Variational Inference for Computational Imaging Inverse Problems

Francesco Tonolini   2402432T@STUDENT.GLA.AC.UK
School of Computing Science, University of Glasgow

Jack Radford   J.RADFORD.1@RESEARCH.GLA.AC.UK
School of Physics and Astronomy, University of Glasgow

Alex Turpin   ALEX.TURPIN@GLASGOW.AC.UK
School of Computing Science, University of Glasgow

Daniele Faccio   DANIELE.FACCIO@GLASGOW.AC.UK
School of Physics and Astronomy, University of Glasgow

Roderick Murray-Smith   RODERICK.MURRAY-SMITH@GLASGOW.AC.UK
School of Computing Science, University of Glasgow

Abstract

Machine learning methods for computational imaging require uncertainty estimation to be reliable in real settings. While Bayesian models offer a computationally tractable way of recovering uncertainty, they need large data volumes to be trained, which in imaging applications implicates prohibitively expensive collections with specific imaging instruments. This paper introduces a novel framework to train variational inference for inverse problems exploiting in combination few experimentally collected data, domain expertise and existing image data sets. In such a way, Bayesian machine learning models can solve imaging inverse problems with minimal data collection efforts. Extensive simulated experiments show the advantages of the proposed framework. The approach is then applied to two real experimental optics settings: holographic image reconstruction and imaging through highly scattering media. In both settings, state of the art reconstructions are achieved with little collection of training data.

Keywords: Inverse Problems, Approximate Inference, Bayesian Inference, Computational Imaging

1. Introduction

Computational imaging (CI) is one of the most important yet challenging forms of algorithmic information retrieval, with applications in Medicine, Biology, Astronomy and more. Bayesian machine learning methods are an attractive route to solve CI inverse problems, as they retain the advantages of learning from empirical data, while improving reliability by inferring uncertainty (Adler and Öktem, 2018; Zhang and Jin, 2019). However, fitting distributions with Bayesian models requires large sets of training examples (Zhang et al., 2018). This is particularly problematic in CI settings, where measurements are often unique to specific instruments, resulting in the necessity to carry out lengthy and expensive acquisition experiments or simulations to collect training data (Lucas et al., 2018; Lee et al., 2017).
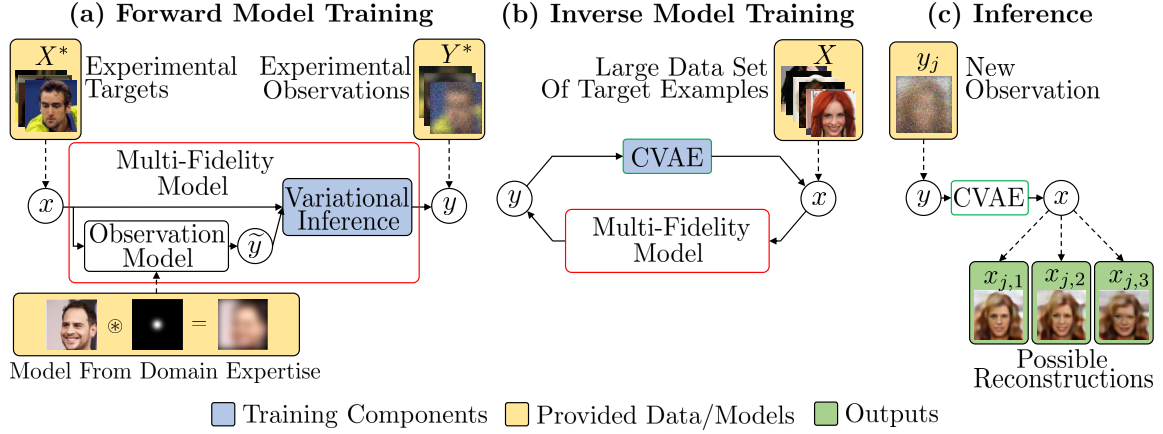
Figure 1: Proposed framework for training variational inference with diverse sources of information accessible in imaging settings. (a) Firstly, a multi-fidelity forward model is built to generate experimental observations. A variational model is trained to reproduce experimental observations Y^* from experimental ground truth targets X^*, exploiting simulated predictions $\widetilde{y}$ given by some analytical observation model defined with domain expertise. (b) A CVAE then learns to solve the inverse problem from a large data set of target examples X with a training loop; target examples x are passed through the previously learned multi-fidelity forward model to generate measurements y, which are then used as conditions for training the CVAE to generate back the targets x. In this way, a large number of ground truth targets can be exploited for learning, without the need for associated experimental measurements. (c) The trained CVAE can then be used to draw different possible solutions $x_{j,i}$ to the inverse problem conditioned on a new observation y_j.

Consider, for example, the task of reconstructing three dimensional environments from LIDAR measurements. Applying machine learning to this task requires data, in particular, paired examples of 3D environments and signals recorded with the particular LIDAR system to be employed. This means that, in principle, each LIDAR system being developed for this task requires its own extensive data set of paired examples to be collected, rendering the use of machine learning extremely impractical.

This paper introduces a novel framework to train conditional variational models for solving CI inverse problems leveraging in combination (i) a minimal amount of experimentally acquired or numerically simulated ground truth target-observation pairs, (ii) an inexpensive analytical model of the observation process from domain expertise and (iii) a large number of unobserved target examples, which can often be found in existing data sets. In such a way, trained variational inference models benefit from all accessible useful data as well as domain expertise, rather than relying solely on specifically collected training inputs and outputs. Recalling the LIDAR example given above, the proposed method would allow

the joint utilisation of limited collections with the specific LIDAR instrument, a physical model of LIDAR acquisition and any number of available examples of 3D environments to train our machine learning models. The framework is derived with Bayesian formulations, interpreting the different sources of information available as samples from or approximations to the underlying hidden distributions.

To address the expected scarcity of experimental data, the novel training strategy adopts a variational semi-supervision approach. Similarly to recent works in semi-supervised VAEs, an auxiliary model is employed to map abundantly available target ground-truths to corresponding measurements, which are, in contrast, scarce (Kingma et al., 2014; Maaløe et al., 2016; Nalisnick et al., 2019). However, the proposed framework introduces two important differences, specifically adapting to CI problems:

i This auxiliary function incorporates a physical observation model designed with domain expertise. In most imaging settings, the Physics of how targets map to corresponding measurements is well understood and described by closed form expressions. These should be used to improve the quality of a reconstruction system.

ii Instead of training the two models simultaneously, a forward model is trained first and then employed as a sampler in training the reconstruction model. This choice is made to avoid that synthetic measurements, i.e. predicted by the auxiliary system, contain more information about the targets than those encountered in reality. While this is not a critical problem for most semi-supervised systems, as auxiliary models often predict low-dimensional conditions such as labels, it very much arises in imaging settings, where these conditions are instead measurements that have comparable or even higher dimensionality than the targets. This high dimensionality of the conditions allows a system training the two models jointly to pass rich information through the synthetic measurements in order to maximise training reconstruction likelihood. By training the forward process separately instead, this auxiliary model is induced to maximise fidelity to real measurements alone, essentially providing an emulator.

Figure 1 schematically illustrate the framework and its components.

The proposed framework is first quantitatively evaluated in simulated image recovery experiments, making use of the benchmark data sets CelebA and CIFAR10 (Liu et al., 2015; Krizhevsky, 2009). In these experiments, different common transformations are applied to the images, including Gaussian blurring, partial occlusion and down-sampling. Image recovery is then performed with variational models. The novel training framework proved significantly advantageous across the range of tested conditions compared with standard training strategies. Secondly, the proposed technique is implemented with experimental imaging systems in two different settings; phase-less holographic image reconstruction (Shechtman et al., 2015; Sinha et al., 2017) and imaging through highly scattering media with temporally resolved sensors (Jiang, 2018; Lyons et al., 2019). In both settings, state of the art results are obtained with the proposed technique, while requiring minimal experimental collection efforts for training.

2. Background and Related Work

Computational imaging (CI) broadly identifies a class of methods in which an image of interest is not directly observed, but rather inferred algorithmically from one or more

measurements (Bertero and Boccacci, 1998). Many image recovery tasks fall within this definition, such as de-convolution (Starck et al., 2003; Bioucas-Dias et al., 2006), computed tomography (CT) (Ritschl et al., 2011) and structured illumination imaging (Magalhães et al., 2011). Traditionally, CI tasks are modelled as inverse problems, where a target signal $x \in \mathbb{R}^n$ is measured through a forward model $y = f(x)$, yielding observations $y \in \mathbb{R}^m$. The aim is then to retrieve the signal x from the observations y (Bertero and Boccacci, 1998; Vogel, 2002). In the following subsections, the main advances in solving imaging inverse problems are reviewed and the relevant background on Bayesian multi-fidelity models is covered.

2.1 Linear models and Hand-crafted Priors

In many CI problems, the forward observation model can be approximately described as a linear operator $A \in \mathbb{R}^{m \times n}$ and some independent noise $\epsilon \in \mathbb{R}^m$ (Bertero and Boccacci, 1998; Daubechies et al., 2004), such that the measurements y are assumed to be generated as

$$y = Ax + \epsilon. \tag{1}$$

The noise ϵ often follows simple statistics, such as Gaussian or Bernoulli distributions, depending on the instruments and imaging settings. This choice of forward model is computationally advantageous for retrieval algorithms, as it can be run efficiently through a simple linear projection, and is often a sufficiently good approximation to the "true" observation process. The difficulty in retrieving the observed signal x from an observation y in this context derives from the fact that in CI inverse problems the operator A is often poorly conditioned and consequentially the resulting inverse problem is ill-posed. Put differently, the inverse A^{-1} is not well-defined and small errors in y result in large errors in the naive estimation $x \simeq A^{-1}y$. To overcome this issue, the classical approach is to formulate certain prior assumptions about the nature of the target x that help in regularising its retrieval.

2.1.1 MAXIMUM A POSTERIORI INFERENCE

A widely adopted framework is that of maximum a posteriori (MAP) inference with analytically defined prior assumptions. The aim is to find a solution which satisfies the linear observations well, while imposing some properties which the target x is expected to retain. Under Gaussian noise assumptions, the estimate of x is recovered by solving a minimisation problem of the form

$$\arg\min_x \quad \frac{1}{2}||Ax - y||^2 + \lambda h(x), \tag{2}$$

where $||\cdot||$ indicates the Euclidean norm, λ is a real positive parameter that controls the weight given to the regularisation and $h(x)$ is an analytically defined penalty function that enforces some desired property in x. In the case of images, it is common to assume that x is sparse in some basis, such as frequency or wavelets, leading to ℓ_1-norm penalty functions (Daubechies et al., 2004; Donoho, 2006; Figueiredo et al., 2007). For such choices of penalty, and other common ones, the objective function of equation 2 is convex. This makes the optimisation problem solvable with a variety of efficient methods (Daubechies et al., 2004; Yang and Zhang, 2011) and provides theoretical guarantees on the recoverability of the solution (Candes et al., 2006).

The aforementioned framework has been widely applied to perform CI. For instance, many image restoration tasks, such as de-blurring, up-sampling and in-painting have been formulated as ill-conditioned linear inverse problems and are solved as described above (Osher et al., 2005). Various more complex sensing models can also be cast as linear operators, leading to the use of constrained optimisation in several CI systems that rely on ill-posed observations, such as sparse CT (Ritschl et al., 2011), single pixel photography (Sun et al., 2013) and imaging of objects hidden from view (Velten et al., 2012).

2.1.2 BAYESIAN INFERENCE

MAP inference aims at recovering a single optimal solution to a given inverse problem. While such retrieval is arguably useful in many settings, it is not a complete description of the solution space. For a given ill-posed inverse problem there may be many solutions that satisfy similarly well the observed measurements and the prior assumptions. To capture the variability of such solution spaces, hence implicitly estimating reconstruction errors, the inverse problem can be cast as a Bayesian inference task; given an observation likelihood $p(y|x)$ and a signal prior $p(x)$, the aim is to retrieve the posterior PDF of solutions $p(x|y) = p(y|x)p(x)/p(y)$. Estimating the full distribution of solutions $p(x|y)$ is generally much harder than simply finding its maximum through MAP inference.

Approximate inference for the aforementioned problem has been approached in different ways. A popular class of methods in settings of limited dimensionality is that of inference through Markov chain Monte Carlo (MCMC) processes, with different choices of conditional sampling having been proposed (Gamerman and Lopes, 2006; Ji et al., 2008; Martin et al., 2012; Marzouk and Xiu, 2009; Malinverno, 2002). Despite their guarantees of convergence to accurate approximations, MCMC methods are often prohibitively expensive for CI problems, as images are rather high dimensional. A second class of approaches is that of variational inference. These methods aim to use a tractable parametric PDF to approximate the true posterior $p(x|y)$ (Mohammad-Djafari, 2013; Tsilifis et al., 2016). Though they do not provide the same guarantees as MCMC methods, these approaches are typically more efficient and have been explored with different PDFs and optimisation techniques in the context of CI (Likas and Galatsanos, 2004; Babacan et al., 2011, 2009).

2.2 Machine Learning for Computational Imaging

The increasing availability of data sets and continuous advancements in learning inference models enabled new possibilities for CI and inverse problems in general. Learning from real examples allows to derive retrieval models directly from empirical experience, instead of relying on analytically defined priors and observation processes. The main classes of machine learning methods for CI are reviewed below.

2.2.1 LEARNING INVERSE MAPPINGS

Most learning approaches can arguably be described as inverse mapping models; with enough example pairs available, a neural network can be trained to directly recover a signal x from observations y (Lucas et al., 2018). Many neural architectures of this type have been developed to perform different image reconstruction tasks. In particular, convolutional neural networks are popular choices due to their ability to capture local pixel dependencies in

images (Egmont-Petersen et al., 2002; McCann et al., 2017). These models are trained solely with paired examples of observed targets X^* as outputs and corresponding observations Y^* as inputs. These target-observation pairs are collected from either experimental acquisitions or numerical simulations. Once the model is trained, a new empirical observation y_j can be mapped to the corresponding target reconstruction estimate x_j (Lucas et al., 2018).

Directly learning inverse mappings retains a number of advantages compared to analytical methods. First, the model is trained with a set of images the target solution is assumed to belong to, implicitly making the signal assumptions more specific than, for example, sparsity in some basis. Second, the observation model $f(x)$ is not constrained to be linear, or even differentiable; so long as a large number of signal-observations pairs is available the network can be trained to perform the inversion. Third, once the model is trained, inference is non-iterative and thus typically much faster, allowing elaborate imaging systems to retrieve reconstructions in real time and even at video rate.

State of the art performance has been demonstrated with specifically designed neural networks models in many common image processing tasks, such as deconvolution and super-resolution (Xu et al., 2014; Ledig et al., 2017), as well as signal recovery from under determined projective measurements (Kulkarni et al., 2016). However, learned inverse mappings for CI retain two main problems. The first is that their accuracy of inference is entirely dependent on the available training targets and observations, leading to the need of carrying out lengthy data collections or numerical simulations to ensure robustness. The second is that it is difficult to asses the reliability of a given reconstruction; the trained neural network returns a deterministic estimate of the target that is usually in the range of the training examples, making it difficult to recognise unsuccessful recovery.

2.2.2 Iterative Inference with Learned Priors

A second class of learning methods that is conceptually closer to analytical techniques is that of MAP inference with learned prior knowledge. The general idea is to exploit a differentiable analytic observation model to maximise the agreement with recorded observations, as in traditional MAP inference, but build the regularising prior empirically, learning from examples of expected signals (Bora et al., 2017). The prior assumptions can be captured and induced in different ways. One option is to train a function $H(x)$ to quantify how much a target x is expected to belong to a given set of training examples. The solution is then found by iteratively solving the minimisation problem

$$\underset{x}{\arg\min} \quad \frac{1}{2}||Ax - y||^2 + \lambda H(x), \tag{3}$$

where A is a linear operator describing the observation process. Different choices of function $H(x)$ have been explored in recent works. One such choice is to train a discriminator $D(x)$ to recognise targets which belong to the training class and then setting $H(x) = s(\log(D(x)))$, where $s(\cdot)$ is a Sigmoid function (Chang et al., 2017). One other popular choice is to train a de-noising function $N(x)$ on the set of expected targets and then use the distance between a target and its de-noised equivalent $H(x) = ||x - N(x)||$ (Chan et al., 2017; Zhang et al., 2017; Aggarwal et al., 2019). Machine learning has also been implemented to train optimisation methods to solve the minimisation of equation 3. In fact, the iterative update of the solution x through the optimisation procedure in these settings is often interpreted as a recurrent

neural network (Chen et al., 2015; Putzky and Welling, 2017). In such a way, the iterative inference precision is empirically adjusted to the specific inversion task, hence gaining in efficiency and accuracy (Adler and Öktem, 2017).

