OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **38** (2022) 075006 (21pp) https://doi.org/10.1088/1361-6420/ac28ec

# Task adapted reconstruction for inverse problems

**Jonas Adler**[1,2], **Sebastian Lunz**[3], **Olivier Verdier**[1,4], **Carola-Bibiane Schönlieb**[3] **and Ozan Öktem**[1,5,*]

[1] Department of Mathematics, KTH–Royal Institute of Technology, 100 44 Stockholm, Sweden
[2] DeepMind, 6 Pancras Square, London N1C 4AG, United Kingdom
[3] Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WA, United Kingdom
[4] Department of Computing, Mathematics and Physics, Western Norway University of Applied Sciences, Bergen, Norway
[5] Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden

E-mail: ozan@kth.se

Received 1 March 2021, revised 9 June 2021
Accepted for publication 21 September 2021
Published 31 May 2022

**Abstract**

The paper considers the problem of performing a post-processing task defined on a model parameter that is only observed indirectly through noisy data in an ill-posed inverse problem. A key aspect is to formalize the steps of reconstruction and post-processing as appropriate estimators (non-randomized decision rules) in statistical estimation problems. The implementation makes use of (deep) neural networks to provide a differentiable parametrization of the family of estimators for both steps. These networks are combined and jointly trained against suitable supervised training data in order to minimize a joint differentiable loss function, resulting in an end-to-end task adapted reconstruction method. The suggested framework is generic, yet adaptable, with a plug-and-play structure for adjusting both the inverse problem and the post-processing task at hand. More precisely, the data model (forward operator and statistical model of the noise) associated with the inverse problem is exchangeable, e.g., by using neural network architecture given by a learned iterative method. Furthermore, any post-processing that can be encoded as a trainable neural network can be used. The approach is demonstrated on joint tomographic

[*] Author to whom any correspondence should be addressed.

1361-6420/22/075006+21$33.00 © 2022 The Author(s). Published by IOP Publishing Ltd Printed in the UK

image reconstruction, classification and joint tomographic image reconstruction segmentation.

Keywords: inverse problems, image reconstruction, tomography, deep learning, feature reconstruction, segmentation, classification

(Some figures may appear in colour only in the online journal)

## 1. Introduction

The overall goal in inverse problems is to determine a model parameter that generates model predictions matching measured data to sufficient accuracy. Such problems arise in science and engineering applications, like imaging/sensing applications. Here, the model parameter to be determined is the image and data represents noisy indirect observations that is acquired using an imaging device, like a tomography scanner or microscope. Perhaps, the most-prominent example is computed tomography (CT) imaging where data is a series of x-ray scans of the patient taken from different directions. The aim in CT imaging is to use the data to computationally recover a 3D image representing the interior anatomy. The technique was introduced into the medical field in late 1960s with the Nobel Prize in Physiology or Medicine awarded shortly afterwards in 1979. Tomography has since then revolutionized health care, allowing doctors to find disease earlier and improve patient outcomes [54].

The inverse problem of reconstructing the model parameter from data is often only one out of many steps in a procedure that ultimately involves decision making. The reconstructed model parameter is typically summarized, either by an expert or automatically, and the resulting task dependent descriptors are used as basis for decision making. As an example, in clinical CT imaging the image is segmented and annotated (semantic segmentation) by a clinical radiologist. Depending on the underlying referral to radiology, the radiologist may also compute specific quantitative measures, like the size/volume of lesions. The above is complemented with qualitative assessment and the findings, bearing in mind the underlying referral, are summarized in a radiology report, that serves as basis for decision making. This workflow for clinical CT, where the recovered model parameter undergoes several down-stream post-processing steps before being integrated into decision making, is typical for applications that involve solving inverse problems as outlined in figure 1.

There could be several disadvantages with a sequential approach in which the various steps of the above pipeline are performed independently from each other. Each single step introduces approximations that are not accounted for by subsequent steps, the reconstruction may not consider the end task, and the feature extraction may not consider how data was acquired. Actually, the task is almost always only accounted for at the final step. It is therefore natural to ask whether one may adapt the reconstruction method for a specific task. *Task adapted reconstruction* refers to methods that integrate the reconstruction procedure with (parts of) the decision making associated with the task. This is sometimes also referred to as 'end-to-end' reconstruction. The numerical examples in the paper compares the task adapted reconstruction against a sequential approach to see whether there are any advantages with the former.

Overview. We start with a brief survey of existing approaches to task adapted reconstruction in the context of tomographic image reconstruction (section 1.1), which also points out the drawbacks that come with these approaches. The section that follows (section 2) introduces the Bayesian view on inverse problems where a reconstruction method is a statistical estimator.

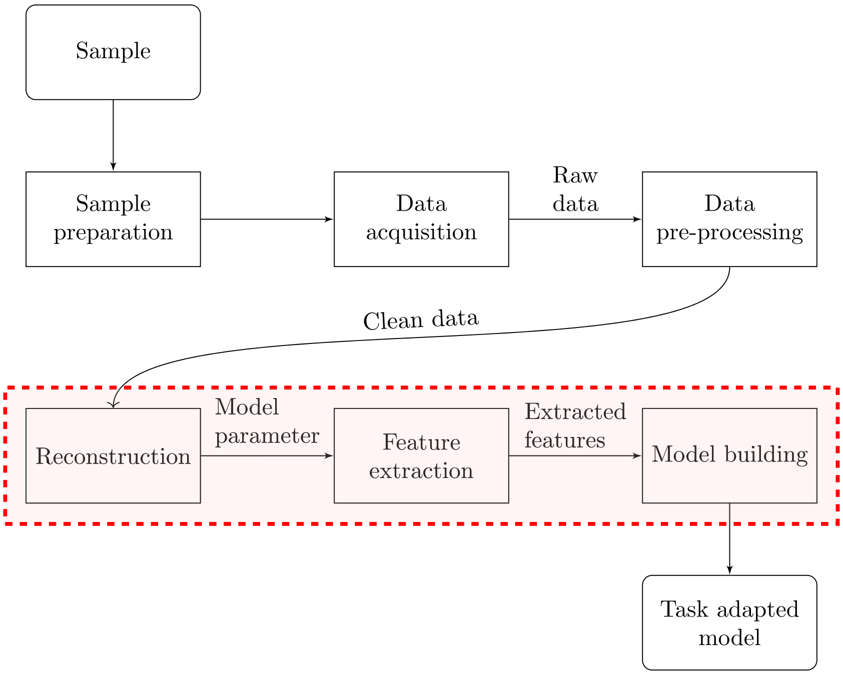


**Figure 1.** Typical workflow involving an inverse problem. The second row represents the data acquisition where raw data is acquired and pre-processed, resulting in cleaned data. In the third row, the cleaned data is used as input to a reconstruction step that recovers the model parameter, which is then post-processed to extract features that are used as input for model building. The final outcome is a task adapted model that can be used for decision making. The dotted part outlines steps that are unified by task adapted reconstruction.

In a similar manner we also formulate post-processing applied to the model parameter space as a statistical estimation problem (section 3). Section 4 starts by introducing task adapted reconstruction in an abstract setting. It then outlines three approaches, all well suited for data driven approaches based on neural networks. This is followed by two applications that are worked out in detail in section 5. Section 6 provides some theoretical considerations regarding regularizing properties and the potential advantage that comes with using a joint approach. Section 7 has a concluding discussion and outlook on future research in this area.

### *1.1. Survey of task adapted reconstruction*

There is an ongoing effort within the inverse problems community to perform reconstruction jointly with various post-processing tasks.

Current approaches are primarily based on a functional analytic/frequentist formalism where the aim is to recover some true (unknown) post-processing feature $d^* \in D$ from data $y \in Y$, which is a *single sample* of the $Y$-valued random variable of $\mathbb{y}$ where

$$d^* = \mathcal{T}(x^*) \quad \text{with} \quad \mathbb{y} = \mathcal{A}(x^*) + \mathbb{e} \quad \text{and} \quad \mathbb{e} \sim \mathsf{P}_{\text{noise}}. \tag{1}$$

In the above, $\mathsf{P}_{\text{noise}}$ is the probability distribution for the noise in measured data, which is assumed to be known. Furthermore, the mappings $\mathcal{A} : X \to Y$ (forward operator) and $\mathcal{T} : X \to D$ (post-processing operator) are also assumed to be known. The forward operator models how a model parameter (an image in tomography) gives rise to data and the post-processing operator extracts the post-processing features of interest from the model parameter.

$$\begin{tikzcd} X \arrow[r, "\mathcal{A}"] \arrow[d, "\mathcal{T}"'] & Y \arrow[dl, dotted] \\ D & \end{tikzcd} \tag{2}$$

Task adapted reconstruction amounts to finding the dotted mapping in (2). Note that the post-processing operator $\mathcal{T} : X \to D$ is often highly non-injective, so it makes no sense to view $\mathcal{A} \circ \mathcal{T}^{-1} : D \to Y$ as a 'new forward operator' in a regular (non-task adapted) reconstruction problem. Approaches for 'solving' the inverse problem formalized in (1) heavily depends on the nature of $\mathcal{T}$. Below are some examples in the context of tomographic imaging.

*1.1.1. Model based approaches.* Most previous work in model based task adapted reconstruction is in tomography. Here, edge recovery by means of Lambda-tomography is one of the earliest examples. This is a non-iterative reconstruction method of filtered backprojection (FBP) type for recovering edges of an image directly from tomographic data. It is based on a microlocal analysis of the reconstruction operator, see [32] for a nice survey. Another non-iterative approach combines the method of approximate inverse with an explicit edge detector [41]. Variational models represent a flexible approach for solving inverse problems where consistency against data is balanced against the need to avoid overfitting. The former is quantified in terms of the negative data log-likelihood whereas the latter is given by a functional that penalises unwanted features (regularizer). When applied to tomographic imaging, variational models do not recover edges in the formal sense. One can, however, choose a regularizer that emphasizes edges. An example is sparsity promoting regularization with respect to an underlying dictionary that is designed to sparsely represent edges (like curvelets, shearlets, beamlets, and bandlets) [34, 55], another is variants of total variation regularization [8].

