [38].

Encoder methods: Other approaches attempt to invert generative models by estimating encoders that map the input images directly into the generator’s latent space [44] in an end-to-end framework. Encoder methods are complimentary to our work, as they can be used to obtain a fast initial estimate of the latent $w$, followed by a slightly longer optimisation process such as ours that can give more accurate reconstructions.

Other inverse problems methods: Approaches similar to ours have also been discussed in inverse problems research. Deep Bayesian Inversion [45] performs image reconstruction using a supervised learning approximation. AmbientGAN [46] builds a GAN model of clean images given noisy observations only, for a specified corruption model. Deep Image Prior (DIP) [47] has shown that the structure of deep convolutional networks captures texture-level statistics that can be used for zero-shot image reconstruction. MimicGAN [48] has shown how to optimise the parameters of the unknown corruption model. Image2StyleGAN [42] finds a projection of a given clean image in the latent space of StyleGAN, while the more updated Image2StyleGAN++ [43] uses the estimated latent projections to demonstrate image manipulations, as well as in-painting. Recent neural radiance fields (NeRF) [49] achieved state-of-the-art results for synthesizing novel views.

2 Method

An overview of our method is shown in Fig. 1. We assume a given generator $G$ can model the distribution of clean images in a given dataset (e.g., human faces), then use a pre-defined forward model $f$ that corrupts the clean image. Given a corrupted input image $I$, we reconstruct it as $G(w^*)$,
where \( w^* \) is the Bayesian MAP estimate over the latent vector \( w \) of \( G \). The graphical model is given in Supp. Fig. 7.

Given an input corrupted image \( I \), we aim to reconstruct the clean image \( I_{CLN} \). In practice, there could be a distribution \( p(I_{CLN}|I) \) of such clean images given a particular input image \( I \), which is estimated using Bayes’ theorem as

\[
p(I_{CLN}|I) \propto p(I_{CLN})p(I|I_{CLN}).
\]

The prior term \( p(I_{CLN}) \) describes the manifold of clean images, restricting the possible reconstructions \( I_{CLN} \) to realistic images. In our context, the likelihood term \( p(I|I_{CLN}) \) describes the corruption process \( f \), which takes a clean image and produces a corrupted image.

### 2.1 The image prior term

The prior model \( p(I_{CLN}) \) has been trained a-priori, before the corruption task is known, hence satisfying the principle of independent mechanisms from causal modelling [50]. In our experiments, \( I_{CLN} = G(w) \), where \( w = [w_1, \ldots, w_{18}] \in \mathbb{R}^{512 \times 18} \) is the latent vector of StyleGAN2 (18 vectors for each resolution level), \( G : \mathbb{R}^{512 \times 18} \rightarrow \mathbb{R}^{n_G \times n_G} \) is the deterministic function given by the StyleGAN2 synthesis network, and \( n_G \times n_G \) is the output resolution of StyleGAN2, in our case 1024x1024 (FFHQ, X-Rays) or 256x256 (brains). Our framework is not specific to StyleGAN2: other generator models such as invertible flows [12, 13] or VAEs [39] can be used, as long as one can flow gradients through the model.

We use the change of variables to express the probability density function over clean images:

\[
p(I_{CLN}) := p(G(w)) = p(w) \left| \frac{\partial G(w)}{\partial w} \right|^{-1}
\]  

(1)

While the traditional change of variables formula assumes that the function \( G \) is invertible, it can be extended to non-invertible\(^1\) mappings\([51, 52]\). In addition, we assume that the Jacobian determinant \( \left| \frac{\partial G(w)}{\partial w} \right| \) is constant for all \( w \), which is a reasonable assumption in StyleGAN2 due to its path length regularization (see Eq. 4 in [4]).

We now seek to instantiate \( p(w) \). Since the latent space of StyleGAN2 consists of many vectors \( w = [w_1, \ldots, w_L] \), where \( L = 18 \) (one for each layer), we need to set meaningful priors for them. While StyleGAN2 assumed that all vectors \( w_i \) are equal, we slightly relax that assumption but set two priors: (i) a cosine similarity prior similar to PULSE [41] that ensures every pair \( w_i \) and \( w_j \) are roughly colinear, and (ii) another prior \( \mathcal{N}(w_i|\mu, \sigma^2) \) that ensures the \( w \) vectors lie in the same region as the vectors used during training. We use the following distribution for \( p(w) \):

\[
p(w) = \prod_i \mathcal{N}(w_i|\mu, \sigma^2) \prod_{i,j} \mathcal{M} \left( \cos^{-1} \frac{w_i w_j^T}{||w_i|| ||w_j||}, 0, \kappa \right)
\]  

(2)

\(^1\)The supp. section of [51] presents an excellent introduction to the generalized change of variable theorem.
where $\cos^{-1} \frac{w_i^T w_j}{||w_i|| ||w_j||}$ is the angle between vectors $w_i$ and $w_j$, and $\mathcal{M}(\cdot, \rho, \kappa)$ is the von Mises distribution with mean zero and scale parameter $\kappa$ which ensures that vectors $w_i$ are aligned. This distribution is analogous to a Gaussian distribution over angles in $[0, 2\pi]$. We compute $\mu$ and $\sigma$ as the mean and standard deviation of 10,000 latent variables passed through the mapping network, like the original StyleGAN2 inversion [4].