A second framework to infer learned properties in iterative MAP inference is that of constrained minimisation with generative models. In these techniques, a generative model, such as a generative adversarial network (GAN) or a variational auto-encoder (VAE), is trained with a data set of expected targets, resulting in a generator $G(z)$ that can synthesise artificial examples x in the range of interest from low-dimensional latent noise variables z. The solution target is then assumed to be synthesised by such generator, resulting in the following minimisation

$$\arg \min_{z} \quad \frac{1}{2}||A \cdot G(z) - y||^2. \tag{4}$$

In such a way, the solution is constrained to be within the domain of the generative model, as the recovered x is by definition generated from z, while at the same time agreement to the measurements is induced by minimising the distance to the observations y. Iterative inference with generative models has been demonstrated for linear observation processes and phase-less linear observation processes (Bora et al., 2017; Mardani et al., 2019; Hand et al., 2018).

MAP inference with learned prior methods do eliminate the problem of data collection, as training is performed using solely examples of targets, while the nature of observations is incorporated through an analytically defined model (Chang et al., 2017). However, compared to learning inverse mappings, it comes with significant drawbacks. Firstly, the target-observations relationship is described entirely by an analytical model, sacrificing the desirable ability of machine learning to generalise mappings from empirical evidence. Secondly, these methods infer a solution to an inverse problem iteratively, excluding real time reconstruction applications. Furthermore, like learned inverse mappings, the solutions returned are deterministic, hence making it difficult to assess the reliability of a reconstruction.

2.2.3 CONDITIONAL GENERATIVE MODELS

A promising direction to overcome the reliability problem is that of conditional generative models; instead of learning a deterministic mapping from observations to single reconstructions, a generative model is trained to generate different targets conditioned on given observations. The generation of multiple solutions from the same measurements can be probabilistically interpreted as sampling from the recovered posterior distribution. From these samples, uncertainty metrics, such as mean and standard deviation, can be inferred. The recovered uncertainty can then be used to asses the reliability of a particular reconstruction or be propagated to automated decisions. Recent advances in variational methods and adversarial models allow to train efficiently approximate inference through generative models that scale to the dimensionalities and numbers of examples typically needed for imaging tasks (Kingma and Welling, 2014; Gulrajani et al., 2016; Mirza and Osindero, 2014). Building upon these advancements, different conditional generative models have been developed in recent years, with the most commonly adopted being conditional Generative Adversarial

Networks (CGANs) and conditional variational auto-encoders (CVAEs) (Isola et al., 2017; Sohn et al., 2015; Nguyen et al., 2017).

Conditional generative models have been applied to perform a range of inference tasks, such as classification (Sohn et al., 2015), generation conditioned on classes (Nguyen et al., 2017; Odena et al., 2017), image-from-text inference (Reed et al., 2016; Yan et al., 2016) and missing value imputation (Pathak et al., 2016; Nazabal et al.). However, within CI, they have been largely restricted to inference from simple deterministic observation, such as missing pixels or down-sampling (Nguyen et al., 2017; Parmar et al., 2018), with the exception of recent work by Adler and Öktem (2018), in which a specifically designed GAN model is used to retrieve medical images from CT scans. In fact, the direct application of conditional generative models in CI is challenging because of the large data volumes requirements. Conditional generative models, in their common form, need a large number of object-condition pairs to train upon. In CI settings, this translates to the need of obtaining a large number of sufficiently accurate target-observation examples, which are unique to imaging instruments and hence expensive to collect.

2.2.4 Semi-Supervised Conditional Generative Models

Another closely related extension of generative models is that of semi-supervised learning with generative models. Similarly to conditional generative models, these methods introduce conditions on their generations, but are able to train with data sets where conditions are only available for a portion of the examples (Kingma et al., 2014; Maaløe et al., 2016; Nalisnick et al., 2019). They achieve this by introducing auxiliary models that map inputs to conditions and are trained jointly with the generator.

In some sense, the presented framework belongs to this class of methods, as the forward model component plays an analogous role to the auxiliary model in these systems, but retains two main distinctions:

i The forward model is built as a multi-fidelity process that incorporates physical observation models and therefore exploits domain expertise to infer conditions (measurements in our setting).

ii Instead of training the auxiliary model and the conditional generator jointly, the proposed method applies a two step procedure, as shown in figure 1.

The reason for the latter difference is that one needs to ensure synthetic measurements do not contain more information about the targets than the real ones. In the imaging setting, training the forward and reconstruction models jointly encourages the former to "hide" additional information about the targets in the generated observations that real ones do not possess. In common semi-supervised settings, this is generally not a problem; the conditions, or auxiliary information, tend to be very low dimensional, e.g. labels, and the arising information bottleneck naturally prevents this effect. This is not true in imaging settings, where recorded measurements easily have more dimensions than the targets.

2.3 Multi-Fidelity Bayesian Models

Multi-fidelity methods exploit both highly accurate but expensive data and less accurate but cheaper data to maximise the accuracy of model estimation while minimising computational cost (Peherstorfer et al., 2018). In multi-fidelity Bayesian inference, the most accurate

predictions, or high-fidelity outputs, are considered to be draws from the underlying true density of interest and the aim is to approximately recover such high-fidelity outputs from the corresponding inputs and low-fidelity outputs of some cheaper computation (Kaipio and Somersalo, 2007). Within Bayesian approaches to solve inverse problems, multi-fidelity models have been used to minimise the cost of estimating expensive forward processes, in particular with MCMC methods to efficiently estimate the likelihood at each sampling step (Christen and Fox, 2005).

In Bayesian optimisation settings, the difference between high and low fidelity predictions is commonly modeled with Gaussian processes, where approximate function evaluations are made cheap by computing low-fidelity estimates and subsequently mapping them to high-fidelity estimates with a Gaussian process (Kennedy and O'Hagan, 2001; Bayarri et al., 2007). In CI settings, Gaussian process multi-fidelity models are difficult to apply, as the available volume of data and the dimensionality of the targets and observations are potentially very large. Recent work by Yang and Perdikaris (2019) proposes to model high-fidelity data with conditional deep generative models, which are instead capable of scaling to the volumes and dimensionalities needed in imaging applications. The multi-fidelity component of the framework presented here follows these ideas and exploits a deep CVAE to model high fidelity data when inferring an accurate forward observation process.

3. Variational Framework for Imaging Inverse Problems

Differently from previous approaches, the proposed variational framework is built to learn from all the useful data and models typically available in CI problems. In the following subsections the problem of Bayesian learning in this context is defined and the components of the proposed variational learning method are derived and motivated.

3.1 Problem Description

3.1.1 THE BAYESIAN INVERSE PROBLEM

The aim of computational imaging is to recover a hidden target $x_j \in \mathbb{R}^N$ from some associated observed measurements $y_j \in \mathbb{R}^M$. In the Bayesian formulation, the measurements y_j are assumed to be drawn from an observation distribution $p(y|x_j)$ and the objective is to determine the posterior $p(x|y_j)$; the distribution of all possible reconstructions. Following Bayes' rule, the form of this posterior is

$$p(x|y_j) = \frac{p(y_j|x)p(x)}{p(y_j)}. \tag{5}$$

The observation distribution $p(y|x)$, often referred to as the data likelihood, describes the observation process, mapping targets to measurements. Given any ground truth target x_i the corresponding measurements that are physically recorded y_i are draws from the data likelihood $y_i \sim p(y|x_i)$. The prior distribution $p(x)$ models the assumed knowledge about the targets of interest. This PDF is the distribution of possible targets prior to carrying out any measurement. Finally, the marginal likelihood $p(y) = \int p(x)p(y|x)dx$ is the distribution of all possible measurements y. The goal of variational inference is to learn a non-iterative approximation to the true intractable posterior distribution of equation 5 for

arbitrary new observations y_j. That is, learning a parametric distribution $r_\theta(x|y)$ which well approximates the true posterior $p(x|y)$ for any new observation $y_j \sim p(y)$ and from which one can non-iteratively draw possible reconstructions $x_{j,i} \sim r_\theta(x|y_j)$.

3.1.2 Information Available

For a given imaging inverse problem of the type described above, there are generally three main sources of information that can be exploited to obtain the best estimate of the target posterior. The first is empirical observations. Physical experiments to collect sets of ground-truth targets $X^* \in \mathbb{R}^{N \times K}$ and associated observations $Y^* \in \mathbb{R}^{M \times K}$ can be recorded with the imaging apparatus of interest. The number of these acquisitions K is normally limited by the time and effort necessary for experimental preparation and collection, or alternatively by computational cost, if these are obtained through numerical simulations. However, empirical target-observation pairs are the most accurate evaluation of the true observation process and therefore can be very informative. A recorded observation $y_k \in Y^*$ obtained when imaging a target $x_k \in X^*$ can be interpreted as a sample from the true data likelihood $y_k \sim p(y|x_k)$.

The second source of information is domain expertise. The measurement process, mapping targets to observations, is described by a physical phenomenon. With knowledge of such phenomenon, one can construct a functional mapping, normally referred to as forward model, which computes observations' estimates $\widetilde{y}$ from targets x. For instance, many observation processes in CI settings can be approximately modelled by a linear transformation and independent Gaussian or Poisson noise (Daubechies et al., 2004). It is clearly infeasible to obtain analytical models that perfectly match reality. However, an analytical forward model can provide inexpensive approximations $\widetilde{y}$ to the true observations y that can be computed for any target x. In the Bayesian formulation, a forward model can be interpreted as a closed form approximation $p(\widetilde{y}|x)$ to the true data likelihood $p(y|x)$.

Lastly, many examples of the targets of interest $X \in \mathbb{R}^{N \times L}$ are often available in the form of unlabelled data sets. Because collection of this type of data is independent of the imaging apparatus, the number of available examples L is expected to be much greater than the number of empirical acquisitions K. In fact, many large image data sets containing relevant targets for CI applications are readily available and easily accessible. These target examples $x_l \in X$ can be interpreted as draws from the prior distribution $x_l \sim p(x)$. In summary, the available sources of information are

- Limited sets of ground-truth targets $X^* = \{x_{k=1:K}\}$ and associated observations $Y^* = \{y_{k=1:K}\}$, the elements of which are point samples of the true data likelihood $y_k \sim p(y|x_k)$.

- An analytical forward model providing a closed form approximation for the true data likelihood $p(\widetilde{y}|x) \approx p(y|x)$.

- A large set of target examples $X = \{x_{l=1:L}\}$ corresponding to prior samples $x_l \sim p(x)$, where $L \gg K$.

The scope of this paper is to design a framework for learning the best possible approximate distribution $r_\theta(x|y)$ by exploiting all the available sources of information described above.

3.2 Multi-Fidelity Forward Modelling

Before training the inversion, an approximate observation distribution $p_\alpha(y|x)$ is trained to fit the true data likelihood $p(y|x)$. Learning this observation distribution first allows effective incorporation of domain expertise, as in CI problems this is usually available in the form of an analytical forward model. Furthermore, training the observation model is expected to require far fewer training input-output pairs than training the corresponding inversion, as most forward models are well-posed, while the corresponding inverse problems are often ill-posed. A good approximation to the data likelihood $p_\alpha(y|x)$ can therefore be learned with a much lower number K of experimental ground-truth targets X^* and measurements Y^* than would be required to train a good approximate posterior $r_\theta(x|y)$ directly.

In order to make use of the analytical approximation $p(\widetilde{y}|x)$, hence incorporating domain expertise in the training procedure, the approximate observation distribution is chosen as

$$p_\alpha(y|x) = \int p(\widetilde{y}|x)p_\alpha(y|x,\widetilde{y})d\widetilde{y}. \tag{6}$$

In such a way, the inference of a measurement y, a high-fidelity prediction, from a target x can exploit the output $\widetilde{y}$ of the analytical forward model $p(\widetilde{y}|x)$, which instead is considered a low-fidelity prediction. The parametric component to be trained is then the conditional $p_\alpha(y|x,\widetilde{y})$, which returns high-fidelity sample measurements y from targets x and low-fidelity predictions $\widetilde{y}$.

To provide flexible inference in the general case, the PDF $p_\alpha(y|x,\widetilde{y})$ is chosen to be a latent variable model of the form

$$p_\alpha(y|x,\widetilde{y}) = \int p_{\alpha_1}(w|x,\widetilde{y})p_{\alpha_2}(y|x,\widetilde{y},w)dw. \tag{7}$$

The two parametric distributions $p_{\alpha_1}(w|x,\widetilde{y})$ and $p_{\alpha_2}(y|x,\widetilde{y},w)$ are chosen to be Gaussian distributions, the moments of which are outputs of neural networks with weights α_1 and α_2 respectively.[1] The model of equation 6 is then trained to fit the sets of experimental ground-truth targets and measurements X^* and Y^*, as these are point samples of the true data likelihood of interest $y_k \sim p(y|x_k)$. The optimisation to be performed is the log likelihood maximisation

$$\underset{\alpha_1,\alpha_2}{\arg\max} \quad \log p_\alpha(Y^*|X^*) = \sum_{k=1}^{K} \log \int p(\widetilde{y}|x_k) \int p_{\alpha_1}(w|x_k,\widetilde{y})p_{\alpha_2}(y_k|x_k,\widetilde{y},w)dwd\widetilde{y}. \tag{8}$$

Due to the integration over latent variables w, the maximisation of equation 8 is intractable to directly perform stochastically. However, problems of this type can be approximately solved efficiently with a variational auto-encoding approach, in which a parametric recognition model is used as a sampling function (Kingma and Welling, 2014; Sohn et al., 2015). The VAE formulation for the multi-fidelity model is presented in detail in supplementary section A.1. Through this approach, training of the parameters $\alpha = \{\alpha_1,\alpha_2\}$ and β can be performed through the following stochastic optimisation:

$$\underset{\alpha_1,\alpha_2,\beta}{\arg\max} \quad \sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{s=1}^{S}\log p_{\alpha_2}(y_k|x_k,\widetilde{y}_v,w_s) - D_{KL}(q_\beta(w|x_k,y_k,\widetilde{y}_v)||p_{\alpha_1}(w|x_k,\widetilde{y}_v))\right]. \tag{9}$$

---

1. The distribution $p_{\alpha_2}(y|x,\widetilde{y},w)$ can alternatively be chosen to match some other noise model if the observation noise is known to be of a particular type, such as Poisson or Bernoulli.
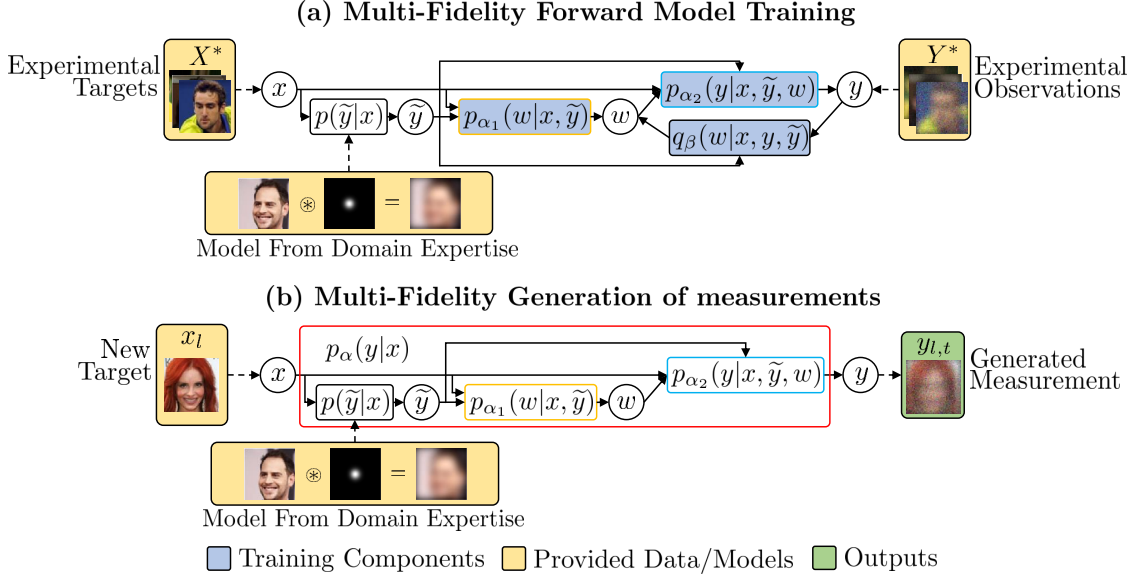
Figure 2: Multi-fidelity forward modelling. (a) The two conditional distributions $p_{\alpha_1}(w|x,\widetilde{y})$ and $p_{\alpha_2}(y|x,\widetilde{y},w)$, parametric components of the multi-fidelity forward model $p_\alpha(y|x)$, are trained with an auto-encoding approach, making use of a recognition model $q_\beta(w|x,y,\widetilde{y})$. These distributions are trained with an analytical forward model defining $p(\widetilde{y}|x)$, experimental ground-truth targets X^* and corresponding observations Y^*. (b) once the parameters $\alpha = \{\alpha_1, \alpha_2\}$ have been trained, the learned distributions can be used to generate multi-fidelity estimates of observations $y_{l,t}$ from a new target x_l. First, a low fidelity estimate $\widetilde{y}_v$ is generated through the analytical observation model $p(\widetilde{y}|x_l)$. Second, this estimate and the corresponding target are used to draw a latent variable from $p_{\alpha_1}(w_s|x_l,\widetilde{y}_v)$. Third, the target x_l, low-fidelity estimate $\widetilde{y}_v$ and latent variable w_s are used to generate a high-fidelity observation's estimate $y_{l,t}$ by sampling from $p_{\alpha_2}(y|x_l,\widetilde{y}_v,w_s)$. Performing these operations in sequence corresponds to running the multi-fidelity forward model $y_{l,t} \sim p_\alpha(y|x)$.