Another example is joint tomographic reconstruction and segmentation where most approaches use a variational model with a suitably chosen regularizer. One is the Mumford–Shah-type of regularizer [11, 26, 31, 50], another is a regularizer given as the negative log-likelihood of a Potts-prior [57]. A further refinement is to consider joint reconstruction and semantic segmentation. A maximum a posteriori estimator with a Gauss–Markov–Potts type of prior leads to a variational model [45] that shows promising results on small-scale examples. Another approach uses a hidden Markov measure field model for estimating the probability of different segmentation classes [52].

The final example refers to tomographic reconstruction in spatiotemporal imaging where one needs to recover an image along with its motion. As surveyed in [22], current approaches are based on variational models where the motion is parametrized, e.g., by a partial differential equation (optical flow or a continuity equation) [10, 42] or by a sequence of time dependent diffeomorphic deformations acting on a template [12], which builds on variational models for joint reconstruction and image registration (indirect registration) [13, 19, 36].

*1.1.2. Data driven approaches.* Model based approaches require access to a handcrafted post-processing operator $\mathcal{T} : X \to D$ that can be interpreted as a feature extraction.[6] It can however be difficult to handcraft the post-processing operator for many natural post-processing tasks, and especially so for tasks involving some kind of classification. Next, state-of-the-art approaches for task adapted reconstruction are mostly based on variational models. These

[6] Such approaches to task adapted reconstruction are often called 'feature reconstruction' [41].

are computationally demanding and involve several hyper-parameters, like when tomographic reconstruction is performed jointly with segmentation or registration.

The above issues seriously limits the usefulness of model based approaches to task adapted reconstruction in applications involving large scale and/or time-critical. An option is therefore to resort to data driven approaches to *simultaneously address both* these limitations. First, there is ample empirical evidence of successfully using machine learning to perform complex post-processing tasks. Some of these, like object detection, image caption generation, and style transfer, are even challenging to formalize, let alone solve, in the context of classical signal processing. Next, once trained, data driven models are also computationally efficient to evaluate, which addresses the issue of computational feasibility. It therefore makes sense to integrate reconstruction with a data driven model for the post-processing task.

The earliest and most commonly occurring example of machine learning in image reconstruction is for denoising/restoration of reconstructed images. In such a sequential approach, the post-processing can only be made aware of the underlying reconstruction method through the choice of training data. As an example, deep learning in CT imaging relies often on training a deep neural network to denoise/restore CT images obtained from a particular reconstruction method (usually FBP) [5, section 5.1.5]. An alternative to such sequential approaches is to perform the denoising/restoration task jointly with reconstruction as outlined in [43, 44].

A more complex task is joint reconstruction and segmentation. In MRI imaging, one can have a unified deep neural network architecture (SegNetMRI) that performs combined Fourier inversion (MRI image reconstruction) and segmentation [58]. Here, one has two neural networks with the same encoder–decoder structure, one for MRI reconstruction consisting of multiple cascaded blocks, each containing an encoder–decoder unit and a data fidelity unit, and the other for segmentation. These are pre-trained and coupled by ensuring they share reconstruction encoders. Another example of data driven approach to joint reconstruction and segmentation is [9], which considers photo-acoustic tomography. The idea is to jointly train a learned iterative scheme for reconstruction [2] with a convolutional neural network for segmentation. A similar approach is taken in [65] for joint reconstruction and abnormality detection, e.g., tumour detection in low-dose CT imaging. Here one jointly trains a learned iterative scheme for reconstruction [2] with a 3D convolutional neural network for detecting the abnormality in the reconstructed images.

A final example of data driven approaches to task adapted reconstruction refers to spatiotemporal imaging. A key part in performing reconstruction jointly with a motion model is the ability to vary the motion model itself through an explicit parametrization [22]. There are many approaches that use machine learning to model motion, but most try to learning the entire motion and therefore they do not admit an explicit parametrization that can be used in reconstruction. What is needed for joint reconstruction and motion modelling is machine learning approaches that learn a motion model that belongs to a pre-defined class of motion models with an explicit parametrization. One example of this is given in [49] for spatiotemporal positron emission tomography (PET) image reconstruction. Here, the ML–EM iterates for updating the PET activity image are intertwined with a parametrized learned diffeomorphic deformations that are given by deep neural networks (VoxelMorph) [7] Another similar approach for indirect registration (joint reconstruction and registration) intertwines iterates for a variational model with learned deformation operators [39]. Finally we point to [56] for and example of reconstruction in MRI with a learned optical flow model for the motion.

### *1.2. Motivation and specific contributions*

The approaches surveyed in section 1.1 involve post-processing tasks from classical image processing, like edge detection, segmentation, and registration. These represent a small (but important) sub-set of the tasks necessary for integrating imaging with decision making. Meanwhile, data driven approaches based on deep learning have made impressive progress in automating a wide range of complex tasks, like object detection, image caption generation, and image translation. Tasks like these typically involve classification and/or generative modelling and they constitute an important part of image guided decision making. Integrating such deep learning based approaches with reconstruction opens up for the ability to perform reconstruction jointly with a wide range of complex post-processing tasks, like those required for in radiomics where the aim is to extract features from images and integrate these into clinical-decision support systems to improve diagnostic, prognostic, and predictive accuracy [35, 61].

The aim here is to develop a generic, yet highly adaptable, framework for task adapted reconstruction that is based on considering both the reconstruction and the post-processing task as statistical estimation problems To the best of our knowledge, this is the first paper that offers an approach where reconstruction can be performed jointly with a range of diverse post-processing tasks in a computationally feasible manner. Statistical decision and learning theory coupled with efficient inference algorithms from deep learning are the guiding principles and the framework can re-use algorithmic components. This opens up for new ways of thinking about machine learning and inverse problems that may ultimately lead to deeper understanding of the possibilities for integrating elements of decision making into the reconstruction. Further novel contributions is usage of the joint loss in (18) and investigating its properties (section 6).

## 2. Reconstruction as statistical estimation

Representing a reconstruction method as a statistical estimator (non-randomized decision rule) requires one to define probability measures on both $X$ (reconstruction space) and $Y$ (data space), i.e., we have measurable spaces $(X, \mathfrak{S}_X)$ and $(Y, \mathfrak{S}_Y)$ where $X$ and $Y$ are separable Banach spaces. Furthermore, $\mathscr{P}_X$ and $\mathscr{P}_Y$ denotes corresponding spaces of measures on $X$ and $Y$. Finally, one has a data model $\mathcal{M} : X \to \mathscr{P}_Y$ that maps a model parameter to the distribution of the data generated by it, i.e., the data model simulates the measurement. Following [16], reconstruction corresponds to a point estimator of $(\mathbb{x} \,|\, \mathbb{y} = y)$ where measured data $y \in Y$ is a single sample of $(\mathbb{y} | \mathbb{x} = x^*) \sim \mathcal{M}(x^*)$ with $x^* \in X$ unknown.

**Remark 1.** It is common to express the data model in terms of a forward operator $\mathcal{A} : X \to Y$ that models how data is generated in absence of noise and a part that represents measurement noise. This corresponds to $\mathcal{M}(x) := \delta_{\mathcal{A}(x)} \circledast \mathsf{P}_{\text{noise}} = \mathsf{P}_{\text{noise}}\left( \cdot\, - \mathcal{A}(x) \right)$ where $\mathsf{P}_{\text{noise}} \in \mathscr{P}_Y$ models measurement noise. Expressed equivalently, $\mathbb{y} = \mathcal{A}(\mathbb{x}) + \mathbb{e}$ where $\mathbb{e} \sim \mathsf{P}_{\text{noise}}$ is the measurement noise. An additional assumption is that $\mathbb{e}$ is independent of $\mathbb{x}$, e.g., as when $\mathsf{P}_{\text{noise}}$ is a Gaussian random measure with a mean and co(variance) independent of $\mathbb{x}$. Another data model is when $\mathcal{M}(x)$ is a Poisson random measure on $Y$ with mean $\mathcal{A}(x)$, which is suitable for imaging in a low-dose setting that relies on counting statistics [25].

One possibility is to use maximum likelihood estimation for reconstruction, i.e., to maximize $x \mapsto \mathcal{M}(x)$. This is however unsuitable for ill-posed inverse problems since it leads to overfitting, e.g., this estimator corresponds to un-regularized least-squares fitting when $\mathbb{e}$ is Gaussian white noise. Statistical decision theory offers a systematic framework for comparing estimators and selecting one that is optimal according to some pre-defined criteria. This can be used for selecting an estimator for reconstruction that is better suited for ill-posed inverse problems. This requires us to formalize reconstruction as Bayesian estimation. To do this, note first that the tuple $\big((Y,\mathfrak{S}_Y),\{\mathcal{M}(x)\}_{x\in X}\big)$ formally defines a statistical model in the sense of [38, equation (1.1)]. If one is given a loss $\ell_X : X\times X\to\mathbb{R}$, then a reconstruction method can be seen as a non-randomized decision rule (estimator) in a statistical estimation problem [38, definition 3.5] with $(X,\mathfrak{S}_X)$ as the decision space. Using a Bayes estimator for reconstruction amounts to selecting the reconstruction method that minimizes the average loss [38, definition 3.36]. Such a reconstruction method is given by a mapping $\widehat{\mathcal{R}} : Y\to X$ that solves the following supervised statistical learning problem

$$\widehat{\mathcal{R}} \in \operatorname*{argmin}_{\mathcal{R}\in\mathscr{D}_{\text{rec}}} \mathbb{E}_{(\mathbb{x},\mathbb{y})\sim\mathsf{P}_X\otimes\mathcal{M}(x)}\left[\ell_X\left(\mathbb{x},\mathcal{R}(\mathbb{y})\right)\right] \text{ for given } \mathsf{P}_X\otimes\mathcal{M}(x)\in\mathscr{P}_{X\times Y}. \tag{3}$$

Here, $\mathscr{D}_{\text{rec}}$ is a given class of measurable $X$-valued mappings on $Y$. Next, $\mathcal{M} : X\to\mathscr{P}_Y$ (data model) is also known, whereas the 'true' $\mathsf{P}_X\in\mathscr{P}_X$ (prior) is often unknown.

Implementation. The expectation in (3) is w.r.t. the joint law $(\mathbb{x},\mathbb{y})\sim\mathsf{P}_X\otimes\mathcal{M}\in\mathscr{P}_{X\times Y}$. Domain specific expertise in inverse problems often provides a reasonably accurate data model $x\mapsto\mathcal{M}(x)\in\mathscr{P}_Y$. In contrast, less is known on how to handcraft a prior that is sufficiently close to the 'true' prior $\mathsf{P}_X\in\mathscr{P}_X$, so the joint law for $(\mathbb{x},\mathbb{y})$ is unknown.

