Learning Regularization Parameter-Maps for Variational Image Reconstruction using Deep Neural Networks and Algorithm Unrolling[*]

Andreas Kofler[1], Fabian Altekrüger[2,3], Fatima Antarou Ba[3], Christoph Kolbitsch[1], Evangelos Papoutsellis[4,5], David Schote[1], Clemens Sirotenko[6], Felix Frederik Zimmermann[1], and Kostas Papafitsoros[7]

[1] *Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany*
[2] *Humboldt-Universität zu Berlin, Department of Mathematics, Berlin, Germany*
[3] *Technische Universität Berlin, Institute of Mathematics, Berlin, Germany*
[4] *Finden Ltd, Rutherford Appleton Laboratory, Harwell Campus, Didcot, United Kingdom*
[5] *Science and Technology Facilities Council, Harwell Campus, Didcot, United Kingdom*
[6] *Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany*
[7] *School of Mathematical Sciences, Queen Mary University of London, United Kingdom*

{*andreas.kofler, felix.zimmermann, david.schote, christoph.kolbitsch*}*@ptb.de*
fabian.altekrueger@hu-berlin.de, fatimaba@math.tu-berlin.de, epapoutsellis@gmail.com,
sirotenko@wias-berlin.de, k.papafitsoros@qmul.ac.uk

August 7, 2023

Abstract

We introduce a method for the fast estimation of data-adapted, spatially and temporally dependent regularization parameter-maps for variational image reconstruction, focusing on total variation (TV)-minimization. The proposed approach is inspired by recent developments in algorithm unrolling using deep neural networks (NNs), and relies on two distinct sub-networks. The first sub-network estimates the regularization parameter-map from the input data. The second sub-network unrolls T iterations of an iterative algorithm which approximately solves the corresponding TV-minimization problem incorporating the previously estimated regularization parameter-map. The overall network is then trained end-to-end in a supervised learning fashion using pairs of clean-corrupted data but crucially without the need of having access to labels for the optimal regularization parameter-maps. We first prove consistency of the unrolled scheme by showing that the unrolled minimizing energy functional used for the supervised learning Γ-converges as T tends to infinity, to the corresponding functional that incorporates the exact solution map of the TV-minimization problem. Then, we apply and evaluate the proposed method on a variety of large scale and dynamic imaging problems with retrospectively simulated measurement data for which the automatic computation of such regularization parameters has been so far challenging using the state-of-the-art methods: a 2D dynamic cardiac MRI reconstruction problem, a quantitative brain MRI reconstruction problem, a low-dose

[*] A small part of the paper referring to MRI reconstruction (1 page abstract submission) has been accepted as a poster presentation to the International Society for Magnetic Resonance in Medicine conference ISMRM May 2023 with slightly different numerics. In addition, the low dose computerized tomography part of this paper has been accepted to the 17th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, July 2023 as a poster presentation.

CT and a dynamic image denoising problem. The proposed method consistently improves the TV-reconstructions using scalar regularization parameters and the obtained regularization parameter-maps adapt well to each imaging problem and data by leading to the preservation of detailed features. Although the choice of the regularization parameter-maps is data-driven and based on NNs the subsequent reconstruction algorithm is interpretable since it inherits the properties (e.g. convergence guarantees) of the respective iterative reconstruction method from which the network is implicitly defined.

1 Introduction

Inverse imaging problems can often be described as

$$\mathbf{z} = \mathbf{A}\mathbf{x}_{\mathrm{true}} + \mathbf{e} \tag{1}$$

where $\mathbf{x}_{\mathrm{true}} \in V^n$ with $V \in \{\mathbb{R}, \mathbb{C}\}$ is the object to be imaged, $\mathbf{A} : V^n \to V^m$ is a linear operator which models the data-acquisition process, $\mathbf{e} \in V^m$ denotes some random noise component and $\mathbf{z} \in V^m$ represents the measured data. The goal is to reconstruct $\mathbf{x}_{\mathrm{true}}$ or at least a good enough approximation of it given the data $\mathbf{z}$. In practice, problem (1) is however ill-posed for various reasons. For example, in Magnetic Resonance Imaging (MRI) which is known to suffer from long acquisition times, the measurement process is often accelerated by undersampling in the raw-data domain, the so-called k-space, leading to an underdetermined system. In low-dose CT, where one reduces the radiation exposure of the patient by reducing the time of exposure to the photons emitted from the X-ray source, the measured data is noisy. Further, different inherent properties of the operator $\mathbf{A}$ also often determine how well-posed the problem is. Therefore, the reconstruction procedure requires the use of regularization methods to be able to obtain high quality images and particularly in medical imaging, images which can be used for diagnostic purposes. A prominent approach is to formulate the reconstruction as a minimization problem

$$\min_{\mathbf{x}} d(\mathbf{A}\mathbf{x}, \mathbf{z}) + \mathcal{R}(\mathbf{x}), \tag{2}$$

where $d(\,\cdot\,,\cdot\,)$ denotes a data-discrepancy measure and $\mathcal{R}(\,\cdot\,)$ a regularization term. Typical choices for $\mathcal{R}$ vary from $\mathcal{R}(\,\cdot\,) = \|\,\cdot\,\|_2^2$ for the well-known Tikhonov regularization [94] or $\mathcal{R}(\,\cdot\,) = \|\mathbf{T}\,\cdot\,\|_1$ for methods enforcing sparsity in some basis [23, 34]. One of the most widely applied methods is the so-called Total Variation (TV) regularization [13, 20, 85, 91]. Remaining in the finite dimensional setting, and choosing the square of the ℓ_2 norm as data discrepancy (appropriate for Gaussian noise), the reconstruction problem is formulated as

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{z}\|_2^2 + \lambda\|\nabla\mathbf{x}\|_1. \tag{3}$$

Here ∇ denotes a finite-differences operator and $\lambda > 0$ is a scalar regularization parameter that balances the effect of the two terms. This means that the regularization imposed on the sought image is given by sparsity in the gradient domain of the image measured with respect to the ℓ_1-norm. One reason for the great success of this method lies in its simple intuition, interpretability as well as its interesting mathematical properties. As a result, in the last decades, it has driven both theoretical as well as applied research fields such as biomedical engineering, inverse problems, optimization and geometric measure theory among others [15, 22, 86], with the complete list of publications in which the approach is investigated for different reconstruction problems in different imaging modalities being quite extensive. In addition, there exist nowadays numerous algorithms with proven convergence guarantees [21, 22, 48, 56, 100, 105] as well as extensions to overcome inherent limitations of the structural properties of the solutions of the problem (3), e.g. the total generalized variation (TGV) [14], for solving for the well-known TV staircasing artefacts (blocky-like, piecewise constant structures).

A crucial aspect which impacts the quality and the usefulness of the images which can be reconstructed by solving problem (3) is the careful choice of the parameter λ. Underestimating λ yields poor regularization, while overestimating it yields too smooth images with artificial "cartoon-like" appearance. Particularly in medical imaging applications, where images are at the basis of diagnostic decisions and therapy planning, a proper choice of any regularization parameter is crucial. There exist quite a few methods regarding the automatic choice of a scalar parameter λ, placed either in the TV term or in the data-discrepancy term, based on the discrepancy principle, L-curve methods and others, e.g. [17, 39, 42, 59].

However, employing one single scalar parameter λ which globally dictates the strength of the regularization over the entire image might seem sub-optimal for various obvious reasons. Depending on the application, it might be desirable to maintain locally higher data-fidelity instead of enforcing visually appealing but rather wrong image features. In that case, one can replace the parameter $\lambda \in \mathbb{R}_+$ in (3) with a spatially varying and pixel/voxel dependent one, denoted now by $\mathbf{\Lambda} \in \mathbb{R}_+^{qn}$, with q denoting the number of directions for which the partial derivatives are computed. Implementation-wise that translates to a stack of diagonal operators which contain a regularization parameter for each single pixel/voxel in the respective gradient domain of the image, resulting in a problem of the form

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{z}\|_2^2 + \|\mathbf{\Lambda}\nabla\mathbf{x}\|_1. \tag{4}$$

An automatic choice for such spatially varying regularization parameter is rather challenging, as the number of its components drastically increases. Towards that task, bilevel optimization techniques have been employed during the last years, which have the following general formulation:

$$\begin{cases} \min\limits_{\mathbf{\Lambda}} \displaystyle\sum_{i=1}^{M} l(\mathbf{x}^i(\mathbf{\Lambda}), \mathbf{x}^i_{\text{true}}) \\ \text{subject to} \quad \mathbf{x}^i(\mathbf{\Lambda}) = \operatorname*{argmin}_{\mathbf{x}} \dfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{z}_i\|_2^2 + \|\mathbf{\Lambda}\nabla\mathbf{x}\|_1, \quad i = 1, \ldots, M. \end{cases} \tag{5}$$

