Unsupervised knowledge-transfer for learned image reconstruction

To cite this article before publication: Riccardo Barbano *et al* 2022 *Inverse Problems* in press https://doi.org/10.1088/1361-6420/ac8a91

Unsupervised Knowledge-Transfer for Learned Image Reconstruction*

Riccardo Barbano† Željko Kereta† Andreas Hauptmann†‡ Simon R. Arridge†

Bangti Jin§

Abstract

Deep learning-based image reconstruction approaches have demonstrated impressive empirical performance in many imaging modalities. These approaches usually require a large amount of high-quality paired training data, which is often not available in medical imaging. To circumvent this issue we develop a novel unsupervised knowledge-transfer paradigm for learned reconstruction within a Bayesian framework. The proposed approach learns a reconstruction network in two phases. The first phase trains a reconstruction network with a set of ordered pairs comprising of ground truth images of ellipses and the corresponding simulated measurement data. The second phase fine-tunes the pretrained network to more realistic measurement data without supervision. By construction, the framework is capable of delivering predictive uncertainty information over the reconstructed image. We present extensive experimental results on low-dose and sparse-view computed tomography showing that the approach is competitive with several state-of-the-art supervised and unsupervised reconstruction techniques. Moreover, for test data distributed differently from the training data, the proposed framework can significantly improve reconstruction quality not only visually, but also quantitatively in terms of PSNR and SSIM, when compared with learned methods trained on the synthetic dataset only.

Keywords: Unsupervised Learning, Test-Time Adaptation, Pretraining, Image Reconstruction, Bayesian Deep Learning, Computed Tomography

1 Introduction

In this work we develop a novel unsupervised knowledge-transfer framework for image reconstruction. The reconstruction of an image is often formulated through a (linear) inverse problem

$$y = Ax + \delta y,$$

where $y \in Y$ is a corrupted measurement, δy is the additive noise, $x \in X$ is the image to be recovered, and the data acquisition is described by a linear forward map $A : X \to Y$, where X and Y are suitable finite-dimensional vector spaces.

*The work of R.B. is substantially supported by the i4health PhD studentship (UK EPSRC EP/S021930/1) and and from The Alan Turing Institute (UK EPSRC EP/N510129/1), and that of Z.K, S.A. and B.J. by UK EPSRC EP/T000864/1, and that of S.A. and B.J. also by UK EPSRC EP/V026259/1. AH acknowledges funding by Academy of Finland Projects 336796, 334817, 338408.

†Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK (riccardo.barbano.19@ucl.ac.uk,z.kereta@ucl.ac.uk, s.arridge@ucl.ac.uk)

‡Research Unit of Mathematical Sciences; University of Oulu, Oulu, Finland (Andreas.Hauptmann@oulu.fi)

§Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, P.R. China (bangti.jin@gmail.com,btjin@math.cuhk.edu.hk)

In the past few years, deep learning (DL)-based image reconstruction techniques have demonstrated remarkable empirical results, often substantially outperforming more conventional methods in terms of both image quality and computational efficiency [7, 44]. In DL-based approaches, image reconstruction can be phrased as the problem of finding a deep neural network (DNN) $\mathrm{F}_\theta : Y \to X$ such that $\mathrm{F}_\theta(y) \approx x$, where the neural network F_θ is parametrised by a parameter vector θ. In supervised learning the optimal parameter vector θ^* is learned from a set of ordered pairs $\mathbb{B} = \{(x_n, y_n)\}_{n=1}^N$ of ground truth images and the corresponding (noisy) measurement data by minimising a suitable loss

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \ell(\mathrm{F}_\theta(y_n), x_n), \tag{1.1}$$

where $\ell(\mathrm{F}_\theta(y_n), x_n)$ measures the discrepancy between the network prediction $\mathrm{F}_\theta(y_n)$ and the corresponding ground truth image x_n, and is often taken to be the mean squared error. Supervised learning has been established as a powerful tool to improve reconstruction quality and speed, rapidly becoming a workhorse in several imaging applications [56].

In order to deliver competitive performance, supervised learning may require many ordered pairs (x_n, y_n), $n = 1, \ldots, N$, which are unfortunately often not available in medical imaging applications since clean ground truth images are either too costly or impossible to collect. Meanwhile, reconstruction methods learned in a scarce-data regime often fail to generalise on instances which belong to a different data distribution [12, 48]. Moreover, even small deviations from the training data distribution can potentially lead to severe reconstruction artefacts (i.e., supervised models can exhibit poor performance even for a small distributional shift). This behaviour is further exacerbated by the presence of structural changes such as rare pathologies; thereby significantly degrading the performance of supervisedly learned reconstruction methods [5]. To make matters worse, such forms of deviation from the training data distribution are ubiquitous in medical imaging, owing to factors such as the change in acquisition protocols. For example, in magnetic resonance imaging (MRI), these factors include echo time, repetition time, flip angle, and inherent hardware variations in the used scanner [31]; in computed tomography (CT), they include the choice of view angles, acquisition time per view, and source-target separation.

Therefore, there is an imperative need to develop learned image reconstruction techniques that do not rely on a large amount of high-quality ordered pairs of training data. In a recent review [56], this issue has been identified as one of the key challenges in the next generation of learned reconstruction techniques. To address this outstanding challenge, in this work we develop a novel unsupervised knowledge-transfer (UKT) strategy to transfer acquired "reconstructive knowledge" across different datasets using the Bayesian framework. It comprises of two phases. The first phase is supervised and is tasked with pretraining a DNN reconstructor on data pairs of ground truth images and corresponding measurement data (which can be either simulated or experimentally collected). The goal of this step is to capture inductive biases of the given reconstruction task using simulated or experimental data. The second phase is unsupervised. It fine-tunes the reconstructor learned in the first phase on clinically-realistic measurement data, using a novel regularised Bayesian loss. This fine-tunes the network to the target reconstruction task while maintaining the prior knowledge learned in the first step. Note that unlike supervised or semi-supervised learning, the proposed framework does not assume any ground truth data from the target domain, and hence it is an unsupervised learning method. Extensive numerical experiments with low-dose and sparse-view CT on two datasets, i.e., FoamFanB [47] and LoDoFanB [37], indicate that the proposed approach is competitive with state-of-the-art methods both quantitatively and qualitatively, and that test-time adaptation can significantly boost performance.

In summary, the development of an unsupervised knowledge-transfer framework for learned

image reconstruction, and its validation on clinically realistic simulated measurement data, represent the main contributions of this work. To the best of our knowledge, this is the first work to propose Bayesian unsupervised knowledge-transfer for test-time adaptation of a learned image reconstruction method. Furthermore, the use of the Bayesian framework allows capturing predictive uncertainty of the obtained reconstructions. Our framework has the following distinct features: (i) adapting to unseen measurement data without the need for ground truth images; (ii) leveraging reconstructive properties learned in the supervised phase for effective feature representation; (iii) providing uncertainty estimates on the reconstructed images. These features make the framework very attractive for performing learned reconstruction without ordered pairs from the target domain, as confirmed by the extensive numerical experiments in Section 4. The Bayesian nature of the framework is noteworthy in the emerging field of scalable uncertainty quantification for image reconstruction, where the heavy computational cost is often deemed as one of the major hurdles [9]. In contrast, the approach presented in this work is highly scalable, by building upon recent advances in variational inference [23], and hence holds significant potential for medical image reconstruction.

1.1 Related Work

The lack of (a sufficient amount of) reference training data has only recently motivated the development of deep learning-based image reconstruction approaches that do not require ground truth images. We identify two main groups of current learned approaches: test-time adaptation, and unsupervised approaches.

Test-time adaptation focuses on learning under differing training and testing distributions. It often consists of fine-tuning a pretrained DNN for a single datum at a time, or for a small set of test instances. In [26, 18] this paradigm is used for MRI reconstruction, where reconstructive properties acquired by a network that has been pretrained on a task for which a large dataset is available, are transferred to a different task where the supervised data is scarce (but still available). The proposed approach extends the aforementioned work from the supervised target reconstruction task to an unsupervised one. In the context of object recognition, Sun et al. [51] propose to adapt only a part of a convolutional neural network (CNN) according to a self-supervised loss defined on the given test image to address distributional shift. The model is then trained via multi-task learning, where shared features are learned jointly over supervised and self-supervised data. Gilton et al. [25] adapt a pretrained image reconstruction network to reconstruct images from a perturbed forward model using only a small collection of measurements, by enforcing the data fidelity while penalising the deviation of the network parameters from the parameters of the pretrained model. Conceptually speaking, our study is complementary to these studies. The proposed approach can be interpreted as conducting unsupervised test-time adaptation for distributional shift of the image data, but within a Bayesian framework. Furthermore, the use of the Bayesian framework brings several distinct advantages: (i) it allows deriving the training loss in a principled manner; (ii) it can boost reconstructive performance; (iii) it simultaneously delivers the predictive uncertainty information associated with the reconstructions.

Meanwhile, deep image prior (DIP) is a representative unsupervised image reconstruction method, which achieves sample-specific performance using DNNs to describe the mapping from latent variables to high-quality images [53]. During inference the network architecture acts as a regulariser for reconstruction [21, 8]. Similarly, Zhang et al. [60] use a U-Net model as the reconstruction network and propose to adapt the model through backpropagation by updating the parameters of a pretrained U-Net under the guidance of data fidelity for each individual test data y, with no supervision, and showcase the approach on under-sampled MRI reconstruction. Despite strong performance, it suffers from slow convergence (often requiring thousands of iterations), and

the need for a well-timed early stopping, otherwise the network may overfit to the noise in the data. The latter issue has motivated the use of an additional stabiliser [8].

Test time adaptation and DIP represent only two approaches that are most closely related to the present work. In recent years, there have been significant advances in unsupervised biomedical imaging reconstruction techniques and we refer interested readers to a recent review [4] on other approaches and references therein, which discusses many promising unsupervised methods.

The rest of the paper is structured as follows. In Section 2 we describe the setting and discuss deep unrolled methods for image reconstruction and Bayesian DL. In Section 3 we develop the proposed two-phase UKT paradigm. In Section 4 we present experimental results for low-dose and sparse-view CT, including several supervised and unsupervised benchmarks, and discuss the results obtained with the two-phase learning paradigm. In Section 5 we add some concluding remarks.

2 Preliminaries

In this section we describe the fundamentals of how unrolled networks are used for image reconstruction. We then describe the Bayesian approach for DNNs, based on which we shall develop the proposed unsupervised knowledge-transfer strategy.

2.1 Unrolled Networks

Unrolling is a popular paradigm for constructing a network F_θ for image reconstruction. The idea is to mimic well-established iterative optimisation algorithms, e.g., (proximal) gradient descent, alternating direction method of multipliers, and primal-dual hybrid gradient method. Namely, unrolled methods use an iterative procedure to reconstruct an image x from the measurement y by combining analytical model components (e.g., the forward map A and its adjoint $A^\top$) with data-driven components that are parameterised by the network parameters θ and learned from the training data. The unrolled nature of the network allows seamlessly integrating the underlying physics of the data acquisition process into the design of the network F_θ, which can enable the development of high-performance reconstructors from reasonably sized training datasets [41]. More specifically, given an initial guess x_0 (e.g., the Filtered Back-Projection (FBP) in CT reconstruction), we recursively compute iterates

$$x_k = \mathrm{F}_{\theta_k}\left(x_{k-1}, \nabla\mathcal{D}_{k-1}\right), \quad k = 1, \ldots, K,$$

with

$$\nabla\mathcal{D}_{k-1} := \nabla\frac{1}{2}\|Ax_{k-1} - y\|^2 = A^\top\left(Ax_{k-1} - y\right),$$

being the gradient of the data fidelity term, where $K \geqslant 1$ is the total number of iterations, F_{θ_k} is the sub-network used at the k-th iteration, and θ_k is the corresponding weight vector. The overall iterative process can then be written as

$$x_K = \mathrm{F}_\theta\left(x_0, \nabla\mathcal{D}_0\right),$$

where x_K is the final reconstruction, and F_θ is the overall network, with parameters $\theta = (\theta_1, \ldots, \theta_K)$, constructed as a concatenation of sub-networks $\mathrm{F}_{\theta_1}, \ldots, \mathrm{F}_{\theta_K}$. In practice, the parameters θ_k of each sub-network F_{θ_k} can be shared across different blocks (i.e., $\theta_1 = \cdots = \theta_K$), a procedure known as weight-tying or weight-sharing. This allows to reduce the total number of trainable parameters, so as to facilitate the training process. By slightly abusing the notation, we denote the shared parameter by θ. In this work, we only consider the case of weights shared across the blocks, but the proposed framework extends straightforwardly to the general case.

2.2 Bayesian Neural Networks

We briefly describe Bayesian neural networks (BNNs), in which network parameters θ are treated as random variables and are learned through a Bayesian framework so as to facilitate uncertainty quantification of the network prediction. Bayesian learning provides a principled yet flexible framework for knowledge integration, and allows quantifying predictive uncertainties associated with a particular point estimate [23, 9]. Bayesian learning is ideally suited for deriving a proper training loss for combining the knowledge across different "domains", to which the framework proposed in Section 3 belongs. Nonetheless, the use of BNNs for medical imaging is still not widespread due to the associated computational challenge.

