VARIATIONAL LEARNING FOR INVERSE PROBLEMS

FRANCESCO TONOLINI

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE

COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

JUNE 2021

Abstract

Machine learning methods for solving inverse problems require uncertainty estimation to be reliable in real settings. While deep variational models offer a computationally tractable way of recovering complex uncertainties, they need large supervised data volumes to be trained, which in many practical applications requires prohibitively expensive collections with specific instruments. This thesis introduces two novel frameworks to train variational inference models for inverse problems, in semi-supervised and unsupervised settings respectively. In the former, a realistic scenario is considered, where few experimentally collected supervised data are available, and analytical models from domain expertise and existing unsupervised data sets are leveraged in addition to solve inverse problems in a semi-supervised fashion. This minimises the supervised data collection requirements and allows the training of effective probabilistic recovery models relatively inexpensively. This novel method is first evaluated in quantitative simulated experiments, testing performance in various controlled settings and compared to alternative techniques. The framework is then implemented in several real world applications, spanning imaging, astronomy and human-computer interaction. In each real world setting, the novel technique makes use of all available information for training, whether this is simulations, data or both, depending on the task. In each experimental scenario, state of the art recovery and uncertainty estimation were demonstrated with reasonably limited experimental collection efforts for training. The second framework presented in this thesis approaches instead the challenging unsupervised situation, where no examples of ground-truths are available. This type of inverse problem is commonly encountered in data pre-processing and information retrieval. A variational framework is designed to capture the solution space of inverse problem by using solely an estimate of the observation process and large ensembles of observations examples. The unsupervised framework is tested on data recovery tasks under the common setting of missing values and noise, demonstrating superior performance to existing variational methods for imputation and de-noising with different real data sets. Furthermore, higher classification accuracy after imputation are shown, proving the advantage of propagating uncertainty to downstream tasks with the new model.

Acknowledgements

Most of the work presented in this thesis was carried out under funding from Amazon and the University of Glasgow together with EPSRC grants EP/M01326X/1, EP/T00097X/1 (*QuantIC*, the *UK Quantum Technology Hub in Quantum Enhanced Imaging*) and EP/R018634/1 (*Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics*). For the work of chapter 3 and the imaging applications of sections 4.1 and 4.2, Daniele Faccio is supported by the Royal Academy of Engineering under the *Chairs in Emerging Technologies* scheme, Jack Radford was supported by the EPSRC *CDT in Intelligent Sensing and Measurement*, Grant Number EP/L016753/1 and Alex Turpin was supported by a Lord Kelvin Adam Smith Fellowship from the University of Glasgow. The work presented in section 5.2 was carried out in a project collaboration together with Google ATAP, separate from my Ph.D. funding, and we acknowledge funding from Google for support for this section.

Table of Contents

1 Introduction 1

1.1 Thesis Overview 2

1.1.1 Publications Summary 4

2 Background and Related Work 5

2.1 Linear models and Hand-crafted Priors 5

2.1.1 Maximum a Posteriori Inference 6

2.1.2 Bayesian Inference 7

2.2 Supervised and Semi-Supervised Settings 8

2.2.1 Iterative Inference with Learned Priors 10

2.2.2 Conditional Generative Models 12

2.2.3 Semi-Supervised Conditional Generative Models 13

2.3 Machine Learning for Solving Inverse Problems in Unsupervised Settings . 14

2.3.1 Unsupervised Bayesian Recovery 14

2.4 Multi-Fidelity Bayesian Models 16

3 Semi-Supervised Variational Inference for Inverse Problems 17

3.1 Problem Description 20

3.1.1 The Bayesian Inverse Problem 20

3.1.2 Information Available in Semi-Supervised Settings 21

3.2 Multi-Fidelity Forward Modelling 22

3.2.1 Variational Inverse Model 25

3.3 Simulated Experiments 30

3.3.1 Qualitative Comparison with Standard training of CVAEs 30

3.3.2 Quantitative Comparison with Standard training of CVAEs 36
3.3.3 Qualitative Comparison with alternative Training 39
3.3.4 Quantitative Comparison with alternative Training 41

4 Enabled Applications in Physics 45

4.1 Holographic image reconstruction 46
4.1.1 Experimental set-up and Data 47
4.1.2 Reconstruction 48
4.2 Imaging Through Highly Scattering Media 49
4.2.1 Physical Experiment 50
4.2.2 Training Data and Models 51
4.2.3 Results 53
4.3 Gravitational Wave Astronomy 55
4.3.1 Method 55
4.3.2 Experiments 56

5 Applications in Human-Computer Interaction 61

5.1 Capacitive Sensing for HCI 62
5.1.1 Experimental Set-up 62
5.1.2 Simulations and Multi-Fidelity Forward Model 63
5.1.3 Inverse Model and Reconstructions 65
5.1.4 Discussion 67
5.2 Radar for Gesture Interaction with Google Soli 73
5.2.1 Experimental Acquisition 73
5.2.2 Analytical Forward Model 75
5.2.3 Data Sets 76
5.2.4 Multi-Fidelity Forward Model 77
5.2.5 Experimental Reconstruction Results 81
5.2.6 Discussion 89

6 Variational Inference for Unsupervised Inverse Problems 90

6.1 Method 92

6.1.1 VAEs and the Posterior Degeneracy Problem 94

6.1.2 Data Modelling View of the Problem 96

6.1.3 Separating Posterior and Prior: The Tomographic Auto-Encoder . . 96

6.2 Experiments 104

6.2.1 Posterior Recovery 105

6.2.2 Posterior Recovery in Different Conditions 107

6.2.3 Downstream Tasks 110

6.2.4 Missing Values in the NYU Depth Maps 111

7 Conclusion 114

7.1 The VICI Framework 114

7.2 The TAE Framework 116

A Semi-Supervised Experiments 118

A.1 Experimental Conditions 118

A.1.1 Holographic Image Reconstruction 118

A.1.2 ToF Diffuse Imaging 118

A.2 Details of the Models' Architectures 119

A.2.1 Multi-Fidelity Forward Model 119

A.2.2 Variational Inverse Model 119

B Unsupervised Experiments 120

B.1 Evaluation ELBO 120

B.2 Experimental Conditions 120

B.2.1 Posterior Recovery 120

B.2.2 Classification Experiments 121

B.2.3 Training Conditions 121

B.2.4 NYU Rooms Experiments 122

Bibliography 123

List of Tables

3.1 Filter banks used in the qualitative experiments of section 3.3.1. 34

3.2 Probabilistic recovery performance evaluation. 43

3.3 Statistical significance of probabilistic performance. 44

5.1 Mean reconstruction error with capacitive data. 66

5.2 Correlation between mean error and retrieved standard deviation. 67

5.3 Reconstruction performance metrics with the coarse hand motions. 85

5.4 Reconstruction performance metrics with the fine hand motions. 85

6.1 Unsupervised missing value imputation and de-noising 107

List of Figures

2.1 Constrained Minimisation. 7

2.2 Learning inverse mappings. 9

2.3 Learning penalty functions for solving inverse problems. 11

2.4 Constrained minimisation with generative models. 12

2.5 Conditional generative models. 13

2.6 Reconstruction in unsupervised settings with VAEs. 15

3.1 The VICI framework for semi-supervised inverse problems. 19

3.2 Multi-fidelity forward modelling. 24

3.3 Variational inverse model. 28

3.4 VICI graphical models. 28

3.5 Forward model convolutional parametric distributions' structures. 32

3.6 Inverse model convolutional parametric distributions' structures. 33

3.7 Qualitative comparison between standard CVAE and proposed framework. . 35

3.8 Forward model fully connected parametric distributions' structures. 37

3.9 Inverse model fully connected parametric distributions' structures. 37

3.10 Quantitative comparison between standard CVAE and proposed framework. 38

3.11 Posterior recovery examples from blurred images. 40

3.12 Mean recovery performance evaluation. 42

4.1 Experimental set up used for holographic image reconstruction. 46

4.2 Reconstructions from experimental Fourier intensity image data. 47

4.3 Image posterior recovery from phase-less measurements. 48

4.4 Testing varying ill-poseness. 50

4.5 Experimental set up for imaging through scattering media. 51

4.6 Simulated and real measurements from the time of flight (ToF) Camera. . . 52
4.7 Reconstructions from experimental ToF videos. 53
4.8 GW reconstruction qualitative results. 58
4.9 GW reconstruction quantitative results. 60

5.1 Capacitive sensing set-up and observation process. 63
5.2 Simulations of capacitive sensing. 64
5.3 Multi-fidelity forward model for capacitive sensing. 64
5.4 Reconstructions of finger poses' parameters. 65
5.5 Reconstructions with standard CVAE and VICI framework. 69
5.6 Uncertainty recovered with standard CVAE. 70
5.7 Histograms of uncertainty recovered with standard CVAE. 71
5.8 Standard CVAE vs. VICI in capacitive sensing. 72
5.9 Soli radar experimental set-up 74
5.10 Passive Optitrack markers on the hand. 74
5.11 Simulated Soli CRDs. 77
5.12 Soli and Optitrack data formats. 78
5.13 Data pre-processing for training the multi-fidelity forward model. 79
5.14 Examples of multi-fidelity generated frames. 80
5.15 Reconstruction samples for the coarse hand motions. 82
5.16 Reconstruction samples for the fine hand motions. 83
5.17 Reconstruction of the hand while pushing down with the index finger. . . . 86
5.18 Reconstruction of the hand while pushing down with the ring finger. . . . 87
5.19 Hand tracking average accuracy. 88
5.20 Individual tracker positioning average accuracy. 88

6.1 Bayesian data recovery with the TAE. 91
6.2 Training LVMs for data recovery. 93
6.3 MNIST data recovery from missing entries and noise. 105
6.4 Recovery from structured missings. 108
6.5 Noiseless missing value imputation. 109

6.6 De-noising performance on MNIST. 109

6.7 Robustness to hyper-parameters. 110

6.8 Propagating uncertainty to a classification task. 110

6.9 Classification accuracy after imputation. 111

6.10 Unsupervised missing value imputation on NYU rooms. 112

Chapter 1

Introduction

Many things that we are interested in exploring, studying and understanding are not directly observable to us. Distant astronomical objects, particles too small to record images of and even our own inner body structure, are all objects we cannot see directly. However, through substantial engineering efforts, we sometimes get partial and indirect observations of these objects, such as emitted radiation intensities or projective measurements. Inverse problems are the processes through which we use these observations to retrieve estimates of the hidden objects. Many problems in computing, scientific research and engineering can be elegantly posed as inverse problems. As a result, solving inverse problems has been a major area of study in mathematics and computer science for decades and is the basis of several technologies, especially for sensing and information retrieval.

Traditionally, inverse problems are approached by constructing analytical priors and observation models, reliant on domain expertise, and finding solutions which satisfy well both observations and prior knowledge [1, 2]. Recently, as machine learning models developed and data became increasingly available, learning approaches to solving inverse problems emerged, providing largely superior performance and entirely new capabilities [3]. With enough data available, one can train an inference model to map observations to hidden objects, drawing information directly from empirical experience, rather than analytically specified models and priors.

However, machine learning has important shortcomings and limitations that may not be captured by common inverse problem solving benchmarks in controlled environments. Their reliance on large data sets makes them expensive to apply in real world scenarios [3, 4, 5, 6, 7]. In addition, they often return compelling retrievals, because they are trained to return realistic reconstructions from the training set [8, 9, 10]. This makes it difficult to identify mistakes, with potentially critical consequences. This thesis studies and addresses these limitations, designing learning frameworks for probabilistic machine learning models in two general scenarios.

1.1 Thesis Overview

In chapter 2, the general modelling of inverse problems is presented and classical methods of solving them are reviewed, summarising the main ideas of iterative maximum a posteriori (MAP) inference and Bayesian inference. The main advances and types of models for solving inverse problems with machine learning methods are then summarised, both in supervised and unsupervised settings.

In chapter 3, a semi-supervised situation is studied, where examples of paired measurements and ground-truth targets are scarce, while examples of targets alone are abundant and some model of the observation process is available. This scenario is rather common in sensing and imaging, as observations are obtained with the instruments and set-up of interest and therefore are unique to the specific task at hand. Conversely, targets of interest tend to be the same across many tasks. Examples include, natural images, skeletal tracking, medical images and others. This means that unlabelled examples are often available, or if they need to be collected, they are relevant to different observation systems and therefore more broadly applicable. In addition, for these physical observation processes, analytical approximate models of how measurements are obtained from targets often exist, or can be built. Such models constitute an additional source of information learning can benefit from. The framework described in chapter 3 draws from all of these sources of information to train a probabilistic learning model that can adequately capture the solution space of an inverse problem with minimal supervised data collection requirements. This method was presented in the article *Variational Inference for Computational Imaging Inverse Problems* [11], co-authored with Jack Radford, Alex Turpin, Daniele Faccio and Roderick Murray-Smith. The novel method is first tested in simulated experiments in chapter 3, where all experimental conditions are controlled to test specific aspects of the frameworks and carry out comparisons with existing methods. In chapters 4 and 5, the proposed framework is then applied to several practical applications, demonstrating the effectiveness of the new techniques and their impact in different domains.

In chapter 4, the framework of chapter 3 is implemented to solve three computational imaging inverse problems in Physics. In section 4.1, holographic image reconstruction is performed with a variational learning model trained with the framework of chapter 3. This task consists in recovering images which went through a physical Fourier transformation from phase-less measurements. In section 4.2, objects embedded within a highly diffusive medium are reconstructed from spatially and temporally resolved time-of-flight measurements at the medium surface. The method of chapter 3 is used to train a variational reconstruction model to recover embedded shapes using two different simulations, with different cost-fidelity trade-offs. In section 4.3, components from the novel framework are used to reconstruct parameters of astronomical bodies' collisions, such as location in the sky and masses, from gravitational waves measurements recorded on earth. In this domain, the simulation of gravitational waves signal is rather accurate and is accepted by the community as a sufficiently high-fidelity approximation. Therefore, in this domain, the inverse model is directly trained with the analytical simulator as forward model. The application of the models here presented to this domain was led by Hunter Gabbard and Chris Messenger and co-authored with them, Ik Siong Heng and Roderick Murray-Smith [12].

In chapter 5, the framework of chapter 3 is applied in two human-computer interaction (HCI) settings, reconstructing users' physical interactions from limited sensor readings. In section 5.1, the scenario of finger pose recovery from capacitive screen reading is investigated. This is a simple, but important and representative HCI example, where capacitive readings on a 10×16 screen are used to infer the position and angle of incidence of the user's finger. This work was led by Roderick Murray-Smith and John H. Williamson, and co-authored with Andrew Ramsay, Simon Rogers and Antoine Loriette [13]. In section 5.2, a more complex setting is investigated, where hand gestures are reconstructed from radar signals recorded with the Google Soli sensor. The framework of chapter 3 is used to train a probabilistic model to reconstruct the gestures from Soli's signals using supervised and unsupervised data, along with a physical model of the radar sensing process. This work was done in collaboration with the Google ATAP Soli team. The physical acquisitions were carried out by Andrew Ramsay and the project was overseen by Roderick Murray-Smith and Nick Gillian (Google).

In chapter 6, the completely unsupervised scenario is studied, where no example of ground-truth is available, but only a large ensemble of observations and a model describing the observation process. This situation occurs often in data cleaning and pre-processing and information retrieval. In chapter 6, it is shown how standard training frameworks for variational models often fail to capture uncertainty in the inverse model. Instead, a novel framework, called the *reduced entropy condition* method, is proposed. The proposed framework is demonstrated to have greatly improved ability to capture the uncertainty of reconstruction and capture the inverse model solution space. The framework is tested in the particular situations of missing value imputation and de-noising, as these are the most commonly encountered in this unsupervised data recovery scenario. This framework was presented in the article *Tomographic Auto-Encoder: Unsupervised Bayesian Recovery of Corrupted Data* [14], co-authored with Pablo G Moreno, Andreas Damianou and Roderick Murray-Smith.

1.1.1 Publications Summary

The material that constitutes this thesis was presented in several publications. These are summarised below.

- [11] Tonolini, F., Radford, J., Turpin, A., Faccio, D., & Murray-Smith, R. (2020). *Variational inference for computational imaging inverse problems*. Journal of Machine Learning Research (JMLR), 21(179), 1-46.
- [12] Gabbard, H., Messenger, C., Heng, I. S., Tonolini, F., & Murray-Smith, R. (2019). *Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy*, Nature Physics, vol. 18, no. 1, pp. 112–117, 2022.
- [14] Tonolini, F., Moreno, P. G., Damianou, A., & Murray-Smith, R. (2021, May). *Tomographic Auto-Encoder: Unsupervised Bayesian Recovery of Corrupted Data*. In International Conference on Learning Representations (ICLR).
- [13] Murray-Smith, R., Williamson, J. H., Ramsay, A., Tonolini, F., Rogers, S. & Loriette, A. (2021) *Forward and Inverse models in HCI:Physical simulation and deep learning for inferring 3D finger pose*, arXiv preprint arXiv:2109.03366, 2021.

Chapter 2

Background and Related Work

An inverse problem is broadly identified as one in which a quantity or object of interest is not directly observed, but rather needs to be inferred algorithmically from one or more measurements [15, 1]. Many recovery tasks fall within this definition, such as de-convolution [16, 17], computed tomography (CT) [18], structured illumination imaging [19] and information retrieval [20, 21]. Traditionally, such retrieval 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 [15, 1, 2]. In the following subsections, the main advances in solving inverse problems are reviewed, starting from linear models and user defined regularisation functions, to then focus on more recent applications of learning based methods.

2.1 Linear models and Hand-crafted Priors

In many inverse problems, the forward observation model can be approximately described as a linear operator $A \in \mathbb{R}^{m \times n}$ and some independent noise $\epsilon \in \mathbb{R}^m$ [1, 22], such that the measurements y are assumed to be generated as

$$y = Ax + \epsilon. \tag{2.1}$$

The noise ϵ is often modelled with simple statistics, such as Gaussian or Bernoulli distributions, depending on the observation setting. 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, as projective measurements are usually close to linear in many different settings, from imaging to matrix completion.

The difficulty in retrieving the observed signal x from an observation y in this context derives from the fact that in many inverse problems of interest the operator A is often poorly conditioned and consequentially the resulting inverse problem is ill-posed. Put differently, the inverse A^{-1}, or pseudo inverse $(A^TA)^-1$ 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

$$\arg\min_x \quad \frac{1}{2}||Ax-y||^2 + \lambda h(x), \tag{2.2}$$

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. For example, 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 of the form $h(x) = ||Wx||_1$, where W is a unitary operator which maps x to a sparse basis [22, 23, 24]. A second notable example is that of low rank in matrix completion settings $h(x) = ||sv(x)||_1$, where $sv(x)$ indicates the singular values of x reshaped in a matrix form [25, 26]. For such choices of penalty, and other common ones, the objective function of equation 2.2 is convex. This makes the optimisation problem solvable with a variety of efficient methods [22, 27] and provides theoretical guarantees on the recoverability of the solution [28].

The aforementioned framework has been widely applied to solve inverse problems. 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 [29]. Various more complex sensing models can also be cast as linear operators, leading to the use of constrained optimisation in several systems that rely on ill-posed observations, such as sparse CT [18], single pixel photography [30] and imaging of objects hidden from view [31]. Figure 2.1 shows an example of reconstruction from a CT scan using constrained minimisation with the total variation (TV) norm as regulariser. The reconstruction greatly improves when inserting domain prior knowledge to regularise reconstruction. In this case, as for many imaging applications, the assumption is that the target image is smooth and

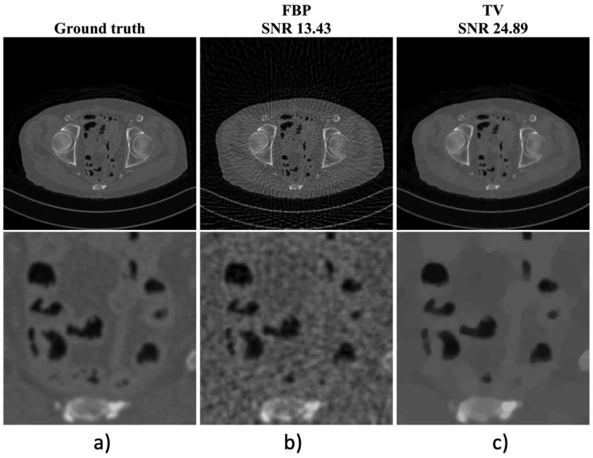


Figure 2.1: Example of reconstruction using constrained minimisation on a CT scan taken from [32]. The bottom row shows enlarged central sections of the images on the top row. a) ground truth, b) unconstrained minimisation reconstruction, c) constrained minimisation with TV norm.

therefore displays a low TV norm.

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

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{2.3}$$

where $p(y) = \int p(y|x)p(x)dx$ is the marginal likelihood of measurements. If the observation likelihood is assumed to be a Gaussian PDF $p(y|x) = \exp(-||Ax-y||^2/2\sigma^2)$ and the prior is chosen to be an exponential distribution $p(x) = \exp(-\lambda h(x))$, the minimisation of equation 2.2 is equivalent to maximising the posterior probability of equation 2.3 with respect to x, which is the definition of MAP inference in Bayesian settings [33]. Estimating the full distribution of solutions $p(x|y)$ is generally much harder than simply finding its maximum through MAP inference. In fact, exact inference of the posterior is intractable for most inverse problems of interest.

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 [34, 33, 35, 36, 37]. In these approaches, the posterior PDF $p(x|y)$ is estimated through chains of samples, where each sample is chosen conditioned on the previous such that the distribution of samples as their number increases approaches the true target one [34]. Several choices of conditional sampling have been proposed for accurate estimation in different types of inverse problems [33, 35, 36, 37]. Despite their guarantees of convergence to accurate approximations, MCMC methods are often prohibitively expensive for many problems, such as imaging or information retrieval, as natural data of interest is often 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 $p(x|y)$ [38, 39]. The parameters of the approximate model are optimised to best match the intractable posterior [38, 39]. 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, especially in the context of computational imaging [40, 41, 42].

2.2 Machine Learning for Solving Inverse Problems in Supervised and Semi-Supervised Settings

The increasing availability of data sets and continuous advancements in learning inference models enabled new possibilities for solving inverse problems. Learning from real examples allows to derive retrieval models directly from empirical experience, instead of relying on analytically defined priors and observation processes. In some cases, this is done over a direct mapping between observations and targets, while in others it is limited to capturing prior knowledge about the target signals. The main classes of machine learning methods for solving inverse problems are reviewed below.

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 [3]. Many neural architectures of this type have been developed to perform different reconstruction tasks. For imaging and image processing tasks, convolutional neural networks are popular choices due to their ability to capture local pixel dependencies in images [43, 32]. 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

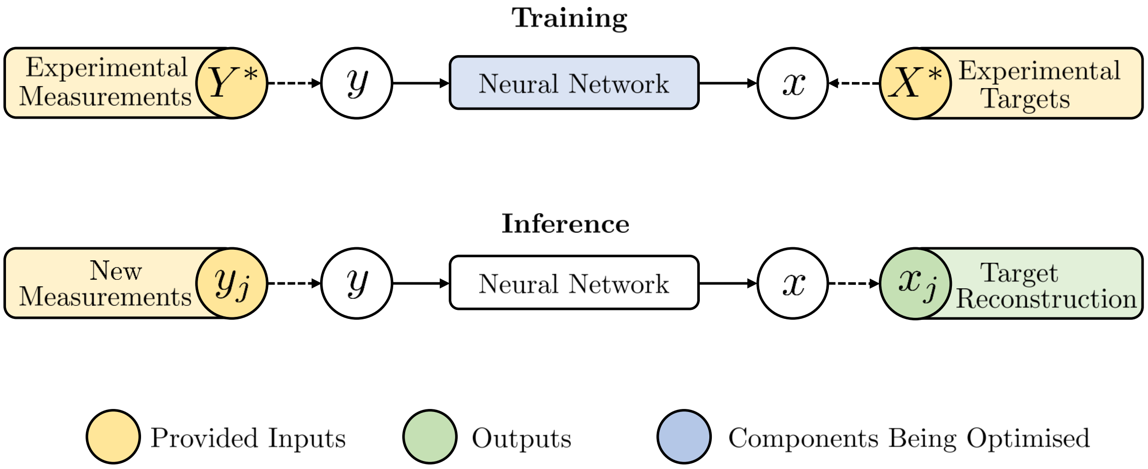


Figure 2.2: Learning inverse mappings. Schematic representation of the training (top) and inference (bottom) procedures typically adopted when directly applying neural networks to solving inverse problems. Empirically collected or simulated targets X^* and corresponding measurements Y^* are used to train a neural network. During inference, new measurements y_j are used as inputs to the trained neural network which returns an estimate of the corresponding target x_j as output.

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 [3]. Figure 2.2 schematically illustrates the general framework.

Directly learning inverse mappings retains a number of advantages compared to analytical methods. First, the model is trained with a set of ground-truths 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 [44, 45], as well as signal recovery from under determined projective measurements [46]. Neural networks have also been used to perform entirely new forms of imaging, learning from measured target-observations pairs in situations where the observation model or signal prior can not be explicitly defined and consequentially analytical methods are not applicable. Notable examples include inferring depth maps from RGB photographs [47], recovering natural images from hand drawings [48] and pose estimation with single cameras or radio reflections [49, 50]. However, learned inverse mappings for solving inverse problems 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.

2.2.1 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 [51]. 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

$$\arg\min_{x} \quad \frac{1}{2}||Ax - y||^2 + \lambda H(x), \tag{2.4}$$

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 [52]. 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 $H(x) = ||x - N(x)||$ [53, 54, 55]. Figure 2.3 schematically shows this framework. Machine learning has also been implemented to train optimisation methods to solve the minimisation of equation 2.4. In fact, the iterative update of the solution x through the optimisation procedure in these settings is often interpreted as a recurrent neural network [56, 57]. In such a way, the iterative inference precision is empirically adjusted to the specific inversion task, hence gaining in efficiency and accuracy [58].

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

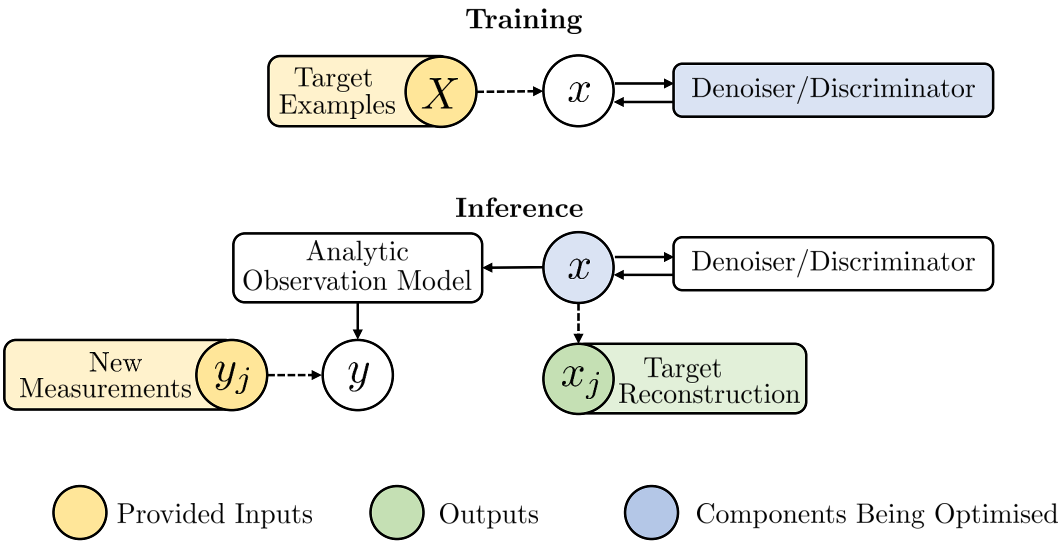


Figure 2.3: Illustration of learning penalty functions for solving inverse problems. During training, a denoiser or discriminator is trained with a data set of expected targets. During inference, the target reconstruction is obtained by optimising x to jointly minimise the cost function of the denoiser/discriminator and discrepancy with the observed measurement through an analytic observation model.

$$\arg\min_{z} \quad \frac{1}{2}||A \cdot G(z) - y||^2. \tag{2.5}$$

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 [51, 59, 60]. Figure 2.4 schematically illustrates this framework.

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 [52]. 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.

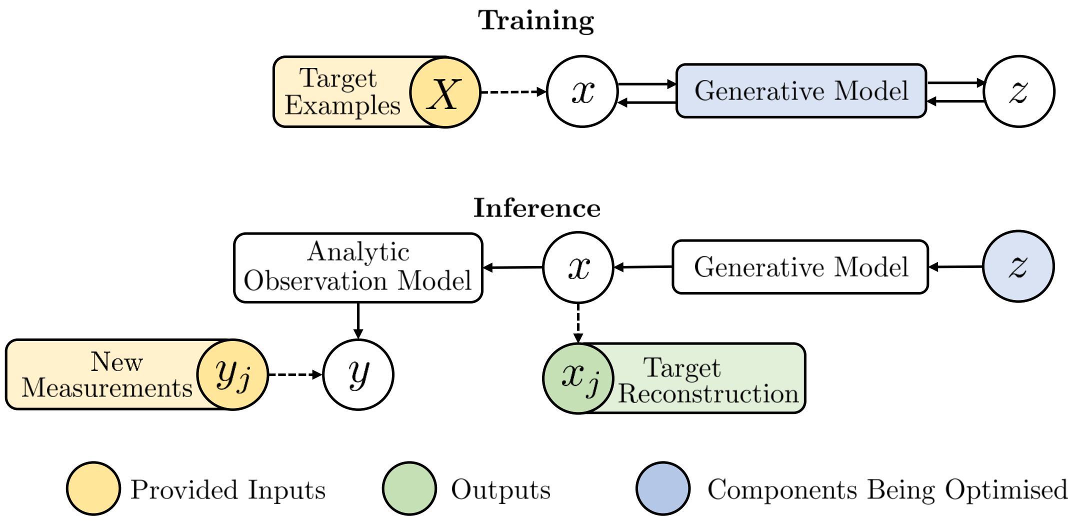


Figure 2.4: Illustration of constrained minimisation with generative models. During training, a latent variable generative model, such as a GAN or VAE, is trained with a data set of expected targets. During inference, the target reconstruction is obtained by optimising the latent variable z to maximise fidelity to the observed measurement through an analytic observation model.

2.2.2 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 and consequentially the expected reconstruction error can be estimated. These measures of uncertainty can be propagated through further processing steps, greatly improving the reliability of automated decisions or visualised in different ways to provide error descriptions upon human inspection. Figure 2.5 schematically illustrates this type of model.

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 tasks of interest, such as image reconstruction [61, 62, 63]. 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) [64, 65, 66].

Conditional generative models have been applied to perform a range of inference tasks, such as classification [65], generation conditioned on classes [66, 67], image-from-text inference [68, 69] and missing value imputation [70, 71]. Within computational imaging, they have

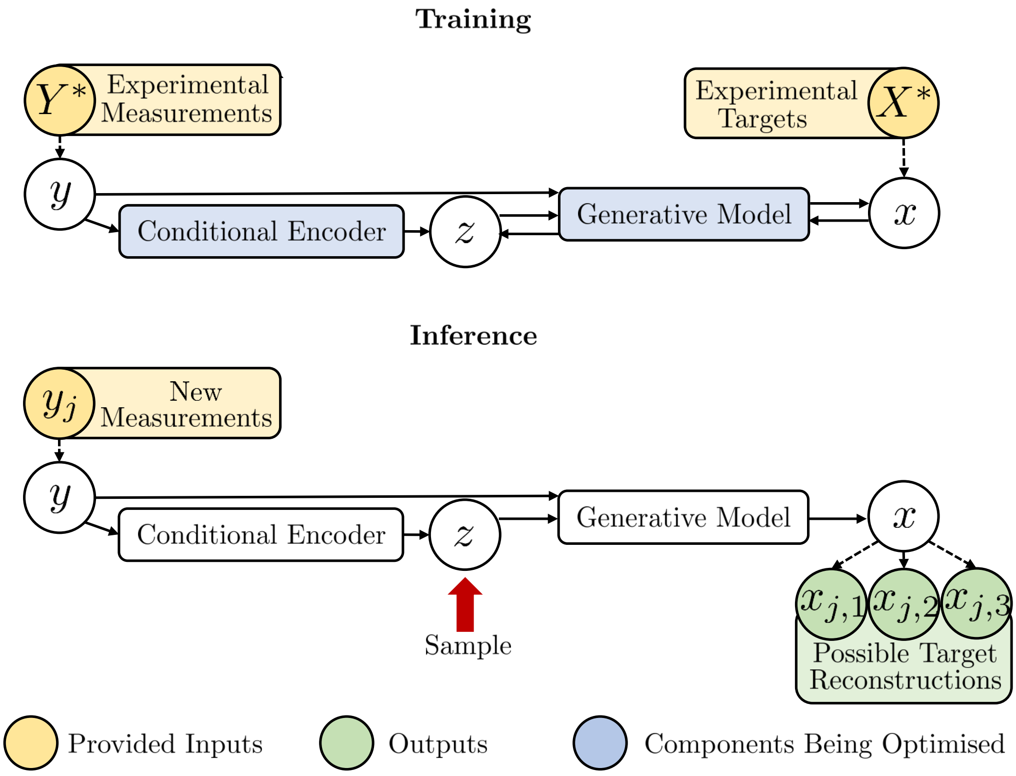


Figure 2.5: Illustration of conditional generative models. During training, the conditional latent variable generative model, such as a GAN or VAE, is trained with a data set of paired targets and observations used as conditions. During inference, the trained conditional encoder outputs a latent distribution. This distribution is sampled to obtain different latent variables z, which are then used as inputs to the generator. For each latent variable, the generator outputs a different realisation of the expected target.

been largely restricted to inference from simple deterministic observation, such as missing pixels or down-sampling [66, 72], with the exception of recent work by [8], in which a specifically designed GAN model is used to retrieve medical images from CT scans. The direct application of conditional generative models for solving inverse problems is challenging because of the large data volumes requirements. Conditional generative models, in their common form, need a large number of target-condition pairs to train upon. In many inverse problems settings, this translates to the need of obtaining a large number of sufficiently accurate target-observation examples, which are often unique to environmental conditions or instrumentation and hence expensive to collect.

2.2.3 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 [73, 74, 75]. They achieve this by introducing auxiliary models that map inputs to conditions and are trained jointly with the generator. The auxiliary model is usually a neural network itself and acts as an extra encoder, mapping targets to corresponding measurements or labels, which in turn are used by the generator as conditions or latent variables to generate samples. The standard encoder and generator are trained with the all data, supervised and unsupervised, as they only rely on the presence of targets. The auxiliary model is instead trained with the labelled portion of the data, as it maps targets to corresponding labels. In most cases, the auxiliary model is a classifier, which is trained jointly with a class-conditional generator.

In some sense, one of the main frameworks presented in this thesis belongs to this class of methods, as the forward model component plays an analogous role to the auxiliary model in these systems, with some important differences that are discussed in section 3.

2.3 Machine Learning for Solving Inverse Problems in Unsupervised Settings

A particularly challenging, yet common setting is that of unsupervised recovery. In these situations, no example of target ground-truths is available, but only large data sets of observations and some knowledge of the observation process. For instance, one such case is that of matrix completion in information retrieval [20, 25, 26]. In this setting, examples of complete vectors are often not available and one can only rely on a large set of vectors with missing values and knowledge of which entries are missing, e.g. the observation process.

2.3.1 Unsupervised Bayesian Recovery

Reconstructing posteriors in the unsupervised case is largely still an open problem. However, several tasks that fall within this definition have been recently approached with Bayesian machine learning methods. Arguably the most investigated is de-noising, i.e., given a noisy data set alone, we wish to train a model to return clean samples. Several works solve this problem by exploiting the natural tendency of neural networks to regularise outputs [76, 77, 78]. Other methods build LVMs that explicitly model the noise process in their decoder, retrieving clean samples upon encoding and generation [79, 80].

A second notable example is that of missing value imputation. Corrupted data corresponds to samples with missing entries. Recent works have explored the use of LVMs to perform imputation, both with GANs [81, 82, 83] and VAEs [84, 85, 86, 87]. In the former, the discriminator of the GAN is trained to distinguish real values from imputed ones, such that the generator is induced to synthesise realistic imputations. In the latter, the encoder of a VAE

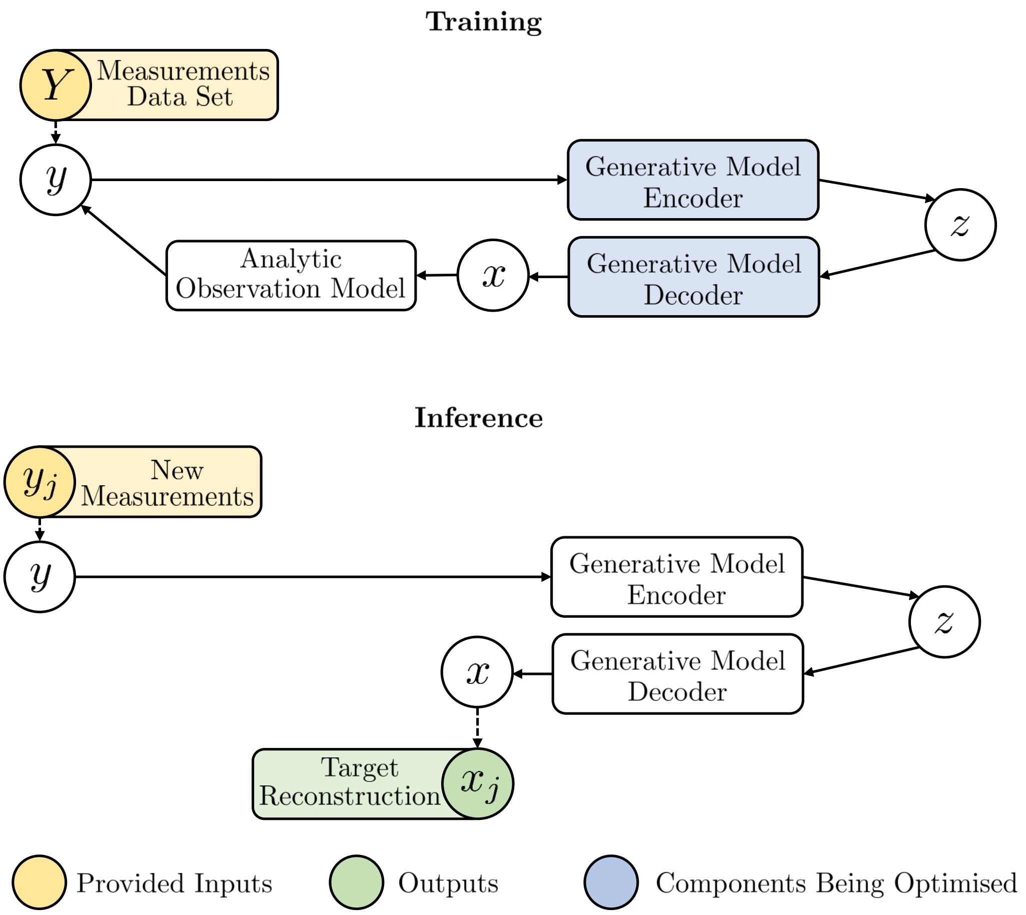


Figure 2.6: Illustration of general framework to perform reconstruction in unsupervised settings (VAE version). During training, the encoder returns latent space distributions from measurements and the decoder takes latent variables z as inputs and generates targets realisations. The generated targets are mapped back to observations through a known observation model, e.g. zeroing out missing entries, in order to optimise data likelihood. During inference, a new observation is passed through the encoder and then decoder to generate a reconstructed target.

maps incomplete samples to a latent space, to then generate complete samples. Successful unsupervised Bayesian missing value imputation has also been demonstrated with neural processes, where a global latent representation is learned to generate input-output models used to impute in each example [88].