Here, $(\mathbf{z}_i, \mathbf{x}^i_{\text{true}})_{i=1}^M$ are M pairs of measured data and corresponding ground truth, and l is a suitable upper level objective. For instance, in the case where $l(x_1, x_2) = l_{\text{PSNR}}(x_1, x_2) := \|x_1 - x_2\|_2^2$, the bilevel problem (5) aims to compute the parameters $\mathbf{\Lambda}$ which are "on the average the best ones" (i.e. PSNR-maximizing), for the given M data pairs. The idea is that, given some new data $\mathbf{z}_{\text{test}}$ that has been measured in a similar way as $(\mathbf{z}_i)_{i=1}^M$, solving (4) (in the "online phase") with the offline-computed $\mathbf{\Lambda}$ will yield a good reconstruction. This scheme has been extensively studied both for scalar and spatially varying regularization parameters. However, in practice it has mainly been applied for image denoising (i.e. $\mathbf{A} = \mathbf{I}_n$) and for scalar or coarse patch-based parameters [16, 24, 29, 31, 57]. An extension for learning the optimal sampling pattern in MRI [89], as well as extensions to non-local and higher order regularizers [30, 33] have been considered as well. Further, unsupervised approaches, employing upper level energies that do not depend on the ground truth $\mathbf{x}_{\text{true}}$, i.e., $l = l(\mathbf{x}(\mathbf{\Lambda}))$ and $M = 1$, have also been considered in a series of works [43, 45–47, 79]. The upper level energy considered there aims to constrain localized versions of the image residuals $\mathbf{A}\mathbf{x} - \mathbf{z}$ within a certain tight corridor around the variance of the (Gaussian) noise $\mathbf{e}$, which is assumed to be known. Even though these bilevel optimization methods are typically accompanied by elegant mathematical theories, there exist limitations on the computational time they require in order to give satisfactory results. For instance, employing these methods for 2D or even 3D dynamic

imaging requires a vast computation effort and as a result, these limitations pose a challenge for the application in modern medical imaging modalities and hence in the clinical routine.
Recently, methods that are based on neural networks (NNs) have been proposed for the task of the estimation of such regularization parameter-maps. In [5], the authors employ a classical supervised learning approach in order to learn the map from the data $\mathbf{z}$ to the optimal scalar regularization parameter λ. The pipeline consists again of an offline and an online phase. More precisely, given again M pairs of measured data and corresponding ground truth images $(\mathbf{z}_i, \mathbf{x}_{\text{true}}^i)_{i=1}^M$, during the first part of the offline phase, a corresponding family of optimal regularization parameters $(\lambda_i)_{i=1}^M$ is computed, e.g. by employing a scheme like (5) separately for each i. Then, in the second part of the offline phase, using the training data $\mathcal{D} = \{(\lambda_i, \mathbf{z}_i)_{i=1}^M\}$, the parameters Θ of a NN $\mathcal{N}_\Theta$ are learned by minimizing

$$\min_{\Theta} \mathcal{L}(\Theta) := \frac{1}{M} \sum_{i=1}^{M} l(\mathcal{N}_\Theta(\mathbf{z}_i), \lambda_i), \tag{6}$$

for a suitable loss function l. Once an estimate of the optimal parameters Θ has been learned, one passes to the online phase, and given some new data $\mathbf{z}_{\text{test}}$, the regularization parameter is simply calculated by applying the learned network to $\mathbf{z}_{\text{test}}$, i.e., $\lambda_\Theta = \mathcal{N}_\Theta(\mathbf{z}_{\text{test}})$ and the classical image reconstruction problem

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{z}_{\text{test}}\|_2^2 + \lambda_\Theta \|\nabla \mathbf{x}\|_1, \tag{7}$$

is solved by an appropriate algorithm. The idea is that, due to the good generalizability and adaptability of NNs on unseen data, the computed regularization parameter $\lambda_\Theta = \mathcal{N}_\Theta(\mathbf{z}_{\text{test}})$ will be better adapted to $\mathbf{z}_{\text{test}}$ than the "average" λ of the (scalar parameter version of) bilevel optimization approach (5). The authors in [5] apply this pipeline to learn scalar regularization parameters for computerized tomography reconstruction and image deblurring. Nevertheless, the computational burden for computing offline the training data as well as solving (7) for high dimensional (3D and dynamic) problems still remains. A similar approach, where the supervised learning problem (6) is performed at the level of small image-patches, was performed in [77] for the image denoising problem.
In this work, inspired by the recent success of unrolled NNs [75], and targeting a variety of inverse problems including dynamic ones, we apply a different strategy for the construction of the regularization parameter-maps. We construct an unrolled NN which corresponds to an implementation of an iterative scheme of finite length to approach the solution of problem (3) assuming a *fixed* regularization parameter-map. Within the unrolled NN, the regularization parameter-map is estimated from the input data and is used throughout the whole reconstruction scheme. To be more precise, given some initial estimate $\mathbf{x}_0$ we work with an iterative scheme

$$\mathbf{x}_T = S^T(\mathbf{x}_0, \mathbf{z}, \mathbf{\Lambda}, \mathbf{A}), \quad T = 0, 1, 2, \ldots \tag{8}$$

for which we know that $\mathbf{x}_T \to S^*(\mathbf{z}, \mathbf{\Lambda}, \mathbf{A})$ as $T \to \infty$ where $S^*(\mathbf{z}, \mathbf{\Lambda}, \mathbf{A})$ is a solution of (4). We note that sometimes, we will drop the dependence of S^T on $\mathbf{x}_0$, $\mathbf{z}$, $\mathbf{A}$, for notational convenience. Then, for some fixed number of iterations $T \in \mathbb{N}$, our unrolled NN reads as follows:

$$\begin{cases} \mathbf{\Lambda}_\Theta = \text{NET}_\Theta(\mathbf{x}_0), \\ \mathbf{x}_k = S^k(\mathbf{x}_0, \mathbf{z}, \mathbf{\Lambda}_\Theta, \mathbf{A}), \quad k = 1, \ldots, T. \end{cases} \tag{9}$$

Here, NET_Θ denotes some convolutional NN (CNN) with learnable parameters Θ. We denote by $\mathcal{N}_\Theta^T$ the overall resulting network, i.e.

$$\mathcal{N}_\Theta^T(\mathbf{x}_0) = S^T(\mathbf{x}_0, \mathbf{z}, \mathbf{\Lambda}_\Theta, \mathbf{A}) = S^T(\mathbf{x}_0, \mathbf{z}, \text{NET}_\Theta(\mathbf{x}_0), \mathbf{A}).$$

The unrolled NN can then be end-to-end trained in a supervised manner on a set of input-target image-pairs. This resulting network can be identified as a pipeline that combines in a sequential way

- the estimation of the regularization parameter-map which is adapted to the data $\mathbf{z}$ (and hence in medical imaging to the new patient) and
- the iterative scheme that solves the image reconstruction problem.

In particular, given a new unseen input data $\mathbf{z}_{\mathrm{test}}$, the regularization parameter-map $\mathbf{\Lambda}_\Theta$ is estimated and stays fixed. Then, the reconstruction problem is solved by unrolling an appropriate algorithm. As such, the resulting method is interpretable and naturally inherits all convergence properties of the initial reconstruction algorithm since the data-driven component merely lies in the choice of the parameter-map.
Our approach can be considered to belong to the family of recently developed image reconstruction methods that combine elements both from *model-based* and *data-driven* regularization approaches. This is a modern and active field of research where interpretability and convergence guarantees from the traditional variational image reconstruction approaches are combined with the flexibility and adaptability of the deep-learning based methods. These combined approaches which aim to bring together the best of both worlds can result from instance by learning the regularization functional $\mathcal{R}$ from data and embed it into a scheme like(2), see [62, 70], by enforcing the reconstruction to be close to an output of a network via $\mathcal{R}$, e.g. $\mathcal{R}(\,\cdot\,) = \|\,\cdot - u_\Theta(\mathbf{s})\|_2^2$ for some network u_Θ with trainable parameters Θ and input $\mathbf{s}$ [35, 50, 53, 87], by substituting proximal operators in classical iterative schemes by learned NN denoisers (in a "plug-and-play" fashion) [73,83], or by using learned iterative schemes [2, 4, 41, 54, 63], see also the review papers [9, 72, 75]. Since one of our choices for the iterative scheme (8) will be the Primal-Dual Hybrid Gradient method (PDHG) of Chambolle and Pock [21], our approach is related to the Learned Primal-Dual method [4], where the proximal operators in the primal and dual step of PDHG are fully substituted by learnable networks. Here, we decrease the complexity but increase the interpretability by keeping the iterations of the iterative scheme untouched and put all the power of NNs in the estimation of the input-dependent regularization parameter-map, given in the first line of (9). As a result, our approach can be regarded as an intermediate approach between [5] and [4]. One of the main reasons we follow this approach is because, apart from the increased interpretability, we also target dynamic imaging applications and we are particularly interested in the interplay between the learned temporal and spatial regularization. As far as we are aware and in contrast to static imaging problems, there are no existing works on automatically computing regularization parameters for dynamic problems that are both spatially and temporally varying. Furthermore, because the "black-box" nature of CNNs in entirely put on the estimation of the regularization parameter-maps, the probability to possibly observe instabilities of the method in the sense of [25] is rather small. From a theoretical point of view, at least for denoising, it can be shown that for smooth regularization parameter-maps, no artefacts (i.e. new discontinuities) can appear in the reconstructions and for rougher weights, any creation of new discontinuities can be controlled [18,44,51]. Moreover, even the worst-case of locally very large produced regularization weights will only result in a locally flat area in the image with controlled values. Further, from a practical point of view, in preliminary experiments, we have observed that even fully random regularization parameter-maps yield reconstructions whose artefacts can be at worst similar to the ones which would result from a locally too low or too strong TV-regularization.

We evaluate the proposed approach on a variety of reconstruction problems such as accelerated cardiac cine MRI, quantitative MRI, dynamic image denoising and low-dose computerized tomography (CT). We show that the proposed approach significantly improves the reconstruction results which can be obtained by the respective methods using only scalar regularization values, and better preserves the fine scale details by adapting the regularization strength to the given data. We finally stress that even though here we focus on TV regularization only, the proposed framework can be in principle adapted to more sophisticated regularization methods

The rest of the paper is organized as follows. In Section 2 we review spatio-temporal TV-based regularization and introduce notation. In Section 3, we present our proposed approach for obtaining a data/patient-adaptive spatial or spatio-temporal regularization parameter-map. We investigate theoretical aspects of the proposed approach in Section 4, focusing on the consistency of the unrolled scheme. In Section 5, we conduct experiments to evaluate the proposed method on different imaging problems. We conclude the work in Section 6 by discussing some aspects of the proposed approach, its limitations and possible future research directions.

2 Spatio-Temporal Variational Regularization Models

In this section we introduce in more detail the different spatio-temporal regularization functionals and we review relevant works from the literature. In parallel, we also fix the different notations for the several regularization parameters (scalar and spatially/spatio-temporally varying) in relationship to the way these are computed, e.g. supervised, unsupervised, ground truth-based, NNs-based.