In a BNN, by placing a prior distribution $p(\theta)$ over the network parameters θ (which is commonly taken to be the standard Gaussian distribution), and by combining it with a likelihood function $p(\mathbb{B}|\theta)$ of the data $\mathbb{B}$ using Bayes' formula, we obtain a posterior distribution $p(\theta|\mathbb{B})$ over the parameters θ, given the data $\mathbb{B}$

$$p(\theta|\mathbb{B}) = Z^{-1} p(\mathbb{B}|\theta) p(\theta),$$

where $Z = \int p(\mathbb{B}|\theta)p(\theta)\mathrm{d}\theta$ is the normalising constant. The likelihood $p(\mathbb{B}|\theta)$ is fully specified upon properly modelling the data noise statistics and the data generation process (e.g., forward operator A). The posterior distribution $p(\theta|\mathbb{B})$ represents the complete Bayesian solution of the learning task.

The posterior $p(\theta|\mathbb{B})$ is often computationally intractable, since the computation of the normalising constant Z involves a high-dimensional integral. To circumvent this computational issue, we adopt variational inference (VI) [30], which employs the Kullback-Leibler (KL) divergence to construct an approximating distribution $q(\theta)$. Recall that the KL divergence $\mathrm{KL}[q(\theta)\|p(\theta)]$ between q and p is defined by

$$\mathrm{KL}\left[q(\theta)\|p(\theta)\right] = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} \mathrm{d}\theta.$$

VI looks for an easy-to-compute approximate posterior distribution q_ψ parametrised by variational parameters ψ. The approximation $q_\psi(\theta)$ is most commonly taken from a variational family consisting of products of independent Gaussians

$$\mathcal{Q} := \left\{ q_\psi(\theta) = \prod_{j=1}^{D} \mathcal{N}(\theta_j; \mu_j, \sigma_j^2) \Big| \psi \in \left(\mathbb{R} \times \mathbb{R}_{\geqslant 0}\right)^D \right\},$$

where the notation $\mathcal{N}(\theta_j; \mu_j, \sigma_j^2)$ denotes a Gaussian distribution with mean μ_j and variance σ_j^2, $\psi = ((\mu_j, \sigma_j^2))_{j=1}^D$ are the variational parameters, and D is the total number of parameters in F_θ. In the literature this is commonly known as the mean field approximation. VI constructs an approximation $q_{\psi*}(\theta)$ within the family $\mathcal{Q}$ by

$$q_{\psi*}(\theta) \in \underset{q_\psi(\theta)\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left[q_\psi(\theta)\|p(\theta|\mathbb{B})\right]. \tag{2.1}$$

Given a learned approximate posterior $q_{\psi*}(\theta)$, the predictive distribution $q_{\psi*}(x|y_q)$ of the target image x for a new query measurement y_q is given by

$$q_{\psi*}(x|y_q) = \int p(x|y_q, \theta) q_{\psi*}(\theta) \mathrm{d}\theta.$$

A point estimate of the image x can then be obtained via Monte Carlo (MC) sampling as

$$\mathbb{E}[x] = \int x q_{\psi*}(x|y_q) \mathrm{d}x \approx \frac{1}{T} \sum_{t=1}^{T} \mathrm{F}_{\theta^t}(x_{q,0}, \nabla \mathcal{D}_{q,0}),$$

with T Monte Carlo samples θ^t, with $t = 1, \ldots, T$, distributed according to $q_{\psi*}(\theta)$.

When the network densities are shared across the iterates, we have

$$\mathrm{F}_{\theta \sim q_\psi^{\otimes K}(\theta)} := \mathrm{F}_{\theta_K \sim q_\psi(\theta)} \circ \cdots \circ \mathrm{F}_{\theta_1 \sim q_\psi(\theta)},$$

with the superscript $\otimes K$ denoting the K-fold product, and the overall iterative process reads

$$x_K = \mathrm{F}_{\theta \sim q_\psi^{\otimes K}(\theta)} \left(x_0, \nabla \mathcal{D}_0 \right).$$

Note that the standard mean field approximation doubles the number of trainable parameters, which brings significant computational challenges. In practice, the training of fully Bayesian models is often non-trivial, and the performance of the resulting network is often inferior to non-Bayesian networks [45]. BNNs are thus still not widely used in learned image reconstruction [9]. To make our approach competitive with non-Bayesian methods, while retaining the benefits of Bayesian modelling, we can adopt the strategy of *being Bayesian only a little bit* [11, 19]. That is, we use VI only on a subset of the parameters θ, and use point estimates for the remaining parameters (or equivalently, a Dirac distribution). This can reduce the number of trainable parameters, and hence greatly facilitate the training process, while maintaining the Bayesian nature of the learning algorithm.

Remark 2.1. *Apart from VI there are other approximate inference schemes, such as MC dropout [24] and Laplace approximation [40, 19]. MC dropout has been widely used for modelling uncertainty, and has also found application in the medical imaging community (e.g., segmentation [52]), due to its computational efficiency and easy implementation, but its approximation accuracy tends to be inferior to VI. For example, MC dropout tends to severely underestimate predictive uncertainty [35]. Laplace approximation [40] has started to attract renewed interest, but has not been explored within medical image reconstruction so far, since the computational cost of approximating the Hessian of the loss with respect to the network parameters θ is often prohibitively high in the context of image reconstruction and a scalable and yet accurate approximation of the Hessian is still under development.*

Remark 2.2. *In light of the decoder-encoder structure of the U-Net that is used below (cf. Fig. 1(b) for a schematic illustration), the idea of "being Bayesian a little bit" resembles a hybridisation of an autoencoder [55] and a variational autoencoder [34], which are used for the decoder and encoder parts, respectively. However, there is a major difference between the two approaches: the formulation we employ for image reconstruction is conditional on the measurement, whereas the standard autoencoder and variational autoencoder formulations are unsupervised in that they access only samples of images.*

3 Two-Phase Learning

In this section we describe our novel two-phase UKT strategy aimed at addressing the challenges associated with the lack of sufficient supervised training data in the target reconstruction task. We systematically develop the learning strategy within a Bayesian framework with a sub-network F_{θ_k} being a downscaled version of a residual U-Net [50] (cf. Fig. 1(b) for a schematic illustration), which is a popular choice in learned image reconstruction [29], and will be used in the experiments in Section 4. The network adopts a multi-scale encoder-decoder structure consisting of an encoding component and a decoding component, whose parameters are denoted respectively by $\theta_\mathrm{e} \in \mathbb{R}^{D_\mathrm{e}}$ and $\theta_\mathrm{d} \in \mathbb{R}^{D_\mathrm{d}}$, and $\theta = (\theta_\mathrm{e}, \theta_\mathrm{d})$. In the derivation of the proposed UKT framework below, we use VI

only on the network parameters θ_{e} of the encoder component, which can be interpreted as choosing an approximate posterior $q_{\psi*}(\theta_{\mathrm{e}})$ for the encoder $p(\theta_{\mathrm{e}}|\mathbb{B}) \approx q_{\psi*}(\theta_{\mathrm{e}})$. The decoder parameters θ_{d} remain deterministic, and are treated as point-estimates. The adaptation of the UKT framework to other network architectures is straightforward.

Now we briefly describe the two phases of the proposed learning strategy. The first phase is supervised, and employs a given training dataset $\mathbb{B}^{\mathrm{s}} = \{(x_n^{\mathrm{s}}, y_n^{\mathrm{s}})\}_{n=1}^{N^{\mathrm{s}}}$ where each pair $(x_n^{\mathrm{s}}, y_n^{\mathrm{s}})$ consists of a ground truth image x_n^{s} and the corresponding (noisy) measurement datum y_n^{s}, which can be either simulated or experimentally collected (if available). The goal of this phase is to pretrain a reconstruction network F_θ by learning the (approximate) posterior distribution $q_{\psi*}(\theta_{\mathrm{e}})$ for the parameters θ_{e} of the encoder, and the optimal deterministic parameters θ_{d}^* of the decoder, in order to assist the unsupervised phase. Specifically, we aim to achieve two objectives: (i) identify a sensible region for the network parameters; (ii) learn robust representations that are not prone to overfitting. Ideally, to facilitate the reconstruction quality this phase should mimic the setting of the target reconstruction task as close as possible in terms of the geometry of image acquisition (e.g., size of images and distribution of image features), and the noise statistics (e.g., distribution and noise level). This phase would allow learning adequate inductive biases and task-specific priors so as to enable successful subsequent unsupervised learned image reconstruction.

The second phase is unsupervised, and has access to a dataset $\mathbb{B}^{\mathrm{u}} = \{y_n^{\mathrm{u}}\}_{n=1}^{N^{\mathrm{u}}}$ which consists of only a few measurements (e.g., clinically-realistic CT sinograms), but with no access to corresponding ground truth images. Moreover, the distribution of the measurement data in $\mathbb{B}^{\mathrm{u}}$ may differ significantly from that in $\mathbb{B}^{\mathrm{s}}$. The aim of this phase is to fine-tune the parameters θ of the reconstruction network F_θ so that it performs well on the data $\mathbb{B}^{\mathrm{u}}$ from the target domain. This is achieved by initialising the parameters $(\psi, \theta_{\mathrm{d}})$ of the reconstruction network F_θ to the optimal configuration $(\psi^*, \theta_{\mathrm{d}}^*)$ found in the first phase, and then minimising a novel loss function, which we shall derive below in the Bayesian framework. Through this phase we address the need for adaptivity to the target reconstruction task due to a potential distributional shift of the data and effectively use the inductive bias to assist the reconstruction of the target task.

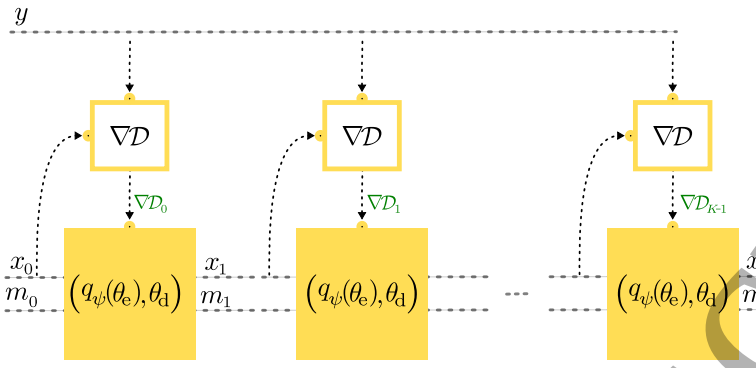
3.1 Pretraining via Supervised Learning

In this first phase, we have access to a training dataset $\mathbb{B}^{\mathrm{s}} = \{(x_n^{\mathrm{s}}, y_n^{\mathrm{s}})\}_{n=1}^{N^{\mathrm{s}}}$ of ordered pairs (which can be either simulated or experimentally collected), and we employ the Bayesian framework described in Section 2.2 to find the optimal distribution $q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})$ for the parameters θ_{e} of the encoder, which approximates the true posterior $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}})$ and the optimal decoder parameters θ_{d}^*. To construct the posterior $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}})$, we first set the prior $p(\theta_{\mathrm{e}})$ over the encoder parameters θ_{e} to the standard Gaussian $\mathcal{N}(\theta_{\mathrm{e}}; 0, I)$, which is a standard practice in the Bayesian DL community. Following the heteroscedastic noise model [43], the likelihood $p(x_n^{\mathrm{s}}|y_n^{\mathrm{s}}, \theta)$ is set to

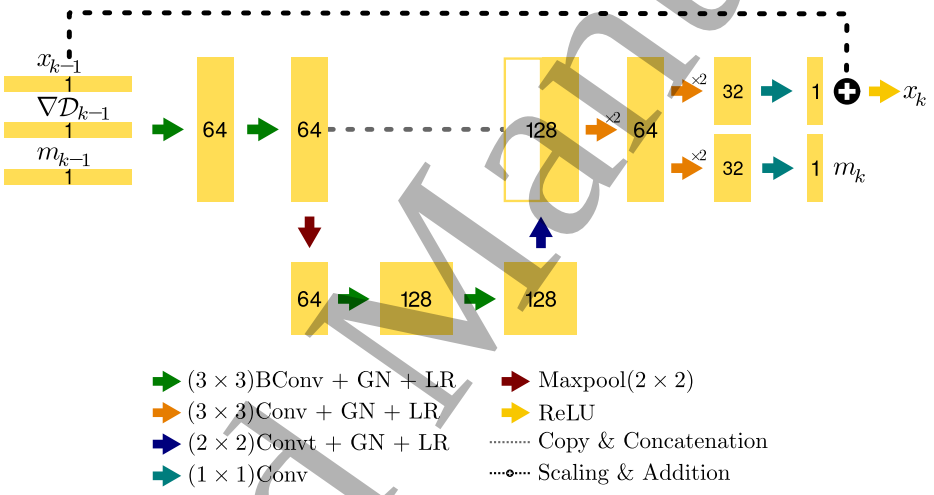
$$p(x_n^{\mathrm{s}}|y_n^{\mathrm{s}}, \theta) = \mathcal{N}\big(y_n^{\mathrm{s}}; \mathrm{F}_\theta^{\mu}(x_{n,0}^{\mathrm{s}}), \hat{\Sigma}_n\big), \tag{3.1}$$

with $\mathrm{F}_\theta^{\mu}(x_{n,0}^{\mathrm{s}}) = \mathrm{F}_\theta^{\mu}(x_{n,0}, \nabla\mathcal{D}_{n,0}, m_{n,0})$ and $\hat{\Sigma}_n = \mathrm{diag}\left(\mathrm{F}_\theta^{\sigma}(x_{n,0}^{\mathrm{s}})\right)$. Analogously, note that $\mathrm{F}_\theta^{\sigma}(x_{n,0}^{\mathrm{s}}) = \mathrm{F}_\theta^{\sigma}(x_{n,0}, \nabla\mathcal{D}_{n,0}, m_{n,0})$. Note that the network F_θ has two outputs: the mean F_θ^{μ}, and the variance $\mathrm{F}_\theta^{\sigma}$. Here $x_{n,0}^{\mathrm{s}}$ denotes the initial guess used by the learned reconstruction method for the n-th training pair $(x_n^{\mathrm{s}}, y_n^{\mathrm{s}})$. For example, in CT reconstruction, we customarily take $x_{n,0}^{\mathrm{s}}$ to be the FBP. We refer readers to Fig. 1(a) for a schematic illustration, where x_k and m_k denote the mean and variance estimates at the k-th iteration, respectively. Up to an additive constant independent of the arguments, we can write

$$\log p(x_n^{\mathrm{s}}|y_n^{\mathrm{s}}, \theta) = -\frac{1}{2}\|\hat{\Sigma}_n^{-\frac{1}{2}}(x_n^{\mathrm{s}} - \mathrm{F}_\theta^{\mu}(x_{n,0}^{\mathrm{s}}))\|^2 - \frac{1}{2}\log(\det(\hat{\Sigma}_n)).$$



(a) Diagram of the reconstructive pipeline. Each sub-network F_{θ_k} receives as input x_{k-1}, $\nabla\mathcal{D}_{k-1}$, m_{k-1}, and outputs x_k and m_k. The gradient of the data fidelity term $\nabla\mathcal{D}_k$ (colour-coded in green) is not an output of the sub-network, and is instead computed using the refined estimate x_k (and the forward and adjoint operators), and then passed as input to the subsequent sub-network together with x_k and m_k.