One alternative to handcrafting a prior is to take a data driven approach. If one has (supervised) training data $(x_i,y_i)\in X\times Y$ generated by $(\mathbb{x},\mathbb{y})\sim\mathsf{P}_X\otimes\mathcal{M}$, then one can replace the joint law $\mathsf{P}_X\otimes\mathcal{M}$ in (3) with its empirical counterpart given by this training data. Inverse problems are often large scale, meaning that elements in $X$ and $Y$ are represented by high dimensional arrays even after clever discretization. The amount of training data is furthermore often limited, so a key aspect is to consider a domain adapted class of estimators $\mathscr{D}_{\text{rec}} = \{\mathcal{R}_\theta\}_{\theta\in\Theta}$ in (3). In the context of inverse problems, this means making use of the *a priori* knowledge that $\mathcal{R}_\theta : Y\to X$ should represent a (regularized) inverse of an explicitly known forward operator $\mathcal{A} : X\to Y$. One option is to consider $\mathcal{R}_\theta$ given by a deep neural network architecture that is obtained by unrolling a suitable iterative scheme, which is originally designed to approximate a regularized inverse of the forward operator, see [5, section 4.9.1] and [46] for further details. These deep neural network architectures have been successfully used in solving large scale inverse problems where there is little training data [2]. To summarize, we use empirical risk minimization to approximate the Bayes estimator in (3):

$$\widehat{\mathcal{R}} \approx \mathcal{R}_{\widehat{\theta}} \quad\text{where}\quad \widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta} \sum_{i=1}^{m} \ell_X\left(x_i,\mathcal{R}_\theta(y_i)\right). \tag{4}$$

In the above, $\mathcal{R}_\theta : Y\to X$ is given by a domain adapted deep neural network architecture that accounts for an explicit forward operator.

The final remark concerns the choice of $\ell_X : X\times X\to\mathbb{R}$ in (4). In imaging one commonly uses the squared $L^2$-distance, so the resulting Bayes estimator becomes the conditional mean (posterior mean) [24]. Other alternatives that are based on image features not expressible as point-wise differences are the Wasserstein distance [3] and perceptual losses [28].

## 3. Post-processing as a decision rule

A wide range of post-processing tasks that act on model parameters in $X$ can be represented by a non-randomized decision rule in a suitably formulated statistical estimation problem.

The underlying statistical model is here $\big((X,\mathfrak{S}_X),\{\mathsf{P}_z\}_{z\in\triangle}\big)$ with $\triangle$ denoting the set parametrizing the probability distributions on $X$. Next, let $(D,\mathfrak{S}_D)$ and $\ell_D : D\times D\to\mathbb{R}$ denote the decision space and loss associated with the post-processing task. Using a Bayes estimator to represent the post-processing task amounts to finding the mapping (post-processing operator) $\widehat{\mathcal{T}} : X \to D$ that solves the following statistical learning problem:

$$\widehat{\mathcal{T}} \in \operatorname*{argmin}_{\mathcal{T}\in\mathscr{D}_{\text{post}}} \mathbb{E}_{(\mathbb{z},\mathbb{x})\sim\eta}\left[\ell_D\left(\tau(\mathbb{z}),\mathcal{T}(\mathbb{x})\right)\right] \quad \text{for some } \eta\in\mathscr{P}_{\triangle\times X}. \tag{5}$$

Here, $\mathscr{D}_{\text{post}}$ is a given suitable class of measurable $D$-valued mappings on $X$ and $\tau : \triangle \to D$ is a given mapping, whereas the joint law $\eta \in \mathscr{P}_{\triangle\times X}$ is usually unknown. This abstract formalization covers a wide range of post-processing tasks, including many relevant for image analysis like classification and segmentation as shown next.

**Remark 2.** It is worth reflecting over the role of the sets $\triangle$ and $D$. Many (complex) tasks can 'naturally' be represented through a hand-crafted mapping $\tau : \triangle \to D$. The issue is that elements in $\triangle$ are not observable, instead observables reside in $X$. The post-processing operator is therefor an estimator $\widehat{\mathcal{T}} : X \to D$ given by the statistical learning problem in (5) while $\tau : \triangle \to D$ remains as the hand-crafted natural representation of the post-processing. The examples in the following sections further clarify these points.

### *3.1. Classification*

The task here is to classify an image into one of $k$ distinct labels, or more precisely, associate an image to a probability distribution over all $k$ labels. This post-processing task can be represented as a statistical estimation problem where the statistical model is $\left((X,\mathfrak{S}_X),\{\mathsf{P}_z\}_{z\in\triangle}\right)$ with $\triangle := \mathbb{Z}_k$, $D := \mathscr{P}_{\triangle}$ (i.e., the decision space is probability distributions over the $k$ labels), and $\tau : \triangle \to D$ as $\tau(z) := \delta_z$ for $z \in \triangle$.

A Bayes estimator is now a mapping $\widehat{\mathcal{T}} : X \to D$ given by (5) where $\mathsf{P}_z \in \mathscr{P}_X$ is the $\triangle$-marginal of $\eta \in \mathscr{P}_{\triangle\times X}$ and $\ell_D : D \times D \to \mathbb{R}$ quantifies similarity between probability measures on $\triangle$. Since elements in $D$ are probability measures on the finite set $\triangle$, one common choice is to use the cross-entropy between probability measures:

$$\ell_D(d, d') := -\sum_{z\in\triangle} d(z) \log\, d'(z) \quad \text{for } d, d' \in D. \tag{6}$$

Implementation. The joint law $(\mathbb{z},\mathbb{x}) \sim \eta \in \mathscr{P}_{\triangle\times X}$ in (5) is unknown, so we approximate it with the empirical measure given by supervised data $(z_i, x_i) \in \triangle \times X$ generated from $(\mathbb{z},\mathbb{x})$. Next, the family of estimators $\mathscr{D}_{\text{post}} = \{\mathcal{T}_\vartheta\}_{\vartheta\in\Xi}$ in (5) is given by a suitable (deep) neural network architecture for classification. An early approach based on a convolutional neural network is in [37], AlexNet [33] and ResNet [23] represent further development along this line. To summarize, the Bayes estimator in (5) is approximated from supervised data by

$$\widehat{\mathcal{T}} \approx \mathcal{T}_{\widehat{\vartheta}} \quad \text{where} \quad \widehat{\vartheta} \in \operatorname*{argmin}_{\vartheta\in\Xi} \left\{ \frac{1}{m}\sum_{i=1}^{m} \left[-\log\left[\mathcal{T}_\vartheta(x_i)(z_i)\right]\right] \right\}. \tag{7}$$

### *3.2. Semantic segmentation*

The aim here is to classify each point in an image into one of $k$ possible labels, so the special case $k = 2$ corresponds to (binary) segmentation. Stated more formally, semantic segmentation applies a mapping that associates each point in an image in $X$ to a probability distribution over all $k$ labels. This post-processing task can be represented as a statistical estimation problem where the statistical model is $\big((X, \mathfrak{S}_X), \{\mathsf{P}_z\}_{z\in\triangle}\big)$ where $\triangle := \mathscr{M}(\Omega, \mathbb{Z}_k)$ is the set of measurable mappings from $\Omega$ to $\mathbb{Z}_k$ and the decision space $D := \mathscr{M}(\Omega, \mathscr{P}_{\mathbb{Z}_k})$ is the set of measurable mappings from $\Omega$ to the class of probability measures on $\mathbb{Z}_k$. Here, $\tau : \triangle \to D$ is given as $\tau(z)(t) := \delta_{z(t)}$ for $z : \Omega \to \mathbb{Z}_k \in \triangle$. Note here that we could have chosen $D = \mathscr{P}_\triangle$ as in classification, but this is intractable for high-dimensional images, so we restrict to a subset of $\mathscr{P}_\triangle$ given by probability distributions for which the label distribution of every pixel is independent.

A Bayes estimator is now a mapping $\widehat{\mathcal{T}} : X \to D$ given by (5) where $\mathsf{P}_z \in \mathscr{P}_X$ is the $\triangle$-marginal of $\eta \in \mathscr{P}_{\triangle\times X}$ and $\ell_D : D \times D \to \mathbb{R}$ quantifies similarity between elements in $D$. Such an element is a function on $\Omega$ that yields a probability measure on the finite set $\mathbb{Z}_k$ at each point $t \in \Omega$. The cross entropy is commonly used for quantifying the similarity between probability distributions, so an option is to define $\ell_D : D \times D \to \mathbb{R}$ as the integral of the point-wise cross entropy of the (point-wise) independent probability measures $d(t), d'(t) \in \mathscr{P}_\triangle$. This translates into

$$\ell_D(d, d') := \int_\Omega \left[ -\sum_{i\in\mathbb{Z}_k} d(t)(i) \log\left[d'(t)(i)\right] \right] \mathrm{d}t \quad \text{for } d, d' : \Omega \to \mathscr{P}_{\mathbb{Z}_k} \in D. \tag{8}$$

Implementation. The joint law $(\mathbb{z}, \mathbb{x}) \sim \eta \in \mathscr{P}_{\triangle\times X}$ in (5) is unknown, so we approximate it with the empirical measure given by supervised data $(z_i, x_i) \in \triangle \times X$ generated from $(\mathbb{z}, \mathbb{x})$. Next, the family of estimators $\mathscr{D}_{\text{post}} = \{\mathcal{T}_\vartheta\}_{\vartheta\in\Xi}$ in (5) is given by a suitable (deep) neural network architecture for (semantic) segmentation, see [21, 40, 48, 60]. In particular, the SegNet architecture has been successful for semantic segmentation of 2D images [6] and for (binary) segmentation one may use the U-net architecture [53]. To summarize, the Bayes estimator in (5) is approximated from supervised data by

$$\widehat{\mathcal{T}} \approx \mathcal{T}_{\widehat{\theta}} \quad \text{where} \quad \widehat{\vartheta} \in \operatorname*{argmin}_{\vartheta\in\Xi} \left\{ \frac{1}{m}\sum_{i=1}^{m} \int_\Omega -\log\left[\mathcal{T}_\vartheta(x_i)(t)\,(z_i(t))\right]\,\mathrm{d}t \right\}. \tag{9}$$

Note that $\mathcal{T}_\vartheta(x)(t)$ in (9) is a probability distribution over $\mathbb{Z}_k$ for any $t \in \Omega$ and $z(t) \in \mathbb{Z}_k$ when $z \in \triangle$, so in particular $\mathcal{T}_\vartheta(x)(t)\,(z(t)) \in [0, 1]$ for any $t \in \Omega$.