### 2.2 The image likelihood term

We instantiate the likelihood term $p(I | IC_{CLN})$ with a potentially probabilistic forward corruption process $f_{IC_{CLN}} : I \rightarrow I_{CLN}$, parameterized by $\psi$. We study two types of corruption processes $f$ as follows:

- Super-resolution: $f_{SR}$ is defined as the forward operator that performs downsampling parameterized by a given kernel $s$. For a high-resolution image $IC_{CLN}$, this produces a low-resolution (corrupted) image $IC_{COR} = (IC_{CLN} \circ k)^s$, where $\circ$ denotes convolution and $^s$ denotes downsampling operator by a factor $s$. The parameters are $\psi = \{k, s\}$

- In-painting with arbitrary mask: $f_{IN}$ is implemented as an operator that performs pixelwise multiplication with a mask $M$. For a given clean image $IC_{CLN}$ and a 2D binary mask $M$, it produces a cropped-out (corrupted) image $IC_{COR} = IC_{CLN} \odot M$, where $\odot$ is the Hadamard product. The parameters of this corruption process are $\psi = \{M\}$ where $M \in \{0, 1\}^{H \times W}$, where $H$ and $W$ are the height and width of the image.

The likelihood model becomes:

$$p(I | IC_{CLN}) = p(I | G(w)) = p(I | f \circ G(w))|J_f (G(w))|^{-1}$$  \hspace{1cm} (3)$$

where $J_f (G(w)) = \frac{\partial f \circ G(w)}{\partial G(w)}$ is the Jacobian matrix of $f$ evaluated at $G(w)$, and is again assumed constant. For the noise model in $p(I | f \circ G(w))$, we consider two types of noise distributions: pixelwise independent Gaussian noise, as well as “perceptual noise”, i.e. independent Gaussian noise in the perceptual VGG embedding space. This yields the following model:

$$p(I | f \circ G(w)) = N(I | f \circ G(w), \sigma_{\text{pixel}}^2 \mathbb{I}_{n_f^2}, \sigma_{\text{percept}}^2 \mathbb{I}_{n_\phi^2})$$  \hspace{1cm} (4)$$

where $\phi : \mathbb{R}^{n_f \times n_f} \rightarrow \mathbb{R}^{n_\phi \times n_\phi}$ is the VGG network, $\sigma_{\text{pixel}}^2 \mathbb{I}_{n_f^2}$ and $\sigma_{\text{percept}}^2 \mathbb{I}_{n_\phi^2}$ are diagonal covariance matrices, $n_f \times n_f$ and $n_\phi \times n_\phi$ are the resolutions of the corrupted images $f \circ G(w)$ as well as perceptual embeddings $\phi \circ f \circ G(w)$. Images $I$, $f(I)$, $f \circ G(w)$ and $\phi \circ f \circ G(w)$ are flattened to 1D vectors, while covariance matrices $\mathbb{I}_{n_f^2}$ and $\mathbb{I}_{n_\phi^2}$ are of dimensions $n_f^2 \times n_f^2$ and $n_\phi^2 \times n_\phi^2$.

### 2.3 Image restoration as Bayesian MAP estimate

The restoration of the optimal clean image $I^*_{CLN}$ given a noisy input image $I$ can be performed through the Bayesian maximum a-posteriori (MAP) estimate:

$$I^*_{CLN} = \arg \max_{I_{CLN}} p(I_{CLN} | I) = \arg \max_{I_{CLN}} p(I_{CLN})p(I | I_{CLN})$$  \hspace{1cm} (5)$$

We now instantiate the prior $p(I_{CLN})$ and the likelihood $p(I | I_{CLN})$ with formulas from Eq. 2 and Eq. 4, and recast the problem as an optimisation over $w$: $w^* = \arg \max_w p(I | w)p(I | w)$. This can be simplified to the following loss function (see Supplementary section A for full derivation):

$$w^* = \arg \min_w \left( \sum_{i} \frac{(w_i - \mu)}{\sigma_i}^2 - 2\kappa \sum_{i,j} \frac{w_i w_j^T}{||w_i|| ||w_j||} \sigma^2_{\text{pixel}} ||I - f \circ G(w)||^2 \sigma^2_{\text{percept}} \right)$$  \hspace{1cm} (6)$$

which can be succinctly written as a weighted sum of four loss terms: $w^* = \arg \min_w \mathcal{L}_w + \lambda_{\text{colin}} \mathcal{L}_{\text{colin}} + \lambda_{\text{pixel}} \mathcal{L}_{\text{pixel}} + \lambda_{\text{percept}} \mathcal{L}_{\text{percept}}$, where $\mathcal{L}_w$ is the prior loss over $w$, $\mathcal{L}_{\text{colin}}$ is the colinearity loss on $w$, $\mathcal{L}_{\text{pixel}}$ is the pixelwise loss on the corrupted images, and $\mathcal{L}_{\text{percept}}$ is the perceptual loss, $\lambda_{\text{colin}} = -2\kappa$, $\lambda_{\text{pixel}} = \sigma^2_{\text{pixel}}$ and $\lambda_{\text{percept}} = \sigma^2_{\text{percept}}$. Given the Bayesian MAP solution $w^*$, the clean image is returned as $I^*_{CLN} = G(w^*)$.

