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

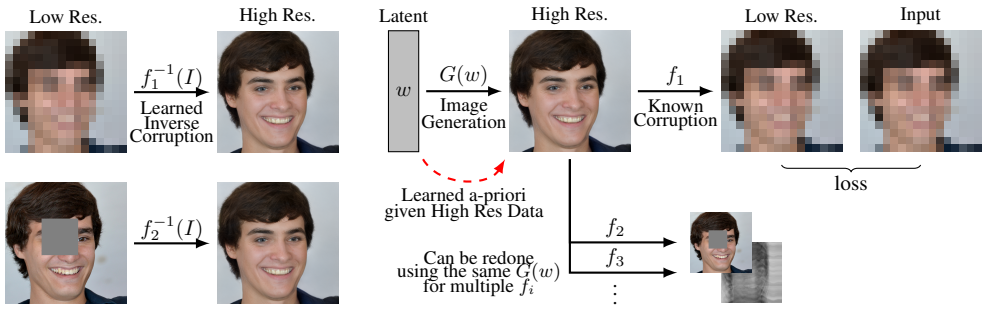
Abstract

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

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications



(a) Task-specific methods

(b) Generative approach

Figure 1: (a) Classical deep-learning methods for image reconstruction learn to invert specific corruption models such as downsampling with a specific kernel or in-painting with rectangular masks. (b) We use a generative approach that can handle arbitrary corruption processes, such as downsampling or in-painting with an arbitrary mask, by optimizing for it on-the-fly at inference time. Given a latent vector w, we use generator G to generate clean images $G(w)$, followed by a corruption model f to generate a corrupted image $f \circ G(w)$. Given an input image I, we find the latent w^* that *generated the input image* using the Bayesian MAP estimate $w^* = \arg\max_w p(w)p(I|f \circ G(w))$, and we use Variational Inference to sample from the posterior $p(w|I)$. 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).

centers. Therefore, the ability to leverage *pre-trained* models for solving downstream prediction or reconstruction tasks becomes crucial.

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 are 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 $w^* = \arg\max_w p(w)p(I|f \circ G(w))$, 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 conditionalized flow models [35], (ii) Langevin dynamics [36, 37] or (iii) Variational Inference (VI) [38], each having several advantages and disadvantages: (i) conditionalized 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 $G(w^*)$,

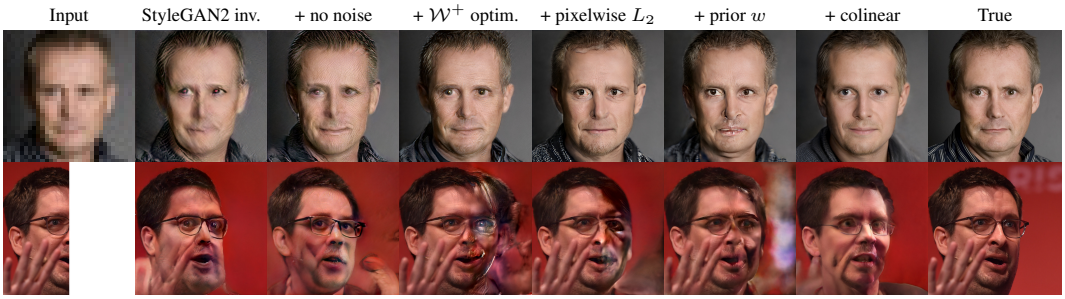


Figure 2: Reconstructions as the loss function evolves from the original StyleGAN2 inversion to our proposed method. Top row shows super resolution, while bottom row shows in-painting. We start from (a) the original StyleGAN2 inversion, and (b) remove noise optimisation, (c) extend optimisation to full $\mathcal{W}^+$ space, (d) add pixelwise L_2 term, (e) add prior on w latent variables and (f) add colinear loss term for 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 $w = [w_1, \ldots, w_{18}] \in \mathbb{R}^{512\times18}$ is the latent vector of StyleGAN2 (18 vectors for each resolution level), $G : \mathbb{R}^{512\times18} \rightarrow \mathbb{R}^{n_G\times n_G}$ 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:

$$p(I_{CLN}) \coloneqq p(G(w)) = p(w)\left|\frac{\partial G(w)}{\partial w}\right|^{-1} \tag{1}$$

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 $p(w)$:

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

[1]The supp. section of [51] presents an excellent introduction to the generalized change of variable theorem.

where $cos^{-1}\frac{w_i w_j^T}{|w_i||w_j|}$ is the angle between vectors w_i and w_j, and $\mathcal{M}(.|0,\kappa)$ is the *von Mises distribution* with mean zero and scale parameter κ which ensures that vectors w_i are aligned. This distribution is analogous to a Gaussian distribution over angles in $[0, 2\pi]$. We compute μ and σ 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 $p(I|I_{CLN})$ with a potentially probabilistic forward corruption process $f(I_{CLN};\psi)$, parameterized[2] by ψ. 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 k. For a high-resolution image I_{CLN}, this produces a low-resolution (corrupted) image $I_{COR} = (I_{CLN} \circledast k) \downarrow^s$, where $\circledast$ denotes convolution and $\downarrow^s$ 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 I_{CLN} and a 2D binary mask M, it produces a cropped-out (corrupted) image $I_{COR} = I_{CLN} \odot M$, 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:

$$p(I|I_{CLN}) = p(I|G(w)) = p(I|f \circ G(w))|J_f\left(G(w)\right)|^{-1} \tag{3}$$

where $J_f\left(G(w)\right) = \frac{\partial f \circ G(w)}{\partial G(w)}$ 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[3]:

$$p(I|f \circ G(w)) = \mathcal{N}(I|f \circ G(w), \sigma^2_{pixel}\mathbb{I}_{n_f^2})\, \mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{percept}\mathbb{I}_{n_\phi^2}) \tag{4}$$

where $\phi : \mathbb{R}^{n_f \times n_f} \rightarrow \mathbb{R}^{n_\phi \times n_\phi}$ is the VGG network, $\sigma^2_{pixel}\mathbb{I}_{n_f^2}$ and $\sigma^2_{percept}\mathbb{I}_{n_\phi^2}$ 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, $\phi(I)$, $f \circ G(w)$ 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:

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

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: $w^* = \arg\max_w p(w)p(I|w)$. This can be simplified to the following loss function (see Supplementary section A for full derivation):

$$w^* = \arg\min_w \underbrace{\sum_i \left(\frac{w_i - \mu}{\sigma_i}\right)^2}_{\mathcal{L}_w} - 2\kappa \underbrace{\sum_{i,j} \frac{w_i w_j^T}{|w_i||w_j|}}_{\mathcal{L}_{colin}} + \sigma^{-2}_{pixel} \underbrace{\|I - f \circ G(w)\|_2^2}_{\mathcal{L}_{pixel}} + \sigma^{-2}_{percept} \underbrace{\|I - \phi \circ f \circ G(w)\|_2^2}_{\mathcal{L}_{percept}} \tag{6}$$

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

[2] Since in our experiments ψ is fixed, we drop the notation of ψ in subsequent derivations.