### 3.3. Further examples

A common trait with the examples in sections 3.1 and 3.2 is that each of the tasks were represented by a Bayes estimator that was approximated by a deep neural network trained against appropriate supervised training data. Many other image processing tasks that can be handled in this way, like edge detection (find edges and their orientations) [4], de-pixelization/super-resolution (upsample images to increase resolution) [14, 51], in-painting (fill out lost parts of images/videos) [66], de-mosaicing (reconstructing a full colour image from the incomplete colour samples) [59], object recognition (localize and classify objects in images) [17, 23], and non-rigid image registration (deform a template image so that it matches a target image) [7, 18, 67]. More complex tasks from image analysis can also be handled similarly, like colorizing (apply colour to grey scale images) [27], image translation (translate between two classes of images of the same object) [64], and image caption generation (generate text that summarizes content of image) [30, 62].

## 4. Task adapted reconstruction

A straightforward approach to perform reconstruction and post-processing is to compose the mappings for these two steps with each other. This constructs the dotted mapping in (2) where reconstruction and post-processing are performed independently from each other. In contrast, the idea in task adapted reconstruction is to perform reconstruction jointly with post-processing.

An important aspect is to retain the possibility to identify the reconstruction and post-processing steps, i.e., we seek a task adapted reconstruction method of the form

$$\widehat{\mathcal{T}} \circ \widehat{\mathcal{R}} : Y \to D \quad \text{for some } \widehat{\mathcal{R}} : Y \to X \text{ and } \widehat{\mathcal{T}} : X \to D. \tag{10}$$

The proposed framework considers task adapted reconstruction methods of the form in (10). The framework is generic but adaptable in the sense that the forward operator (in the reconstruction step) and the post-processing task are replaceable in a plug-and-play manner. As such, it differs from approaches surveyed in section 1.1 that are tailored to a specific post-processing task (some allow for replacing the forward operator in a plug-and-play manner).

All approaches are based on statistical decision theory where both $\widehat{\mathcal{R}} : Y \to X$ (reconstruction operator) and $\widehat{\mathcal{T}} : X \to D$ (post-processing operator) in (10) are represented as statistical estimators. This requires one to extend the formalism in (1) by assuming that the (unknown) true model parameter $x^* \in X$ and (unknown) true post-processing feature $d^* \in D$ are generated by $X$- and $D$-valued random variables $\mathbb{x}$ and $\mathbb{d} := \tau(\mathbb{z})$, respectively.

In the following sections, we outline different approaches for constructing a task adapted reconstruction method of the form in (10). The implementations consider estimators for post-processing and reconstruction that are parametrized by fixed domain adapted classes $\mathscr{D}_{\text{post}} = \{\mathcal{T}_{\vartheta}\}_{\vartheta \in \Xi}$ and $\mathscr{D}_{\text{rec}} = \{\mathcal{R}_{\theta}\}_{\theta \in \Theta}$, respectively. Hence, task adapted reconstruction is given as in $\widehat{\mathcal{T}} \circ \widehat{\mathcal{R}} : Y \to D$ where

$$\widehat{\mathcal{T}} \approx \mathcal{T}_{\widehat{\vartheta}} : X \to D \quad \text{and} \quad \widehat{\mathcal{R}} \approx \mathcal{R}_{\widehat{\theta}} : Y \to X. \tag{11}$$

In the above, $(\widehat{\theta}, \widehat{\vartheta}) \in \Theta \times \Xi$ are computed from training data. Note that $\widehat{\theta}$ corresponds to using training data to learn a reconstruction method. Likewise, $\widehat{\vartheta}$ corresponds to a learned post-processing method.

### *4.1. Sequential approach*

Here one first determines the reconstruction operator (as a Bayes estimator) independently of the post-processing. This is then used in defining the estimator representing the post-processing, i.e., the post-processing accounts for the reconstruction. Task adapted reconstruction in (10) is then given as

$$\begin{cases} \widehat{\mathcal{R}} \in \underset{\mathcal{R} \in \mathscr{D}_{\text{rec}}}{\operatorname{argmin}} \mathbb{E}_{(\mathbb{x},\mathbb{y}) \sim \mathsf{P}_X \otimes \mathcal{M}} \left[ \ell_X \left( \mathbb{x}, \mathcal{R}(\mathbb{y}) \right) \right] \\ \widehat{\mathcal{T}} \in \underset{\mathcal{T} \in \mathscr{D}_{\text{post}}}{\operatorname{argmin}} \mathbb{E}_{(\mathbb{z},\mathbb{x}) \sim \eta} \left[ \ell_D \left( \tau(\mathbb{z}), \mathcal{T}(\mathbb{x}) \right) | \mathbb{x} = \widehat{\mathcal{R}}(\mathbb{y}) \right] \end{cases} \tag{12}$$

In the above, the joint law $(\mathbb{z}, \mathbb{x}) \sim \eta \in \mathscr{P}_{\triangle \times X}$ needs to be consistent with the joint law $(\mathbb{x}, \mathbb{y}) \sim \mathsf{P}_X \otimes \mathcal{M} \in \mathscr{P}_{X \times Y}$, e.g., by requiring that the $\triangle$-marginal of $\eta \in \mathscr{P}_{\triangle \times X}$, denoted by $\mathsf{P}_z \in \mathscr{P}_X$, is the push forward of $\mathcal{M} \in \mathscr{P}_Y$ through the reconstruction operator $\widehat{\mathcal{R}} : Y \to X$. Alternatively, if one is given a 'task prior' $\mathsf{P}_{\triangle} \in \mathscr{P}_{\triangle}$ so $\eta = \mathsf{P}_{\triangle} \otimes \mathsf{P}_z$, then the requirement is that $\mathsf{P}_X = \mathsf{P}_z$ where $\mathsf{P}_z \in \mathscr{P}_X$ denotes the $\triangle$-marginal of $\eta \in \mathscr{P}_{\triangle \times X}$.

Implementation. The data driven approach for approximating (12) assumes access to coupled supervised training data of the form: $(z_i, x_i, y_i) \in \triangle \times X \times Y$ for $i = 1, \ldots, m$ where $(x_i, y_i)$ and $(z_i, x_i)$ are generated by the marginal distribution $\mathsf{P}_X \otimes \mathcal{M}$ and $\eta$, respectively. One needs to ensure that training data is random samples of distributions that are consistent in the sense stated above. In particular, a post-processing operator trained in $(z_i, x_i)$ is only well defined for input taken from the support of its training data, i.e., it may fail when applied to data it has never seen like when new data has a different statistical assumption.

Task adapted reconstruction is then approximated as in (11) where $(\widehat{\theta}, \widehat{\vartheta}) \in \Theta \times \Xi$ solves

$$\begin{cases} \widehat{\theta} \in \underset{\theta\in\Theta}{\operatorname{argmin}} \left\{ \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \ell_X\left(x_i, \mathcal{R}_\theta(y_i)\right) \right\} \\ \widehat{\vartheta} \in \underset{\vartheta\in\Xi}{\operatorname{argmin}} \left\{ \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \ell_D\left(\tau(z_i), \mathcal{T}_\vartheta(x'_i)\right) \right\} \quad \text{with } x'_i := \mathcal{R}_{\widehat{\theta}}(y_i). \end{cases} \tag{13}$$

Note that $\widehat{\vartheta} \in \Xi$ in (13) is not learned from training data $(z_i, x_i)$. Instead, we make use of training data $(z_i, x'_i)$ where $x'_i := \mathcal{R}_{\widehat{\theta}}(y_i)$. This is a pragmatic way for ensuring that the range of the learned reconstruction operator is contained in the support of the elements in $X$ used for learning the post-processing. In particular, it ensures that the statistical assumptions on training data are consistent.

### *4.2. End-to-end approach*

Here one only uses the post-processing related loss when defining the Bayes estimator for task adapted reconstruction, i.e., one considers (10) with

$$(\widehat{\mathcal{T}}, \widehat{\mathcal{R}}) \in \underset{\substack{\mathcal{T}\in\mathscr{D}_{\text{post}}\\ \mathcal{R}\in\mathscr{D}_{\text{rec}}}}{\operatorname{argmin}} \mathbb{E}_{(\mathbb{y},\mathbb{z})\sim\nu}\left[\ell_D\left(\tau(\mathbb{z}), \mathcal{T}\circ\mathcal{R}(\mathbb{y})\right)\right] \quad \text{for some } \nu \in \mathscr{P}_{Y\times\triangle}. \tag{14}$$

Implementation. One data driven approach is to approximate (14) using supervised training data of the form:

$$(y_i, z_i) \in \triangle \times X \quad \text{generated by} (\mathbb{y}, \mathbb{z}) \sim \nu \text{ for } i = 1, \ldots, m. \tag{15}$$

Task adapted reconstruction is then approximated as in (11) where $(\widehat{\theta}, \widehat{\vartheta}) \in \Theta \times \Xi$ solves

$$(\widehat{\vartheta}, \widehat{\vartheta}) \in \underset{\substack{\vartheta\in\Xi\\ \theta\in\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_D\left(\tau(z_i), \mathcal{T}_\vartheta \circ \mathcal{R}_\theta(y_i)\right) \right\}. \tag{16}$$

### *4.3. Joint approach*

Task adapted reconstruction is here given by (10) with

$$(\widehat{\mathcal{T}}, \widehat{\mathcal{R}}) \in \underset{\substack{\mathcal{T}\in\mathscr{D}_{\text{post}}\\ \mathcal{R}\in\mathscr{D}_{\text{rec}}}}{\operatorname{argmin}} \mathbb{E}_{(\mathbb{x},\mathbb{y},\mathbb{z})\sim\sigma}\left[\ell_{X\times D}\left(\left(\mathbb{x}, \tau(\mathbb{z})\right), \left(\mathcal{R}(\mathbb{y}), \mathcal{T}\circ\mathcal{R}(\mathbb{y})\right)\right)\right] \tag{17}$$

for some $(\mathbb{z}, \mathbb{x}, \mathbb{y}) \sim \sigma \in \mathscr{P}_{X\times Y\times\triangle}$ and $\ell_{X\times D} : (X \times D) \times (X \times D) \to \mathbb{R}$. This defines a Bayes estimator for $(\mathbb{z}, \mathbb{x} \mid \mathbb{y} = y)$ in the joint statistical estimation (inverse problem) that is obtained by assuming that $x^* \in X$ and $d^* \in D$ in (1) are generated by random variables $\mathbb{x}$ and $\mathbb{d} := \tau(\mathbb{z})$.