(b) The architecture of F_{θ_k} is a downscaled version of a residual U-Net with two scales of 64 and 128 channels. Each box corresponds to a multi-channel feature map, with the number of channels indicated inside. The inputs x_{k-1}, $\nabla\mathcal{D}_{k-1}$, and m_{k-1} go through a contractive path of repeated applications of two Bayesian convolutional layers (BConv), each followed by group normalisation (GN) [58] and leaky ReLU (LR), with a maxpool operation in between. Maxpool halves the feature channels resulting in a coarser scale. The expansive path consists of a transposed convolution (Convt) with stride length 2, which doubles the number of feature channels. The resulting feature map is then concatenated with the feature map from the contracting path, which is further processed through a convolutional pipeline. The architecture then bifurcates into two identical convolutional pipelines with feature maps reduced to a single channel. The output of the first pipeline is added as a residual update to the initial input iterate, and projected onto the positive set to produce a new iterate x_k. The second output is the intermediate estimate of the variance. At the final iteration, we have $x_K \to \mathrm{F}_\theta^\mu$, and $m_K \to \mathrm{F}_\theta^\sigma$. The arrows denote different operations, and the ones which have a symbol "×2" next to the arrow imply that the operation in question is repeated twice.

Figure 1: (a) Schematic illustration of the overall iterative reconstructive process, and (b) the architecture of each sub-network.

The minimisation of KL divergence in (2.1) can be recast as the minimisation of the following

loss over the admissible set $\mathbb{R}^{D_\mathrm{d}} \times \mathcal{Q}$

$$\mathcal{L}^\mathrm{s}(\theta_\mathrm{d}, q_\psi(\theta_\mathrm{e})) = -\frac{1}{N^\mathrm{s}} \sum_{n=1}^{N^\mathrm{s}} \mathbb{E}_{q_\psi(\theta_\mathrm{e})} \left[\log p\left(x_n^\mathrm{s} | y_n^\mathrm{s}, \theta\right)\right] + \beta \mathrm{KL}\left[q_\psi(\theta_\mathrm{e}) \| p(\theta_\mathrm{e})\right],$$

where $\beta > 0$ is a regularisation parameter. This loss coincides with the negative value of the Evidence Lower BOund (ELBO) in VI (when $\beta = 1$). Upon expanding the terms, fixing the prior at $p(\theta_\mathrm{e}) = \mathcal{N}(\theta_\mathrm{e}; 0, I)$, and ignoring additive constants independent of θ_d and $q_\psi(\theta_\mathrm{e})$, we can rewrite the loss as (recall that D_e denotes the dimensionality of the encoder parameter θ_e)

$$\begin{aligned}\mathcal{L}^\mathrm{s}(\theta_\mathrm{d}, q_\psi(\theta_\mathrm{e})) = \frac{1}{N^\mathrm{s}} \sum_{n=1}^{N^\mathrm{s}} \mathbb{E}_{q_\psi(\theta_\mathrm{e})} \left[\frac{1}{2} \|\hat{\Sigma}_n^{-\frac{1}{2}} (x_n^\mathrm{s} - \mathrm{F}_\theta^\mu(x_{n,0}^\mathrm{s}))\|^2 + \frac{1}{2} \log(\det(\hat{\Sigma}_n))\right] \\ + \beta \sum_{j=1}^{D_\mathrm{e}} \left[-\log \sigma_j + \frac{1}{2}(\sigma_j^2 + \mu_j^2)\right], \end{aligned} \tag{3.2}$$

where the vector $\psi = ((\mu_j, \sigma_j^2))_{j=1}^{D_\mathrm{e}}$ refers to variational parameters of the approximate distribution $q_\psi(\theta_\mathrm{e})$, where μ_j and σ_j^2 are respectively the mean and the variance of the j-th component of the encoder parameters θ_e. Note that the term $\mathrm{KL}\left[q_\psi(\theta_\mathrm{e}) \| p(\theta_\mathrm{e})\right]$ affects only the encoder parameters θ_e, whereas the decoder parameters θ_d are treated deterministically (without any explicit penalty). In order to minimise the loss $\mathcal{L}^\mathrm{s}$ with respect to the variational parameters ψ, we need to compute the gradient $\nabla_\psi \mathcal{L}^\mathrm{s}$ of the loss $\mathcal{L}^\mathrm{s}$ with respect to ψ. This can be done efficiently using the local reparametrisation trick [33], which employs a deterministic dependence of the ELBO with respect to ψ.

The combination of the unrolled network with Bayesian neural networks allows quantifying the uncertainty over the reconstructed image by unrolling methods, and we have termed the resulting approach (when trained in a greedy manner) as Bayesian deep gradient descent (BDGD) in prior works [11, 10]. BDGD provides natural means to quantify not only the predictive uncertainty associated with a given reconstruction, but also to disentangle the sources from which the predictive uncertainty arises. Uncertainty is typically categorised into aleatoric and epistemic uncertainties [32, 9]. Epistemic uncertainty arises from the uncertainty over the network parameters, and is captured by the posterior $q_\psi(\theta_\mathrm{e})$ [13, 32]. Aleatoric uncertainty is instead caused by the randomness in the data acquisition process. To account for this, in the loss (3.1) we employ a heteroscedastic noise model [43], which sets the likelihood $p(x_n^\mathrm{s} | y_n^\mathrm{s}, \theta)$ to be a Gaussian distribution, with both its mean F_θ^μ and variance F_θ^σ predicted by the network F_θ. Accordingly, we adjust the network architecture by bifurcating the decoder output. Namely, sub-network outputs $\mathrm{F}_{\theta_k}^\mu$ are used to update the estimate x_k, whilst the intermediate term m_k, which embodies a form of "information transmission", is given by $\mathrm{F}_{\theta_k}^\sigma$. At the final iteration m_K provides an estimate of the variance component of the likelihood; see Fig. 1(a) again for a schematic illustration on the overall workflow of the network F_θ.

Following [20], we can decompose the (entry-wise) predictive variance $\mathrm{Var}[x]$ into a sum of aleatoric ($\Delta_\mathrm{A}[y_q]$) and epistemic ($\Delta_\mathrm{E}[y_q]$) uncertainties using the law of total variance as follows

$$\mathrm{Var}[x] = \mathbb{E}_{q_{\psi*}(\theta_\mathrm{e})}[\mathrm{Var}(x|y_q, \theta)] + \mathrm{Var}_{q_{\psi*}(\theta_\mathrm{e})}[\mathbb{E}(x|y_q, \theta)] =: \Delta_\mathrm{A}[y_q] + \Delta_\mathrm{E}[y_q].$$

Upon denoting the initial guesses for the mean and the variance for a query data y_q by $x_{q,0}$ and $m_{q,0}$, respectively, and abbreviating $\mathrm{F}_{\theta^t}^\sigma(x_{q,0}, \nabla\mathcal{D}_{q,0}, m_{q,0})$ as $\mathrm{F}_{\theta^t}^\sigma(x_{q,0})$, and $\mathrm{F}_{\theta^t}^\mu(x_{q,0}, \nabla\mathcal{D}_{q,0}, m_{q,0})$ as

$\mathrm{F}^{\mu}_{\theta^t}(x_{q,0})$, we estimate $\Delta_{\mathrm{A}}[y_q]$ and $\Delta_{\mathrm{E}}[y_q]$ by $T \geqslant 1$ Monte Carlo samples $\{\theta^t_{\mathrm{e}}\}_{t=1}^T \sim q^{\otimes K}_{\psi *}(\theta_{\mathrm{e}})$ as

$$\Delta_{\mathrm{A}}[y_q] \approx \frac{1}{T}\sum_{t=1}^{T} \mathrm{F}^{\sigma}_{\theta^t}(x_{q,0}) \quad \text{and} \quad \Delta_{\mathrm{E}}[y_q] \approx \frac{1}{T}\sum_{t=1}^{T} \mathrm{F}^{\mu}_{\theta^t}(x_{q,0})^2 - \Big(\frac{1}{T}\sum_{t=1}^{T} \mathrm{F}^{\mu}_{\theta^t}(x_{q,0})\Big)^2,$$

where all the operations on vectors are understood entry-wise.

Remark 3.1. *There are at least two alternative loss functions that can be derived from the Bayesian loss (3.2). The first option is to set the parameters θ as fully deterministic, which gives rise to the following non-Bayesian loss*

$$\mathcal{L}^{\mathrm{s}}(\theta) = \frac{1}{N^{\mathrm{s}}}\sum_{n=1}^{N^{\mathrm{s}}} \big[\frac{1}{2}\|\hat{\Sigma}_n^{-\frac{1}{2}}(x^{\mathrm{s}}_n - \mathrm{F}^{\mu}_{\theta}(x^{\mathrm{s}}_{n,0}))\|^2 + \frac{1}{2}\log(\det(\hat{\Sigma}_n))\big] + \frac{\beta}{2}\|\theta_{\mathrm{e}}\|^2.$$

Note that this loss does not penalise the decoder parameters θ_{d}, as in the Bayesian formulation, whereas it penalises the encoder parameters θ_{e} by the standard weight decay, which corresponds directly to the standard Gaussian prior $p(\theta_{\mathrm{e}}) = \mathcal{N}(\theta_{\mathrm{e}}; 0, I)$ on the encoder parameters θ_{e}. The presence of the log-determinant $\log(\det(\hat{\Sigma}_n))$ is due to heteroscedastic noise modelling [43], and accordingly the network F_{θ} has two outputs, one for the mean and the other for the variance. The second option is to fix the output noise variance as $\hat{\Sigma}_n = \sigma^2 I$ (with known σ) in the heteroscedastic noise modelling. This leads to the following loss

$$\mathcal{L}^{\mathrm{s}}(\theta) = \frac{1}{N^{\mathrm{s}}}\sum_{n=1}^{N^{\mathrm{s}}} \frac{1}{2\sigma^2}\|x^{\mathrm{s}}_n - \mathrm{F}^{\mu}_{\theta}(x^{\mathrm{s}}_{n,0})\|^2 + \frac{\beta}{2}\|\theta_{\mathrm{e}}\|^2.$$

This is essentially identical to the loss in (1.1) (modulo weight decay), which is arguably the most popular loss for obtaining supervised end-to-end DL-based image reconstruction algorithms.