Finally, Bayesian LVM methods have been used on other unsupervised tasks that can be cast as special cases of data recovery problems. Amongst these, we find Multi-view generation [89, 90], where the target clean data includes all views for each samples, but the observed data only presents subsets. Blind source separation can also be cast as a recovery problem and has been approached with GANs and VAEs [91, 92]. Figure 2.6 schematically show the general framework with a VAE architecture.

These models proved to be successful at reconstructing data in their specific domain. However, as part of the work in this thesis, it is shown in section 6.1.1 how exploiting a standard VAE inference structure, similarly to several of the aforementioned methods, often leads to posteriors of clean data that collapse on single estimates, sacrificing the probabilistic capability of LVMs.

2.4 Multi-Fidelity Bayesian Models

An important aspect of solving inverse problems with empirical learning methods treated throughout this thesis is the modelling of accurate forward observation processes. These are often estimated with some degree of accuracy with analytical models, but are better represented by the empirical data, where available. The availability of limited but very accurate forward process realisations through empirical data and readily available, but less accurate ones through analytical models is a setting often approach with a class of methods called multi-fidelity 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 [93]. 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 [94]. 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 [95].

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 [96, 97]. In many inverse problems settings explored throught this thesis, 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 [98] proposes to model high-fidelity data with conditional deep generative models, which are instead capable of scaling to the volumes and dimensionalities needed in many of these 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.

Chapter 3

Semi-Supervised Variational Inference for Inverse Problems

Solving inverse problems is one of the most important yet challenging forms of algorithmic information retrieval, with applications in Medicine, human-computer interaction (HCI), Astronomy and more. Bayesian machine learning methods are an attractive route to solve inverse problems, as they retain the advantages of learning from empirical data, while improving reliability by inferring uncertainty [8, 9]. However, fitting distributions with Bayesian models requires large sets of training examples [10]. This is particularly problematic in imaging and HCI 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 [3, 4]. 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 chapter introduces a novel framework to train conditional variational models for solving inverse problems leveraging in combination:

1. A minimal amount of experimentally acquired or numerically simulated ground truth target-observation pairs.
2. An inexpensive analytical model of the observation process from domain expertise.
3. 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 machine learning models. In this chapter, the novel 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 [73, 74, 75]. The framework is named *variational inference for computational imaging* (VICI), as it was initially developed and demonstrated for imaging problems in [11]. However, as it is demonstrated in chapters 4 and 5, it is broadly applicable to different types of inverse problems. Figure 3.1 schematically illustrates the framework and its components.

Compared with previous work on semi-supervised generative models inference [73, 74, 75], the proposed framework introduces two important differences, specifically adapting to inverse problems with high-dimensional targets and observations:

i The auxiliary function incorporates a physical observation model designed with domain expertise. In many settings, especially within imaging and HCI, the Physics of how targets map to corresponding measurements is well understood and described by closed form expressions. These are 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.

In section 3.3, the proposed framework is quantitatively evaluated in simulated image recov-

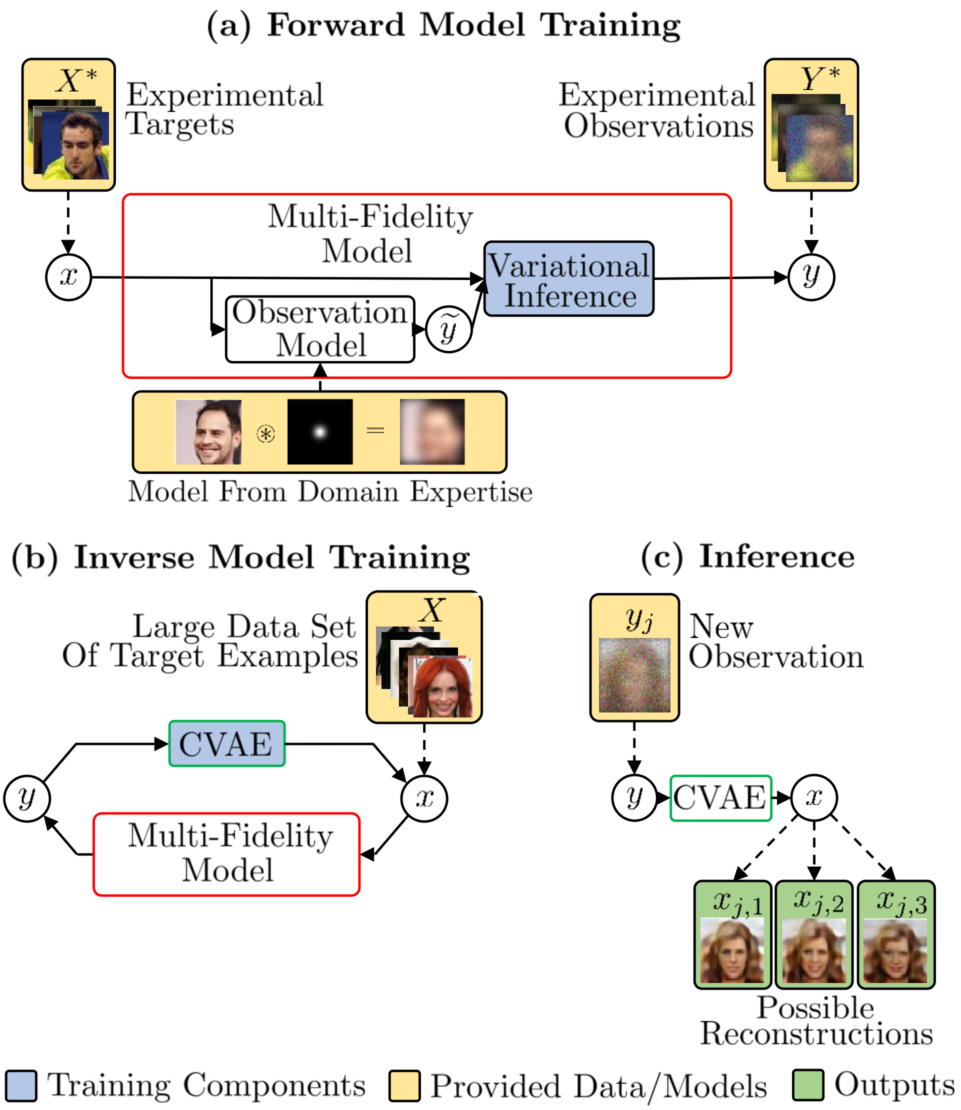


Figure 3.1: Proposed framework for training variational inference with diverse sources of information accessible for solving inverse problems, illustrated with a de-blurring example.
a) Firstly, a multi-fidelity forward model is built to generate experimental observations. A variational model is trained 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 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 used as conditions for training the CVAE to generate back the targets x. 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.

ery experiments, making use of the benchmark data sets CelebA and CIFAR10 [99, 100]. 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. Furthermore, reconstructions were found to consistently benefit from both improved analytical models and increasing number of simulated experimental acquisitions, demonstrating the ability of the novel framework to exploit in combination different types of data and domain expertise. Throughout chapter 4, the proposed technique is implemented in various real applications, making use of experimentally collected imaging and HCI data.

Differently from previous approaches, the proposed variational framework is built to learn from all the useful data and models typically available in applied inverse 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 solving an inverse problem 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

$$p(x|y_j) = \frac{p(y_j|x)p(x)}{p(y_j)}. \tag{3.1}$$

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 3.1 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 in Semi-Supervised Settings

For inverse problems relying on instrumentation specific measurements, such as imaging and HCI, 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 imaging settings can be approximately modelled by a linear transformation and independent Gaussian or Poisson noise [22]. 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. For example, many large image data sets containing relevant targets for imaging 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 framework described throughout this chapter aims at 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 inverse 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

$$p_\alpha(y|x) = \int p(\widetilde{y}|x) p_\alpha(y|x, \widetilde{y}) d\widetilde{y}. \tag{3.2}$$

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 $\widetilde{y}$.

In many inverse problems, especially in imaging and HCI, measurements are high dimensional and can present complicated posteriors that cannot be well captured by simple distributions, such as Gaussians. To provide flexible inference in the general case, while retaining efficiency of computation, the PDF $p_\alpha(y|x, \widetilde{y})$ is chosen to be a latent variable model of the form

$$p_\alpha(y|x, \widetilde{y}) = \int p_{\alpha_1}(w|x, \widetilde{y}) p_{\alpha_2}(y|x, \widetilde{y}, w) dw. \tag{3.3}$$

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 3.2 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

[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.

likelihood of interest $y_k \sim p(y|x_k)$. The optimisation to be performed is the log likelihood maximisation

$$\arg\max_{\alpha_1,\alpha_2} \quad \log p_\alpha(Y^*|X^*) = \sum_{k=1}^{K} \log \int p(\widetilde{y}|x_k) \int p_{\alpha_1}(w|x_k,\widetilde{y}) p_{\alpha_2}(y_k|x_k,\widetilde{y},w) dw d\widetilde{y}. \tag{3.4}$$

Due to the integration over latent variables w, the maximisation of equation 3.4 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 [61, 65]. The VAE formulation for the multi-fidelity model is presented in detail in supplementary section A.2.1. Through this approach, training of the parameters $\alpha = \{\alpha_1, \alpha_2\}$ and β can be performed through the following stochastic optimisation:

$$\arg\max_{\alpha_1,\alpha_2,\beta} \quad \sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{s=1}^{S} \log p_{\alpha_2}(y_k|x_k,\widetilde{y}_{k,v},w_s) - D_{KL}(q_\beta(w|x_k,y_k,\widetilde{y}_{k,v})||p_{\alpha_1}(w|x_k,\widetilde{y}_{k,v}))\right]. \tag{3.5}$$

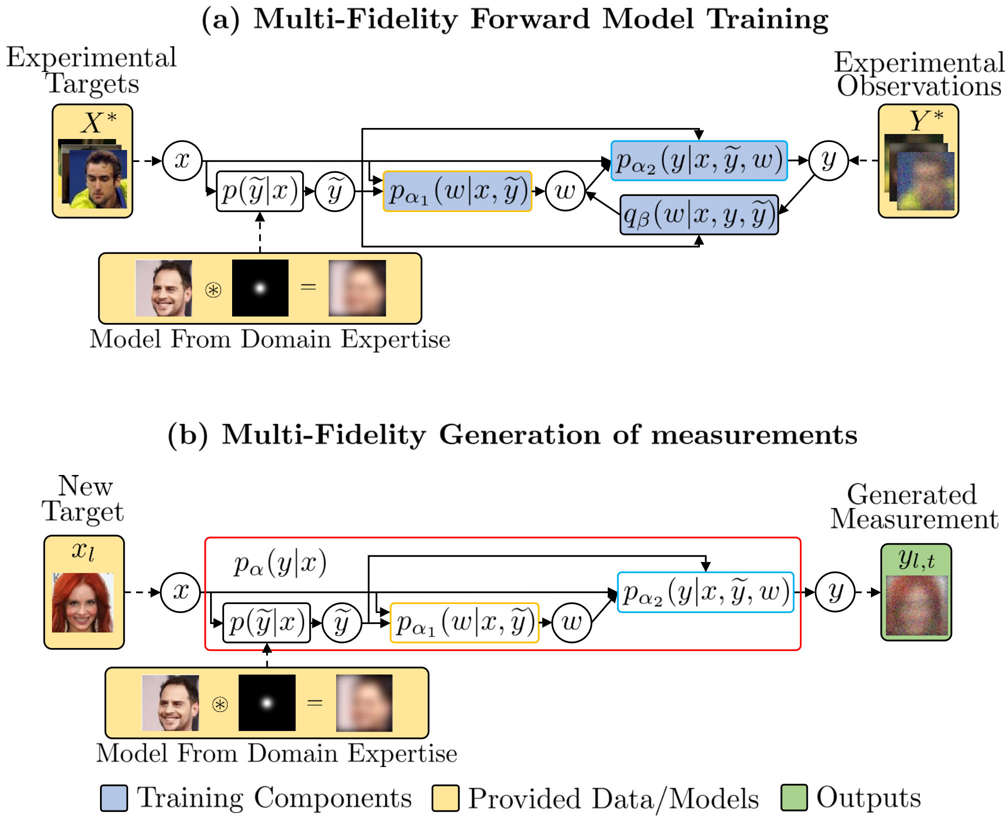
This bound is maximised stochastically by drawing samples from the approximate distribution $p(\widetilde{y}|x)$ and subsequently from a recognition model $q_\beta(w|x_k, y_k, \widetilde{y})$, which is chosen as an isotropic Gaussian distribution, the moments of which are outputs of a neural network 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. Through Jensen's inequality, a lower bound for the parametric distribution $p_\alpha(y|x)$ can be defined as

$$\begin{aligned}\log p_\alpha(y_k|x_k) &\geq \int p(\widetilde{y}|x_k) \int q_\beta(w|x_k,y_k,\widetilde{y}) \log\left[\frac{p_{\alpha_1}(w|x_k,\widetilde{y})}{q_\beta(w|x_k,y_k,\widetilde{y})} p_{\alpha_2}(y_k|x_k,\widetilde{y},w)\right] dw d\widetilde{y} \\ &= \int p(\widetilde{y}|x_k)\left[\int q_\beta(w|x_k,y_k,\widetilde{y}) \log p_{\alpha_2}(y_k|x_k,\widetilde{y},w) dw - D_{KL}(q_\beta||p_{\alpha_1})\right] d\widetilde{y},\end{aligned} \tag{3.6}$$

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

$$D_{KL}(q_\beta||p_{\alpha_1}) = \int q_\beta(w|x_k,y_k,\widetilde{y}) \log \frac{q_\beta(w|x_k,y_k,\widetilde{y})}{p_{\alpha_1}(w|x_k,\widetilde{y})} dw. \tag{3.7}$$


(a) Multi-Fidelity Forward Model Training

(b) Multi-Fidelity Generation of measurements

Experimental Targets
X^*
x
$p(\widetilde{y}|x)$
$\widetilde{y}$
$p_{\alpha_1}(w|x,\widetilde{y})$
w
$p_{\alpha_2}(y|x,\widetilde{y},w)$
$q_\beta(w|x,y,\widetilde{y})$
y
Experimental Observations
Y^*
⊛ =
Model From Domain Expertise
New Target
x_l
x
$p_\alpha(y|x)$
$p(\widetilde{y}|x)$
$\widetilde{y}$
$p_{\alpha_1}(w|x,\widetilde{y})$
w
$p_{\alpha_2}(y|x,\widetilde{y},w)$
y
Generated Measurement
$y_{l,t}$
⊛ =
Model From Domain Expertise
Training Components
Provided Data/Models
Outputs

Figure 3.2: 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 $y_{l,t} \sim p_\alpha(y|x)$.

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 [61]. Pseudo-code for the multi-fidelity forward model training is given in algorithm 1.

Once the weights α have been trained through the maximisation of equation 3.5, 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

$$y_{l,t} \sim p_{\alpha_2}(y|x_l,\widetilde{y}_v,w_s), \quad \text{where} \quad \widetilde{y}_v \sim p(\widetilde{y}|x_l) \quad \text{and} \quad w_s \sim p_{\alpha_1}(w|x_l,\widetilde{y}_v). \tag{3.8}$$

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

Inputs: Analytical forward model from domain expertise $p(\widetilde{y}|x)$; set of measured Groundtruths $X^* = \{x_{k=1:K}\}$; 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.

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

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.

3.2.1 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 [101]. The target posterior of equation 3.1 is intractable to directly evaluate or to draw samples from. This is because, as described in section 3.1.1, the prior $p(x)$ and likelihood $p(y|x)$ are not directly accessible, but only samples are available in the form of data sets. As a result, to find an approximate non-iterative solution PDF to the inverse problem, a parametric model $r_\theta(x|y)$ is trained to approximate the true intractable posterior $p(x|y)$ over the distribution of expected measurements $p(y)$.

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,

$$\arg\min_{\theta} \ \mathbb{E}_{p(y)} H[p(x|y), r_\theta(x|y)] = \arg\max_{\theta} \ \mathbb{E}_{p(y)} \int p(x|y) \log r_\theta(x|y) dx. \tag{3.9}$$

The optimisation of Equation 3.9 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

$$\begin{aligned} \mathbb{E}_{p(y)} \int p(x|y) \log r_\theta(x|y) dx &= \iint p(y) \frac{p(y|x)p(x)}{p(y)} \log r_\theta(x|y) dx dy \\ &= \int p(x) \int p(y|x) \log r_\theta(x|y) dy dx. \end{aligned} \tag{3.10}$$

In order to stochastically estimate and maximise the expression of equation 3.10, 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 3.10 to be maximised can then be approximated as

$$\mathbb{E}_{p(y)} \int p(x|y) \log r_\theta(x|y) dx \simeq \int p(x) \int p_\alpha(y|x) \log r_\theta(x|y) dy dx. \tag{3.11}$$

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.

CVAE as Approximate Posterior

As target images and observations often lie on complicated manifolds, the approximate distribution $r_\theta(x|y)$ needs to be of considerable capacity in order to accurately capture the variability of solution spaces in inverse problems, such as image reconstruction and pose estimation. For example, pixel values in a distribution of natural images are highly correlated in complicated ways. Therefore, a parametric function aiming to capture reconstruction PDFs of natural images needs to be sufficiently expressive to express these correlations. To this end, and in order to retain computational efficiency, the approximating distribution $r_\theta(x|y)$ is chosen as a conditional latent variable model

$$r_\theta(x|y) = \int r_{\theta_1}(z|y) r_{\theta_2}(x|z,y) dz. \tag{3.12}$$

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 [62], where each generated pixel in the recovered image is conditioned on z and y, but also on the previously generated neighbouring pixels.

Latent variable models of this type have been proven to be powerful conditional image generators [65, 66] 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

$$\arg\max_{\theta_1,\theta_2} \int p(x) \int p_\alpha(y|x) \log \int r_{\theta_1}(z|y) r_{\theta_2}(x|z,y) dz dy dx. \tag{3.13}$$

As for the likelihood structure in the multi-fidelity forward modelling, directly performing the maximisation of equation 3.13 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.2.2. Making use of this lower bound, we can define the objective function for the inverse model as

$$\arg\max_{\theta_1,\theta_2,\phi} \sum_{l=1}^{L} \sum_{t=1}^{T} \left[\sum_{s=1}^{S} \log r_{\theta_2}(x_l|z_s, y_{l,t}) - D_{KL}(q_\phi(z|x_l, y_{l,t}) || r_{\theta_1}(z|y_{l,t})) \right], \tag{3.14}$$

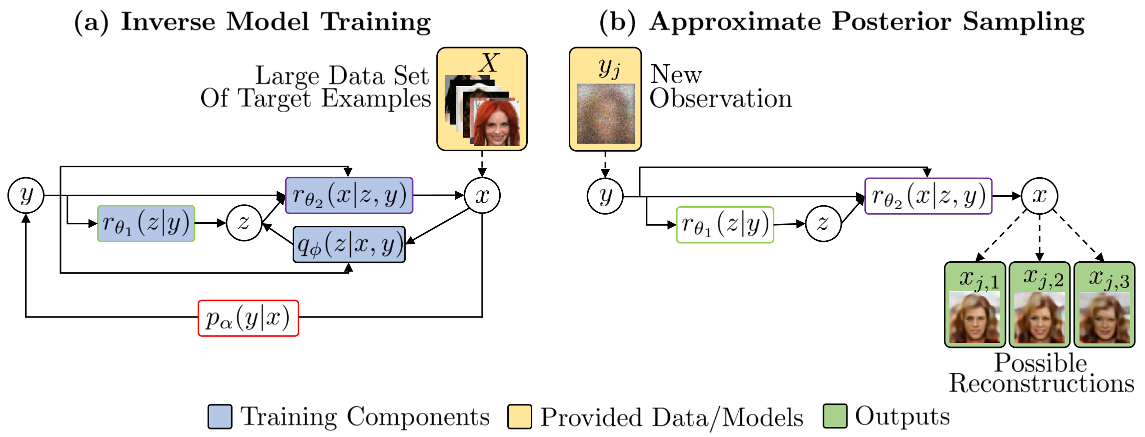


Figure 3.3: Variational inverse model. (a) The model is trained to maximise the evidence lower bound on the likelihood of targets x conditioned on observations y. The posterior components $r_{\theta_1}(z|y)$ and $r_{\theta_2}(x|z,y)$ are trained along with the auxiliary recognition model $q_\phi(z|x,y)$. 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 $x_{j,i} \sim r_{\theta_2}(x|z_{j,i}, y_j)$.

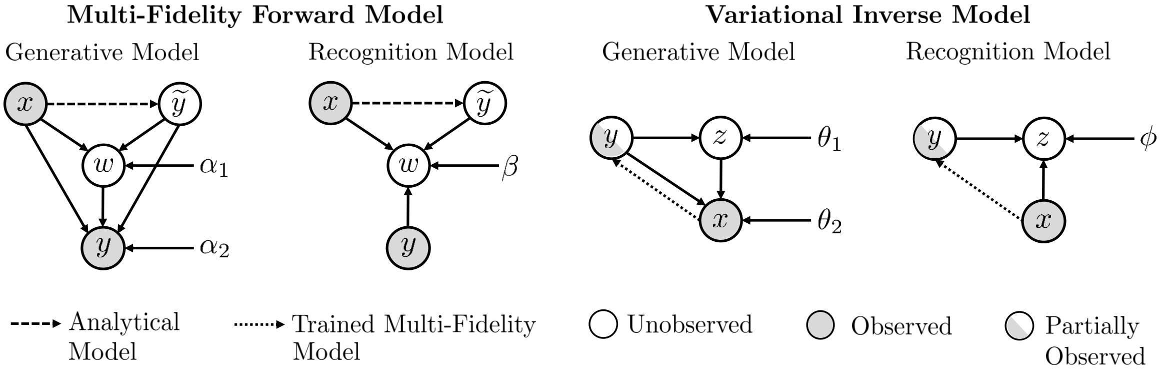


Figure 3.4: Graphical models for training of the multi-fidelity forward model and the variational inverse model.

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 [61]. The variational approximate posterior $r_\theta(x|y)$ is trained by performing the maximisation of equation 3.14 through steepest ascent. The training procedure is schematically shown in Figure 3.3(a) and detailed as a pseudo-code in algorithm 2. 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 3.4.

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.

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

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 interpreted as different possible solutions to the inverse problem and can be used in different ways to extract information of interest. For instance, in the case of images, one can compute per-pixel marginal means and standard deviations, in order to visualise the expected mean values and marginal uncertainty on the retrieved targets. Figure 3.3(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 [65]. 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.

3.3 Simulated Experiments

To quantitatively test the proposed framework in fully controlled settings, simulated experiments are performed with common image processing tasks, often used as inverse model examples. Experiments include Gaussian blurring, down-sampling and partial occlusion with both CelebA and CIFAR examples [99, 100]. The test images are corrupted with the given transformation and additive Gaussian noise. Variational models are then used to perform reconstructions, with the aim of capturing the posterior of solutions to the resulting inverse problem.

Conditional generative models for solving this type of problems are typically trained by applying the known degradation to the whole training set and then using the degraded images as conditions to train the model [66, 72]. With real imaging systems, obtaining degraded images for training sets containing tens of thousands of examples is extremely expensive and often unfeasible, as these need to be physically acquired or simulated with very high accuracy to match the real measurements that will be encountered upon testing.

As a result, in the experiments presented here, to simulate real 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 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. For example, in the case of Gaussian blurring, the inaccurate observation model is given different Gaussian point spread function (PSF) width and noise level compared to the true transformation encountered upon testing.

3.3.1 Qualitative Comparison with Standard training of CVAEs

Experimental Details

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.

First, CVAEs are trained directly, using $K = 3,000$ 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 $K = 3,000$ paired examples, but including the whole training set of $L = 100,000$ unobserved targets from the CelebA data set and the inaccurate blurring model.

The first set of experiments shown in figure 3.7 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 (top row in figure 3.7), 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 (middle row in figure 3.7), 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 (bottom row in figure 3.7), the true blurring Gaussian PSF 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 composed of convolutional and fully connected networks and the precise configuration of its components is illustrated in figure 3.5. The inverse model, inferring reconstructed images from blurred observations, is also the convolutional version shown in figure 3.6, both for the proposed training method and for the CVAE standard training. The sizes of the filter banks W used are reported in table 3.1.

From the trained CVAE models, we extract four different types of reconstructions: i) a pseudo-maximum reconstruction (Pmax in figure 3.7), reconstructed from the mean latent variable of the conditional encoder's distribution, ii) a mean reconstruction, averaging reconstructions resulting from multiple draws of the conditional encoder's distribution, iii) a per-pixel standard deviation of these multiple reconstructions and iv) examples of these individual multiple reconstructions (Draws from Posterior in figure 3.7).

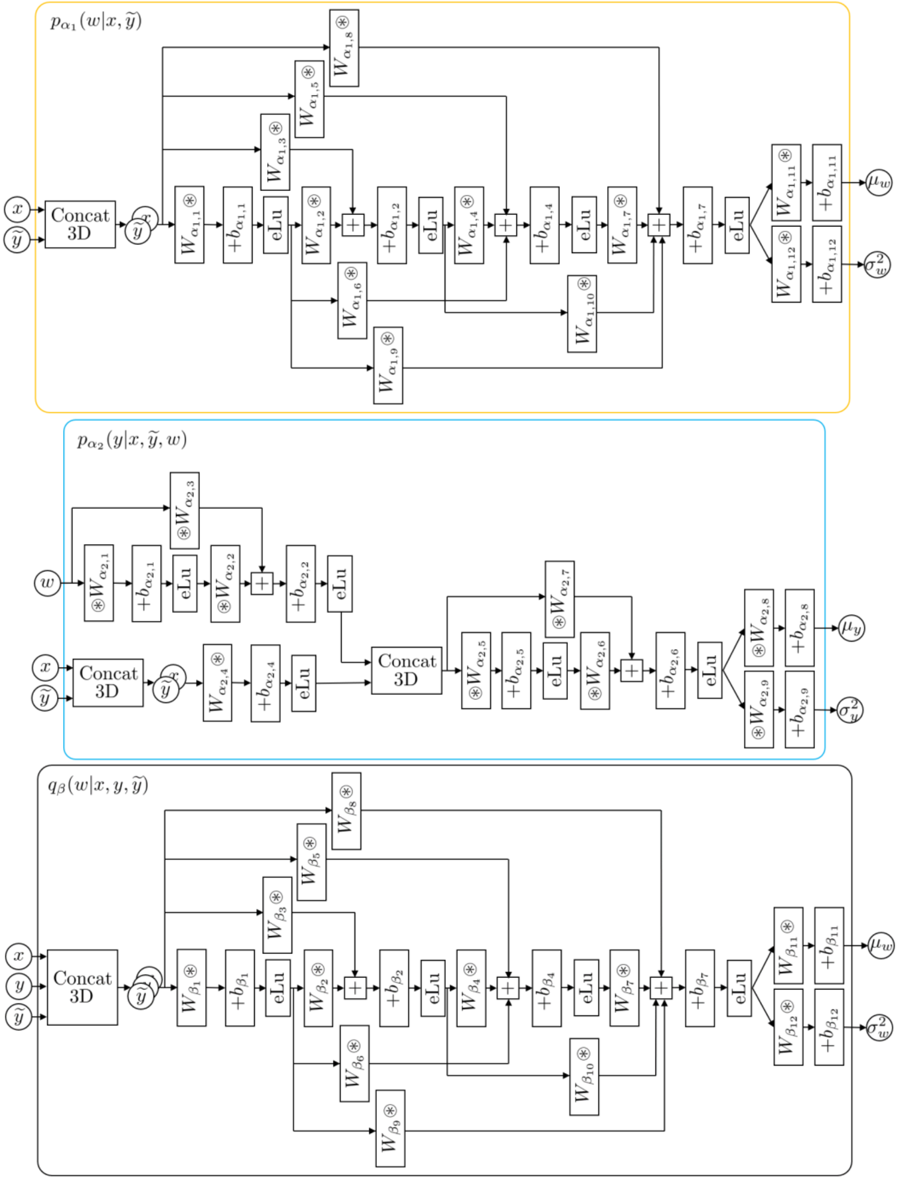

$p_{\alpha_1}(w|x,\tilde{y})$
x
$\tilde{y}$
Concat 3D
$W_{\alpha_{1,1}}\circledast$
$+b_{\alpha_{1,1}}$
eLu
$W_{\alpha_{1,2}}\circledast$
$W_{\alpha_{1,3}}\circledast$
$W_{\alpha_{1,5}}\circledast$
$W_{\alpha_{1,8}}\circledast$
$+b_{\alpha_{1,2}}$
$W_{\alpha_{1,4}}\circledast$
$W_{\alpha_{1,6}}\circledast$
$W_{\alpha_{1,9}}\circledast$
$+b_{\alpha_{1,4}}$
$W_{\alpha_{1,7}}\circledast$
$W_{\alpha_{1,10}}\circledast$
$+b_{\alpha_{1,7}}$
$W_{\alpha_{1,11}}\circledast$
$W_{\alpha_{1,12}}\circledast$
$+b_{\alpha_{1,11}}$
$+b_{\alpha_{1,12}}$
μ_w
σ^2_w
$p_{\alpha_2}(y|x,\tilde{y},w)$
w
$\circledast W_{\alpha_{2,1}}$
$+b_{\alpha_{2,1}}$
$\circledast W_{\alpha_{2,2}}$
$\circledast W_{\alpha_{2,3}}$
$+b_{\alpha_{2,2}}$
$W_{\alpha_{2,4}}\circledast$
$+b_{\alpha_{2,4}}$
$\circledast W_{\alpha_{2,5}}$
$+b_{\alpha_{2,5}}$
$\circledast W_{\alpha_{2,6}}$
$\circledast W_{\alpha_{2,7}}$
$+b_{\alpha_{2,6}}$
$\circledast W_{\alpha_{2,8}}$
$\circledast W_{\alpha_{2,9}}$
$+b_{\alpha_{2,8}}$
$+b_{\alpha_{2,9}}$
μ_y
σ^2_y
$q_\beta(w|x,y,\tilde{y})$
y
$W_{\beta_1}\circledast$
$+b_{\beta_1}$
$W_{\beta_2}\circledast$
$W_{\beta_3}\circledast$
$W_{\beta_5}\circledast$
$W_{\beta_8}\circledast$
$+b_{\beta_2}$
$W_{\beta_4}\circledast$
$W_{\beta_6}\circledast$
$W_{\beta_9}\circledast$
$+b_{\beta_4}$
$W_{\beta_7}\circledast$
$W_{\beta_{10}}\circledast$
$+b_{\beta_7}$
$W_{\beta_{11}}\circledast$
$W_{\beta_{12}}\circledast$
$+b_{\beta_{11}}$
$+b_{\beta_{12}}$

Figure 3.5: 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.

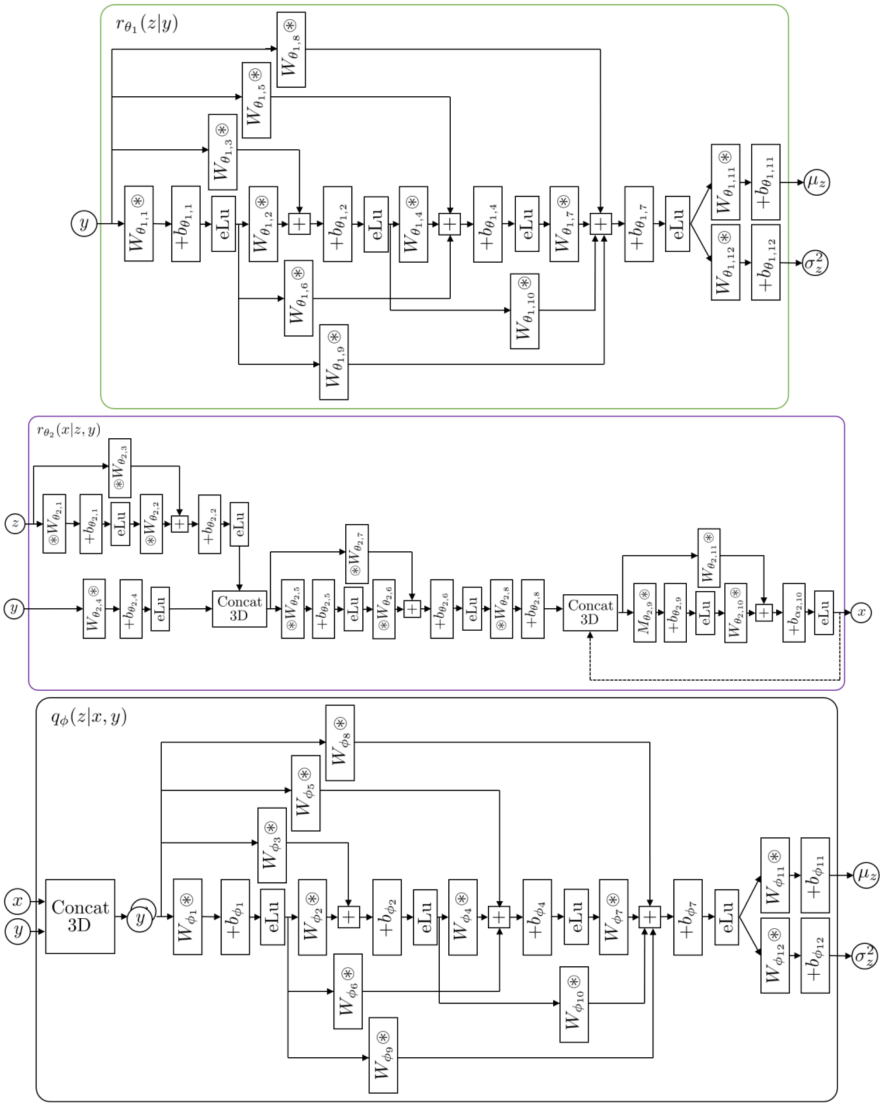

$r_{\theta_1}(z|y)$
y
$W_{\theta_{1,1}}\circledast$
$+b_{\theta_{1,1}}$
eLu
$W_{\theta_{1,2}}\circledast$
$W_{\theta_{1,3}}\circledast$
$+b_{\theta_{1,2}}$
eLu
$W_{\theta_{1,4}}\circledast$
$W_{\theta_{1,5}}\circledast$
$W_{\theta_{1,6}}\circledast$
$+b_{\theta_{1,4}}$
eLu
$W_{\theta_{1,7}}\circledast$
$W_{\theta_{1,8}}\circledast$
$W_{\theta_{1,9}}\circledast$
$W_{\theta_{1,10}}\circledast$
$+b_{\theta_{1,7}}$
eLu
$W_{\theta_{1,11}}\circledast$
$+b_{\theta_{1,11}}$
$W_{\theta_{1,12}}\circledast$
$+b_{\theta_{1,12}}$
μ_z
σ_z^2
$r_{\theta_2}(x|z,y)$
z
$\circledast W_{\theta_{2,1}}$
$+b_{\theta_{2,1}}$
eLu
$\circledast W_{\theta_{2,2}}$
$\circledast W_{\theta_{2,3}}$
$+b_{\theta_{2,2}}$
eLu
y
$W_{\theta_{2,4}}\circledast$
$+b_{\theta_{2,4}}$
eLu
Concat 3D
$\circledast W_{\theta_{2,5}}$
$+b_{\theta_{2,5}}$
eLu
$\circledast W_{\theta_{2,6}}$
$\circledast W_{\theta_{2,7}}$
$+b_{\theta_{2,6}}$
eLu
$\circledast W_{\theta_{2,8}}$
$+b_{\theta_{2,8}}$
Concat 3D
$M_{\theta_{2,9}}\circledast$
$+b_{\theta_{2,9}}$
eLu
$W_{\theta_{2,10}}\circledast$
$W_{\theta_{2,11}}\circledast$
$+b_{\alpha_{2,10}}$
eLu
x
$q_\phi(z|x,y)$
x
y
Concat 3D
y
$W_{\phi_1}\circledast$
$+b_{\phi_1}$
eLu
$W_{\phi_2}\circledast$
$W_{\phi_3}\circledast$
$+b_{\phi_2}$
eLu
$W_{\phi_4}\circledast$
$W_{\phi_5}\circledast$
$W_{\phi_6}\circledast$
$+b_{\phi_4}$
eLu
$W_{\phi_7}\circledast$
$W_{\phi_8}\circledast$
$W_{\phi_9}\circledast$
$W_{\phi_{10}}\circledast$
$+b_{\phi_7}$
eLu
$W_{\phi_{11}}\circledast$
$+b_{\phi_{11}}$
$W_{\phi_{12}}\circledast$
$+b_{\phi_{12}}$
μ_z
σ_z^2

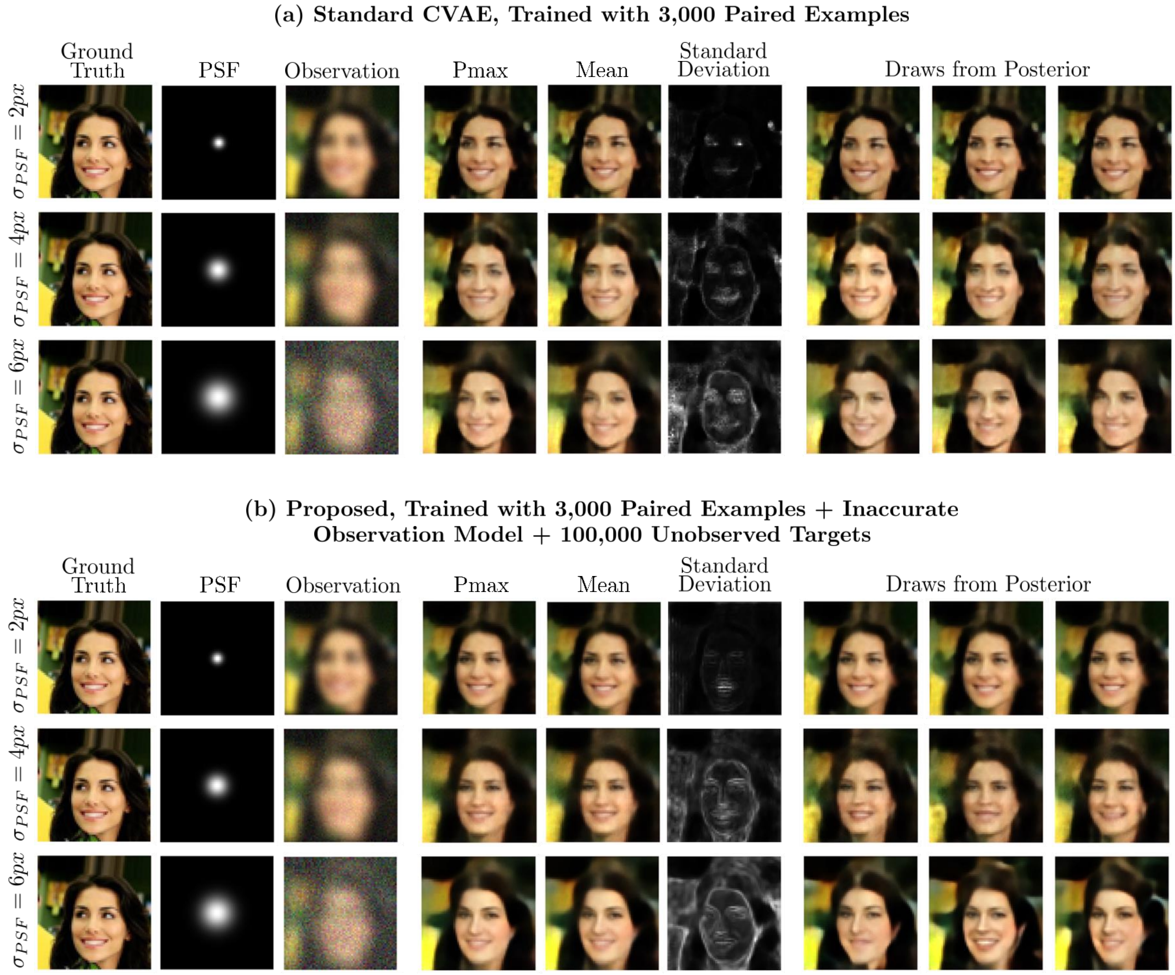
Figure 3.6: Parametric distributions' structures for the convolutional version of the multi-fidelity forward model. $W\circledast$ indicates a convolution with filter bank W, while $\circledast W$ indicates a transpose convolution. M indicates the masked convolution part of the PixelVAE model (see [62]).