2.1 Spatio-Temporal Total Variation

Setting $V^n := V^{n_x \times n_y \times n_t}$, $n_x, n_y, n_t \in \mathbb{N}$, we define the discrete spatio-temporal gradient operator $\nabla : V^n \to (V \times V \times V)^n$ as

$$\nabla \mathbf{x}(z) = [\nabla_x \mathbf{x}(z), \nabla_y \mathbf{x}(z), \nabla_t \mathbf{x}(z)]^{\mathsf{T}}, \quad \mathbf{x} \in V^n, \tag{10}$$

where $\nabla_x, \nabla_y, \nabla_t$ are finite difference operators along the corresponding dimension. Here, $z \in [1, \ldots, n_x] \times [1, \ldots, n_y] \times [1, \ldots, n_t] := \mathcal{I}$ denotes the set of indices. In this work we employ the anisotropic version for the total variation of $\mathbf{x} \in V^n$, i.e., the $\ell_{1,1}$-norm of $\nabla \mathbf{x}$

$$\mathrm{TV}(\mathbf{x}) = \|\nabla \mathbf{x}\|_1 = \sum_{z \in \mathcal{I}} |\nabla \mathbf{x}(z)|_1 := \sum_{z \in \mathcal{I}} |\nabla_x \mathbf{x}(z)| + |\nabla_y \mathbf{x}(z)| + |\nabla_t \mathbf{x}(z)|, \tag{11}$$

Note that in the case $V = \mathbb{C}$, then we identify this space with $\mathbb{R}^2$ endowed with the $|\cdot|_1$ norm, and ∇, TV are defined analogously.

2.2 Notations on the Different Regularization Weights and Corresponding Spatio-Temporal Total Variation Functionals

In general, we will denote scalar and spatially (and/or temporally) varying regularization parameters with λ and $\mathbf{\Lambda}$, respectively. Whenever such a parameter is the output of a NN (with weights Θ), the subindex Θ will be used, i.e., λ_Θ or $\mathbf{\Lambda}_\Theta$. If such a parameter produces the best corresponding TV-reconstruction with respect to the PSNR for some given data $\mathbf{z}$, it will be denoted by λ_{P} or $\mathbf{\Lambda}_{\mathrm{P}}$. For instance, these would be the optimal parameters that are solutions to the bilevel scheme (5) when the upper level energy l_{PSNR} is used and $M = 1$. In that case, the training and the test image coincide. If the "best" is understood as "on average" based on some training data, i.e. bilevel scheme (5) with l_{PSNR} and $M > 1$, we denote these parameters as $\lambda_{\tilde{\mathrm{P}}}$ or $\mathbf{\Lambda}_{\tilde{\mathrm{P}}}$. In that case the test image is not part of the training data.

On the other hand, we use superindices to index whether the different components of the regularization parameters that correspond to the different dimensions are the same or not. For instance,

$$\lambda^{x,y,t} = (\lambda^x, \lambda^y, \lambda^t) \in \mathbb{R}^3_+, \tag{12}$$

$$\lambda^{xy,t} = (\lambda^{xy}, \lambda^{xy}, \lambda^t) \in \mathbb{R}^3_+, \tag{13}$$

$$\lambda^{xyt} = (\lambda^{xyt}, \lambda^{xyt}, \lambda^{xyt}) \in \mathbb{R}^3_+, \tag{14}$$

denote parameters that weight all the components differently, weight only the spatial components equally, and weight all the components equally respectively. For instance, the use of (12) leads to the following version of weighted TV:

$$\mathrm{TV}_{\lambda^{x,y,t}}(\mathbf{x}) := \|\lambda^{x,y,t}\nabla\mathbf{x}\|_1 = \sum_{z\in\mathcal{I}} \lambda^{x,y,t}|\nabla\mathbf{x}(z)|_1 := \sum_{z\in\mathcal{I}} \lambda^x|\nabla_x\mathbf{x}(z)| + \lambda^y|\nabla_y\mathbf{x}(z)| + \lambda^t|\nabla_t\mathbf{x}(z)|. \tag{15}$$

In contrast, $\lambda^{xy,t} = (\lambda^{xy}, \lambda^{xy}, \lambda^t)$ denotes a parameter where the spatial components x and y are weighted equally. Analogously, we define the spatio-temporally varying versions, generally denoted by $\mathbf{\Lambda} \in \mathbb{R}^{qn}_+$. In particular, we define

$$\mathbf{\Lambda}^{x,y,t} = (\mathbf{\Lambda}^x, \mathbf{\Lambda}^y, \mathbf{\Lambda}^t) \in (\mathbb{R}^n_+)^3, \tag{16}$$

$$\mathbf{\Lambda}^{xy,t} = (\mathbf{\Lambda}^{xy}, \mathbf{\Lambda}^{xy}, \mathbf{\Lambda}^t) \in (\mathbb{R}^n_+)^3, \tag{17}$$

$$\mathbf{\Lambda}^{xyt} = (\mathbf{\Lambda}^{xyt}, \mathbf{\Lambda}^{xyt}, \mathbf{\Lambda}^{xyt}) \in (\mathbb{R}^n_+)^3, \tag{18}$$

with (17), for instance, leading to the following version of weighted TV

$$\mathrm{TV}_{\mathbf{\Lambda}^{xy,t}}(\mathbf{x}) := \|\mathbf{\Lambda}^{xy,t}\nabla\mathbf{x}\|_1 = \sum_{z\in\mathcal{I}} |\mathbf{\Lambda}^{xy,t}(z)\nabla\mathbf{x}(z)|_1 \tag{19}$$

$$:= \sum_{z\in\mathcal{I}} \mathbf{\Lambda}^{xy}(z)|\nabla_x\mathbf{x}(z)| + \mathbf{\Lambda}^{xy}(z)|\nabla_y\mathbf{x}(z)| + \mathbf{\Lambda}^t(z)|\nabla_t\mathbf{x}(z)|. \tag{20}$$

Here the multiplication of $\mathbf{\Lambda}^{xy,t}$ and $\nabla\mathbf{x}$ is considered component-wise. The full notations of the type, e.g. $\mathbf{\Lambda}^{xy,t}_\Theta$, $\mathbf{\Lambda}^{xy,t}_\mathrm{P}$, $\lambda^{xy,t}_\mathrm{P}$ have the obvious meaning. Of particular interest will be the comparison of the reconstructions that correspond to the above parameters to the ones that correspond to the parameters $\mathbf{\Lambda}^{x,y,t}_\Theta/\mathbf{\Lambda}^{xy,t}_\Theta/\mathbf{\Lambda}^{xyt}_\Theta$ that we learn through our unrolled scheme. We also note that the quantities which correspond to spatial regularization only, i.e., for static imaging tasks, are defined straightforwardly by omitting the temporal component t.

2.3 Related Literature on Spatio-Temporal Total Variation-type Functionals

There have been quite a few related works in the dynamic inverse problems literature that employ regularization functionals of the type (15), or higher order extensions. Even though, we will also use our approach for static tasks, we briefly review these works since the literature on computing spatio-temporal regularization parameters for dynamic problems is essentially void. In [49], the authors use a regularization functional defined as an infimal convolution of functionals of the type (15) for video denoising and decompression, an approach which splits the image sequence into two components with little change in space and time respectively. A bilevel approach for dynamic denoising is considered in [11]. A higher extension of the approach, applied to dynamic MRI was investigated in [88] and

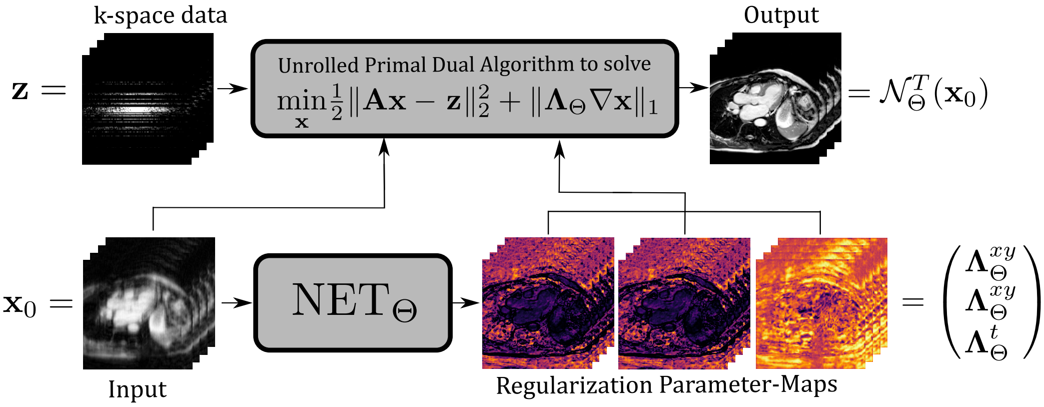
k-space data
$\mathbf{z} =$
Unrolled Primal Dual Algorithm to solve
$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{z}\|_2^2 + \|\mathbf{\Lambda}_\Theta \nabla \mathbf{x}\|_1$
Output
$= \mathcal{N}_\Theta^T(\mathbf{x}_0)$
$\mathbf{x}_0 =$
NET_Θ
$= \begin{pmatrix} \mathbf{\Lambda}_\Theta^{xy} \\ \mathbf{\Lambda}_\Theta^{xy} \\ \mathbf{\Lambda}_\Theta^{t} \end{pmatrix}$
Input
Regularization Parameter-Maps

Figure 1: Illustration of the proposed network architecture for a dynamic cardiac MR image reconstruction problem. The network consists of a sub-network which estimates the regularization parameter-maps and a sub-network which reconstructs the image using the PDHG algorithm described in Algorithm 1. First, a spatio-temporal parameter-map $\mathbf{\Lambda}_\Theta$, here as in (17), is estimated by applying NET_Θ to an input image $\mathbf{x}_0$. The regularization parameter-map is then used within the reconstruction network which assumes the parameter-map to be fixed. The regularization parameter-map is trained such that the output of the PDHG algorithm is close to a reference image.

for dynamic PET in [12]. Regularization of the type (15) has been also considered for dynamic tomographic imaging [80] and dynamic cardiac MRI [101]. We stress however that in all these works all regularization parameters are scalar and they are manually selected. Here we allow for better flexibility in the regularization by automatically computing regularization parameters that are both spatially and temporally dependent, and we let these parameters guide the decoupling to static and moving parts in the image sequence.