[3] Model is equivalent to $p(I|f \circ G(w)) = \mathcal{N}\left(\begin{bmatrix} I \\ \phi(I) \end{bmatrix} \middle| \begin{bmatrix} f \circ G(w) \\ \phi \circ f \circ G(w) \end{bmatrix}, \begin{bmatrix} \sigma^2_{pixel}\mathbb{I}_{n_f^2} & 0 \\ 0 & \sigma^2_{percept}\mathbb{I}_{n_\phi^2} \end{bmatrix}\right)$

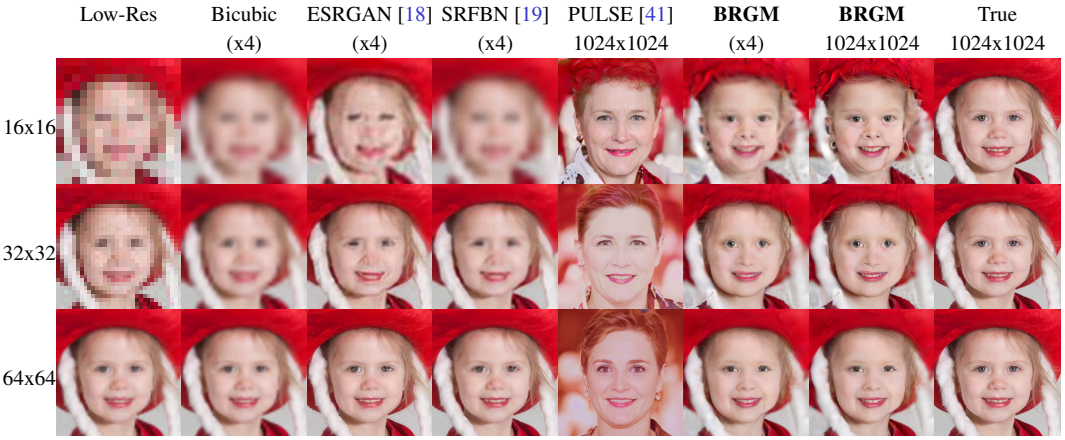


Figure 3: Qualitative evaluation on FFHQ at different input resolutions. Left column shows low resolution inputs, while right column shows true high-quality images. ESRGAN and SRFBN show clear distortion and blurriness, while PULSE does not recover the true image due to strong priors. BRGM shows significant 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*, 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 θ 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:

$$\theta^* = \arg\min_{\theta} KL\left[q(w|\theta)||p(w|I)\right] = \arg\min_{\theta} \int q(w|\theta)\ log\ \frac{q(w|\theta)}{p(w)p(I|w)} dw$$

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 $q(w|\theta)$:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} log\ q(w^{(i)}|\theta) - log\ p(w^{(i)}) - log\ p(I|w^{(i)}) \tag{7}$$

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 ϵ, and then shifting it by the variational mean μ_v and variational standard deviation σ_v. To ensure σ_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 σ_v, to give the samples more variability. We therefore optimise the following:

$$\theta^* = \arg\min_{\theta} - log\ p(\theta) + \sum_{i=1}^{n} log\ q(w^{(i)}|\theta) - log\ p(w^{(i)}) - log\ p(I|w^{(i)}) \tag{8}$$

where $p(\theta) = f(\sigma_v; \alpha, \beta)$ is an inverse gamma distribution on the variational parameter σ_v, with concentration α and rate β. 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

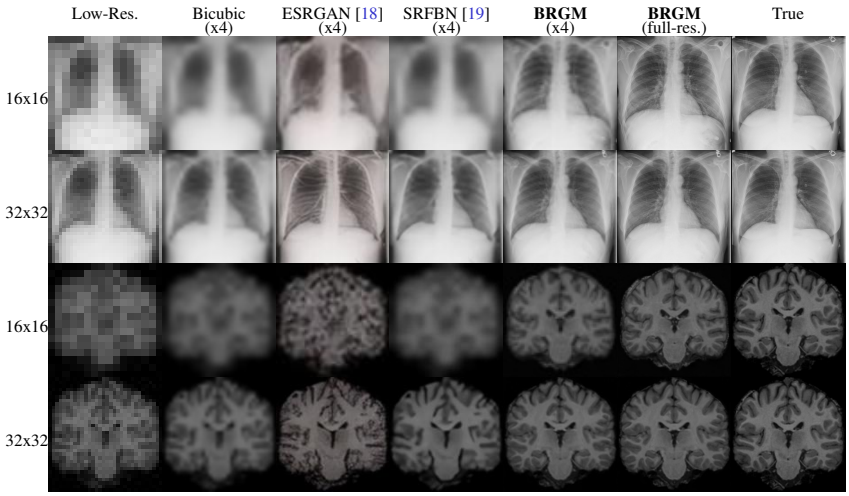


Figure 4: Qualitative evaluation on medical datasets at different resolutions. The left column shows input images, while the right column shows the true high-quality images. BRGM shows improved quality of reconstructions across all resolution levels and datasets. We used the exact same setup as in FFHQ in Fig. 3, without any dataset-specific parameter tuning.

don't optimise σ_v directly, rather we optimise ρ_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 λ_{colin}, λ_{pixel} and $\lambda_{percept}$, α and β 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 $\mathcal{W}^+$, where each resolution-specific vector w_1, ... , 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 1024^2 resolution, 240,000 frontal-view chest X-ray image from MIMIC III [2] at 1024^2 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: Comparison between our method and SN-PatchGAN [21] on in-painting. SN-PatchGAN fails on large masks, while our method can still recover the high-level structure.

three datasets, see Supp. Section B. For a description of the inference times of BRGM MAP and VI, see Supp. Section C.

We compare our method with the state-of-the-art methods in super-resolution (ESRGAN [18] and SRFBN [19]) as well as in-painting (SN-PatchGAN [21]). We additionally compared with PULSE [41] due to its closeness to our method, as well as the use of the state-of-the-art StyleGAN model. For these methods, we downloaded the pre-trained models. We could not compare with NeRF [49] as it requires multiple views of the same object. Since DIP [47] uses statistics in the input image only and cannot handle large masks or large super-resolution factors, we did not include it in the performance evaluation, although we show results with DIP in the supplementary material. For PULSE [41], we only tested it on FFHQ, as for the medical datasets it required re-training of the StyleGAN2 generator in their own PyTorch implementation, that was different from the official StyleGAN2 implementation. We release our code with the CC-BY license.

3 Results

We applied BRGM and the other models on super-resolution at different resolution levels (Fig. 3 and Fig. 4). On all three datasets, our method performs considerably better than other models, in particular at lower input resolutions: ESRGAN yields jittery artifacts, SRFBN gives smoothed-out results, while PULSE generates very high-resolution images that don't match the true image, likely due to the hard projection of their optimized latent to S^{d-1}, the unit sphere in d-dimensions, as opposed to a soft prior term such as $\mathcal{L}_w$ in our case. Moreover, as opposed to ESRGAN and SRFBN, both our model as well as PULSE can perform more than x4 super-resolution, going up to 1024x1024. Without changing any hyper-parameters, we observe similar trends on the other two medical datasets.

Fig. 5 illustrates our method's performance on in-painting with arbitrary as well as rectangular masks, as compared to 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 λ_{pixel}, $\lambda_{percept}$ and λ_{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.

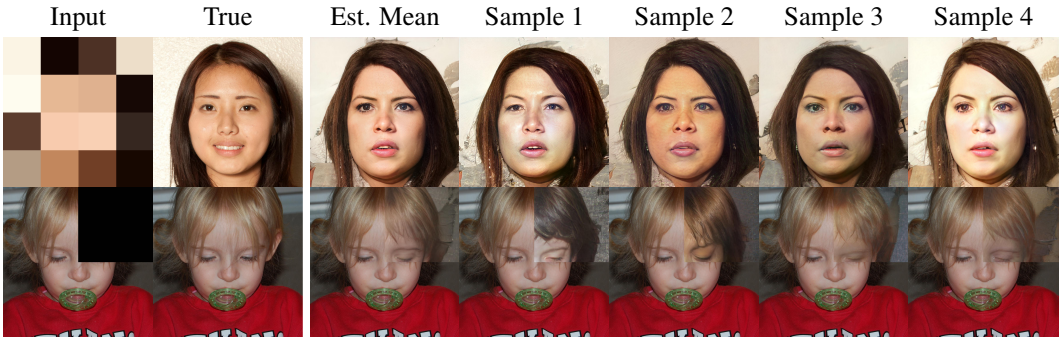


Figure 6: Sampling using Variational Inference. Given the input image (left column), we show the estimated mean image $G(\mu_v)$ (third column), alongside samples around the mean $G(\mu_v + \sigma_v \epsilon)$.