---

\*2Since in our experiments $\psi$ is fixed, we drop the notation of $\psi$ in subsequent derivations.

\*3Model is equivalent to $p(I | f \circ G(w)) = N\left( \left[ \begin{array}{c} I \\ \phi(I) \\ \phi \circ f \circ G(w) \end{array} \right], \left[ \begin{array}{c} 0 \\ \sigma^2_{\text{pixel}} \mathbb{I}_{n_f^2} \\ 0 \end{array} \right] \right)$
Figure 3: Qualitative evaluation on FFHQ at different input resolutions. Left column shows low resolution inputs, while right column shows true high-quality images. ESRGAN and SRFBN show clear distortion and blurriness, while PULSE does not recover the true image due to strong priors. BRGM shows significant improvements, especially at low resolutions.

2.4 Sampling multiple reconstructions using Variational Inference

To sample multiple image reconstructions from the posterior distribution $p(w|I)$, we use Variational Inference. We use an approach similar to the Variational Auto-encoder [39, 53], where for each data-point we estimate a Gaussian distribution of latent vectors, with the main difference that we do not use an encoder network, but instead optimise the mean and covariance directly. Thus, our approach is also similar to Bayes-by-Backprop (BBB) [40], but we estimate a Gaussian distribution over the latent vector $w$, instead of the network weights as in their case.

Variational inference (Hinton and Van Camp 1993, Graves 2011) aims to find a parametric approximation $q(w|\theta)$, where $\theta$ are parameters to be learned, to the true posterior $p(w|I)$ over the latent inputs $w$ to the generator network. We seek to minimize:

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log p(w^{(i)}) - \log p(I|w^{(i)})$$

(7)

Using the same approach as in Bayes-by-Backprop [40], we approximate the expected value over $q(w|\theta)$ using Monte Carlo samples $w^{(i)}$ taken from $q(w|\theta)$:

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log p(w^{(i)}) - \log p(I|w^{(i)})$$

(7)

We parameterize $q(w|\theta)$ as a Gaussian distribution, although in practice we can choose any parametric form for $q$ (e.g. mixture of Gaussians) due to the Monte Carlo approximation. We sample the Gaussian by first sampling unit Gaussian noise $\epsilon$, and then shifting it by the variational mean $\mu_v$ and variational standard deviation $\sigma_v$. To ensure $\sigma_v$ is always positive, we re-parameterize it as $\sigma_v = \log(1 + \exp(\rho_v))$. The variational posterior parameters are $\theta = [\mu_v, \rho_v]$. The prior and likelihood models, $p(w^{(i)})$ and $p(I|w^{(i)})$ are as defined in Eq. 2 and Eq. 4.

While the role of the entropy term $\sum_{i=1}^{n} q(w^{(i)}|\theta)$ is to regularize the variance of $q$ and ensure there is no mode-collapse, we found it useful to also add a prior over the variational parameter $\sigma_v$, to give the samples more variability. We therefore optimise the following:

$$\theta^* = \arg\min_{\theta} -\log p(\theta) + \frac{1}{n} \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log p(w^{(i)}) - \log p(I|w^{(i)})$$

(8)

where $p(\theta) = f(\sigma_v; \alpha, \beta)$ is an inverse gamma distribution on the variational parameter $\sigma_v$, with concentration $\alpha$ and rate $\beta$. This prior, although optional, encourages larger standard deviations, which ensure that as much of the posterior as possible is covered. Note that, even with this prior, we
Figure 4: Qualitative evaluation on medical datasets at different resolutions. The left column shows input images, while the right column shows the true high-quality images. BRGM shows improved quality of reconstructions across all resolution levels and datasets. We used the exact same setup as in FFHQ in Fig. 3, without any dataset-specific parameter tuning.

don’t optimise $\sigma_v$ directly, rather we optimise $\rho_v$. We compute the gradients using the same approach as in Bayes-by-Backprop [40].

To sample the posterior $p(w|I)$, we also tried Stochastic Gradient Langevin Dynamics (SGLD) [54] and Variational Adam [55], which is equivalent to Variational Online Gauss-Newton (VOGN) [55] in our case when the batch size is 1. However, we could not get these methods to work in our setup: SGLD was adding noise of too-high magnitude and the optimisation quickly diverged, while Variational Adam produced little variability between the samples.

### 2.5 Model Optimisation

We optimise the loss in Eq. 6 using Adam [56] with learning rate of 0.001, while fixing $\lambda_{colin}$, $\lambda_{pixel}$ and $\lambda_{percept}$, $\alpha$ and $\beta$ a-priori. On our datasets, we found the following values to give good results: $\lambda_{colin} = 0.03$, $\lambda_{pixel} = 10^{-5}$, $\lambda_{percept} = 0.01$, $\alpha = 0.1$ and $\beta = 0.95$. In Fig 2, we show image super-resolution and in-painting starting from the original StyleGAN2 inversion, and gradually modify the loss function and optimisation until we arrive at our proposed solution. The original StyleGAN2 inversion results in line artifacts for super-resolution, while for in-painting it cannot reconstruct well. After removing the optimisation of noise layers from the original StyleGAN2 inversion[4] and switching to the extended latent space $W^+$, where each resolution-specific vector $w_1, \ldots, w_{18}$ is independent, the image quality improves for super-resolution, while for in-painting the existing image is recovered well, but the reconstructed part gets even worse. More improvements are observed by adding the pixelwise $L_2$ loss, mostly because the perceptual loss only operates at 256x256 resolution. Adding the prior on $w$ and the cosine loss produces smoother reconstructions with less artifacts, especially for in-painting.