3 Proposed Unrolled Network Structure

As described in (9), our network architecture $\mathcal{N}_\Theta^T$ consists of two parts. The first part of the network is concerned with the determination of the regularization parameter-map $\mathbf{\Lambda}_\Theta$ which is subsequently fed into the second part. We describe this procedure in more detail in Section 3.1. Assuming the regularization parameter-map $\mathbf{\Lambda}_\Theta$ is fixed in the second module of the network, $\mathbf{\Lambda}_\Theta$ is fed into an unrolled iterative scheme of length T which, if run until convergence, exactly solves

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{z}\|_2^2 + \|\mathbf{\Lambda}_\Theta \nabla \mathbf{x}\|_1. \tag{21}$$

For the latter we choose T iterations of the PDHG algorithm [21], which we briefly recall here.

By denoting $X = V^n$, $Z = V^m, Q = V^{qn}$, the image reconstruction problem (4) can be equivalently formulated as

$$\min_{\mathbf{x} \in X} f(\mathbf{K}\mathbf{x}) + g(\mathbf{x}), \tag{22}$$

with $f : Y = Z \times Q \to \mathbb{R}$ where

$$f(\mathbf{y}) = f(\mathbf{p}, \mathbf{q}) := f_1(\mathbf{p}) + f_2(\mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{z}\|_2^2 + \|\mathbf{\Lambda}_\Theta\, \mathbf{q}\|_1, \quad \mathbf{K} := \begin{bmatrix} \mathbf{A} \\ \nabla \end{bmatrix}, \quad g(\mathbf{x}) := \mathbf{0}. \tag{23}$$

Here, the variables $\mathbf{p}, \mathbf{q}$ belong to the finite dimensional Euclidean spaces Z and Q that correspond to the specificities, e.g. dimensions, of each problem, and $\mathbf{K} : X \to Y$.

The PDHG algorithm for solving problems of the general form (22), i.e., with f and g convex as well as $\mathbf{K}$ bounded and linear is described in Algorithm 1. Recall that, for a convex function h and scalar $\sigma > 0$, the proximal operator $\mathrm{prox}_{\sigma h}$ is defined as

$$\mathrm{prox}_{\sigma h}(\overline{\mathbf{y}}) := \operatorname*{argmin}_{\mathbf{y}} \frac{1}{2}\|\mathbf{y} - \overline{\mathbf{y}}\|_2^2 + \sigma h(\mathbf{y}), \tag{24}$$

while the convex conjugate of h is defined as

$$h^*(\overline{\mathbf{y}}) := \max_{\mathbf{y}} \langle \mathbf{y}, \overline{\mathbf{y}} \rangle - h(\mathbf{y}). \tag{25}$$

In order to be consistent with our purposes, we have stated the Algorithm 1 such that it terminates in T iterations with an output $\mathbf{x}_T$. However, from standard convergence analysis it holds that $\mathbf{x}_T \to \mathbf{x}^*$ as $T \to \infty$, where $\mathbf{x}^*$ solves (4).

Algorithm 1 Unrolled PDHG algorithm [21]

```
Input: $L = \|\mathbf{K}\|$,  $\sigma\tau < 1/L^2$,  $\theta = 1$,  initial guess $\mathbf{x}_0$
Parameters: number of iterations $T > 0$
Output: reconstructed image $\mathbf{x}_T$
$\bar{\mathbf{x}}_0 = \mathbf{x}_0$
$\mathbf{y}_0 = \mathbf{0}$
for $k = 0, \ldots, T-1$ do
    $\mathbf{y}_{k+1} = \mathrm{prox}_{\sigma f^*}(\mathbf{y}_k + \sigma \mathbf{K}\bar{\mathbf{x}}_k)$
    $\mathbf{x}_{k+1} = \mathrm{prox}_{\tau g}(\mathbf{x}_k - \tau \mathbf{K}^{\mathsf{T}}\mathbf{y}_{k+1})$
    $\bar{\mathbf{x}}_{k+1} = \mathbf{x}_{k+1} + \theta(\mathbf{x}_{k+1} - \mathbf{x}_k)$
end for
```

Remark 1. *Later we recall the precise form of Algorithm* (1) *for the problem* (4)*. Here, we only mention that the* $\mathrm{prox}_{\sigma f^*}$ *for* f *as in* (23)*, decouples to* $\mathrm{prox}_{\sigma f_1^*}$ *and* $\mathrm{prox}_{\sigma f_2^*}$*, with the latter acting as a pointwise projection ("clipping") onto the bilateral set* $[-\mathbf{\Lambda}_i, \mathbf{\Lambda}_i]$*,* $i = 1, \ldots, qn$*. In particular, the map* $\mathbf{\Lambda} \mapsto \mathrm{prox}_{\sigma f_2^*}(\mathbf{q})$ *is Lipschitz with constant one, for every* $\mathbf{q}$*. We remark that this is the only place where the parameter* $\mathbf{\Lambda}$ *appears in the version of Algorithm* (1) *for the problem* (4)*.*

Remark 2. *We mention that for the low-dose CT application which we consider in Section 5.5, we will be using a generalization of PDHG, namely the PD3O algorithm [106] which is better adapted for the Kullback-Leibler divergence fidelity term used there. We give more details later in that section.*

3.1 Obtaining the Regularization Parameter-Map Via a CNN

In our set-up, $\mathbf{\Lambda}_\Theta$ is the output of a CNN with parameters Θ, denoted by NET_Θ, which takes as an input an initial image $\mathbf{x}_0$, i.e., $\mathbf{\Lambda}_\Theta = \mathrm{NET}_\Theta(\mathbf{x}_0)$. Depending on the structure of the considered imaging problem, we can explore different possibilities for the construction of the latter. For instance, for a dynamic imaging problem, i.e., 2D + time, we might prefer to attribute equal importance to the x- and y-direction, but use a different parameter-map for the temporal component resulting in

$$\mathbf{\Lambda}_\Theta = (\mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{t}). \tag{26}$$

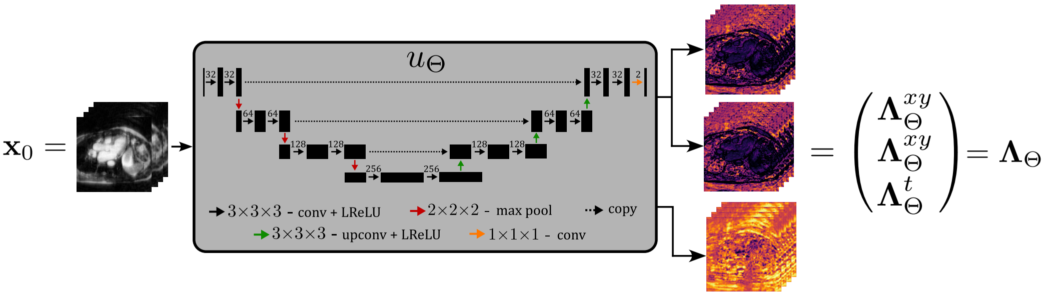
$\mathbf{x}_0 =$
u_Θ
32
32
64
64
128
128
256
256
128
128
64
64
32
32
2
3×3×3 - conv + LReLU
2×2×2 - max pool
copy
3×3×3 - upconv + LReLU
1×1×1 - conv
$= \begin{pmatrix} \mathbf{\Lambda}_\Theta^{xy} \\ \mathbf{\Lambda}_\Theta^{xy} \\ \mathbf{\Lambda}_\Theta^{t} \end{pmatrix} = \mathbf{\Lambda}_\Theta$

Figure 2: The CNN-block u_Θ is a U-Net which we hyper-parametrize by the number of encoding stages, the number of convolutional layers and the initial number of filters applied to the input-image. The filter-sizes as well as the number of output channels of u_Θ depend on the considered application. The network shown here is a 3D U-Net with four encoding stages, two convolutional layers per stage and 32 initially applied filters. Here, the CNN is constructed to yield two different components of the regularization parameter-map, i.e. $\mathbf{\Lambda}_\Theta^{xy}$ and $\mathbf{\Lambda}_\Theta^{t}$, which are then used to construct the final spatio-temporal regularization parameter-map $\mathbf{\Lambda}_\Theta^{xy,t}$ according to (17), i.e. by $\mathbf{\Lambda}_\Theta^{xy,t} = (\mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{t})$.

This choice is motivated by the later shown cardiac cine MRI reconstruction problem. There, the temporal dimension is the one which on the one hand exhibits the largest correlation to be exploited by the TV-method, but on the other hand also the one which contains the diagnostic information and therefore requires special care to ensure that important features are preserved.

For a 3D imaging problem, one could for example attribute equal importance to all spatial-directions or opt for a construction as in (26), if for example the z-direction has a different resolution than the x- and y-directions. Moreover, for complex-valued images, it seems intuitive to share the same regularization map across the real and the imaginary parts of the images.

The core of the overall network, denoted by u_Θ, consists of a (sub-)CNN with high expressive capabilities such as the U-Net [84]. To constrain the regularization parameter-maps to be strictly positive, we then apply a softplus activation function ϕ and, as last operation, we multiply the output by a positive parameter $t > 0$. Empirically, we have experienced that the network's training benefits in terms of faster convergence if the order of the scale of the output is properly set depending on the application. This can be achieved either by accordingly initializing the weights of the network u_Θ, or in a simpler way, as we do here by scaling the output of the CNN. Summarizing, given an input image $\mathbf{x}_0$, we estimate the corresponding regularization parameter-map by

$$\mathbf{\Lambda}_\Theta = \mathrm{NET}_\Theta(\mathbf{x}_0) = t\,\phi(u_\Theta(\mathbf{x}_0)). \tag{27}$$

We finally recall that the overall network has the form

$$\mathcal{N}_\Theta^T(\mathbf{x}_0) = S^T(\mathbf{x}_0, \mathbf{z}, \mathrm{NET}_\Theta(\mathbf{x}_0), \mathbf{A}). \tag{28}$$

Remark 3. *Note that in our set-up we use the same quantity $\mathbf{x}_0$ as the input for the CNN-block $\mathrm{NET}_\Theta(\mathbf{x}_0)$ as well as the initialization for the unrolled PDHG $S^T(\mathbf{x}_0, \ldots)$. According to our experience, this produces satisfactory results, see also the discussion in Section 5.1. However this is not a hard constraint of the method and one could also further experiment with having different values for these variables.*