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 out 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

Super-resolution (LPIPS↓ / RMSE↓)

Dataset	**BRGM**	PULSE [41]	ESRGAN [18]	SRFBN [19]
FFHQ 16^2	**0.24**/25.66	0.29/27.14	0.35/29.32	0.33/**22.07**
FFHQ 32^2	0.30/18.93	0.48/42.97	0.29/23.02	**0.23/12.73**
FFHQ 64^2	0.36/16.07	0.53/41.31	0.26/18.37	**0.23/9.40**
X-ray 16^2	**0.18/11.61**	-	0.32/14.67	0.37/12.28
X-ray 32^2	0.23/10.47	-	0.32/12.56	**0.21/6.84**
X-ray 64^2	0.31/10.58	-	0.30/8.67	**0.22/5.32**
Brains 16^2	**0.12/12.42**	-	0.34/22.81	0.33/12.57
Brains 32^2	**0.17**/11.08	-	0.31/14.16	0.18/**6.80**

In-painting

	BRGM				SN-PatchGAN [21]			
Dataset	LPIPS	RMSE	PSNR	SSIM	LPIPS	RMSE	PSNR	SSIM
FFHQ	**0.19**	**24.28**	**21.33**	**0.84**	0.24	30.75	19.67	0.82
X-ray	**0.13**	**13.55**	**27.47**	**0.91**	0.20	27.80	22.02	0.86
Brains	**0.09**	**8.65**	**30.94**	**0.88**	0.22	24.74	21.47	0.75

Human evaluation (proportion of votes for best image)

Dataset	**BRGM**	PULSE [41]	ESRGAN [18]	SRFBN [19]
FFHQ 16^2	**42%**	32%	11%	15%
FFHQ 32^2	39%	2%	12%	**47%**
FFHQ 64^2	14%	8%	32%	**45%**

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 re-construction in the forced-choice pairwise comparison test. Bold numbers show best performance.

lung images, but many other lung conditions are not covered, while for the brain dataset it contains healthy brains as well as Alzheimer's and Parkinson's, but does not cover rarer brain diseases. Before deployment in the real-world, further work is required to make the method robust to diverse inputs, in order to avoid negative impact on the users.

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

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[2] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[3] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[5] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[7] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. 2016.

[9] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[10] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[11] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[12] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[14] Andrei Nikolaevich Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.

[15] Mário AT Figueiredo and Robert D Nowak. A bound optimization approach to wavelet-based image deconvolution. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–782. IEEE, 2005.

[16] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696, 2009.

[17] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[18] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[19] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[20] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[21] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[24] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 860–867. IEEE, 2005.

[25] David H Brainard and William T Freeman. Bayesian color constancy. *JOSA A*, 14(7):1393–1411, 1997.

[26] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics.

[27] William T Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368 (6471):542–545, 1994.

[28] Marshall F Tappen Bryan C Russell and William T Freeman. Exploiting the sparse derivative prior for super-resolution and image demosaicing.

[29] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[30] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.

[31] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *European Conference on Computer Vision*, pages 262–277. Springer, 2020.

[32] Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *International Conference on Machine Learning*, pages 399–409. PMLR, 2020.

[33] Paul Hand, Oscar Leong, and Vladislav Voroninski. Phase retrieval under a generative prior. *arXiv preprint arXiv:1807.04261*, 2018.

[34] Varun A Kelkar and Mark A Anastasio. Prior image-constrained reconstruction using style-based generative models. *arXiv preprint arXiv:2102.12525*, 2021.

[35] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

[36] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jonathan I Tamir. Robust compressed sensing mri with deep generative priors. *arXiv preprint arXiv:2108.01368*, 2021.

[37] Ajil Jalal, Sushrut Karmalkar, Alexandros G Dimakis, and Eric Price. Instance-optimal compressed sensing via posterior sampling. *arXiv preprint arXiv:2106.11438*, 2021.

[38] Jay Whang, Erik Lindgren, and Alex Dimakis. Composing normalizing flows for inverse problems. In *International Conference on Machine Learning*, pages 11158–11169. PMLR, 2021.

[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[40] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[41] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020.

[42] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019.

[43] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.

[44] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.

[45] Jonas Adler and Ozan Öktem. Deep bayesian inversion. *arXiv preprint arXiv:1811.05910*, 2018.

[46] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. *ICLR*, 2(5):3, 2018.

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[48] Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, and Peer-Timo Bremer. Mimicgan: Robust projection onto image manifolds with corruption mimicking. *International Journal of Computer Vision*, pages 1–19, 2020.

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

[50] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017.

[51] Milan Cvitkovic and Günther Koliander. Minimal achievable sufficient statistic learning. In *International Conference on Machine Learning*, pages 1465–1474. PMLR, 2019.

[52] Steven G Krantz and Harold R Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.

[53] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[54] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[55] Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *International Conference on Machine Learning*, pages 2611–2620. PMLR, 2018.

[56] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[57] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[58] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

[59] Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M Tanner, John Q Trojanowski, et al. The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477, 2018.

[60] Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, et al. The australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009.

[61] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17:16–23, 2018.

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[63] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, pages 2478–2491. Wiley Online Library, 2012.

[64] Petru-Daniel Tudosiu, Thomas Varsavsky, Richard Shaw, Mark Graham, Parashkev Nachev, Sebastien Ourselin, Carole H Sudre, and M Jorge Cardoso. Neuromorphologicaly-preserving volumetric data encoding using VQ-VAE. *arXiv preprint arXiv:2002.05692*, 2020.

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. $n_G \times n_G$, $n_f \times n_f$ 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:

$$
\begin{aligned}
p(w|I) \propto p(w)p(I|w) = &\prod_i \mathcal{N}(w_i|\mu, \sigma^2) \prod_{i,j} \mathcal{M}(cos^{-1}\frac{w_i w_j^T}{|w_i||w_j|}|0, \kappa) \\
&\mathcal{N}(I|f \circ G(w), \sigma^2_{pixel}\mathbb{I}_{n_f^2})\, \mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{percept}\mathbb{I}_{n_\phi^2})
\end{aligned}
\tag{9}
$$

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 κ, and $\sigma^2_{pixel}\mathbb{I}_{n_f^2}$ and $\sigma^2_{percept}\mathbb{I}_{n_\phi^2}$ 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:

$$
\begin{aligned}
&w^* = \arg\max_w p(w)p(I|w) = \arg\max_w \prod_i \mathcal{N}(w_i|\mu, \sigma^2) \prod_{i,j} \mathcal{M}(cos^{-1}\frac{w_i w_j^T}{|w_i||w_j|}|0, \kappa) \\
&\mathcal{N}(I|f \circ G(w), \sigma^2_{pixel}\mathbb{I}_{n_f^2})\mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{percept}\mathbb{I}_{n_\phi^2})
\end{aligned}
\tag{10}
$$