### 2.6 Model training and evaluation

We train our model on data from three datasets: (i) 70,000 images from FFHQ [1] at 1024² resolution, 240,000 frontal-view chest X-ray image from MIMIC III [2] at 1024² resolution, as well as 7,329 middle coronal 2D slices from a collection of 5 brain datasets: ADNI [57], OASIS [58], PPMI [59], AIBL [60] and ABIDE [61]. We obtained ethical approval for all data used. All brain images were pre-registered rigidly. For all experiments, we trained the generator, in our case StyleGAN2, on 90% of the data, and left the remaining 10% for testing. We did not use the pre-trained StyleGAN2 on FFHQ as it was trained on the full FFHQ. Training was performed on 4 Titan-Xp GPUs using StyleGAN2 config-e, and was performed for 20,000,000 images shown to the discriminator (20,000 kimg), which took almost 2 weeks on our hardware. For a description of the generator training on all
Figure 5 illustrates our method’s performance on in-painting with arbitrary as well as rectangular masks, as compared to the leading in-painting model SN-PatchGAN [21]. Our method produces considerably better results than SN-PatchGAN. In particular, SN-PatchGAN lacks high-level semantics in the reconstruction, and cannot handle large masks. For example, in the first figure, when the mother is cropped out, SN-PatchGAN is unable to reconstruct the ear. Our method on the other hand is able to reconstruct the ear and the jawline. One reason for the lower performance of SN-PatchGAN could be that it was trained on CelebA, which has lower variation than FFHQ. In Supplementary Figs. 9, 10 and 11, we show further in-painting examples with our method as well as SN-PatchGAN [21], on all three datasets, and for different types of arbitrary masks.

In Fig. 6, we show samples from the variational posterior $q(w|\theta)$, for both super-resolution and in-painting. For super-resolution, we show an extreme downsampling example (x256) going from 1024x1024 to 4x4, in order to clearly see the potential variability in the reconstructions. The Variational Inference method gives samples of reasonably high variability and fidelity, although in harder cases (Supp. Fig. 13 and 14) it overfits the posterior.

In supplementary section E, we show in an ablation study over the hyper-parameters $\lambda_{\text{pixel}}$, $\lambda_{\text{percept}}$ and $\lambda_{\text{colin}}$ that our method is not sensitive to the choice of these parameters, as there are multiple levels of magnitudes giving good results. In addition, in supplementary section F, we show that BRGM-reconstructed images can be used for downstream processing, through an example of edema severity prediction on the Chest X-Ray images.
Input True Est. Mean Sample 1 Sample 2 Sample 3 Sample 4

Figure 6: Sampling using Variational Inference. Given the input image (left column), we show the estimated mean image $G(\mu_v)$ (third column), alongside samples around the mean $G(\mu_v + \sigma_v \epsilon)$.

3.1 Quantitative evaluation

Table 1 (left) reports performance metrics of super-resolution on 100 unseen images at different resolution levels. At low $16^2$ input resolutions, our method outperforms all other super-resolution methods consistently on all three datasets. However, at resolutions of $32^2$ and higher, SRFBN [19] achieves the lowest LPIPS[62] and root mean squared error (RMSE), albeit the qualitative results from this method showed that the reconstructions are overly smooth and lack detail. The performance degradation of our model is likely because the StyleGAN2 generator $G$ cannot easily generate these unseen images at high resolutions, although this is expected to change in the near future given the fast-paced improvements in such generator models. However, compared to those methods, our method is more generalisable as it is not specific to a particular type of corruption, and can increase the resolution by a factor higher than 4x. In Supplementary Table 2, we additionally provide PSNR, SSIM and MAE scores, which show a similar behavior to LPIPS and RMSE.

For quantitative evaluation on in-painting, we generated 7 masks similar to the setup of [43], and applied them in cyclical order to 100 unseen images from the test sets of each dataset. In Table 1 (top-right), we show that our method consistently outperforms SN-PatchGAN [21] with respect to all performance measures.

To account for human perceptual quality, we performed a forced-choice pairwise comparison test, which has been shown to be most sensitive and simple for users to perform [63]. Twenty raters were each shown 100 test pairs of the true image and the four reconstructed images by each algorithm, and raters were asked to choose the best reconstruction (see supplementary section D for more information on the design). We opted for this paired test instead of the mean opinion score (MOS) because it also accounts for fidelity of the reconstruction to the true image. This is important in our setup, because a method such as PULSE can reconstruct high-resolution faces that are nonetheless of a different person (see Fig. 3). In Table 1 (bottom-right), the results confirm that our method is the best at low $16^2$ resolution and second-best at $32^2$ resolution, with lower performance at $64^2$ resolution.