3.2 Network Training

Training the network $\mathcal{N}_\Theta^T$ refers to minimizing a chosen energy-function $\mathcal{L}$ over a set of input-target training pairs $\mathcal{D} = \{(\mathbf{x}_0^i, \mathbf{x}_{\text{true}}^i)_{i=1}^M : \mathbf{x}_0^i = \mathbf{A}^\ddagger \mathbf{z}_i,\, \mathbf{z}_i := \mathbf{A}\mathbf{x}_{\text{true}}^i + \mathbf{e}_i\}$. Here $\mathbf{A}^\ddagger$ denotes some reconstruction operator, e.g. the pseudo-inverse of $\mathbf{A}$, which provides the inputs $\mathbf{x}_0^i$ and $\mathbf{e}_i$ are noisy terms. Using an appropriate loss function l and potentially some regularization function r for the weights Θ, we end up with the following energy-function:

$$\mathcal{L}(\Theta) = \frac{1}{M} \sum_{i=1}^{M} l\big(\mathcal{N}_\Theta^T(\mathbf{x}_0^i)\,,\, \mathbf{x}_{\text{true}}^i\big) + r(\Theta). \tag{29}$$

Note that r can also encode some constraints on Θ by being an indicator of some set. Note that the network is trained end-to-end from the initial reconstruction to its estimate. Therefore, the set Θ is adjusted such that the estimated parameter-map $\mathbf{\Lambda}_\Theta$ is appropriate for a subsequent reconstruction using a suitable reconstruction algorithm as for example, the primal-dual method in Algorithm 1. This conceptually highly differs from approaches as in [5], in which a network is trained to estimate the best scalar regularization parameter which is previously obtained by a time-consuming grid-search. First of all, in [5] the learning procedure is entirely decoupled from the employed reconstruction algorithm. Second, opposed to our approach, the method requires access to a target regularization parameter, meaning that a generalization of [5] to regularization parameter-maps would require access to entire target regularization parameter-maps which can typically only obtained by even more time-consuming approaches. Our approach, in contrast, allows to implicitly learn the regularization parameter-maps by unrolling the reconstruction algorithm and thus only requires access to ground truth target images.

4 Consistency Analysis of the Unrolled Scheme

In addition to the practical advantages of the proposed method which will be highlighted in Section 5, we want to discuss some of the emerging theoretical questions and in particular some consistency results when we let the number of unrolled iterations $T \to \infty$. We note that there are papers that study hyperparameter search with bilevel optimization and unrolled optimization methods, see e.g. [65,66,78]. Although some of the latter articles provide consistency analysis in different contexts, we think that none of the techniques presented there can be applied to our problem.

We will be using the space and operator notation X, Y, Z and $\mathbf{K}$ as these were defined in the previous section and we will also set $\mathcal{V} := X \times Y$. For simplicity here we work with the real-valued case, i.e. $V = \mathbb{R}$. Recall that the solution of the convex variational problem (4) and the corresponding T-th iterate of the unrolled algorithm are denoted by $\mathbf{x}^* = S^*(\mathbf{z}, \mathbf{\Lambda})$ and $\mathbf{x}_T = S^T(\mathbf{x}_0, \mathbf{z}, \mathbf{\Lambda})$. Recall also that for the ease of notation we sometimes suppress the dependence of S^T on the initialization $\mathbf{x}_0$ of Algorithm 1, as well as the one of the dual variable $\mathbf{y}_0$. We then have $S^T(\mathbf{x}_0, \mathbf{z}, \mathbf{\Lambda}) \to S^*(\mathbf{z}, \mathbf{\Lambda})$ as $T \to \infty$. Furthermore, for this section we consider a more general fidelity term d, such that $f_1(\cdot) := d(\cdot, \mathbf{z})$.

Let us now consider the learning framework as presented in Section 3

$$\min_{\Theta \in \mathbb{R}^\ell} \mathcal{L}^T(\Theta) := \frac{1}{M} \sum_{i=1}^{M} l(\mathcal{N}_\Theta^T(\mathbf{x}_0^i)\,,\, \mathbf{x}_{\text{true}}^i) + r(\Theta), \tag{30}$$

as well as the corresponding training scheme where no unrolling is taking place, i.e.,

$$\min_{\Theta\in\mathbb{R}^{\ell}} \mathcal{L}^{*}(\Theta) := \frac{1}{M}\sum_{i=1}^{M} l(\mathcal{N}_{\Theta}^{*}(\mathbf{x}_0^i)\,,\, \mathbf{x}_{\text{true}}^i) + r(\Theta), \tag{31}$$

where we used analogously the notation $\mathcal{N}_{\Theta}^{*}(\mathbf{x}_0) := S^{*}(\mathbf{z}, \mathbf{\Lambda}_{\Theta}(\mathbf{x}_0))$. Our target will be to show convergence of (ϵ)-minimizers of (30) to minimizers of (31) as $T \to \infty$ under appropriate conditions via a Γ-convergence argument.

Naturally, in order to guarantee existence of minimizers for the problems (30) and (31), the functionals $\mathcal{L}^{T}$ and $\mathcal{L}^{*}$ must be coercive, in addition to the standard lower semicontinuity assumptions. However, it is not so clear if this can be achieved without imposing coercivity via the regularization function r, which can be the case when e.g. r is some norm in $\mathbb{R}^{\ell}$ or an indicator function of a bounded set. Even though strictly speaking, it is not needed for our main consistency result Corollary 8, we will assume that the minimization problems (30) and (31) indeed admit solutions. Of course in deep learning practice, one does not compute minimizers for these problems, but rather it is aimed that the energy is decreased up to some degree based on a validation set, in order to guarantee generalizability. However, the analysis presented here can serve as a starting point to further show consistency in the level of stationary points and/or energy decrease using validation sets.

Below we summarize a series of assumptions which we will need next:

Assumption 4. *We assume that the following hold:*

(i) The operator $\mathbf{A} : X \to Z$ is injective.

(ii) The fidelity term $d(\cdot, \mathbf{z})$ is $\mu_{\mathbf{z}}$-strongly convex and Lipschitz continuously differentiable for every $\mathbf{z} \in Z$. We denote by $L_{\mathbf{z}}$ the corresponding Lipschitz constant.

(iii) The parameters $\sigma, \tau > 0$ in Algorithm 1 are small enough such that $\sigma\tau < \|\mathbf{K}\|^2$.

(iv) The regularization function $r : \mathbb{R}^{\ell} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is proper and lower semicontinuous.

(v) The loss function $l : X \times X \to \mathbb{R}$ is continuous.

(vi) The activation functions in the U-Net u_{Θ} are continuous.

Note that if *(iii)* above is satisfied, then the matrix

$$\mathbf{M} = \begin{pmatrix} \frac{1}{\tau}\mathbf{I} & -\mathbf{K}^{\mathsf{T}} \\ -\mathbf{K} & \frac{1}{\sigma}\mathbf{I} \end{pmatrix} \tag{32}$$

is symmetric, positive definite and thus defines a norm in $\mathcal{V}$. Then there exist $c, C > 0$ such that

$$c\|\mathbf{v}\|_2 \le \|\mathbf{v}\|_{\mathbf{M}} := \sqrt{\langle \mathbf{M}\mathbf{v}, \mathbf{v}\rangle} \le C\|\mathbf{v}\|_2 \quad \text{for every } \mathbf{v} \in \mathcal{V}. \tag{33}$$

Remark 5. *The injectivity of the operator $\mathbf{A}$, together with the strong convexity of d, is used in order to ensure that $\mathbf{x} \mapsto d(\mathbf{A}\mathbf{x}, \mathbf{z})$ is strongly convex. This indeed guarantees uniqueness of the solution for the variational problem, and in particular the map $\mathbf{\Lambda} \mapsto S^{*}(\mathbf{z}, \mathbf{\Lambda})$ is well-defined and single-valued. For the applications we will consider in Section 5, i.e., denoising, MRI with multiple receiver coils and CT with enough angular views and detectors, this injectivity assumption is satisfied. We note however it might be possible to drop this injectivity assumption following [96], [97], or [107].*

We start with the following Proposition 6 showing Lipschitz continuity of the iterates $S^T(\mathbf{x}_0, \mathbf{z}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{\Lambda}$ as well as the equicontinuity property $S^T(\mathbf{x}_0, \mathbf{z}, \boldsymbol{\Lambda}_T) \to S^*(\mathbf{z}, \boldsymbol{\Lambda})$ whenever $\boldsymbol{\Lambda}_T \to \boldsymbol{\Lambda}$. Note that the convergence $\boldsymbol{\Lambda}_T \to \boldsymbol{\Lambda}$ as $T \to \infty$, is merely part of a technical condition and it is not associated to the structure of our unrolled scheme where, as we have pointed out, the CNN-output $\boldsymbol{\Lambda}_\Theta = \mathrm{NET}_\Theta(\mathbf{x}_0)$ remains unchanged.

Proposition 6. *Assuming (i)-(iii) of Assumption 4, the following statements hold:*

(i) The solution map $\boldsymbol{\Lambda} \mapsto S^*(\mathbf{z}, \boldsymbol{\Lambda})$ is Lipschitz continuous for every $\mathbf{z} \in Z$. In particular the following bound holds for every $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \in \mathbb{R}^{qn}_+$,

$$\|S^*(\mathbf{z}, \boldsymbol{\Lambda}_1) - S^*(\mathbf{z}, \boldsymbol{\Lambda}_2)\|_2 \leq \frac{2\|\nabla\|}{\lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A})\mu_\mathbf{z}} \|\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\|_2. \tag{34}$$

(ii) The map $\boldsymbol{\Lambda} \mapsto S^T(\mathbf{x}_0, \mathbf{z}, \boldsymbol{\Lambda})$ is Lipschitz continuous for every $\mathbf{z} \in Z$, $\mathbf{x}_0 \in X$ and $T \in \mathbb{N}$.