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:

$$
\begin{aligned}
&w^* = \arg\max_w \sum_i log\, \mathcal{N}(w_i|\mu, \sigma^2) + \sum_{i,j} log\, \mathcal{M}(cos^{-1}\frac{w_i w_j^T}{|w_i||w_j|}|0, \kappa) + \\
&log\, \mathcal{N}(I|f \circ G(w), \sigma^2_{pixel}\mathbb{I}_{n_f^2}) + log\, \mathcal{N}(\phi(I)|\phi \circ f \circ G(w), \sigma^2_{percept}\mathbb{I}_{n_\phi^2})
\end{aligned}
\tag{11}
$$

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

$$
\begin{aligned}
&w^* = \arg\max_w \sum_i C_1 - \frac{(w_i - \mu)^2}{2\sigma_i^2} + \sum_{i,j} C_2 + \kappa cos(cos^{-1}\frac{w_i w_j^T}{|w_i||w_j|}) \\
&+ C_1 - \frac{1}{2}(I - f \circ G(w))^T(\sigma^{-2}_{pixel}\mathbb{I}_{n_f^2})(I - f \circ G(w)) \\
&+ C_2 - \frac{1}{2}(I - \phi \circ f \circ G(w))^T(\sigma^{-2}_{percept}\mathbb{I}_{n_\phi^2})(I - \phi \circ f \circ G(w))
\end{aligned}
\tag{12}
$$

where $C_1 = log\ (2\pi\sigma_i^2)^{-\frac{1}{2}}$, $C_2 = log\ (-2\pi I_0(\kappa))$, $C_3 = log\ ((2\pi)^{n^2}|\sigma_{pixel}\mathbb{I}_{n_f^2}|)^{-\frac{1}{2}}$ and $C_4 = log\ ((2\pi)^{m^2}|\sigma_{percept}\mathbb{I}_{n_\phi^2}|)^{-\frac{1}{2}}$ are constants with respect to w, so we can ignore them.

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

[4] The von-Mises distribution is the analogous of the Gaussian distribution over angles $[0 - 2\pi]$. $\mathcal{M}(.|\mu, \kappa)$ is analogous to $\mathcal{N}(.|\mu, \sigma)$, where $\kappa^{-1} = \sigma^2$

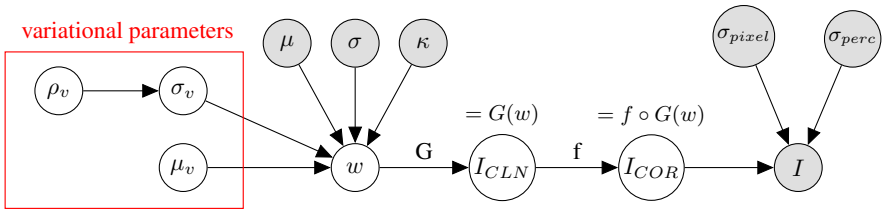


Figure 7: Graphical model of our method. In gray shade are known observations or parameters: the input corrupted image I, the parameters μ, σ and κ defining the prior on latent vector w, and σ_{pixel}, $\sigma_{percept}$, the parameters defining the noise model over I. In the red box are the variational parameters μ_v, σ_v and ρ_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).

$$
\begin{aligned}
w^* = \arg\min_w \sum_i \left(\frac{w_i - \mu}{\sigma_i} \right)^2 - 2\kappa \sum_{i,j} \frac{w_i w_j^T}{|w_i||w_j|} \\
+ \sigma_{pixel}^{-2} \|I - f \circ G(w)\|_2^2 + \sigma_{percept}^{-2} \|I - \phi \circ f \circ G(w)\|_2^2
\end{aligned}
\tag{13}
$$

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 (μ, ρ) 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	**BRGM**			PULSE [41]			ESRGAN [18]			SRFBN [19]		
	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓
FFHQ 16^2	20.13	0.74	17.46	19.51	0.68	19.20	18.91	0.69	20.01	21.43	0.76	15.01
FFHQ 32^2	22.74	0.74	12.52	15.37	0.35	33.13	21.10	0.72	14.53	26.28	0.89	7.46
FFHQ 64^2	24.16	0.70	10.63	15.74	0.37	31.54	23.14	0.72	10.94	28.96	0.90	5.21
X-ray 16^2	27.14	0.91	7.45	-	-	-	25.17	0.87	10.14	26.88	0.92	7.63
X-ray 32^2	27.84	0.84	6.77	-	-	-	26.44	0.81	8.36	31.80	0.95	3.71
X-ray 64^2	27.62	0.79	6.63	-	-	-	29.47	0.87	5.33	33.91	0.95	2.47
Brains 16^2	26.33	0.84	7.29	-	-	-	21.06	0.60	14.27	26.21	0.77	8.62
Brains 32^2	27.30	0.81	6.54	-	-	-	25.23	0.78	8.35	31.60	0.93	3.86

Table 2: Additional performance metrics (PSNR, SSIM and MAE) for the super-resolution evaluation.

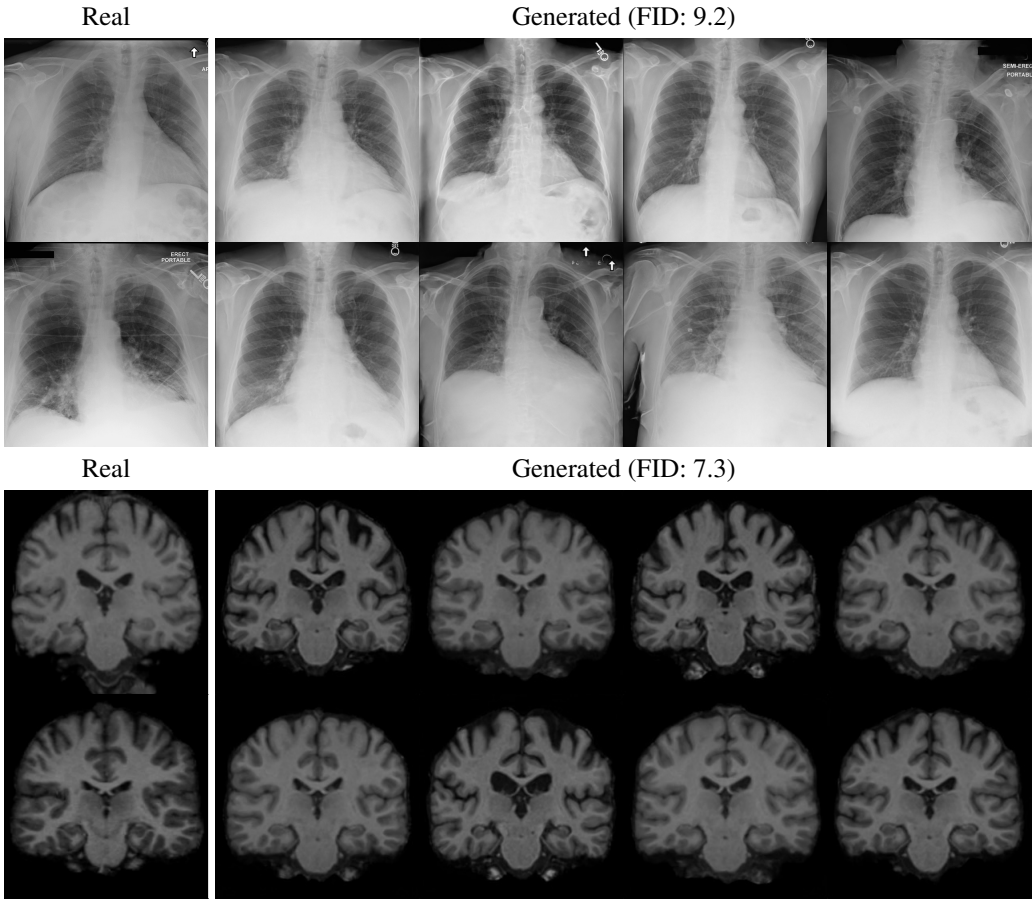


Figure 8: Uncurated images generated by our StyleGAN2 generator trained on the chest X-ray dataset (MIMIC III) (top) and the brain dataset (bottom). Left images are random examples of real images from the actual datasets, while the right-side images are generated. The image quality is relatively good,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