This bound is maximised stochastically by drawing samples from the approximate distribution $p(\widetilde{y}|x)$ and subsequently from a recognition model $q_\beta(w|x_k,y_k,\widetilde{y})$, which is chosen as an isotropic Gaussian distribution, the moments of which are outputs of a neural network taking as inputs targets x, high-fidelity measurements y and low-fidelity measurements $\widetilde{y}$.

Sampling from the approximate likelihood $\widetilde{y}_v \sim p(\widetilde{y}|x_k)$ is equivalent to running the analytical forward observation model. For instance, in the case of a linear observation model, the samples $\widetilde{y}_v$ are computed as $\widetilde{y}_v = Ax_k + \epsilon_v$, where A is the linear mapping given by the model and ϵ_v is drawn from the noise process characteristic of the apparatus of interest. Pseudo-code for the multi-fidelity forward model training is in supplementary section B.

Once the weights α have been trained through the maximisation of equation 9, it is possible to inexpensively compute draws $y_{l,t}$ from the multi-fidelity data likelihood estimate $p_\alpha(y|x_l)$ given a new target x_l as

$$y_{l,t} \sim p_{\alpha_2}(y|x_l, \widetilde{y}_v, w_s), \quad \text{where} \quad \widetilde{y}_v \sim p(\widetilde{y}|x_l) \quad \text{and} \quad w_s \sim p_{\alpha_1}(w|x_l, \widetilde{y}_v). \tag{10}$$

Computing a forward model estimate with the trained multi-fidelity likelihood consists of three consecutive computations. First, a low-fidelity estimate $\widetilde{y}_v$ is computed by running the analytical forward model. Second, a latent variable w_s is drawn from the latent distribution $p_{\alpha_1}(w|x_l, \widetilde{y}_v)$. Lastly, the high-fidelity measurement estimate $y_{l,t}$ is drawn from the conditional $p_{\alpha_2}(y|x_l, \widetilde{y}_v, w_s)$. As all of these operations are computationally inexpensive, running the resulting multi-fidelity forward model is also inexpensive.

3.3 Variational Inverse Model

To learn an inversion model, the approximate posterior distribution $r_\theta(x|y)$ is trained to recover targets from observations, exploiting the learned PDF $p_\alpha(y|x)$ to generate measurements from the large data set of target examples X. In such a way, training of the approximate posterior $r_\theta(x|y)$ can exploit the large number $L \gg K$ of target examples X, even though no corresponding measurements are available, as estimates of these are generated implicitly during training through the learned forward model $p_\alpha(y|x)$. sampling synthetic measurements from $p_\alpha(y|x)$ also introduces variation in the training inputs to $r_\theta(x|y)$, improving generalisation in a similar way to noise injection strategies (Matsuoka, 1992).

The aim of this training stage is to train a parametric distribution $r_\theta(x|y)$ to match the true posterior $p(x|y)$. To this end, the expectation of the cross entropy $H[p(x|y), r_\theta(x|y)]$ under the measurements' distribution $p(y)$ is minimised with respect to the model's variational parameters θ,

$$\arg\min_\theta \; \mathbb{E}_{p(y)} H[p(x|y), r_\theta(x|y)] = \arg\max_\theta \; \mathbb{E}_{p(y)} \int p(x|y) \log r_\theta(x|y) dx. \tag{11}$$

The optimisation of Equation 11 is equivalent to fitting $r_\theta(x|y)$ to the true posterior $p(x|y)$ over the distribution of measurements that are expected to be observed $p(y)$. This objective function can be simplified to give

$$\begin{aligned} \mathbb{E}_{p(y)} \int p(x|y) \log r_\theta(x|y) dx &= \iint p(y) \frac{p(y|x)p(x)}{p(y)} \log r_\theta(x|y) dx dy \\ &= \int p(x) \int p(y|x) \log r_\theta(x|y) dy dx. \end{aligned} \tag{12}$$

In order to stochastically estimate and maximise the expression of equation 12, drawing samples from the prior $x_l \sim p(x)$ and from the likelihood $y_{l,t} \sim p(y|x_l)$ needs to be realizable and inexpensive. In the case of the former, a large ensemble of samples is readily available from the data set of target examples X. Therefore, to approximately sample from the prior one only needs to sample from this data set. On the other hand, sampling from the likelihood $p(y|x_l)$ is not possible, as the form of the true forward observation model is not accessible. However, the previously learned multi-fidelity forward model $p_\alpha(y|x)$, described

in subsection 3.2, offers a learned approximation to the data likelihood from which it is inexpensive to draw realisations. The objective of equation 12 to be maximised can then be approximated as

$$\mathbb{E}_{p(y)} \int p(x|y) \log r_\theta(x|y) dx \simeq \int p(x) \int p_\alpha(y|x) \log r_\theta(x|y) dy dx. \tag{13}$$

In this form, stochastic estimation is inexpensive, as prior samples $x_l \sim p(x)$ can be drawn from the data set X and draws from the approximate likelihood $y_{l,t} \sim p_\alpha(y|x_l)$ can be computed by running the multi-fidelity forward model as described in subsection 3.2.

3.3.1 CVAE as Approximate Posterior

The approximate distribution $r_\theta(x|y)$ needs to be of considerable capacity in order to accurately capture the variability of solution spaces in imaging inverse problems. To this end, the approximating distribution $r_\theta(x|y)$ is chosen as a conditional latent variable model

$$r_\theta(x|y) = \int r_{\theta_1}(z|y) r_{\theta_2}(x|z, y) dz. \tag{14}$$

The latent distribution $r_{\theta_1}(z|y)$ is an isotropic Gaussian distribution $\mathcal{N}(z; \mu_z, \sigma_z^2)$, where its moments μ_z and σ_z^2 are inferred from a measurement y by a neural network. The neural networks may be convolutional or fully connected, depending on the nature of the observed signal from which images need to be reconstructed. The likelihood distribution $r_{\theta_2}(x|z, y)$ can take different forms, depending on the nature of the images to be recovered and requirements on the efficiency of training and reconstruction. In the experiments presented here, the distribution $r_{\theta_2}(x|z, y)$ was set to either an isotropic Gaussian with moments determined by a fully connected neural network, taking concatenated z and y as input, or a convolutional pixel conditional model analogous to that of a pixelVAE (Gulrajani et al., 2016).

Latent variable models of this type have been proven to be powerful conditional image generators (Sohn et al., 2015; Nguyen et al., 2017) and therefore are expected to be suitable variational approximators for posteriors in imaging problems. With this choice of approximate posterior $r_\theta(x|y)$, the objective function for model training is

$$\arg\max_{\theta_1, \theta_2} \int p(x) \int p_\alpha(y|x) \log \int r_{\theta_1}(z|y) r_{\theta_2}(x|z, y) dz dy dx. \tag{15}$$

As for the likelihood structure in the multi-fidelity forward modelling, directly performing the maximisation of equation 15 is intractable due to the integral over the latent space variables z. However, using Jensen's inequality, a tractable lower bound for this expression can be derived with the aid of a parametric recognition model $q_\phi(z|x, y)$.

As for the forward multi-fidelity model, the recognition model $q_\phi(z|x, y)$ is an isotropic Gaussian distribution in the latent space, with moments inferred by a neural network, taking as input both example targets x and corresponding observations y. This neural network may be fully connected, partly convolutional or completely convolutional, depending on the nature of the targets x and observations y. The VAE formulation for the Variational inverse problem is presented in detail in supplementary section A.2. Making use of this lower bound,
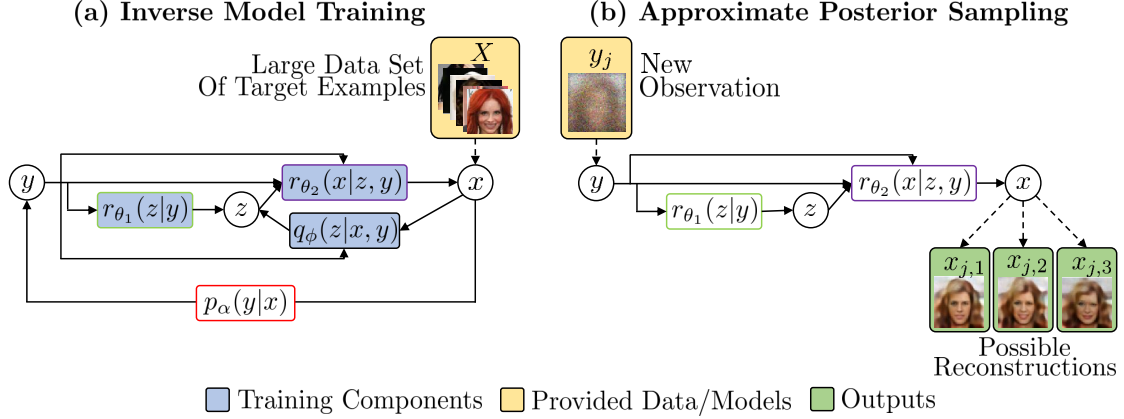
Figure 3: Variational inverse model. (a) The model is trained to maximise the evidence lower bound on the likelihood of targets x conditioned on observations y. The posterior components $r_{\theta_1}(z|y)$ and $r_{\theta_2}(x|z,y)$ are trained along with the auxiliary recognition model $q_\phi(z|x,y)$. Instead of training on paired targets and conditions, as for standard CVAEs, the model is given target examples X alone and generates training conditions y stochastically through the previously learned multi-fidelity forward model $p_\alpha(y|x)$. (b) Given new observations y_j, samples from the approximate posterior $r_\theta(x|y_j)$ can be non-iteratively generated with the trained model by first drawing a latent variable $z_{j,i} \sim r_{\theta_1}(z|y_j)$ and subsequently generating a target $x_{j,i} \sim r_{\theta_2}(x|z_{j,i}, y_j)$.

we can define the objective function for the inverse model as

$$\underset{\theta_1, \theta_2, \phi}{\arg\max} \; \sum_{l=1}^{L} \sum_{t=1}^{T} \left[\sum_{s=1}^{S} \log r_{\theta_2}(x_l | z_s, y_{l,t}) - D_{KL}(q_\phi(z|x_l, y_{l,t}) || r_{\theta_1}(z|y_{l,t})) \right], \qquad (16)$$

where target examples are drawn from the large data set as $x_l \sim X$, measurements are generated with the multi-fidelity model as $y_{l,t} \sim p_\alpha(y|x_l)$ and latent variables are drawn from the recognition model as $z_s \sim q_\phi(z|x_l, y_{l,t})$, using the reparametrisation trick presented in Kingma and Welling (2014). The variational approximate posterior $r_\theta(x|y)$ is trained by performing the maximisation of equation 16 through steepest ascent. The training procedure is schematically shown in Figure 3(a) and detailed as a pseudo-code in supplementary B. The models employed during training of the multi-fidelity forward model and the variational inverse model are both summarised in the graphical models of figure 4.

3.3.2 INFERENCE

Once the variational parameters $\theta = \{\theta_1, \theta_2\}$ have been trained, the learned approximate posterior can be used to generate draws $x_{j,i} \sim r_\theta(x|y_j)$ conditioned on new measurements y_j. Draws from the posterior are obtained by first drawing a latent variable $z_{j,i} \sim r_{\theta_1}(z|y_j)$ and subsequently generating a target $x_{j,i} \sim r_{\theta_2}(x|z_{j,i}, y_j)$. Such generated samples can be

Multi-Fidelity Forward Model

Generative Model   Recognition Model

Variational Inverse Model
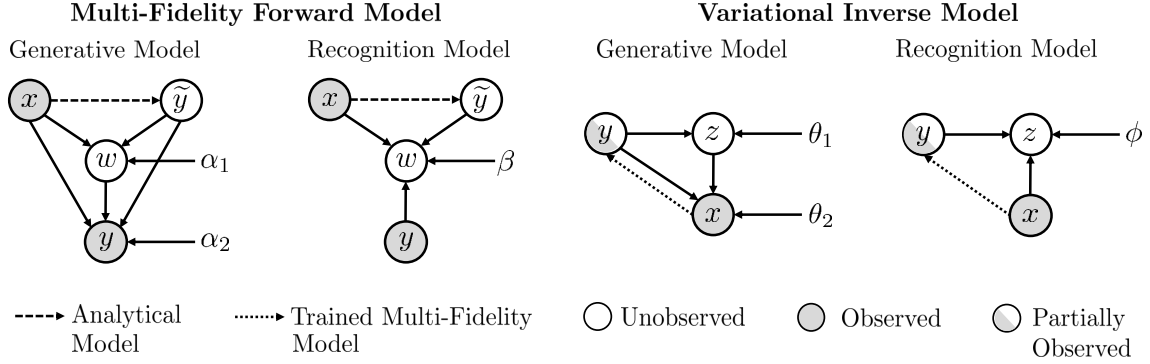
Generative Model   Recognition Model

Figure 4: Graphical models for training of the multi-fidelity forward model and the variational inverse model.

interpreted as different possible solutions to the inverse problem and can be used in different ways to extract information of interest. For instance, one can compute per-pixel marginal means and standard deviations, in order to visualise the expected mean values and marginal uncertainty on the retrieved images. Figure 3(b) schematically illustrates the approximate posterior sampling procedure.

It may be of interest to also estimate a single best retrieval x_j^* given the observed measurements y_j, which would be the image yielding the highest likelihood $r_\theta(x_j^*|y_j)$. This retrieval can be performed iteratively, by maximising $r_\theta(x|y_j)$ with respect to x, as proposed by Sohn et al. (2015). As the focus of this work is non-iterative inference, a pseudo-maximum non-iterative retrieval is instead used. Such retrieval is performed by considering the point of maximum likelihood of the conditional Gaussian distribution in the latent space $r_{\theta_1}(z|y_j)$, which is by definition its mean $\mu_{z,j}$. The pseudo-maximum reconstruction x_j^* is then the point of maximum likelihood of $r_{\theta_2}(x|\mu_{z,j}, y_j)$, which is also its mean $\mu_{x,j}$. This pseudo-maximum estimate adds the ability to retrieve an inexpensive near-optimal reconstruction, analogous to that recovered by deterministic mappings.

4. Experiments

The proposed framework is tested both in simulation and with real imaging systems. Firstly, quantitative simulated experiments are performed to compare the proposed training framework to other strategies using the same Bayesian networks. Secondly, the new approach is applied to phase-less holographic image reconstruction and imaging through highly scattering media, comparing reconstructions to the most recent state of the art in the respective fields.