Table 3.1: Filter banks of the multi-fidelity forward model and variational inverse model used in the experiments of section 3.3.1. The table reports the filter bank name used in the architectures shown in figures 3.5 and 3.6, 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

Results and Discussion

Reconstruction examples are shown in Figure 3.7. With a limited amount of training examples available, the CVAE results into over-fitting and under-estimation of the posterior variance, failing to explore the variability of the solution space (3.7(a)). This effect can be observed particularly in the draws from the posterior for the most degraded example in the bottom row figure 3.7(a). This is because 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. Contrarily, by exploiting all available data and models with the proposed framework, the trained CVAE model adequately fits to the posterior of possible solutions to the inverse model. As shown in figure 3.7(b), draws from the recovered posterior diversify more as the inverse problem becomes increasingly ill posed, reflecting a more accurate representation of the solution space variance.


(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 3.7: 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. (b) Posterior recovery obtained by training the CVAE with the proposed framework, exploiting all sources of information available.

3.3.2 Quantitative Comparison with Standard training of CVAEs

Experimental Details

To test the proposed framework in different conditions, multiple experiments analogous to those illustrated in Figure 3.7, with different relative model errors, are performed varying the number K of available image-observation pairs. This set of experiments was 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 PSNR serves as a measure of deterministic performance, giving an index of similarity between the ground truth and the most likely reconstruction. It is defined as follows:

$$PSNR = 20\log_{10}\frac{I_{max}}{|I-R|},$$

where I is the ground-truth image, R is the reconstructed image and I_{max} is the maximum pixel value in the ground-truth I. The PSNR is often use to quantify image reconstruction quality because it approximates our understanding of how image quality is perceived by humans [102]. The latter is a measure of probabilistic performance, as it approximates the log 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 log likelihood $\log(p(x|y))$ is a natural way to evaluate the probabilistic performance of models, as it measures how likely the model is to generate the ground-truth x when given y.

The forward multi-fidelity model used in the proposed methods is built with the simple fully connected structures of figure 3.8. The comparison CVAE and the inverse models are identical and are built with the fully connected structures of figure 3.9. 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 $2,500$ hidden units in the deterministic layers and latent variables z of 800 dimensions.

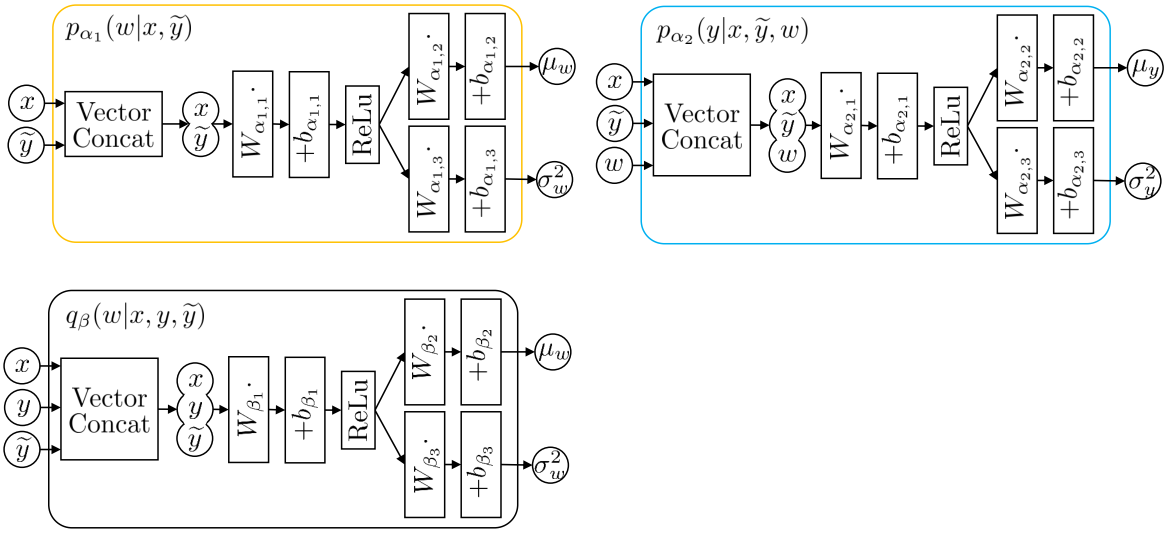


Figure 3.8: Parametric distributions' structures for the fully connected version of the multi-fidelity forward model. The output variables are sampled from Gaussian distributions having the corresponding output moments shown.

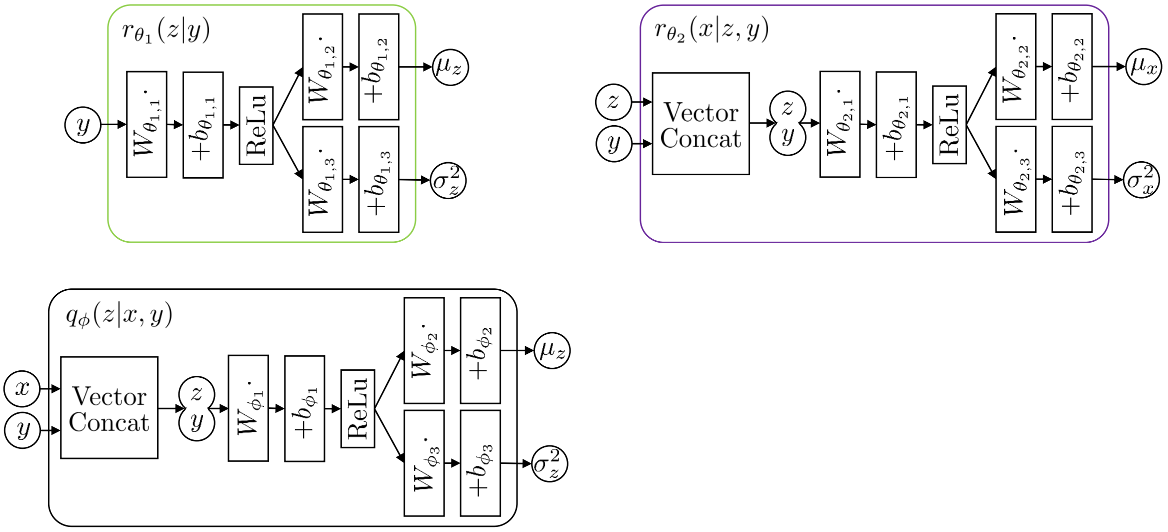


Figure 3.9: Parametric distributions' structures for the fully connected version of the inverse model. The output variables are sampled from Gaussian distributions having the corresponding output moments shown.

Results and Discussion

As shown in Figure 3.10, 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

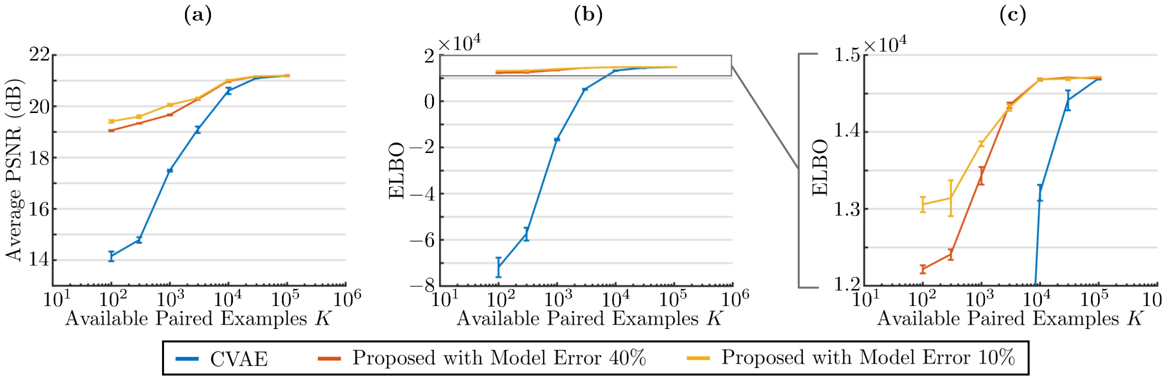


Figure 3.10: Posterior reconstruction from blurred CelebA images at varying number K of paired training examples. (a) Average PSNR between reconstructed pseudo-maximum and ground truth images. (b-c) ELBO assigned to the test set by the trained models.

cheap sources of information, the proposed framework displays appreciable performance gains, as evaluated by PSNR and ELBO, even with very scarce paired image-observation examples.

This experiment particularly highlights the role of the analytical observation model in the proposed framework. The use 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. This means that the accuracy of the analytical observation model significantly affects the recovery when availability of empirical evidence is low, but is progressively less influential as more data becomes available. This behaviour is desired when designing inverse model reconstruction systems, as the model can significantly use the information provided by domain expertise when empirical evidence is scarce, but is then able to progressively abandon it in favour of real-world observations as more measurements become available.

3.3.3 Qualitative Comparison with alternative Training

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.

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.

Experimental Details

These experiments were performed 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. In all cases, the CVAEs used for reconstructions are identical and have the convolutional form shown in figure 3.6, with the filter banks parameters listed in table 3.1. For the proposed VICI framework, the forward model was of the convolutional form shown in figure 3.5, with the filter structures reported in table 3.1.

Results and Discussion

Reconstructions obtained by training CVAEs with the three baseline approaches are compared to the proposed method in Figure 3.11. When training with too few experimental examples only (figure 3.11(a)), the CVAE over-fits and draws from the posterior are all very similar. This is a symptom of the inability of the trained model to accurately capture uncertainty in the reconstruction task. When using the available observation model only, ignoring any empirical measurements (figure 3.11(b)), reconstructions display noticeable artefacts. This is because the available observation model used for training does not match the true one encountered upon testing and therefore test samples are out of distribution compared to the training data. If the smaller portion of empirical measurements is added to the training set as-is (figure 3.11(c)), the presence of real measurements in the training set improves reconstruction marginally, but artefacts are still largely present, as the model cannot distinguish between high fidelity real data and low fidelity simulations at training time. When using

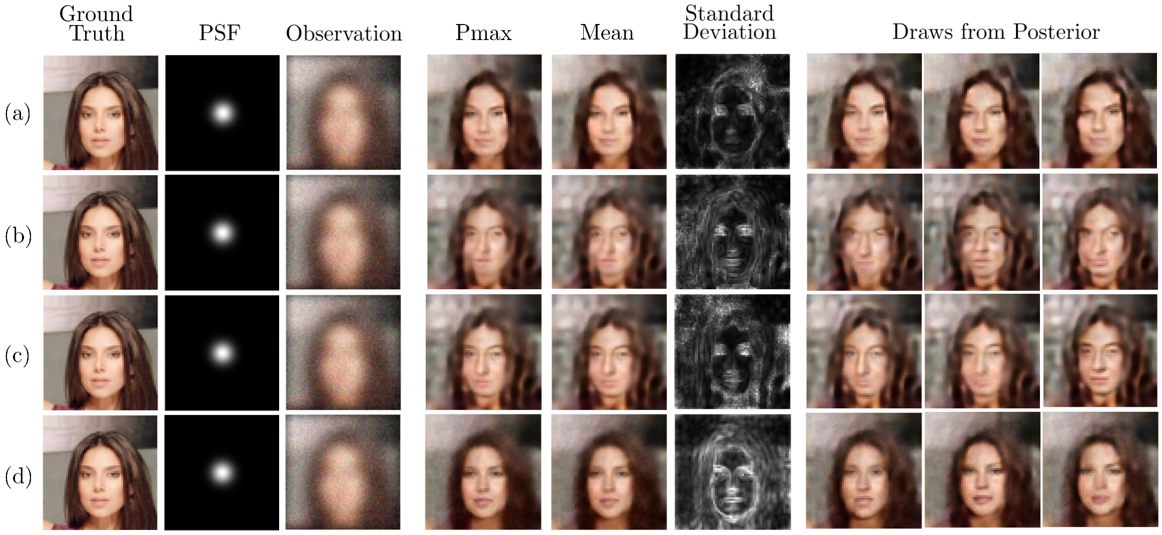


Figure 3.11: Posterior recovery from blurred. (a) CVAE trained with $K = 3,000$ paired examples alone. (b) CVAE trained with $L = 100,000$ target examples and corresponding simulated observations from the inaccurate observation model. (c) CVAE trained with $K = 3,000$ paired examples in combination with $L = 100,000$ target examples and corresponding inaccurately simulated observations. (d) CVAE trained with proposed variational framework.

the novel VICI framework to train a CVAE for reconstruction (figure 3.11(d)), the different sources of information are exploited in a principled way, resulting in accurate posterior recovery; different draws explore various plausible reconstructions.

3.3.4 Quantitative Comparison with alternative Training

Experimental Details

Experiments analogous to that described in the previous section were 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. Four different degradation conditions where tested, applying the following four degradations:

- **$\times$ 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 3.8. All inversion models, competitive and proposed, are identical and were implemented with the fully connected version of the inverse model given in figure 3.9. 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. Reconstructions are performed with $2,000$ test examples and average PSNR and ELBO are computed. Each experiment is repeated 5 times with a different random seed to obtain error bars and measures of statistical significance.

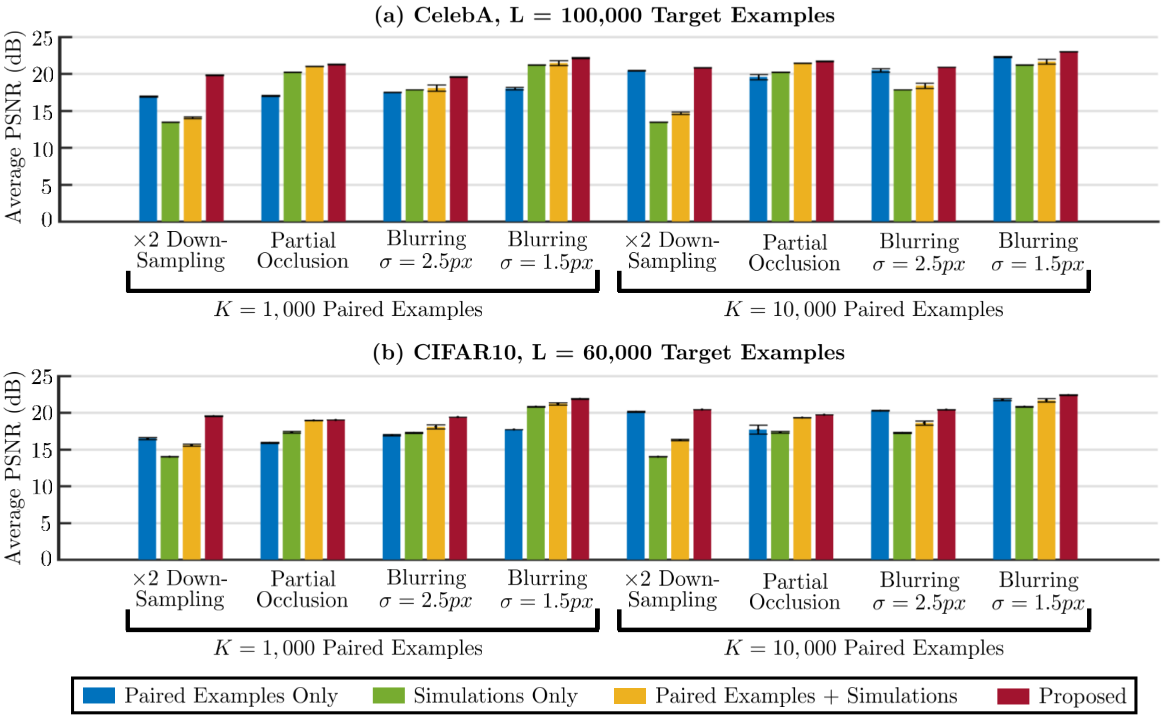


Figure 3.12: average test PSNR between ground truth images and reconstructed pseudo-maxima for (a) CelebA and (b) CIFAR10 images.

Results and Discussion

Figure 3.12 shows the average PSNR, while recovered ELBO values are reported in table 6.1. As the ELBO values are the primary focus of this study, the statistical significance of these over the 5 repeats of the experiments is assessed by computing the two-sample t-test p-values between the ELBO values obtained with the proposed VICI framework and each of the competing training strategy. These p-values are shown in table 3.3.

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.

Table 3.2: Test set evidence lower bound (ELBO) for the proposed framework compared to alternative methods of using the same information to train a CVAE.

	Paired Examples	Simulations	Paired+ Simulations	Proposed
CelebA, K=1,000 $\times 2$ 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 $\times 2$ 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 $\times 2$ 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 $\times 2$ 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	$\mathbf{14540 \pm 21}$

Table 3.3: Two-sample t-test p values between distribution of ELBO values obtained with the proposed VICI framework and each of the baseline training methods. The p-values are computed from distributions of 5 repeats of identical experiments with different random seeds.

	VICI with Paired Examples	VICI with Simulations	VICI with Paired + Simulations
CelebA, K=1,000 $\times 2$ Down-sampling	3.35×10^{-17}	2.46×10^{-16}	2.60×10^{-09}
CelebA, K=1,000 Partial Occlusion	3.75×10^{-14}	4.80×10^{-15}	5.28×10^{-10}
CelebA, K=1,000 Blurring $\sigma = 2.5px$	1.04×10^{-21}	4.13×10^{-12}	7.44×10^{-04}
CelebA, K=1,000 Blurring $\sigma = 1.5px$	4.63×10^{-14}	2.00×10^{-11}	9.94×10^{-04}
CelebA, K=10,000 $\times 2$ Down-sampling	3.48×10^{-10}	2.11×10^{-16}	1.27×10^{-06}
CelebA, K=10,000 Partial Occlusion	6.63×10^{-07}	1.48×10^{-15}	7.31×10^{-08}
CelebA, K=10,000 Blurring $\sigma = 2.5px$	6.28×10^{-15}	2.47×10^{-23}	1.65×10^{-07}
CelebA, K=10,000 Blurring $\sigma = 1.5px$	3.29×10^{-07}	5.15×10^{-15}	5.52×10^{-09}
CIFAR10, K=1,000 $\times 2$ Down-sampling	5.15×10^{-13}	3.72×10^{-13}	5.60×10^{-05}
CIFAR10, K=1,000 Partial Occlusion	8.48×10^{-13}	1.42×10^{-07}	2.30×10^{-11}
CIFAR10, K=1,000 Blurring $\sigma = 2.5px$	2.08×10^{-18}	1.84×10^{-12}	5.03×10^{-02}
CIFAR10, K=1,000 Blurring $\sigma = 1.5px$	3.62×10^{-19}	2.60×10^{-10}	9.48×10^{-02}
CIFAR10, K=10,000 $\times 2$ Down-sampling	9.88×10^{-12}	3.39×10^{-13}	6.54×10^{-07}
CIFAR10, K=10,000 Partial Occlusion	5.78×10^{-05}	9.35×10^{-09}	4.84×10^{-08}
CIFAR10, K=10,000 Blurring $\sigma = 2.5px$	3.4×10^{-08}	1.28×10^{-17}	6.74×10^{-08}
CIFAR10, K=10,000 Blurring $\sigma = 1.5px$	8.00×10^{-09}	2.83×10^{-12}	7.2210^{-07}

Chapter 4

Enabled Applications in Physics

In this chapter, the variational learning technique presented in chapter 3 is applied to several real Physics applications, where recovering objects of interest requires solving particularly challenging inverse problems. In these examples, it is shown how the new framework is able to provide rich descriptions of the solutions space to the corresponding inverse problem, capturing the complex uncertainty of these tasks and providing the ability to generate diverse realisations, corresponding to the different possible solutions consistent with the inverse problem setting.

In sections 4.1 and 4.2, two experimental optics applications are explored, performing computational imaging in holography and imaging through scattering media. In these scenarios, it is shown how the technique presented in chapter 3 is able to successfully solve the inverse problems and capture the uncertainty in the resulting complex posteriors, without needing prohibitively extensive data collections. These results open the possibilities for these complex systems to be scaled up beyond controlled laboratory settings to real world applications, as data collection requirements are greatly reduced compared to typical supervised machine learning approaches and uncertainty in the recoveries is well captured by the proposed models.

In section 4.3, a reduced version of the framework of chapter 3 is used to capture probability densities of astronomical parameters for black hole collisions from gravitational wave measurements in simulation. Capturing these probability densities is of great importance in gravitational wave astronomy, as they provide information needed for tuning instruments to better recover ongoing events, e.g. turning telescopes towards the right sky location. However, current methods built on MCMC samplers are often too slow to provide accurate estimates in time to promptly observe events. Using components from the method of chapter 3, probability densities of interest are captured in sub-second times.

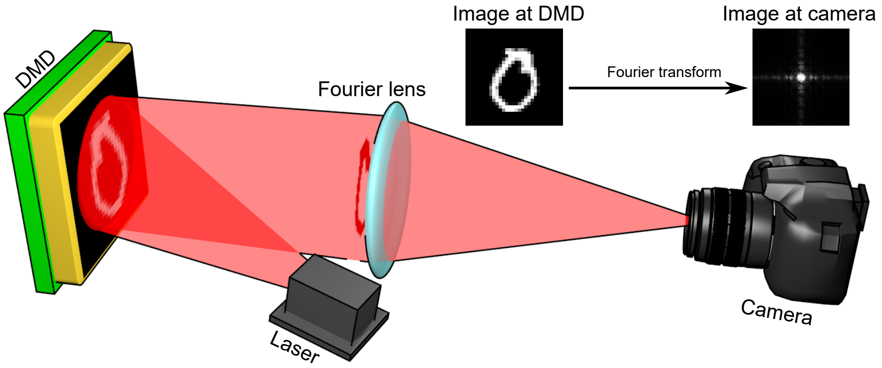


Figure 4.1: Experimental set up used for holographic image reconstruction. A binary amplitude image is projected by the Digital Micromirror Device (DMD) and a lens placed at the focal distance from the DMD display produces the corresponding Fourier image at the camera.

4.1 Holographic image reconstruction

Sensor arrays, such as Charge Coupled Device (CCD) or complementary metal-oxide semiconductor (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 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 [103, 104, 105].

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 [106, 107, 108]. 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, the variational framework proposed in chapter 3 is used 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 [109, 110], optical trapping [111], and microscopy [112].

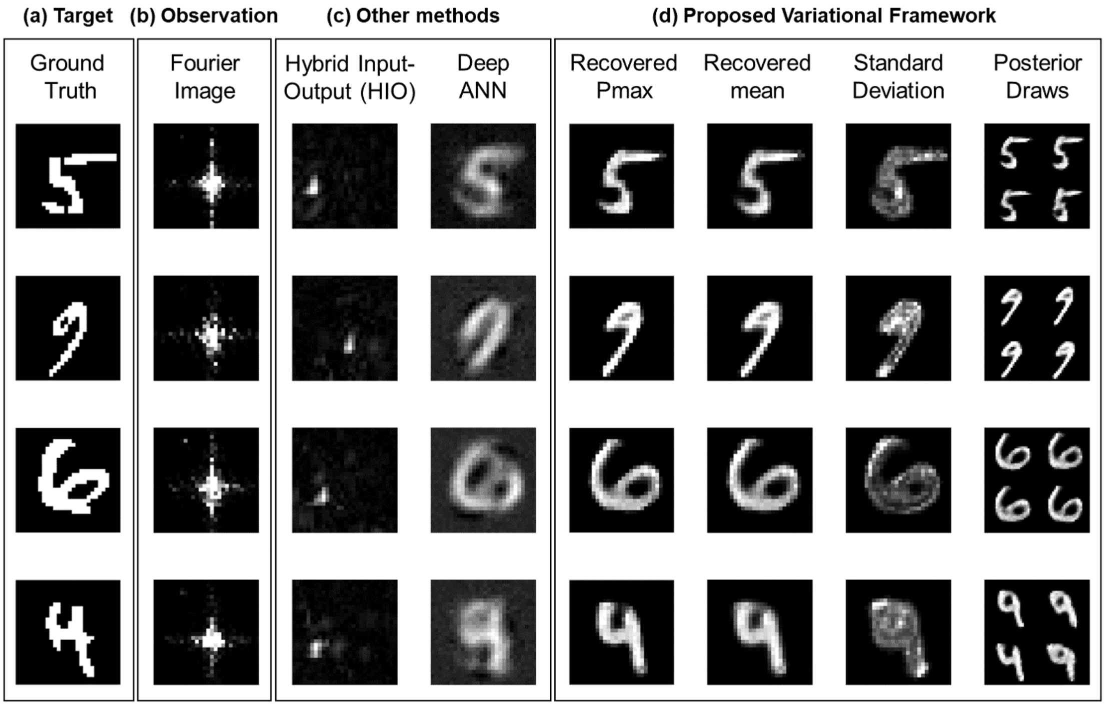


Figure 4.2: Reconstructions from experimental Fourier intensity image data. (a) Target images projected by the DMD, (b) intensity Fourier image observed at the camera, (c) reconstructions using other techniques (d) reconstructions obtained with the proposed variational method, including pseudo-maximum, pixel-marginal mean, pixel-marginal standard deviation and examples of draws from the recovered posterior.

4.1.1 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 4.1. 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, the system operates in a saturated condition, i.e. the intensities received by the sensors are higher than what can be registered, resulting in blinding, and with extremely low-resolution images. $9,600$ MNIST digits are displayed on the DMD and the corresponding camera observations are recorded. This data is used as high fidelity paired ground truths X^* and measurements Y^*. The remaining $50,400$ 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.

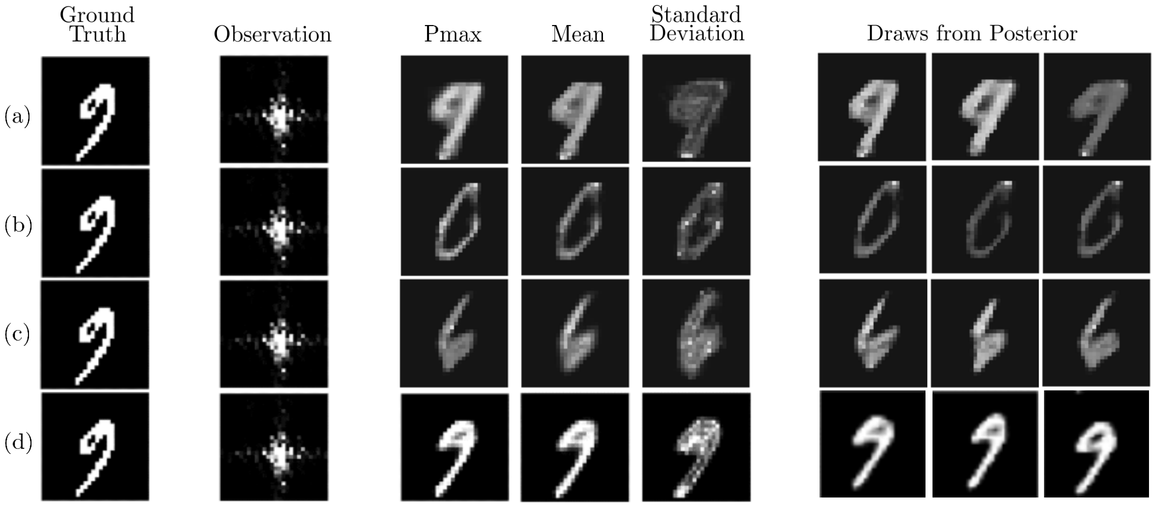


Figure 4.3: Image posterior recovery from phase-less measurements. (a) CVAE trained with the available $K = 9,600$ paired examples alone. The size of this training set is too small to obtain accurate posteriors. (b) CVAE trained with $L = 50,400$ target examples and corresponding simulated observations 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.

4.1.2 Reconstruction

Differently from the analysis presented in chapter 3, this section additionally aims at comparing the novel framework to current state of the art holographic image reconstruction techniques. For this reason, the baseline are chosen to be i) the Hybrid Input-Output (HIO) complex light-field retrieval algorithm, typically used in classical approaches [103, 104, 105], and ii) a 4-layer deep Artificial Neural Network (deep ANN), similar to the simple deep learning approaches proposed more recently to tackle the problem [106, 107, 108]. Figure 4.2 shows, with a few examples, a qualitative comparison of the considered baselines and proposed framework.

The HIO algorithm was tested only on the four images shown, as this iterative method, in its standard form tested here, requires considerable time to reconstruct each image. Both the ANN and the proposed framework are instead tested on 100 test images, for which we compute and compare the PSNR. Using the ANN, the reconstruction PSNR is $\mathbf{11.15 \pm 1.56}$, whereas using the proposed method the PSNR is $\mathbf{13.12 \pm 2.21}$. The difference was found to be statistically significant, with a two-sample p-value of 9.98×10^{-12}.

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 supplementary A.1.1 for details). This is precisely what can be observed in Figure 4.2(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 $9,600$ experimental images. As shown in Figure 4.2(d), highly accurate reconstructed 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 $50,400$ 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 3.11 for this physical experiment. An example is shown in figure 4.3. Analogously to the simulated experiments, using the experimental training set alone gives results of limited quality (figure 4.3(a)). This is because the size of this experimental training set is too small to obtain accurate posteriors. Using simulations in naive ways completely disrupts reconstructions, whether these are used alone or combined with the experimental data (figure 4.3(b-c)), 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 and good representation of the solution space, as shown in the drawn examples (figure 4.3(d)).

Figure 4.4 illustrates further this posterior exploration capability. When progressively down-sampling the resolution of experimentally measured observations, the pseudo-max 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.

4.2 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 [113]. 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

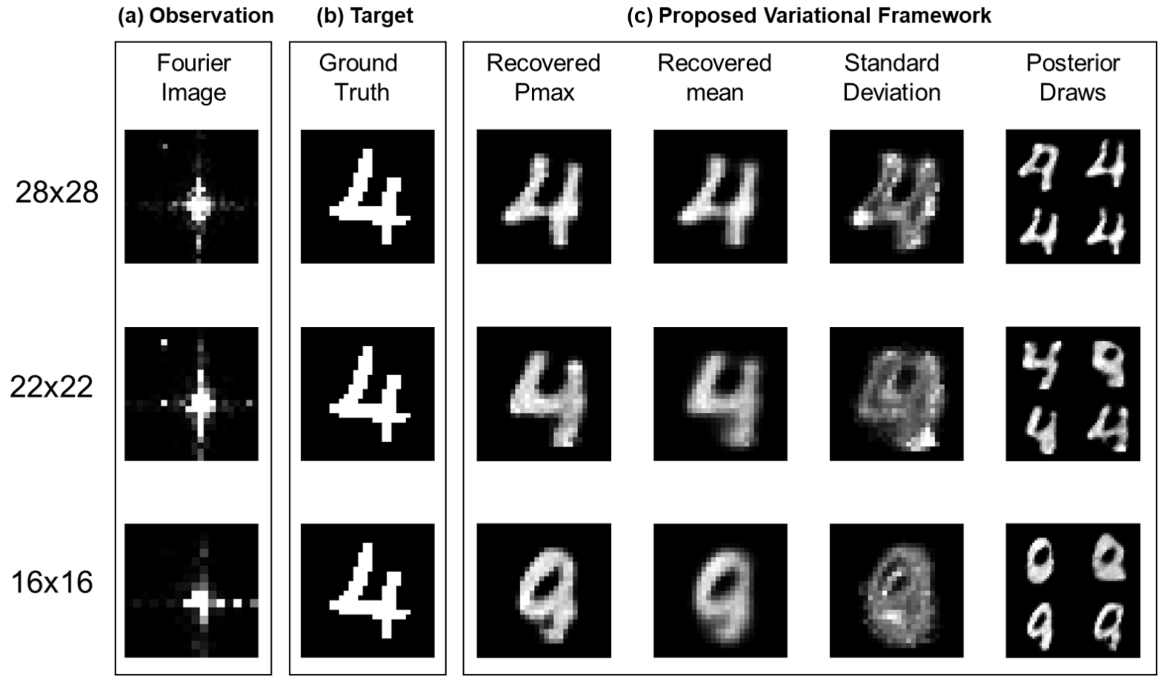


Figure 4.4: (a) Experimental Fourier intensity image data down-sampled to (28×28), (22×22) and (16×16) (top to bottom) for (b) the same target image. (c) The proposed variational framework which shows the reconstructed image quality degrades with decreasing resolution of the measured data. As expected, 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).

a hidden object from observations at the scattering medium's surface is the inverse problem that will be addressed in this section.

4.2.1 Physical Experiment

Following the experimental implementation presented by [114], 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 $\mu_a = 0.09 \text{ cm}^{-1}$ and $\mu_s = 16.5 \text{ cm}^{-1}$ respectively. A schematic representation and a photograph of the set up are shown in Figure 4.5(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

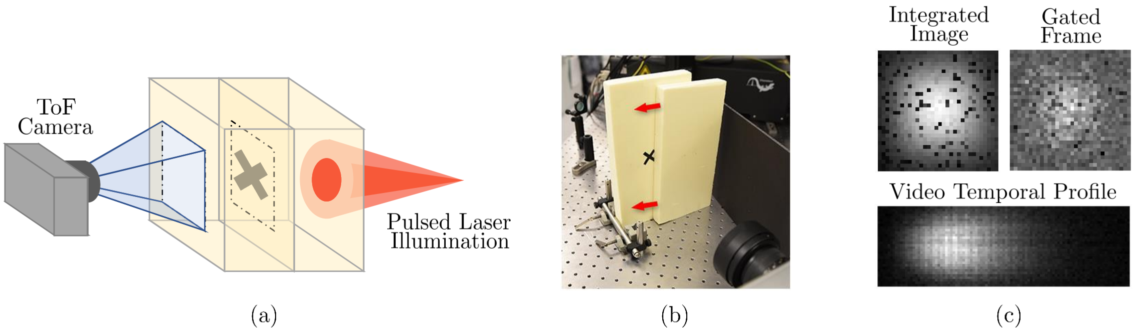


Figure 4.5: Experimental set up for imaging through scattering media. (a) Schematic representation of the experiment. A target object is embedded between two $2.5\ \mathrm{cm}$-thick slabs of diffusing material, with absorption and scattering properties comparable to those of biological tissue. One exposed 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 line of pixels at different times.

estimated difference in light intensity caused by the hidden object. An example of such videos is shown in Figure 4.5(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 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.

4.2.2 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 [115]. $86,400$ 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 [114, 116]. The propagation of photons under this assumption is described by the following differential equation

$$c^{-1}\frac{\partial \Phi(\vec{r},t)}{\partial t} + \mu_a \Phi(\vec{r},t) - D\nabla \cdot [\nabla \Phi(\vec{r},t)] = S(\vec{r},t), \tag{4.1}$$

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 4.6(c-d).

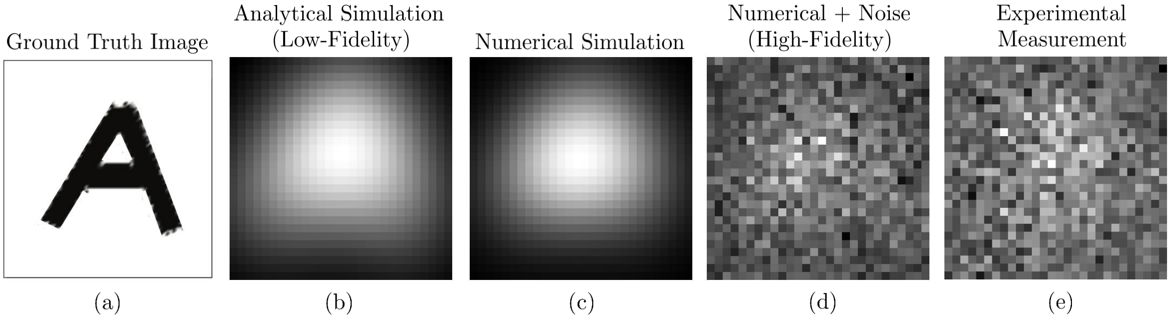


Figure 4.6: Simulated and real measurements from the time of flight (ToF) Camera. Images are single frames from the camera videos. (a) Image of the hidden object. (b) Simulated measurement using the analytical solution from the linear approximation, taken as low-fidelity estimate. (c) Simulation obtained 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 [114, 116]. 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:

$$\Phi(\vec{r}, t; \vec{r}', t') = \frac{c}{[4\pi Dc(t - t')]^{3/2}} \times \exp\left[-\frac{|\vec{r} - \vec{r}'|^2}{4Dc(t - t')}\right] \exp\left[-\mu_a c(t - t')\right]. \tag{4.2}$$

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 4.2 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 [114]. An example of such analytical estimates of the measurements is shown if Figure 4.6(b). These computations are much less expensive to perform than propagating numerically the diffusion equation, requiring less than $100\ \mathrm{ms}$ per sample to run on a TitanX GPU. However, they introduce 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 $\widetilde{y}$.