(iii) For $\|\boldsymbol{\Lambda}\|_2 \leq \overline{\boldsymbol{\Lambda}}$ we obtain the following sub-linear rate, for $\mathbf{v}_0 := (\mathbf{x}_0, \mathbf{y}_0)$ being the initial iterates of Algorithm 1

$$\|S^T(\mathbf{x}_0, \mathbf{z}, \boldsymbol{\Lambda}) - S^*(\mathbf{z}, \boldsymbol{\Lambda})\|_2 \leq \frac{3C_{\mathbf{z},\mathbf{A}}}{T^{1/4}} \left(1 + \|\mathbf{v}_0 - \mathbf{v}^*(\boldsymbol{\Lambda}, \mathbf{z})\|_\mathbf{M}\right), \tag{35}$$

where

$$C_{\mathbf{z},\mathbf{A}} := \frac{\max\left(CL_\mathbf{z}\|\mathbf{A}\|\,,\, 4C\overline{\boldsymbol{\Lambda}}\,,\, 2\,,\, \lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A})\mu_\mathbf{z}\right)}{\lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A})\mu_\mathbf{z}}, \tag{36}$$

with $\lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A})$ denoting the smallest eigenvalue of $\mathbf{A}^\mathsf{T}\mathbf{A}$.

(iv) Whenever $\boldsymbol{\Lambda}_T \to \boldsymbol{\Lambda}$ as $T \to \infty$, it holds $S^T(\mathbf{x}_0, \mathbf{z}, \boldsymbol{\Lambda}_T) \to S^*(\mathbf{z}, \boldsymbol{\Lambda})$ for every $\mathbf{x}_0 \in X$, $\mathbf{z} \in Z$.

Proof. (i) This statement is proved similarly to e.g. in [32, Theorem 4.1] and it is strongly based on the $\mu_\mathbf{z}$-strong convexity of the map $d(\cdot, \mathbf{z})$ and the injectivity of $\mathbf{A}$. The statement (ii) can also be seen easily since the only dependence of $\boldsymbol{\Lambda}$ in the unrolled PDHG scheme is via the pointwise projection onto $[-\boldsymbol{\Lambda}_i, \boldsymbol{\Lambda}_i]$ which is a Lipschitz map, recall Remark 1. As a result, the map $\boldsymbol{\Lambda} \mapsto S^T(\mathbf{x}_0, \mathbf{z}, \boldsymbol{\Lambda})$ is Lipschitz, as a composition of Lipschitz functions.

The proof of (iii) is more involved. We fix a ball of radius $\overline{\boldsymbol{\Lambda}}$ centered at the origin, denoted by $B_{\bar{\boldsymbol{\Lambda}}} \subset \mathbb{R}^{qn}_+$ and let $\boldsymbol{\Lambda} \in B_{\overline{\boldsymbol{\Lambda}}}$ be arbitrary. In what follows, we initially suppress the dependence of all variables on $\boldsymbol{\Lambda}$. Define the primal-dual gap

$$L(\mathbf{x}, \mathbf{y}) := \langle \mathbf{K}\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) + g(\mathbf{x}), \tag{37}$$

and denote by $\mathbf{v}_T := (\mathbf{x}_T, \mathbf{y}_T)$ the iterates of the Algorithm 1 and by $\mathbf{v}^* = (\mathbf{x}^*, \mathbf{y}^*)$ the corresponding limits. Then, the following estimate holds, see [68, Corollary 1] for a proof,

$$L(\mathbf{x}_T, \mathbf{y}) - L(\mathbf{x}, \mathbf{y}_T) \leq \frac{1}{\sqrt{T}} \left(\|\mathbf{v}_0 - \mathbf{v}^*\|^2_\mathbf{M} + \|\mathbf{v}_0 - \mathbf{v}^*\|_\mathbf{M}\|\mathbf{v} - \mathbf{v}^*\|_\mathbf{M}\right), \tag{38}$$

where $\mathbf{v} = (\mathbf{x}, \mathbf{y})$ is arbitrary. We can thus take the supremum over $\mathbf{y} \in \partial f(K\mathbf{x}_T)$ in both sides in (38) and estimate the left hand side as follows

$$\sup_{\mathbf{y}\in\partial f(\mathbf{K}\mathbf{x}_T)} L(\mathbf{x}_T, \mathbf{y}) - L(\mathbf{x}, \mathbf{y}_T) \geq \sup_{\mathbf{y}\in\partial f(\mathbf{K}\mathbf{x}_T)} \langle \mathbf{K}\mathbf{x}_T, \mathbf{y}\rangle - f^*(\mathbf{y}) - \langle \mathbf{K}\mathbf{x}, \mathbf{y}_T\rangle + f^*(\mathbf{y}_T) \geq f(\mathbf{K}\mathbf{x}_T) - f(\mathbf{K}\mathbf{x}), \tag{39}$$

where we used the fact that $\langle \mathbf{K}\mathbf{x}_T, \mathbf{y}\rangle - f^*(\mathbf{y}) = f(\mathbf{K}\mathbf{x}_T)$ if and only if $\mathbf{y} \in \partial f(\mathbf{K}\mathbf{x}_T)$. By setting $\mathbf{x} = \mathbf{x}^*$, using the $\mu_{\mathbf{z}}$-strong convexity of $f_1(\cdot) = d(\cdot, \mathbf{z})$, the convexity of f_2, together with $\mathbf{y}^* \in \partial f(\mathbf{K}\mathbf{x}^*)$ we deduce

$$f(\mathbf{K}\mathbf{x}_T) - f(\mathbf{K}\mathbf{x}^*) \geq \langle \mathbf{K}^*\mathbf{y}^*, \mathbf{x}_T - \mathbf{x}^*\rangle + \frac{\mu_{\mathbf{z}}}{2}\|\mathbf{A}\mathbf{x}_T - \mathbf{A}\mathbf{x}^*\|_2^2. \tag{40}$$

Taking into account that $\mathbf{K}^*\mathbf{y}^* = 0$ (taking limits at line 6 of Algorithm 1, using the fact that $\operatorname{prox}_{\tau g} = Id$), using the injectivity of $\mathbf{A}$, we infer from (39) and (40)

$$\|\mathbf{x}_T - \mathbf{x}^*\|_2^2 \leq \frac{2}{\lambda_{\min}(\mathbf{A}^{\mathbf{T}}\mathbf{A})\mu_{\mathbf{z}}\sqrt{T}} \sup_{\mathbf{y}\in\partial f(\mathbf{K}\mathbf{x}_T)} \left(\|\mathbf{v}_0 - \mathbf{v}^*\|_{\mathbf{M}}^2 + \|\mathbf{v}_0 - \mathbf{v}^*\|_{\mathbf{M}}\|\mathbf{v} - \mathbf{v}^*\|_{\mathbf{M}}\right). \tag{41}$$

We proceed by estimating the last term in (41) again making the dependence of $\mathbf{\Lambda}$ explicit. Thus, recalling that $\mathbf{v} = (\mathbf{x}^*(\mathbf{\Lambda}), \mathbf{y})$ with $(\mathbf{p}, \mathbf{q}) =: \mathbf{y} \in \partial f(\mathbf{K}\mathbf{x}_T(\mathbf{\Lambda}))$ arbitrary, we have

$$\begin{aligned}
\|\mathbf{v} - \mathbf{v}^*(\mathbf{\Lambda})\|_{\mathbf{M}} &\leq C\sqrt{\|\mathbf{x}^*(\mathbf{\Lambda}) - \mathbf{x}^*(\mathbf{\Lambda})\|_2^2 + \|\mathbf{p} - \mathbf{p}^*(\mathbf{\Lambda})\|_2^2 + \|\mathbf{q} - \mathbf{q}^*(\mathbf{\Lambda})\|_2^2} \\
&= C\sqrt{\|\mathbf{p} - \mathbf{p}^*(\mathbf{\Lambda})\|_2^2 + \|\mathbf{q} - \mathbf{q}^*(\mathbf{\Lambda})\|_2^2} \\
&\leq C\sqrt{\|\nabla f_1(\mathbf{A}\mathbf{x}_T(\mathbf{\Lambda})) - \nabla f_1(\mathbf{A}\mathbf{x}^*(\mathbf{\Lambda}))\|_2^2 + 4\overline{\mathbf{\Lambda}}^2} \\
&\leq C\left(L_{\mathbf{z}}\|\mathbf{A}\|\|\mathbf{x}_T(\mathbf{\Lambda}) - \mathbf{x}^*(\mathbf{\Lambda})\|_2 + 2\overline{\mathbf{\Lambda}}\right),
\end{aligned} \tag{42}$$

where the last inequality used the fact that $\sqrt{a^2 + b^2} \leq a+b$ for $a, b \geq 0$. We also used the relationship $\mathbf{p}^*(\mathbf{\Lambda}) = \nabla f_1(\mathbf{A}\mathbf{x}^*(\mathbf{\Lambda}))$, the Lipschitz continuity of ∇f_1, as well as $\mathbf{q}^*(\mathbf{\Lambda}) \in \partial f_2(\nabla \mathbf{x}^*(\mathbf{\Lambda}))$ which implies that $\mathbf{q}^*(\mathbf{\Lambda}) \in B_{\overline{\mathbf{\Lambda}}}$. By combining (41), (49) and (42), and by defining

$$r_0(\mathbf{\Lambda}) := \|\mathbf{v}_0 - \mathbf{v}^*(\mathbf{\Lambda})\|_{\mathbf{M}},$$

we end up to

$$\|\mathbf{x}_T(\mathbf{\Lambda}) - \mathbf{x}^*(\mathbf{\Lambda})\|_2^2 \leq \frac{2}{\lambda_{\min}(\mathbf{A}^{\mathbf{T}}\mathbf{A})\mu_{\mathbf{z}}\sqrt{T}} \left(r_0(\mathbf{\Lambda})^2 + r_0(\mathbf{\Lambda})CL_{\mathbf{z}}\|\mathbf{A}\|\|\mathbf{x}_T(\mathbf{\Lambda}) - \mathbf{x}^*(\mathbf{\Lambda})\|_2 + 2Cr_0(\mathbf{\Lambda})\overline{\mathbf{\Lambda}}\right).$$