4.1 Simulated Experiments

Simple image restoration tasks are performed in simulation. Experiments include Gaussian blurring, down-sampling and partial occlusion with both CelebA and CIFAR examples (Liu et al., 2015; Krizhevsky, 2009). The test images are corrupted with the given transformation and additive Gaussian noise. Variational models are then used to perform reconstructions,

16

(a) Standard CVAE, Trained with 3,000 Paired Examples

(b) Proposed, Trained with 3,000 Paired Examples + Inaccurate Observation Model + 100,000 Unobserved Targets
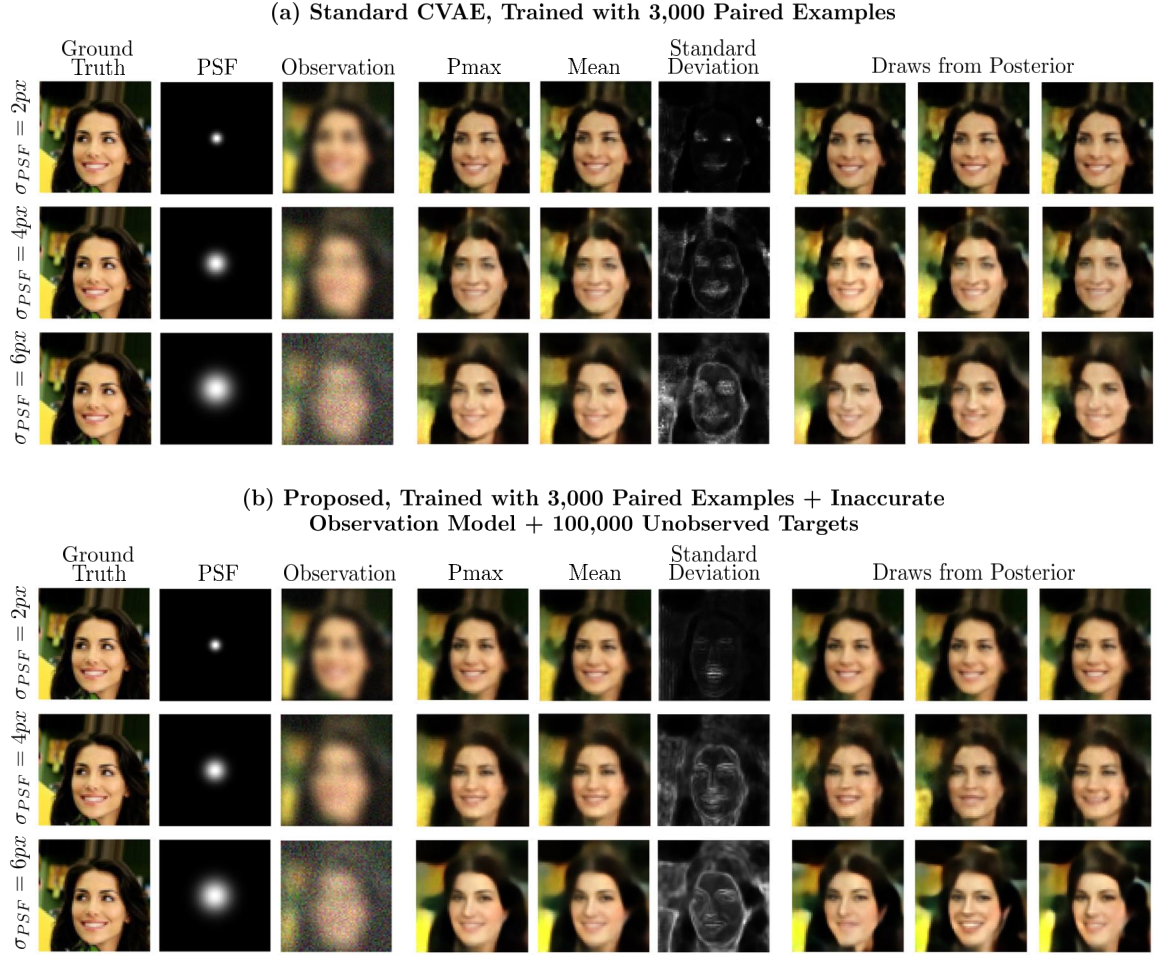

Figure 5: Comparison between standard CVAE trained with paired examples and proposed framework. (a) Posterior recovery obtained with a CVAE trained on $3,000$ available image-observation pairs. The number of available paired examples is not sufficient to train a CVAE capable of capturing the variability of the solution space and the model over-fits; draws from the posterior are all very similar independently of how ill-posed the de-convolution inverse problem is. (b) Posterior recovery obtained by training the CVAE with the proposed framework, exploiting all sources of information available. This model adequately captures the variability of the different solution spaces; as blurring and noise become more severe and the corresponding inverse problem more ill-posed, the draws from the posterior increasingly diversify, exploring the possible faces that could lie behind the associated observation.

with the aim of capturing the posterior of solutions to the resulting inverse problem. To simulate typical CI conditions, only a small subset of images degraded with the true transformation is made accessible. However, the whole training set of ground truth images remains available, as this does not rely on the particular imaging instrument and can be
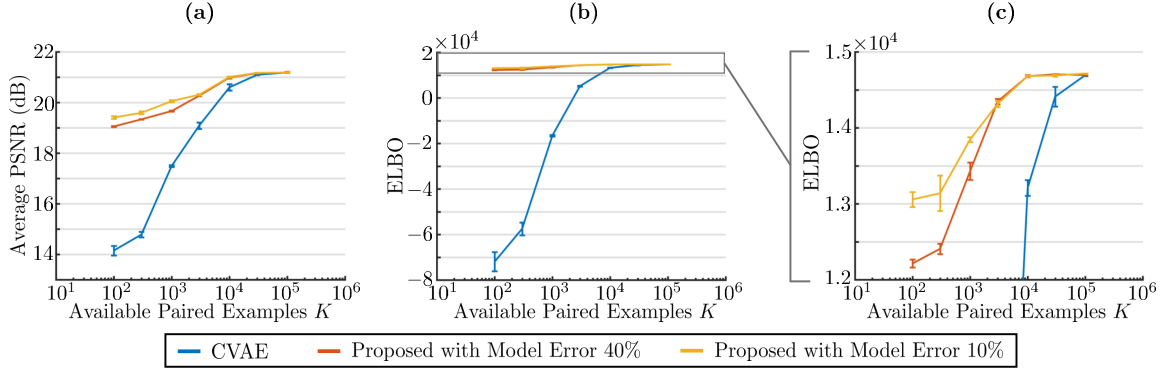
Figure 6: Posterior reconstruction from blurred CelebA images at varying number K of paired training examples. (a) Average PSNR between reconstructed pseudo-maximum and ground truth images. (b-c) ELBO assigned to the test set by the trained models. A standard CVAE requires a large number K of paired images and observations to obtain accurate mean reconstructions (high PSNR values), and tens of thousands of examples before yielding distributions that approximately match the true posterior (high ELBO values). For many imaging tasks these would be prohibitively expensive to collect. By incorporating additional cheap sources of information, the proposed framework achieves high performance with far fewer paired examples, minimising data collection costs. Furthermore, training with a more accurate observation model yields superior performance when fewer paired examples are available, proving the ability of the framework to benefit from improved domain expertise.

sourced independently. In addition, an inaccurate degradation function is provided to simulate domain expertise. Inaccuracies compared to the true transformation are simulated with errors on the transformation's parameters.

4.1.1 Comparison with Standard training of CVAEs

As a first example, three different levels of Gaussian blurring and additive noise degradation conditions are considered with 64×64 images. The models are given $K = 3,000$ paired examples generated with the true transformation to train upon. The inaccurate observation model exploited by the proposed framework under-estimates the point spread function (PSF) width and noise standard deviation by 25% compared to the true transformation. Reconstruction examples are shown in Figure 5. More experimental details are given in appendix D.1.

With only $3,000$ image-observation pairs to train upon, the CVAE is not able to properly capture the variability of the solution space. The model returns compelling mean and pseudo-max reconstructions, but fails to explore the variation of possible solutions; different draws from the recovered posterior remain very similar to each other and do not properly represent the range of faces that could generate the observed blurred image. Contrarily, by including the additional data and model with the proposed framework, the CVAE is
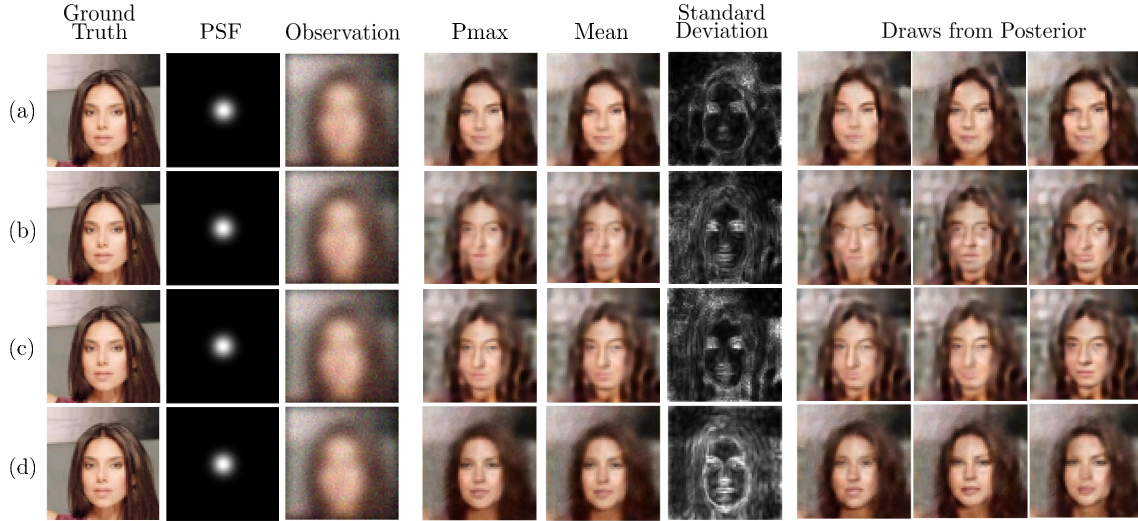
Figure 7: Posterior recovery from blurred. (a) CVAE trained with $K = 3,000$ paired examples alone. With too few examples to train with, the model over-fits and draws from the posterior are all very similar. (b) CVAE trained with $L = 100,000$ target examples and corresponding simulated observations from the inaccurate observation model. Because the observation model does not match the true one encountered upon testing, reconstructions display noticeable artefacts. (c) CVAE trained with $K = 3,000$ paired examples in combination with $L = 100,000$ target examples and corresponding inaccurately simulated observations. The presence of real measurements in the training set improves reconstruction marginally, but artefacts are still largely present. (d) CVAE trained with proposed variational framework. The sources of information are exploited in a principled way, resulting in accurate posterior recovery; different draws explore various plausible reconstructions.

adequately trained. Draws from the posterior increasingly diversify as blurring and noise intensify, reflecting the increasing variance of the solution space.

To test the proposed framework in different conditions, multiple experiments analogous to those illustrated in Figure 5, with different relative model errors, are performed varying the number K of available image-observation pairs. More experimental details are given in appendix D.1. As shown in Figure 6, a standard CVAE yields very low peak signal to noise ratio (PSNR) if the number K of available paired training data is below a few thousands, indicating poor mean performance. The behaviour of the ELBO is even more dramatic, essentially suggesting complete inability to capture the posterior of solutions with less than a few tens of thousands paired examples. In many imaging settings, collecting such a high number of image-observation pairs would be extremely expensive. Instead, by incorporating additional cheap sources of information, the proposed framework displays appreciable PSNR and ELBO, even with very scarce paired image-observation examples. Furthermore, the use

(a) CelebA, L = 100,000 Target Examples

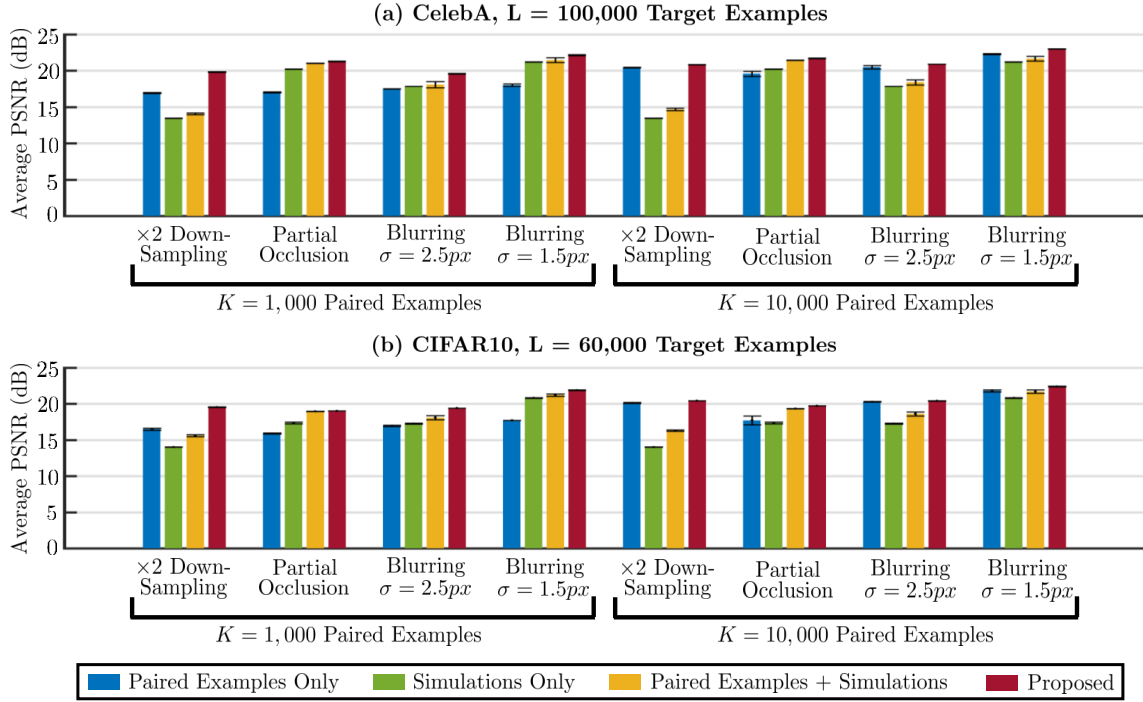**(b) CIFAR10, L = 60,000 Target Examples**

Figure 8: average test PSNR between ground truth images and reconstructed pseudo-maxima for (a) CelebA and (b) CIFAR10 images. The proposed framework consistently outperforms other training methods that exploit part of or all of the same sources of information.

of a more accurate observation model was found to sensibly improve reconstructions at low numbers K of available paired examples, to then converge towards similar performance as this was increased. Such results indicate that the proposed framework is able to make better use of empirical data and domain expertise than the naive baselines; the accuracy of the analytical observation model affects the recovery when availability of empirical evidence is low, but is progressively less influential as more data becomes available.

4.1.2 Comparison with alternative Training Methods

Given the small number K of paired training data, a large number L of target examples and an inaccurate observation model, one can conceive different naive ways to train a conditional generative model for inversion:

  i Standard conditional training; discard the availability of target examples and domain expertise and train solely on K empirical target-observation pairs.

 ii Use of domain expertise only; simulate a large number L of measurements from all available targets through the analytical model and use these as pairs to train the model.
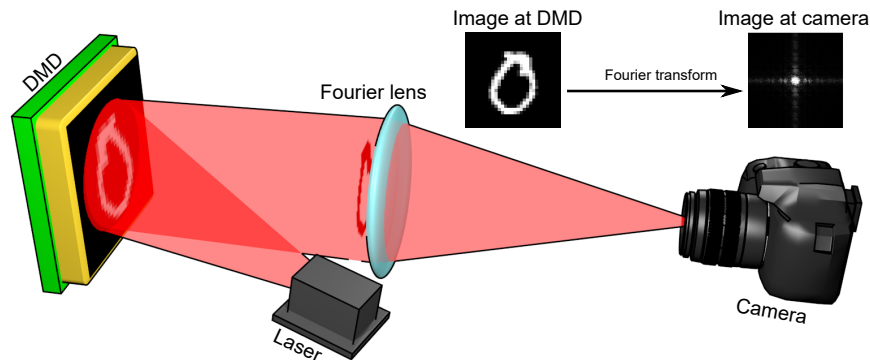
Figure 9: Experimental set up used for holographic image reconstruction. A binary amplitude image is projected by the Digital Micromirror Device (DMD) and a lens placed at the focal distance from the DMD display produces the corresponding Fourier image at the camera.

iii Combining the previous two approaches; the K targets for which empirical measurements are available are paired with them, while the L unobserved targets are paired with simulated measurements.

These three approaches are compared to the proposed method in Figure 7. Details of the experimental conditions can be found in appendix D.1. Similar experiments are then performed to reconstruct 32×32 images from the CelebA data set and the CIFAR10 data set. Different degradation conditions are tested, including blurring, down-sampling and partial occlusion. For each case, models are trained with $K = 1,000$ and $K = 10,000$ available training pairs. More details about these experiments can be found in appendix D.1. Reconstructions are performed with $2,000$ test examples. Figure 8 shows the average PSNR, while recovered ELBO values are reported in table 1.

The proposed framework proved advantageous across all tested conditions, both with respect to the mean reconstruction quality, given by the mean PSNR values, and the recovered posterior density matching, approximately measured by the ELBO values. It is also noticeable how the choice of optimal approach amongst the three naive strategies is far from obvious; which training method yields best performance is highly dependent on available number K of image-observation pairs and type of transformation. In contrast, the proposed framework consistently gives the best results, proving its ability to better exploit the provided information, independently of the particular conditions.

4.2 Holographic image reconstruction

Sensor arrays, such as CCD or CMOS cameras, are a ubiquitous technology that obtain a digital image of a scene. However, cameras are only able to retrieve the intensity of the light field at every point in space, computational techniques and additional elements in imaging set-ups are required to obtain the full information of the light field, i.e. both amplitude and phase. Unfortunately, it is not always possible to include the additional experimental

Table 1: Test set evidence lower bound (ELBO) for the proposed framework compared to alternative methods of using the same information to train a CVAE. The proposed framework consistently returns higher ELBO values, indicating a more accurate posterior recovery.