4.2.3 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 [114], 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 4.7.

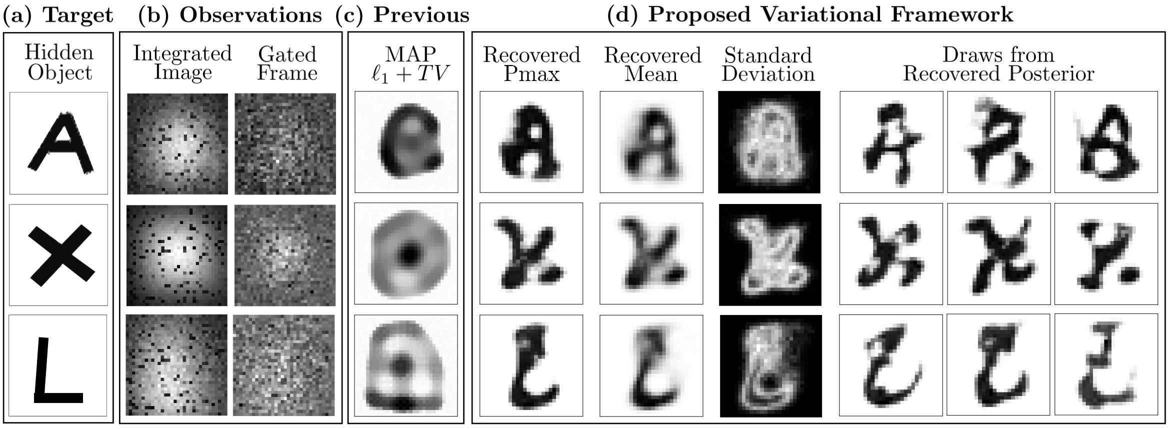


Figure 4.7: Reconstructions from experimental ToF videos. (a) Target images embedded in the scattering medium, (b) integrated and gated frames from the ToF camera videos, constituting the observed measurements, (c) reconstruction obtained using constrained optimisation with ℓ_1-norm and total variation 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 4.7(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.

4.3 Gravitational Wave Astronomy

Gravitational wave (GW) detection is now commonplace [117, 118] and as the sensitivity of the global network of GW detectors improves, $\mathcal{O}(100)$s of transient GW events will be observed per year [119]. In GW detection, the aim is to retrieve physical parameters of collision events, usually collisions between black holes or neutron stars, in order to tune instruments on earth, e.g. radio telescopes, and record such events, expanding our knowledge of the astronomical bodies involved. For example, the most intuitive of these parameters is the sky location, as shown in figure 4.8 on the top right. From GW signals, it is possible to estimate a probability density of location in the sky where the event is taking place and turn our radio telescopes towards these locations.

The main challenge in effectively using GW to carry out the aforementioned process is speed. The current methods used to estimate their source parameters employ optimally sensitive [120] but computationally costly Bayesian inference approaches [121] where typical analyses have taken between 6 hours and 5 days [122]. Binary neutron stars (BNS) and neutron stars black hole (NSBH) systems prompt counterpart electromagnetic (EM) signatures are expected on timescales of 1 second – 1 minute and the current fastest method for alerting EM follow-up observers [123], can provide estimates in $\mathcal{O}(1)$ minute, on a limited range of key source parameters. This means that, in many cases, by the time posteriors are formed using standard Bayesian methods, it is too late to tune recording instruments towards the event.

In this section, part of the framework of chapter 3 is used to train a variational model to recover these parameters of astronomical events from GW sensing. The proposed framework sacrifices the theoretical guarantees of MCMC-based methods, such as those described above, but provides several orders of magnitudes improvement in the computational speed, being able to render posteriors of collision's parameters in fractions of a second instead of hours. This means that, with the new framework, we can estimate PDFs of parameters, such as sky location, fast enough to tune our instruments and record collisions. An extended version of this application, with higher dimensional distributions and additional evaluations, is presented in [12].

4.3.1 Method

In this application, the forward model is taken as a simulation of the gravitational wave sensing process alone, instead of a multi-fidelity model as described in chapter 3 and adopted in the other applications. The inverse variational model employed has the general structure described in section 3.2.1, but was built with a number of specific features that were included in order to specifically tailor the analysis to GW signals and event parameters recovery. These modifications are as follows:

- **Physically motivated decoder distributions:** The decoder distributions $r_{\theta_2}(x_i|z, y)$ were designed to match the nature of the parameters x_i to be modelled. The sky location parameters α and δ were modelled with von Mises-Fisher distributions, which is a close approximation to a 2D Gaussian wrapped on the surface of a sphere. Analogously, periodic parameters, such as binary inclination Θ_{jn} and polarisation angle ϕ, were modelled with a von Mises distribution, which is the equivalent of the above for a 1D circle. Finally, continuous parameters for which physical limits are known, such as the masses of the colliding bodies m_1 and m_2, were modelled with truncated Gaussians.

- **1D convolutional nets to process input signals:** GW signals from different detectors are taken as distinct channels containing time a time series. Therefore, all models taking signals y as inputs ($r_{\theta_1}, r_{\theta_2}$, and q_ϕ) were designed to have a 1D multi-channel convolutional network component, mapping GW signals to intermediate vectors, which are then concatenated with other inputs before being mapped to outputs x or latent variables z.

- **Mixture of Gaussians Prior:** PDFs of collision's parameters are expected to be multimodal in many cases. While in principle a standard CVAE ca ngenerally approximate multi-mdality, in practice it is quite difficult to obtain decoder networks which maps single Gaussians distributions in the latent space to distinct distributions in sample space. For this reason, the encoder model $r_{\theta_1}(z|y)$ is constructed as a mixture of Gaussian, encoding signals y into $M = 16$ distinct Gaussian distributions.

The framework described above was named `VItamin`.

4.3.2 Experiments

Results are presented on 256 multi-detector GW test BBH waveforms in simulated advanced detector noise [124] from the LIGO Hanford, Livingston and Virgo detectors. Variants of the existing Bayesian approaches are compared to the variational model trained with the simulator as described in section 3.2.1. Posteriors produced by the `Bilby` inference library [125] are used as a benchmark in order to assess the efficiency and quality of the novel approach with the existing methods for posterior sampling.

For the benchmark analysis 9 parameters are assumed to be unknown[1]: the component masses m_1, m_2, the luminosity distance d_{L}, the sky position α, δ, the binary inclination Θ_{jn},

[1] This analysis omits the 6 additional parameters required to model the spin of each BBH component mass.

the GW polarisation angle ψ, the time of coalescence t_0, and the phase at coalescence ϕ_0. For each parameter a uniform prior is used, with the exception of the declination and inclination parameters for which priors uniform in $\cos\delta$ and $\sin\Theta_{jn}$ respectively are employed. The sampling frequency is 256 Hz, a time-series duration of 1 second, and the waveform model used is `IMRPhenomPv2` [126] with a minimum cutoff frequency of 20Hz. For each input test waveform the benchmark analysis is run using multiple sampling algorithms available within `Bilby`. For each run and sampler $\mathcal{O}(10^4)$ samples are extracted from the posterior on the 9 physical parameters.

The training process of chapter 3 is implemented using as input 10^7 whitened waveforms corresponding to parameters drawn from the same priors as assumed for the benchmark analysis. The waveforms are also of identical duration, sampling frequency, and use the same waveform model as in the benchmark analysis. The posterior results are produced by passing each of the 256 whitened noisy testing set of GW waveforms as input into the testing path of the trained inverse variational model. For each input waveform, the posterior is sampled until 10^4 posterior samples have been generated on 7 physical parameters $x = (m_1, m_2, d_{\mathrm{L}}, t_0, \Theta_{jn}, \alpha, \delta)$.

In figure 4.8, the posterior reconstruction for a test example is shown and qualitatively compared to two `Bilby` samplers, namely `Dynesty` and `ptemcee`. Two and one-dimensional marginalised posteriors generated using the output samples from `VItamin` and the `Bilby` samplers (`Dynesty` in blue, and `ptemcee` in green) are shown on the bottom-left half of figure 4.8. The different two-dimensional plots show marginalised distributions over all but two variables. From these plots, one can see that the distributions to be captured in this application are of varied nature and in many cases multi-modal and therefore impossible to capture adequately with parametric models. In this example, one can also see that `VItamin` generates a multi-dimensional, multi-modal distribution which approximates very well that returned by the more robust, but orders of magnitude slower MCMC samplers.

To quantitatively evaluate the proposed technique, `VItamin` and all other baseline samplers were run over 256 generated test parameters-GW signals pairs. For each test example and for each method samples are drawn to represent the respective distributions. The agreement between samplers is then calculated by computing the KL divergence between distributions with a k-nearest-neighbours method [127]. The KL divergences between `VItamin` and baseline samplers are histogrammed in figure 4.9. The coloured histograms are generated by computing the KL divergence between `VItamin` and the baseline sampler that is being compared. These histograms show the distribution of disagreement between `VItamin` and the baselines. The grey histograms in each plot are generated by computing the KL divergences between the baseline being compared in each case and all other baselines. These histograms show the distribution of disagreement amongst baselines. These result show that `VItamin` reconstructs PDFs of collisions' parameters that are in good agreement with the

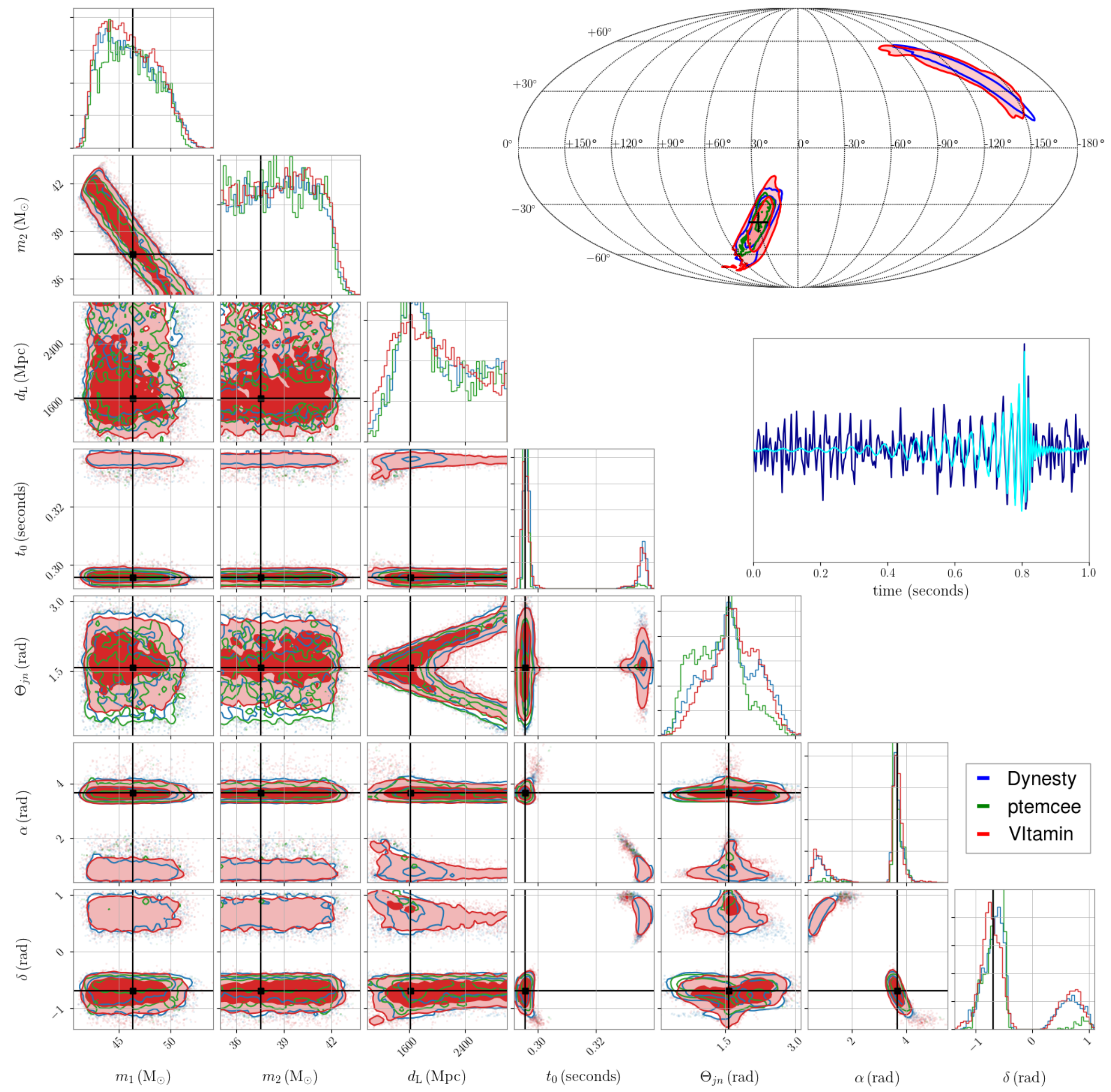


Figure 4.8: Corner plot showing one and two-dimensional marginalised posterior distributions on the GW parameters for one test example. Filled red contours represent the two-dimensional joint posteriors obtained with `VItamin`. Solid blue and green contours are the corresponding posteriors output from benchmark analyses using the `Dynesty` and `ptemcee` samplers respectively. In each case, the contour boundaries enclose $68, 90$ and 95% probability. One dimensional histograms of the posterior distribution for each parameter from both methods are plotted along the diagonal. Black vertical and horizontal lines denote the true parameter values of the simulated signal. A Mollweide projection of the sky location posteriors from all three analyses are included at the top-right of the figure. All results presented in this section correspond to a three-detector configuration but for clarity only the H1 whitened noisy time-series y and the noise-free whitened signal (in blue and cyan respectively) are plotted to the right of the figure.

slower more accurate methods, as it disagrees with baselines similarly to how much the baselines disagree with each other.

The significance of these results is most evident in the orders of magnitude increase in speed over existing algorithms; All of the baseline samplers take in the order of hours to retrieve PDFs of collision parameters, while VItamin can generate sufficient samples from test inputs in the order of 10^{-2}-10^{-1} seconds, depending on hardware resources. This speed improvement comes at little accuracy cost, as both qualitative and quantitative results show that the proposed framework can produce PDFs of equivalent quality to the competing methods.

The approach was demonstrated using BBH signals but with additional work to increase sample rate and signal duration, the method can also be extended for application to signals from BNS mergers (e.g., GW170817 [118], and GW190425 [128]) and NSBH systems where improved low-latency alerts will be especially pertinent. Current Bayesian analyses limit the amount of lead time it is possible to give EM partners in order to slew their telescopes to the optimal location in the sky. By using the variational approach, parameter estimation speed will no longer be a limiting factor in observing the prompt EM emission expected on shorter time scales than is achievable with existing LIGO-Virgo Collaboration analysis tools such as Bayestar [123].

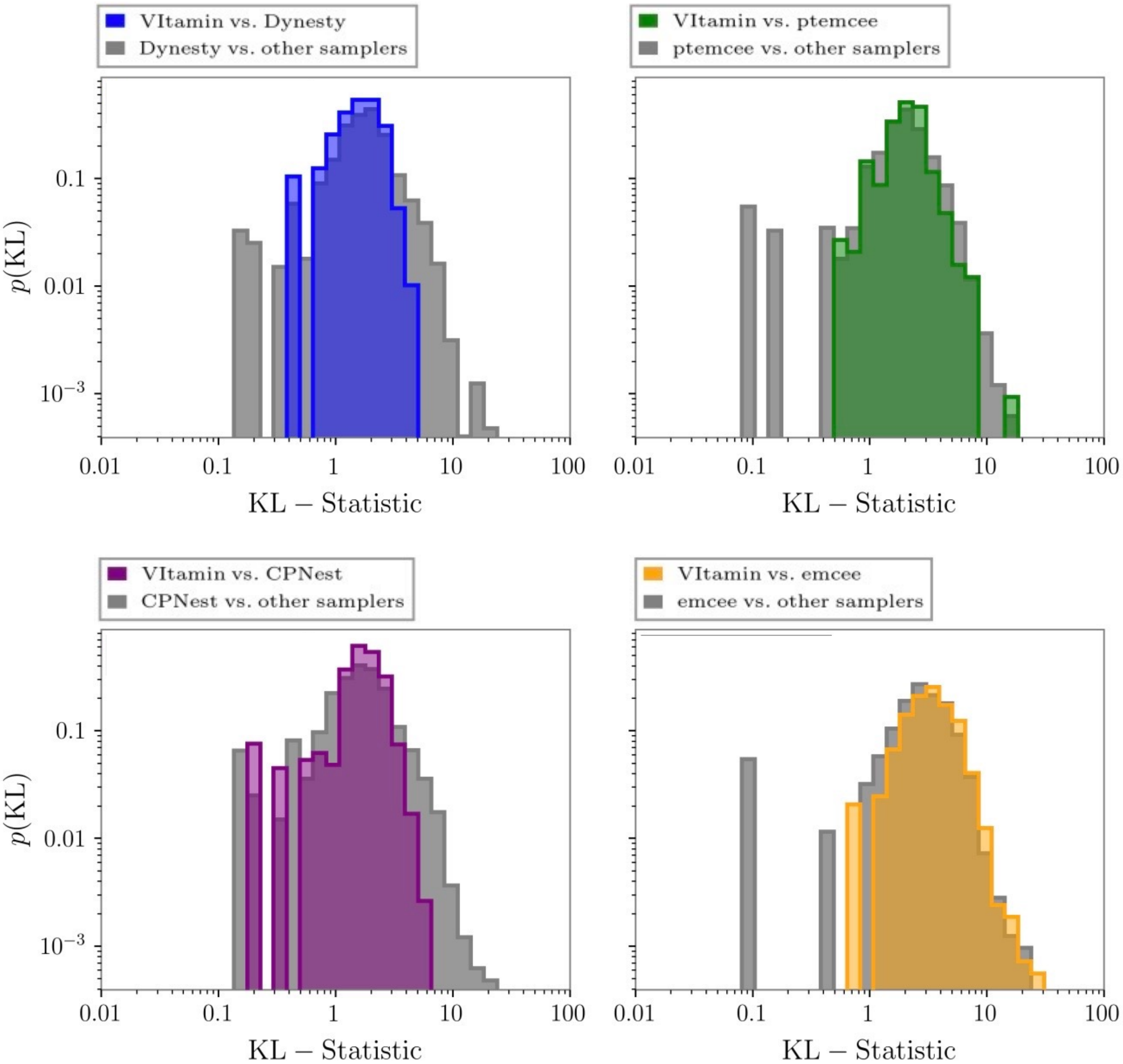


Figure 4.9: Histograms of KL divergence between `VItamin` and baseline samplers over 256 test simulations. In each plot, the grey histograms aggregate the KL divergences of the baseline being compared and the other three baselines.

Chapter 5

Applications in Human-Computer Interaction

In this chapter, the framework of chapter 3 is applied to two human-computer interaction (HCI) settings. In many novel deep learning enabled interaction modalities proposed for consumer products, inexpensive sensors are employed to record signals from users' interactions and, ultimately, infer intention. A prime example is that of gesture interaction, where a device has to record and interpret the user's gestures, typically with non-video recording sensors, such as sonar or radar. In these settings, recovering gestures and poses from limited measurements is often an ill-posed problem, where the physical sensing process does not allow to analytically uniquely reconstruct targets from measurements. Deep learning offers the possibilities to use recorded data to map signals from embedded sensors to gestures and poses. However, the two main issues identified in chapter 3 persist: i) without accurate measures of uncertainty, systems are susceptible to inference errors and ii) large training sets are required to build models for each specific system, adding considerable cost to the development of new devices. In this chapter, these two problems are addressed in HCI settings with the framework of chapter 3.

In sections 5.1, the reconstruction of finger pose from a capacitive screen's measurements is performed with the novel framework. Simulation of capacitive fields are used to generate low fidelity measurements in simulation, which are then combined with a smaller amount of real recordings with associated ground-truths in order to train the variational mdels. In section 5.2, a more complex setting of hand gestures reconstruction is considered. one-second long hand gestures are reconstructed from radar measurements recorded with the Google Soli sensor. The advantage in reconstruction accuracy and uncertainty estimation when using the novel method is demonstrated.

5.1 Capacitive Sensing for HCI

In this section, the VICI framework of chapter 3 is applied to a simple HCI application, demonstrating the advantage of modelling both the forward model with the multi-fidelity system and the inverse, making use of simulated and real data. The resulting system constitutes the sensing component of a general modelling practice for human interaction with sensors. This section is partially extracted from recent work conducted by Roderick Murray-Smith, John H. Williamson, Andrew Ramsay, Simon Rogers and Antoine Loriette [129].

Interactive systems must be able to sense and interpret human actions to infer their intentions. HCI research continually explores the use of novel sensors to enable novel forms of interaction, but lacks a coherent, consistent framework for characterising this process with incrementally improving precision for different sensors and different human behaviours. Common practice tends to be to hand-craft features and associated thresholds for specific use-cases. This can be time-consuming, especially as the dimension of the sensors increases. Furthermore, the thresholds for one application might not be appropriate for another (e.g. touch typing vs continuous gestures).

For the field to make steady progress, a more general, formal framework for characterisation of the pathway from human intent to sensor state is needed. This pathway can include formal, computational models of human elements such as cognition and physiological processes, as well as purely technical elements such as the characterisation of the physical processes of the sensor.

In this section, the framework of chapter 3 is implemented to model and invert the latter for the problem of finger pose estimation from capacitive sensing. Progress in design of capacitive screen technology has led to the ability to sense the user's fingers up to several centimetres above the screen. However, the inference of position and pose, given only the readings from the two-dimensional capacitive sensor pads, is a classic example of an ill-posed inverse problem, making the solutions inherently uncertain.

5.1.1 Experimental Set-up

A touch screen of dimensions 6.1×9.7 cm was used, with a prototype transparent capacitive sensor with an extended depth range of between $0\,\mathrm{cm}$ to $5\,\mathrm{cm}$ from the screen (although accuracy decreases with height) and a resolution of 10×16 pads and a refresh rate of 120 Hz. Sampling is performed at $60\,\mathrm{Hz}$. It was embedded in a functional mobile phone (7.3 × 13.7cm size), with an external casing making the prototype slightly deeper than normal modern smartphones. The screen was active during the measurements. It is a self-capacitance (as opposed to the more common mutual capacitance) touch screen, with a checkerboard electrode layout. Ground-truth finger pose was recorded using the Optitrack system and passive markers on the finger. Figure 5.1 shows a photo of the experimental set-up, along with a schematic representation of the sensing process. The target to be reconstructed in this setting is the finger posed, composed of the three spatial coordinates of finger tip position (x,y,z) and the two inclination angles of the finger (pitch and yaw). With this set-up, a total of $40,428$ experimental acquisitions of finger poses paired with measured capacitive matrices were collected.

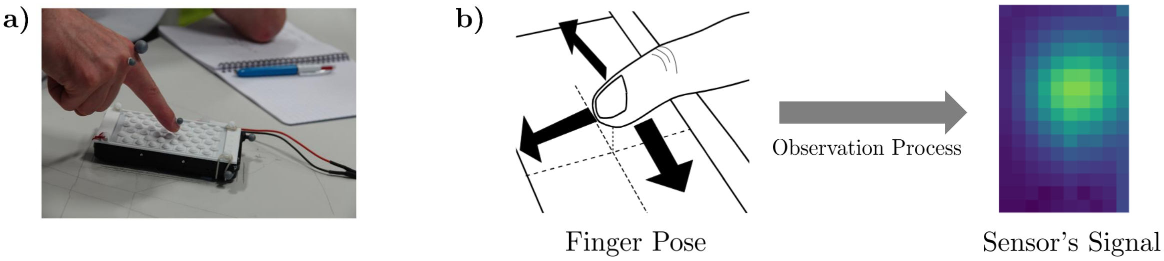


Figure 5.1: **a)** Photo of the experimental set-up, where the user holds the index finger above the capacitive screen, with the passive grey Optitrack trackers on their finger. **b)** Schematic representation of the observation process and example of capacitive sensor's readings, constituting a 10×16 intensity field.

5.1.2 Simulations and Multi-Fidelity Forward Model

A simulation of capacitive sensing was used as the low fidelity forward model within the learning framework of chapter 3. The simulation was performed using a 3D finger object created by attaching a hemisphere to one end of a cylinder. The default diameters of the cylinder and hemisphere were both $9\,\mathrm{mm}$, while the total length of the finger object was $10\,\mathrm{cm}$ (although all of these dimensions were defined as script parameters and could be easily changed). This script runs finite element method (FEM) simulation and parses the meshed charge values into total charge per plate values. These values represent the charge/capacitance matrix. Figure 5.2 shows an example of simulation along with a simulated 10×16 capacitive measurement.

The forward model component of chapter 3 was then trained with the available $40,428$ physically acquired data to reproduce physical measurements, using the simulated ones described above as low-fidelity examples. samples of multi-fidelity forward model generated measurements compared to simulated ones are shown in figure 5.3. As shown, the multi-fidelity forward model reproduces real measurements to a higher accuracy than the analytical simulations.

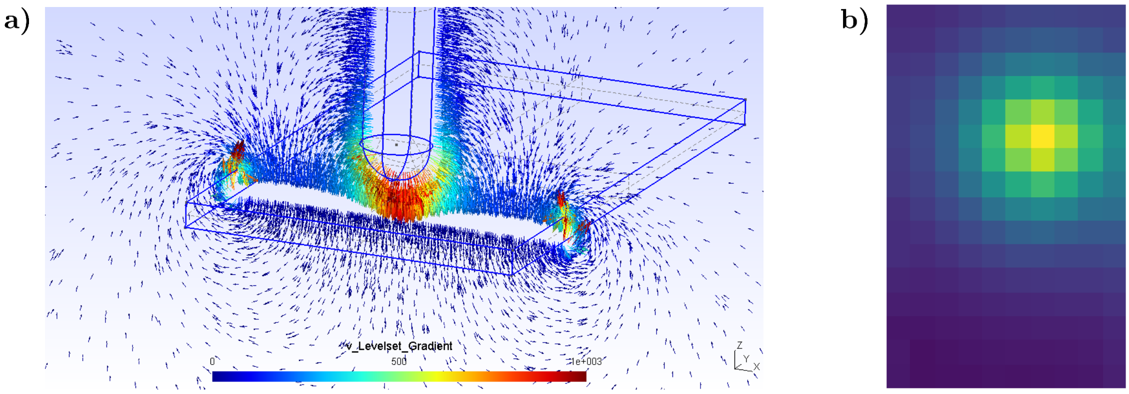


Figure 5.2: **a)** Example of FEM simulation. **b)** Example of resulting simulated capacitive measurements.

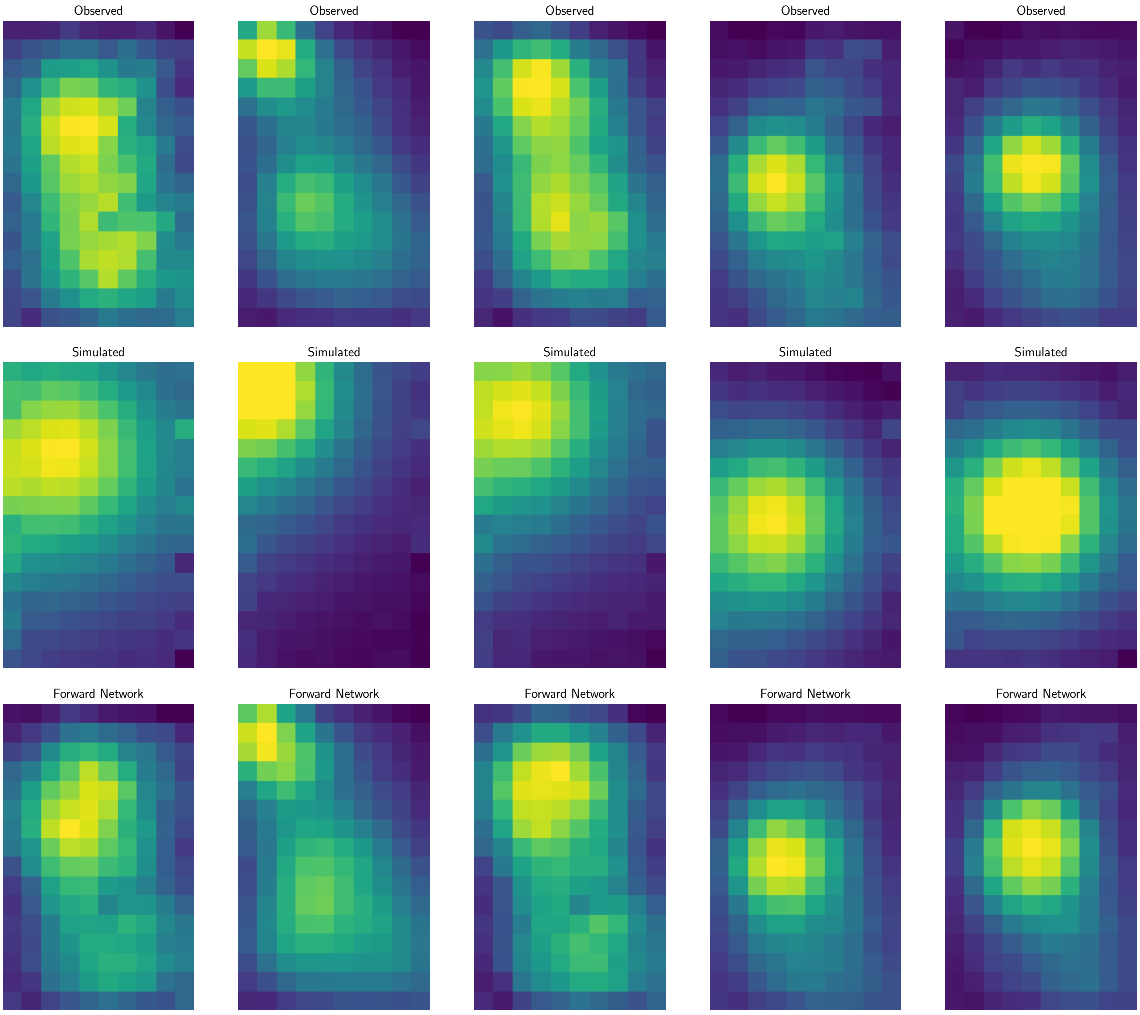


Figure 5.3: Multi-fidelity forward model generated measurements (lower row), compared to real experimental ones (upper row) and simulated measurements (middle row).

5.1.3 Inverse Model and Reconstructions

Having learned a suitable multi-fidelity forward model using the supervised available training data, this is integrated in the framework of chapter 3 to train an inverse, making use in addition of $82,231$ unsupervised hand poses, for which only the generated fingers parameters and simulated capacitive field images are available. For this unsupervised portion of the data, measurements are generated using the trained multi-fidelity forward model throughout the training of the inverse model, as described in chapter 3. Examples of test reconstructions obtained with the trained inverse model are shown in figure 5.4. The model is able to recover the finger's parameters with a good degree of accuracy and the standard deviation computed with its draws suitably adjusts to the reconstruction error in each degree of freedom, meaning that uncertainty is adequately captured.

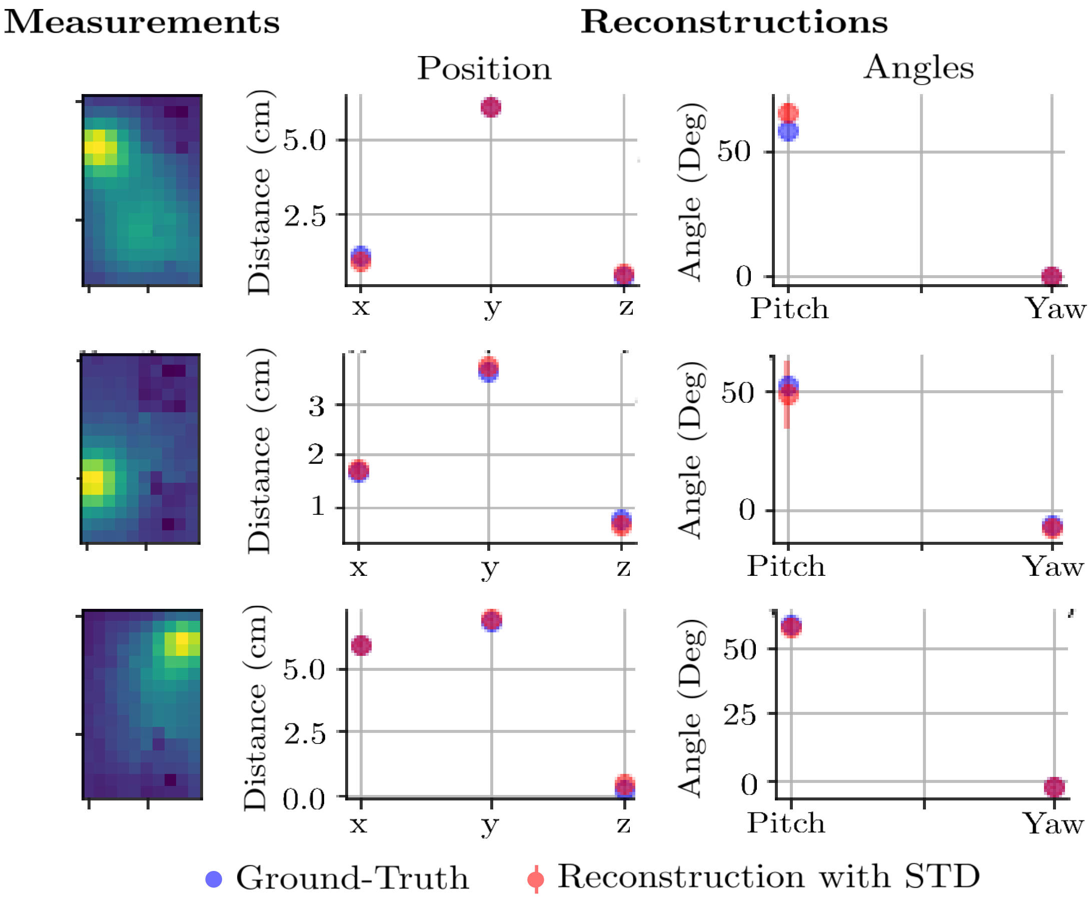


Figure 5.4: Reconstructions of finger poses' parameters using the VICI framework of chapter 3. On the left, measurements recorded with the capacitive screen are shown, while on the right the corresponding reconstructed finger pose parameters are shown with the standard deviation recovered by the model.

To compare the novel learning framework to standard approaches, the experimental portion of the data is used to train a conditional VAE having the same structure as the inverse model

Table 5.1: Average RMSE to ground-truth of reconstructions with standard CVAE and VICI framework for each coordinate and angle. Two-sample p-values are also shown for statistical significance

Coordinates	CVAE	VICI	p-value
x	0.1785cm	0.1841cm	5.21×10^{-5}
y	0.0891cm	0.0786cm	9.00×10^{-46}
z	0.1308cm	0.1098cm	9.39×10^{-144}
Pitch	9.968^o	6.291^o	$< 10^{-300}$
Yaw	7.669^o	7.151^o	9.63×10^{-24}

directly, without making use of the simulations. For this CVAE and the proposed framework, reconstructions are computed over the test set. Figure 5.5 shows reconstructed coordinates and finger's angles plotted against the corresponding ground truth. The red line in each plot corresponds to the exact reconstruction diagonal, i.e. the closer data points are to this diagonal, the more accurate is their reconstruction. The average errors for each coordinate and angle are reported in table 5.1, along with two-sample p-values to asses the statistical significance of the observed difference in mean performance between the standard CVAE and the VICI framework. Using the proposed framework results in improved mean performance in all but the x position, which remains competitive, given the small difference in mean performance. In particular, the error on the Pitch angle is greatly reduced.

The second aspect to evaluate and compare is the retrieval of uncertainty. Figure 5.6 shows the empirical standard deviation, calculated on each degree of freedom through sampling 100 times from the models, plotted against the mean reconstruction error. Points above the red line correspond to errors within one model's standard deviation, while points above the green line correspond to points within two standard deviations. Figure **??** Shows instead histograms of error distributions compared to recovered standard deviations distributions for both the baseline CVAE and the proposed VICI framework.

The CVAE model trained solely on experimental data is often over-confident about its inferences. In figure 5.6, errors fall outwith two standard deviations more often than the expected rate. This can be seen in the histograms of figure 5.7 too, where in all dimensions there is a tail in the actual errors not covered by the standard deviation distribution. The inverse model trained with the novel framework returns instead more accurate uncertainties and inaccuracies are largely present in over-estimates of uncertainty, rarely returning over-confident reconstructions. Table 5.2 reports the correlation between error to the mean and retrieved empirical standard deviation for each model. The VICI model presents a higher correlation in all variables, except the y coordinate, where the CVAE presents slightly better correlation.

Table 5.2: Correlations between mean error and retrieved empirical standard deviation for the standard CVAE and VICI models. All correlations were found to have a p-value of approximately zero, meaning they are statistically significant.

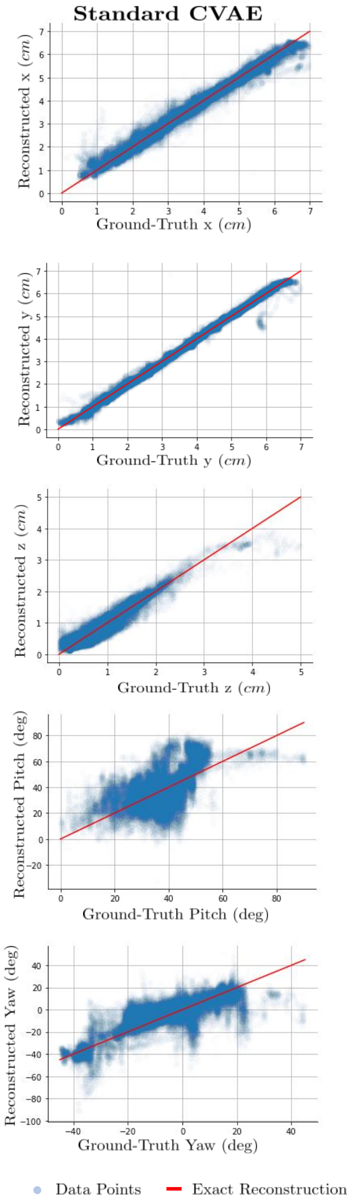
Coordinates	CVAE	VICI
x	0.487	0.498
y	0.455	0.445
z	0.410	0.462
Pitch	0.213	0.282
Yaw	0.425	0.493

The fact that the uncertainty retrieved by VICI, on average, correlates better with its error than the standard CVAE means that using the extra information available through this model results in better error calibration.