By setting $r_T(\mathbf{\Lambda}) := \|\mathbf{x}_T(\mathbf{\Lambda}) - \mathbf{x}^*(\mathbf{\Lambda})\|_2$ and

$$C_1 := \frac{CL_{\mathbf{z}}\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}^{\mathbf{T}}\mathbf{A})\mu_{\mathbf{z}}} \quad C_2 := \frac{4C\overline{\mathbf{\Lambda}}}{\lambda_{\min}(\mathbf{A}^{\mathbf{T}}\mathbf{A})\mu_{\mathbf{z}}} \quad C_3 := \frac{2}{\lambda_{\min}(\mathbf{A}^{\mathbf{T}}\mathbf{A})\mu_{\mathbf{z}}}, \tag{43}$$

we infer

$$r_T(\mathbf{\Lambda})^2 - \frac{2C_1}{\sqrt{T}} r_T(\mathbf{\Lambda}) r_0(\mathbf{\Lambda}) + \frac{C_1^2 r_0(\mathbf{\Lambda})^2}{T} \leq \frac{2}{\sqrt{T}}\left(C_3 r_0(\mathbf{\Lambda})^2 + C_2 r_0(\mathbf{\Lambda})\right) + \frac{C_1^2 r_0(\mathbf{\Lambda})^2}{T}.$$

After applying the binomial formula, this yields

$$\begin{aligned}
r_T(\mathbf{\Lambda}) &\leq \frac{C_1 r_0(\mathbf{\Lambda})}{\sqrt{T}} + \sqrt{\frac{2}{\sqrt{T}}\left(C_3 r_0(\mathbf{\Lambda})^2 + C_2 r_0(\mathbf{\Lambda})\right) + \frac{C_1^2 r_0(\mathbf{\Lambda})^2}{T}} && (44) \\
&\leq \frac{C_{\mathbf{z},\mathbf{A}} r_0(\mathbf{\Lambda})}{\sqrt{T}} + \sqrt{\frac{C_{\mathbf{z},\mathbf{A}}}{\sqrt{T}}\left(r_0(\mathbf{\Lambda})^2 + r_0(\mathbf{\Lambda})\right) + \frac{C_{\mathbf{z},\mathbf{A}}^2 r_0(\mathbf{\Lambda})^2}{T}} && (45) \\
&\leq \frac{3C_{\mathbf{z},\mathbf{A}}}{T^{1/4}}(r_0(\mathbf{\Lambda}) + 1), && (46)
\end{aligned}$$

where the last inequality uses basic estimates, like $\sqrt{T} \leq T$, again $\sqrt{a^2+b^2} \leq a+b$ for $a, b \geq 0$ and the fact that $C_{\mathbf{z},\mathbf{A}} \geq 1$ by its definition (36). This proves (iii).

To show (iv) let $\mathbf{\Lambda}_T \to \mathbf{\Lambda}$ and fix $\overline{\mathbf{\Lambda}} := \sup_{T\in\mathbb{N}} \|\mathbf{\Lambda}_T\|_2$. By (iii) we have that

$$\|\mathbf{x}_T(\mathbf{\Lambda}_T) - \mathbf{x}^*(\mathbf{\Lambda}_T)\|_2 \leq \frac{3C_{\mathbf{z},\mathbf{A}}}{T^{1/4}} \left(1 + \|\mathbf{v}_0 - \mathbf{v}^*(\mathbf{\Lambda}_T)\|_{\mathbf{M}}\right), \tag{47}$$

where we can further estimate by norm-equivalence

$$\|\mathbf{v}_0 - \mathbf{v}^*(\mathbf{\Lambda}_T)\|_{\mathbf{M}} \leq C\sqrt{\|\mathbf{x}_0 - \mathbf{x}^*(\mathbf{\Lambda}_T)\|_2^2 + \|\mathbf{p}_0 - \mathbf{p}^*(\mathbf{\Lambda}_T)\|_2^2 + \|\mathbf{q}_0 - \mathbf{q}^*(\mathbf{\Lambda}_T)\|_2^2}. \tag{48}$$

Using again the boundedness of $(\mathbf{\Lambda}_T)_{T\in\mathbb{N}}$, the continuity of ∇f_1 and the relationships $\mathbf{p}^*(\mathbf{\Lambda}) = \nabla f_1(\mathbf{A}\mathbf{x}^*(\mathbf{\Lambda}))$ and $\mathbf{q}^*(\mathbf{\Lambda}) \in \partial f_2(\nabla \mathbf{x}^*(\mathbf{\Lambda}))$, we conclude that there exists a constant $\hat{C} > 0$ independent of T, such that

$$\|\mathbf{v}_0 - \mathbf{v}^*(\mathbf{\Lambda}_T)\|_{\mathbf{M}} \leq \hat{C}, \quad \text{for every } T \in \mathbb{N}. \tag{49}$$

Thus we deduce

$$\|\mathbf{x}_T(\mathbf{\Lambda}_T) - \mathbf{x}^*(\mathbf{\Lambda}_T)\|_2 \to 0, \quad \text{as } T \to \infty.$$

We finally use the triangle inequality to obtain

$$\|\mathbf{x}_T(\mathbf{\Lambda}_T) - \mathbf{x}^*(\mathbf{\Lambda})\|_2 \leq \|\mathbf{x}_T(\mathbf{\Lambda}_T) - \mathbf{x}^*(\mathbf{\Lambda}_T)\|_2 + \|\mathbf{x}^*(\mathbf{\Lambda}_T) - \mathbf{x}^*(\mathbf{\Lambda})\|_2 \to 0,$$

as $T \to \infty$, where we have also used (i). □

We can now proceed with our main result.

Theorem 7. *Let the Assumption 4 hold, let the training set $\mathcal{D}$ be fixed and consider the sequence of functionals $\mathcal{L}^T : \mathbb{R}^\ell \to \overline{\mathbb{R}}$, $T \in \mathbb{N}$, as well as $\mathcal{L}^* : \mathbb{R}^\ell \to \overline{\mathbb{R}}$ defined as in* (30) *and* (31)*. Then we have that $\mathcal{L}^T$ Γ-converges to $\mathcal{L}^*$ as $T \to \infty$.*

Proof. It suffices to check the conditions in the definition of Γ-convergence [28], i.e.,:

(i) For all $\Theta_T \to \Theta$, it holds $\mathcal{L}^*(\Theta) \leq \liminf_{T\to\infty} \mathcal{L}^T(\Theta_T)$.

(ii) For all $\Theta \in \mathbb{R}^\ell$, there exists $\Theta_T \to \Theta$ such that $\limsup_{T\to\infty} \mathcal{L}^T(\Theta_T) \leq \mathcal{L}^*(\Theta)$.

The first condition holds due to the lower semicontinuity of r, the continuity of the map $\Theta \mapsto \mathbf{\Lambda}_\Theta$, Proposition 6 (iv) and the continuity of the loss function l. The fact that the map $\Theta \mapsto \mathbf{\Lambda}_\Theta$ is continuous, follows by the continuity of all constituent functions, in particular, from the continuity of the activation functions of the U-Net u_Θ. The second condition follows similarly, setting $\Theta_T := \Theta$ for all $T \in \mathbb{N}$ and using the convergence of the iterative scheme, i.e. $S^T(\mathbf{x}_0, \mathbf{z}, \mathbf{\Lambda}) \to S^*(\mathbf{z}, \mathbf{\Lambda})$ as $T \to \infty$, as well as the continuity of the other involved functions. □

The following consistency result follows directly from the Γ-convergence, see [28, Corollary 7.20] for a proof.

Corollary 8 (Consistency of the unrolled scheme)**.** *Let the Assumption 4 hold, let the training set $\mathcal{D}$ be fixed and let $\epsilon_T \to 0$. Suppose that Θ_T is an ϵ_T-minimizer of $\mathcal{L}^T$ i.e. $\mathcal{L}^T(\Theta_T) \leq \inf_{\Theta\in\mathbb{R}^\ell} \mathcal{L}^T(\Theta) + \epsilon_T$. Then, if Θ is an accumulation point of $(\Theta_T)_{T\in\mathbb{N}}$ it is a minimizer of $\mathcal{L}^*$ and $\mathcal{L}^*(\Theta) = \limsup_{T\to\infty} \mathcal{L}^T(\Theta_T)$.*

5 Applications

In the following, we apply our proposed method to several different imaging problems to demonstrate its versatility. The considered imaging problems differ in terms of the operator $\mathbf{A}$ and, more importantly, on the number of dimensions, e.g. 2D, 3D or 2D+time as well as on the specific role the dynamic component plays in the respective problem. For all considered problems, the measurement data was retrospectively simulated from ground-truth images according to (1) by assuming the forward operator to be known. All images were evaluated in terms of PSNR, normalized root mean-squared error (NRMSE), structural similarity index measure [102] (SSIM) and blur effect [26]. `Python` code is available at `github.com/koflera/LearningRegularizationParameterMaps`.

5.1 Initialization for the Unrolled PDHG

In general, an initial image for the PDHG can be directly reconstructed from the measured data by applying the adjoint of the forward operator, i.e., $\mathbf{x}_0 := \mathbf{A}^{\mathsf{H}}\mathbf{z}$. Often, the set-up for realistic imaging problems is that $\mathbf{A}$ is given by a tall operator, i.e., $m > n$. Therefore, to obtain a better estimate of the unknown image from which one can estimate $\mathbf{\Lambda}_\Theta$ by applying the CNN NET_Θ, one can consider the normal equation

$$\mathbf{A}^{\mathsf{H}}\mathbf{A}\mathbf{x} = \mathbf{A}^{\mathsf{H}}\mathbf{z}, \tag{50}$$

and approximately solve it. As $\mathbf{A}$ is typically constructed such that the normal operator $\mathbf{A}^{\mathsf{H}}\mathbf{A}$ is invertible, in the absence of noise, i.e., $\mathbf{z} \in \mathrm{range}(\mathbf{A})$, solving (50) using an iterative scheme to approximate $\mathbf{x}^\dagger := (\mathbf{A}^{\mathsf{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{H}}\mathbf{z}$ would allow for a perfect reconstruction of the ground truth image. However, in the presence of noise, early stopping is required to avoid a noise amplification during the iterations. An approximate solution of (50) can then be used as a better initial estimate for the PDHG method as well as the image from which the CNN NET_Θ estimates the different components of the regularization map $\mathbf{\Lambda}_\Theta$. For the case that the considered imaging problem is not overdetermined, i.e., $m \leq n$, e.g. for image denoising, one simply uses $\mathbf{x}_0 := \mathbf{A}^{\mathsf{H}}\mathbf{z}$ as the input of NET_Θ.