	Paired Examples	Simulations	paired+ Simulations	Proposed
CelebA, K=1,000 ×2 Down-sampling	-19827 ± 771	-208 ± 403	11300 ± 454	$\mathbf{14553 \pm 20}$
CelebA, K=1,000 Partial Occlusion	-21390 ± 1655	14124 ± 33	14751 ± 42	$\mathbf{15134 \pm 18}$
CelebA, K=1,000 Blurring $\sigma = 2.5px$	-16264 ± 122	10581 ± 21	12371 ± 532	$\mathbf{13365 \pm 201}$
CelebA, K=1,000 Blurring $\sigma = 1.5px$	-13872 ± 1298	13152 ± 62	13805 ± 221	$\mathbf{14189 \pm 63}$
CelebA, K=10,000 ×2 Down-sampling	13450 ± 149	-208 ± 403	10303 ± 1192	$\mathbf{14763 \pm 2}$
CelebA, K=10,000 Partial Occlusion	12902 ± 556	14124 ± 33	15043 ± 17	$\mathbf{15187 \pm 32}$
CelebA, K=10,000 Blurring $\sigma = 2.5px$	13265 ± 53	10581 ± 21	12635 ± 437	$\mathbf{14672 \pm 9}$
CelebA, K=10,000 Blurring $\sigma = 1.5px$	13502 ± 310	13152 ± 62	13936 ± 136	$\mathbf{14842 \pm 11}$
CIFAR10, K=1,000 ×2 Down-sampling	-21846 ± 2128	-3059 ± 987	12005 ± 921	$\mathbf{14247 \pm 19}$
CIFAR10, K=1,000 Partial Occlusion	-23358 ± 2188	12890 ± 57	14118 ± 61	$\mathbf{14702 \pm 64}$
CIFAR10, K=1,000 Blurring $\sigma = 2.5px$	-18683 ± 51	10051 ± 82	12924 ± 296	$\mathbf{13212 \pm 195}$
CIFAR10, K=1,000 Blurring $\sigma = 1.5px$	-14390 ± 40	13008 ± 105	13869 ± 174	$\mathbf{13988 \pm 25}$
CIFAR10, K=10,000 ×2 Down-sampling	13496 ± 69	-3059 ± 987	12096 ± 577	$\mathbf{14415 \pm 23}$
CIFAR10, K=10,000 Partial Occlusion	12171 ± 925	12890 ± 57	14427 ± 37	$\mathbf{14789 \pm 38}$
CIFAR10, K=10,000 Blurring $\sigma = 2.5px$	13134 ± 219	10051 ± 82	13094 ± 312	$\mathbf{14348 \pm 30}$
CIFAR10, K=10,000 Blurring $\sigma = 1.5px$	13402 ± 177	13008 ± 105	13974 ± 141	$\mathbf{14540 \pm 21}$

components to the set-up and therefore algorithms have been adapted to use only intensity images. Retrieving the full light field information from intensity-only measurements is a very important inverse problem that has been studied exhaustively during the last 40 years (Gerchberg, 1972; Fienup, 1982; Shechtman et al., 2015).

Machine learning methods have been proposed in this context to learn either phase or amplitude of images/light fields from intensity-only diffraction patterns recorded with a camera (Sinha et al., 2017; Rivenson et al., 2018, 2019). Such an ability is desirable because the intensity images can be recorded with cheap digital cameras, instead of expensive and delicate phase-sensitive instruments. Following these recent advances, we aim to use our proposed variational framework approach to solve the following problem: Given the camera intensity image of the diffraction pattern at the Fourier plane, what is the amplitude of the corresponding projected image? This apparently simple problem has multiple applications in areas such as material science, where X-rays are used to infer the structure of a molecule from its diffraction pattern (Marchesini et al., 2003; Barmherzig et al., 2019), optical trapping (Jesacher et al., 2008), and microscopy (Faulkner and Rodenburg, 2004).

4.2.1 Experimental set-up and Data

The experiment consisted of an expanded laser beam incident onto a Digital Micromirror Device (DMD) which displays binary patterns, as shown in Figure 9. DMDs consist of an array of micron-sized mirrors that can be arranged into two angles that correspond to "on" and "off" states of the micromirror. Consequently, the amplitude of the light is binarised by the DMD pattern and propagates toward a single lens. The lens, placed at the focal distance from the DMD display, will cause the rays to form the Fourier image of the MNIST digit at the camera.

To make the problem even harder, we work under saturation conditions, i.e. assuming blinding on the camera, and with extremely low-resolution images. We display 9600 MNIST digits on the DMD and record the corresponding camera observations. This data is used as high fidelity paired ground truths X^* and measurements Y^*. The remaining 50400 MNIST examples are used as the large set of unobserved ground truth signals X. The analytical observation model $p(\widetilde{y}|x)$ is built as a simple intensity Fourier transform computation, to which we add artificial saturation.

4.2.2 Reconstruction

Image reconstruction from the image at the Fourier plane is performed with the proposed variational framework and compared with a Hybrid Input-Output (HIO) complex light-field retrieval algorithm and a 4-layer deep Artificial Neural Network (deep ANN) as shown in Figure 10. On the one hand, given that the HIO retrieval algorithm is an iterative method that uses the light intensity pattern recorded by the camera at the Fourier plane at each iteration, it is not expected to operate well in conditions of saturation and/or down-sampling (see appendix D.2 for details). This is precisely what we observe in Figure 10(c), where the HIO algorithm simply predicts spots at some positions.

The results of a deep ANN show that a more accurate solution can be found. However, the accuracy of the deep ANN to reconstruct ground truth is hindered by the limited training set of 9600 experimental images. As shown in Figure 10(d), highly accurate reconstructed
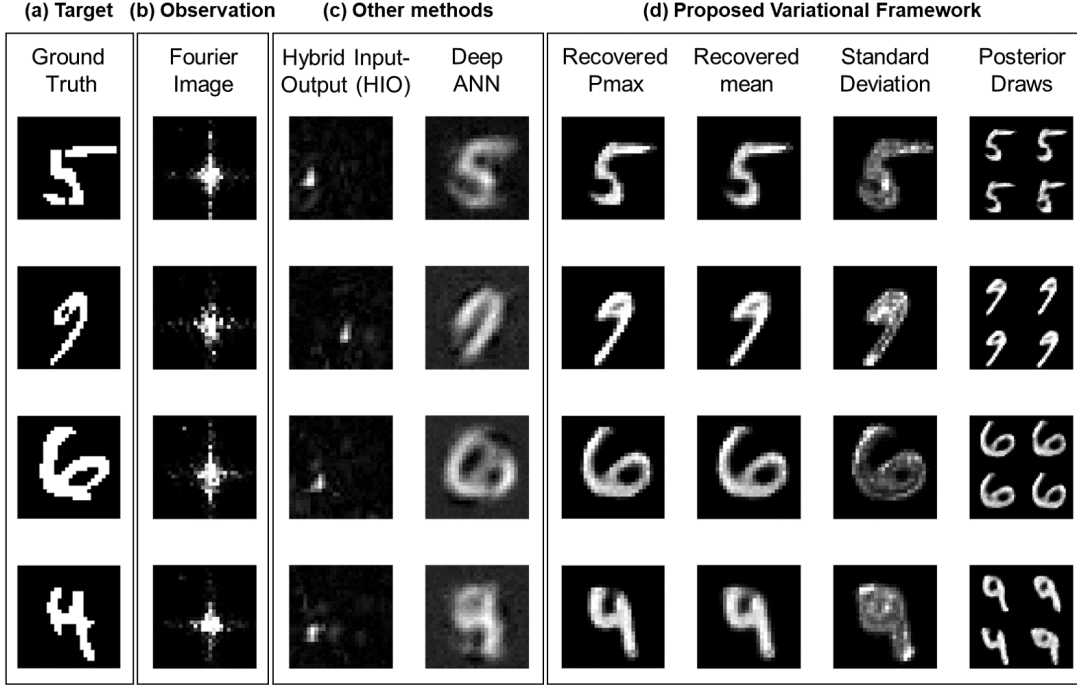
Figure 10: Reconstructions from experimental Fourier intensity image data. (a) Target images projected by the DMD, (b) intensity Fourier image observed at the camera, (c) reconstructions using other techniques (d) reconstructions obtained with the proposed variational method, including pseudo-maximum, pixel-marginal mean, pixel-marginal standard deviation and examples of draws from the recovered posterior.

images are achieved with the proposed variational method which exploits the generative multi-fidelity forward model to train the inverse model using the additional unobserved 50400 examples. Furthermore, the proposed method retrieves full posterior densities, from which we can draw to explore different possible reconstructions as a result of the ill-posed nature of the inverse problem.

In order to demonstrate the advantage of employing the proposed framework compared to naive strategies in a real scenario, we repeat the evaluation of figure 7 for this physical experiment. An example is shown in figure 11. Analogously to the simulated experiments, using the experimental training set alone gives results of limited quality. Combining the simulations and real data in naive ways completely disrupts reconstructions, as in this experiments the simulations are significantly different from real measurements. However, they are far from useless, as including them in a principled way through the proposed framework gives significant improvement in reconstruction quality.

Figure 12 illustrates further this posterior exploration capability. When progressively down-sampling the resolution of experimentally measured observations, the pseudo-max
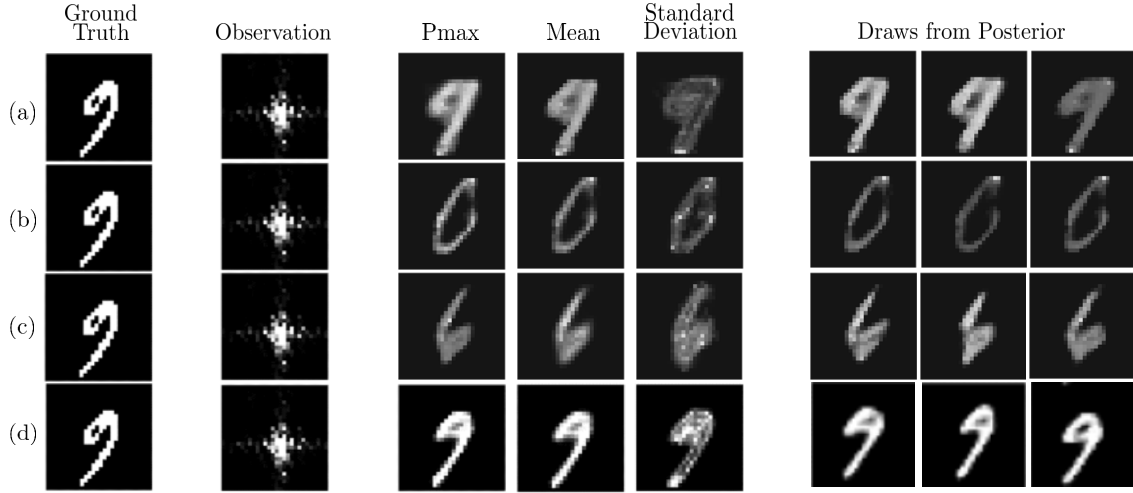
Figure 11: Image posterior recovery from phase-less measurements. (a) CVAE trained with the available $K = 9,600$ paired examples alone. The size of this training set is too small to obtain accurate posteriors. (b) CVAE trained with $L = 50,400$ target examples and corresponding simulated observations from the inaccurate observation model. Because the observation model, i.e. a simple Fourier transform, does not match the true one encountered upon testing, the image is not well recovered. (c) CVAE trained with $K = 9,600$ paired examples in combination with $L = 50,600$ target examples and corresponding inaccurately simulated observations. The presence of real measurements in the training gives more realistic MNIST-like shapes, but the reconstruction is still inaccurate. (d) CVAE trained with proposed variational framework. The sources of information are exploited in a principled way, resulting in accurate posterior recovery.

reconstructed image quality degrades and the range of possible solutions, visualised through the different draws, extends. When down-sampling the experimental images to a resolution of (16×16), the inverse problem becomes critically ill-posed such that the solution space becomes too varied to accurately recover the ground truth image.

4.3 Imaging Through Highly Scattering Media

Imaging through strongly diffusive media remains an outstanding problem in optical CI, with applications in biological and medical imaging and imaging in adverse environmental conditions (Jiang, 2018). Visible or near-infrared light does propagate in turbid media, such as biological tissue or fog, however, its path is strongly affected by scattering, leading to the loss of any direct image information after a short propagation length. The reconstruction of a hidden object from observations at the scattering medium's surface is the inverse problem that will be addressed in this section.
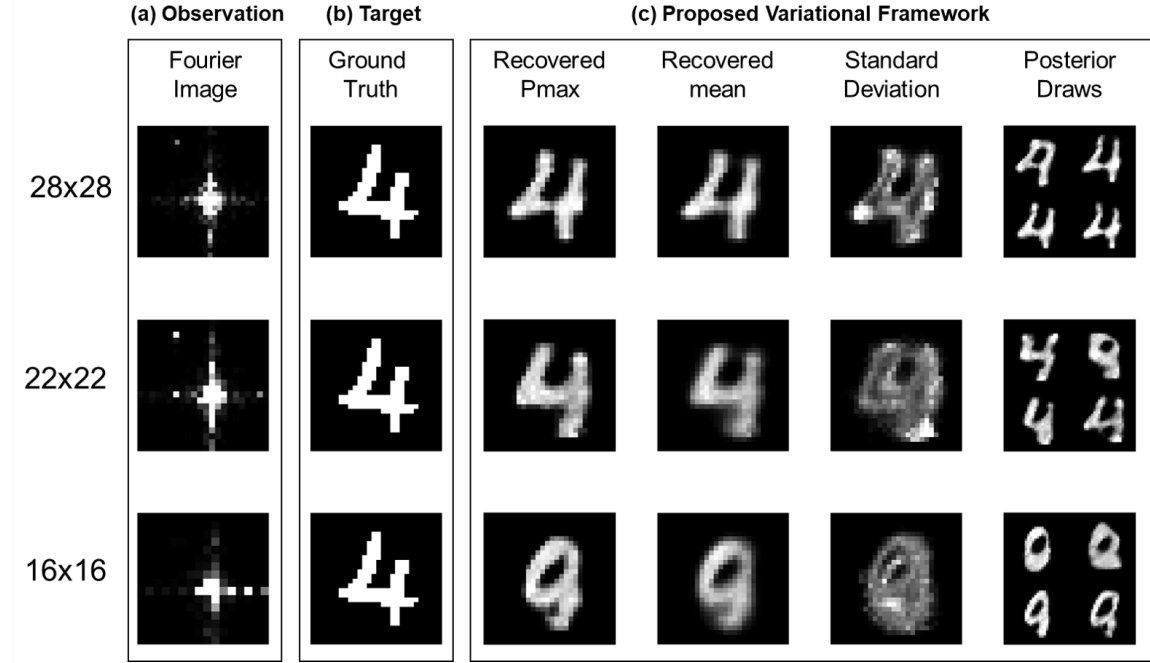

Figure 12: (a) Experimental Fourier intensity image data down-sampled to (28×28), (22×22) and (16×16) (top to bottom) for (b) the same target image. (c) The proposed variational framework which shows the reconstructed image quality degrades with decreasing resolution of the measured data. As expected, the standard deviation and samples from the recovered posterior show high variability to the solution when reaching the critically ill-posed resolution limit of (16×16).

4.3.1 PHYSICAL EXPERIMENT

Following the experimental implementation presented by Lyons et al. (2019), imaging is performed with a 130 fs near-infrared pulsed laser and a single photon sensitive time of flight (ToF) camera with a temporal resolution of 55 ps to perform transmission diffuse imaging. In these experiments, different cut-out shapes of alphabetic letters were placed between two identical 2.5 cm thick slabs of diffusive material, with measured absorption and scattering coefficients of $\mu_a = 0.09$ cm^{-1} and $\mu_s = 16.5$ cm^{-1} respectively. A schematic representation and a photograph of the set up are shown in Figure 13(a-b).

A pulse of light from the laser propagates through the diffusing material, reaches the hidden object, which partially absorbs it, and then propagates further through the medium to the imaged surface. The ToF camera records a video of the light intensity as a function of time as it exits the medium. A video recorded with an empty piece of material is used as background and subtracted to that obtained with the object present, thereby obtaining a video of the estimated difference in light intensity caused by the hidden object. An example of such videos is shown in Figure 13(c). At this depth, more than 40 times longer than the photon's mean free path, the diffusion effect is so severe that even basic shapes are not
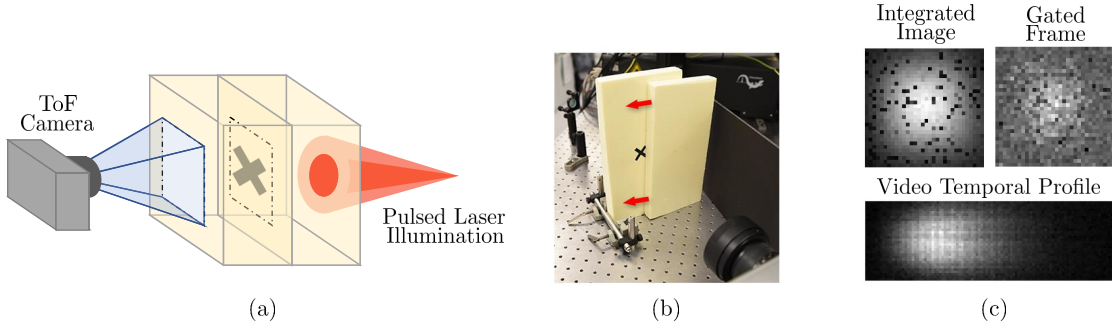
Figure 13: Experimental set up for imaging through scattering media. (a) Schematic representation of the experiment. A target object is embedded between two 2.5 cm-thick slabs of diffusing material, with absorption and scattering properties comparable to those of biological tissue. One exposed face is illuminated with a pulsed laser and the opposite face is imaged with the ToF camera. (b) A photograph of the same experimental set up. (c) Example of the video recorded by the ToF camera as light exits the medium's surface. Images show the integration over all time frames (i.e. the image a camera with no temporal resolution would acquire), a single frame of the video gated in time and the intensity profile of a pixels' line at different times.

distinguishable directly from the videos. Furthermore, the measurements experience low signal-to-noise ratio due to the low light intensity that reaches the imaged surface and the low fill factor of the ToF camera, which is about 1%. Achieving accurate reconstructions with simple objects in this settings, is a first important step towards achieving imaging through biological tissue with near-infrared light and hence non-ionising radiation.