The difference in performance between a model trained using solely the available experimental data and one trained with the novel VICI framework of chapter 3 is even more evident when the number of available experimental training examples is low. Figure 5.8 shows reconstruction performance metrics obtained with the two training methods at varying number of available experimental training data between $5,000$ and $15,000$ examples. For the VICI framework, the number of simulated examples incorporated in the training procedure is kept constant as before at $82,231$. In low experimental data regimes, the mean performance of the standard CVAE is significantly lower (resulting in higher RMSE) than the VICI framework until the number of experimental samples exceeds $12,000$, at which point the performance of the two frameworks is comparable. The probabilistic performance, measured by the ELBO, is consistently better and much less sample size dependent for the VICI framework. These results show how the proposed VICI framework provide sensibly improved performance when experimental data is scarce and makes models less dependent on its availability, reducing the need to perform physical experiments in order to train the model.

5.1.4 Discussion

This section demonstrated the application of the VICI framework of chapter 3 to the HCI scenario of finger pose reconstruction from capacitive measurements. The results shown with this simple interaction example outline two important advantages of applying the VICI framework in HCI. Firstly, the retrieval of uncertainty measures. As reconstructions of physical user interactions from sensors readings are often ill-posed, simply applying standard machine learning techniques can easily yield compelling reconstructions that are far from the ground-truth targets. Accurate quantification of uncertainty is essential to machine learning enabled interaction, as systems need to propagate such uncertainty to make robust decisions. The framework demonstrated above is capable of recovering such uncertainty quite accurately, an therefore can be integrated as the physical interaction component of an HCI system robust to ill posed inversions. Secondly, the proposed framework allows to incorporate physical models of the sensing process and simulated gesture into the learning procedure, providing improved performance with no extra collection efforts. As often HCI system rely on sensing processes the physics of which is well understood, this capability can enable machine learning to be applied to HCI system with limited physical data collection. This is especially relevant in HCI where physical collections need to be acquired with multiple human users and therefore this process is generally time consuming.



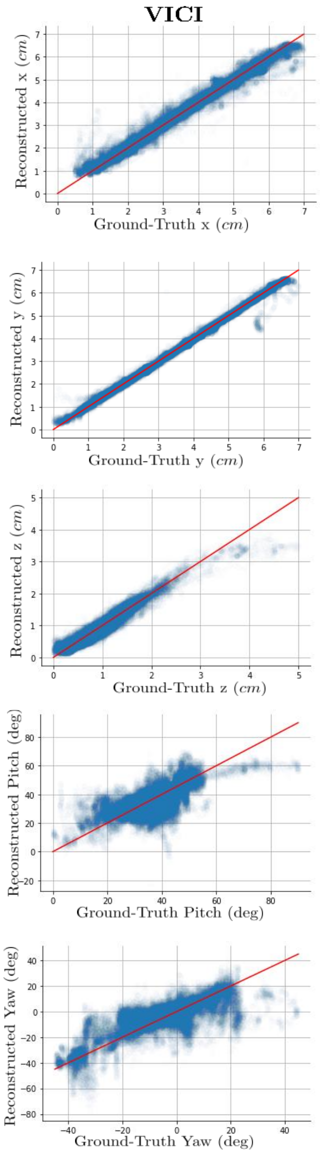


Figure 5.5: Mean reconstruction of each degree of freedom plotted against ground-truths for the CVAE trained on paired experimental acquisitions only and incorporating simulations using the VICI framework.

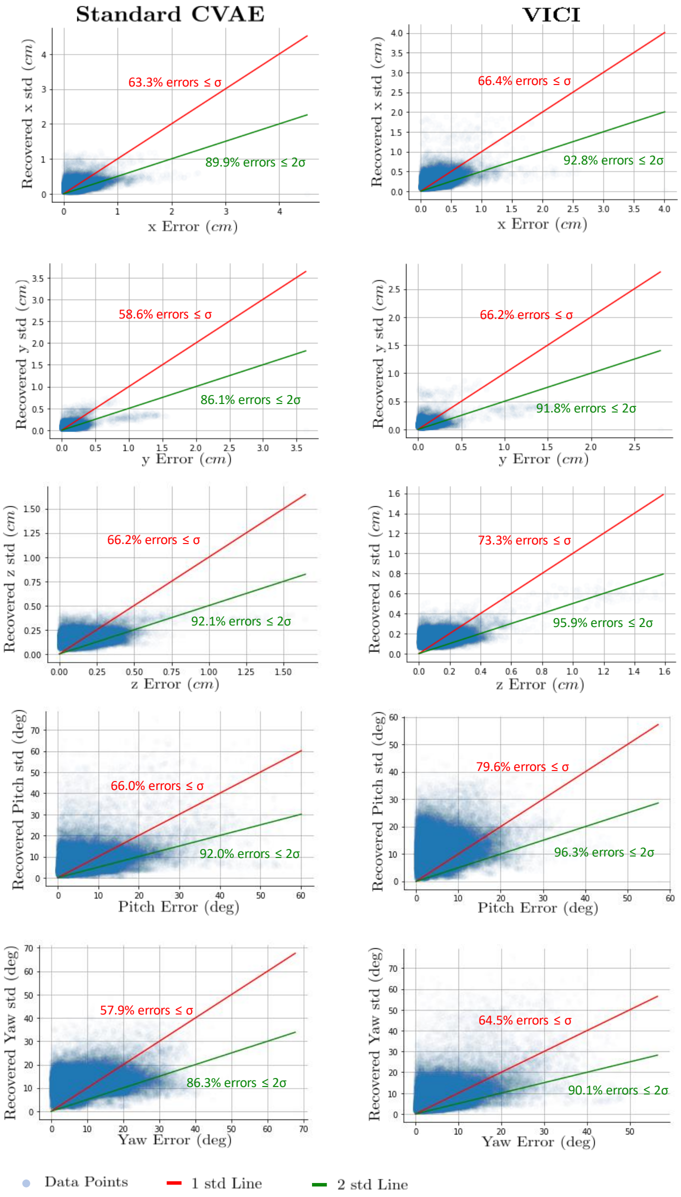


Figure 5.6: Empirical standard deviation recovered plotted against reconstruction error. On the left, the results obtained with the standard CVAE are shown, while on the right, the results obtained with the VICI framework are shown.

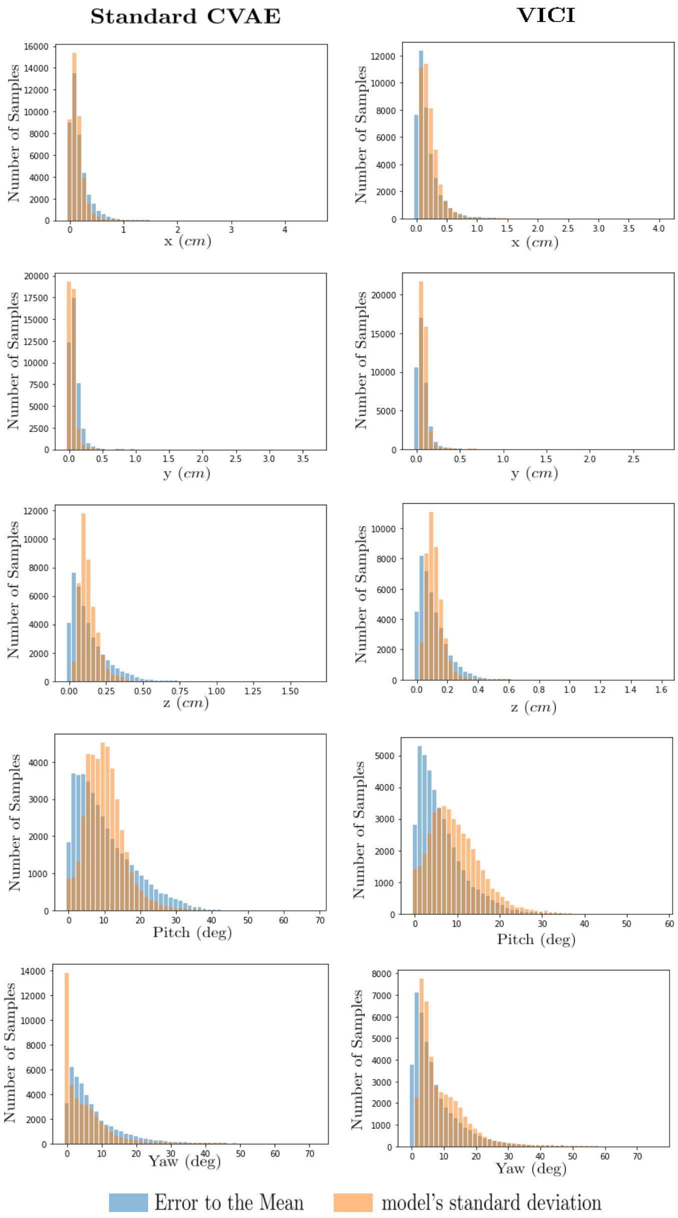


Figure 5.7: Distribution of mean errors compared to distribution of standard deviation retrieved with each model.

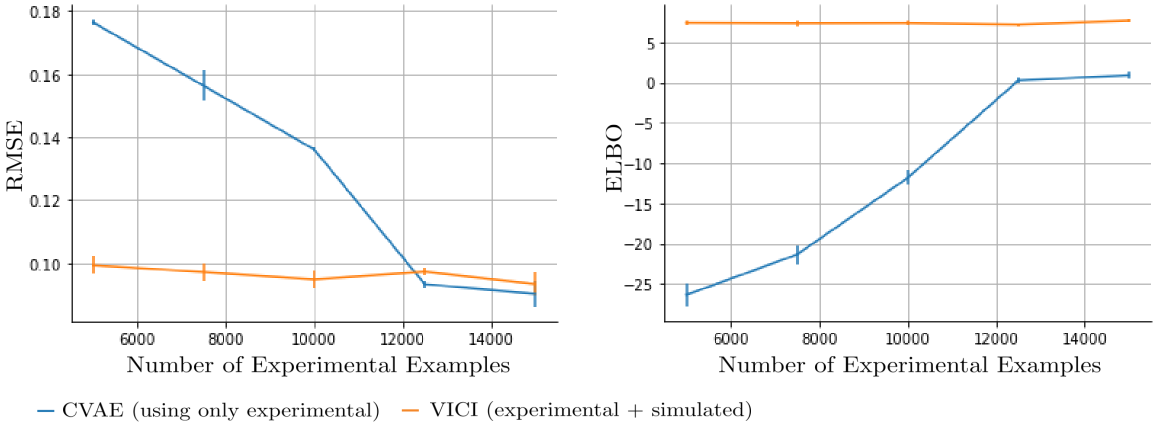


Figure 5.8: Comparing the performance of a variational model trained with a standard CVAE architecture using experimental data only and experimental and simulated data in combination using the VICI framework. The graphs show two different performance metrics at varying number of available experimental examples. On the left, the RMSE to the recovered mean and on the right the ELBO. Each experiment is repeated four times in order to obtain error bars.

5.2 Radar for Gesture Interaction with Google Soli

The framework presented in chapter 3 was applied and tested with the Google Soli millimetre wave radar sensor [130, 131] for hand gesture inference. The Google Soli sensor is a 60GHz, 1Tx 3Rx radar, built to sense motions through frequency modulated continuous wave (FMCW) Doppler sensing. The high frequency allows to achieve a resolution in the order of centimeters and velocities on the order of a tenth of a meter per second. In addition, the three receivers allow to recover coarse angle of incidence through phase differences. The resulting signal is processed into complex range doppler (CRD) images of size 64×8 (8 range bins and 64 velocity bins). Inferring hand gestures from a time series of these CRD images is a challenging inverse problem, as scattered signals from different parts of the hand can super-impose in the same time-velocity bins.

This is an important and representative inverse problem in HCI, where a relatively inexpensive sensor, such as Soli, registers signals associated with with the human interaction, but from which it is far from trivial to recover it. This application outlines the importance of two features of the framework presented in chapter 3:

1. Capturing uncertainty in inferences. As many inverse problems of interest, mapping radar signals to hand gestures is an ill-posed problem, meaning that for a given signal there may be many distinct possible hand gestures. In this setting, it is also difficult to predict which gestures can be recovered with good uncertainty and which are instead difficult to distinguish for the Soli sensor. The framework of chapter 3 recovers complex PDFs of possible reconstruction, hence characterising this uncertainty in the inferred hand motions.

2. Incorporating domain expertise and unsupervised data to reduce data collection. In order to train a machine learning model to reconstruct hand motions from the Soli signals, fairly large training data sets of paired hand motions and Soli recordings are needed. These acquisitions are specific to the particular Soli hardware configuration and if the sensors improves or is embedded in a different device they would need to be completely re-acquired. The framework of chapter 3 trains such models augmenting the training set with unsupervised data and domain expertise, minimising the paired acquisitions requirements.

5.2.1 Experimental Acquisition

The physical experiments to acquire paired gestures and Soli signals was performed by Andrew Ramsey and the following set-up description is partially extracted by a technical report to which he contributed. All data gathered for this section used the Soli hardware and a

NaturalPoint OptiTrack motion tracking system with passive infrared markers to acquire hand gestures ground truths. 3 Prime 13W cameras were positioned around the workspace and linked to a PC running the NaturalPoint "Motive" application (version 2.2.0, running on Windows 10). The cameras were calibrated using the recommended process, and all IR-reflective objects in the fields of view were removed to avoid extraneous markers being detected. The frame rate for all cameras was set to 100Hz. Figure 5.9 shows the physical recording setup.

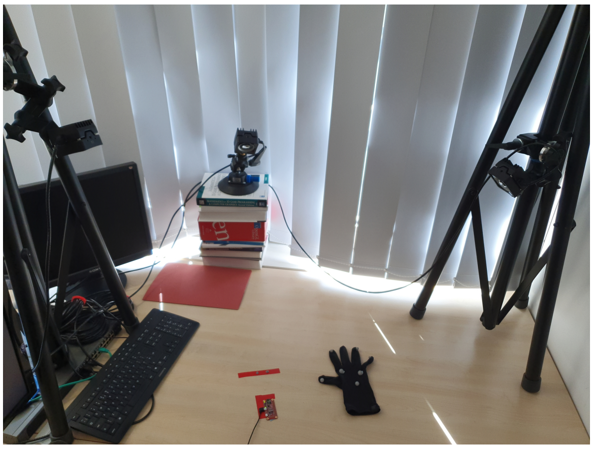


Figure 5.9: The OptiTrack recording setup showing the 3 cameras positioned around the workspace and the Soli board on the desk.

A total of 7 passive IR markers were used to track hand motion. The marker locations were: tip of thumb, tip of index finger, knuckle of index finger, tip of ring finger, tip of little finger, knuckle of little finger, and back of palm. Figure 5.10 shows the markers attached to a glove to help ensure consistent positioning across sessions. More markers could have been used, but this can increase the chances of them becoming temporarily obscured during certain movements. This in turn significantly increases the amount of time-consuming, manual post-processing required.

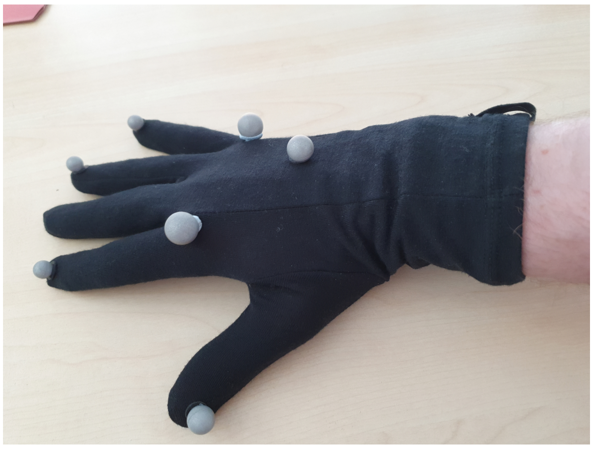


Figure 5.10: The 7 passive IR markers used to track hand motions.

With the set-up described above, two types of hand gestures were performed and recorded and used to build two distinct data sets:

1. **Coarse gestures:** The first set of data recorded consisted of motions using the whole hand, held in a mostly rigid posture with fingers parallel to the palm and thumb naturally held at the side in the same plane. Motions used include horizontal swipes, vertical bounces, circling over/around the Soli, and random movement within the 3D space above the Soli.

2. **Fine motion gestures:** Subsequent data recording sessions focused on acquiring finer-grained finger movements. A total of 5 types of gestures were recorded as part of this data set: i) hand held horizontal over the Soli, with fingers together, while moving thumb around, ii) starting from a horizontal and rigid hand position, bend all 4 fingers back towards the palm at the middle knuckle while keeping the thumb extended to the side, iii) continually drumming fingers in mid-air over the sensor, as if on a table, iv) holding the hand horizontally while splaying all fingers in and out repeatedly and v) keeping each finger straight and bending them down from the first knuckle individually.

For each type of gesture described above, time series of Optitrack markers' locations and paired Soli signals are recorded. The Soli signals are processed to obtain complex range-doppler (CRD) 2D fields at each time frame, as it is standard for frequency-modulated continuous wave (FMCW) radar [131]. The CRD is a complex 2D field, where the horizontal axis corresponds to source velocity, the vertical axis to source distance and the phase in each pixel to the source angle of incidence. The resulting data consists of time series of spatial coordinates for each marker and paired CRD videos (2D fields at each time frame).

5.2.2 Analytical Forward Model

The cheap low fidelity measurements in this context are analytically simulated Soli CRDs, given a hand gesture. As the Physics of FMCW radar sensing is well understood, it is possible to efficiently simulate these in closed form. Segments connecting the tracked markers on the hand are filled with an arbitrary number of scattering points, which are assigned a linear combination of diffusive and reflecting property. For each time frame in the Optitrack time series of markers locations, the position and velocity of each scattering point are computed and used to simulate the corresponding CRD.

The radar observation model is assumed to be linear, meaning that a CRD is computed for each scattering point and the final simulated CRD is obtained by summing the individual CRDs computed for each point. A CRD for a single scattering point is computed as follows:

1. The distance r from the Soli sensor and its derivative with respect to time v are computed.

2. A Gaussian $2D$ field of dimensions $samples\_per\_chirp/2 \quad \times \quad chirps\_per\_burst$ is generated, having the Gaussian mean in each dimension as $\mu = (r/Dr, v/Dv)$ and standard deviation as $\sigma = (1, 1)$, where $Dr = \frac{c}{2B}$ and $Dv = \frac{cPRF}{2f_c l}$. The Gaussian is then multiplied by an intensity coefficient $I = \frac{a}{r^2} + \frac{b}{r^4}$, where a and b are coefficients determining the diffusive/reflective behaviour. This is the simulated amplitude range-Doppler (ARD) field.

3. The angles of arrival in the xz and yz planes θ and ϕ are computed.

4. θ and ϕ are converted to phase differences in the rx signals as $\Delta\theta_{2,0} = 2\pi(a/\lambda)\sin(\theta)$ and $\Delta\theta_{2,1} = 2\pi(a/\lambda)\sin(\phi)$.

5. two phase matrices are created, each of dimensions $samples\_per\_chirp/2 \times$ $chirps\_per\_burst$. The first, is set to $\Delta\theta_{2,0}$ where $ARD > 0.1 * \max(ARD)$ and 0 otherwise. Similarly, the second is set to $\Delta\theta_{2,1}$ where $ARD > 0.1 * \max(ARD)$ and 0 otherwise.

Figure 5.11 shows some examples of analytically simulated CRDs compared to experimentally recorded ones.

5.2.3 Data Sets

From the data acquisitions described in section 5.2.1 two distinct data sets are built; one for coarse hand motions and one for the finer finger movements. From the time series recorded in each setting, one second long time series of paired CRDs and Optitrack data are extracted. These time series are obtained by stepping a one second long window by 25ms through the recordings. Each one second long hand motion series is then used to generate low fidelity CRDs using the simulation described in section 5.2.2. The resulting data sets are as follows:

1. a data set of 4100 one second long coarse hand motions, along with corresponding experimental CRDs and analytically simulated CRDs.

2. a data set of 4700 one second long fine motion gestures, along with corresponding experimental CRDs and analytically simulated CRDs.

In both cases, the data has the same format; the hand motion are 100×21 1D time series with 21 channels, each corresponding to one coordinate of an Optitrack tracker (x, y, z for each of the 7 trackers). The CRDs, both experimental and simulated, are $8 \times 64 \times 25 \times 3$ 3D volumetric

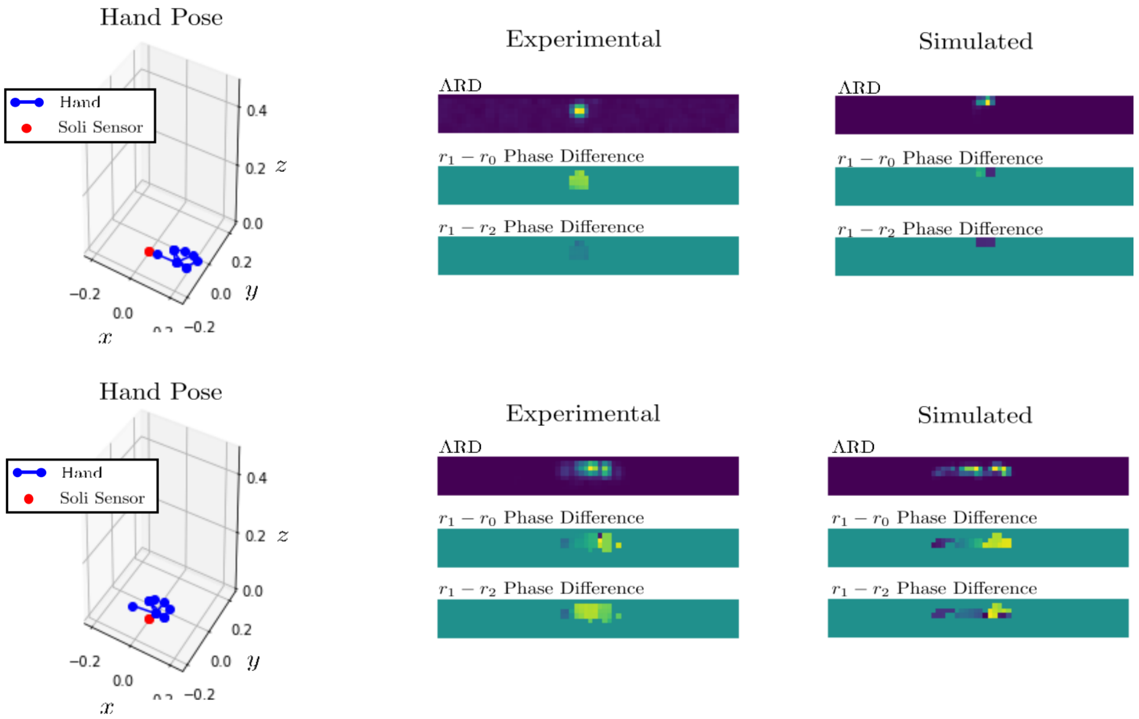


Figure 5.11: Example of simulated Soli CRD frames compared to experimentally recorded ones. Real hand gestures recorded with the Optitrack system are used to generate simulated Soli CRDs, which are compared to those experimentally recorded with the Soli sensor.

samples, having 8 range bins, 64 velocity bins, 25 frames and 3 channels, corresponding to ARD and the two phase differences between antennas. As the velocities observe with the hand motions are limited, before using them as inputs/outputs, the CRD sizes are reduced by taking only the central 32 velocity bins. This makes learning faster without losing significant information. Before being used as inputs/outputs for the framework of chapter 3, the CRDs are converted to two real and two imaginary fields. The structure of the data is shown in figure 5.12.

To test the framework of chapter 3, these data sets are split in three parts; i) a supervised training set, ii) an unsupervised training set and iii) a test set. For the supervised training set (i), all data is taken, assuming this is the data collected synchronising Optitrack and Soli acquisitions. For the unsupervised training set (ii), only the hand motions and simulated CRDs are taken, as for these it is assumed that only hand models are available, but no synchronised Soli acquisitions. For the test set (iii), a small amount of hand motions and paired experimental CRDs is set aside to test reconstruction.

5.2.4 Multi-Fidelity Forward Model

The multi-fidelity forward model is built to recover high fidelity estimates of CRDs from Optitrack hand motions and simulated CRDs, which are considered to be low-fidelity estimates.

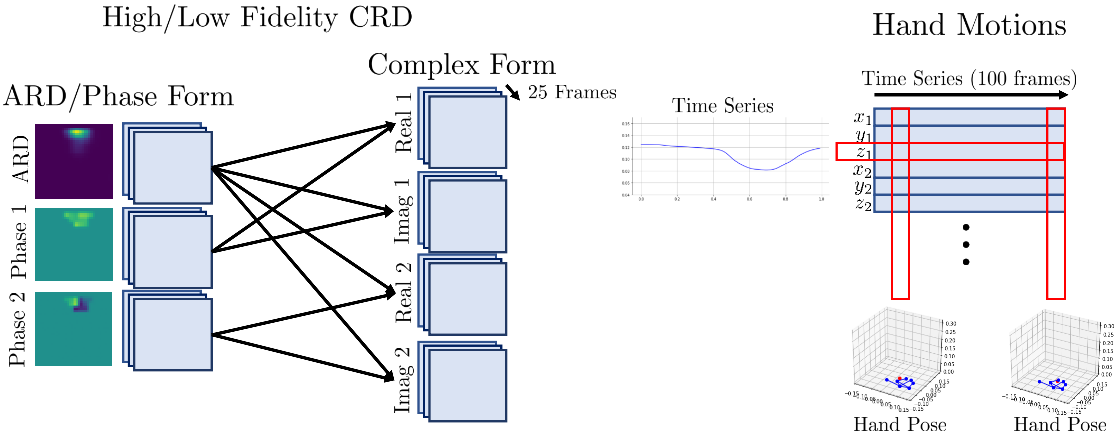


Figure 5.12: Data formats used to train the framework of chapter 3 in the Soli setting. The samples consist of one second long time series of hand motions paired to Soli sensor's CRDs, in the complex format shown in the figure, where the four CRD channels correspond to real and imaginary components of complex fields having as amplitude the ARD and as phases the phase differences between the central antenna and the other two.

This forward model is built to operate on a frame-by-frame basis, instead of processing entire one-second-long time series. Data is extracted from samples of the format described in section 5.2.3 as shown in figure 5.13. The forward model was built with fully connected networks for its components.

Two different multi-fidelity forward models are trained using the supervised components of each data set described in section 5.2.3 respectively. Examples of hand poses, analytically simulated CRDs and muti-fidelity generated CRDs for two different motions are shown in figure 5.14. With the multi-fidelity forward model, Soli CRDs can be generated with high fidelity and can therefore be used to train the inverse, augmenting the supervised portion of the data by providing CRDs for the unsupervised portion of the set.

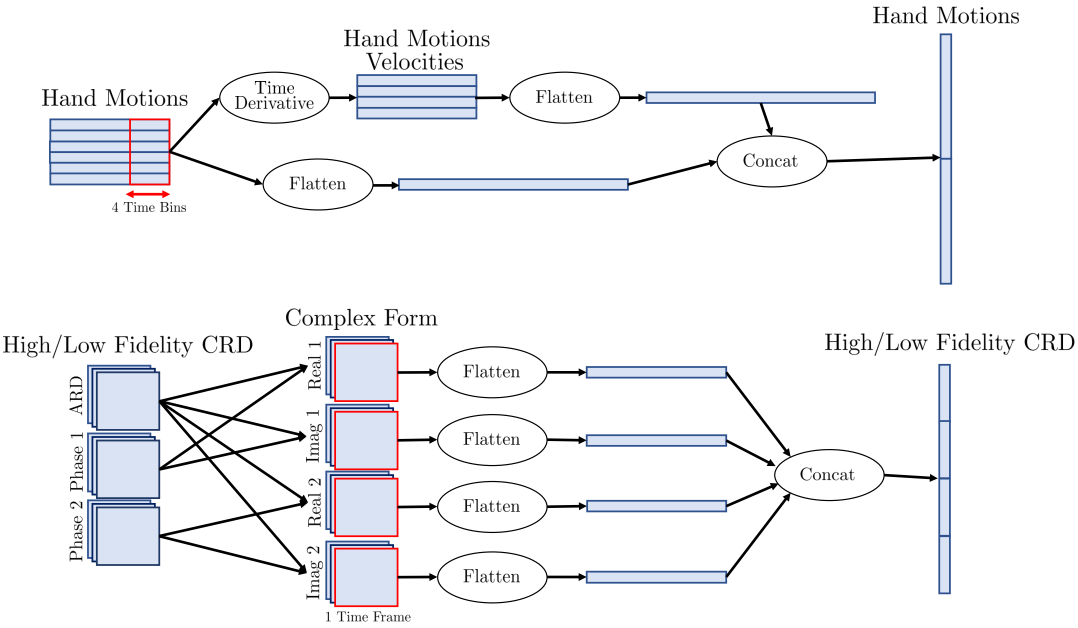


Figure 5.13: Data pre-processing used for training the multi-fidelity forward model architecture. Four time bins of markers' positions are extracted from the longer time series and their velocities are computed with simple element wise gradient. The positions and velocities are flattened and concatenated to give a single vector. Single time frames are extracted from the CRDs in the complex form, flattened and concatenated to form a single vector. The two vectors constitute an input–output pair for the forward model.

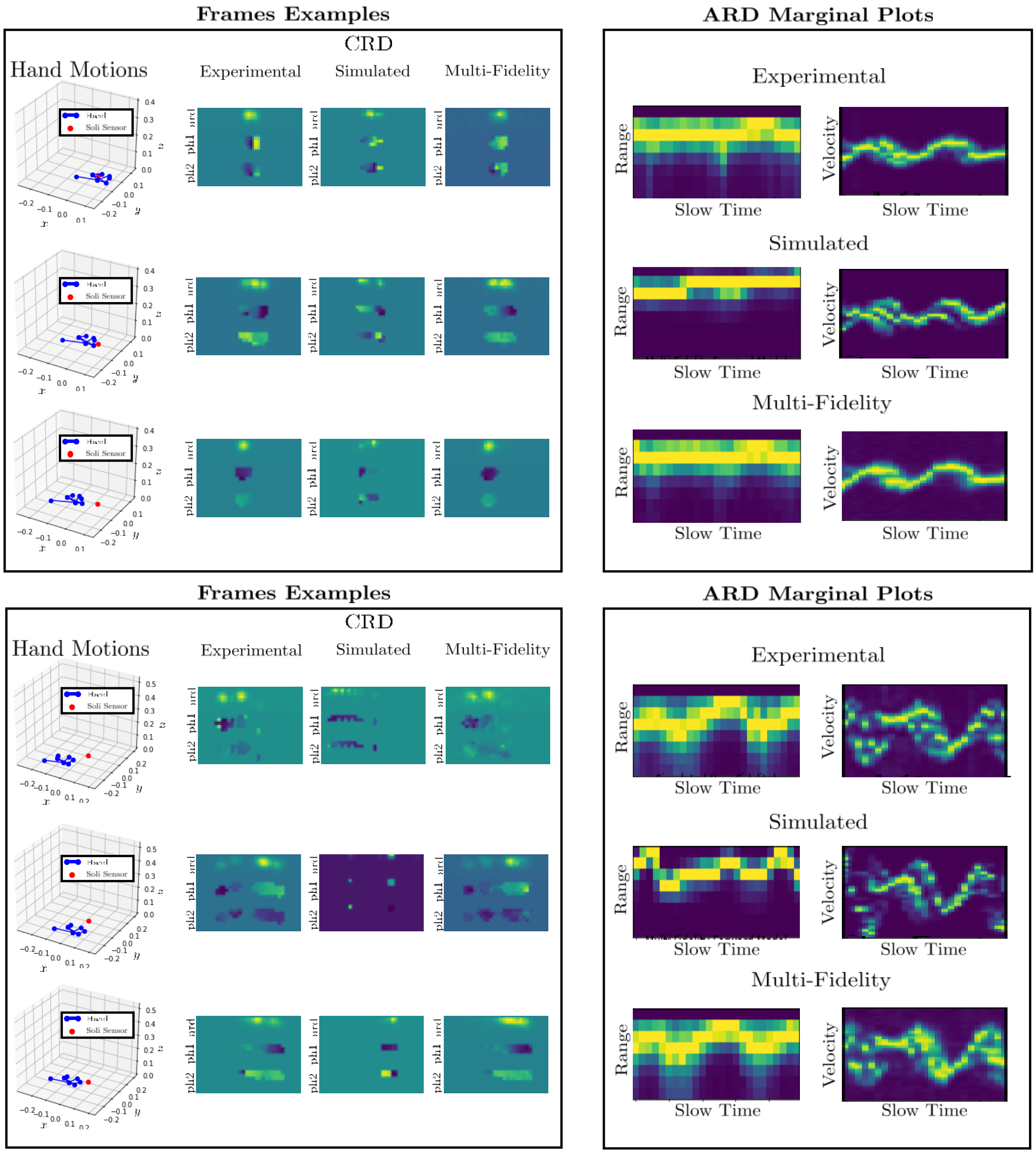


Figure 5.14: Examples of multi-fidelity generated frames for one second long hand motions using the multi-fidelity forward model. Simulated here refers to analytically simulated data and multi-fidelity refers to data simulated using the trained multi-fidelity forward model. The multi-fidelity model is able to generate both amplitudes and phase differences with much higher fidelity than analytical simulations.

5.2.5 Experimental Reconstruction Results

in these experiments, the supervised sets, for both coarse and fine hand motion sets, are composed of 2000 examples. The unsupervised sets are composed of 2000 and 2600 examples respectively. Finally, 100 examples are reserved for each set as test. With these, two sets of experiments are performed, each using either coarse or fine motion data. As a baseline to compare performance, a conditional VAE is trained to reconstruct hand motions from CRDs using only the supervised portion of the set. To test the novel framework, the model is trained as described in chapter 3, making use of both supervised and unsupervised portions of the data, as well as the multi-fidelity forward model, previously trained on the supervised set as described in section 5.2.4. Testing is performed by sampling from the trained models using as inputs the test CRDs and hence form distributions of possible reconstructed hand motions. Figures 5.15 and 5.16 show examples of reconstructions' drawn from the recovered posteriors with coarse and fine hand motions respectively.

These qualitative examples highlight the advantage of augmenting the data during training using the VICI framework of chapter 3. The VICI recoveries are visibly more accurate and the different draws explore more adequately the uncertainty space of the retrieval task, while training with 2000 paired examples alone yields less accurate results overall, but also collapsed draws that are all similar to each other, failing to capture the uncertainty in the recovery.

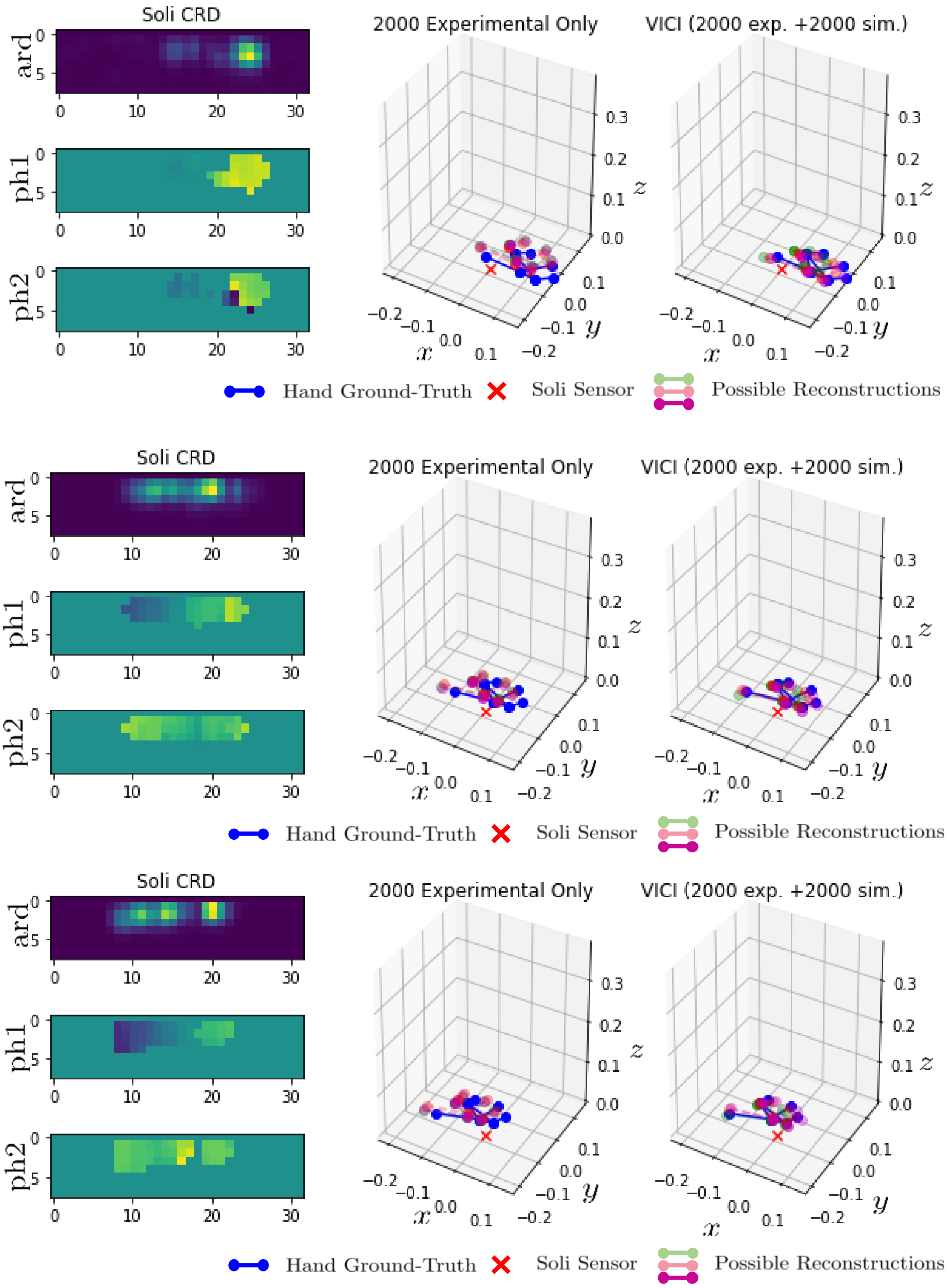


Figure 5.15: Example of reconstruction samples for the coarse hand motions by a CVAE, first trained using only the supervised portion of the set and then trained with the VICI framework of chapter 3, making use additionally of the forward model and unsupervised data.