3.2 Unsupervised Knowledge-Transfer

In the second phase we use the Bayesian framework to integrate the knowledge learned in the first phase to new imaging data for which we don't have access to paired training data but only to noisy observations. Note that the knowledge of the trained network (on the supervised data $\mathbb{B}^{\mathrm{s}}$) is encoded indirectly in the posterior distribution $q^{\mathrm{s}}_{\psi *}(\theta_{\mathrm{e}})$ and in the optimal parameters θ^*_{d}. The goal of the second phase is to approximate the true posterior $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$, and to find the updated optimal decoder parameters θ^*_{d} given the measurement data $\mathbb{B}^{\mathrm{u}}$ and the supervised data $\mathbb{B}^{\mathrm{s}}$ from the first phase. This can be achieved as follows. By Bayes' formula, the posterior distribution $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$ is given by

$$p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}}) = (Z^{\mathrm{u}})^{-1} p(\mathbb{B}^{\mathrm{u}}|\theta_{\mathrm{e}}) p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}).$$

Here $p(\mathbb{B}^{\mathrm{u}}|\theta_{\mathrm{e}})$ is the likelihood at test-time (i.e., the likelihood of the measurement data $\mathbb{B}^{\mathrm{u}}$ from the target reconstruction task), and the normalising constant $Z^{\mathrm{u}} = \int p(\mathbb{B}^{\mathrm{u}}|\theta_{\mathrm{e}}) p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}) \mathrm{d}\theta_{\mathrm{e}}$ is the marginal likelihood of the total observed data $(\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$. We approximate the posterior $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}})$ (from the supervised phase) by the estimated optimal posterior $q^{\mathrm{s}}_{\psi *}(\theta_{\mathrm{e}})$, which is learned in the first phase, thus encapsulating the "proxy" knowledge we have acquired from the supervised dataset $\mathbb{B}^{\mathrm{s}}$. An approximation $q^{\mathrm{u}}_{\psi *}(\theta_{\mathrm{e}})$ to the true posterior $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$ for the combined data $(\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$ can then be obtained using VI as

$$(\theta_{\mathrm{d}}, q^{\mathrm{u}}_{\psi *}(\theta_{\mathrm{e}})) \in \operatorname*{argmin}_{\theta_{\mathrm{d}} \in \mathbb{R}^{D_{\mathrm{d}}}, q_{\psi}(\theta_{\mathrm{e}}) \in \mathcal{Q}} \mathcal{L}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_{\psi}(\theta_{\mathrm{e}})),$$

where the objective function is given by

$$\mathcal{L}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_\psi(\theta_{\mathrm{e}})) := \mathrm{KL}\left[q_\psi(\theta_{\mathrm{e}}) \| (Z^{\mathrm{u}})^{-1} p(\mathbb{B}^{\mathrm{u}}|\theta_{\mathrm{e}}) q^{\mathrm{s}}_{\psi*}(\theta_{\mathrm{e}})\right]. \tag{3.3}$$

The approximate posterior $q^{\mathrm{s}}_{\psi*}(\theta_{\mathrm{e}})$ over the supervised dataset $\mathbb{B}^{\mathrm{s}}$ is by construction used as a prior in the second phase. It remains to construct the likelihood $p(\mathbb{B}^{\mathrm{u}}|\theta_{\mathrm{e}})$ for the unsupervised dataset $\mathbb{B}^{\mathrm{u}}$. For any measurement datum $y^{\mathrm{u}} \in \mathbb{B}^{\mathrm{u}}$, the likelihood $p(y^{\mathrm{u}}|\theta_{\mathrm{e}})$ is set to

$$p(y^{\mathrm{u}}|\theta_{\mathrm{e}}) = \mathcal{N}(y^{\mathrm{u}}; A\mathrm{F}^{\mu}_{\theta}(x^{\mathrm{u}}_0), \sigma^2 I).$$

Upon letting $\bar{y}^{\mathrm{u}} = A\mathrm{F}^{\mu}_{\theta}(x^{\mathrm{u}}_0)$, we have

$$\log p(y^{\mathrm{u}}|\theta_{\mathrm{e}}) = -\frac{1}{2\sigma^2}\|\bar{y}^{\mathrm{u}} - y^{\mathrm{u}}\|^2 - \frac{m}{2}\log(2\pi\sigma^2).$$

Note that unlike in (3.2), this likelihood would exert no influence on the component $\mathrm{F}^{\sigma}_{\theta}$ of the network output F_θ (arising from the heteroscedastic modelling). To address this, we shall, inspired by the bias variance decomposition, replace the log-likelihood $\log p(y^{\mathrm{u}}|\theta_{\mathrm{e}})$ with a suitable modification. For $p(\tilde{x}^{\mathrm{u}}) = \mathcal{N}\big(x^{\mathrm{u}}; \mathrm{F}^{\mu}_{\theta}(x^{\mathrm{u}}_0), \hat{\Sigma}\big)$, using the standard bias-variance decomposition, we obtain

$$\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\|A\tilde{x}^{\mathrm{u}} - y^{\mathrm{u}}\|^2] = \|A\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\tilde{x}^{\mathrm{u}}] - y^{\mathrm{u}}\|^2 + \mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\|A\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\tilde{x}^{\mathrm{u}}] - A\tilde{x}^{\mathrm{u}}\|^2].$$

By the definition of $\bar{y}^{\mathrm{u}}$, the first term can be rewritten as $\|\bar{y}^{\mathrm{u}} - y^{\mathrm{u}}\|^2$. Meanwhile, for a random vector w with mean $\mathbb{E}[w] = 0$ and covariance $\mathrm{Cov}(w)$, we have

$$\mathbb{E}[\|w\|^2] = \mathbb{E}[w^\top w] = \mathrm{trace}(\mathrm{Cov}(w)).$$

Since $A\tilde{x}^{\mathrm{u}} - A\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\tilde{x}^{\mathrm{u}}]$ is a zero mean random vector with covariance $\mathrm{Cov}(w) = A\hat{\Sigma}A^\top$, we have

$$\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\|A\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}[\tilde{x}^{\mathrm{u}}] - A\tilde{x}^{\mathrm{u}}\|^2] = \mathrm{trace}(A\hat{\Sigma}A^\top).$$

Consequently,

$$\mathbb{E}_{p(\tilde{x}^{\mathrm{u}})}\|A\tilde{x}^{\mathrm{u}} - y^{\mathrm{u}}\|^2 = \|A\hat{x}^{\mathrm{u}} - y^{\mathrm{u}}\|^2 + \mathrm{trace}(A\hat{\Sigma}A^\top), \quad \text{with } \hat{x}^{\mathrm{u}} = \mathrm{F}^{\mu}_{\theta}(x^{\mathrm{u}}_0)$$

This will be used in the loss function in the modified log-likelihood. In practice, the term $\mathrm{trace}(A\hat{\Sigma}A^\top)$ can be approximated using randomised trace estimators (e.g., the Hutchinson's estimator [15]). The computation of the optimal variational parameters ψ^* and the optimal decoder parameter θ^*_{d} by minimising the negative value of the ELBO proceeds analogously to the supervised phase, but with the key changes outlined above.

In addition to enforcing data fidelity, we also include a regularisation term to the loss in (3.3),

$$\tilde{\mathcal{L}}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_\psi(\theta_{\mathrm{e}}) = \mathcal{L}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_\psi(\theta_{\mathrm{e}})) + \gamma\mathbb{E}_{q_\psi(\theta_{\mathrm{e}})}\left[\mathrm{TV}(\mathrm{F}^{\mu}_{\theta}(x^{\mathrm{u}}_0))\right],$$

where as a regulariser we take the total variation seminorm $\mathrm{TV}(u) = \|\nabla u\|_1$, and $\gamma > 0$ is the regularisation parameter. This incorporates prior knowledge over expected images by penalising unlikely or undesirable solutions. TV is widely used in image reconstruction, due to its edge-preserving properties [14], and has also been applied to learned reconstruction [8, 16]. Intuitively, without the TV term, optimising the loss is akin to minimising the fidelity, and thus the training process is prone to overfitting, especially when the neural network is over-parameterised, necessitating the use of early stopping (which also has a regularising effect). The numerical experiments indicate that incorporating this term can help stabilise the training process and lead to improved

reconstructions, which agrees with earlier observations [8, 16]. In summary, the loss at the second phase reads

$$\tilde{\mathcal{L}}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_\psi(\theta_{\mathrm{e}})) = -\mathbb{E}_{q_\psi(\theta_{\mathrm{e}})}\left[\log p\left(\mathbb{B}^{\mathrm{u}}|\theta_{\mathrm{e}}\right) - \gamma \mathrm{TV}(\mathrm{F}_\theta^\mu(x_0^{\mathrm{u}}))\right] + \beta \mathrm{KL}\left[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})\right],$$

which upon expansion, relabelling, and the aforementioned modifications, leads to the loss

$$\begin{aligned}\tilde{\mathcal{L}}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_\psi(\theta_{\mathrm{e}})) = \frac{1}{N^{\mathrm{u}}} \sum_{n=1}^{N^{\mathrm{u}}} \mathbb{E}_{q_\psi(\theta_{\mathrm{e}})}\big[\frac{1}{2}\|y_n^{\mathrm{u}} - A\mathrm{F}_\theta^\mu(x_{n,0}^{\mathrm{u}})\|^2 + \mathrm{trace}(A\hat{\Sigma}A^\top) + \gamma \mathrm{TV}(\mathrm{F}_\theta^\mu(x_{n,0}^{\mathrm{u}}))\big] \\ + \beta \mathrm{KL}\left[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})\right]. \qquad (3.4)\end{aligned}$$

Since $q_\psi(\theta_{\mathrm{e}})$ and $q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})$ are constructed as the products of independent Gaussians (i.e., mean field approximation), the term $\mathrm{KL}[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})]$ has a closed-form expression given by

$$\mathrm{KL}\left[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})\right] = \sum_{j=1}^{D_{\mathrm{e}}} \left[\log \frac{\sigma_j^{\mathrm{s}}}{\sigma_j} + \frac{\sigma_j^2 + (\mu_j - \mu_j^{\mathrm{s}})^2}{2(\sigma_j^{\mathrm{s}})^2} - \frac{1}{2}\right],$$

where $\psi = ((\mu_j, \sigma_j))_{j=1}^{D_{\mathrm{e}}}$ refers to variational parameters of the approximate distribution $q_\psi(\theta_{\mathrm{e}})$, where μ_j and σ_j are the mean and the variance of the j-th component of θ_{e}, and σ_j^{s} and μ_j^{s} are the optimal variational parameters learned in the first phase (and thus fixed during the second phase). Note that the loss in (3.4) represents only one possibility for unsupervised knowledge transfer, and there are alternatives. In the appendix, we derive an alternative training loss, by constructing the likelihood $p(y^{\mathrm{u}}|\theta_{\mathrm{e}})$ differently, which also allows interpreting the loss $\tilde{\mathcal{L}}^{\mathrm{u}}$ as an approximate Bayesian loss.

It is instructive to interpret the terms in the loss $\tilde{\mathcal{L}}^{\mathrm{u}}$ in the lens of more familiar variational regularisation [22, 28]. The first term in (3.2) enforces data fidelity, which encourages the learned network F_θ to be close to the right-inverse of A (i.e., the action of the forward map A on the output of $\mathrm{F}_\theta(x_0^{\mathrm{u}})$ is close to the measurement data y^{u}). The second term, $\mathrm{trace}(A\hat{\Sigma}A^\top)$, controls the growth of the variance component, and along with the first term arises naturally when performing approximate VI (with a Gaussian likelihood) on the posterior distribution $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$; see the appendix for further discussions. Note that this term does not appear if one considers only the usual maximum a posteriori (MAP) estimator to the posterior distribution $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$. The third term, the TV regulariser, plays a crucial role in stabilising the learning process [8]. The fourth term $\mathrm{KL}[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})]$ forces the posterior $q_\psi(\theta_{\mathrm{e}})$ to be close to the prior $q_{\psi*}^{\mathrm{s}}(\theta_{\mathrm{e}})$ of the unsupervised phase (which is the posterior obtained during the supervised phase). These properties together give rise to a highly flexible UKT paradigm: the adaptation can be done individually for each query image datum (which is natural for streaming data) or for a whole batch of measurement data. The regularisation parameters $\gamma > 0$ and $\beta > 0$ control the strength of the related penalty terms. In practice, it is important to choose the regularisation parameters β and γ suitably, as in any inverse technique. In our experiments β and γ are chosen on a validation set.

Remark 3.2. *The loss $\tilde{\mathcal{L}}^{\mathrm{u}}$ in (3.4) can be viewed as a generalisation of the more conventional non-Bayesian approaches for domain adaptation*

$$\mathcal{L}^{\mathrm{u}}(\theta) = \frac{1}{N^{\mathrm{u}}} \sum_{n=1}^{N^{\mathrm{u}}} \big[\frac{1}{2}\|y_n^{\mathrm{u}} - A\mathrm{F}_\theta(x_{n,0}^{\mathrm{u}})\|^2 + \gamma \mathrm{TV}(\mathrm{F}_\theta(x_{n,0}^{\mathrm{u}}))\big] + \frac{\beta}{2}\|\theta_{\mathrm{e}} - \theta_{\mathrm{e}}^{\mathrm{s}}\|^2, \qquad (3.5)$$

where $\theta_{\mathrm{e}}^{\mathrm{s}}$ is the optimal encoder network parameter learned at the supervised phase. This loss encourages the network output $\mathrm{F}_\theta(x_0^{\mathrm{u}})$ to be close to piece-wise constant, and meanwhile, the

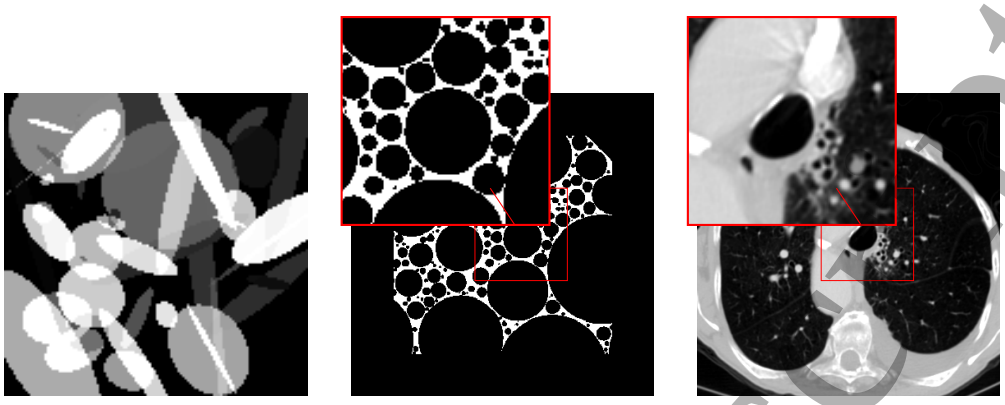


Figure 2: Representative ground truth images from Ellipses (left), FoamFanB (middle) and LoDoFanB (right) datasets. The window of the LoDoFanB dataset is set to a Hounsfield unit (HU) range $\approx$ [-1000, 400].