Of particular interest is to choose

$$\ell_{X\times D}\left((x,d),(x',d')\right) := (1-C)\ell_D(d,d') + C\ell_X(x,x') \quad \text{for fixed } 0 \leqslant C \leqslant 1, \tag{18}$$

where $\ell_X : X \times X \to \mathbb{R}$ and $\ell_D : D \times D \to \mathbb{R}$ are given. This represents a 'middle-way' between the sequential and end-to-end approaches. To see this, start by decomposing the measure $\sigma \in \sigma \in \mathscr{P}_{X\times Y\times\triangle}$ in terms of conditional probabilities using the chain rule: $\mathrm{d}\sigma(z,x,y) = \mathrm{d}\mathcal{M}(x)(y)\,\mathrm{d}\eta(z,x)$. This follows by inserting $(\mathbb{z},\mathbb{x}) \sim \eta$ and $(\mathbb{y}\,|\,\mathbb{x}=x) \sim \mathcal{M}(x)$ and making use of the additional assumption that $\mathrm{dP}(y\,|\,z,x) = \mathrm{dP}(y\,|\,x)$ (which holds if $\mathbb{x}$ is a sufficient statistic for $\mathbb{y}$). In the limit $C \to 1$, the joint approach reduces to the sequential one.

Note also that it may seem sufficient to only consider the loss $\ell_D$ in (17), i.e., to set $C = 0$ in (18). Then (17) becomes equivalent to the end-to-end approach in (14). There is however an issue problem with non-uniqueness in this case since the loss does not consider reconstruction space at all, so any model parameter can be reconstructed as long as the downstream post-processing operator is adopted appropriately. Stated formally,

$$(\widehat{\mathcal{R}},\widehat{\mathcal{T}})\,\text{solves}\,(17) \;\Rightarrow\; (\mathcal{B}^{-1}\circ\widehat{\mathcal{R}},\widehat{\mathcal{T}}\circ\mathcal{B})\,\text{solves}\,(17)\;\text{for any invertible}\;\mathcal{B}: X\to X.$$

This non-uniqueness does not arise when $C > 0$, so incorporating a loss term associated with the reconstruction may act as a regularizer. This also indicates that the limit $C \to 0$ in (17) does not necessarily coincide with the case $C = 0$ in (17).

Implementation. The data driven approach for approximating (17) is based on supervised training data that jointly involves the reconstruction and task:

$$(x_i, y_i, z_i) \in X \times Y \times \triangle \quad \text{generated by}\,(\mathbb{x},\mathbb{y},\mathbb{z}) \sim \sigma \;\text{for}\; i = 1,\ldots,m. \tag{19}$$

Task adapted reconstruction is then approximated as in (11) where $(\widehat{\theta},\widehat{\vartheta}) \in \Theta \times \Xi$ solves

$$(\widehat{\theta},\widehat{\vartheta}) \in \underset{(\theta,\vartheta)\in\Theta\times\Xi}{\operatorname{argmin}} \left\{\frac{1}{m}\sum_{i=1}^{m}\ell_{X\times D}\left((x_i,\tau(z_i)),\left(\mathcal{R}_\theta(\mathbb{y}_i),\mathcal{T}_\vartheta\circ\mathcal{R}_\theta(\mathbb{y}_i)\right)\right)\right\} \tag{20}$$

where $\ell_{X\times D} : (X \times D) \times (X \times D) \to \mathbb{R}$ is the joint loss in (18). Note that one may *in addition* have access to separate sets of training data generated by $(\mathbb{x},\mathbb{y}) \sim \mathsf{P}_X \otimes \mathcal{M}$ and $(\mathbb{z},\mathbb{x}) \sim \eta$. In such case, it is possible to first *pre-train* by approximately solving (5) and (4) separately, and use the resulting outcomes to initialize an algorithm for solving (20).

## 5. Applications

In the following we demonstrate performance of the task adapted reconstruction schemes outlined in section 4. All cases involve reconstruction from tomographic 2D parallel beam data, so elements in $X$ are compactly supported real-valued functions representing 2D images. As post-processing tasks, we consider MNIST classification and segmentation.

### 5.1. Joint tomographic reconstruction and classification

**Post-processing task**. Recover probabilities that a 2D grey scale MNIST image is a `0`, `1`, ..., `9` from noisy parallel beam tomographic data (see section 3.1).

**Data**. Elements in $Y$ are real-valued functions representing samples of a Poisson random variable with mean equal to the exponential of the parallel beam ray transform and an intensity corresponding to 60 photons/line. The ray transform is digitized by sampling the angular variable at 5 uniformly sampled points in $[0,\pi]$ with 25 lines/angle.

**Table 1.** In both the pre-training and sequential approaches, the reconstruction and post-processing operators are trained separately. In the sequential approach, the post-processing operator is then further trained on the output of the trained reconstruction operator. In the end-to-end approach ($C = 0$ in (18)) the reconstruction operator is pre-trained with $L^2$-loss. Finally, the joint approach uses the full loss (18). We see that the classification accuracy (explained in 'decision space' in section 5.1) improves when using a joint approach. In fact, using a 'suitable' $C$ (figure 2(a)) yields an accuracy of 97.00% that is quite close to the upper limit of 97.76%, which is the accuracy of the classifier when trained on true images.

| Approach | Accuracy | $L^2$-loss | Cross entropy |
|---|---|---|---|
| Pre-training | 93.61% | 9.0 | 0.643 |
| Sequential | 96.01% | 9.0 | 0.124 |
| End-to-end | 96.70% | 19.7 | 0.118 |
| Joint with $C = 0.01$ | 96.74% | 12.8 | 0.108 |
| Joint with $C = 0.5$ | 97.00% | 9.2 | 0.100 |
| Joint with $C = 0.999$ | 96.61% | 9.0 | 0.108 |
| Classification on true images | 97.76% | | 0.088 |

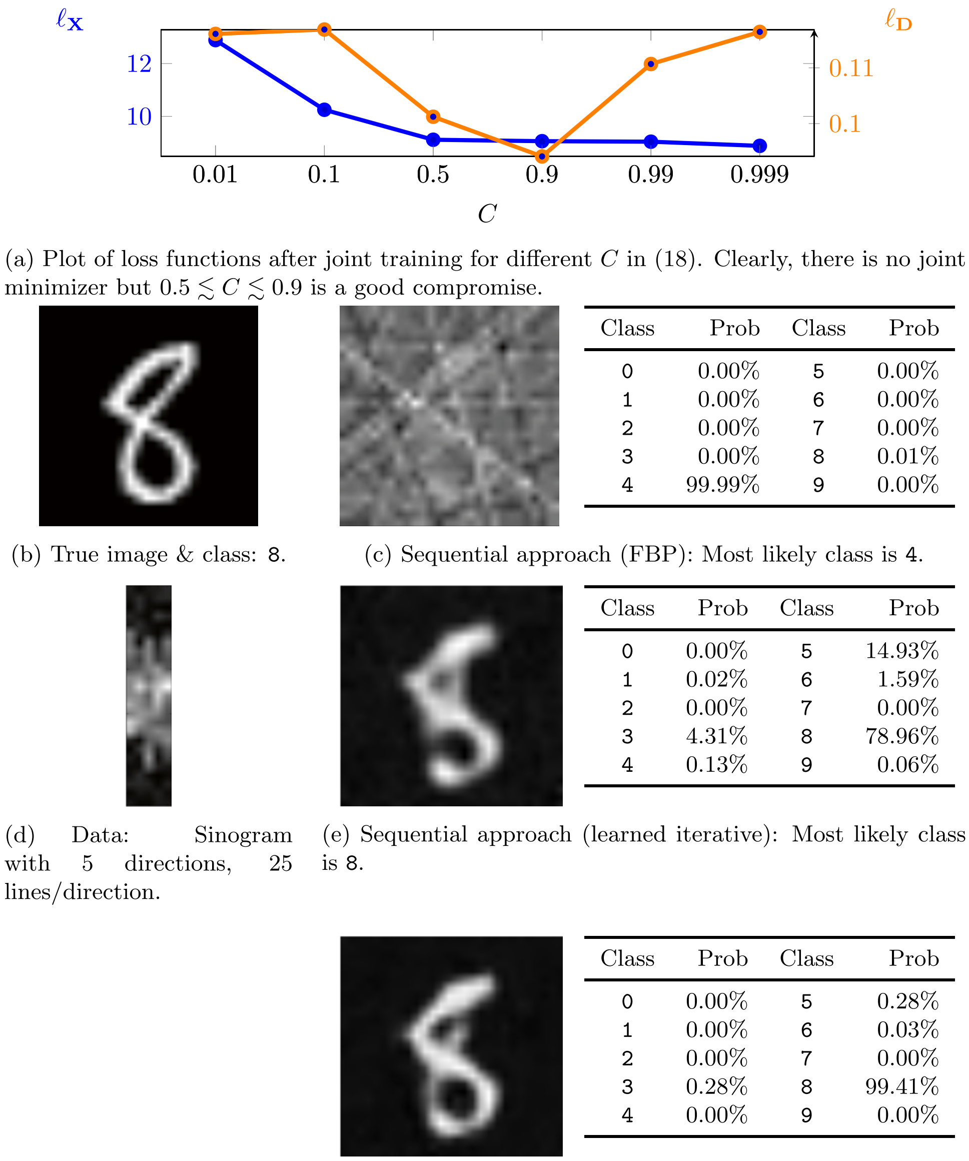
**Model parameter space**. Elements in $X$ are functions representing images supported on a fixed rectangular region $\Omega \subset \mathbb{R}^2$, so $X := L^2(\Omega, \mathbb{R})$. These are discretized by sampling on a uniform $28 \times 28$ grid. The loss $\ell_X : X \times X \to \mathbb{R}$ is the squared $L^2$-distance on $X$.