3.2 Method limitations and potential negative societal impact

In Supp. Fig. 17, we show failure cases on the super-resolution task. The reason for the failures is likely due to the limited generalisation abilities of the StyleGAN2 generator to such unseen images. We particularly note that, as opposed to the simple inversion of Image2StyleGAN [42], which relies on latent variables at high resolution to recover the fine details, we cannot optimize these high-resolution latent variables, thus having to rely on the proper ability of StyleGAN2 to extrapolate from lower-level latent variables. Another limitation of our method is the inconsistency between the downsampled input image and the given input image, which we exemplify in Supp. Figs 15 and 16. We attribute this again to the limited generalisation of the generator to these unseen images. The same inconsistency also applies to in-painting, as shown in Fig. 2.

By leveraging models pre-trained on FFHQ, our methodology can be potentially biased towards images from people that are over-represented in the dataset. On the medical datasets, we also have biases in disease labels. For example, the MIMIC dataset contains both healthy and pneumonia
Table 1: (left) Evaluation on (x4) super-resolution at different input resolution levels. Reported are LPIPS/RMSE scores. (top-right) Evaluation of BRGM and SN-PatchGAN on in-painting. (bottom-right) Human evaluation showing the proportion of votes for the best super-resolution reconstruction in the forced-choice pairwise comparison test. Bold numbers show best performance.

4 Conclusion

We proposed a simple Bayesian framework for performing different reconstruction tasks using deep generative models such as StyleGAN2. We estimate the optimal reconstruction as the Bayesian MAP estimate, and use Variational Inference to sample from an approximate posterior of all possible solutions. We demonstrated our method on two reconstruction tasks, and on three distinct datasets, including two challenging medical datasets, obtaining competitive results in comparison with state-of-the-art models. Future work can focus on jointly optimizing the parameters of the corruption models, as well as extending to more complex corruption models.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] Our main theoretical contributions are the Bayesian MAP and variational inference framework, while on the application side, our contribution are the application of the methods to three datasets, including two challenging medical datasets. We highlight this in the abstract and intro.
(b) Did you describe the limitations of your work? [Yes] Yes, see section 3.2.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] Yes, see section 3.2.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We confirm our paper fits in the ethics guidelines.

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes, we included all assumptions, e.g. that the jacobians were assumed constant and that we used the generalized change of variable formula for non-invertible functions.

(b) Did you include complete proofs of all theoretical results? [Yes] Yes, see supplementary section A.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See abstract

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See sections 2.5 and 2.6.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] While we could not include the standard deviation of performance metrics in the main manuscript due to lack of space, we will include it in the supplementary material.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 2.6.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Yes, we cited StyleGAN2 and StyleGAN-ADA as the main models we built on.

(b) Did you mention the license of the assets? [Yes] The StyleGAN2 model is released with a non-commercial license. Our model is released with CC-BY license. See end of section 2.6.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We released our code as a URL, see abstract.

(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] See section 2.6. We have been granted access to all brain datasets, as well as the MIMIC dataset.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All medical data used is already de-identified. For FFHQ, the people tagged granted access for their photographs to be used in computer vision research.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Yes, included screenshot of human evaluation in Supp. Fig. 12.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] Yes, see supp. section D.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Yes, see supp. section D.
Supplementary Material

A Derivation of loss function for the Bayesian MAP estimate

We assume $w = [w_1, \ldots, w_{18}] \in \mathbb{R}^{512 \times 18}$ is the StyleGAN2 latent vector, $I \in \mathbb{R}^{n \times n}$ is the corrupted input image, $G : \mathbb{R}^{512 \times 18} \rightarrow \mathbb{R}^{n_G \times n_G}$ is the StyleGAN2 generator network function, $f : \mathbb{R}^{n_G \times n_G} \rightarrow \mathbb{R}^{n_f \times n_f}$ is the corruption function, and $\phi : \mathbb{R}^{n_f \times n_f} \rightarrow \mathbb{R}^{n_\phi \times n_\phi}$ is a function describing the perceptual network. $n_f$, $n_G$, $n_f \times n_f$ and $n_\phi \times n_\phi$ are the resolutions of the clean image $G(w)$, corrupted image $f \circ G(w)$ and of the perceptual embedding $\phi \circ f \circ G(w)$. The full Bayesian posterior $p(w|I)$ of our model is proportional to:

$$p(w|I) \propto p(I)p(I|w) = \prod_i N(w_i|\mu, \sigma^2) \prod_{i,j} \mathcal{M}(\cos^{-1} \frac{w_i w_j^T}{|w_i||w_j|}, 0, \kappa) \mathcal{N}(I|f \circ G(w), \sigma^2_{\text{pixel}} \mathbb{I}_{n_f^2}) \mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{\text{percept}} \mathbb{I}_{n_\phi^2})$$

(9)

where $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}$ are means and standard deviations of the prior on $w_i$, $\mathcal{M}(.|0, \kappa)$ is the von Mises distribution$^4$ with mean zero and scale parameter $\kappa$, and $\sigma^2_{\text{pixel}} \mathbb{I}_{n_f^2}$ and $\sigma^2_{\text{percept}} \mathbb{I}_{n_\phi^2}$ are identity matrices scaled by variance terms.