corresponding network should not deviate too much from $\theta_{\mathrm{e}}^{\mathrm{s}}$. *Due to the use of the Bayesian framework, the UKT loss* (3.4) *involves extra terms that are related to the variance of the parameters. Formally, the loss in* (3.5) *can also be obtained by considering the MAP estimator of the posterior distribution* $p(\theta_{\mathrm{e}}|\mathbb{B}^{\mathrm{s}}, \mathbb{B}^{\mathrm{u}})$, *concurring with the well-known connection between the MAP estimator and the posterior distribution. Nonetheless, even if considering the loss* (3.5) *alone, the Bayesian framework elucidates the standing assumptions for obtaining the loss. The loss in* (3.5) *is closely connected to the loss*

$$\mathcal{L}^{\mathrm{u}}(\theta) = \frac{1}{N^{\mathrm{u}}} \sum_{n=1}^{N^{\mathrm{u}}} \frac{1}{2} \|y_n^{\mathrm{u}} - A\mathrm{F}_\theta(x_{n,0}^{\mathrm{u}})\|^2 + \frac{\beta}{2} \|\theta - \theta^{\mathrm{s}}\|^2, \tag{3.6}$$

which also penalises the deviation of decoder parameters θ_{d} *from the pretrained parameters* $\theta_{\mathrm{d}}^{\mathrm{s}}$. *This is essentially the training loss employed in [25]. The other major difference between* (3.5) *and* (3.6) *lies in use of the TV penalty on the network output* $\mathrm{F}_\theta(x_0^{\mathrm{u}})$.

It is also worth noting that $\hat{\Sigma}$ *affects the loss* $\tilde{\mathcal{L}}^{\mathrm{u}}$ *in* (3.4) *only via the term* $\mathrm{trace}(A\hat{\Sigma}A^\top)$, *controlling the covariance of the estimate, but not via the data fidelity term. This is due to the simplified derivation of the loss. The genuine Bayesian loss in* (A.1) *to be derived in the appendix does incorporate* $\hat{\Sigma}$ *into noise covariance and consequently it does enter into the noise weighting matrix in the data fidelity. See the appendix for further discussions.*

4 Experiments and Results

In this section we present numerical experiments on simulated data to showcase the performance of the proposed UKT framework.

4.1 Experimental Settings

First we describe the experimental setting, including datasets, data generation, benchmark methods and training details.

Datasets. In the experiments we use three datasets: Ellipses, FoamFanB and LoDoFanB. The Ellipses dataset consists of random phantoms of overlapping ellipses, and is commonly used for inverse problems in imaging [2]. The intensity of the background is taken to be 0, the intensity

of each ellipse is taken randomly between 0.1 and 1, and the intensities are added up in regions where ellipses overlap. The phantoms are of size 128×128; see Fig. 2 for a representative phantom. The training set contains 32 000 pairs of phantoms and sinograms, while the test set consists of 128 pairs. This dataset is used for the training of all the methods that involve supervised training. The FoamFanB dataset is constructed using a cylindrical foam phantom containing 100 000 randomly-placed non-overlapping bubbles. The phantom consists of 100 slices of size 1024×1024 and is generated with the open-source `foam_ct_phantom` package [47]. Analytic projection images of the phantoms were also computed using the package. Each slice is then cropped into four 256×256 square sections, which are zero padded with 50 pixels in all four directions. Out of the resulting 400 slices, we randomly retain half; see Fig. 2 for a representative slice. The intensity of the pixels are either 0 or 0.5, which allows retaining finer structures in the images. The LoDoFanB dataset [37] is more medically realistic, and consists of 223 human chest CTs, in which the (original) slices from the LIDC/IDRI Database [6] have been pre-processed, and the resulting images are of size 362×362; see Fig. 2 for a representative slice. The FoamFanB and LoDoFanB datasets are used in the unsupervised phase, where we assume to know only the sinograms. The ground truth images are only used to evaluate the performance of all the studied methods, unless otherwise specified.

Data generation. For the forward map A, taken to be the Radon transform, we employ a two-dimensional fan-beam geometry with 600 angles for the low-dose CT setting, and 100 angles for the sparse-view CT setting. Source-to-axis and axis-to-detector distances are both set to 500 mm. For both datasets we apply a corruption process given by $\lambda \exp(-\mu Ax)$, where $\lambda \in \mathbb{R}^+$ is the mean number of photons per pixel and is fixed at 8000 (corresponding to low-dose CT), and $\mu \in \mathbb{R}^+$ is the attenuation coefficient, set to 0.2. We linearise the forward model by applying the transformation $-\log(\cdot)/\mu$. We can then use $\frac{1}{2}\|Ax - y\|^2$ as the data fidelity term since post-log measurements of low-dose CT approximately follow a Gaussian distribution [57, 38].

Benchmark methods. We compare the proposed BDGD+UKT approach with several unsupervised and supervised benchmarks. Unsupervised methods include FBP (using a Hann filter with a low-pass cut-off 0.6), (isotropic) TV regularisation, and DIP+TV [8]. Supervised methods include U-Net based post-processing (FBP+U-Net) [17], two learned iterative schemes: learned gradient descent (LGD) [2] and learned primal dual (LPD) [3], and BDGD (i.e., without UKT) [11, 10]. U-Net is widely used for post-processing (e.g., denoising and artefact removal), including FBPs [29], and our implementation follows [8] using a slightly down-scaled version of the standard U-Net. LGD and LPD are widely used, with the latter often seen as the standard benchmark for supervised deep tomographic reconstruction. BDGD exhibits competitive performance while being a Bayesian method [11, 10].

Training, hyper-parameters, and implementation. All supervised methods are first trained on the Ellipses dataset, and then tested on Ellipses, FoamFanB and LoDoFanB datasets separately. Unless otherwise stated, the learned models are not adapted to the FoamFanB and LoDoFanB datasets, but perform reconstruction directly on a given sinogram. The methods were implemented in PyTorch, and trained on a GeForce GTX 1080 Titan GPU. All operator-related components (e.g., forward operator, adjoint, and FBP) are implemented using the Operator Discretisation Library [1] with the `astra_gpu` backend [54].

For all the unsupervised methods (FBP, TV, DIP+TV), the hyperparameters (frequency scaling in FBP and regularisation parameter in TV and DIP+TV) are selected to maximise the PSNR on a subset of the dataset consisting of 5 images. DIP+TV adopts a U-Net architecture proposed in [8] (accessible in the DIVal library [36]): a 5-scale U-Net without skip connections for the Ellipses dataset, and a 6-scale U-Net with skip connections only at the last two scales for FoamFanB and LoDoFanB datasets. For both architectures the number of channels is set to 128 at every scale. In

Table 1: Reconstruction methods used in this work. For each method, the number of learnable parameters is indicated, as well as approximate runtime for both low-dose CT and sparse-view CT on the LoDoFanB dataset, is reported.

Methods		Parameters	Runtime
Unsupervised	FBP	1	38ms/7ms
	TV	1	20s/10s
	DIP+TV	$2.9 \cdot 10^6$	20min/18min
Supervised	FBP+U-Net	$6.1 \cdot 10^5$	5ms
	LGD	$1.3 \cdot 10^5$	89ms/34ms
	LPD	$2.5 \cdot 10^5$	180ms/55ms
	BDGD	$8.8 \cdot 10^5$	7s/6s
BDGD+UKT		$8.8 \cdot 10^5$	7s/6s

Table 1 we report the number of parameters used for the LoDoFanB dataset.

All learned reconstruction methods were trained until convergence on the Ellipses dataset. FBP+U-Net implements a down-sized U-Net architecture with 4 scales and skip connections at each scale. LGD is implemented as in [3], where the parameters of the reconstructor are not shared across the iterates, and we use $K = 10$ unrolled iterations. LPD follows the implementation in [3]. We train FBP+U-Net, LGD and LPD by minimising the loss in (1.1) using the Adam optimiser and a learning rate schedule according to cosine annealing [39]. All models are trained for 30 epochs. BDGD uses a multi-scale convolutional architecture (cf. Fig. 1), with $K = 3$ unrolled iterations. Furthermore, the UKT phase is initialised with parameters $(\psi*, \theta_{\mathrm{d}}^*)$, which are obtained at the end of the supervised training on the Ellipses dataset. For the FoamFanB dataset, the regularisation parameter γ is set to $5 \cdot 10^{-5}$ for the low-dose setting and to $1 \cdot 10^{-4}$ for the sparse-view setting. Analogously, for the LoDoFanB dataset, the regularisation parameter γ is set to $1 \cdot 10^{-4}$ for the low-dose setting and to $5 \cdot 10^{-4}$ for the sparse-view setting. On both datasets, β is set to $1 \cdot 10^{-4}$ for both settings. $T = 10$ Monte Carlo samples are used to reconstruct the point estimate, and to compute the associated uncertainty estimates. A Pytorch implementation of the proposed approach is publicly available at https://github.com/rb876/unsupervised_knowledge_transfer to reproduce the numerical experiments.

4.2 Experimental Results

In Table 2 we report PSNR and SSIM values for the studied datasets. We observe that unsupervised methods give higher PSNR and SSIM values on FoamFanB and LoDoFanB datasets than on the Ellipses dataset, with FBP on FoamFanB being the exception. The converse is true for supervised methods. Moreover, TV and DIP+TV outperform supervised reconstruction methods in both low-dose and sparse-view CT settings for FoamFanB and LoDoFanB datasets. The results for BDGD+UKT and BDGD indicate that adapting the parameters on the given dataset allows achieving a noticeable improvement in reconstruction quality in both low-dose and sparse-view CT settings. Note also that BDGD+UKT outperforms all supervised reconstruction methods, while performing on par with DIP+TV (but the corresponding computation time is only a small fraction of that for the latter). This last observation is not surprising, since the test data (FoamFanB and LoDoFanB) are distributed differently from the synthetic training data (Ellipses). As a result, the performance of supervised reconstruction methods deteriorates significantly.

Table 1 reports also the approximate runtime for all the methods under consideration. All

Table 2: Comparison of reconstruction methods for the Ellipses, FoamFanB, and LoDoFanB datasets by average PSNR and SSIM. All supervised methods are trained on the Ellipses dataset. Learned models are then tested on the FoamFanB and LoDoFanB datasets. In the table, the two best performing methods are highlighted in bold case.

Methods		Low-Dose CT			Sparse-View CT		
		Ellipses	FoamFanB	LoDoFanB	Ellipses	FoamFanB	LoDoFanB
Unsupervised	FBP	28.50/0.844	20.73/0.629	33.01/0.842	26.74/0.718	16.34/0.174	29.10/0.593
	TV	33.41/0.878	36.39/0.939	36.55/0.869	30.98/0.869	27.53/0.832	34.74/0.833
	DIP+TV	34.53/0.957	**38.42/0.997**	**39.32/0.896**	32.02/0.931	**31.99/0.987**	**36.80/0.866**
Supervised	FBP+U-Net	37.05/0.970	30.26/0.723	32.13/0.820	32.13/0.936	20.09/0.347	27.22/0.694
	LGD	40.73/0.985	31.37/0.909	33.42/0.862	33.72/0.952	22.86/0.687	28.49/0.507
	LPD	**44.27/0.994**	28.09/0.918	33.21/0.866	**36.19/0.970**	24.86/0.886	34.60/0.838
	BDGD	**43.60/0.994**	30.72/0.974	35.91/0.877	**35.36/0.971**	19.44/0.406	34.16/0.824
BDGD+UKT		–	**40.72/0.997**	**38.40/0.899**	–	**30.07/0.966**	**35.67/0.855**

Table 3: "Upper-bounds" obtained via supervised fine-tuning on LoDoFanB.

Methods	Low-Dose CT	Sparse-View CT
FBP+U-Net	36.05/0.879	34.47/0.828
LGD	38.33/0.894	36.00/0.855
LPD	**39.85/0.914**	**37.59/0.876**
BDGD+SKT	**40.14/0.909**	**37.71/0.877**

Table 4: Comparison between BDGD+UKT and UL.

Methods	Low-Dose CT	Sparse-View CT
BDGD+UKT	**38.33/0.895**	35.67/**0.853**
UKT w/o TV	27.65/0.549	22.86/0.354
U-BDGD	36.64/0.870	**35.68**/0.852

learned methods (i.e., LGD, LPD, BDGD) require multiple calls of the forward operator A, and thus they are slower at test time than the methods that do not (e.g., FBP+U-Net, which only post-processes the FBP reconstruction). In addition, BDGD and BDGD+UKT use 10 Monte Carlo samples to obtain a single reconstruction, leading to a slightly longer reconstruction time of approximately 7s per image. However, all learned methods are found to be significantly faster than the TV reconstruction. Meanwhile, DIP+TV is much slower than TV taking approximately 20 minutes to reconstruct a single instance of the LoDoFanB dataset. The runtimes for the FoamFanB dataset were almost identical and are thus not included.