4.3.2 TRAINING DATA AND MODELS

As target objects in these experiments are character-like shapes, the training images are taken from the NIST data set of hand-written characters (Johnson, 2010). $86,400$ NIST images are used as the large data set of unobserved target examples X. Because of experimental preparation, it is infeasible to perform a large number of physical acquisitions to build a training set. However, the process of light propagation through a highly scattering medium can be accurately described with the diffusion approximation, commonly adopted in these settings (Lyons et al., 2019; Yoo et al., 1990). The propagation of photons under this assumption is described by the following differential equation

$$c^{-1}\frac{\partial \Phi(\vec{r},t)}{\partial t} + \mu_a \Phi(\vec{r},t) - D\nabla \cdot [\nabla \Phi(\vec{r},t)] = S(\vec{r},t), \tag{17}$$

where c is the speed of light in the medium, $\vec{r}$ is the spatial position, t is the temporal coordinate, $\Phi(\vec{r},t)$ is the photons flux, $S(\vec{r},t)$ is a photon source, here the illumination at the surface, and $D = \big(3(\mu_a + \mu_s)\big)^{-1}$. The measurements recorded by the ToF camera in

the experiment described above can be accurately simulated by numerically propagating the photon flux $\Phi(\vec{r}, t)$ in space and time with appropriate boundary conditions at the edges of the medium and a high absorption coefficient μ_a assigned to the object voxels. These simulations are accurate, but expensive. To simulate the experiments of interest here they take in the order of a few minutes per example to run on a TitanX GPU. Obtaining paired inputs and outputs for tens of thousands of experiments is expensive. Instead, only $1,000$ examples of the $84,400$ training targets were generated in this way and were taken as high-fidelity measurement estimates Y^* from corresponding ground truth images X^*. An example of such simulations for one of the test characters is shown in Figure 14(c-d).

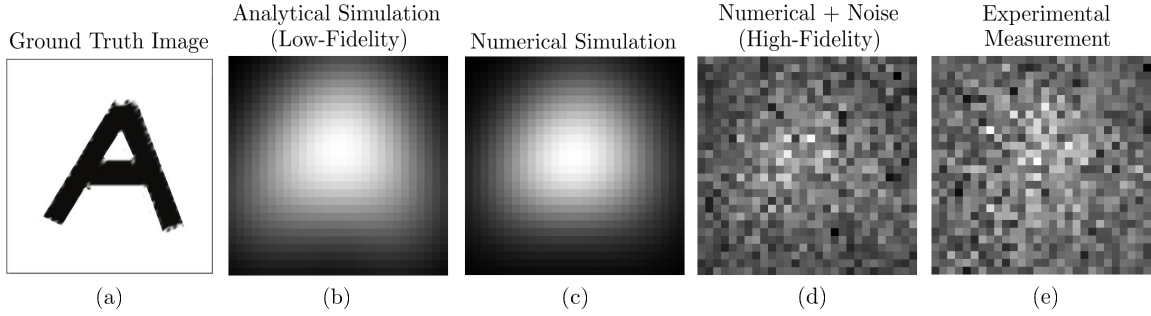| Ground Truth Image	Analytical Simulation (Low-Fidelity)	Numerical Simulation	Numerical + Noise (High-Fidelity)	Experimental Measurement
(a)	(b)	(c)	(d)	(e)

Figure 14: Simulated and real measurements from the time of flight (ToF) Camera. Images are single frames from the camera videos. (a) Image of the hidden object. (b) Simulated measurement using the analytical solution from the linear approximation, taken as low-fidelity estimate. (c) Simulation obtained by numerically propagating the diffusion equation, which is accurate, but expensive. (d) Numerical simulation with added noise, used as high-fidelity estimates of the measurements. (e) The real measurements recorded by the ToF camera for this object.

In order to simulate measurements at a lower computational cost, a linear approximation of the observation process can be exploited (Lyons et al., 2019; Yoo et al., 1990). For a delta function initial illumination $S(\vec{r}, t) = \delta(\vec{r} = \vec{r}', t = t')$ and an infinite uniform scattering medium, an analytical solution for $\Phi(\vec{r}, t)$ exists:

$$\Phi(\vec{r}, t; \vec{r}', t') = \frac{c}{[4\pi Dc(t - t')]^{3/2}} \times \exp\left[-\frac{|\vec{r} - \vec{r}'|^2}{4Dc(t - t')}\right] \exp\left[-\mu_a c(t - t')\right]. \tag{18}$$

This solution constitutes a point spread function with which an analytical estimate of the measurements can be computed through two consecutive convolutions. First, the illumination at the entering surface is convolved in 2D and time with the PSF of equation 18 to obtain an estimate of the illumination at the object plane. Second, this estimate multiplied by the object image at each time frame is convolved again with the PSF to estimate the intensity field at the exiting surface, imaged by the ToF camera (Lyons et al., 2019). An example of such analytical estimates of the measurements is shown if Figure 14(b). These computations are much less expensive to perform than propagating numerically the diffusion equation, requiring less than 100 ms per sample to run on a TitanX GPU. However, they introduce

28

approximations which sacrifice the accuracy of the simulated measurements. In particular, they don't take into account any boundary condition and assume that the observation process is linear, whereas in reality the light absorbed by some part of the object will affect the illumination at some other part. This analytical observation model is taken as the approximate likelihood $p(\widetilde{y}|x)$ generating low-fidelity measurement's estimates $\widetilde{y}$.

4.3.3 RESULTS

The ToF videos recorded for three different shapes embedded in the scattering medium were used to perform reconstructions. Firstly, the recovery is performed using the method presented by Lyons et al. (2019), consisting of a constrained minimisation with ℓ_1 and total variation regularisation. Secondly, retrieval is performed with a CVAE trained with the proposed framework and using the sources of information described above. Results are shown in Figure 15.
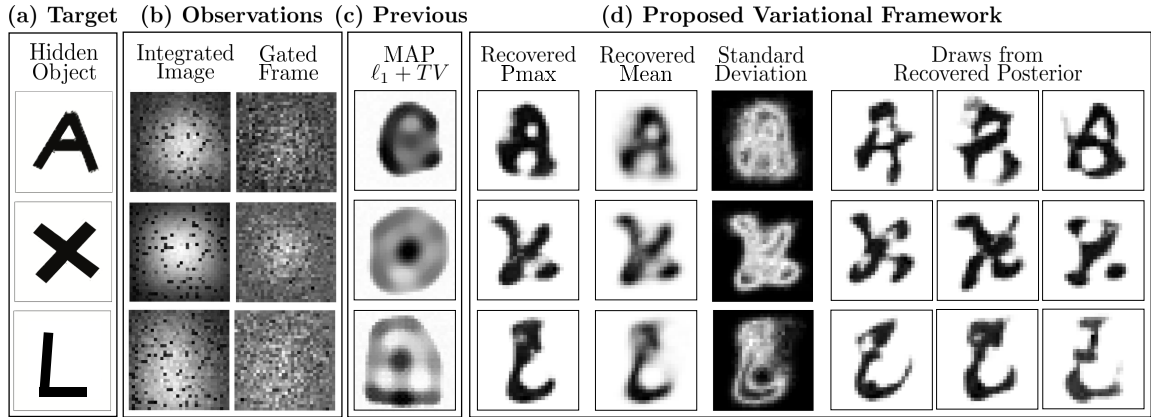

Figure 15: Reconstructions from experimental ToF videos. (a) Target images embedded in the scattering medium, (b) integrated and gated frames from the ToF camera videos, constituting the observed measurements, (c) reconstruction obtained using constrained optimisation with ℓ_1-norm and total variation regularisation and (d) reconstructions obtained with the proposed variational method, including pseudo-maximum, pixel-marginal mean, pixel-marginal standard deviation and examples of draws from the recovered posterior. The proposed framework recovers arguably more accurate images compared to the state of the art, while also allowing exploration of the manifold of possible solutions to the inverse problem.

The prior method is capable of retrieving general features of the objects embedded in the scattering medium, but sometimes results in severe artefacts that make the images unrecognisable. Furthermore, to obtain the displayed results, it is necessary to carefully tune the penalty coefficients of the constrained optimisation for each example, making such retrieval highly dependent on human supervision. Exploiting a more specific empirical prior,

the proposed framework allows to retrieve more accurate reconstructions, where the different letters are clearly recognisable. Moreover, this particularly ill-posed inverse problem example highlights the importance of using a Bayesian approach; the solution space given a diffuse ToF video is rather variable and, unlike constrained optimisation and other single estimate methods, through the approximate posterior such variability can be captured by empirically estimating uncertainty and visualising different drawn samples, as shown in Figure 15(d). Note that, thanks to the proposed framework, the model was successfully trained with very limited effort and resources; the large data set of targets was readily available independently of the application of interest, while only $1,000$ expensive simulations were used, requiring just a few tens of hours of simulation time on a single GPU to be generated.

5. Conclusion

This paper introduced a novel framework to train variational inference in imaging inverse problems, utilising different types of data and domain expertise in combination. As a result, Bayesian machine learning is rendered accessible for a broad range of imaging applications, where empirical training data is typically scarce or expensive to collect. The method was derived from a Bayesian formulation of inverse problems and interpreting accessible sources of information as approximations to or samples from underlying distributions, providing theoretical foundation. Simulated experiments thoroughly tested the proposed technique in a range of different conditions, proving its ability to better exploit all sources of information available. The method was then applied to real imaging systems, demonstrating the first successful application of Bayesian machine learning in both phase-less holographic image reconstruction and imaging through scattering media. In both settings, state of the art reconstructions were achieved, while requiring little training collection efforts, whereas before Bayesian methods would have required prohibitively large volumes of data.

Appendix A. Details of ELBO Formulation

A.1 VAE Formulation for Multi-Fidelity Forward Model

Through Jensen's inequality, a lower bound for the parametric distribution $p_\alpha(y|x)$ can be defined as

$$
\log p_\alpha(y_k|x_k) \geq \int p(\widetilde{y}|x_k) \int q_\beta(w|x_k, y_k, \widetilde{y}) \log \left[\frac{p_{\alpha_1}(w|x_k, \widetilde{y})}{q_\beta(w|x_k, y_k, \widetilde{y})} p_{\alpha_2}(y_k|x_k, \widetilde{y}, w) \right] dw d\widetilde{y}
$$
$$
= \int p(\widetilde{y}|x_k) \left[\int q_\beta(w|x_k, y_k, \widetilde{y}) \log p_{\alpha_2}(y_k|x_k, \widetilde{y}, w) dw - D_{KL}(q_\beta||p_{\alpha_1}) \right] d\widetilde{y}, \tag{19}
$$

where $q_\beta(w|x, y, \widetilde{y})$ is the recognition model, chosen as an isotropic Gaussian distribution, the moments of which are outputs of a neural network taking as inputs targets x, high-fidelity measurements y and low-fidelity measurements $\widetilde{y}$. $D_{KL}(q_\beta||p_{\alpha_1})$ is the KL divergence between the distributions q_β and p_{α_1} defined as

$$
D_{KL}(q_\beta||p_{\alpha_1}) = \int q_\beta(w|x_k, y_k, \widetilde{y}) \log \frac{q_\beta(w|x_k, y_k, \widetilde{y})}{p_{\alpha_1}(w|x_k, \widetilde{y})} dw. \tag{20}
$$

As both $p_{\alpha_1}(w|x_k, \widetilde{y})$ and $q_\beta(w|x_k, y_k, \widetilde{y})$ are isotropic Gaussian distributions, a closed form solution for the KL divergence exists and can be exploited in computing and optimising the lower bound (Kingma and Welling, 2014).

A.2 VAE Formulation for Variational Inverse Model

using Jensen's inequality, a tractable lower bound for the expression of equation 15 can be derived with the aid of a parametric recognition model $q_\phi(z|x, y)$ as

$$
\int p(x) \int p_\alpha(y|x) \log \int r_{\theta_1}(z|y) r_{\theta_2}(x|z, y) dz dy dx \geq
$$
$$
\int p(x) \int p_\alpha(y|x) \int q_\phi(z|x, y) \log \left[\frac{r_{\theta_1}(z|y)}{q_\phi(z|x, y)} r_{\theta_2}(x|z, y) \right] dz dy dx = \tag{21}
$$
$$
\int p(x) \int p_\alpha(y|x) \left[\int q_\phi(z|x, y) \log r_{\theta_2}(x|z, y) dz - D_{KL}(q_\phi(z|x, y)||r_{\theta_1}(z|y)) \right] dy dx.
$$

The recognition model $q_\phi(z|x, y)$ is an isotropic Gaussian distribution in the latent space, with moments inferred by a neural network, taking as input both example targets x and corresponding observations y. This neural network may be fully connected, partly convolutional or completely convolutional, depending on the nature of the targets x and observations y. As both $q_\phi(z|x, y)$ and $r_{\theta_1}(z|y)$ are isotropic Gaussians in the latent space, their KL divergence $D_{KL}(q_\phi(z|x, y)|r_{\theta_1}(z|y))$ has an analytical solution. All the remaining integrals can be estimated stochastically, leading to the maximisation of equation 16.

Appendix B. Algorithms

In this supplementary section we detail the training procedure for the forward model $p_\alpha(y|x)$ and inverse model $r_\theta(x|y)$. The following pseudo-code details the training of the two models:

Algorithm 1 Training the Forward Model $p_\alpha(y|x)$

Inputs: Analytical forward model from domain expertise $p(\widetilde{y}|x)$; set of measured Ground-truths $X^* = \{x_{k=1:K}\}$; corresponding set of measurements $Y^* = \{y_{k=1:K}\}$; user-defined number of iterations N_{iter}; batch zise $K_b \leq K$; Initialised weights $\{\alpha_1^{(0)}, \alpha_2^{(0)}, \beta^{(0)}\}$; user-defined latent dimensionality, J_w.

1: **for** *the n'th iteration* **in** $[0 : N_{iter}]$
2:    **for** *the k'th example* **in** $[0 : K_b]$
3:       $\widetilde{y}_k \sim p(\widetilde{y}|x_k)$
4:       *compute moments of* $p_{\alpha_1^{(n)}}(w|x_k, \widetilde{y}_k)$
5:       *compute moments of* $q_{\beta^{(n)}}(w|x_k, y_k, \widetilde{y}_k)$
6:       $w_k \sim q_{\beta^{(n)}}(w|x_k, y_k, \widetilde{y}_k)$
7:       *compute moments of* $p_{\alpha_2^{(n)}}(y|x_k, \widetilde{y}_k, w_k)$
8:    **end**
9:    $\mathbf{L}^{(n)} \leftarrow \frac{1}{K_b} \sum_k^{K_b} \log p_{\alpha_2^{(n)}}(y|x_k, \widetilde{y}_k, w_k) - D_{KL}(q_{\beta^{(n)}}(w|x_k, y_k, \widetilde{y}_k)||p_{\alpha_1^{(n)}}(w|x_k, \widetilde{y}_k))$
10:    $\alpha_1^{(n+1)}, \alpha_2^{(n+1)}, \beta^{(n+1)} \leftarrow \arg\max(\mathbf{L}^{(n)})$
11: **end**

Algorithm 2 Training the Inverse Model $r_\theta(x|y)$

Inputs: Trained multi-fidelity forward model $p_\alpha(y|x)$; set of unobserved ground-truths $X = \{x_{l=1:L}\}$; user-defined number of iterations N_{iter}; batch zise $L_b \leq L$; Initialised weights $\{\theta_1^{(0)}, \theta_2^{(0)}, \phi^{(0)}\}$; user-defined latent dimensionality, J_z.

1: **for** *the n'th iteration* **in** $[0 : N_{iter}]$
2:    **for** *the k'th example* **in** $[0 : K_b]$
3:       $y_l \sim p_\alpha(y|x_l)$
4:       *compute moments of* $r_{\theta_1^{(n)}}(z|y_l)$
5:       *compute moments of* $q_{\phi^{(n)}}(z|x_l, y_l)$
6:       $z_l \sim q_{\phi^{(n)}}(z|x_l, y_l)$
7:       *compute moments of* $r_{\theta_2^{(n)}}(x|z_l, y_l)$
8:    **end**
9:    $\mathbf{L}^{(n)} \leftarrow \frac{1}{L_b} \sum_l^{L_b} \log r_{\theta_2^{(n)}}(x|z_l, y_l) - D_{KL}(q_{\phi^{(n)}}(z|x_l, y_l)||r_{\theta_1^{(n)}}(z|y_l))$
10:    $\theta_1^{(n+1)}, \theta_2^{(n+1)}, \phi^{(n+1)} \leftarrow \arg\max(\mathbf{L}^{(n)})$
11: **end**

Appendix C. Details of the Models' Architectures

The different architectures for each inference distribution implemented in the presented experiments are described here.