**Decision space**. $\triangle := \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{9}\}$ is the set of labels and $D$ is probability distributions over $\triangle$ with a loss function $\ell_D : D \times D \to \mathbb{R}$ given by the cross entropy in (6). In addition to cross entropy, we employ *classification accuracy* to measure performance. Given a probability distribution $d \in D$ over $\triangle$, the single label prediction is defined to be the element in $\triangle$ that is assigned the highest probability, i.e. $\arg\max_{z\in\triangle} d(z)$. The percentage of images in the evaluation data set for which the predicted label coincides with the real one is reported as classification accuracy.

**Reconstruction and post-processing operators**. Reconstruction $\mathcal{R}_\theta : Y \to X$ is given by an unrolled gradient descent scheme [1] and post-processing $\mathcal{T}_\vartheta : X \to D$ is a MNIST classifier given by a standard convolutional neural net classifier with three convolutional layers, each followed by $2 \times 2$ max pooling for segmentation. The activation functions used were ReLUs, layers had 32, 64 and 128 channels, respectively. The final layer is dense and transforms the output of the last convolutional layer to a logit layer of size 10, with the last activation function being a softmax.

**Joint training**. Joint supervised data is given as 512 000 triplets $(x_i, y_i, z_i)$ where $z_i \in \triangle$ is the label corresponding to the MNIST labels. We also performed pre-training for both the reconstruction and post-processing operator (classifier). The reconstruction operator was pre-trained using 256 000 pairs $(x_i, y_i)$ with 8000 steps with a batch size of 64 and the post-processing operator (classifier) was pre-trained until 97.7% accuracy. Note here that we use about 60 000 entries from the MNIST database, so the above supervised data is not statistically independent.

Example outcomes, which are summarized in table 1 and figure 2, show that a joint approach outperforms a sequential one when considering the classification accuracy. Besides an improved classification accuracy we also see a significant improvement regarding interpretability. The reconstructed image part can in the joint setting actually be used as a benchmark to assess the reconstructed classification probabilities. On the other hand, the sequential approach results in classification probabilities that deviates from this intuitive observation. We also see



(a) Plot of loss functions after joint training for different $C$ in (18). Clearly, there is no joint minimizer but $0.5 \lesssim C \lesssim 0.9$ is a good compromise.

| Class | Prob | Class | Prob |
|---|---|---|---|
| 0 | 0.00% | 5 | 0.00% |
| 1 | 0.00% | 6 | 0.00% |
| 2 | 0.00% | 7 | 0.00% |
| 3 | 0.00% | 8 | 0.01% |
| 4 | 99.99% | 9 | 0.00% |

(b) True image & class: 8.

(c) Sequential approach (FBP): Most likely class is 4.

| Class | Prob | Class | Prob |
|---|---|---|---|
| 0 | 0.00% | 5 | 14.93% |
| 1 | 0.02% | 6 | 1.59% |
| 2 | 0.00% | 7 | 0.00% |
| 3 | 4.31% | 8 | 78.96% |
| 4 | 0.13% | 9 | 0.06% |

(d) Data: Sinogram with 5 directions, 25 lines/direction.

(e) Sequential approach (learned iterative): Most likely class is 8.

| Class | Prob | Class | Prob |
|---|---|---|---|
| 0 | 0.00% | 5 | 0.28% |
| 1 | 0.00% | 6 | 0.03% |
| 2 | 0.00% | 7 | 0.00% |
| 3 | 0.28% | 8 | 99.41% |
| 4 | 0.00% | 9 | 0.00% |

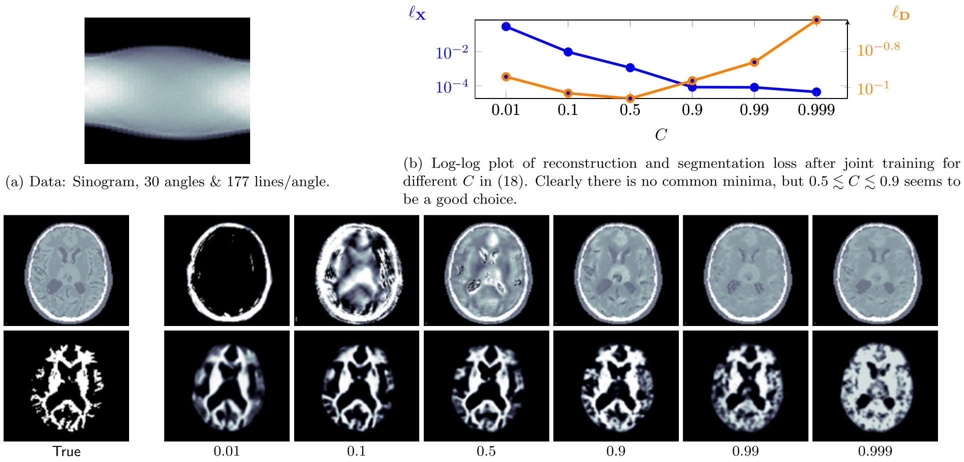
(f) Joint approach with $C = 0.5$. Most likely class is 8.

**Figure 2.** Joint tomographic reconstruction and classification of MNIST images described in section 5.1. Training data is to the left and reconstructed image with classification probabilities are on the right. Table 1 summarizes the overall performance (accuracy) on the entire MNIST dataset.

that in both cases, the classification probabilities are unnaturally concentrated on a single label, but this is a know phenomena also for regular for MNIST classification [20].

### *5.2. Joint tomographic reconstruction and segmentation*

**Post-processing task**. Recover the probability map for segmentation of a grey scale image (see section 3.2 with $k = 2$) from noisy parallel beam tomographic data. In this specific example,



(a) Data: Sinogram, 30 angles & 177 lines/angle.

(b) Log-log plot of reconstruction and segmentation loss after joint training for different $C$ in (18). Clearly there is no common minima, but $0.5 \lesssim C \lesssim 0.9$ seems to be a good choice.

(c) The segmentation is a normalized grey-scale image denoting the probability that a point belongs to the segmented structure. The choice $C = 0.9$ seems to be a good compromise for a good reconstruction *and* segmentation (see (b)). Note also that $C \to 1$ gives the sequential approach, so $C = 0.999$ may serve as a proxy for it. Reconstructions take a few milliseconds on a desktop workstation.

**Figure 3.** Joint tomographic reconstruction and segmentation as described in section 5.2 for different values of $C$ in (18).

we consider segmenting the grey matter of CT images of the brain, which is relevant in imaging of neurodegerantive diseases like Alzheimers' disease.

**Data**. $Y$ contains real-valued functions on lines representing parallel beam tomographic data.These are digitized by sampling the angular variable uniformly at 30 points in $[0, \pi]$ with 183 lines/angle. We also add 0.1% additive Gaussian noise to data.

**Model parameter space**. $X$ contains functions representing images supported on a fixed rectangular region $\Omega \subset \mathbb{R}^2$, so $X := L^2(\Omega, \mathbb{R})$. These are discretized by uniform sampling on a $128 \times 128$ grid. The loss $\ell_X : X \times X \to \mathbb{R}$ is the squared $L^2$-distance.

**Decision space**. Elements in $D$ are point-wise probability distributions over binary images on $\Omega$, which can be represented by grey-scale images with values in [0, 1] that gives the probability that a point is part of the segmented object. Hence, $D = \mathscr{M}(\Omega, [0, 1])$ with the loss function $\ell_D : D \times D \to \mathbb{R}$ as in (8).

**Reconstruction and post-processing operators**. $\mathcal{R}_\theta : Y \to X$ is given by unrolling a primal-dual scheme [2] and $\mathcal{T}_\vartheta : X \to D$ is given by an 'off the shelf' U-net convolutional neural net for segmentation [53].

**Joint training**. Joint supervised data is given as 100 triplets $(x_i, y_i, d_i)$ where $d_i$ is the segmentation (binary image). We extend joint training data by data argumentation (±5 pixel translation and ±10° rotation). There was no pre-training.

Figure 3 summarizes example outcome, note that image in figure 3(c) for $C = 0.99$ can be seen as the outcome from a sequential approach since $C = 1$ corresponds to the sequential approach. Clearly, a joint approach with $C$ between 0.1 and 0.5 outperforms a sequential one.

Next, as $C$ decreases the reconstruction becomes more adapted to the segmentation task. In the limit $C \to 0$ this post-processed image is viable but the reconstruction part is useless, which is to be expected. In the other direction, as $C$ decreases the reconstructed part becomes less adapted to the post-processing task and the latter becomes increasingly challenging due to the low contrast between white and grey matter.

Finally, using $C > 0$ not only reduces the non-uniqueness as explained in (17), it further regularizes in the sense that information from the reconstruction guides the segmentation, which otherwise would amount to learning the segmented image directly from the noisy sinogams. Intuitively there seems to be an 'information exchange' between the task of reconstruction and that of segmentation, which when properly balanced by choosing $C$ acts as a regularizer for the segmentation, e.g., the white/grey matter contrast in the reconstruction is overemphasized for small $C$. This improves the interpretability since it shows how the reconstructed image 'helps' in interpreting why a certain segmentation is taken.

## 6. Theoretical considerations

The task adapted reconstruction in (17) is a Bayes estimator, so it is natural to investigate whether theory of Bayesian inversion [15, 29, 47] can be used to analyse its regularizing properties. This theory however does not apply since results are based on too many restrictive assumptions.

Another line of investigation considers the potential advantage that comes with using a joint approach over a sequential one. Since the reconstruction and post-processing operators are trained separately in a sequential approach, some information is inevitably lost when applying a regularized reconstruction operator. In contrast, the joint approach involves simultaneously training both reconstruction and post-processing operators, so there is a better chance of preserving the information. We therefore expect a joint approach to perform better. This is also supported by the observation in (17) and the empirical evidence in section 5.

Albeit convincing, the above heuristic argument is flawed! As stated by proposition 1, it is surprisingly difficult to prove that a joint approach outperforms a sequential one in a non-parametric setting where one has access to all of data. The reason is that many standard operators from data space $Y$ to model parameter space $X$ are formally information conserving. The adjoint of a (linear) forward operator, its Moore–Penrose pseudo-inverse, and even some regularized reconstruction operators like the usual Tikhonov regularization are information conserving under standard Gaussian noise. Another example in planar tomography is the FBP reconstruction operator (with a filter that is strictly non-zero in frequency space). The next results provides the theoretical foundation for the above claims.