Dataset	**BRGM**		PULSE [41]		ESRGAN [18]		SRFBN [19]	
	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓	LPIPS↓	RMSE↓
FFHQ 16^2	0.24 ± 0.07	25.66 ± 6.13	0.29 ± 0.07	27.14 ± 4.08	0.35 ± 0.07	29.32 ± 6.52	0.33 ± 0.05	22.07 ± 4.47
FFHQ 32^2	0.30 ± 0.07	18.93 ± 3.95	0.48 ± 0.08	42.97 ± 3.78	0.29 ± 0.05	23.02 ± 5.48	0.23 ± 0.04	12.73 ± 3.08
FFHQ 64^2	0.36 ± 0.06	16.07 ± 3.21	0.53 ± 0.07	41.31 ± 3.57	0.26 ± 0.05	18.37 ± 5.06	0.23 ± 0.04	9.40 ± 2.48
FFHQ 128^2	0.34 ± 0.05	15.84 ± 3.23	0.57 ± 0.06	34.89 ± 2.21	0.15 ± 0.05	15.84 ± 4.83	0.09 ± 0.02	7.55 ± 2.30
X-ray 16^2	0.18 ± 0.05	11.61 ± 3.22	-	-	0.32 ± 0.07	14.67 ± 4.48	0.37 ± 0.04	12.28 ± 4.39
X-ray 32^2	0.23 ± 0.05	10.47 ± 2.04	-	-	0.32 ± 0.05	12.56 ± 3.34	0.21 ± 0.03	6.84 ± 2.01
X-ray 64^2	0.31 ± 0.04	10.58 ± 1.81	-	-	0.30 ± 0.03	8.67 ± 1.86	0.22 ± 0.02	5.32 ± 1.44
X-ray 128^2	0.27 ± 0.03	10.53 ± 1.91	-	-	0.20 ± 0.02	7.19 ± 1.34	0.07 ± 0.01	4.33 ± 1.30
Brains 16^2	0.12 ± 0.03	12.42 ± 1.71	-	-	0.34 ± 0.04	22.81 ± 3.26	0.33 ± 0.03	12.57 ± 1.51
Brains 32^2	0.17 ± 0.03	11.08 ± 1.29	-	-	0.31 ± 0.03	14.16 ± 2.36	0.18 ± 0.03	6.80 ± 1.14

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].

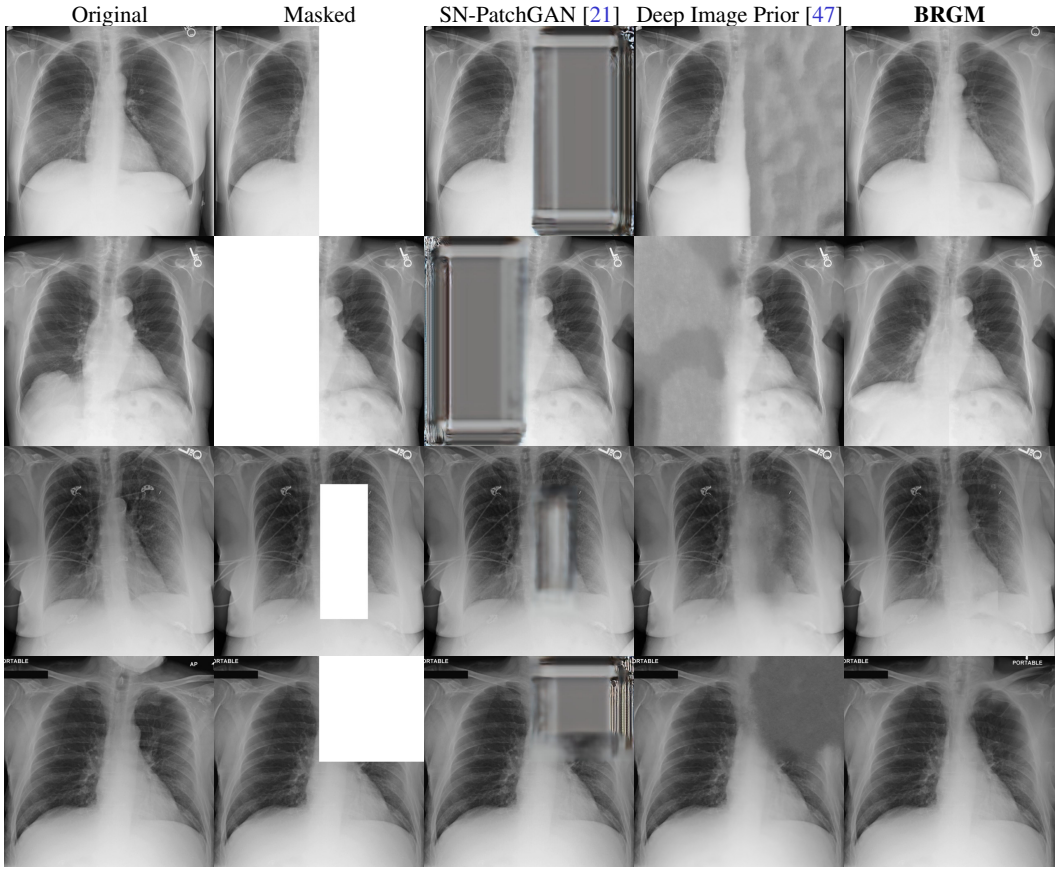


Figure 10: Uncurated in-painting examples by BRGM on the Chest X-ray dataset, compared against SN-PatchGAN [21] and Deep Image Prior [47].