Example reconstructed images are shown in Figs. 3 and 4, for the sparse-view FoamFanB and the low-dose LoDoFanB CT settings, respectively. We observe that BDGD+UKT significantly reduces background noise in the reconstructions while faithfully capturing finer details, particularly in the low-dose setting. Overall, DIP+TV and BDGD+UKT produce reconstructions with similar properties. However, DIP+TV, LPD and BDGD+UKT tend to suffer from slight over-smoothing. Meanwhile, TV reconstruction suffers from patchy artefacts, which is a well-known drawback of the TV penalty [14], and also retains more background noise.

The sparse-view setting in Fig. 3 is numerically more challenging and the reconstructions are susceptible to streak artefacts, which are especially pronounced in the FBP but are still discernible in reconstructions obtained by other methods. Nonetheless, best performing methods (DIP+TV and BDGD+UKT) can achieve an excellent compromise between smoothing and the removal of streak artefacts. Interestingly, in Fig 4, the learned methods, including BDGD+UKT, suffer from some undesirable over-smoothing inside the lung cavity.

BDGD+UKT is good at recovering fine structures that are present in the FoamFanB data, which

are poorly reconstructed by BDGD. For example, in the last row of Fig. 3, the smaller circles are smoothed out and thus not discernible in BDGD reconstructions, but they are well reconstructed with BDGD+UKT. Similarly, Fig. 4 shows that BDGD+UKT better captures fine details in the human torso; for example the zoomed-in region shows an improvement over the overly-smoothed reconstruction produced by BDGD. These observations clearly indicate that the unsupervised fine-tuning is highly beneficial in improving the quality of the reconstructed image.

We further evaluate the learned methods by first pretraining them on the Ellipses dataset, and then fine-tuning them on one half of the LoDoFanB dataset but with ground truth data included. The remaining half of the LoBaFanB dataset is used for testing. We thus operate under the assumption that we have access to only one half of the ground truth images from the LoDoFanB dataset. This is intended to benchmark the reconstructive properties of the unsupervised fine-tuning against a more popular supervised adaptation, and may serve as the "upper-bound" on the reconstructive performance of the proposed method. The quantitative results of this controlled setting are presented in Tables 3 and 4. The notation SKT stands for the supervised knowledge-transfer: the fine-tuning is conducted via (3.2) on one half of the LoDoFanB dataset including ground truth data. Unsupervised (U)-BDGD refers to BDGD trained via (3.4) by completely omitting the pretraining in the first phase. It is observed that U-BDGD shows subpar reconstructive properties only for the low-dose CT setting, but surprisingly, it matches the performance obtained by BDGD+UKT for the sparse-view CT setting. However, we observe that pretraining helps to considerably speed up the convergence of BDGD+UKT. It takes only a few epochs to converge, whilst U-BDGD leads to a more unstable and lengthy learning (up to 100 epochs). This behaviour is also observed with the fine-tuning of other learned benchmark methods. This indicates the need of an adaptation phase, in the presence of distributional shift, and the beneficial effect of pretraining. Moreover, Table 3 shows that using supervised data pairs from the target domain to adapt the network to the target task can significantly improve the reconstructive properties of all the learned methods. Nonetheless, the degree of improvement depends strongly on both the used method and the problem setting. The proposed BDGD+UKT approach dramatically improves the performance and mitigates the performance drop due to the distributional shift. Table 4 also shows the influence of the TV in the fine-tuning stage. Setting $\gamma = 0$ leads to overfitting to the noise after 10 epochs (i.e., approx. 2000 gradient updates), and even with careful early stopping the performance is still subpar when compared with the approach employing TV regularisation. Therefore, the TV term plays an important role in the proposed framework.

It is worth noting that BDGD+UKT also provides useful predictive uncertainty information on the reconstructions. In Figs. 5 and 6, we present the uncertainty estimates along with pixel-wise errors for the FoamFanB and LoDoFanB CT settings, respectively. The overall predictive uncertainty largely concentrates around the edges: the reconstruction of sharp edges exhibits a higher degree of uncertainty. This agrees well with the intuition that edges are more challenging to accurately resolve than smooth regions, and thus are more prone to reconstruction errors. Note that aleatoric and epistemic uncertainties have different sources, one is due to inherent data noise, and the other due to the model uncertainty, arising from the lack of a sufficient amount of training data. To ascertain the sources, we apply the decomposition (3.1). Interestingly, we observe that in both the low-dose and the sparse-view CT settings, epistemic uncertainty appears to be dominating within the (overall) predictive uncertainty. Nonetheless, the two types of uncertainty share a similar shape, and in either case, the overall shape closely resembles the pixel-wise error, indicating that the uncertainty estimate can potentially be used as an error indicator, concurring with existing empirical measurement data [52]. It is also instructive to compare the uncertainty estimates obtained by BDGD and BDGD+UKT. Figs. 5 and 6 show that the estimates obtained by BDGD result in larger magnitudes, with the aleatoric component overshadowing the epistemic one. Visually,

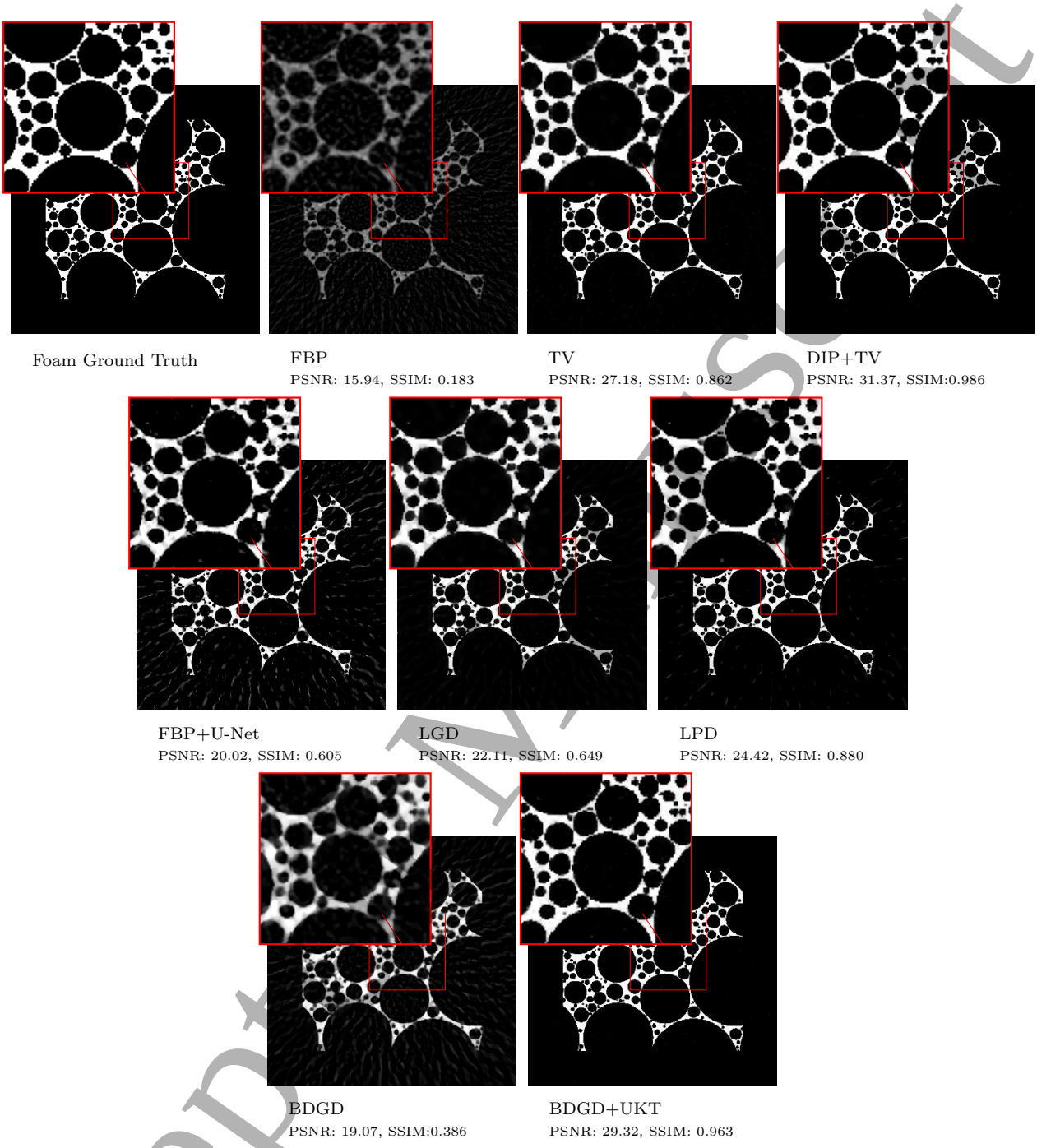


Figure 3: Sparse-view reconstruction of the FoamFanB dataset along with a zoomed-in region indicated by a small square.

the unsupervised adaptation phase ameliorates the epistemic estimate: the pixel-wise predictive epistemic uncertainty obtained with BDGD+UKT is better at capturing the edges of the anatomical structures present in the reconstructed image.

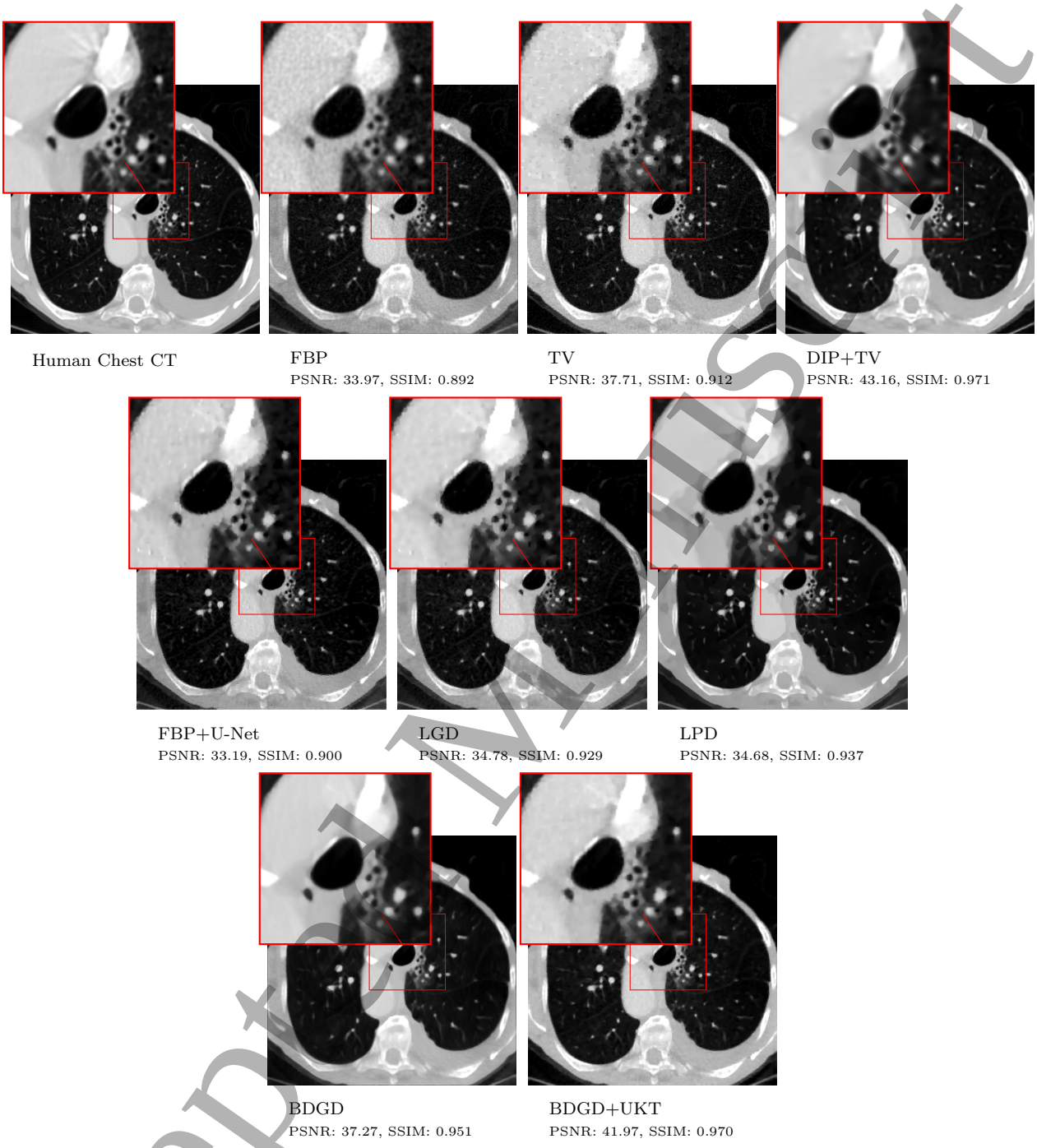


Figure 4: Low-dose human chest CT reconstruction within the LoDoFanB dataset along with a zoomed-in region indicated by a small square. The window is set to a HU range of ≈ [-1000, 400].