The Bayesian MAP estimate is the vector $w^*$ that maximizes Eq. 9, and provides the most likely vector $w$ that could have generated input image $I$:

$$w^* = \arg \max_w p(w)p(I|w) = \arg \max_w \prod_i N(w_i|\mu, \sigma^2) \prod_{i,j} \mathcal{M}(\cos^{-1} \frac{w_i w_j^T}{|w_i||w_j|}, 0, \kappa) \mathcal{N}(I|f \circ G(w), \sigma^2_{\text{pixel}} \mathbb{I}_{n_f^2}) \mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{\text{percept}} \mathbb{I}_{n_\phi^2})$$

(10)

Since logarithm is a strictly increasing function that won’t change the output of the $\arg \max_w$ operator, we take the logarithm to simplify Eq. 10 to:

$$w^* = \arg \max_w \sum_i \log N(w_i|\mu, \sigma^2) + \sum_{i,j} \log \mathcal{M}(\cos^{-1} \frac{w_i w_j^T}{|w_i||w_j|}, 0, \kappa) + \log \mathcal{N}(I|f \circ G(w), \sigma^2_{\text{pixel}} \mathbb{I}_{n_f^2}) + \log \mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{\text{percept}} \mathbb{I}_{n_\phi^2})$$

(11)

We expand the probability density functions of each distribution to get:

$$w^* = \arg \max_w \sum_i C_1 - \frac{(w_i - \mu)^2}{2\sigma_i^2} + \sum_{i,j} C_2 + \kappa \cos(\cos^{-1} \frac{w_i w_j^T}{|w_i||w_j|}) + \frac{1}{2}(I - f \circ G(w))^T (\sigma^{-2}_{\text{pixel}} \mathbb{I}_{n_f^2})(I - f \circ G(w)) + \frac{1}{2}(I - \phi \circ f \circ G(w))^T (\sigma^{-2}_{\text{percept}} \mathbb{I}_{n_\phi^2})(I - \phi \circ f \circ G(w))$$

(12)

where $C_1 = \log (2\pi \sigma_i^2)^{-\frac{1}{2}}$, $C_2 = \log (2\pi \sigma_i^2)^{-\frac{1}{2}}$, $C_3 = \log ((2\pi)^n |\sigma_{\text{pixel}} \mathbb{I}_{n_f^2}|)^{-\frac{1}{2}}$, $C_4 = \log ((2\pi)^m |\sigma_{\text{percept}} \mathbb{I}_{n_\phi^2}|)^{-\frac{1}{2}}$ are constants with respect to $w$, so we can ignore them.

We remove the constants, multiply by $(-2)$, which requires switching to the $\arg \min_w$ operator, to get:

$\text{The von-Mises distribution is the analogous of the Gaussian distribution over angles } [0 - 2\pi]. \ \mathcal{M}(.|\mu, \kappa) \text{ is analogous to } \mathcal{N}(.|\mu, \sigma).$
Figure 7: Graphical model of our method. In gray shade are known observations or parameters: the input corrupted image $I_{COR}$, the parameters $\mu$, $\sigma$ and $\kappa$ defining the prior on latent vector $w$, and $\sigma_{pixel}$, $\sigma_{percept}$, the parameters defining the noise model over $I$. In the red box are the variational parameters $\mu_{v}$, $\sigma_{v}$ and $\rho_{v}$, defining an approximated Gaussian posterior over $w$ (section 2.4). Unknown latent variables (in white), to be estimated, are $w$, the latent vectors of StyleGAN, $I_{CLN}$, the clean image, and $I_{COR}$, the corrupted image simulated through the pipeline. Transformation $G$ is modelled by the StyleGAN2 generator, while $f$ by a known corruption model (e.g. downsampling with a known kernel).

$$w^* = \arg\min_w \sum_i \left( \frac{w_i - \mu}{\sigma_i} \right)^2 - 2\kappa \sum_{i,j} \frac{w_i w_j^T}{|w_i||w_j|} \right) + \sigma_{pixel}^2 \|I - f \circ G(w)\|_2^2 + \sigma_{percept}^2 \|I - \phi \circ f \circ G(w)\|_2^2 \right)$$

(13)

This is equivalent to Eq. 6, which finishes our proof.

**B Training StyleGAN2**

In Fig. 8, we show uncurated images generated by the cross-validated StyleGAN2 trained on our medical datasets, along with a few real examples. For the high-resolution X-rays, we notice that the image quality is very good, although some artifacts are still present: some text tags are not properly generated, some bones and rib contours are wiggly, and the shoulder bones show less contrast. For the brain dataset, we do not notice any clear artifacts, although we did not assess distributional preservation of regional volumes as in [64]. For the cross-validated FFHQ model, we obtained an FID of 4.01, around 0.7 points higher than the best result of 3.31 reported for config-e [4].

**C Inference times of our method**

The inference time of our method is as follows: for MAP inference, it takes between 32-34 seconds for 500 iterations on a 1024x1024 image, while for fitting the variational posterior parameters ($\mu$, $\rho$) it takes approximately 2.5 minutes for 500 iterations. Once the variational posterior is fit, the model can generate any arbitrary number of samples instantaneously.