C.1 Multi-Fidelity Forward Model

The multi-fidelity forward model includes three parametric distributions, the parameters of which are optimised during training (see figure 2); $p_{\alpha_1}(w|x,\widetilde{y})$, $p_{\alpha_2}(y|x,\widetilde{y},w)$ and $q_\beta(w|x,y,\widetilde{y})$. Two versions of the multi-fidelity forward model were implemented. In the first, the parametric distributions consist of fully connected layers mapping inputs to outputs' Gaussian moments, from which samples are drawn upon training and inference. These structures are schematically represented in figure 16. In the second, the parametric distributions consist of deeper convolutional recurrent layers, again mapping mapping inputs to outputs' Gaussian moments, from which samples are drawn upon training and inference. These structures are instead shown in figure 17.
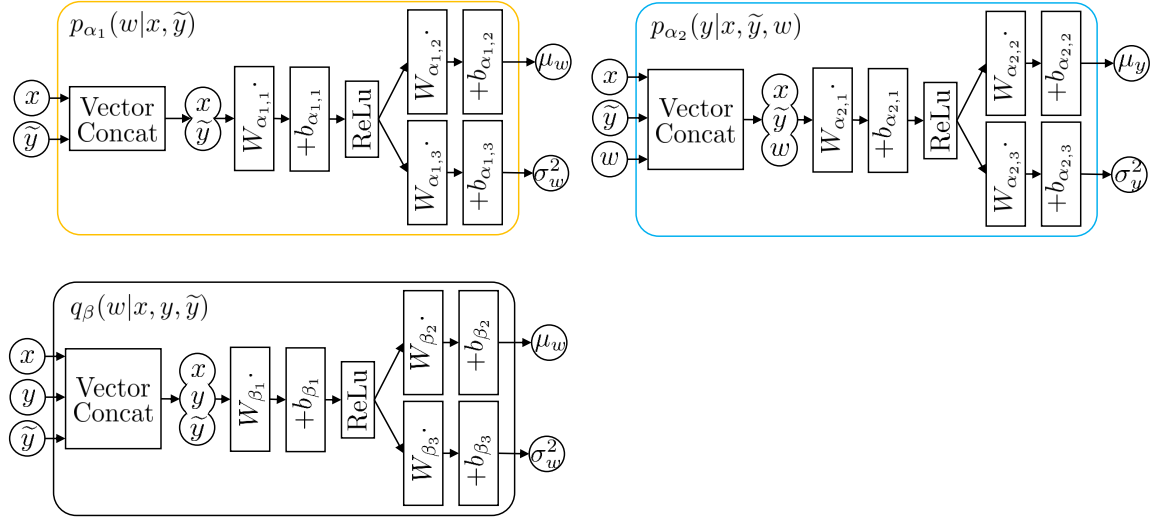

Figure 16: Parametric distributions' structures for the fully connected version of the multi-fidelity forward model. The output variables are sampled from Gaussian distributions having the corresponding output moments shown.

C.2 Variational Inverse Model

Like the multi-fidelity forward model, the inverse model includes three parametric distributions (see figure 3); $p_{\theta_1}(z|y)$, $p_{\theta_2}(x|y,z)$ and $q_\phi(z|x,y)$. As before, two versions of the inverse model were implemented. In the first, the parametric distributions consist of fully connected layers mapping inputs to outputs' Gaussian moments, from which samples are drawn upon training and inference. These structures are schematically represented in figure 18. In
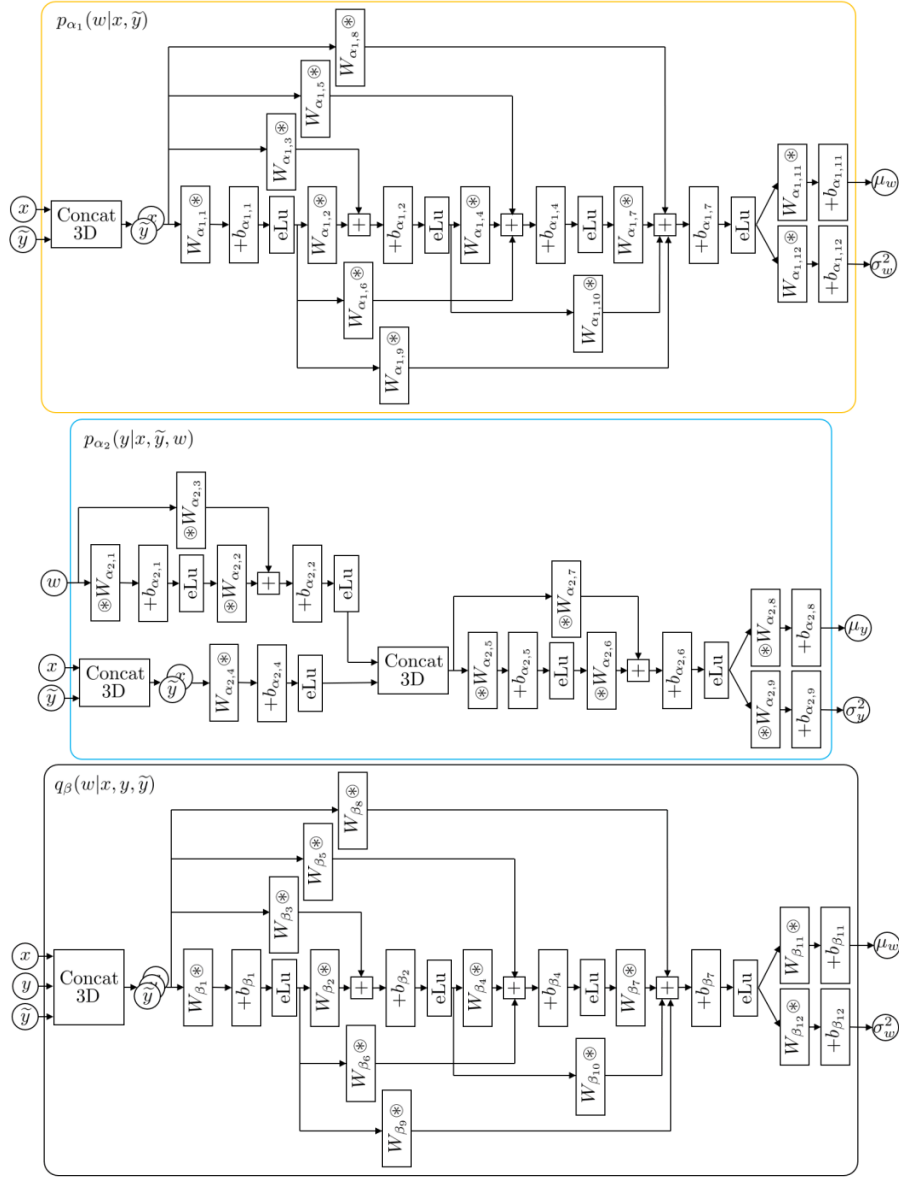
Figure 17: Parametric distributions' structures for the convolutional version of the multi-fidelity forward model. $W\circledast$ indicates a convolution with filter bank W, while $\circledast W$ indicates a transpose convolution.

the second, $p_{\theta_1}(z|y)$ and $q_\phi(z|x,y)$ consist of deeper convolutional recurrent layers, again mapping inputs to outputs' Gaussian moments, from which samples are drawn upon training and inference. $p_{\theta_2}(x|y,z)$ is similarly built with convolutional layers, but the generation of the final images is performed conditioning on previously predicted adjacent pixels with a masked convolution as described in Gulrajani et al. (2016). These structures are instead shown in figure 19.
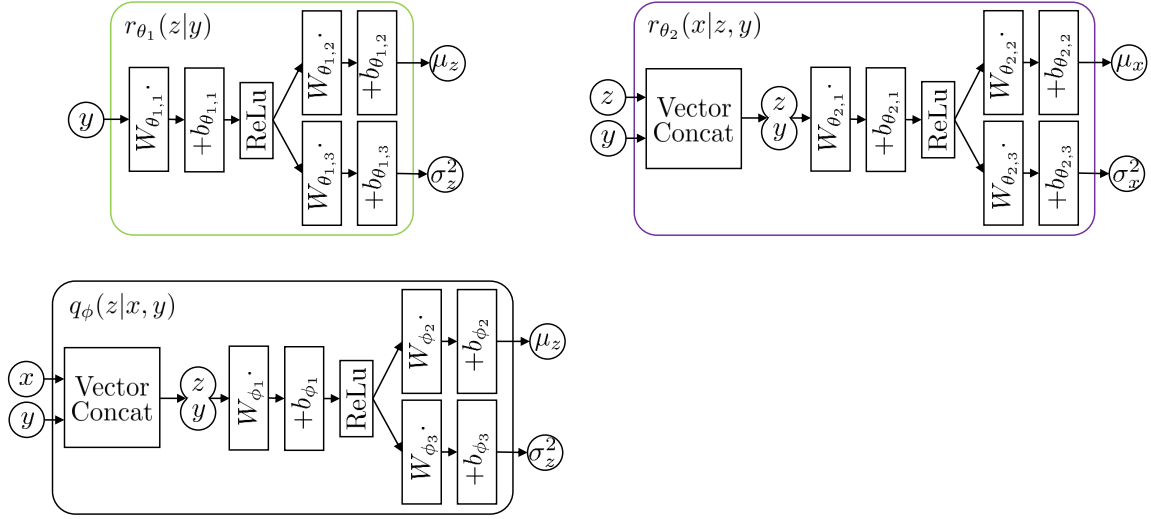
Figure 18: Parametric distributions' structures for the fully connected version of the inverse model. The output variables are sampled from Gaussian distributions having the corresponding output moments shown.

Appendix D. Details of Experiments

D.1 Simulated Experiments

D.1.1 QUALITATIVE COMPARISON WITH STANDARD CVAE

Variational models are trained to reconstruct images of faces from blurred and noisy observations. First, CVAEs are trained directly, using K available images and observations as training targets and conditions respectively. Second, the same CVAE models are trained with the proposed framework, making use of the same K paired examples, but including the whole training set of $L = 100,000$ unobserved targets from the CelebA data set and the inaccurate observation model as described in section 3.

The first set of experiments shown in figure 5 was carried out with a 64×64 down-sampled and centered version of the CelebA data set. Three Gaussian blurring conditions were tested, with increasing PSF width and noise standard deviation. In each case, the PSF and noise where chosen differently for the true transformation, applied to the small set of paired examples and the test data, and an inaccurate observation model, used instead as the low-fidelity model from domain expertise. In the first experiment, the true blurring Gaussian PSF was set to have standard deviation $\sigma_{PSF} = 2px$ and signal to noise ratio (SNR) of $25dB$, while the low-fidelity model was given $\sigma_{PSF} = 1.5px$ and $SNR = 28dB$. In the second experiment, the true blurring Gaussian PSF was set to have standard deviation $\sigma_{PSF} = 4px$ and signal to noise ratio (SNR) of $16dB$, while the low-fidelity model was given $\sigma_{PSF} = 3px$ and $SNR = 20dB$. In the third experiment, the true blurring Gaussian PSF
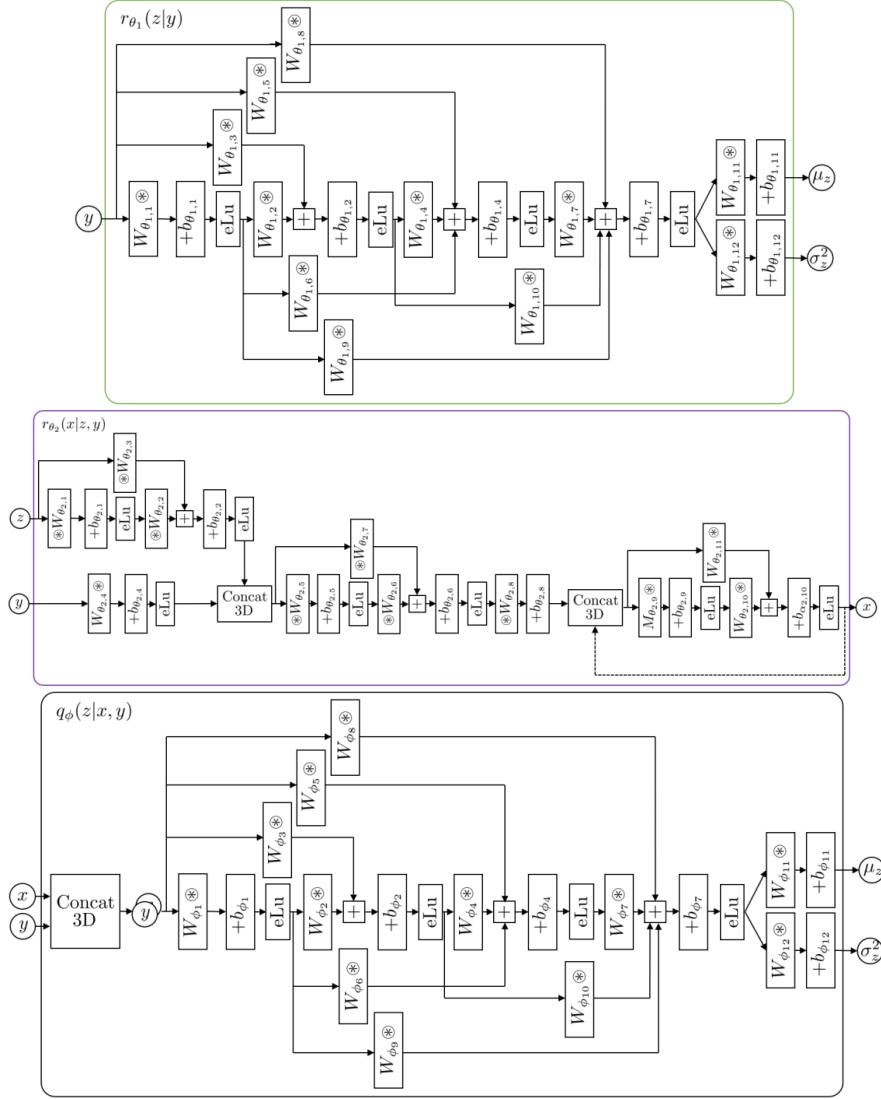
Figure 19: Parametric distributions' structures for the convolutional version of the multi-fidelity forward model. $W\circledast$ indicates a convolution with filter bank W, while $\circledast W$ indicates a transpose convolution. M indicates the masked convolution part of the PixelVAE model (see Gulrajani et al. (2016)).

was set to have standard deviation $\sigma_{PSF} = 6px$ and signal to noise ratio (SNR) of $8dB$, while the low-fidelity model was given $\sigma_{PSF} = 4px$ and $SNR = 12dB$.

The multi-fidelity forward model used in these experiment is the convolutional version, the components of which are illustrated in figure 17. The inverse model, inferring reconstructed images from blurred observations, is also the convolutional version shown in figure 19, both for the proposed training method and for the CVAE standard training. The sizes of the filter banks W used are reported in table 2.

Table 2: Filter banks of the multi-fidelity forward model and variational inverse model used in the experiments of figure 5. The table reports the filter bank name used in the architectures shown in figures 17 and 19, filters height×width×number of channels and strides of the convolutions.