4.3 Discussion

The experimental results in Tables 2 and 4 have several implications for image reconstruction. First, they show that while supervised iterative methods (FBP+U-Net, LGD, and LPD) can deliver impressive results when trained and tested on imaging datasets that come from the same distribution,

but fail when applied directly to data from a different distribution. Specifically, on the Ellipses dataset they vastly outperform the traditional FBP and TV, but on the LoDoFanB dataset the difference between learned methods and FBP nearly vanishes (particularly in the low-dose setting), and the standard TV actually outperforms the supervised methods. This behaviour might be due to a form of bias-variance trade-off, where training with a large dataset allows improving the performance in the supervised case, but which has a negative effect on the generalisation property. The performance degrades significantly when the distribution of the testing measurement data deviates from that of the training data. This results in a loss of flexibility, and underwhelming performance, when reconstructing an image of a different type. Thus, adjusting the training regiment, or further adapting the network parameters to data from a different distribution, can be beneficial for improving the reconstruction quality. The results in Table 3 indicate that all learned methods, including BDGD+UKT, can benefit greatly from the supervised data from the target domain.

Overall, the results show that Bayesian neural networks with VI can deliver strong performance that is competitive with deterministic reconstruction networks, when equipped with the strategy of *being Bayesian only a little bit*. This can be first observed on the Ellipses dataset. Table 1 shows that BDGD performs on par with (or often better than) all the unsupervised and supervised methods under consideration, which is in agreement with previous experimental findings [10]. The results also show the potential of the Bayesian UKT framework for medical image reconstruction in the more challenging setting where ground truth images are not available. Namely, adapting the model through the described framework allows achieving a significant performance boost on both the FoamFanB and LoDoFanB datasets. Moreover, BDGD+UKT shows roughly the same performance as DIP+TV, while being significantly faster in terms of runtime, cf. Table 1. This observation is consistent with existing studies using pretraining in other contexts [27, 49]. Indeed, all the learned methods are significantly faster than TV and DIP+TV reconstructions. In addition, BDGD+UKT can deliver uncertainty estimates on the reconstructions, with their sources quantified into aleatoric and epistemic ones. It is observed that for the studied settings, the epistemic uncertainty dominates the aleatoric one, and both uncertainty estimates correlate well with the pixel-wise error of the reconstructions. Nonetheless, the calibration of these estimates remains to be validated, like nearly all DL-based uncertainty quantification techniques [9].

The extensive experimental results indicate that UKT shows great promise in the unsupervised setting. The results clearly show the need for adapting data-driven approaches to structural changes in the data, its distribution and size, and for incorporating the insights observed in the available supervised data to update the reconstruction model [46, 42]. Though only conducted on labelling tasks, recent studies show that transfer learning through pretraining exhibits good results when the difference between data distributions is small [59]. Moreover, one needs to ensure that pretraining does not result in overfitting the data from the first task. Both requirements seem to be satisfied in the studied setting. Further investigation is needed to examine how does the performance of a reconstruction network change with respect to the size and type of data that the pretraining dataset consists of, as well as with respect to changes in the physical setting (e.g., forward operators and noise statistics).

5 Concluding Remarks

The use of a full Bayesian treatment for learned medical image reconstruction methods is still largely under development, due to the associated training challenges [9]. The proposed BDGD+UKT is very promising in the following aspects: (i) it is easy to train due to the adoption of the strategy *being Bayesian only a little bit*; (ii) the performance of the obtained point estimates is competitive

with benchmark methods; (iii) it also delivers predictive uncertainty. In particular, the numerical results indicate that the predictive uncertainty can be visually used as a reliable error indicator. In this work we have presented a novel two-phase learning framework, termed UKT, for addressing the lack of a sufficiently large amount of paired training data in learned image reconstruction techniques. The framework consists of two learning phases, both within a Bayesian framework. It first pretrains a learned iterative reconstructor on (simulated) ordered pairs and then at test-time, it fine-tunes the model to realise sample-wise adaptation using only noisy clinically realistic measurements. Extensive experiments on low-dose and sparse-view CT reconstructions show that the approach is indeed very promising. It can achieve competitive performance with several state-of-the-art supervised and unsupervised approaches both qualitatively and quantitatively.

Acknowledgements

The authors are grateful to two anonymous referees for their constructive comments, and Johannes Leuschner (University of Bremen) for helpful discussions.

A Alternative Loss for Unsupervised Knowledge-Transfer

The derivation of the training loss (3.4) is not very principled in the sense that the trace term $\mathrm{trace}(A\hat{\Sigma}A^\top)$ does not arise naturally from the likelihood function $p(y^{\mathrm{u}}|\theta_{\mathrm{e}})$ using Bayes' rule. One can though derive alternative losses by slightly modifying the construction of the likelihood $p(y^{\mathrm{u}}|\theta_{\mathrm{e}})$. Below we give one such construction based on hierarchical modelling. For any measurement datum $y^{\mathrm{u}} \in \mathbb{B}^{\mathrm{u}}$, corresponding to the unknown image x^{u}, the likelihood $p(y^{\mathrm{u}}|x^{\mathrm{u}})$ is set to

$$p(y^{\mathrm{u}}|x^{\mathrm{u}}) = \mathcal{N}(y^{\mathrm{u}}; Ax^{\mathrm{u}}, \sigma^2 I),$$

where σ^2 dictates the strength of the measurement noise. The likelihood $p(y^{\mathrm{u}}|x^{\mathrm{u}})$ can be obtained under the standard assumption that the noise corruption to the exact data Ax^{u} (for the unknown image x^{u}) follows a Gaussian distribution with zero mean and variance $\sigma^2 I$. Meanwhile, under the heteroscedastic noise modelling, the unknown image x^{u} (which in turn depends on the network parameter θ) is assumed to follow a Gaussian distribution

$$p(x^{\mathrm{u}}|\theta_{\mathrm{e}}) = \mathcal{N}\big(x^{\mathrm{u}}; \mathrm{F}_\theta^\mu(x_0^{\mathrm{u}}), \hat{\Sigma}\big),$$

with the mean $\mathrm{F}_\theta^\mu(x_0^{\mathrm{u}})$ and covariance $\hat{\Sigma}$ being the two outputs of the neural network F_θ. Consequently, assuming that the data noise and the unknown image x^{u} are independent, combining the last two identities using Bayes' rule leads to

$$p(y^{\mathrm{u}}|\theta_{\mathrm{e}}) = \mathcal{N}(y^{\mathrm{u}}; A\mathrm{F}_\theta^\mu(x_0^{\mathrm{u}}), A\hat{\Sigma}A^\top + \sigma^2 I).$$

Upon letting $\bar{y}^{\mathrm{u}} = A\mathrm{F}_\theta^\mu(x_0^{\mathrm{u}})$, we then have

$$\log p(y^{\mathrm{u}}|\theta_{\mathrm{e}}) = -\frac{1}{2}(\bar{y}^{\mathrm{u}} - y^{\mathrm{u}})^\top (A\hat{\Sigma}A^\top + \sigma^2 I)^{-1}(\bar{y}^{\mathrm{u}} - y^{\mathrm{u}}) - \frac{1}{2}\log(\det(A\hat{\Sigma}A^\top + \sigma^2 I)) - \frac{m}{2}\log(2\pi).$$

In addition to enforcing data fidelity, we may also include the total variation penalty into the loss (to stabilise the training process). Finally, after expanding, relabelling, and ignoring the constant

terms (in θ), we obtain the following alternative loss at the second phase

$$\tilde{\mathcal{L}}^{\mathrm{u}}(\theta_{\mathrm{d}}, q_\psi(\theta_{\mathrm{e}})) = \frac{1}{N^{\mathrm{u}}} \sum_{n=1}^{N^{\mathrm{u}}} \mathbb{E}_{q_\psi(\theta_{\mathrm{e}})}\Big[\frac{1}{2}(A\mathrm{F}_\theta^\mu(x_{n,0}^{\mathrm{u}}) - y^{\mathrm{u}})^\top (A\hat{\Sigma}_n A^\top + \sigma^2 I)^{-1}(A\mathrm{F}_\theta^\mu(x_{n,0}^{\mathrm{u}}) - y^{\mathrm{u}}) + \frac{1}{2}\log(\det(A\hat{\Sigma}_n A^\top + \sigma^2 I)) + \gamma \mathrm{TV}(\mathrm{F}_\theta^\mu(x_{n,0}^{\mathrm{u}}))\Big] + \beta \mathrm{KL}\left[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi *}^{\mathrm{s}}(\theta_{\mathrm{e}})\right]. \quad (A.1)$$

Note that the term $\mathrm{KL}[q_\psi(\theta_{\mathrm{e}}) \| q_{\psi *}^{\mathrm{s}}(\theta_{\mathrm{e}})]$ has a closed-form expression.

The loss in (A.1) differs from that in (3.4) only in the construction of the likelihood $p(y^{\mathrm{u}}|\theta_{\mathrm{e}})$. However, the former is computationally less convenient, due to the presence of the factor $(A\hat{\Sigma}A^\top + \sigma^2 I)^{-1}$ in the data consistency term, as well as the log-determinant $\log(\det(A\hat{\Sigma}A^\top + \sigma^2 I))$. Indeed, in view of the following well-known matrix directional derivative formulas

$$\frac{\mathrm{d}\log(\det(X))}{\mathrm{d}X}[H] = \mathrm{trace}(X^{-1}H) \quad \text{and} \quad \frac{\mathrm{d}X^{-1}}{\mathrm{d}X}[H] = -X^{-1}HX^{-1},$$

for any symmetric positive definite X, and admissible direction H, the gradient evaluation requires solving multiple linear systems, with the matrices given only implicitly. This can be computationally demanding for large-scale image restoration tasks such as CT reconstruction. In practice, the derivative of the log-determinant can be efficiently approximated using randomised trace estimators (e.g., the Hutchinson's estimator [15], which again involves multiple linear solves).

The next result shows that the loss in (3.4) is actually a computationally more tractable approximation to the genuine Bayesian loss in (A.1), under the condition $A\hat{\Sigma}A^\top \ll \sigma^2 I$ (i.e., the matrix $\sigma^{-2}A\hat{\Sigma}A^\top$ has a small operator norm). This result provides a more principled Bayesian interpretation of the loss (3.4).

Proposition A.1. *Under the condition $A\hat{\Sigma}A^\top \ll \sigma^2 I$, the loss in (3.4) is an approximation to the Bayesian loss in (A.1).*

Proof. Let $r = A\mathrm{F}_\theta^\mu(x_0^{\mathrm{u}}) - y^{\mathrm{u}}$ be the residual. It follows directly from the preceding matrix derivative formulas that

$$\begin{aligned} r^\top (A\hat{\Sigma}A^\top + \sigma^2 I)^{-1} r &\approx r^\top \sigma^{-2} I r - r^\top \sigma^{-4} A\hat{\Sigma}A^\top r, \\ \log(\det(A\hat{\Sigma}A^\top + \sigma^2 I)) &= \log(\det(\sigma^2 I)) + \log(\det(I + \sigma^{-2}A\hat{\Sigma}A^\top)) \\ &\approx m\log\sigma^2 + \mathrm{trace}(\sigma^{-2}A\hat{\Sigma}A^\top). \end{aligned}$$

Note that the approximation is good under the given conditions. Then substituting these approximations into (A.1), ignoring the constant term and relabelling, we obtain the loss in (3.4). This shows the desired assertion. □

References

[1] J. Adler, H. Kohr, and O. Öktem. Operator discretization library (ODL). Software available from https://github.com/odlgroup/odl, 2017.

[2] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[3] J. Adler and O. Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018.

[4] M. Akçakaya, B. Yaman, H. Chung, and J. C. Ye. Unsupervised deep learning methods for biological image reconstruction. *IEEE Signal Processing Magazine*, 39(2):28–44, 2022.

[5] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48):30088–30095, 2020.

[6] S. G. Armato III, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, et al. Lung image database consortium: Developing a resource for the medical imaging research community. *Radiology*, 232(3):739–748, 2004.

[7] S. Arridge, P. Maaß, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[8] D. O. Baguer, J. Leuschner, and M. Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9):094004, 2020.

[9] R. Barbano, S. Arridge, B. Jin, and R. Tanno. Uncertainty quantification for medical image synthesis. In N. Burgos and D. Svoboda, editors, *Biomedical Image Synthesis and Simulation: Methods and Applications*, pages 601–641. Elsevier, 2022.

[10] R. Barbano, Ž. Kereta, C. Zhang, A. Hauptmann, S. Arridge, and B. Jin. Quantifying sources of uncertainty in deep learning-based image reconstruction. NeurIPS 2020 Workshop on Deep Learning and Inverse Problems, 2020.

[11] R. Barbano, C. Zhang, S. Arridge, and B. Jin. Quantifying model-uncertainty in inverse problems via Bayesian deep gradient descent. In *International Conference on Pattern Recognition*, pages 1392–1399, 2021.

[12] S. Bickel. *Learning under Differing Training and Test Distributions*. PhD thesis, Universität Potsdam, 2008.

[13] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[14] K. Bredies and M. Holler. Higher-order total variation approaches and generalisations. *Inverse Problems*, 36(12):123001, 128, 2020.

[15] Z. Bujanovic and D. Kressner. Norm and trace estimation with random rank-one vectors. *SIAM Journal on Matrix Analysis and Applications*, 42(1):202–223, 2021.