### Dataset Performance Metrics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BRGM</th>
<th>PULSE [41]</th>
<th>ESRGAN [18]</th>
<th>SRFBN [19]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
<td>MAE↓</td>
<td>PSNR↑</td>
</tr>
<tr>
<td>FFHQ 16^2</td>
<td>20.13</td>
<td>0.74</td>
<td>17.46</td>
<td>19.51</td>
</tr>
<tr>
<td>FFHQ 32^2</td>
<td>22.74</td>
<td>0.74</td>
<td>12.52</td>
<td>15.37</td>
</tr>
<tr>
<td>FFHQ 64^2</td>
<td>24.16</td>
<td>0.70</td>
<td>10.63</td>
<td>15.74</td>
</tr>
<tr>
<td>X-ray 16^2</td>
<td>27.14</td>
<td>0.91</td>
<td>7.45</td>
<td>-</td>
</tr>
<tr>
<td>X-ray 32^2</td>
<td>27.84</td>
<td>0.84</td>
<td>6.77</td>
<td>-</td>
</tr>
<tr>
<td>X-ray 64^2</td>
<td>27.62</td>
<td>0.79</td>
<td>6.63</td>
<td>-</td>
</tr>
<tr>
<td>Brains 16^2</td>
<td>26.33</td>
<td>0.84</td>
<td>7.29</td>
<td>-</td>
</tr>
<tr>
<td>Brains 32^2</td>
<td>27.30</td>
<td>0.81</td>
<td>6.54</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Additional performance metrics (PSNR, SSIM and MAE) for the super-resolution evaluation.
Figure 8: Uncurated images generated by our StyleGAN2 generator trained on the chest X-ray dataset (MIMIC III) (top) and the brain dataset (bottom). Left images are random examples of real images from the actual datasets, while the right-side images are generated. The image quality is relatively good, albeit some anatomical artifacts are still observed, such as incomplete labels, wiggly bones or discontinuous wires.

Table 3: Performance metrics for super-resolution as in Table 1 (left), but additionally including the standard deviation of scores across the 100 test images.
Figure 9: Uncurated in-painting examples by BRGM on the FFHQ dataset, compared against SN-PatchGAN [21] and Deep Image Prior [47].
Figure 10: Uncurated in-painting examples by BRGM on the Chest X-ray dataset, compared against SN-PatchGAN [21] and Deep Image Prior [47].
Figure 11: Uncurated in-painting examples by BRGM on the brain dataset, compared against SN-PatchGAN [21] and Deep Image Prior [47].
D Additional Evaluation Results

To evaluate human perceptual quality, we performed a forced-choice pairwise comparison test as shown in Fig. 12. Each rater is shown a true, high-quality image on the left, and four potential reconstructions they have to choose from. For each input resolution level (16\(^2\), 32\(^2\), . . . ), we ran the human evaluation on 20 raters using 100 pairs of 5 images each (total of 500 images per experiment shown to each rater). We launched all human evaluations on Amazon Mechanical Turk. We paid $346 for the crowdsourcing effort, which gave workers an hourly wage of approximately $9. We obtained IRB approval for this study from our institution.

E Sensitivity to hyper-parameters

To understand how sensitive our model is to the hyper-parameters \(\lambda_{\text{pixel}}, \lambda_{\text{percept}}, \text{ and } \lambda_{\text{colin}}\), we performed in Table 4 an ablation analysis and computed the perceptual distance (LPIPS) and root mean squared error (RMSE) between our reconstructions and the true images. We performed this ablation on 64x super-resolution on 5 FFHQ test images, where we varied only one parameter at a time, keeping the others fixed to the following values: \(\lambda_{\text{pixel}} = 10^{-5}\), \(\lambda_{\text{percept}} = 10^{5}\) and \(\lambda_{\text{colin}} = 10^{-2}\). The results show that our method is not overly-sensitive to the choice of the lambda hyper-parameters, as there are multiple values over which results are satisfactory: \(\lambda_{\text{pixel}} \geq 10^{-6}\), \(10^{5} \leq \lambda_{\text{percept}} \leq 10^{6}\), and \(\lambda_{\text{colin}} \leq 10^{-2}\).

F Downstream processing of BRGM-restored images

In order to show that the BRGM-reconstructed images are useful for downstream tasks, we ran a ResNet-16 that predicts lung-edema severity on BRGM-restored Chest X-Ray images. Edema severity was predicted as either (1) no edema, (2) mild – vascular congestion, (3) moderate – interstitial edema and (4) severe – alveolar edema. The confusion matrix is shown in Table 5 (top table). We ran the model on 100 normal images (denoted as True), as well as the same 100 images but with the left-half masked and then reconstructed by our BRGM method (denoted as Reconstructed). We obtained a precision of 0.58 and a recall of 0.58. Most (73/100) of the severity scores are either preserved in the reconstructed images (diagonal), or change to an adjacent score (just above/below the diagonal). We note that, in general, the classification between adjacent severity levels (e.g. No edema vs Mild, Mild vs Moderate, etc ...) is a very hard problem, where the ResNet itself has an score of 45.03% (random guessing would give 25% on a balanced dataset).