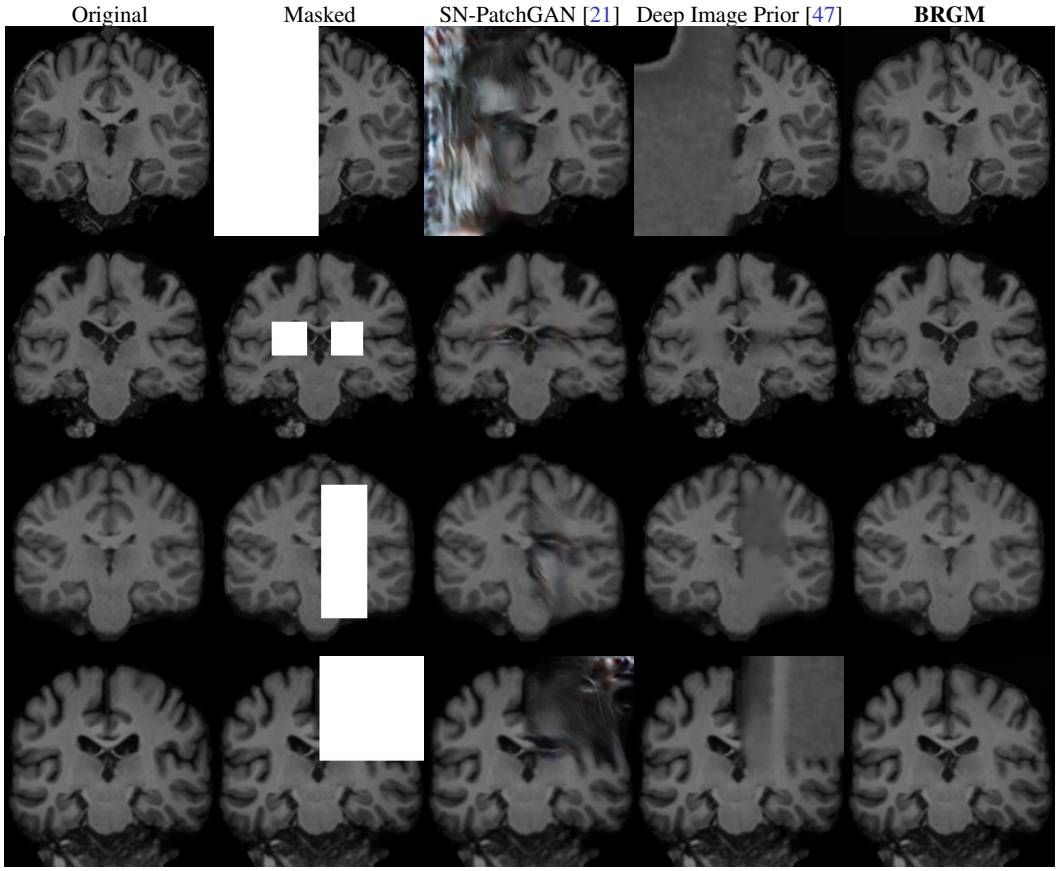


Figure 11: Uncurated in-painting examples by BRGM on the brain dataset, compared against SN-PatchGAN [21] and Deep Image Prior [47].

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 λ_{pixel}, $\lambda_{percept}$ and λ_{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: $\lambda_{pixel} = 10^{-5}$, $\lambda_{percept} = 10^5$ and $\lambda_{colin} = 10^{-2}$. 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: $\lambda_{pixel} \geq 10^{-6}$, $10^5 \leq \lambda_{percept} \leq 10^6$, and $\lambda_{colin} \leq 10^{-2}$.

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.

λ_{pixel}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
LPIPS/RMSE	0.65/49.48	0.57/36.33	0.59/36.73	0.58/35.87	0.58/36.77
$\lambda_{percept}$	10^3	10^4	10^5	10^6	10^7
LPIPS/RMSE	0.60/36.42	0.59/35.19	0.57/34.67	0.56/33.56	0.60/38.86
λ_{colin}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^0
LPIPS/RMSE	0.56/31.57	0.57/34.38	0.58/35.35	0.64/44.83	0.68/63.95

Table 4: Results of ablation study over hyper-parameters λ_{pixel}, $\lambda_{percept}$ and λ_{colin}. Reported are perceptual distance (LPIPS) and pixelwise root mean squared errors (RMSE) between the true images and the reconstructed images.

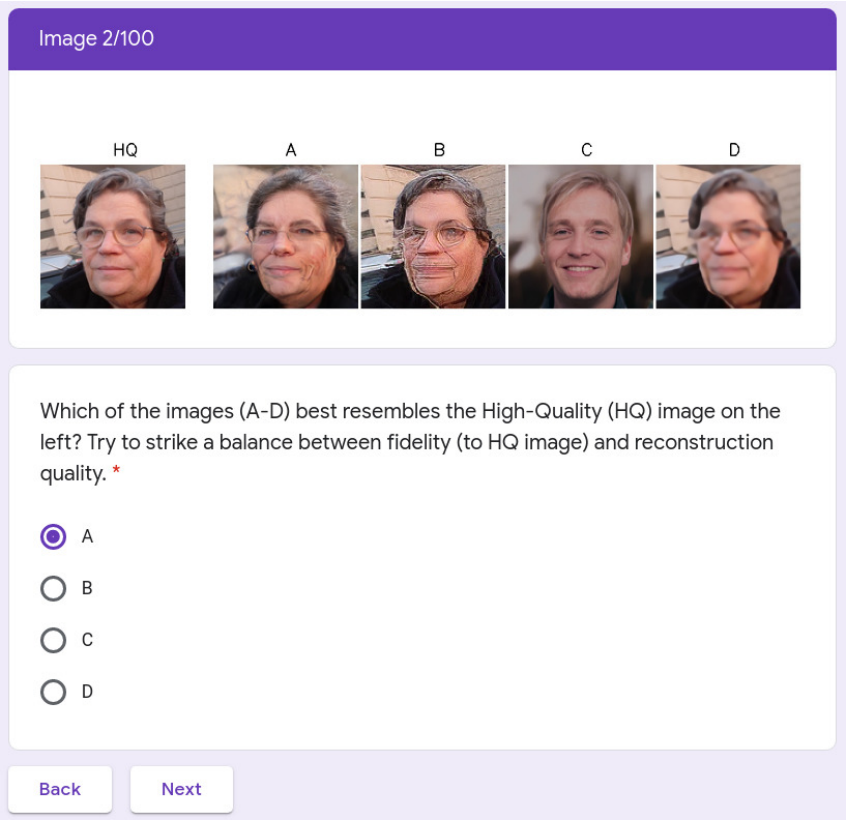


Figure 12: Setup of our human study, using a forced-choice pairwise comparison design. Each rater is shown a true, high-quality image on the left, and four potential reconstructions (A-D) by different algorithms. They have to select which reconstruction best resembled the HQ image.

Mask 50% of image (left half of image)

True Reconstructed	No edema	Mild	Moderate	Severe
No edema	53	2	11	1
Mild	5	0	1	0
Moderate	9	3	5	0
Severe	3	3	4	0

Mask 25% of image (top-right quadrant only)

True Reconstructed	No edema	Mild	Moderate	Severe
No edema	66	2	8	1
Mild	3	1	3	0
Moderate	5	0	4	1
Severe	0	1	1	4

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.