[16] P. Cascarano, A. Sebastiani, M. C. Comes, G. Franchini, and F. Porta. Combining weighted total variation and deep image prior for natural and medical image restoration via ADMM. arXiv preprint arXiv:2009.11380, 2020.

[17] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, 2017.

[18] S. U. H. Dar, M. Özbey, A. B. Çatlı, and T. Çukur. A transfer-learning approach for accelerated MRI using deep neural networks. *Magnetic Resonance in Medicine*, 84(2):663–685, 2020.

[19] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *Proceedings of the 38th International Conference on Machine Learning, PMLR*, volume 139, pages 2510–2521, 2021.

[20] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193, 2018.

[21] S. Dittmer, T. Kluth, P. Maaß, and D. Otero Baguer. Regularization by architecture: A deep prior approach for inverse problems. *Journal of Mathematical Imaging and Vision*, 62(3):456–470, 2020.

[22] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[23] Y. Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[24] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning, PMLR*, volume 48, pages 1050–1059, 2016.

[25] D. Gilton, G. Ongie, and R. Willett. Model adaptation in biomedical image reconstruction. In *International Symposium on Biomedical Imaging*, pages 1223–1226, 2021.

[26] Y. Han, J. Yoo, H. H. Kim, H. J. Shin, K. Sung, and J. C. Ye. Deep learning with domain adaptation for accelerated projection-reconstruction MR. *Magnetic Resonance in Medicine*, 80(3):1189–1205, 1998.

[27] K. He, R. Girshick, and P. Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4918–4927, 2019.

[28] K. Ito and B. Jin. *Inverse Problems: Tikhonov Theory and Algorithms*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2015.

[29] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[30] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[31] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021.

[32] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Neural Information Processing Systems*, pages 5580–5590, 2017.

[33] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Neural Information Processing Systems*, pages 2575–2583, 2015.

[34] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. Preprint, arXiv:1312.6114, 2013.

[35] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier. Recalibration of aleatoric and epistemic regression uncertainty in medical imaging. *arXiv preprint arXiv:2104.12376*, 2021.

[36] J. Leuschner, M. Baltazar, and D. Erzmann. Deep inversion validation library. *Software available from https://github.com/jleuschn/dival*, 2019.

[37] J. Leuschner, M. Schmidt, D. O. Baguer, and P. Maaß. The LoDoPaB-CT dataset: A benchmark dataset for low-dose CT reconstruction methods. *Scientific Data*, 8:109, 2021.

[38] T. Li, X. Li, J. Wang, J. Wen, H. Lu, J. Hsieh, and Z. Liang. Nonlinear sinogram smoothing for low-dose X-ray CT. *IEEE Transactions on Nuclear Science*, 51(5):2505–2513, 2004.

[39] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[40] D. J. C. Mackay. *Bayesian Methods for Adaptive Models*. PhD thesis, Pasadena, California, 1992. UMI Order No. GAX92-32200.

[41] V. Monga, Y. Li, and Y. C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[42] M. Mundt, Y. W. Hong, I. Pliushch, and V. Ramesh. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. Preprint, arXiv:2009.01797, 2020.

[43] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *International Conference on Neural Networks*, volume 1, pages 55–60. IEEE, 1994.

[44] G. Ongie, A. Jalal, R. G. Baraniuk, C. A. Metzler, A. G. Dimakis, and R. Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, pages 39–56, 2020.

[45] K. Osawa, S. Swaroop, M. E. E. Khan, A. Jain, R. Eschenhagen, R. E. Turner, and R. Yokota. Practical deep learning with Bayesian principles. In *Neural Information Processing Systems*, 2019.

[46] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[47] D. M. Pelt, K. J. Batenburg, and J. A. Sethian. Improving tomographic reconstruction from limited data using mixed-scale dense convolutional neural networks. *Journal of Imaging*, 4(11):128, 2018.

[48] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, Massachussets, 2009.

[49] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3347–3357, 2019.

[50] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015.

[51] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248, 2020.

[52] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.

[53] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[54] W. Van Aarle, W. J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers. The ASTRA toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*, 157:35–47, 2015.

[55] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[56] G. Wang, J. C. Ye, and B. De Man. Deep learning for tomographic image reconstruction. *Nature Machine Intelligence*, 2(12):737–748, 2020.

[57] J. Wang, T. Li, H. Lu, and Z. Liang. Noise reduction for low-dose single-slice helical CT sinograms. *IEEE Transactions on Nuclear Science*, 53(3):1230–1237, 2006.

[58] Y. Wu and K. He. Group normalization. In *European Conference on Computer Vision*, pages 3–19, 2018.

[59] X. Yang, X. He, Y. Liang, Y. Yang, S. Zhang, and P. Xie. Transfer learning or self-supervised learning? A tale of two pretraining paradigms. arXiv preprint arXiv:2007.04234, 2020.

[60] J. Zhang, Z. Liu, S. Zhang, H. Zhang, P. Spincemaille, T. D. Nguyen, M. R. Sabuncu, and Y. Wang. Fidelity imposed network edit (FINE) for solving ill-posed image reconstruction. *NeuroImage*, 211:116579, 2020.

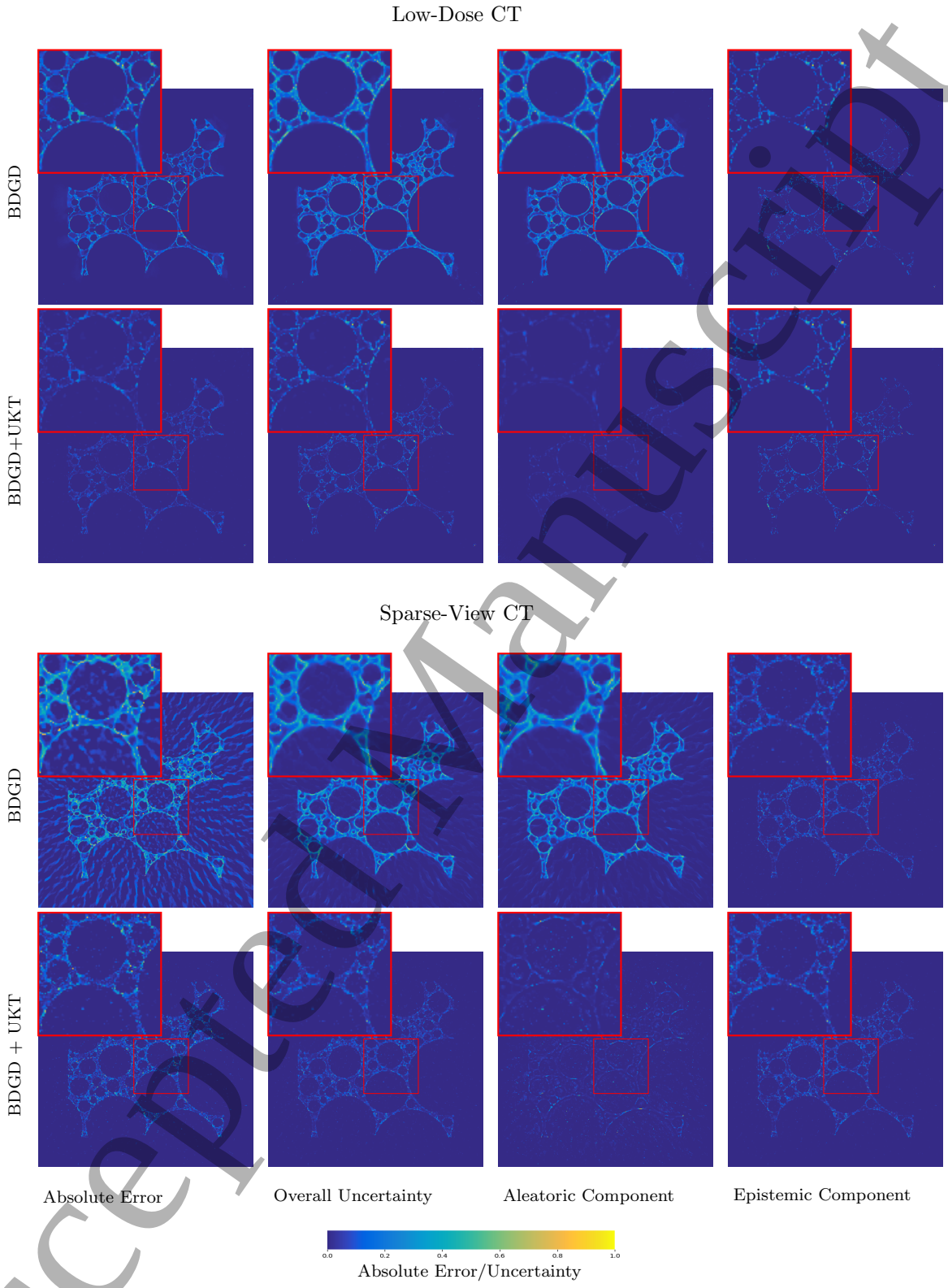


Figure 5: Qualitative uncertainty analysis on the FoamFanB dataset. The pixel-wise absolute reconstruction error, (max-min normalised across low-dose and sparse-view CT settings) pixel-wise predictive uncertainty, and its decomposition into the aleatoric and epistemic constituent components for low-dose and sparse-view CT obtained by BDGD and BDGD+UKT.

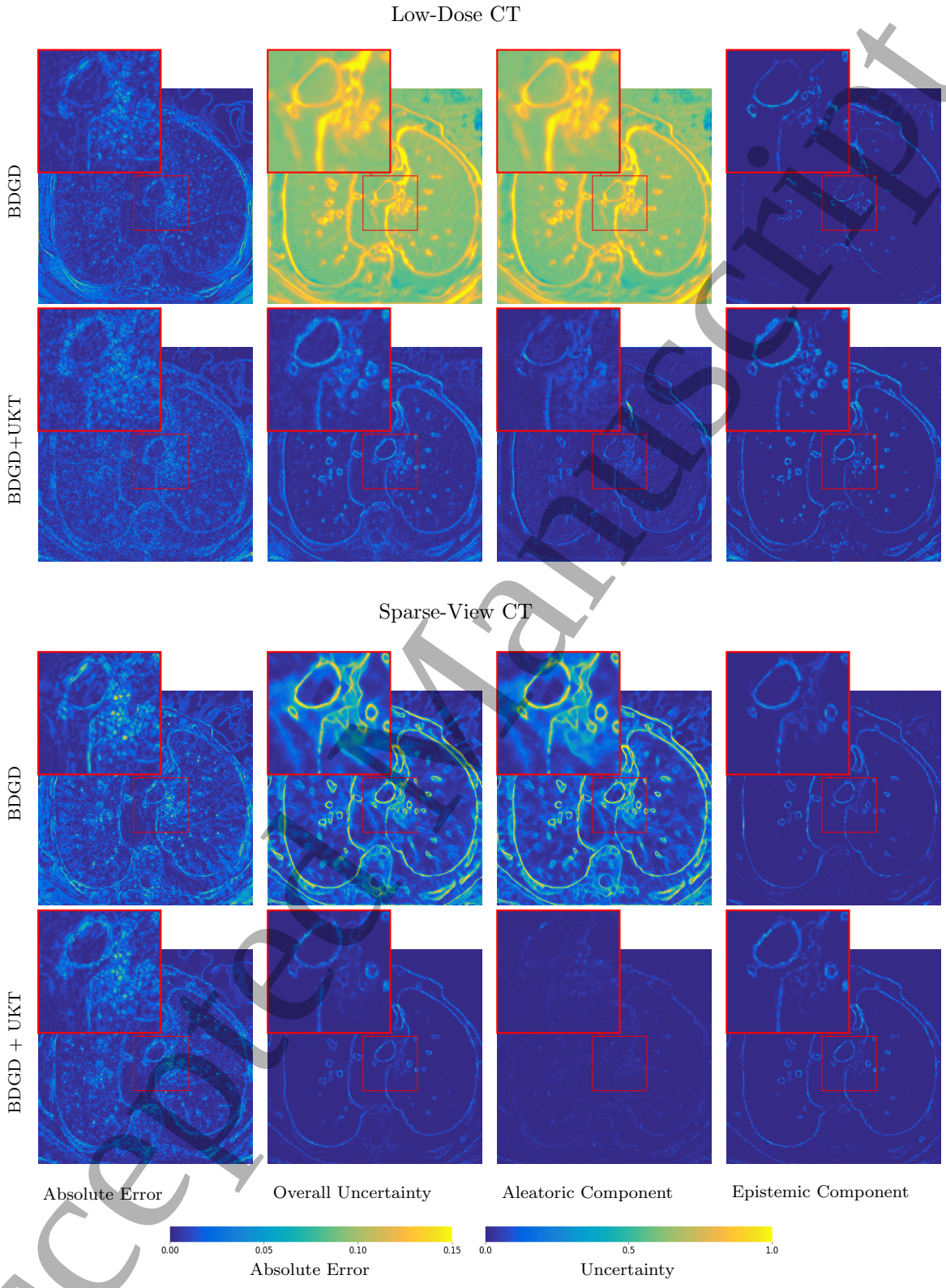


Figure 6: Qualitative uncertainty analysis on the LoDoFanB dataset. The pixel-wise absolute reconstruction error, (max-min normalised across low-dose and sparse-view CT settings) pixel-wise predictive uncertainty, and its decomposition into the aleatoric and epistemic constituent components for low-dose and sparse-view CT obtained by BDGD and BDGD+UKT.