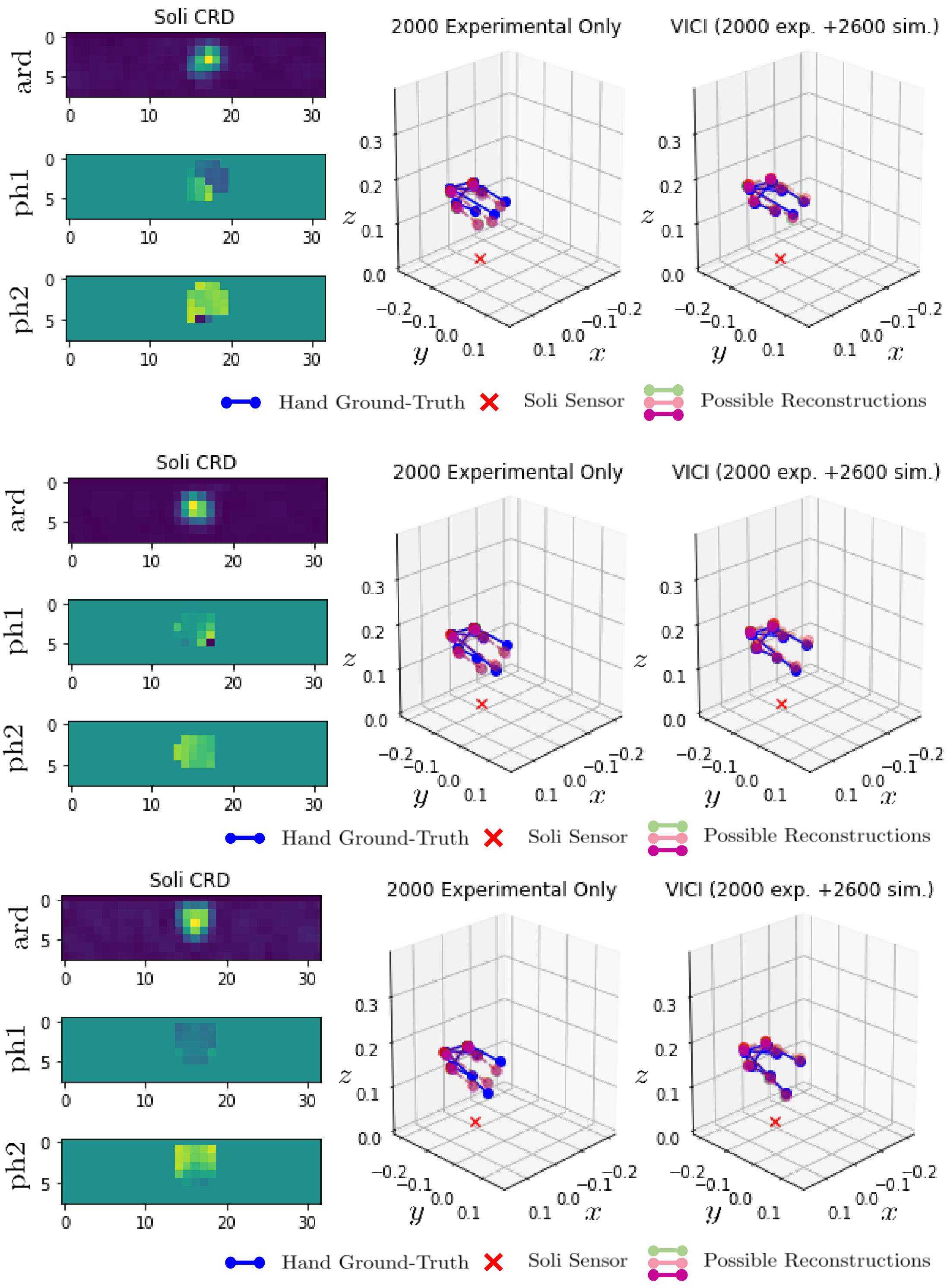


Figure 5.16: Example of reconstruction samples for the fine hand motions by a CVAE, first trained using only the supervised portion of the set and then trained with the VICI framework of chapter 3, making use additionally of the forward model and unsupervised data.

Quantitative Evaluation

To quantitatively assess the performance of the VICI framework, different reconstruction metrics are computed and compared between the two methods. These are as follows:

1. **Per-frame RMSE:** The root mean square error (RMSE) between the mean of the recovered hand poses and the ground truth is computed in each frame in each test motion. This is a metric which measures how close the mean reconstruction is to the target, and therefore it is a deterministic measure.

2. **Gaussian Fit Log Marginal Likelihood:** All entries in the reconstructed hand motions are taken and a Gaussian is fit on each of them using the samples from the model. The joint log likelihood these Gaussian assign to the ground truth test motions is then evaluated. This is a probabilistic measure, but only of the marginal distributions per sample, as it takes account of the variance of reconstruction in each entry, but discards correlations.

3. **ELBO:** The ELBO assigned to the test set by the models is computed. This is the most commonly used measure of performance for VAEs. It is an approximation (a lower bound) to the true data log likelihood. Compared to the previous measure, it can capture the full complexity of the true PDFs, but it is less robust.

Each of these metrics is evaluated with the two training strategies and compared. Results are shown in tables 5.3 and 5.4.

Both qualitative and quantitative comparisons demonstrate that, in the situation where experimental training data is scarce, using the framework of chapter 3 provides considerable advantage. The novel framework manages to incorporate into the training of models both unsupervised hand motion examples, which can be simulated or generated with generative models, and physical models of the Soli sensing process.

Table 5.3: Reconstruction performance metrics with the coarse hand motions for direct training of a CVAE with 2000 supervised examples and training with the VICI framework of chapter 3, making use of 2000 supervised examples, available unsupervised examples and physical models of the Soli sensor (incorporated in the multi-fidelity forward model).

	Per-Frame RMSE	Gaussian Fit Log Marginal	ELBO
Exp. Only (2000 Examples)	0.088775	cm -4124.207	3197.079
VICI Framework (2000 Sup + 2000 Unsup)	$\mathbf{0.073691}$cm	$\mathbf{-1826.493}$	$\mathbf{6679.111}$
Two-sample p-values	6.1×10^{-37}	2.1×10^{-167}	5.8×10^{-277}

Table 5.4: Reconstruction performance metrics with the fine hand motions for direct training of a CVAE with 2000 supervised examples and training with the VICI framework of chapter 3, making use of 2000 supervised examples, available unsupervised examples and physical models of the Soli sensor (incorporated in the multi-fidelity forward model).

	Per-Frame RMSE	Gaussian Fit Log Marginal	ELBO
Exp. Only (2000 Examples)	0.042554cm	-1659.566	6393.266
VICI Framework (2000 Sup + 2600 Unsup)	$\mathbf{0.035248}$cm	$\mathbf{-949.092}$	$\mathbf{7344.824}$
Two-sample p-values	5.1×10^{-14}	1.3×10^{-187}	1.5×10^{-144}

Examples of Reconstructing Fine Motions

When training the model with the novel approach of chapter 3, combining limited supervised and unsupervised data, fine motions can be reconstructed with a good degree of accuracy, recovering the location in time of individual trackers on the hand with centimeter accuracy with a wide variety of gestures. In figures 5.17 and 5.18 the particular case of pressing down with individual fingers is examined in detail. The model trained with the novel framework is able to recognise and reconstruct the motions of individual fingers pushing down, while also returning accurate measures of uncertainty. These are results from the model trained with the full variety of fine motions, and therefore it is trained to reconstruct a wider range of gestures and not just those shown in figures 5.17 and 5.18.

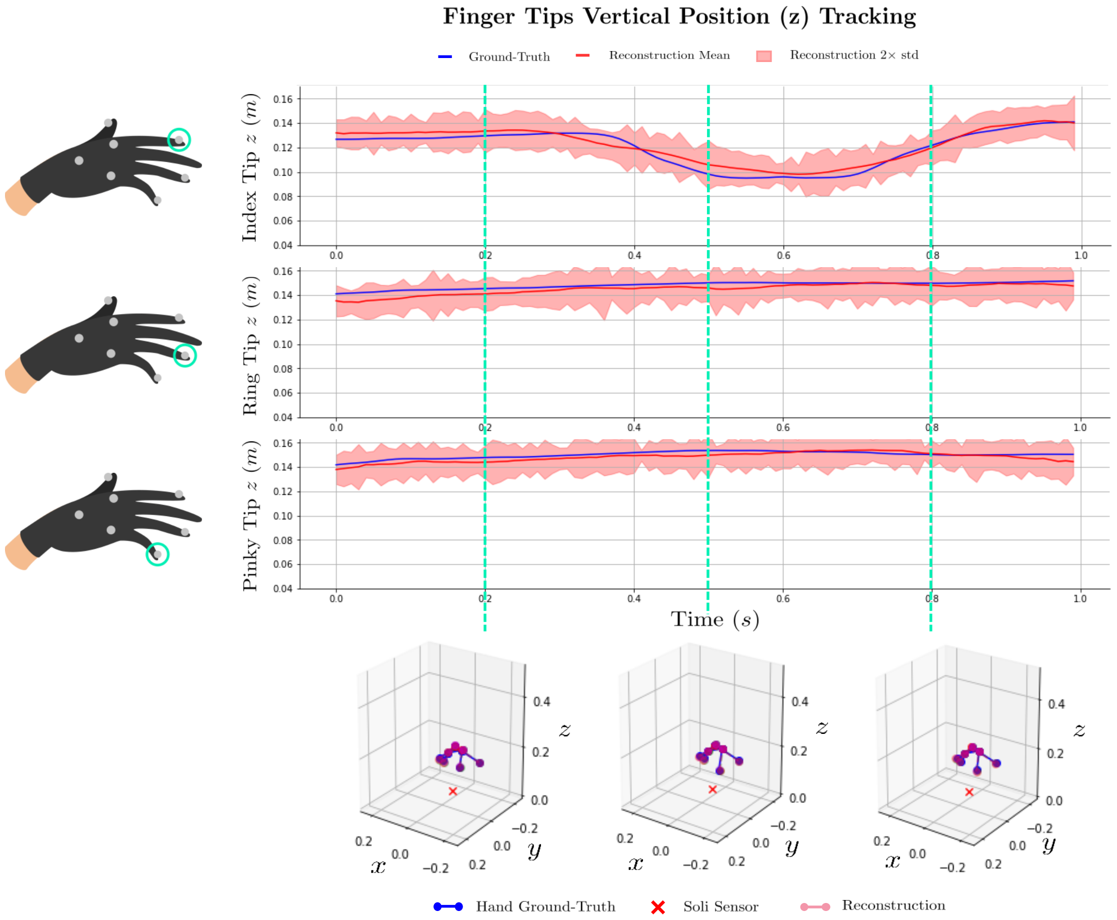


Figure 5.17: Reconstruction of the hand while pushing down with the index finger. The model trained with the novel framework is able to accurately track the location of the index finger pushing down, while recognising the other fingers are not moving. It also returns a suitable measure of uncertainty for each marker.

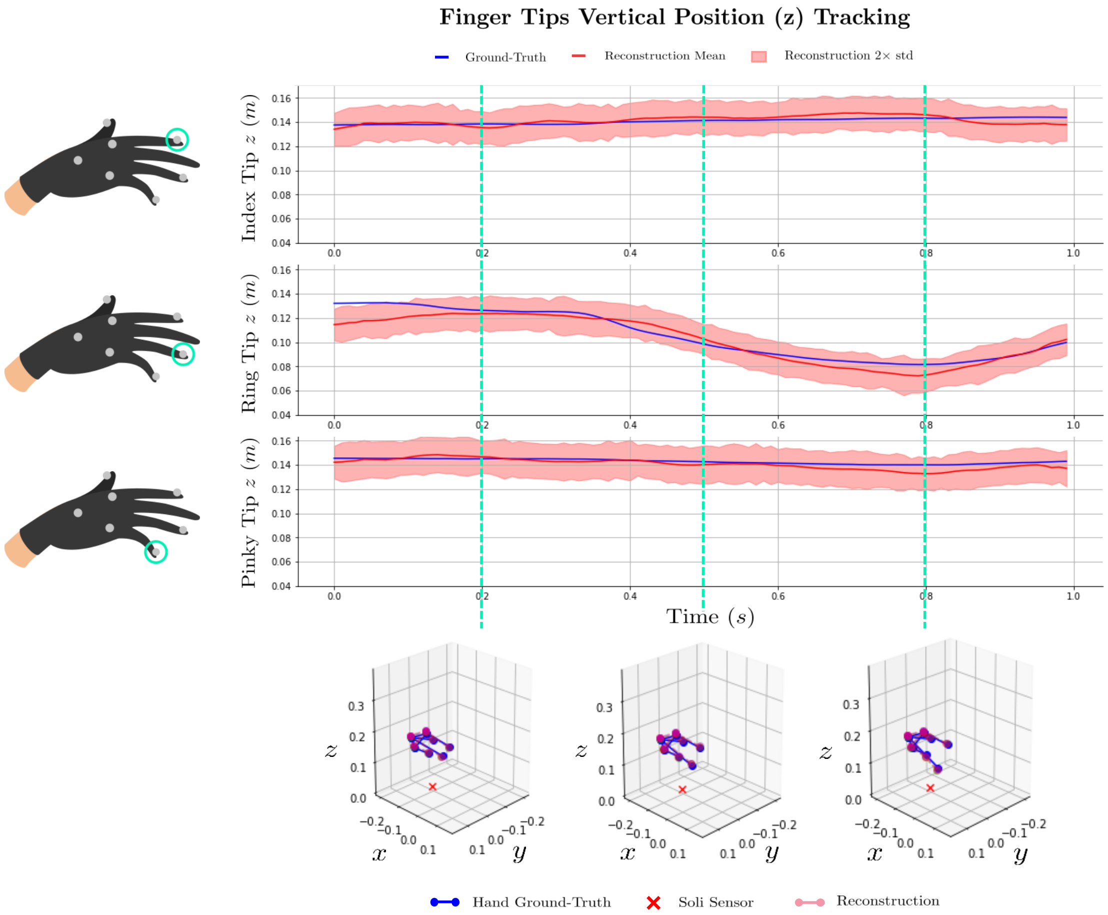


Figure 5.18: Reconstruction of the hand while pushing down with the ring finger. As in figure 5.17, the model trained with the novel framework is tracking the full hand motion, also returning a measure of uncertainty. Here too, the model accurately tracks the movement of the finger being pushed down, while recognising the other markers are fairly stationary.

Examples of Marginal Plots

Once a suitable probabilistic recovery model has been trained, one can distill any statistical measure of interest from the model's samples. For instance, figure 5.19 shows a marginal map of hand position accuracy as a function of distance and inclination angle from the sensor. Figure 5.20 shows an analogous map of tracker accuracy. These maps were built with unseen data from the test set only. The model's own estimate of uncertainty (right-hand plots) correlates well with the true recovery precision (middle plots). This means that measures such as these ones can be recovered even without ground-truth data, but simply by acquiring data with the Soli sensor alone, in normal operation.

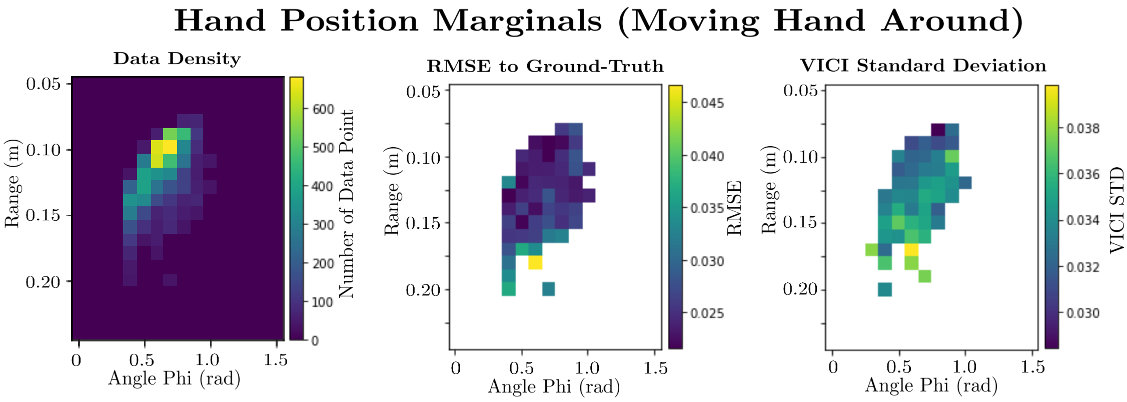


Figure 5.19: Plots of hand tracking average accuracy as a function of range and inclination angle (Phi) from the Soli sensor. On the left, the test data density is plotted over the same axes, in order to show how much data was available to compute this measure in each quantised location. In the middle, the empirical tracking precision is shown, measured by the RMSE to the ground-truth. On the right, the average standard deviation is shown, computed using draws from the trained model only, hence without the need to use ground-truth hand motions.

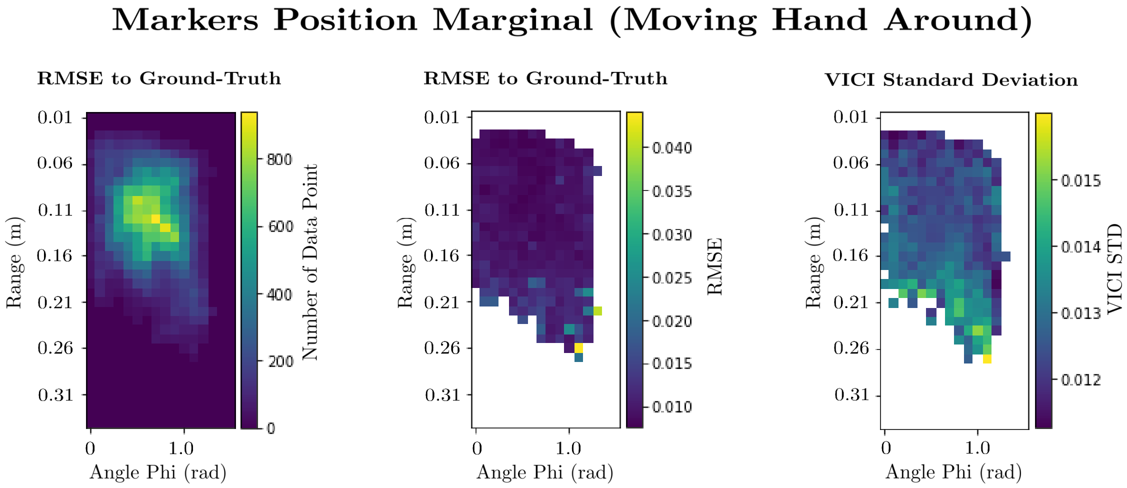


Figure 5.20: Plots of tracker positioning average accuracy as a function of range and inclination angle (Phi) from the Soli sensor. On the left, the test data density is plotted over the same axes, in order to show how much data was available to compute this measure in each quantised location. In the middle, the empirical tracking precision is shown, measured by the RMSE to the ground-truth. On the right, the average standard deviation is shown, computed using draws from the trained model only, hence without the need to use ground-truth hand motions.

5.2.6 Discussion

With the application of the novel framework for semi-supervised inverse problems presented in chapter 3, new capabilities for the Soli sensing system were demonstrated. Firstly, in section 5.2.4, the multi-fidelity forward model was demonstrated to provide accurate simulations that leverage physical models in combination to empirical data. With either of these improving, the forward model, and the whole system as a result, are expected to improve even further and extend to the general case of any hand motion. This would allow for the use of simulated environments for the Soli sensor, which can be used to investigate interaction designs and test performance entirely in software, other than augmenting training data for inversion.

Second, the results of section 5.2.5 demonstrated that including models and unsupervised data with the framework of chapter 3 in the training of a reconstruction model gives appreciable improvement compared to only training with the available labelled data. In the experiments presented here, using a few thousand gesture examples paired to Soli signal, using additional unlabelled data and a forward model allowed to train a rather accurate system, which otherwise tends to over-fit and give inadequate reconstructions, as shown in figures 5.15 and 5.16.

Third, section 5.2.5 demonstrated that the our method is capable of recovering fine motions with good accuracy, including tracking the movements of individual fingers, as shown in figures 5.17 and 5.18. These results also show how the novel framework recovers good measures of uncertainty, with which one can extract different types of measure and informative visualisations, such as those shown in section 5.2.5.

Chapter 6

Variational Inference for Unsupervised Inverse Problems

The framework presented in chapter 3 is built to solve inverse problems in which examples of targets are available, giving examples of what reconstructions should look like in some sense. However, there is many settings, particularly within information retrieval, where one desires to retrieve targets solely from large data sets of partial or corrupted observations. These type of problems can be considered as unsupervised inverse problems and commonly occur in the early stages of a learning pipeline.

In fact, data sets are rarely clean and ready to use when first collected. More often than not, they need to undergo some form of pre-processing before analysis, involving expert human supervision and manual adjustments [132, 133, 134]. Filling missing entries, correcting noisy samples, filtering collection artefacts and other similar tasks are some of the most costly and time consuming stages in the data modeling process and pose an enormous obstacle to machine learning at scale [135].

Traditional data cleaning methods rely on some degree of supervision in the form of a clean dataset or some knowledge collected from domain experts. However, the exponential increase of the data collection and storage rates in recent years, makes any supervised algorithm impractical in the context of modern applications that consume millions or billions of datapoints. This chapter introduces a novel variational framework to perform automated data cleaning and recovery without any example of clean data or prior signal assumptions.

The Tomographic auto-encoder (TAE), is named in analogy with standard tomography. Classical tomographic techniques for signal recovery aim at reconstructing a target signal, such as a 3D image, by algorithmically combining different incomplete measurements, such as 2D images from different view points, subsets of image pixels or other projections [136]. The TAE extends this concept to the reconstruction of data manifolds; the target signal is a

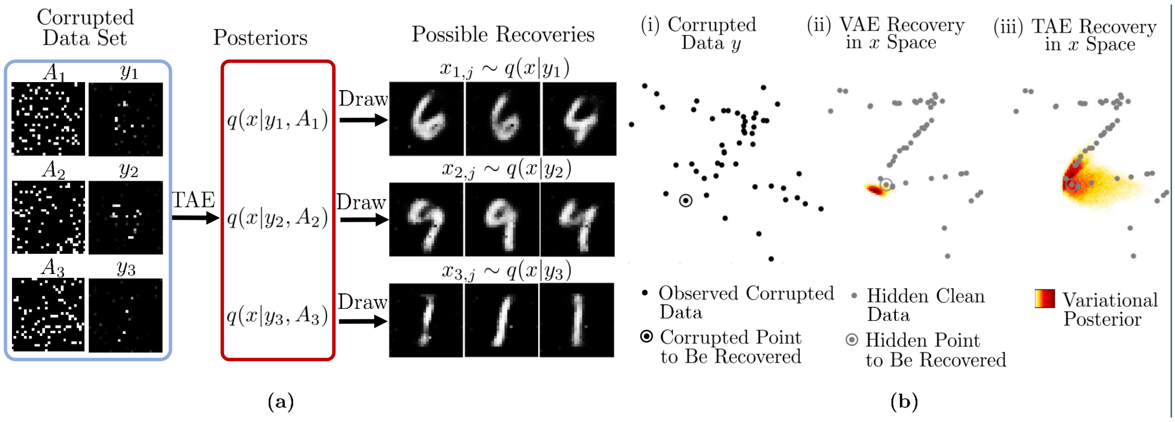


Figure 6.1: **(a)** Example of Bayesian recovery from corrupted data with a Tomographic Auto-Encoder (TAE) on corrupted MNIST. The TAE recovers posterior probability densities $q(x|y_i, A_i)$ for each corrupted sample y_i and known measurement A_i. We can draw from these to explore different possible clean solutions. **(b)** Two dimensional Bayesian recovery experiment. (i) Observed set of corrupted data Y, with the point we are inferring from y_i highlighted. (ii) Ground truth hidden clean data with the target point x_i highlighted, along with the posterior $q(x|y_i)$ (in this experiment $A_i = \mathbf{1}$ in all observations) reconstructed by a VAE. (iii) Posterior $q(x|y_i)$ recovered with the TAE.

clean data set, where corrupted data is interpreted as incomplete measurements. The aim is to combine these to reconstruct the clean data.

More specifically, as for the method presented in chapter 3 for semi-supervised situations, the task of interest is performing Bayesian recovery, where degraded samples are not simply transformed into clean ones, but probabilistic functions are recovered, with which diverse clean signals can be generated, exploring the whole range of possible solutions to the inverse problem and capturing uncertainty. As for other inverse problems, uncertainty is considerably important when cleaning data. If one is over-confident about specific solutions, errors are easily ignored and passed on to downstream tasks. For instance, in the example of figure 6.1(a), some corrupted observations y_i and corresponding known measurements A_i are consistent with multiple digits. If a single possibility was to be imputed for each sample, the true underlying solution may be ignored early on in the modeling pipeline and the digit will be consistently mis-classified. If instead accurate probability densities are recovered, one can remain adequately uncertain in any subsequent processing task.

Several variational auto-encoder (VAE) models have been proposed for applications that can be considered special cases of this problem [79, 84, 90] and, in principle, they are capable of performing Bayesian reconstruction, as they are latent variable models which are able to sample from their implicit PDFs. However, in this chapter, it is shown that surrogating variational inference (VI) in a latent space with VAEs results in collapsed distributions that do not explore the different possibilities of clean samples, but only return single estimates. This pathology precludes the capability which is the focus of the work in this thesis; the ability to perform Bayesian reconstruction and retrieve all different possible solutions to an inverse problem. To overcome this issue, the TAE performs approximate VI in the space of recovered data instead, through the novel *reduced entropy condition* method, proposed and detailed in section 6.1.3. The resulting posteriors adequately explore the manifold of possible clean samples for each corrupted observation and, therefore, adequately capture the uncertainty of the task.

Figure 6.1(a) shows inference examples using the TAE framework on MNIST with missing values, where only 10% of pixels are observed with additive Gaussian noise. In this example, the coordinates of the observed pixels are known and incrporated into the TAE framework through observation masks A_i. The TAE algorithm is also compared to a standard VAE approach in a simple 2D experiment shown in figure 6.1(b). In this experiment, 50 data points lie along a randomly generated sinusoidal curve, shown in grey. Gaussian noise is added to the data points to generate the observations, meaning that, in this case, the observation masks A_i contain all ones. Both a standard VAE and the TAE are then used to recover a probability density of an underlying clean data point x_i using the corresponding corrupted observation y_i as input. Both models have identical architectures, with 2-dimensional latent spaces and fully connected networks for their components. While the VAE posterior collapses to a single point, the TAE reconstructs a rich posterior that adjusts to the data manifold.

The experiments the TAE framework is tested with presented in sections 6.2 and 6.2.4 focus on data recovery from noisy samples and missing entries. This is one of the most common data corruption settings being encountered in a wide range of domains with different types of data [137, 138]. By testing the novel approach in this prevalent scenario, it can be closely compared with recently proposed VAE approaches [84, 139, 85]. The experiments of section 6.2 show how the existing VAE models exhibit the posterior collapse problem while the TAE produces rich posteriors that capture the underlying uncertainty. The TAE is then tested on classification subsequent to imputation, demonstrating superior performance to existing methods in these downstream tasks. In section 6.2.4, the TAE is used to perform automated missing values imputation on raw depth maps from the NYU rooms data set [140].

6.1 Method

In order to frame the problem and understand the issues with standard variational methods in this context, the data recovery task is viewed from a signal reconstruction prospective. The final scope of a Bayesian data recovery method is to build and train a parametric probability density function (PDF) $q(x|y, A)$, which takes as inputs corrupted measurements y and known observation parameters A, where available, and generates different possible corresponding clean data $x \sim q(x|y, A)$ through sampling. For the rest of the chapter, the known observation parameters A are left out of the modelling for simplicity and all poste-

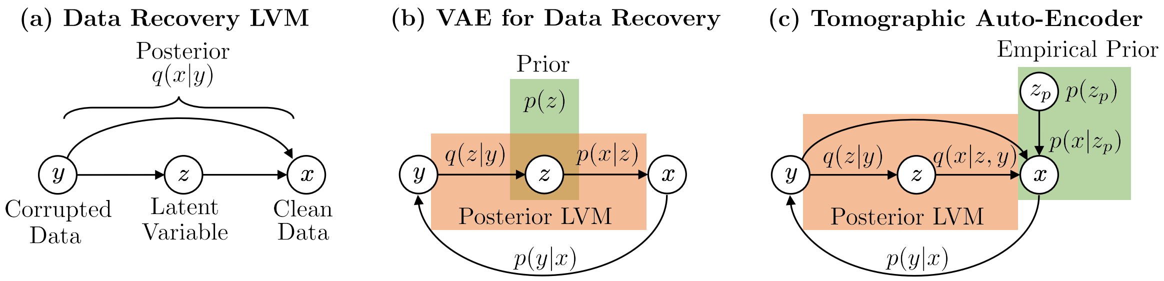


Figure 6.2: Training LVMs for data recovery. **(a)** Structure of the reconstruction LVM used to infer approximate posteriors $q(x|y)$ of clean data x from corrupted observations y as conditional inputs. **(b)** Training of $q(x|y)$ using a VAE. A prior in the latent space z is introduced as a regulariser, however no explicit regularisation is imposed in x. **(c)** Training of $q(x|y)$ using our TAE model. An empirical prior $p(x) = \int p(z_p)p(x|z_p)dz_p$ is instead introduced in clean data space x.

riors $q(x|y, A)$ will be written as $q(x|y)$. In the construction of this posterior, there are two aspects that need to be designed: i) the structure of this conditional PDF and ii) the way it will be trained to perform the recovery task.

Regarding the former, as natural data often lies on highly non-linear manifolds, the conditional PDF needs to capture complicated modalities, e.g. the distribution of plausible images consistent with one of the corrupted observations in figure 6.1(a). A suitable recovery PDF $q(x|y)$ needs to be able to capture such complexity. As in the semi-supervised case, a natural choice to achieve high capacity and tractability is to construct $q(x|y)$ as a conditional latent variable model (LVM). As described in section 2.2.2 and also demonstrated by the novel method presented in chapter 3, conditional LVM neural networks have achieved efficient and expressive variational inference in many recovery settings, capturing complex solution spaces in high dimensional problems, such as image reconstruction [66, 63, 8]. The conditional LVM consists of a first conditional distribution $q(z|y)$ mapping input corrupted data y to latent variables z, and a second inference $q(x|z, y)$ mapping latent variables to output clean data x. The resulting PDF is $q(x|y) = \int q(z|y)q(x|z, y)dz$, where both $q(z|y)$ and $q(x|z, y)$ are simple distributions, such as isotropic Gaussians, whose moments are inferred by neural networks taking the respective conditional arguments as inputs. Figure 6.2(a) shows a graphical model for the conditional LVM.

While the choice of structure is fairly straightforward, the main difficulty lies in training the recovery LVM in the absence of clean ground truths x. In the supervised and semi-supervised cases, several established methods exist, including the novel one presented in chapter 3. With even partial supervision, the observed distributions of clean data x conditioned on paired observations y can be matched by parametric ones through a VAE or GAN training strategy [65, 8, 11]. The focus of this chapter is instead the unsupervised situation, where only corrupted data $Y = \{y_{1:N}\}$ is accessible, along with a functional form for the corrupted data likelihood $p(y|x)$, e.g., missing values and additive noise. Training a conditional LVM to fit posteriors without any ground truth examples x is rather challenging, as there is no target data to encode from, in the case of VAE architectures, or adversarially compare with, in the case of GAN models.

6.1.1 VAEs and the Posterior Degeneracy Problem

Variational auto-encoders (VAEs) have been proposed for several problems within this definition of unsupervised reconstruction [139, 79, 90, 87]. These methods lead to good single estimates of the underlying targets. However, they easily over-fit their posteriors resulting in collapsed PDFs $q(x|y)$. Put differently, they are often unable to explore different possible solutions to the recovery problem and return single estimates instead. Figure 6.1(b-ii) shows this pathology in a two dimensional experiment.

The reason for this can be explained considering what the reconstruction LVM $q(x|y)$ is and how it is trained when directly employing a VAE in the unsupervised recovery scenario. The VAE encodes latent vectors z from corrupted observations y with an encoder $q(z|y)$ and reconstructs clean data x with a decoder $p(x|z)$. These two functions constitute the reconstruction LVM $q(x|y) = \int q(z|y)p(x|z)dz$. As there are no clean ground truths x, data likelihood is maximised by mapping reconstructed clean samples x back to corrupted samples y with a corruption process likelihood $p(y|x)$, e.g. zeroing out missing entries, to maximise reconstruction of the observations y. Concurrently, regularisation in the latent space is induced with a user defined prior $p(z)$ (e.g. a unit Gaussian). The resulting lower bound to be maximised during training can be expressed as follows:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q(z|y)} \log p(y|z) - KL(q(z|y)||p(z)), \tag{6.1}$$

where the observations likelihood is $p(y|z) = \int p(x|z)p(y|x)dx$ and in some cases, such as for missing values and additive noise, it is analytical. This bound for missing value imputation using a VAE can be derived as follows. The aim is to maximise the log likelihood of the observed corrupted data y:

$$\log p(y) = \log \int_x \underbrace{\int_z p(z)p(x|z)dz}_{p(x)}\, p(y|x)dx. \tag{6.2}$$

To obtain a tractable approximation to this likelihood, one can introduce a variational distribution in both clean data space and latent space $q(x,z|y)$ and define a lower bound as

$$\begin{aligned}\log p(y) \geq & \int_x \int_z q(x,z|y) \log \frac{p(z)p(x|z)dz}{q(x,z|y)} dzdx \\ & + \int_x \int_z q(x,z|y) \log p(y|x) dzdx.\end{aligned} \tag{6.3}$$

To obtain the VAE ELBO used in data recovery settings, the choice of this variational posterior is $q(x,z|y) = q(z|y)p(x|z)$. The ELBO can then be simplified to give

$$\begin{aligned}\log p(y) \geq & \int_x \int_z q(z|y)p(x|z) \log \frac{p(z)p(x|z)dz}{q(z|y)p(x|z)} dzdx \\ & + \int_x \int_z q(z|y)p(x|z) \log p(y|x) dzdx \\ = & \underbrace{\int_x p(x|z)dx}_{=1} \int_z q(z|y) \log \frac{p(z)dz}{q(z|y)} dz \\ & + \int_x \int_z q(z|y)p(x|z)dz \log p(y|x) dx.\end{aligned} \tag{6.4}$$

For situations in which the observations' likelihood $p(y|z) = \int_x p(x|z)p(y|x)dx$ has a closed form, such as additive noise and missing entries, a tighter bound to the likelihood can be defined by moving the integral in x in the second term inside the logarithm:

$$\begin{aligned}\log p(y) \geq & \int_z q(z|y) \log \frac{p(z)dz}{q(z|y)} dz \\ & + \int_z q(z|y) \log \left[\int_x p(y|x)p(x|z)dx \right] dz \\ = & - KL(q(z|y)||p(z)) + \mathbb{E}_{q(z|y)} \log p(y|z) dx.\end{aligned} \tag{6.5}$$

Note that, because $p(x|z)$ simplifies in the KL term, this ELBO avoids variational inference in the space of clean data x.

Viewing the VAE training from a signal reconstruction prospective, where the reconstruction model is $q(x|y) = \int q(z|y)p(x|z)dz$, one can notice that no prior is directly introduced on the hidden targets x, but only in the LVM latent space z. While regularising only in z may be computationally desirable, if the decoder $p(x|z)$ is of sufficient capacity, the model can learn to collapse regularised distributions in z to single estimates in x, failing to capture the uncertainty in the solution space of the inverse problem. In fact, this is induced by the objective function of equation 6.1; the model finds broad distributions in the latent space $q(z|y)$, which minimise the KL divergence with $p(z)$, but the generator $p(x|z)$ can learn to collapse them back to single maximum likelihood solutions in x, maximising $\mathbb{E}_{q(z|y)} \log \int p(x|z)p(y|x)dx$. This effect may be counteracted by reducing the capacity of $p(x|z)$ or the dimensionality of

z, but doing so also reduces the capacity of the reconstruction model $q(x|y)$, resulting in an undesirable coupling between regularisation and posterior capacity.

6.1.2 Data Modelling View of the Problem

An alternative way to view the task and understand the posterior collapse problem is from a data modelling prospective, which is perhaps more familiar to readers acquainted with generative models and VAEs. In this view, the *a priori* goal is to build a model $p(y)$ to capture the observable data points y_i. The model is constructed as a latent variable model $p(y) = \int_x \int_z p(z)p(x|z)dzp(y|x)dx$ where $p(y|x)$ is completely or partially fixed by the observation model. To train this generative model using a VI approach, a recognition model $q(x, z|y)$ is introduced to train an ELBO on the log likelihood of the observations y_i. The standard VAE approach introduces a factorisation assumption in the construction of this recognition model which is improper and results in the posterior collapse problem. Namely, as described above through the data recovery prospective, the VAE assumes that $q(x, z|y) = q(z|y)p(x|z)$. Instead, given the generative structure described above, the formally correct factorisation is $q(x, z|y) = q(x|y)q(z|x)$. The TAE framework presented here applies this factorisation and models $q(x|y)$ as a conditional LVM itself. To reconcile this way of viewing the problem and the resulting starting point to model the TAE with the data recovery ones presented above, one simply needs to rename z above to z_p and formalise the recognition model's LVM component as $q(x|y) = \int q(z|y)q(x|z)dz$, where now z is its own latent variable. With this renaming of variables, the TAE development presented below follows from this second perspective as well.