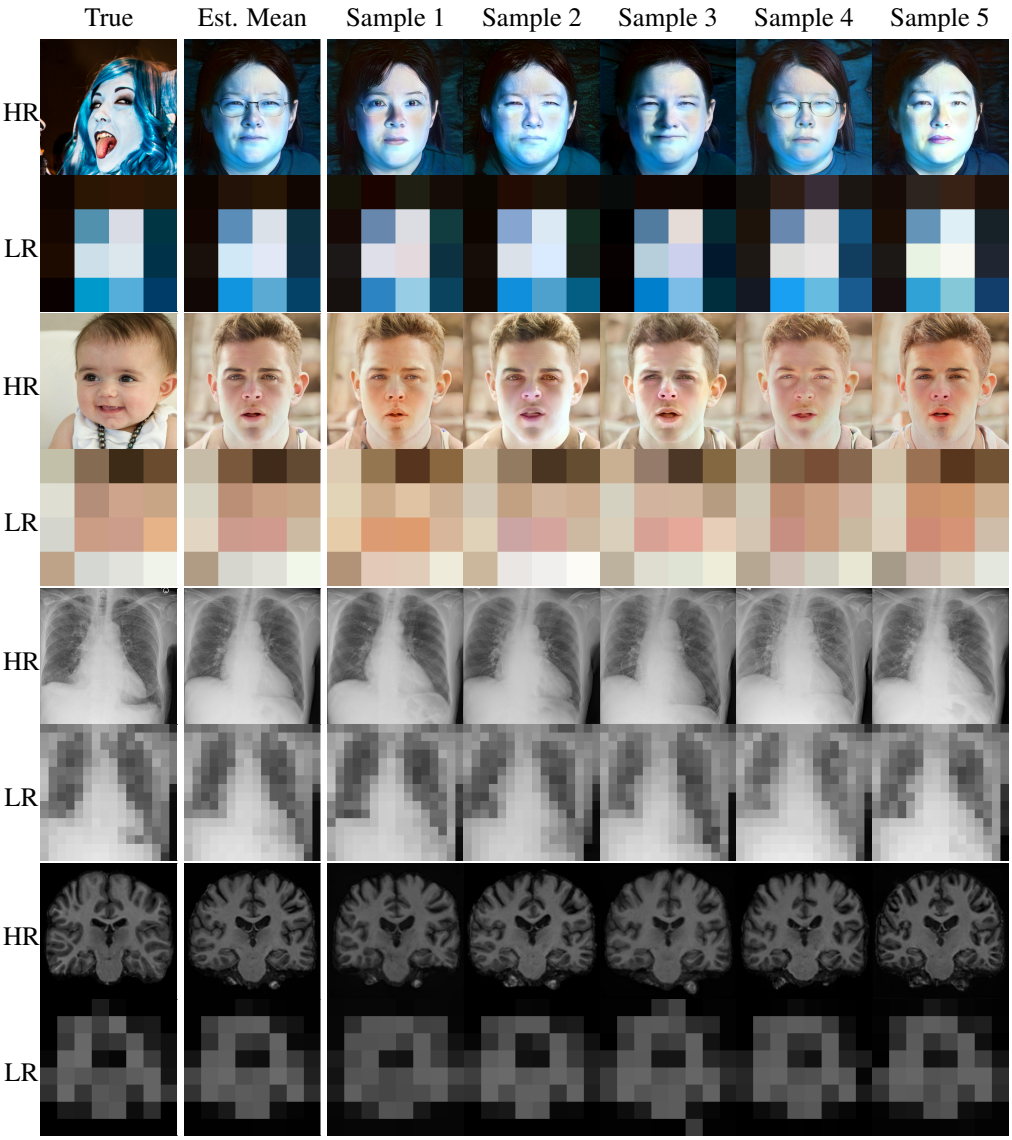


Figure 13: Sampling of multiple reconstructions using Variational Inference on super-resolution tasks with varying factors. From left, we show the true image, the estimated variational mean, alongside five random samples around that mean. For each high-resolution (HR) image, we show the corresponding 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.

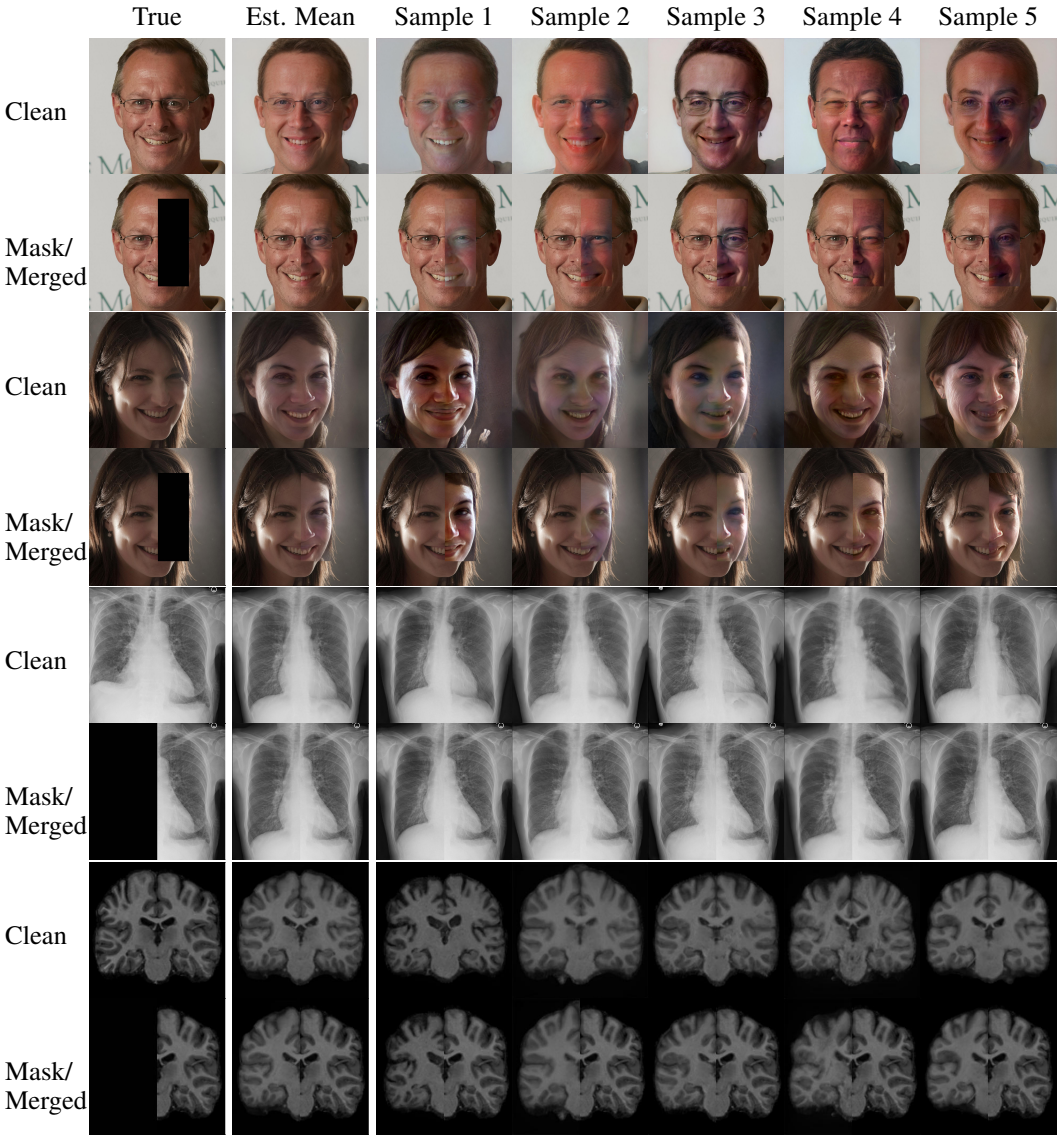


Figure 14: Sampling of multiple reconstructions using Variational Inference on in-painting tasks. From left, we show the true image, the estimated variational mean, alongside five random samples around that mean. For the mean and for each sample, we show both the clean image, as well as the true image with the in-painted area from the sample.