**Proposition 1.** *Let $\mathbb{x}$ and $\mathbb{y}$ be a $X$- and $Y$-valued random variables that are both defined on the same probability space. Next, $\Pi : Y \to Y$ is a measurable operator with closed range and $\mathcal{B}$ is an arbitrary measurable map on $Y$ that is injective when restricted to ran$(\Pi)$. Then*

$$\mathbb{E}\left[f(\mathbb{x})|\mathbb{y}\right] = \mathbb{E}\left[f(\mathbb{x})|\mathcal{B}(\Pi\mathbb{y})\right] \text{ for } f : X \to \mathbb{R} \text{ measurable} \iff \mathbb{x} \perp\!\!\!\perp (\mathbb{y} - \Pi\mathbb{y})|\Pi\mathbb{y}. \tag{21}$$

The proof follows directly from inserting the definitions. When applied to an inverse problem, $\Pi$ typically represents an orthogonal projection onto the closure of the range of $\mathcal{A}$. From proposition 1 we get that $(\mathbb{x} \mid \mathbb{y})$ and $\left(\mathbb{x}|\mathcal{B}(\Pi\mathbb{y})\right)$ have the same distribution if and only if, given the knowledge of $\Pi\mathbb{y}$, the 'noise' $\mathbb{y} - \Pi\mathbb{y}$ in the null space of $\Pi$ is independent of $\mathbb{x}$.

**Corollary 1.** *Consider the setting in section 4 for task adapted reconstruction and assume in particular that $\mathbb{y}$ and $\mathbb{z}$ are conditionally independent given $\mathbb{x}$. Finally, let $\mathcal{B}$ satisfy the assumptions in proposition 1; we also assume that the equality in proposition 1 holds, that is $\pi(x \mid y) = \pi(x \mid \mathcal{B}(y))$. Then, $\mathsf{P}(z \mid y) = \mathsf{P}\left(z \mid \mathcal{B}(y)\right)$.*

**Proof.** The conditional independence assumption can be written as $\pi(z \mid x, y) = \pi(z \mid x)$. Using this, we compute $\mathsf{P}(x, y, z) = \mathsf{P}(z \mid x, y)\mathsf{P}(x, y) = \mathsf{P}(z \mid x)\mathsf{P}(x, y)$, which yields

$$\mathsf{P}(x, z \mid y) = \mathsf{P}(z \mid x)\mathsf{P}(x \mid y). \tag{22}$$

Notice that $\mathcal{B}(\mathbb{y})$ and $\mathbb{z}$ are also conditionally independent given $\mathbb{x}$, so we similarly obtain

$$\mathsf{P}(x, z \mid \mathcal{B}(y)) = \mathsf{P}(z \mid x)\mathsf{P}(x \mid \mathcal{B}(y)). \tag{23}$$

Now, (22) and (23) imply in particular that

$$\begin{aligned} \mathsf{P}(z \mid y) &= \int \mathsf{P}(z, x \mid y)\,\mathrm{d}x = \int \mathsf{P}(z \mid x)\mathsf{P}(x \mid y)\,\mathrm{d}x \\ \mathsf{P}\left(z \mid \mathcal{B}(y)\right) &= \int \mathsf{P}\left(z, x \mid \mathcal{B}(y)\right)\,\mathrm{d}x = \int \mathsf{P}(z \mid x)\mathsf{P}\left(x \mid \mathcal{B}(y)\right)\,\mathrm{d}x. \end{aligned} \tag{24}$$

Our assumption is that $\mathsf{P}(x \mid y) = \mathsf{P}\left(x \mid \mathcal{B}(y)\right)$, which combined with (24) proves the claim. □

Corollary 1 implies that $(\mathbb{z} \mid \mathbb{y} = y)$ and $(\mathbb{z} \mid \mathbb{x} = \mathcal{B}(y))$ are equally distributed as $\triangle$-valued random variables. In particular, a task adapted reconstruction method (sequential or joint) is equivalent to first performing reconstruction by some $\mathcal{B} : Y \to X$ followed by the learned measurable map $\mathcal{C} : X \to D$ that is given as $\mathcal{C} := \mathcal{T} \circ \mathcal{R} \circ \mathcal{B}^{-1}$. Here, $\mathcal{C}$ defines a non-randomized decision rule that in principle serves as a 'task' operator.

To conclude, theoretical arguments for proving superiority of the joint approach over a post-processing approach cannot be based on a 'information bottleneck' type of argument. Hence, the reason must be related to either the choice of architecture or the training protocol. Unfortunately, both these are related to the central open problem, namely why deep learning 'works'. Another argument in favour of a joint approach is that it is highly non-trivial to select an appropriate architecture for parametrizing $\mathcal{C}$, whereas $\mathcal{T}$ and $\mathcal{R}$ are easier to parametrize by means of neural networks. Furthermore, evaluating $\mathcal{B}$ or its inverse $\mathcal{B}^{-1}$ may not be stable. Finally, as we have seen from the examples, using knowledge about the reconstruction may in fact act as a regularizer, either by improving the trainability or the generalization properties.

## 7. Discussion and outlook

A key aspect for the implementation of the joint task adapted reconstruction method in (17) is that both reconstruction and post-processing are given by trainable neural networks, which after joint training forms a single intertwined neural network.

As argued for in section 3, a wide range of post-processing tasks can be represented by an Bayes estimator. The abstract framework for task adapted reconstruction (section 4) works with *any* task that can be represented by a neural network as long as the parametrization and the loss functions are differentiable, like those listed in section 3. Hence, our approach opens up for *truly task adapted reconstruction that goes well beyond performing reconstruction jointly with simple feature extraction*. In particular, more advanced post-processing tasks, like image caption generation or image-processing steps in radiomics can be performed jointly with reconstruction. The potential for task adapted reconstruction in radiomics is also emphasized in [63], which introduces the term *rawdiomics* (on p 1294).

An important advantage that comes with a joint approach is increased robustness. Advanced post-processing, like radiomics, commonly rely on deep neural networks that are trained on images in a supervised setting. Such images are however computed from measured data, so contrast and texture may depend on the instrumentation used to acquiring the data and the reconstruction method used for computing the images. Hence, a neural network that has been trained against images acquired from a particular equipment, or obtained using a particular reconstruction method, may generalize poorly when either of these factors change. This is especially the case for tasks involving elements of visual classification, such as semantic segmentation, that can be sensitive to variations in texture and contrast. In contrast, task adapted reconstruction acts on measured data instead of images (model parameters). Using a reconstruction step that incorporates a physics guided model for how measured data is generated results in a joint approach that is more robust against variations in how data is acquired. This is e.g. essential if radiomics is to be part of clinical-decision support systems for improving diagnostic, prognostic, and predictive accuracy.

Another advantage relates to computationally feasibility. Scalability is a serious issue with the model based approaches in section 1.1, e.g., state-of-the-art methods for joint reconstruction and segmentation quickly become computationally infeasible. In contrast, once trained, neural networks for task adapted reconstruction scale to large scale problems. The examples in figure 3 for joint reconstruction and segmentation obtained by using section 5.2 take a few milliseconds on a desktop gaming PC.

Finally, examples involving tomographic image reconstruction (section 5) support the claim that a joint approach outperforms a sequential one. Understand this theoretically (section 6) is however an open problem. In particular, there is currently no theory motivating using a joint loss of the type in (18), even though empirical evidence suggests such a choice outperforms the end-to-end and sequential approaches.

## Acknowledgments

The work by Jonas Adler, Olivier Verdier, and Ozan Öktem has been supported by the Swedish Foundation for Strategic Research grant AM13-0049, Industrial PhD grant ID14-0055, and Elekta AB Carola-Bibiane Schönlieb and Sebastian Lunz acknowledge support from the Engineering and Physical Sciences Research Council (EPSRC) 'EP/K009745/1', the EPSRC Grant 'EP/M00483X/1', the EPSRC centre 'EP/N014588/1', the Leverhulme Trust project 'Breaking the non-convexity barrier', the CHiPS (Horizon 2020 RISE project grant), the Cantab Capital Institute for the Mathematics of Information, and the Alan Turing Institute 'TU/B/000071'.

## Data availability statement

The data that support the findings of this study will be openly available following an embargo at the following URL/DOI: https://repository.cam.ac.uk. Data will be available from 01 August 2021.

## ORCID iDs

Jonas Adler https://orcid.org/0000-0001-9928-3407
Carola-Bibiane Schönlieb https://orcid.org/0000-0003-0099-6306
Ozan Öktem https://orcid.org/0000-0002-1118-6483

## References

[1] Adler J and Öktem O 2017 Solving ill-posed inverse problems using iterative deep neural networks *Inverse Problems* **33** 124007

[2] Adler J and Öktem O 2018 Learned primal-dual reconstruction *IEEE Trans. Med. Imaging* **37** 1322–32

[3] Adler J, Ringh A, Öktem O and Karlsson J 2017 Learning to solve inverse problems using Wasserstein loss (arXiv:1710.10898)

[4] Andrade-Loarca H, Kutyniok G, Öktem O and Petersen P C 2019 Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks *SIAM J. Imaging Sci.* **12** 1936–66

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

[6] Badrinarayanan V, Kendall A and Cipolla. R 2017 SegNet: a deep convolutional encoder-decoder architecture for image segmentation *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 2481–95

[7] Balakrishnan G, Zhao A, Sabuncu M R, Guttag J and Dalca A V 2019 VoxelMorph: a learning framework for deformable medical image registration *IEEE Trans. Med. Imaging* **38** 1788–800

[8] Benning M and Burger M 2018 Modern regularization methods for inverse problems *Acta Numer.* **27** 1–111

[9] Boink Y E, Manohar S and Brune C 2019 A partially-learned algorithm for joint photo-acoustic reconstruction and segmentation *IEEE Trans. Med. Imaging* **39** 129–39

[10] Burger M, Dirks H and Schönlieb C-B 2018 A variational model for joint motion estimation and image reconstruction *SIAM J. Imaging Sci.* **11** 94–128