6.1.3 Separating Posterior and Prior: The Tomographic Auto-Encoder

The premise of the model presented in this chapter to address the aforementioned problem is simple: Introduce a prior $p(x)$ in the hidden clean signal space. In particular, the proposed method uses an empirical prior, having itself the form of an LVM $p(x) = \int p(z_p)p(x|z_p)dz_p$. An empirical prior, i.e. a prior which itself is trained with data, is chosen in order to better capture the nature of clean data x, as this is a distribution over natural data. $p(z_p)$ is a unit Gaussian and $p(x|z_p)$ is a neural network, the parameters of which are trained along those of the posterior. In this way, approximate variational inference is performed in clean data space x, instead of being surrogated to the reconstruction function's latent space z. By doing so, the capacity of the prior $p(x)$ can be controlled to induce regularisation independently of the capacity of the reconstruction model $q(x|y) = \int q(z|y)q(x|z, y)dz$. In fact, the capacity of this prior model is kept limited to avoid over-fitting, i.e. $p(x|z_p)$ has many fewer parameters than $q(x|z)$, ensuring that the prior does not over-fit. For this framework, An ELBO can be formulated as follows. The likelihood to be maximise is

$$\log p(y) = \log \int_x \underbrace{\int_{z_p} p(z_p)p(x|z_p)dz_p}_{p(x)} p(y|x)dx. \tag{6.6}$$

Similarly to the VAE ELBO case, a variational posterior $q(x, z_p|y)$ is used to find a lower bound

$$\begin{aligned}\log p(y) \geq & \int_x \int_{z_p} q(x, z_p|y) \log \frac{p(z_p)p(x|z_p)dz_p}{q(x, z_p|y)} dz_p dx \\ & + \int_x \int_{z_p} q(x, z_p|y) \log p(y|x)dz_p dx.\end{aligned} \tag{6.7}$$

However, the TAE model does not make the assumption that the variational posterior has the special form described in section 6.1.1 and instead set it to have the form $q(x, z_p|y) = q(x|y)q(z_p|x)$, separating posterior inference from observations y to clean data x and inference of prior latent variables z_p. The resulting lower bound is

$$\begin{aligned}\log p(y) \geq & \int_x \int_{z_p} q(x|y)q(z_p|x) \log \frac{p(z_p)p(x|z_p)}{q(x|y)q(z_p|x)} dz_p dx + \int_x \int_{z_p} q(x|y)q(z_p|x) \log p(y|x)dz_p dx \\ = & \int_x q(x|y) \underbrace{\int_{z_p} q(z_p|x) \log \frac{p(z_p)p(x|z_p)}{q(z_p|x)} dz_p}_{\geq \log p(x)} dx + \int_x \underbrace{\int_{z_p} q(z_p|x)}_{=1} dz_p q(x|y) \log p(y|x)dx \\ & - \int_x \underbrace{\int_{z_p} q(z_p|x)dz_p}_{=1} q(x|y) \log q(x|y)dx \\ = & \mathbb{E}_{q(x|y)} \left[\mathbb{E}_{q(z_p|x)} \log p(x|z_p) - KL(q(z_p|x)||p(z_p)) \right] + \mathbb{E}_{q(x|y)} \log p(y|x) + H(q(x|y)).\end{aligned} \tag{6.8}$$

The main technical challenge and focus of this chapter is how to compute and maximise the self entropy of the approximate posterior $H(q(x|y))$, as this conditional distribution is an LVM of the form $q(x|y) = \int q(z|y)q(x|z, y)dz$.

Reduced Entropy Condition

Direct computation of the entropy of an LVM model $q(x|y) = \int_z q(z|y)q(x|z, y)dz$ is intractable in the general case. [141] proposed an approximate inference method to compute the gradient of the LVM's entropy for variational inference in latent spaces. However, this involves multiple samples to be drawn and evaluated with the LVM, which is expected to scale in complexity as the dimensionality and capacity of the target distribution increase.

In the TAE model, the aim is to approximately compute and optimise the entropy $H(q(x|y))$ for a distribution capturing natural data, which can be high-dimensional and lie on complicated manifolds. In order to maintain efficiency in the entropy estimation, the TAE introduces a new strategy; a class of LVM posteriors for which the entropy reduces to a tractable form is identified and the posterior is approximately constrained to such a class in the ELBO optimisation. The main result is summarized in the following theorem:

Theorem 1 *If $\frac{q(z|x,y)}{q(z|y)} = B\delta(z - g(x,y))$, where $\delta(\,\cdot\,)$ is the Dirac Delta function, B is a real positive parameter and $g(x,y)$ is a deterministic function, then $H(q(x|y)) = H(q(z|y)) + \mathbb{E}_{q(z|y)}H(q(x|z,y))$.*

We note that, in principle, given the clean data x, the The proof is as follows.

$$\begin{aligned}
&\frac{q(z|x,y)}{q(z|y)} = B\delta(z - g(x,y)) \implies \frac{q(z|x,y)}{q(z|y)}\frac{q(z'|x,y)}{q(z'|y)} = 0, \quad \forall x, z \neq z' \\
&\implies \frac{q(x|z,y)}{q(x|y)}\frac{q(x|z',y)}{q(x|y)} = 0, \quad \forall x, z \neq z' \\
&\implies q(x|z,y)q(x|z',y) = 0, \quad \forall x, z \neq z' \\
&\implies q(x|z',y) = 0, \quad \forall x \sim q(x|z,y), z \neq z'
\end{aligned} \tag{6.9}$$

Using the result of equation 6.9, the form of the entropy $H(q(x|y))$ for this special case can be derived as the following:

$$\begin{aligned}
H(q(x|y)) &= -\int_x \left[\int_z q(z|y)q(x|z,y)dz\right] \cdot \log\left[\int_{z'} q(z'|y)q(x|z',y)dz'\right]dx \\
&= -\int_x\int_z q(z|y)q(x|z,y) \cdot \log\left[\int_{z'=z} q(z'|y)q(x|z',y)dz'\right. \\
&\quad \left. + \underbrace{\int_{z'\neq z} q(z'|y)q(x|z',y)dz'}_{eq.6.9 \implies =0}\right]dzdx \\
&\quad = -\int_z\int_x q(z|y)q(x|z,y)\log\left[q(z|y)q(x|z,y)\right]dxdz \\
&\quad = -\int_z\int_x q(z|y)q(x|z,y)\log q(z|y)dxdz - \int_z\int_x q(z|y)q(x|z,y)\log q(x|z,y)dxdz \\
&\quad = -\int_z q(z|y)\log q(z|y)dz - \int_z q(z|y)\int_x q(x|z,y)\log q(x|z,y)dxdz \\
&\quad = H(q(z|y)) + \mathbb{E}_{q(z|y)}H(q(x|z,y)),
\end{aligned} \tag{6.10}$$

obtaining the reduced tractable form of the entropy stated in theorem 1. Theorem 1 states that if the posterior over latent variables $q(z|x,y)$ is infinitely more localised than the latent conditional $q(z|y)$, then the LVM entropy $H(q(x|y))$ has the tractable form given above. This condition imposes the LVM posterior to present non-overlapping conditionals $q(x|z,y)$ for different latent variables z, but does not impose any explicit restriction to the capacity of the model. The reduced entropy condition can be re-formulated as follows.

$$\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} = C, \quad C \to \infty. \tag{6.11}$$

This equivalence can be proven by proving that one is both sufficient and necessary condition for the other:

proof of necessary condition:

$$\begin{aligned}
\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} &= \int_z q(z|y) \int_x q(x|z,y) \log \frac{q(z|x,y)}{q(z|y)} dxdz \\
&= \int_x q(x|y) \int_z q(z|x,y) \log \frac{q(z|x,y)}{q(z|y)} dzdx \\
&= \int_x q(x|y) \int_z q(z|x,y) \log q(z|x,y) dzdx \\
&- \int_x q(x|y) \int_z q(z|x,y) \log q(z|y) dzdx \\
&= \int_x q(x|y) \int_z q(z|x,y) \log q(z|x,y) dzdx \\
&- \underbrace{\int_x q(x|z,y)dx}_{=1} \int_z q(z|y) \log q(z|y) dz \\
&= \mathbb{E}_{q(x|y)} \underbrace{\int_z q(z|x,y) \log q(z|x,y) dz}_{-H(q(z|x,y))} \\
&- \underbrace{\int_z q(z|y) \log q(z|y) dz}_{-H(q(z|y))}.
\end{aligned} \tag{6.12}$$

If the above expression tends to infinity, either $H(q(z|x,y)) \to -\infty$ or $H(q(z|y)) \to \infty$, meaning that either $q(z|x,y) \to$ a Delta function, or $q(z|y) \to$ uniform. Either condition implies $\frac{q(z|x,y)}{q(z|y)} = B\delta(z - g(x,y))$.

proof of sufficient condition:

$$\begin{aligned}
\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} &= \int_x q(x|y) \int_z q(z|x,y) \log \frac{q(z|x,y)}{q(z|y)} dzdx \\
&= \int_x q(x|y) \int_z q(z|y) \frac{q(z|x,y)}{q(z|y)} \log \frac{q(z|x,y)}{q(z|y)} dzdx.
\end{aligned} \tag{6.13}$$

Now we set $\frac{q(z|x,y)}{q(z|y)} = B\delta(z - g(x,y))$:

$$\begin{aligned}&\int_x q(x|y) \int_z q(z|y) B\delta(z - g(x,y)) \log B\delta(z - g(x,y)) dz dx \\ = &\int_x q(x|y) q(g(x,y)|y) \log B \underbrace{\delta(g(x,y) - g(x,y))}_{\to\infty, \forall x} dx.\end{aligned} \tag{6.14}$$

Therefore, $\frac{q(z|x,y)}{q(z|y)} = B\delta(z - g(x,y))$ is a sufficient condition for $\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} \to \infty$.

To train the posterior $q(x|y)$, the ELBO $\mathcal{L}_{TAE}$ is maximised with the reduced entropy, while enforcing the condition of equation 6.11:

$$\begin{aligned}&\arg\max \quad \mathbb{E}_{q(x|y)} \log p(y|x) + \mathbb{E}_{q(x|y)} \left[\mathbb{E}_{q(z_p|x)} \log p(x|z_p) - KL(q(z_p|x)||p(z_p)) \right] \\ &+ H(q(z|y)) + \mathbb{E}_{q(z|y)} H(q(x|z,y)), \quad s.t. \quad \mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} = C, \quad C \to \infty.\end{aligned} \tag{6.15}$$

While the ELBO is now amenable to stochastic optimization, the constraint is intractable since $C \to \infty$ and the posterior $q(z|x,y)$ is intractable.

Relaxed Constraint

To render the constraint tractable, C is first relaxed to be a positive hyper-parameter. The higher the value of C, the more localised $q(z|x,y)$ is imposed to be compared to $q(z|y)$ and the closest the reduced entropy is to the true one.

To address the intractability of the posterior $q(z|x,y)$, a variational approximation with a parametric function $r(z|x,y)$ is used. In fact, for any valid probability density $r(z|x,y)$, it can be proven that

$$\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} \geq \mathbb{E}_{q(x,z|y)} \log \frac{r(z|x,y)}{q(z|y)}. \tag{6.16}$$

The proof is as follows.

$$
\begin{aligned}
\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} = & \int_z \int_x q(x,z|y) \log q(z|x,y) dz dx \\
& - \int_z \int_x q(x,z|y) \log q(z|y) dz dx \\
& = \int_x q(x|y) \int_z q(z|x,y) \log q(z|x,y) dz dx \\
& - \int_z \int_x q(x,z|y) \log q(z|y) dz dx \\
& \geq \int_x q(x|y) \int_z q(z|x,y) \log r(z|x,y) dz dx \\
& - \int_z \int_x q(x,z|y) \log q(z|y) dz dx \\
& = \mathbb{E}_{q(x,z|y)} \log \frac{r(z|x,y)}{q(z|y)},
\end{aligned}
\tag{6.17}
$$

Where the inequality derives from the positivity of the KL divergence $KL(q(z|x,y)||r(z|x,y))$. The above bound implicates that

$$
\mathbb{E}_{q(x,z|y)} \log \frac{r(z|x,y)}{q(z|y)} = C \Rightarrow \mathbb{E}_{q(x,z|y)} \log \frac{q(z|x,y)}{q(z|y)} \geq C.
$$

This means that imposing the condition with a parametric distribution $r(z|x,y)$, which is trained along with the rest of the model, ensures deviation from the set condition only by excess. As the exact condition is met only at $\mathbb{E}_{q(x,z|y)} \log \frac{q(z|x)}{q(z|y)} \rightarrow \infty$, the constraint can never be relaxed more than already set by the finite value of C.

The TAE Objective Function

Having defined a tractable ELBO and a tractable condition, one needs to perform the constrained optimisation

$$
\begin{aligned}
\arg\max \quad & \mathbb{E}_{q(x|y)} \log p(y|x) + \mathbb{E}_{q(x|y)} \left[\mathbb{E}_{q(z_p|x)} \log p(x|z_p) - KL(q(z_p|x)||p(z_p)) \right] \\
& + H(q(z|y)) + \mathbb{E}_{q(z|y)} H(q(x|z,y)), \quad s.t. \quad \mathbb{E}_{q(x,z|y)} \log \frac{r(z|x,y)}{q(z|y)} = C.
\end{aligned}
\tag{6.18}
$$

To do this, the commonly adopted penalty function method [142, 143] is adopted and equation 6.18 is relaxed to an unconstrained optimisation with the use of a positive hyper-parameter

λ:

$$\begin{aligned}\arg\max \quad & \mathbb{E}_{q(x|y)} \log p(y|x) + \mathbb{E}_{q(x|y)} \left[\mathbb{E}_{q(z_p|x)} \log p(x|z_p) - KL(q(z_p|x)||p(z_p)) \right] \\ & + H(q(z|y)) + \mathbb{E}_{q(z|y)} H(q(x|z,y)) - \lambda \left| \mathbb{E}_{q(z,x|y)} \log \frac{r(z|x,y)}{q(z|y)} - C \right| . \end{aligned} \tag{6.19}$$

To train the model, equation 6.19 is maximised using the ADAM optimiser. Once the model is trained, diverse reconstructions from a corrupt observation y_i can be generated by sampling from the posterior $q(x|y_i)$.

The objective of equation 6.19 is simplified to aid understanding. In the general case, the corruption process $p(y|x)$, mapping clean data x to degraded samples y, is controlled by parameters that differ from sample to sample. We can distinguish these into observed parameters α and unobserved parameters β. For example, in the case of missing values and noise, the indexes of missing entries in each sample are often observed parameters, while the noise level is an unobserved parameter. The complete form of the corruption likelihood for a clean sample x_i is then $p(y|x_i, \alpha_i, \beta_i)$.

Explicitly showing the parameters to be optimised, the objective function maximised to train the TAE is the following

$$\begin{aligned}\arg\max_{\theta,\phi} \quad & \mathbb{E}_{q_\phi(x,\beta|y,\alpha)} \log p(y|x,\alpha,\beta) \\ & + \gamma \mathbb{E}_{q_\phi(x|y,\alpha)} \left[\mathbb{E}_{q_{\phi_3}(z_p|x)} \log p_\theta(x|z_p) - KL(q_{\phi_3}(z_p|x)||p(z_p)) \right] \\ & + H(q_{\phi_1}(z|y,\alpha)) + \gamma \mathbb{E}_{q_{\phi_1}(z|y,\alpha)} H(q_{\phi_2}(x|z,y,\alpha)) \\ & - \lambda \left| \mathbb{E}_{q_\phi(z,x|y,\alpha)} \log \frac{r_{\phi_4}(z|x)}{q_{\phi_1}(z|y,\alpha)} - C \right| , \end{aligned} \tag{6.20}$$

where $q_\phi(x,\beta|y,\alpha) = \int_z q_{\phi_1}(z|y,\alpha) q_{\phi_2}(x|z,y,\alpha) q_{\phi_5}(\beta|z,y,\alpha) dz$, $q_\phi(x|y,\alpha)$ $= \int_z q_{\phi_1}(z|y,\alpha) q_{\phi_2}(x|z,y,\alpha) dz$, $q_\phi(z,x|y,\alpha) = q_{\phi_1}(z|y,\alpha) q_{\phi_2}(x|z,y,\alpha)$, $\phi = \{\phi_{1:5}\}$ are the parameters of the inference models and θ are the parameter of the prior model.

Training

All expectations in the above expression are computed and optimised by sampling the corresponding conditional distributions using the re-parametrisation trick characteristic of VAEs.

Because the prior LVM $p(x) = \int p(z_p) p(x|z_p) dz_p$ is training entirely with samples from the posterior LVM, which is also training, the model can easily obtain high values for the prior ELBO by generating collapsed samples x with the posterior and get stuck in an unfavourable local minimum. To avoid this, a warm up strategy is employed. a positive parameter γ is defined, which multiplies the expectation of the prior ELBO and the entropy $H(x|z,y)$:

$$\begin{aligned}\arg\max \quad & \mathbb{E}_{q(x|y)} \log p(y|x) + \gamma \mathbb{E}_{q(x|y)} \big[\underbrace{\mathbb{E}_{q(z_p|x)} \log p(x|z_p) - KL(q(z_p|x)||p(z_p))}_{Prior \quad ELBO, \quad \geq p(x)} \big] \\ & + H(q(z|y)) + \gamma \mathbb{E}_{q(z|y)} H(q(x|z,y)) - \lambda \left| \mathbb{E}_{q(z,x|y)} \log \frac{r(z|x,y)}{q(z|y)} - C \right| .\end{aligned} \tag{6.21}$$

The value of γ is initially set to zero. After a set number of iterations it is linearly increased to reach one and kept constant for the remaining training iterations. A pseudo-code for the TAE training procedure is given in algorithm 3.

Algorithm 3 Training the TAE Model

Inputs: Corrupted observations $Y = \{y_{1:N}\}$; Observed Parameters $A = \{\alpha_{1:N}\}$ initial model parameters, $\{\theta^{(0)}, \phi^{(0)}\}$; user-defined posterior latent dimensionality, J; user-defined prior latent dimensionality, J_p; user-defined condition strength λ; user-defined condition parameter C; user-defined latent prior $p(z_p)$; user-defined initial warm-up coefficient γ_0; user-defined final warm-up coefficient γ_f; warm-up start N_{w0}; warm-up end N_{wf}; user-defined number of iterations, N_{iter}.

```
$\gamma^{(k=0)} \leftarrow \gamma_0$
for the $k$'th iteration in $[0 : N_{iter} - 1]$
    for the $i$'th observation
        $z_i \sim q_{\phi_1^{(k)}}(z|y_i, \alpha_i)$
        $x_i \sim q_{\phi_2^{(k)}}(x|z_i, y_i, \alpha_i)$
        $\beta_i \sim q_{\phi_5^{(k)}}(\beta|z_i, y_i, \alpha_i)$
        $z_{p,i} \sim q_{\phi_3^{(k)}}(z_p|x_i)$
        $\mathbf{E}_i^{(k)} \leftarrow \log p(y_i|x_i, \beta_i)$
        $\mathbf{P}_i^{(k)} \leftarrow \log p_{\theta^{(k)}}(x_i|z_{p,i})$
        $\mathbf{K}_i^{(k)} \leftarrow D_{KL}(q_{\phi_3^{(k)}}(z_p|x_i)||p(z_p))$
        $\mathbf{Hz}_i^{(k)} \leftarrow H(q_{\phi_1^{(k)}}(z|y_i, \alpha_i))$
        $\mathbf{Hx}_i^{(k)} \leftarrow H(q_{\phi_2^{(k)}}(x|z_i, y_i, \alpha_i))$
        $\mathbf{R}_i^{(k)} \leftarrow \log r_{\phi_4^{(k)}}(z_i|x_i, y_i, \alpha_i)$
        $\mathbf{Q}_i^{(k)} \leftarrow \log q_{\phi_1^{(k)}}(z_i|y_i, \alpha_i)$
    end
    $\mathbf{F}^{(k)} = \sum_i \Big( \mathbf{E}_i^{(k)} + \gamma^{(k)} \Big[ \mathbf{P}_i^{(k)} - \mathbf{K}_i^{(k)} + \mathbf{Hx}_i^{(k)} \Big]$
        $+ \mathbf{Hz}_i^{(k)} - \lambda \Big| \mathbf{R}_i^{(k)} - \mathbf{Q}_i^{(k)} - C \Big| \Big)$
    $\theta^{(k+1)}, \phi^{(k+1)} \leftarrow \arg\max(\mathbf{F}^{(k)})$
    if $k > N_{w0}$ and $k < N_{wf}$
        $\gamma^{(k+1)} \leftarrow \gamma^{(k)} + (\gamma_f - \gamma_0)/(N_{wf} - N_{w0})$
    else
        $\gamma^{(k+1)} \leftarrow \gamma^{(k)}$
    end
end
=0
```

6.2 Experiments

The TAE framework is tested in a series of controlled experiments with different data sets. First, in section 6.2.1, the quality of the recovered posteriors is evaluated through the ELBO measured by an independent model, in order to obtain a quantitative measure of how well

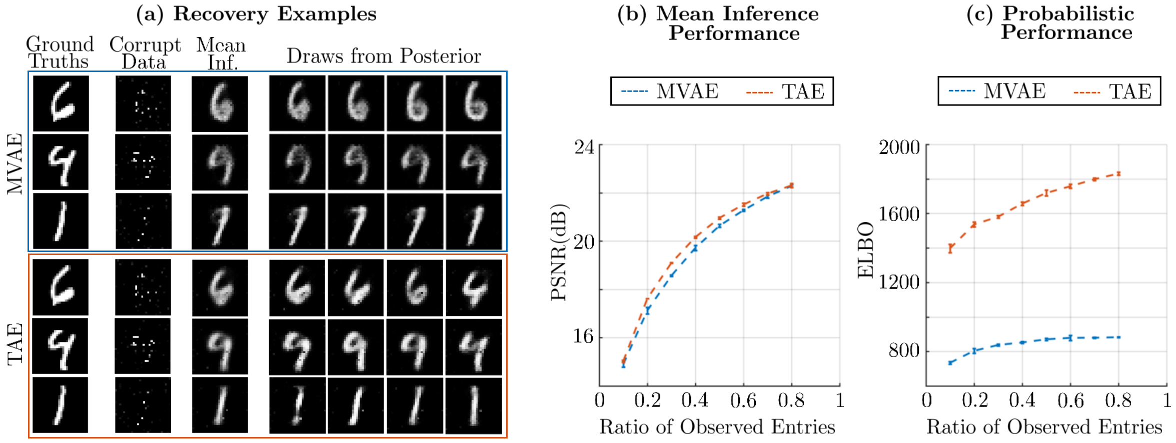


Figure 6.3: MNIST data recovery from missing entries and noise. **(a)** Recoveries using an MVAE and the TAE, showing average reconstruction and samples from the trained posteriors. **(b)** PSNR between ground truths and mean reconstruction. **(c)** ELBO assigned by the recovered posteriors to the ground truth data.

latent variable models can capture the inverse problem solution space in completely unsupervised settings. Secondly, in section 6.2.3, the recovery models are combined with a classification network in order to form a complete data pipeline and evaluate the TAE compared to standard variational approaches with respect to typical down-stream tasks of interest.

6.2.1 Posterior Recovery

Experimental Conditions

All posterior recovery experiments presented in this subsection are performed on samples that have been re-scaled from 0 to 1. In all cases, the sets are injected with additive Gaussian noise having standard deviation 0.1. Subsequently, random binary masks are generated to block out some entries, resulting in missing values. The proportion of missing entries in the masks was varied to test the TAE and competing strategies in different limits.

In the experiment of figure 6.3, The MNIST dataset [144] is corrupted by introducing missing values and additive Gaussian noise on the observed entry, as described above. Both a missing value imputation VAE (MVAE), analogous to those presented in [84] and [139], and our TAE model are trained with the corrupted data sets. The VAE and TAE are constructed such that the structure of their posteriors $q(x|y)$, i.e. the functions mapping corrupted data to distributions of clean data at test time, are exactly the same. In this way, it is ensured that differences in performance are due to the variational inference method employed and not the choice of posterior model. In particular, both $q(z|y)$ and $q(x|z)$ are fully connected networks with four layers, giving as outputs moments of Gaussian distributions. The posterior latent space z is 20-dimensional. For the TAE, the prior networks $p(x|z_p)$ and $q(z_p|y)$ are also fully connected layers, but their capacity is lower, having only two layers and a latent space z_p

of 8 dimensions. Figure 6.3(a) shows examples of mean reconstruction and posterior draws from a situation with 90% missing values, while the plots in figure 6.3(b-c) show the RMSE and ELBO respectively at varying ratio of available entries. Error bars are computed from five repetitions of each experiment, each time performing a new random sub-sampling.

In the experiments of table 6.1, the TAE ELBO is evaluated further with Fashion-MNIST – 28×28 grey-scale images of clothing [145], and the UCI HAR dataset, which consists of filtered accelerometer signals from mobile phones worn by different people during common activities [146]. As before, the recovery of these data sets from observations affected by missing values and additive noise is tested. In addition to the MVAE baseline, performance is compared against the recently proposed missing values importance weighted auto encoder (MIWAE) [85]. This model modifies the VAE ELBO by sampling multiple points from the recognition model and building the log reconstruction term with a sum of contributions, instead of a single decoding. The architectures of the models are identical to those described above from the previous experiments with MNIST. In this set of experiments, the MIWAE was trained with 20 samples of the latent space per encoding. The ELBO for each model in different missing values situations is computed and reported in table 6.1. As before, uncertainties are recovered from five repeats of the experiments.

To compute the ELBO in both sets of experiments described above, and in all further experiments presented in this chapter, noiseless and complete test ground truths are emplied. An ELBO evaluation strategy which is commonly adopted in fully unsupervised settings [147, 148, 85] is employed. After each model is trained unsupervisedly, a posterior of the form $q(x|y) = \int q(z|y)q(x|z)dz$ is obtained, where for the MVAE and MIWAE, $q(x|z) = p(x|z)$. Given a test set of paired clean and corrupted samples x_t and y_t, a new parametric recognition model is constructed, which encodes latent distributions from ground-truths $q_\eta(z|x)$. The following objective is then optimised:

$$\arg\max_{\eta} \quad \mathbb{E}_{q_\eta(z|x_t)} \log q(x_t|z) + KL(q_\eta(z|x_t)|q(z|y_t)). \tag{6.22}$$

The above is a conditional VAE ELBO with conditional prior $q(z|y)$ and is a lower bound to the test likelihood of interest $q(x_t|y_t)$. Note that the optimisation is performed over η only, therefore the new recognition model $q(z|x)$ is the only one which is affected by this optimisation and the components of our reconstruction model $q(z|y)$ and $q(x|z)$ remain the same as trained with the unsupervised training set. As a result, this new optimisation only tightens the bound, rather than maximising the likelihood, As the aim is to use the hidden groundtruths only to evaluate the likelihood and not improve it.

Table 6.1: Bayesian recovery from noisy data with different percentages of missing entries. Table shows the ELBO assigned by the retrieved posteriors to the ground truth clean data. Two-sample p-values between the TAE results and the best of the competing strategies are also reported for statistical significance.

	MNIST		Fashion-MNIST		UCI HAR	
	50%	80%	50%	80%	50%	80%
MVAE	870 ± 6	803 ± 15	757 ± 1	723 ± 7	585 ± 4	471 ± 10
MIWAE	917 ± 4	780 ± 6	800 ± 7	766 ± 8	613 ± 6	584 ± 8
TAE	$\mathbf{1719 \pm 7}$	$\mathbf{1536 \pm 14}$	$\mathbf{1326 \pm 7}$	$\mathbf{1094 \pm 13}$	$\mathbf{1014 \pm 6}$	$\mathbf{854 \pm 52}$
p-values	1×10^{-16}	6×10^{-13}	1×10^{-14}	1×10^{-11}	2×10^{-15}	2×10^{-8}

Discussion

As shown in figure 6.3(b-c), the mean inference performance on the MNIST data set at varying ratio of missing values is very similar for the two models (PSNR values), while the probabilistic performance (ELBO values) is significantly higher for the TAE model. We can see qualitative evidence of this difference in the reconstruction examples. The MVAE and TAE return similarly adequate mean solutions, but the MVAE posterior's draws are all very similar, suggesting that the posterior has collapsed on a particular reconstruction. Contrarily, the posteriors returned by the TAE explore different possible solutions that are consistent with the associated corrupted observation. The ELBO values recovered with other data sets shown in table 6.1 also show the same behaviour; the TAE posteriors return significantly improved ELBO values, suggesting that the PDFs of reconstructed solutions capture more adequately the true underlying solution space.

6.2.2 Posterior Recovery in Different Conditions

Structured Missings

A TAE is tested in a situation analogous to that shown in figure 6.3, but with structured missing values instead of randomly missing ones. Note that this situation is different from the missing not-at-random case, where the pattern of missing entries is dependent on the clean sample. This condition still falls within the missing at random case, as the structure of the missing entries is specific, but independent of the clean data. For each sample in MNIST, only a small window of 10×10 pixels is made visible, randomly placed in each example, while the rest of the images remain hidden. In addition, the values in the observed window are subject to additive Gaussian noise, similarly to the missing-at-random case. Reconstructions with the comparative MVAE and our TAE are shown in figure 6.4.

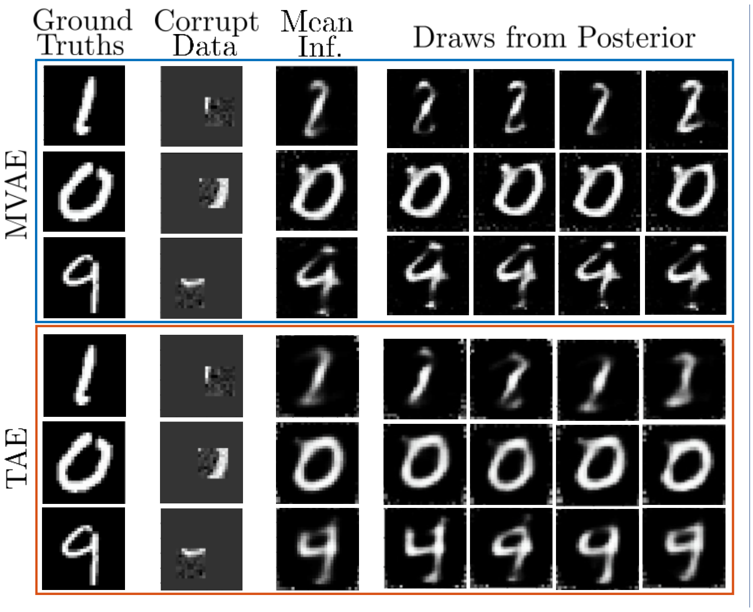


Figure 6.4: Examples of Bayesian reconstructions with MVAE and TAE on structured missing values.

Similarly to the missing-at-random case, the MVAE collapses on single solutions, giving draws from the posterior that are all very similar to each other. Contrarily, the TAE gives more variation in the possible solutions exploring more appropriately the uncertainty in the solution space. The MVAE ELBO over the clean data for this problem is 428, while the TAE one is 638. The performance improvement provided by the TAE is analogous to that observed with missing-at-random experiments.

Imputation Without Noise and De-noising

Experiments on MNIST analogous to those shown in figure 6.3 are performed, but, firstly, in the absence of noise, in order to test performance on imputation alone, and secondly, testing fully observed images at different levels of noise, in order to test de-noising alone. Each tested ratio of observed entries, or noise levels, is repeated three times, re-generating the patterns of missings each repeat in order to obtain error bars. Results are shown in figures 6.5 and 6.6 respectively.

In the case of imputation on clean data, shown in figure 6.5, the PSNR values between the MVAE and the TAE are very similar. The TAE presents significantly superior ELBO values at low ratios of observed entries, but in this case, the gap is reduced as more entries are observed. This is because in the noiseless case, the solution space when most entries are observed is much more localised than in the noisy case, and therefore the MVAE collapsed posteriors do not fail as much to capture it. In the case of denoising (figure 6.6), as in the missing value impution case, the $MVAE$ and TAE perform very similarly in their mean reconstructions, but the TAE presents significantly better performance in capturing the distributions of clean solutions, as the test ELBO values are higher.

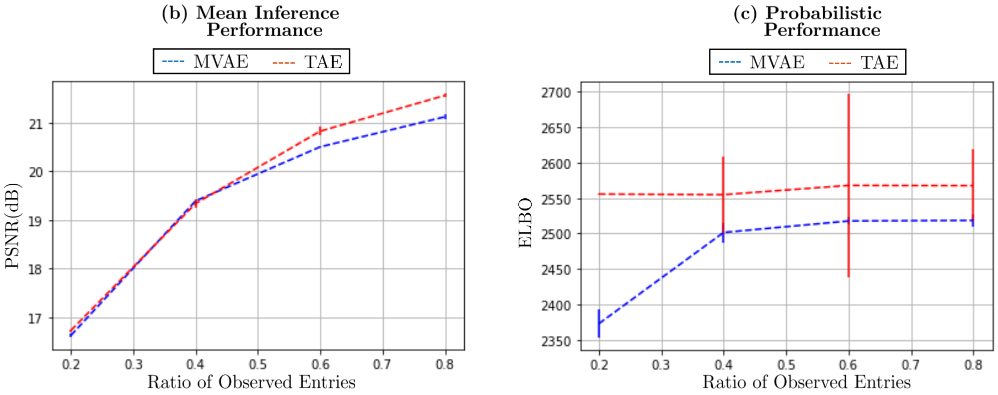


Figure 6.5: Missing value imputation performance on MNIST in the absence of noise.

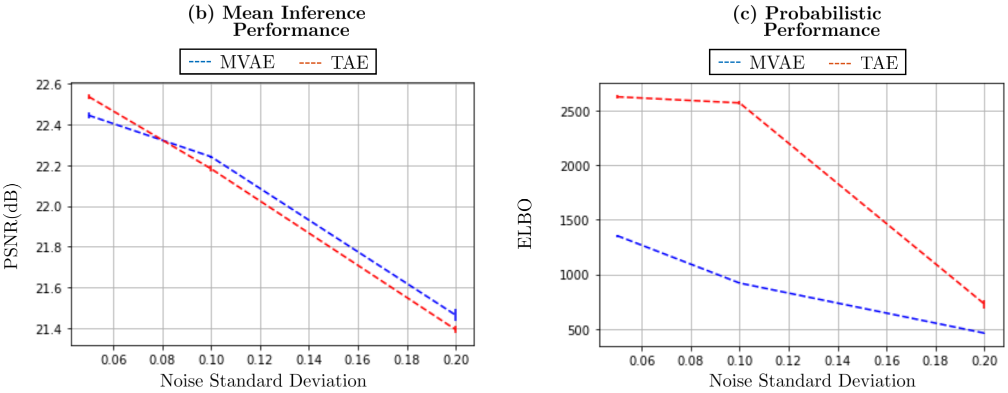


Figure 6.6: De-noising performance on MNIST.

Robustness to Hyper-parameters Choice

C and λ in equation 6.20 are hyper-parameters of our inference algorithm and need to be user defined. In the experiments using the TAE, the optimal values are determined by cross-validation, as described in supplementary B.2.1. However, the performance of TAEs was found to be rather stable with respect to the choice of these hyper-parameters. Figure 6.7 shows a cross validation study where the TAE ELBO for MNIST is measured with 90% missing values and additive noise.

As shown in figure 6.7, the performance of TAEs is robust to variations in hyper-parameters C and λ over a broad range of values. If the values are too large, the model collapses during optimisation, making such situations easy to diagnose. The two parameters also have an intuitive meaning that helps in their selection. In practice, C controls the final value of localisation and is desirable to be as high as stability of the optimisation allows. λ controls how fast one is imposing the model to approach C.

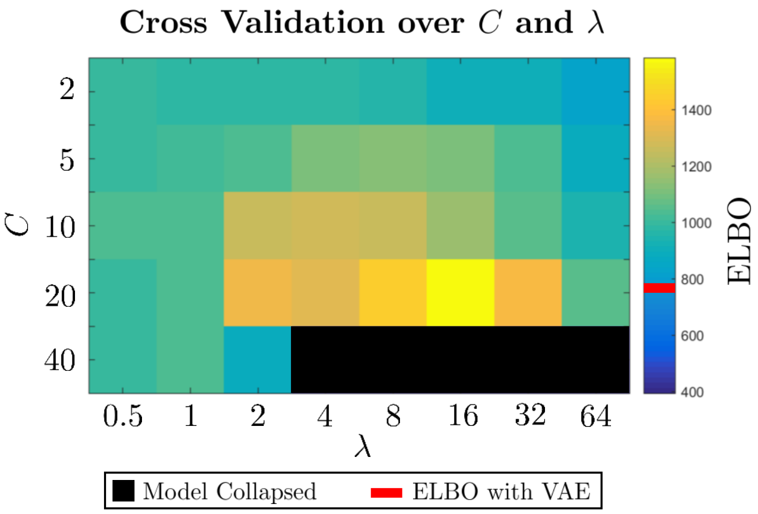


Figure 6.7: ELBO for MNIST with 90% missing values and additive noise as a function of chosen hyper-parameters C and λ (in log scale).

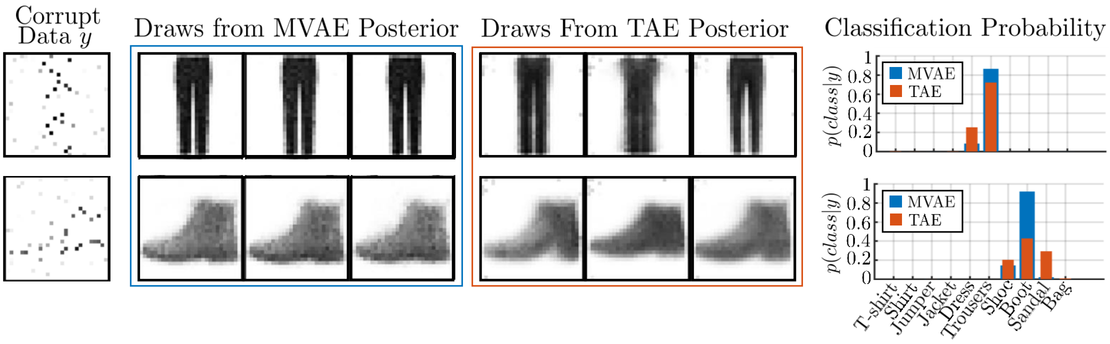


Figure 6.8: Propagating uncertainty to a classification task.

6.2.3 Downstream Tasks

To investigate the advantage of capturing complex uncertainties with the TAE model, performance is tested in downstream tasks. classification performance is tested on subsets of the MNIST and Fashion-MNIST data sets, after recovery with the TAE. With both sets, experiments are performed in situations where 10.000 examples are available, but corrupted with missing entries and noise. $1,000$ of these are labelled with one of 10 possible classes and the aim is to classify the remaining $9,000$. To do so, the TAE model is first trained on the full set, then the recovered posteriors are used to generate multiple possible cleaned data for the labelled sub-set and these are finally used to train a classifier.

To perform classification on the $9,000$ remaining examples, multiple possible cleaned data samples are generated with the variational posteriors. Then, for each posterior sample, classification is performed and the results are used to build histograms. Examples are shown in figure 6.8. Draws from the MVAE posterior are all very similar to each other. As a result, the imputed images are almost always classified in the same way and the uncertainty of the task is underestimated. The TAE posterior explores varied possible solutions to the recovery task. These can be recognised as different classes, resulting in less concentrated distributed probabilities that better reflect the associated uncertainty.

To evaluate the performance, the class with the largest histogram is taken as the inferred one. This experiment is performed for different ratios of missing values and several repetitions, varying the subsets of labelled and unlabelled data used. Classification accuracy results are shown in figure 6.9. Classifying using TAE imputations gives an advantage in this downstream task over using raw corrupted data and MVAE imputations, especially when the number of missing entries is high. This is because the MVAE collapses on single imputations, while the TAE generates diverse samples for each corrupted observation. The TAE classifier trains with data augmentations consistent with observed corrupted images, instead of single estimates.