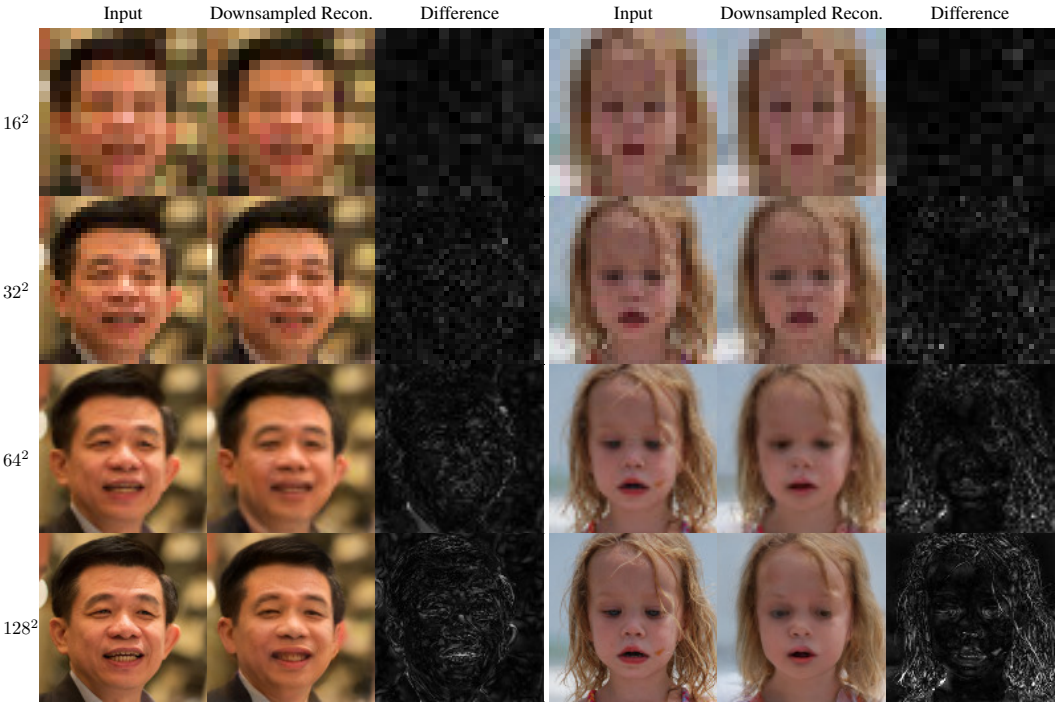


Figure 15: Inconsistency of our method on FFHQ, across different resolution levels, using uncurated example pictures. Left columns shows input images, and the middle columns show downsampled reconstructions (i.e. $f \circ G(w^*)$) where images were 4x super-resolved, then downsampled by 4x to match again 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 $\mathcal{L}_{cosine}$, 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.

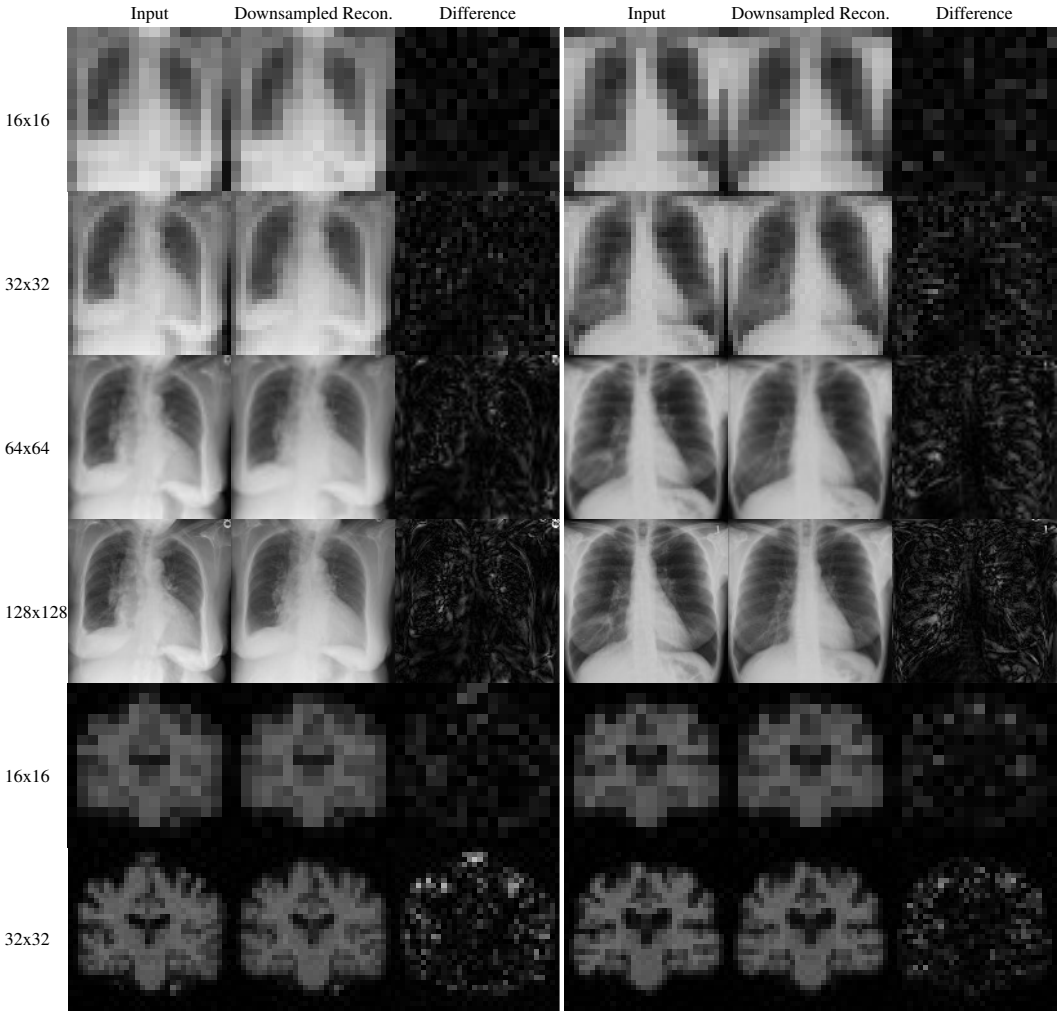


Figure 16: Inconsistency of our method on the medical datasets, using uncurated examples. Same setup as in Fig. 15.

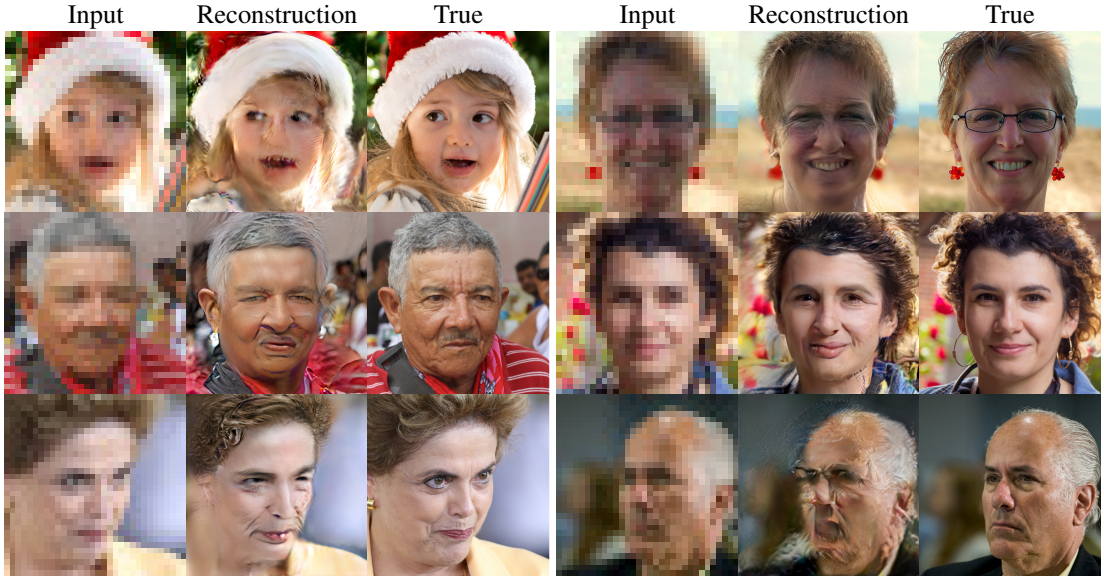


Figure 17: Failure cases of our method.