[11] Burger M, Rossmanith C and Zhang X 2016 Simultaneous reconstruction and segmentation for dynamic spect imaging *Inverse Problems* **32** 104002

[12] Chen C, Gris B and Öktem O 2019 A new variational model for joint image reconstruction and motion estimation in spatiotemporal imaging *SIAM J. Imaging Sci.* **12** 1686–719

[13] Chen C and Öktem O 2018 Indirect image registration with large diffeomorphic deformations *SIAM J. Imaging Sci.* **11** 575–617

[14] Dahl R, Norouzi M and Shlens J 2017 Pixel recursive super resolution *2017 IEEE Int. Conf. Computer Vision ICCV*

[15] Dashti M and Stuart A M 2017 The Bayesian approach to inverse problems *Handbook of Uncertainty Quantification* ed R Ghanem, D Higdon and H Owhadi (New York: Springer) chapter 10

[16] Evans S N and Stark P B 2002 Inverse problems as statistics *Inverse Problems* **18** R1–R55

[17] Farabet C, Couprie C, Najman L and LeCun Y 2013 Learning hierarchical features for scene labeling *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1915–29

[18] Ghosal S and Ray N 2017 Deep deformable registration: enhancing accuracy by fully convolutional neural net *Pattern Recognit. Lett.* **94** 81–6

[19] Gris B, Chen C and Öktem O 2020 Image reconstruction through metamorphosis *Inverse Problems* **36** 025001

[20] Guo C, Pleiss G, Sun Y and Weinberger K O 2017 On calibration of modern neural networks *Proc. 34th Int. Conf. on Machine Learning PMLR 70*

[21] Guo Y, Liu Y, Georgiou T and Lew M S 2018 A review of semantic segmentation using deep neural networks *Int. J. Multimed. Inf. Retr.* **7** 87–93

[22] Hauptmann A, Öktem O and Schönlieb C-B 2020 Image reconstruction in dynamic inverse problems with temporal models *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging* ed K Chen, C-B Schönlieb, X-C Tai and L Younes (Berlin: Springer)

[23] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *2016 IEEE Conf. Computer Vision and Pattern Recognition CVPR* pp 770–8

[24] Helin T and Burger M 2015 Maximum *a posteriori* probability estimates in infinite-dimensional Bayesian inverse problems *Inverse Problems* **31** 085009

[25] Hohage T and Werner F 2016 Inverse problems with Poisson data: statistical regularization theory, applications and algorithms *Inverse Problems* **32** 093001

[26] Hohm K, Storath M and Weinmann A 2015 An algorithmic framework for Mumford–Shah regularization of inverse problems in imaging *Inverse Problems* **31** 115011

[27] Iizuka S, Simo-Serra E and Ishikawa H 2016 Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification *ACM Trans. Graph.* **35** Proceedings of ACM SIGGRAPH 2016 1–11

[28] Johnson J, Alahi A and Fei-Fei L 2016 Perceptual losses for real-time style transfer and super-resolution *14th European Conf. Computer Vision (ECCV 2016)* (*Lecture Notes in Computer Science* vol 9906) ed B Leibe, J Matas, N Sebe and M Welling (Berlin: Springer) pp 694–711

[29] Kaipio J P and Somersalo E 2005 *Statistical and Computational Inverse Problems* (*Applied Mathematical Sciences* vol 160) (Berlin: Springer)

[30] Karpathy A and Fei-Fei L 2017 Deep visual-semantic alignments for generating image descriptions *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 664–76

[31] Klann E, Ramlau R, Ramlau R and Ring W 2011 A Mumford–Shah level-set approach for the inversion and segmentation of SPECT/CT data *Inverse Problems Imaging* **5** 137–66

[32] Krishnan V P and Quinto E T 2015 Microlocal analysis in tomography *Handbook of Mathematical Methods in Imaging* 2nd edn ed O Scherzer (Berlin: Springer) pp 847–902

[33] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *25th Conf. on Neural Information Processing Systems (NIPS 2012)* pp 1097–105

[34] Kutyniok G and Labate D (ed) 2012 *Shearlets: Multiscale Analysis for Multivariate Data* (Berlin: Springer)

[35] Lambin P *et al* 2017 Radiomics: the bridge between medical imaging and personalized medicine *Nat. Rev. Clin. Oncol.* **14**

[36] Lang L F, Neumayer S, Öktem O and Schönlieb C-B 2019 Template-based image reconstruction from sparse tomographic data *Appl. Math. Optim.* **82** 1081–109

[37] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324

[38] Liese F and Miescke K-J 2008 *Statistical Decision Theory: Estimation, Testing, and Selection* (*Springer Series in Statistics*) (New York: Springer)

[39] Liu J, Aviles-Rivero A I, Ji H and Schönlieb C-B 2019 Rethinking medical image reconstruction via shape prior, going deeper and faster: deep joint indirect registration and reconstruction (arXiv:1912.07648)

[40] Long J, Shelhamer E and Darrell T 2015 Fully convolutional networks for semantic segmentation *IEEE Conf. Computer Vision and Pattern Recognition CVPR* pp 3431–40

[41] Louis A K 2011 Feature reconstruction in inverse problems *Inverse Problems* **27** 065010

[42] Lucka F, Huynh N, Betcke M, Zhang E, Beard P, Cox B and Arridge S 2018 Enhancing compressed sensing 4D photoacoustic tomography by simultaneous motion estimation *SIAM J. Imaging Sci.* **11** 2224–53

[43] Lunz S, Öktem O and Schönlieb C-B 2018 Adversarial regularizers in inverse problems *32nd Conf. Neural Information Processing Systems NeurIPS*

[44] Mataev G, Elad M and DeepRED P M 2019 Deep image prior powered by RED *ICCV 2019 Workshop on Learning for Computational Imaging*

[45] Mohammad-Djafari A 2009 Gauss–Markov–Potts priors for images in computer tomography resulting to joint optimal reconstruction and segmentation *Int. J. Tomogr. Stat.* **11** 76–92 http://www.ceser.in/ceserp/index.php/ijts/article/view/139

[46] Monga V, Li Y and Eldar Y C 2019 Algorithm unrolling: interpretable, efficient deep learning for signal and image processing (arXiv:1912.10557)

[47] Nickl R 2017 On Bayesian inference for some statistical inverse problems with partial differential equations *Bernoulli News* **24** 5–9 http://www.bernoulli-society.org/files/BernoulliNews2017-2.pdf

[48] Noh H, Hong S and Han B 2015 Learning deconvolution network for semantic segmentation *2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* pp 1520–8

[49] Pouchol C, Verdier O and Öktem O 2019 Spatiotemporal PET reconstruction using ML-EM with learned diffeomorphic deformation *Machine Learning for Medical Image Reconstructio (MLMIR 2019)* (*Lecture Notes in Computer Science* vol 11905) ed F Knoll, A Maier, D Rueckert and J C Ye (Berlin: Springer) pp 151–62

[50] Ramlau R and Ring W 2007 A Mumford–Shah level-set approach for the inversion and segmentation of x-ray tomography data *J. Comput. Phys.* **221** 539–57

[51] Romano Y, Isidoro J and Milanfar P 2017 RAISR: rapid and accurate image super resolution *IEEE Trans. Comput. Imaging* **3** 110–25

[52] Romanov M, Dahl A B, Dong Y and Hansen P C 2016 Simultaneous tomographic reconstruction and segmentation with class priors *Inverse Problems Sci. Eng.* **24** 1432–53

[53] Ronneberger O, Fischer P, Brox T and U-Net 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer Assisted Intervention (MICCAI 2015)* (*Lecture Notes in Computer Science* vol 9351) ed N Navab, J Hornegger, W M Wells and A F Frangi (Berlin: Springer) pp 234–41

[54] Rubin G D 2014 Computed tomography: revolutionizing the practice of medicine for 40 years *Radiology* **273** 45–74

[55] Rubinstein R, Bruckstein A M and Elad M 2010 Dictionaries for sparse representation modeling *Proc. IEEE* **98** 1045–57

[56] Schmoderer T, Aviles-Rivero A I, Corona V, Debroux N and Schönlieb C-B 2020 Learning optical flow for fast MRI reconstruction (arXiv:2004.10464)

[57] Storath M, Weinmann A, Frikel J and Unser M 2015 Joint image reconstruction and segmentation using the Potts model *Inverse Problems* **31** 025003

[58] Sun L, Fan Z, Huang Y, Ding X and Paisley J 2018 Joint CS-MRI reconstruction and segmentation with a unified deep network *Information Processing in Medical Imaging (IPMI 2019)* (*Lecture Notes in Computer Science* vol 11492) pp 492–504

[59] Syu N-S, Chen Y-S and Chuang Y-Y 2018 Learning deep convolutional networks for demosaicing (arXiv:1802.03769)

[60] Thoma M 2016 A survey of semantic segmentation (arXiv:1602.06541)

[61] van Timmeren J E, Cester D, Tanadini-Lang S, Alkadhi H and Baessler B 2020 Radiomics in medical imaging-'how-to' guide and critical reflection *Insights Imaging* **11** 91

[62] Vinyals O, Toshev A, Bengio S and Erhan D 2015 Show and tell: a neural image caption generator *IEEE Conf. Computer Vision and Pattern Recognition* pp 3156–64

[63] Wang G, Ye J C, Mueller K and Fessler J A 2018 Image reconstruction is a new frontier of machine learning *IEEE Trans. Med. Imaging* **37** 1289–96

[64] Wolterink J M, Dinkla A M, Savenije M H F, Seevinck P R, van den Berg C A T and Išgum I 2017 Deep MR to CT synthesis using unpaired data *Simulation and Synthesis in Medical Imaging* (*Lecture Notes in Computer Science* vol 10557) ed S Tsaftaris, A Gooya, A Frangi and J Prince SASHIMI (Cham: Springer) pp 14–23

[65] Wu D, Kim K, Dong B and Li Q 2017 End-to-end abnormality detection in medical imaging (arXiv:1711.02074)

[66] Xie J, Xu L and Chen E 2012 Image denoising and inpainting with deep neural networks *25th Conf. Neural Information Processing Systems (NIPS 2012)* pp 341–9

[67] Yang X, Kwitt R, Styner M and Niethammer M 2017 Quicksilver: fast predictive image registration–a deep learning approach *NeuroImage* **158** 378–96