Filters	h × w × c	Strides	Filters	h × w × c	Strides
$W_{\alpha_{1,1}}$	$12 \times 12 \times 10$	2×2	$W_{\alpha_{1,2}}$	$12 \times 12 \times 10$	1×1
$W_{\alpha_{1,3}}$	$12 \times 12 \times 10$	2×2	$W_{\alpha_{1,4}}$	$12 \times 12 \times 10$	2×2
$W_{\alpha_{1,5}}$	$12 \times 12 \times 10$	4×4	$W_{\alpha_{1,6}}$	$12 \times 12 \times 10$	2×2
$W_{\alpha_{1,7}}$	$12 \times 12 \times 10$	2×2	$W_{\alpha_{1,8}}$	$12 \times 12 \times 10$	8×8
$W_{\alpha_{1,9}}$	$12 \times 12 \times 10$	2×2	$W_{\alpha_{1,10}}$	$12 \times 12 \times 10$	4×4
$W_{\alpha_{1,11}}$	$12 \times 12 \times 3$	1×1	$W_{\alpha_{1,12}}$	$12 \times 12 \times 3$	1×1
$W_{\alpha_{2,1}}$	$12 \times 12 \times 10$	2×2	$W_{\alpha_{2,2}}$	$12 \times 12 \times 10$	2×2
$W_{\alpha_{2,3}}$	$12 \times 12 \times 10$	4×4	$W_{\alpha_{2,4}}$	$12 \times 12 \times 10$	2×2
$W_{\alpha_{2,5}}$	$12 \times 12 \times 10$	1×1	$W_{\alpha_{2,6}}$	$12 \times 12 \times 10$	2×2
$W_{\alpha_{2,7}}$	$12 \times 12 \times 10$	2×2	$W_{\alpha_{2,8}}$	$12 \times 12 \times 3$	1×1
$W_{\alpha_{2,9}}$	$12 \times 12 \times 3$	1×1			
W_{β_1}	$12 \times 12 \times 10$	2×2	W_{β_2}	$12 \times 12 \times 10$	1×1
W_{β_3}	$12 \times 12 \times 10$	2×2	W_{β_4}	$12 \times 12 \times 10$	2×2
W_{β_5}	$12 \times 12 \times 10$	4×4	W_{β_6}	$12 \times 12 \times 10$	2×2
W_{β_7}	$12 \times 12 \times 10$	2×2	W_{β_8}	$12 \times 12 \times 10$	8×8
W_{β_9}	$12 \times 12 \times 10$	4×4	$W_{\beta_{10}}$	$12 \times 12 \times 10$	2×2
$W_{\beta_{11}}$	$12 \times 12 \times 3$	1×1	$W_{\beta_{12}}$	$12 \times 12 \times 3$	1×1
$W_{\theta_{1,1}}$	$9 \times 9 \times 30$	2×2	$W_{\theta_{1,2}}$	$9 \times 9 \times 30$	1×1
$W_{\theta_{1,3}}$	$9 \times 9 \times 30$	2×2	$W_{\theta_{1,4}}$	$9 \times 9 \times 30$	2×2
$W_{\theta_{1,5}}$	$9 \times 9 \times 30$	4×4	$W_{\theta_{1,6}}$	$9 \times 9 \times 30$	2×2
$W_{\theta_{1,7}}$	$9 \times 9 \times 30$	2×2	$W_{\theta_{1,8}}$	$9 \times 9 \times 30$	8×8
$W_{\theta_{1,9}}$	$9 \times 9 \times 30$	4×4	$W_{\theta_{1,10}}$	$9 \times 9 \times 30$	2×2
$W_{\theta_{1,11}}$	$9 \times 9 \times 3$	1×1	$W_{\theta_{1,12}}$	$9 \times 9 \times 3$	1×1
$W_{\theta_{2,1}}$	$9 \times 9 \times 30$	2×2	$W_{\theta_{2,2}}$	$9 \times 9 \times 30$	2×2
$W_{\theta_{2,3}}$	$9 \times 9 \times 30$	4×4	$W_{\theta_{2,4}}$	$9 \times 9 \times 30$	2×2
$W_{\theta_{2,5}}$	$9 \times 9 \times 30$	1×1	$W_{\theta_{2,6}}$	$9 \times 9 \times 30$	2×2
$W_{\theta_{2,7}}$	$9 \times 9 \times 30$	2×2	$W_{\theta_{2,8}}$	$9 \times 9 \times 3$	1×1
$M_{\theta_{2,9}}$	$9 \times 9 \times 10$	1×1	$W_{\theta_{2,10}}$	$9 \times 9 \times 3$	1×1
W_{ϕ_1}	$9 \times 9 \times 30$	2×2	W_{ϕ_2}	$9 \times 9 \times 30$	1×1
W_{ϕ_3}	$9 \times 9 \times 30$	2×2	W_{ϕ_4}	$9 \times 9 \times 30$	2×2
W_{ϕ_5}	$9 \times 9 \times 30$	4×4	W_{ϕ_6}	$9 \times 9 \times 30$	2×2
W_{ϕ_7}	$9 \times 9 \times 30$	2×2	W_{ϕ_8}	$9 \times 9 \times 30$	8×8
W_{ϕ_9}	$9 \times 9 \times 30$	4×4	$W_{\phi_{10}}$	$9 \times 9 \times 30$	2×2
$W_{\phi_{11}}$	$9 \times 9 \times 3$	1×1	$W_{\phi_{12}}$	$9 \times 9 \times 3$	1×1

D.1.2 PSNR and ELBO versus Number of Examples

The set of experiments giving results shown in figure 6 is carried out on a 32×32 down-sampled version of the CelebA data set. Images are blurred with a Gaussian PSF having a standard deviation of 2 pixels. As before, the standard CVAEs are trained with the K image-observation pairs alone. The proposed framework is then applied in each condition, exploiting the same K paired images and observations, $L = 100,000$ unobserved target examples and an inaccurate observation model. Two different inaccurate observation models are used; a more accurate one with 10% under-estimation of PSF width and noise level and a less accurate one, having 40% under-estimation. After training each model, reconstructions are performed with $2,000$ test examples and two quantitative measures are extracted: (i) the average peak signal to noise ration (PSNR) between the pseudo-maximum reconstructions and the original images and (ii) the evidence lower bound (ELBO). The former serves as a measure of deterministic performance, giving an index of similarity between the ground truth and the most likely reconstruction. The latter is a measure of probabilistic performance, as it approximates the likelihood assigned by the model to the ground truths and consequentially is an index of how well the distribution of solutions to the inverse model is captured.

The forward multi-fidelity model used in the proposed methods is built with the simple fully connected structures of figure 16. The comparison CVAE and the inverse models are identical and are built with the fully connected structures of figure 18. Multi-fidelity forward models were built to have 300 hidden units in all deterministic layers, while the latent variable w was chosen to be 100-dimensional. The inverse models, both for the proposed framework and the comparative CVAE, were built with 2500 hidden units in the deterministic layers and latent variables z of 800 dimensions.

D.1.3 Qualitative Comparison with Alternative Training Strategies

The results of figure 7 were obtained by reconstructing from blurred 64×64 CelebA images, blurred with a Gaussian PSF having standard deviation of 4 pixels and additive gaussian noise, corresponding to a SNR of $16dB$. The inaccurate observation model was instead given a PSF with standard deviation of 3 pixels and no additive noise.

The multi-fidelity forward model used in these experiments was of the convolutional form shown in figure 17, with the filter structures reported in table 2.

D.1.4 Quantitative Comparison with Alternative Training Strategies

Four different degradation conditions where tested, applying the following four degradations to 32×32 CelebA and CIFAR10 images:

- **$\times$ 2 Down-Sampling.** The true transformation applied to the K observed images consists of a $\times 2$ down-sampling of the images in each dimension and a subsequent blurring with a PSF having standard deviation 1.4 pixels. The low-fidelity accessible model down-samples by 2, but does not apply any blurring afterwards (i.e. the source of inaccuracy in the known forward model derives from ignoring blurring).

- **Partial Occlusion.** In the true transformation applied to the K observed images, a rectangular section of 8×11 pixels is set to zero in a given position in all images.

The low-fidelity model places instead a 5×15 at random with a difference in central position of $dy = 2$ and $dx = -2$.

- **Gaussian Blurring, $\sigma_{PSF} = 2.5px$.** The true transformation blurs the images with a PSF having standard deviation $\sigma_{PSF} = 2.5$ pixels and additive Gaussian noise at $12dB$. The low-fidelity analytical model instead blurs the images with a PSF having standard deviation $\sigma_{PSF} = 1.5$ pixels and does not add any noise.

- **Gaussian Blurring, $\sigma_{PSF} = 1.5px$.** The true transformation blurs the images with a PSF having standard deviation $\sigma_{PSF} = 1.5$ pixels and additive Gaussian noise at $16dB$. The low-fidelity analytical model instead blurs the images with a PSF having standard deviation $\sigma_{PSF} = 1$ pixels and does not add any noise.

The forward multi-fidelity model used in the proposed methods is built with the simple fully connected structures of figure 16. All inversion models, competitive and proposed, were implemented with the fully connected version of the inverse model given in figure 18. Multi-fidelity forward models were built to have 300 hidden units in all deterministic layers, while the latent variable w was chosen to be 100-dimensional. The inverse models, both for the proposed framework and the comparative training methods, were built with 2500 hidden units in the deterministic layers and latent variables z of 800 dimensions.

D.2 Holographic Image Reconstruction

We provide here more details on the HIO algorithm. The HIO algorithm is a Fourier transform-based method for holographic reconstruction where some constraints are used as support. In our case we have access to the amplitude at the camera plane and we assume that the phase at the DMD plane is uniform accross all micromirrors. The HIO algorithm starts with a random guess of the phase of the recorded image at the camera, performs an inverse Fourier transform to obtain a guess of both amplitude and phase at the DMD plane, and replaces the obtained phase with a uniform phase (one of our constraints). At this point, further constraints are added e.g. there is only image information at the central $N \times M$ pixels of the image (with N, M being arbitrary). After that, a forward Fourier transform is performed and the corresponding amplitude is replaced by the image recorded by the camera. This process is repeated iteratively. The problem is that if the recorded image is saturated and down-sampled, the iterative process breaks after the first iteration. As a consequence, it is impossible for the algorithm to converge towards a solution close to the ground truth. This is precisely what we observe in Figure 10(c), where the HIO algorithm simply predicts spots at some positions.

D.3 ToF Diffuse Imaging

The comparative iterative method was taken from Lyons et al. (2019), reproducing exactly the main results therein. For the proposed variational method, only the first 15 frames of the recorded experimental video were used as observation, as most of the information is contained in the rising front of the signal and around the peak. Consequentially, the corresponding frames in the two simulations (high and low fidelity) were used to train the model. The forward multi-fidelity model for the proposed variational method was built

with the fully connected structures shown in figure 16, with all deterministic intermediate layers having 3000 hidden units and latent variables w having 100 dimensions. The inverse model was also constructed using fully connected structures, as shown in figure 18, with all deterministic intermediate layers having 1500 hidden units and latent variables z having 800 dimensions.

References

Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

Jonas Adler and Ozan Öktem. Deep Bayesian inversion. *arXiv preprint arXiv:1811.05910*, 2018.

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on medical imaging*, 38(2): 394–405, 2019.

S. Derin Babacan, Rafael Molina, and Aggelos K. Katsaggelos. Variational Bayesian blind deconvolution using a total variation prior. *IEEE Transactions on Image Processing*, 18 (1):12–26, 2009.

S. Derin Babacan, Rafael Molina, and Aggelos K. Katsaggelos. Variational Bayesian super resolution. *IEEE Transactions on Image Processing*, 20(4):984–999, 2011.

David A. Barmherzig, Ju Sun, Po-Nan Li, Thomas Joseph Lane, and Emmanuel J. Candès. Holographic phase retrieval and reference design. *Inverse Problems*, 2019.

Maria J. Bayarri, James O. Berger, Rui Paulo, Jerry Sacks, John A. Cafeo, James Cavendish, Chin-Hsu Lin, and Jian Tu. A framework for validation of computer models. *Technometrics*, 49(2):138–154, 2007.

Mario Bertero and Patrizia Boccacci. *Introduction to inverse problems in imaging*. CRC press, 1998.

José M. Bioucas-Dias, Mario A.T. Figueiredo, and João Pedro Oliveira. Total variation-based image deconvolution: a majorization-minimization approach. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 2, pages II–II. IEEE, 2006.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. *Proceedings of the 34th International Conference on Machine Learning, PMLR*, pages 537–546, 2017.

Emmanuel J. Candes, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8): 1207–1223, 2006.

Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2017.

Jen-Hao Rick Chang, Chun-Liang Li, Barnabas Poczos, B. V. K. Vijaya Kumar, and Aswin C. Sankaranarayanan. One network to solve them all-solving linear inverse problems using deep projection models. In *ICCV*, pages 5889–5898, 2017.

Yunjin Chen, Wei Yu, and Thomas Pock. On learning optimized reaction diffusion processes for effective image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5261–5269, 2015.

J. Andrés Christen and Colin Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical statistics*, 14(4):795–810, 2005.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

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

Michael Egmont-Petersen, Dick de Ridder, and Heinz Handels. Image processing with neural networks – a review. *Pattern recognition*, 35(10):2279–2301, 2002.

Helen Mary Louise Faulkner and J.M. Rodenburg. Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm. *Physical review letters*, 93(2):023903, 2004.

James R. Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

Mário A.T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007.

Dani Gamerman and Hedibert F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

Ralph W. Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pages 9154–9164, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Alexander Jesacher, Christian Maurer, Andreas Schwaighofer, Stefan Bernet, and Monika Ritsch-Marte. Full phase and amplitude control of holographic optical tweezers with high efficiency. *Optics express*, 16(7):4479–4486, 2008.

Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346, 2008.

Huabei Jiang. *Diffuse optical tomography: principles and applications.* CRC press, 2018.

Sherena G. Johnson. NIST Special Database 30. 2010.

Jari Kaipio and Erkki Somersalo. Statistical inverse problems: discretization, model reduction and inverse crimes. *Journal of computational and applied mathematics*, 198(2):493–504, 2007.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Durk P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 449–458, 2016.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017.

June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

Aristidis Likas and Nikolas P. Galatsanos. A variational approach for Bayesian blind image deconvolution. *IEEE Transactions on Signal Processing*, 52(8):2222–2233, 2004.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

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

Ashley Lyons, Francesco Tonolini, Alessandro Boccolini, Audrey Repetti, Robert Henderson, Yves Wiaux, and Daniele Faccio. Computational time-of-flight diffuse optical tomography. *Nature Photonics*, 13:575–579, 2019.

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Filipe Magalhães, Francisco M. Araújo, Miguel V. Correia, Mehrdad Abolbashari, and Faramarz Farahi. Active illumination single-pixel camera based on compressive sensing. *Applied optics*, 50(4):405–414, 2011.

Alberto Malinverno. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3):675–688, 2002.

S Marchesini, H.N. Chapman, S.P. Hau-Riege, R.A. London, A. Szoke, H. He, M.R. Howells, H. Padmore, R. Rosen, J.C.H. Spence, and U. Weierstall. Coherent X-ray diffractive imaging: applications and limitations. *Optics Express*, 11(19):2344–2353, 2003.

M. Mardani, E. Gong, J. Y. Cheng, S. S. Vasanawala, G. Zaharchuk, L. Xing, and J. M. Pauly. Deep generative adversarial networks for compressed sensing automates MRI. *IEEE Transactions on Medical Imaging*, 32(1):167–179, Jan 2019.

James Martin, Lucas C Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

Youssef Marzouk and Dongbin Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847, 2009.

Kiyotoshi Matsuoka. Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):436–440, 1992.

Michael T. McCann, Kyong Hwan Jin, and Michael Unser. A review of convolutional neural networks for inverse problems in imaging. *IEEE Signal Processing Magazine*, 34(6):85–95, November 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Ali Mohammad-Djafari. Bayesian inference tools for inverse problems. In *AIP Conference Proceedings*, volume 1553, pages 163–170. AIP, 2013.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Hybrid models with deep and invertible features. In *Proc. 36th International Conference on Machine Learning (ICML), Long Beach, California, PMLR 97.*, 2019.

Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition, 107501*, 107, Nov .

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, volume 2, page 7, 2017.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learnin (ICML), PMLR 70*, pages 2642–2651, 2017.

Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80*, 2018.

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.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

Patrick Putzky and Max Welling. Recurrent inference machines for solving inverse problems. *arXiv preprint arXiv:1706.04008*, 2017.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Ludwig Ritschl, Frank Bergner, Christof Fleischmann, and Marc Kachelrieß. Improved total variation-based CT image reconstruction applied to clinical data. *Physics in Medicine & Biology*, 56(6):1545, 2011.

Yair Rivenson, Yibo Zhang, Harun Günaydın, Da Teng, and Aydogan Ozcan. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light: Science & Applications*, 7(2):17141, 2018.

Yair Rivenson, Yichen Wu, and Aydogan Ozcan. Deep learning in holography and coherent imaging. *Light: Science & Applications*, 8(1):1–8, 2019.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

Jean-Luc Starck, Mai K. Nguyen, and Fionn Murtagh. Wavelets and curvelets for image deconvolution: a combined approach. *Signal processing*, 83(10):2279–2283, 2003.

Baoqing Sun, Matthew P. Edgar, Richard Bowman, Liberty E. Vittert, Stuart Welsh, Adrian Bowman, and Miles J. Padgett. 3D computational imaging with single-pixel detectors. *Science*, 340(6134):844–847, 2013.

Panagiotis Tsilifis, Ilias Bilionis, Ioannis Katsounaros, and Nicholas Zabaras. Computationally efficient variational approximations for Bayesian inverse problems. *Journal of Verification, Validation and Uncertainty Quantification*, 1(3):031004, 2016.

Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G. Bawendi, and Ramesh Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature communications*, 3:745, 2012.

Curtis R. Vogel. *Computational methods for inverse problems*, volume 23. SIAM, 2002.

Li Xu, Jimmy S.J. Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

Junfeng Yang and Yin Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1):250–278, 2011.

Yibo Yang and Paris Perdikaris. Conditional deep surrogate models for stochastic, high-dimensional, and multi-fidelity systems. *Computational Mechanics*, 64:417–434, 2019.

K.M. Yoo, Feng Liu, and R.R. Alfano. When does the diffusion approximation fail to describe photon transport in random media? *Physical review letters*, 64(22):2647, 1990.

Chen Zhang and Bangti Jin. Probabilistic residual learning for aleatoric uncertainty in image restoration. *arXiv preprint arXiv:1908.01010*, 2019.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017.