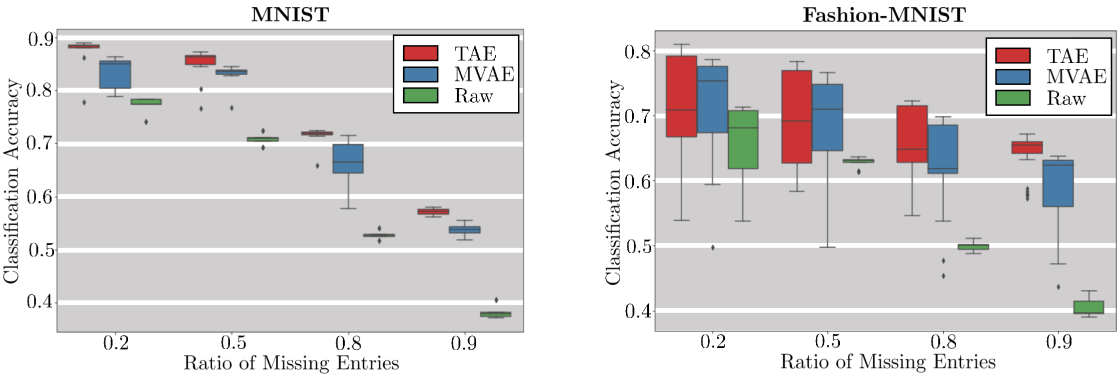


Figure 6.9: Classification accuracy after imputation.

6.2.4 Missing Values in the NYU Depth Maps

In order to test the TAE architecture on a real unsupervised recovery problem, involving large natural images, a convolutional version of the model is used to perform structured missing value imputation on depth maps of indoors rooms collected with a Kinect depth sensor. Missing entries are very common in depth maps recorded with such structured light sensors [149], as semi-reflective surfaces facing away from the imaging instrument often deflect light away from the sensor. Raw depth data from the NYU rooms dataset is used. This data set is commonly used to test various computer vision systems [140, 150, 151, 152] exploiting 3D information. This data set is composed of both RGB and matched depth images of indoors rooms, 608×480 pixels in size. A small sub-set of the depth maps has been corrected by imputing the missing entries and is popularly employed to train and test various learning systems [150, 151, 152].

A large portion of the set is available only as raw data, which presents missing entries. These are especially concentrated around objects' edges and reflecting surfaces, breaking the common assumption of missing at random, making this task particularly challenging. The TAE model is trained with a subset of this raw data set to perform imputation. In this experiments, there is no ground-truth data with which one can quantitatively test the performance, as the data employed comes from a real unsupervised imputation scenario, where there is no examples of targets at all. This experiment demonstrates how the TAE framework can be applied with relative ease to large real problems with high dimensional data (608×480 images) and give qualitatively reasonable results which avoid the problems of other tested approaches. Examples are shown in figure 6.10.

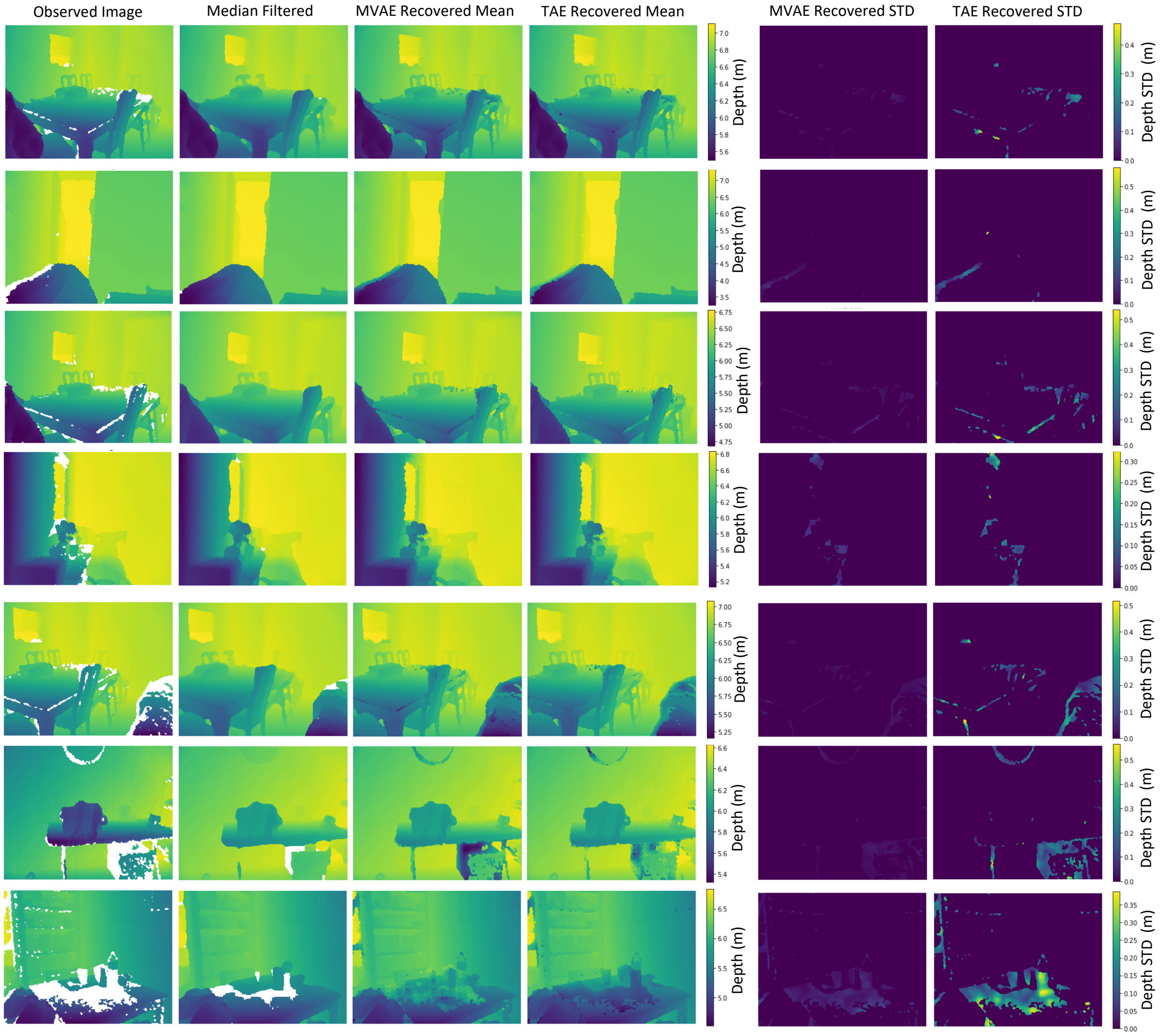


Figure 6.10: Unsupervised missing value imputation with the TAE model on raw depth maps from the NYU rooms data set, compared with a median filter approach and the standard MVAE. Missing pixels in the observed images are in white.

The median filter results in overly smoothed images and is unable to fill pixels that are surrounded by large missing areas. The MVAE returns adequate reconstructions, however, it over-fits to inaccurate solutions in certain locations, returning low uncertainty. The TAE returns good reconstructions and assigns high uncertainty to locations where reconstructions are most inaccurate, as shown by the marginal standard deviations. The TAE generates possibilities for the imputed pixels, which can be aggregated to recover a mean and standard deviation to quantify uncertainty in the retrieved imputations. The imputation of this data set was carried out with no signal prior, no particular domain expertise, temporal or structural assumptions and no associated RGB images or examples of complete data. A convolutional version of the TAE algorithm was simply run on the raw depth maps to impute the missing values.

Chapter 7

Conclusion

Probabilistic machine learning models hold a lot of potential for solving inverse problems in many scientific and technological domains. However, there are several scalability obstacles that make the application of these models impractical for real world settings. In this thesis, these practical issues were considered and addressed by modelling the learning process incorporating all available data, supervised and unsupervised, and domain expertise about the observation process, often available in the form of closed form mappings. In particular, this thesis proposed two novel frameworks for training probabilistic recovery models for inverse problems in two typical situations:

1. When acquiring paired ground-truth and measurements is expensive, and therefore supervised data is scarce.
2. when there is no examples of ground-truths, but only observations.

The novel techniques proved successful at training probabilistic recovery models in these settings and were demonstrated in a number of technically challenging, practical applications from the forefront of science and engineering. The new capabilities are expected to broaden the applicability of learning methods to more practical scenarios and allow for the development of machine learning enabled systems which are capable to self-assess their own accuracy and be deployed safely and robustly, without excessive data collection costs.

7.1 The VICI Framework

Capturing uncertainty when solving inverse problems is critical, especially for machine learning retrieval methods, where reconstructions are often compelling, even when not accurate. Whether the recoveries are interpreted by humans or passed on to further computations, recovering accurate estimates of uncertainty provides essential robustness, as systems can self-evaluate the confidence in their retrievals and be deployed safely. Through the experiments of chapters 3, 4 and 5 it was demonstrated how learning accurate probabilistic models that return uncertainties requires large amounts of supervised data, which in many practical applications, such as those considered in this thesis, is impractical, as it adds large costs to the development of reconstruction systems, because of the associated data collections. The VICI framework of section 3 greatly mitigates this limitation, as it offers a principled way to incorporate in the learning process unsupervised examples of targets and physical models of the observation process, which are typically cheap sources of information. Both in the simulated controlled experiments of chapter 3 and the application experiments of chapters 4 and 5, probabilistic models trained with this framework were proven to give accurate reconstructions and associated measures of uncertainty with limited data from physical acquisitions.

Chapters 4 and 5 demonstrated the advantages of applying the VICI framework of chapter 3 in several practical applications. The applications of these chapters demonstrated how applying the novel framework provides reconstruction systems of high accuracy and adequate uncertainty quantification with limited experimental acquisitions. These results suggest that using the VICI framework would allow the application of deep learning for inverse problems in real scenarios more systematically. While with standard supervised learning, data would need to be collected for each newly developed sensing system, with the framework of VICI, a large data set of ground-truths can be continuously collected and used for any new reconstruction or interaction system being developed, requiring only small amounts of data with the new apparatus to be incorporated as high-fidelity samples. Adopting this framework is expected to provide robust and accurate systems with greatly reduced data collection costs in imaging, astronomy and HCI.

In general, the VICI framework allows for machine learning models applied to inverse problems to be less reliant on supervised data, compared to standard implementations, and instead draw training information from various sources, e.g. unsupervised data and domain expertise. This means that, in the development and continuous improvement of reconstruction systems, developers have a choice of how to spend time and resources in the most efficient way in order to build and improve their models. In some situations, collecting more paired ground-truths and measurements can be relatively inexpensive and is the best source to invest in, while in others it can be extremely time consuming and spending resources on improving analytical models or gathering ground-truth data alone is comparably much more efficient. The VICI framework allows these choices to be made and to combine any and all of the available sources of information in a principled way, adapting to corner cases with little to no structural change.

The main limitation of the VICI framework is the difficulty in evaluating uncertainty accurately for inputs which are very different from training ones. The models constituting the learning framework are generally good at capturing aleatoric uncertainty, i.e. the inherent variability of the forward or inverse model, but struggle to capture epistemic uncertainty, i.e. the variability corresponding to model uncertainty. This is a limitation that broadly affects deep neural networks and is currently intensively studied in machine learning. What this means for the VICI framework, is that it can successfully train models with sparse supervised examples, as demonstrated in chapters 3, 4 and 5, but these examples need to adequately span the space of targets of interest. This still provides a way to greatly limit collection costs, but makes it difficult to adapt models to completely new tasks without any experimental data. An interesting future direction to address this limitation is that of adapting recent advances in approximate Bayesian neural networks, such as dropout or deep ensembles, to the models composing the VICI framework. In this way, the uncertainty returned by both forward and inverse models would be more accurate, both near and far away from training data, opening the attractive possibility to adapt probabilistic reconstruction models to new domains with entirely simulated data.

A second attractive possibility for future research directions, allowed by the novel capability of VICI to incorporate unsupervised ground-truth examples, is that of using unsupervised generative models during the training process. Few available examples of ground-truths, whether paired to measurements or not, can be used to train a generative model, such as a VAE or a GAN. The trained generative model can then be incorporated in the training of the inverse by continuously generating ground-truth examples and inferring corresponding measurements with the multi-fidelity forward model. This would add a further level of data augmentation to the training of the inverse model, expectedly improving performance and robustness further.

7.2 The TAE Framework

Chapter 6 addresses the particular situation of unsupervised training of reconstruction systems, where only examples of observations are available, along with some model of the observation process. This situation is commonly encountered in data pre-processing and information retrieval, where the quantities of interest are only partially available, with missing entries and noise, or other corruptions. In these settings capturing uncertainty is extremely important. If there are errors in data processing stages, these will be unavoidably passed on to subsequent analysis and cause mistakes that are difficult to avoid or even diagnose. If instead uncertainty is captured properly at the pre-processing stage, it can be propagated to these subsequent tasks and provide accurate error estimates downstream. Application of existing variational frameworks for these pre-processing tasks often results in a collapse of the uncertainty estimation. The problem is analysed in detail in chapter 6 and a novel approach, which avoids this pathology, is presented. The novel method allows to recover accurate posterior densities of clean data, given corrupted ones and result in more accurate and adequately uncertain inferences in subsequent tasks.

The TAE framework allows unsupervised data cleaning and recovery to be performed with adequate uncertainty estimation. In future work, the method could be expanded to properly model the missing-not-at-random situation, where the observation process is dependent on the hidden clean data, e.g. entries more likely to be missing if the underlying entry takes certain values. A similar strategy to that used by the TAE to impose a prior for clean data and recover its PDF could be adapted to jointly model the observation model and its dependency on the clean data itself. This would allow the extension of TAEs to the missing-not-at-random case and extract useful information about the clean data from the observation parameters, e.g. the pattern of missings.

Appendix A

Semi-Supervised Experiments

A.1 Experimental Conditions

A.1.1 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 4.2(c), where the HIO algorithm simply predicts spots at some positions.

A.1.2 ToF Diffuse Imaging

The comparative iterative method was taken from [114], 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 3.8, 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 3.9, with all deterministic intermediate layers having 1500 hidden units and latent variables z having 800 dimensions.

A.2 Details of the Models' Architectures

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

A.2.1 Multi-Fidelity Forward Model

The multi-fidelity forward model includes three parametric distributions, the parameters of which are optimised during training (see figure 3.2); $p_{\alpha_1}(w|x,\widetilde{y})$, $p_{\alpha_2}(y|x,\widetilde{y},w)$ 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 3.8. In the second, the parametric distributions consist of deeper convolutional recurrent layers, again mapping inputs to outputs' Gaussian moments, from which samples are drawn upon training and inference. These structures are instead shown in figure 3.5.

A.2.2 Variational Inverse Model

Like the multi-fidelity forward model, the inverse model includes three parametric distributions (see figure 3.3); $p_{\theta_1}(z|y)$, $p_{\theta_2}(x|y,z)$ 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 3.9. In 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 [62]. These structures are instead shown in figure 3.6.

Appendix B

Unsupervised Experiments

B.1 Evaluation ELBO

To evaluate the probabilistic performance of our method compared to others, we compute an evaluation ELBO which relies on test ground truths. After each model is trained unsupervisedly, we obtain a posterior of the form $q(x|y) = \int q(z|y)q(x|z)dz$, where for the MVAE and MIWAE, $q(x|z) = p(x|z)$. Given the a test set of paired clean and corrupted samples x_t and y_t, we construct a new parametric recognition model, which encodes latent distributions from ground-truths $q_\eta(z|x)$. We then optimise the following:

$$\arg\max_{\eta} \quad \mathbb{E}_{q_\eta(z|x_t)} \log q(x_t|z) + KL(q_\eta(z|x_t)|q(z|y_t)). \tag{B.1}$$

The above is a conditional VAE ELBO with conditional prior $q(z|y)$ and is a lower bound to the test likelihood we are interested in $q(x_t|y_t)$. Note that we optimise over η only, therefore the new recognition model $q(z|x)$ is the only one which is affected by this optimisation and the components of our reconstruction model $q(z|y)$ and $q(x|z)$ remain the same as trained with the unsupervised training set. As a result, this new optimisation only tightens the bound, rather than maximising the likelihood, which we want to evaluate as trained previously.

B.2 Experimental Conditions

B.2.1 Posterior Recovery

All posterior recovery experiments, with each of the three data sets tested, are performed on samples that have been re-scaled from 0 to 1. In all cases, the sets are injected with additive Gaussian noise having standard deviation 0.1. Subsequently, random binary masks are generated to block out some entries, resulting in missing values. The proportion of missing entries in the masks was set as described in the main body in each case.

Experiments were repeated with re-generated binary masks 5 times. The means and error bars shown in figure 4 and the uncertainty reported in table 1 were computed from these. The MIWAE was trained with 20 weights per sample. After training, all posteriors $q(x|y)$ have identical structure and are tested in the same way, by training an inference network on the test set to compute the ELBO values.

B.2.2 Classification Experiments

The TAE models for the MNIST and Fashion-MNIST experiments were trained in the conditions described above. In each case, a random subset of $10,000$ samples is taken from the corrupted set and the TAE and MVAE models are trained with it. A random subset of $1,000$ of these is selected and ground-truth lables for these samples are made available.

A classifier consisting in a single fully connected layer with leaky ReLu non-linearity is trained to perform classification on this subset. For each stochastic training iteration of this classifier, we generate samples associated with the corrupted observations and provide the associated labels. After the classifier is trained, we test classification performance on the remaining $9,000$ examples, by running the train classifier 400 times per sample, each time generating clean data from a corrupted observation with the TAE and the MVAE. The histograms shown in figure 5 are built by aggregating the resulting classification.

The above procedure is repeated 15 times. The resulting means and standard deviations of the tested classification performance are shown in figure 6.

B.2.3 Training Conditions

Hyper-parameters of optimisation for the models were cross validated with the MNIST data set at a proportion of missing entries of 0.9. Hyper-parameters common to all models were determined by obtaining best performance with the MVAE model. Hyper- parameters specific to the TAE model were obtained by fixing the common parameters and cross validating these. The resulting optimal hyper parameters were then used in all other experiments of section 4.1 and 4.2, including those with different data sets. Common parameters are as follows: $500,000$ iterations with the ADAM optimiser in Tensorflow, an initial training of 2^{-4} and batch size of 20. The hyper-parameters specific to the TAE are instead: γ initially set to 0.01 and then linearly increased to 1 between $50,000$ and $100,000$ iterations, $\lambda = 2$ and $C = 10$. All experiments were performed using a TitanX GPU.

B.2.4 NYU Rooms Experiments

For these experiments, we take a subset of 3612 depth maps from the NYU raw data set. We slightly crop these in one dimension to be 480×608 images. The convolutional TAE and MVAE to obtain the results of figure 7, were trained for $100,000$ iterations using the ADAM optimiser in Tensorflow, with a batch size of 20 images and an initial training rate of 2×10^{-2}. For the warm up, we initially set $\gamma = 0.01$ and linearly increase it to 1 between $10,000$ and $20,000$ iterations.

Bibliography

[1] M. Bertero and P. Boccacci, *Introduction to inverse problems in imaging*. CRC press, 1998.

[2] C. R. Vogel, *Computational methods for inverse problems*. SIAM, 2002, vol. 23.

[3] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, “Using deep neural networks for inverse problems in imaging: beyond analytical methods,” *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 20–36, 2018.

[4] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, “Deep learning in medical imaging: general overview,” *Korean Journal of Radiology*, vol. 18, no. 4, pp. 570–584, 2017.

[5] S. Alghunaim and H. H. Al-Baity, “On the scalability of machine-learning algorithms for breast cancer prediction in big data context,” *IEEE Access*, vol. 7, pp. 91 535–91 546, 2019.

[6] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[7] B. Collins, J. Deng, K. Li, and L. Fei-Fei, “Towards scalable dataset construction: An active learning approach,” in *European conference on computer vision*. Springer, 2008, pp. 86–98.

[8] J. Adler and O. Öktem, “Deep Bayesian inversion,” *arXiv preprint arXiv:1811.05910*, 2018.

[9] C. Zhang and B. Jin, “Probabilistic residual learning for aleatoric uncertainty in image restoration,” *arXiv preprint arXiv:1908.01010*, 2019.

[10] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, “Advances in variational inference,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[11] F. Tonolini, J. Radford, A. Turpin, D. Faccio, and R. Murray-Smith, "Variational inference for computational imaging inverse problems," *Journal of Machine Learning Research*, vol. 21, no. 179, pp. 1–46, 2020.

[12] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, "Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy," *Nature Physics*, vol. 18, no. 1, pp. 112–117, 2022.

[13] R. Murray-Smith, J. H. Williamson, A. Ramsay, F. Tonolini, S. Rogers, and A. Loriette, "Forward and Inverse models in HCI: Physical simulation and deep learning for inferring 3D finger pose," *arXiv preprint arXiv:2109.03366*, 2021.

[14] F. Tonolini, P. G. Moreno, A. Damianou, and R. Murray-Smith, "Tomographic autoencoder: Unsupervised bayesian recovery of corrupted data," *International Conference on Learning Representations (ICLR)*, 2021.

[15] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[16] J.-L. Starck, M. K. Nguyen, and F. Murtagh, "Wavelets and curvelets for image deconvolution: a combined approach," *Signal processing*, vol. 83, no. 10, pp. 2279–2283, 2003.

[17] J. M. Bioucas-Dias, M. A. Figueiredo, and J. P. 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*, vol. 2. IEEE, 2006, pp. II–II.

[18] L. Ritschl, F. Bergner, C. Fleischmann, and M. Kachelrieß, "Improved total variation-based CT image reconstruction applied to clinical data," *Physics in Medicine & Biology*, vol. 56, no. 6, p. 1545, 2011.

[19] F. Magalhães, F. M. Araújo, M. V. Correia, M. Abolbashari, and F. Farahi, "Active illumination single-pixel camera based on compressive sensing," *Applied optics*, vol. 50, no. 4, pp. 405–414, 2011.

[20] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[21] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[22] I. Daubechies, M. Defrise, and C. 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*, vol. 57, no. 11, pp. 1413–1457, 2004.

[23] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[24] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE Journal of selected topics in signal processing*, vol. 1, no. 4, pp. 586–597, 2007.

[25] P. Jain, P. Netrapalli, and S. Sanghavi, “Low-rank matrix completion using alternating minimization,” in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 2013, pp. 665–674.

[26] B. Vandereycken, “Low-rank matrix completion by riemannian optimization,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1214–1236, 2013.

[27] J. Yang and Y. Zhang, “Alternating direction algorithms for ℓ_1-problems in compressive sensing,” *SIAM journal on scientific computing*, vol. 33, no. 1, pp. 250–278, 2011.

[28] E. J. Candes, J. K. Romberg, and T. 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*, vol. 59, no. 8, pp. 1207–1223, 2006.

[29] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, “An iterative regularization method for total variation-based image restoration,” *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 460–489, 2005.

[30] B. Sun, M. P. Edgar, R. Bowman, L. E. Vittert, S. Welsh, A. Bowman, and M. J. Padgett, “3D computational imaging with single-pixel detectors,” *Science*, vol. 340, no. 6134, pp. 844–847, 2013.

[31] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, “Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging,” *Nature communications*, vol. 3, p. 745, 2012.

[32] M. T. McCann, K. H. Jin, and M. Unser, “A review of convolutional neural networks for inverse problems in imaging,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 85–95, November 2017.

[33] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, p. 2346, 2008.

[34] D. Gamerman and H. F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[35] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1460–A1487, 2012.

[36] Y. Marzouk and D. Xiu, "A stochastic collocation approach to Bayesian inference in inverse problems," *Communications in Computational Physics*, vol. 6, no. 4, pp. 826–847, 2009.

[37] A. Malinverno, "Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem," *Geophysical Journal International*, vol. 151, no. 3, pp. 675–688, 2002.

[38] A. Mohammad-Djafari, "Bayesian inference tools for inverse problems," in *AIP Conference Proceedings*, vol. 1553, no. 1. AIP, 2013, pp. 163–170.

[39] P. Tsilifis, I. Bilionis, I. Katsounaros, and N. Zabaras, "Computationally efficient variational approximations for Bayesian inverse problems," *Journal of Verification, Validation and Uncertainty Quantification*, vol. 1, no. 3, p. 031004, 2016.

[40] A. Likas and N. P. Galatsanos, "A variational approach for Bayesian blind image deconvolution," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2222–2233, 2004.

[41] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.

[42] ——, "Variational Bayesian blind deconvolution using a total variation prior," *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 12–26, 2009.

[43] M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networks – a review," *Pattern recognition*, vol. 35, no. 10, pp. 2279–2301, 2002.

[44] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[45] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution

using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[46] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 449–458.

[47] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5162–5170.

[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, vol. abs/1611.07004, 2017.

[49] T. H. H. Maung, "Real-time hand tracking and gesture recognition system using neural networks," *World Academy of Science, Engineering and Technology*, vol. 50, pp. 466–470, 2009.

[50] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 7356–7365.

[51] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," *Proceedings of the 34th International Conference on Machine Learning, PMLR*, pp. 537–546, 2017.

[52] J.-H. R. Chang, C.-L. Li, B. Poczos, B. V. K. V. Kumar, and A. C. Sankaranarayanan, "One network to solve them all-solving linear inverse problems using deep projection models." in *ICCV*, 2017, pp. 5889–5898.

[53] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 84–98, 2017.

[54] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3929–3938.

[55] H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Transactions on medical imaging*, vol. 38, no. 2, pp. 394–405, 2019.

[56] Y. Chen, W. Yu, and T. Pock, "On learning optimized reaction diffusion processes for effective image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5261–5269.

[57] P. Putzky and M. Welling, "Recurrent inference machines for solving inverse problems," *arXiv preprint arXiv:1706.04008*, 2017.

[58] J. Adler and O. Öktem, "Solving ill-posed inverse problems using iterative deep neural networks," *Inverse Problems*, vol. 33, no. 12, p. 124007, 2017.

[59] 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*, vol. 32, no. 1, pp. 167–179, Jan 2019.

[60] P. Hand, O. Leong, and V. Voroninski, "Phase retrieval under a generative prior," in *Advances in Neural Information Processing Systems*, 2018, pp. 9154–9164.

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

[62] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "PixelVAE: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[63] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[64] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[65] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[66] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space." in *CVPR*, vol. 2, no. 5, 2017, p. 7.

[67] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the 34th International Conference on Machine Learnin (ICML), PMLR 70*, 2017, pp. 2642–2651.

[68] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[69] X. Yan, J. Yang, K. Sohn, and H. Lee, “Attribute2image: Conditional image generation from visual attributes,” in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

[70] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[71] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, “Handling incomplete heterogeneous data using VAEs,” *Pattern Recognition, 107501*, vol. 107, Nov.

[72] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran, “Image transformer,” *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80*, pp. 4055–4064, 2018.

[73] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[74] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, “Auxiliary deep generative models,” in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[75] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, “Hybrid models with deep and invertible features,” in *Proc. 36th International Conference on Machine Learning (ICML), Long Beach, California, PMLR 97.*, 2019.

[76] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, “Noise2noise: Learning image restoration without clean data,” *arXiv preprint arXiv:1803.04189*, 2018.

[77] A. Krull, T.-O. Buchholz, and F. Jug, “Noise2void-learning denoising from single noisy images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2129–2137.

[78] A. Krull, T. Vicar, and F. Jug, “Probabilistic Noise2Void: unsupervised content-aware denoising,” *arXiv preprint arXiv:1906.00651*, 2019.

[79] D. I. J. Im, S. Ahn, R. Memisevic, and Y. Bengio, "Denoising criterion for variational auto-encoding framework," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[80] A. Creswell and A. A. Bharath, "Denoising adversarial autoencoders," *IEEE Transactions on neural networks and learning systems*, vol. 30, no. 4, pp. 968–984, 2018.

[81] S. C.-X. Li, B. Jiang, and B. Marlin, "MisGAN: Learning from incomplete data with generative adversarial networks," *arXiv preprint arXiv:1902.09599*, 2019.

[82] J. Yoon, J. Jordon, and M. Van Der Schaar, "Gain: Missing data imputation using generative adversarial nets," *arXiv preprint arXiv:1806.02920*, 2018.

[83] Y. Luo, X. Cai, Y. Zhang, J. Xu *et al.*, "Multivariate time series imputation with generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 1596–1607.

[84] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using VAEs," *arXiv preprint arXiv:1807.03653*, 2018.

[85] P.-A. Mattei and J. Frellsen, "MIWAE: deep generative modelling and imputation of incomplete data sets," in *International Conference on Machine Learning*, 2019, pp. 4413–4423.

[86] C. Ma, S. Tschiatschek, K. Palla, J. M. Hernández-Lobato, S. Nowozin, and C. Zhang, "EDDI: Efficient dynamic discovery of high-value information with partial VAE," *arXiv preprint arXiv:1809.11142*, 2018.

[87] M. Collier, A. Nazabal, and C. K. Williams, "VAEs in the presence of missing data," *arXiv preprint arXiv:2006.05301*, 2020.

[88] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. J. Rezende, and S. A. Eslami, "Conditional neural processes," in *ICML*, 2018.

[89] C. Shang, A. Palmer, J. Sun, K.-S. Chen, J. Lu, and J. Bi, "VIGAN: Missing view imputation with generative adversarial networks," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 766–775.

[90] S. K. Ainsworth, N. J. Foti, and E. B. Fox, "Disentangled VAE representations for multi-aspect and missing data," *arXiv preprint arXiv:1806.09060*, 2018.

[91] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," *arXiv preprint arXiv:1808.00892*, 2018.

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

[93] B. Peherstorfer, K. Willcox, and M. Gunzburger, “Survey of multifidelity methods in uncertainty propagation, inference, and optimization,” *SIAM Review*, vol. 60, no. 3, pp. 550–591, 2018.

[94] J. Kaipio and E. Somersalo, “Statistical inverse problems: discretization, model reduction and inverse crimes,” *Journal of computational and applied mathematics*, vol. 198, no. 2, pp. 493–504, 2007.

[95] J. A. Christen and C. Fox, “Markov chain Monte Carlo using an approximation,” *Journal of Computational and Graphical statistics*, vol. 14, no. 4, pp. 795–810, 2005.

[96] M. C. Kennedy and A. O’Hagan, “Bayesian calibration of computer models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[97] M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu, “A framework for validation of computer models,” *Technometrics*, vol. 49, no. 2, pp. 138–154, 2007.

[98] Y. Yang and P. Perdikaris, “Conditional deep surrogate models for stochastic, high-dimensional, and multi-fidelity systems,” *Computational Mechanics*, vol. 64, pp. 417–434, 2019.

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

[100] A. Krizhevsky, “Learning multiple layers of features from tiny images,” Master’s thesis, University of Toronto, 2009.

[101] K. Matsuoka, “Noise injection into inputs in back-propagation learning,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 436–440, 1992.

[102] J. C. Russ, J. R. Matey, A. J. Mallinckrodt, and S. McKay, “The image processing handbook,” *Computers in Physics*, vol. 8, no. 2, pp. 177–178, 1994.

[103] R. W. Gerchberg, “A practical algorithm for the determination of phase from image and diffraction plane pictures,” *Optik*, vol. 35, pp. 237–246, 1972.

[104] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[105] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[106] A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," *Optica*, vol. 4, no. 9, pp. 1117–1125, 2017.

[107] Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light: Science & Applications*, vol. 7, no. 2, p. 17141, 2018.

[108] Y. Rivenson, Y. Wu, and A. Ozcan, "Deep learning in holography and coherent imaging," *Light: Science & Applications*, vol. 8, no. 1, pp. 1–8, 2019.

[109] S. Marchesini, H. Chapman, S. Hau-Riege, R. London, A. Szoke, H. He, M. Howells, H. Padmore, R. Rosen, J. Spence, and U. Weierstall, "Coherent X-ray diffractive imaging: applications and limitations," *Optics Express*, vol. 11, no. 19, pp. 2344–2353, 2003.

[110] D. A. Barmherzig, J. Sun, P.-N. Li, T. J. Lane, and E. J. Candès, "Holographic phase retrieval and reference design," *Inverse Problems*, 2019.

[111] A. Jesacher, C. Maurer, A. Schwaighofer, S. Bernet, and M. Ritsch-Marte, "Full phase and amplitude control of holographic optical tweezers with high efficiency," *Optics express*, vol. 16, no. 7, pp. 4479–4486, 2008.

[112] H. M. L. Faulkner and J. Rodenburg, "Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm," *Physical review letters*, vol. 93, no. 2, p. 023903, 2004.

[113] H. Jiang, *Diffuse optical tomography: principles and applications*. CRC press, 2018.

[114] A. Lyons, F. Tonolini, A. Boccolini, A. Repetti, R. Henderson, Y. Wiaux, and D. Faccio, "Computational time-of-flight diffuse optical tomography," *Nature Photonics*, vol. 13, pp. 575–579, 2019.

[115] S. G. Johnson, "NIST Special Database 30," 2010.

[116] K. Yoo, F. Liu, and R. Alfano, "When does the diffusion approximation fail to describe photon transport in random media?" *Physical review letters*, vol. 64, no. 22, p. 2647, 1990.

[117] B. P. Abbott *et al.*, “Binary black hole mergers in the first advanced LIGO observing run,” *Phys. Rev. X*, vol. 6, p. 041015, Oct 2016. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.6.041015

[118] ——, “Gw170817: Observation of gravitational waves from a binary neutron star inspiral,” *Phys. Rev. Lett.*, vol. 119, p. 161101, Oct 2017. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.119.161101

[119] ——, “Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA,” *Living Reviews in Relativity*, vol. 21, no. 1, p. 3, Apr 2018.

[120] A. C. Searle, P. J. Sutton, and M. Tinto, “Bayesian detection of unmodeled bursts of gravitational waves,” *Classical and Quantum Gravity*, vol. 26, no. 15, p. 155017, Aug 2009.

[121] J. Veitch, V. Raymond, B. Farr, W. M. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin, W. D. Pozzo, F. Feroz, J. Gair, C.-J. Haster, V. Kalogera, T. Littenberg, I. Mandel, R. O’Shaughnessy, M. Pitkin, C. Rodriguez, C. Röver, T. Sidery, R. Smith, M. V. D. Sluys, A. Vecchio, W. Vousden, and L. Wade, “Robust parameter estimation for compact binaries with ground-based gravitational-wave observations using the lalinference software library,” *Physical Review D*, 2014.

[122] “GraceDB — gravitational-wave candidate event database (ligo/virgo o3 public alerts),” https://gracedb.ligo.org/superevents/public/O3/, accessed: 2019-09-16.

[123] L. P. Singer and L. R. Price, “Rapid Bayesian position reconstruction for gravitational-wave transients,” *Physical Review D*, vol. 93, no. 2, p. 024013, Jan 2016.

[124] “Advanced LIGO sensitivity design curve,” https://dcc.ligo.org/LIGO-T1800044/public, accessed: 2019-06-01.

[125] G. Ashton, M. Huebner, P. D. Lasky, C. Talbot, K. Ackley, S. Biscoveanu, Q. Chu, A. Divarkala, P. J. Easter, B. Goncharov, F. H. Vivanco, J. Harms, M. E. Lower, G. D. Meadors, D. Melchor, E. Payne, M. D. Pitkin, J. Powell, N. Sarin, R. J. E. Smith, and E. Thrane, “Bilby: A user-friendly Bayesian inference library for gravitational-wave astronomy,” *Astrophysical Journal Supplement Series*, 2018.

[126] S. Khan, K. Chatziioannou, M. Hannam, and F. Ohme, “Phenomenological model for the gravitational-wave signal from precessing binary black holes with two-spin effects,” 2018.

[127] Q. Wang, S. R. Kulkarni, and S. Verdú, “Divergence estimation for multidimensional densities via k-nearest-neighbor distances,” *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[128] B. P. Abbott *et al.*, “GW190425: Observation of a Compact Binary Coalescence with Total Mass 3.4 M,” *Astrophysical Journal Letters*, vol. 892, no. 1, p. L3, Mar. 2020.

[129] J. H. W. Roderick Murray-Smith and F. Tonolini, “Human–computer interaction design inverse problems,” in *Bayesian Methods for Interaction and Design*. Cambridge University Press, 2021.

[130] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, “Soli: Ubiquitous gesture sensing with millimeter wave radar,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[131] E. Hayashi, J. Lien, N. Gillian, L. Giusti, D. Weber, J. Yamanaka, L. Bedal, and I. Poupyrev, “Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[132] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[133] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, “Data cleaning: Overview and emerging challenges,” in *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2016, pp. 2201–2206.

[134] A. Nazabal, C. K. Williams, G. Colavizza, C. R. Smith, and A. Williams, “Data engineering for data analytics: a classification of the issues, and case studies,” *arXiv preprint arXiv:2004.12929*, 2020.

[135] M. A. Munson, “A study on the importance of and time spent on different modeling steps,” *ACM SIGKDD Explorations Newsletter*, vol. 13, no. 2, pp. 65–71, 2012.

[136] L. L. Geyer, U. J. Schoepf, F. G. Meinel, J. W. Nance Jr, G. Bastarrika, J. A. Leipsic, N. S. Paul, M. Rengo, A. Laghi, and C. N. De Cecco, “State of the art: iterative CT reconstruction techniques,” *Radiology*, vol. 276, no. 2, pp. 339–357, 2015.

[137] I. R. White, P. Royston, and A. M. Wood, “Multiple imputation using chained equations: issues and guidance for practice,” *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[138] S. K. Kwak and J. H. Kim, “Statistical data preparation: management of missing values and outliers,” *Korean journal of anesthesiology*, vol. 70, no. 4, p. 407, 2017.

[139] A. V. Dalca, J. Guttag, and M. R. Sabuncu, “Unsupervised data imputation via variational inference of deep subspaces,” *arXiv preprint arXiv:1903.03503*, 2019.

[140] N. Silberman and R. Fergus, “Indoor scene segmentation using a structured light sensor,” in *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011.

[141] M. K. Titsias and F. Ruiz, “Unbiased implicit variational inference,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 167–176.

[142] W. I. Zangwill, “Non-linear programming via penalty functions,” *Management science*, vol. 13, no. 5, pp. 344–358, 1967.

[143] M. Phuong, M. Welling, N. Kushman, R. Tomioka, and S. Nowozin, “The mutual autoencoder: Controlling information in latent code representations,” in *ICLR*, 2018.

[144] L. Deng, “The MNIST database of handwritten digit images for machine learning research [best of the web],” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[145] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms,” *arXiv:1708.07747*, 2017.

[146] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine,” in *Proceedings of the International Workshop on Ambient Assisted Living*. Springer, 2012, pp. 216–223.

[147] C. Cremer, X. Li, and D. Duvenaud, “Inference suboptimality in variational autoencoders,” in *Proc. 35th Inter. Conference on Machine Learning, PMLR 80*, 2018, pp. 1078–1086.

[148] P.-A. Mattei and J. Frellsen, “Leveraging the exact likelihood of deep latent variable models,” in *Advances in Neural Information Processing Systems*, 2018, pp. 3855–3866.

[149] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.

[150] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGB-D images,” in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[151] P. Dollár and C. L. Zitnick, “Structured forests for fast edge detection,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1841–1848.

[152] A. Chang, A. Dai, T. A. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3D: Learning from RGB-D data in indoor environments,” in *7th IEEE International Conference on 3D Vision, 3DV 2017*. Institute of Electrical and Electronics Engineers Inc., 2018, pp. 667–676.