We also note that these results depend on how much of the image is reconstructed. If we re-run the analysis with a smaller mask (only a quadrant masked and reconstructed instead of half of the image – see Table 5 bottom), we obtain significantly better results in the confusion matrix, and a precision of 0.75 and recall of 0.75. In particular, we note that the true severe cases improved their diagnosis.

<table>
<thead>
<tr>
<th>(\lambda_{\text{pixel}})</th>
<th>10(^{-7})</th>
<th>10(^{-6})</th>
<th>10(^{-5})</th>
<th>10(^{-4})</th>
<th>10(^{-3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPIPS/RMSE</td>
<td>0.65/49.48</td>
<td>0.57/36.33</td>
<td>0.59/36.73</td>
<td>0.58/35.87</td>
<td>0.58/36.77</td>
</tr>
<tr>
<td>(\lambda_{\text{percept}})</td>
<td>10(^3)</td>
<td>10(^4)</td>
<td>10(^5)</td>
<td>10(^6)</td>
<td>10(^7)</td>
</tr>
<tr>
<td>LPIPS/RMSE</td>
<td>0.60/36.42</td>
<td>0.59/35.19</td>
<td>0.57/34.67</td>
<td>0.56/33.56</td>
<td>0.60/38.86</td>
</tr>
<tr>
<td>(\lambda_{\text{colin}})</td>
<td>10(^{-4})</td>
<td>10(^{-3})</td>
<td>10(^{-2})</td>
<td>10(^{-1})</td>
<td>10(^0)</td>
</tr>
<tr>
<td>LPIPS/RMSE</td>
<td>0.56/31.57</td>
<td>0.57/34.38</td>
<td>0.58/35.35</td>
<td>0.64/44.83</td>
<td>0.68/63.95</td>
</tr>
</tbody>
</table>

Table 4: Results of ablation study over hyper-parameters \(\lambda_{\text{pixel}}, \lambda_{\text{percept}}\) and \(\lambda_{\text{colin}}\). Reported are perceptual distance (LPIPS) and pixelwise root mean squared errors (RMSE) between the true images and the reconstructed images.
Figure 12: Setup of our human study, using a forced-choice pairwise comparison design. Each rater is shown a true, high-quality image on the left, and four potential reconstructions (A-D) by different algorithms. They have to select which reconstruction best resembled the HQ image.

<table>
<thead>
<tr>
<th>True</th>
<th>No edema</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>No edema</td>
<td>53</td>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Mild</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Moderate</td>
<td>9</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Severe</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5: Results of downstream processing on BRGM-reconstructed images vs true images. (Top table) We show the confusion matrix of a ResNet-16 at edema prediction on BRGM-inpainted images vs True images. For the inpainted images, the mask covered 50% of the pixels (left-half). (Bottom table) As above, but only 25% of image is masked and then reconstructed.
Figure 13: Sampling of multiple reconstructions using Variational Inference on super-resolution tasks with varying factors. From left, we show the true image, the estimated variational mean, alongside five random samples around that mean. For each high-resolution (HR) image, we show the corresponding low-resolution (LR) image below. While for some of the images, the reconstructions don’t match the true image, the downsampled low-resolution images do match with the true image. We chose such extreme super-resolution in order to obtain a wide posterior distribution.
Figure 14: Sampling of multiple reconstructions using Variational Inference on in-painting tasks. From left, we show the true image, the estimated variational mean, alongside five random samples around that mean. For the mean and for each sample, we show both the clean image, as well as the true image with the in-painted area from the sample.
Figure 15: Inconsistency of our method on FFHQ, across different resolution levels, using uncurated example pictures. Left columns shows input images, and the middle columns show downsampled reconstructions (i.e. $f \circ G(w^*)$) where images were 4x super-resolved, then downsampled by 4x to match again the input. Right columns show difference between input and the downsampled reconstructions. For higher resolution inputs (128x128), the method cannot accurately reconstruct the input image, likely because the generator has limited generalisability to such unseen faces from FFHQ (our method was trained not on the entire FFHQ, but on a training subset). The difference maps, representing $x3$ scaled mean absolute errors, show that certain regions in particular are not well reconstructed, such as the hair of the girl on the right.

G Method Inconsistency

One caveat of our method is that it can create reconstructions that are inconsistent with the input data. We highlight this in Figs. 15 and 16. This is because our method relies on the ability of a pre-trained generator to generate any potential realistic image as input. In addition to that, in Eq. (6), our method optimizes the pixelwise and perceptual loss terms between the input image and the downsampled reconstruction. As we show in Figs. 15 and 16, while there are little differences between the input and the reconstruction at low 16x16 resolutions, at higher 128x128 resolutions, these differences become larger and more noticeable. Another aspect that contributes to this issue is the extra prior term $L_{\text{cosine}}$, which is however required to ensure better reconstructions (see Fig 2). Nevertheless, we believe that the inconsistency is fundamentally caused by limitations of the generator $G$, that will be solved in the near future with better generator models that offer improved generalisability to unseen images.
Figure 16: Inconsistency of our method on the medical datasets, using uncurated examples. Same setup as in Fig. 15.
Figure 17: Failure cases of our method.