5.2 Dynamic Cardiac MR Image Reconstruction

Here, we apply the proposed NN to a dynamic cardiac MR image reconstruction problem. The problem consists of a set of independent 2D problems from which static images of the heart can be reconstructed. By stacking the different images along time, one can obtain a sequence of images which cover the entire cardiac cycle, also referred to as cardiac cine MRI. In clinical practice, cardiac cine MRI can be used to assess the cardiac function, see e.g. [98]. Due to the structure of the problem, the temporal dimension is the one which offers the greatest potential to exploit the sparsity of the image in its gradient domain. However, a careful choice of the regularization parameter-map is required to ensure that the cardiac motion as well as smaller diagnostic image details are well-preserved after the reconstruction.

5.2.1 Problem Formulation

For a complex-valued dynamic 2D MR image with vector representation $\mathbf{x} \in \mathbb{C}^N$ with $n = n_x \cdot n_y \cdot n_t$, the forward operator in (1) is given as

$$\mathbf{A} := (\mathbf{I}_{n_c} \otimes \mathbf{E})\mathbf{C}, \tag{51}$$

where $\mathbf{I}_{n_c}$ denotes the $n_c \times n_c$-sized identity-operator with n_c being the number of receiver coils used for the data acquisition. The operator $\mathbf{C}$ is a tall operator which contains the coil-sensitivity maps, i.e. $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_{n_c}]^\mathsf{T}$ with $\mathbf{C}_k = \operatorname{diag}(\mathbf{c}_k)$ and $\mathbf{c}_k \in \mathbb{C}^n$, $k = 1, \ldots, n_c$. Let $\mathbf{E}_{I_t} := \mathbf{S}_{I_t}\mathbf{F}$ with $\mathbf{F}$ being the FFT be an operator which acquires the k-space data of a static 2D MR image $\mathbf{x}_t$ at time-point t by sampling the k-space coefficients indexed by the set $I_t \subset J$, where $J = \{1, \ldots, n_{xy}\}$ with $n_{xy} := n_x \cdot n_y$. Thereby, the mask $\mathbf{S}_{I_t} \in \{0,1\}^{m_t \times n_{xy}}$ with $m_t < n_{xy}$ for all $t = 1, \ldots, n_t$ models the undersampling process. Undersampling the Fourier-space data is employed in order to accelerate the data acquisition process which usually takes place during a breathhold of the patient. Finally, the encoding operator $\mathbf{E}$ is given by

$$\mathbf{E} := \begin{pmatrix} \mathbf{E}_{I_1} & \mathbf{0}_{m_1 \times n_{xy}} & \mathbf{0}_{m_1 \times n_{xy}} & \cdots & \mathbf{0}_{m_1 \times n_{xy}} \\ \mathbf{0}_{m_2 \times n_{xy}} & \mathbf{E}_{I_2} & \mathbf{0}_{m_2 \times n_{xy}} & \cdots & \mathbf{0}_{m_2 \times n_{xy}} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0}_{m_{n_t} \times n_{xy}} & \mathbf{0}_{m_{n_t} \times n_{xy}} & \mathbf{0}_{m_{n_t} \times n_{xy}} & \cdots & \mathbf{E}_{I_{n_t}} \end{pmatrix}, \tag{52}$$

where $\mathbf{0}_{m_t \times n_{xy}} \in \{0\}^{m_t \times n_{xy}}$ denotes a $m_t \times n_{xy}$-sized zero-matrix. Note that in our simulations the matrix $\mathbf{C}$ has full rank and thus our consistency analysis holds, see Remark 5. In that case $\mathbf{x}^\dagger$ is also well-defined and can be used as initialization. However, we mention that in real-world scenarios, $\mathbf{C}$ might be close to being rank-deficient by having singular values of small magnitude. The coil-sensitivity maps used in this work were simulated using the function `mrisensesim.py` in the `Python` package `Torch Kb-Nufft`, version 0.3.4 [76].

5.2.2 PDHG for Dynamic Multi-Coil MRI

For the sake of completeness, we briefly summarize the PDHG-algorithm based on the identification mentioned in (23). Recall the definition of f_2 from (23). Since

$$\big(\operatorname{prox}_{\tau f_2^*}(\mathbf{q})\big)_i = \begin{cases} -(\mathbf{\Lambda}_\Theta)_i, & \mathbf{q}_i \in \big(-\infty, -(\mathbf{\Lambda}_\Theta)_i\big) \\ \mathbf{q}_i, & \mathbf{q}_i \in \big[-(\mathbf{\Lambda}_\Theta)_i, (\mathbf{\Lambda}_\Theta)_i)\big] \\ (\mathbf{\Lambda}_\Theta)_i, & \mathbf{q}_i \in \big((\mathbf{\Lambda}_\Theta)_i, \infty\big) \end{cases}, \tag{53}$$

the proximal operator $\operatorname{prox}_{\tau f_2^*}$ acts by "clipping" each entry in the vector $\mathbf{q}$ if its magnitude exceeds the corresponding entry in $\mathbf{\Lambda}_\Theta$ and we therefore abbreviate it as $\operatorname{prox}_{\tau f_2^*} := \operatorname{clip}_{\mathbf{\Lambda}_\Theta}$ to emphasize its dependence on the regularization parameter-map $\mathbf{\Lambda}_\Theta$. The algorithm is summarized in Algorithm 2.

5.2.3 Results

For a detailed description of the experimental set-up we refer to the Section A.1 at the Appendix. Figure 3 shows an example of a single frame of the reconstructed MR image sequences for an acceleration factor of $R = 6$ using several approaches. We show the reconstructions that correspond to the single scalar parameter λ_P^{xyt} as well as to the scalar parameter pair (one spatial and one temporal) $\lambda_\mathrm{P}^{xy,t} = (\lambda_\mathrm{P}^{xy}, \lambda_\mathrm{P}^{xy}, \lambda_\mathrm{P}^{t})$ which are the parameters that maximize the PSNR of entire cine MR image and are obtained via a grid search by making use of the corresponding ground truth image. We also show the results that correspond to the parameters $\lambda_{\bar{\mathrm{P}}}^{xyt}$ and $\lambda_{\bar{\mathrm{P}}}^{xy,t}$ which are respectively the single and the pair of scalar parameters that on average maximize the PSNR over the training set. These were obtained by treating the scalar regularization parameters as trainable parameters and training them by minimizing (29). We finally show the results for our estimated parameter-map $\mathbf{\Lambda}_\Theta^{xy,t}$ with the proposed method. As observed, for all choices of the regularization parameters, the error with respect to the target image was significantly reduced compared to the initial zero-filled

Algorithm 2 Unrolled PDHG algorithm for general linear operator $\mathbf{A}$ with $d(\,\cdot\,,\,\cdot\,) = \frac{1}{2}\|\cdot - \cdot\|_2^2$ and *fixed* regularization parameter-map $\mathbf{\Lambda}_\Theta$ (adapted from [90])

```
Input: $L = \|[\mathbf{A}, \nabla]^{\mathsf{T}}\|$,  $\tau = 1/L$,  $\sigma = 1/L$,  $\theta = 1$,  initial guess $\mathbf{x}_0$
Parameters: number of iterations $T > 0$
Output: reconstructed image $\mathbf{x}_{\mathrm{TV}}$
$\bar{\mathbf{x}}_0 = \mathbf{x}_0$
$\mathbf{p}_0 = \mathbf{0}$
$\mathbf{q}_0 = \mathbf{0}$
for $k = 0, \ldots, T-1$ do
    $\mathbf{p}_{k+1} = \big(\mathbf{p}_k + \sigma(\mathbf{A}\bar{\mathbf{x}}_k - \mathbf{y})\big)/(1+\sigma)$
    $\mathbf{q}_{k+1} = \mathrm{clip}_{\mathbf{\Lambda}_\Theta}(\mathbf{q}_k + \sigma\nabla\bar{\mathbf{x}}_k)$
    $\mathbf{x}_{k+1} = \mathbf{x}_k - \tau\mathbf{A}^{\mathsf{H}}\mathbf{p}_{k+1} - \tau\nabla^{\mathsf{T}}\mathbf{q}_{k+1}$
    $\bar{\mathbf{x}}_{k+1} = \mathbf{x}_{k+1} + \theta(\mathbf{x}_{k+1} - \mathbf{x}_k)$
end for
$\mathbf{x}_{\mathrm{TV}} = \mathbf{x}_T$
```

reconstruction. Further, we can see how the use of the estimated parameter-map yields the most accurate reconstruction and the best preservation of image details.

Figure 4 summarizes the results obtained over the test set with the help of box-plots. Compared to the initial zero-filled reconstruction, an improvement is clearly visible for all choices of the regularization parameter with respect to all reported measures and for all acceleration factors. In addition we see how allowing the temporal direction to be differently regularized than the two spatial dimensions positively influences the results compared to having one global parameter λ (orange vs blue). Last, we see how using the proposed method to estimate an entire spatio-temporal parameter-map $\mathbf{\Lambda}_\Theta$ further surpasses the scalar regularization parameter-maps (green vs orange and blue), especially in terms of SSIM. Table 1 lists the mean and the standard deviation of all TV-reconstructions. The results are consistent with the ones from the box-plots.

Figure 5 shows an example of a spatio-temporal regularization parameter-map which was estimated using the proposed approach for an acceleration factor of $R = 6$. The network u_Θ estimates the regularization parameter-map to be pointwise relatively consistenly higher than the spatially required regularization. This result is in fact expected as the temporal dimension is the one for which the gradients of the images are the sparsest because of the high temporal correlation. Further, we see how the network consistently predicts both the spatial regularization as well as the temporal regularization to be less strong in the area where most of the movement is expected, i.e. in the cardiac region.

Remark 9. *From Algorithm 2, we see that the considered PDHG algorithm for solving problem* (21) *involves the repeated separate application of the forward and the adjoint operators* $\mathbf{A}$ *and* $\mathbf{A}^{\mathsf{H}}$*. Depending on the considered problem - more precisely, on the operator of the data-acquisition - this aspect can be problematic. For example, for non-Cartesian sampling trajectories in MRI, i.e. each* $\mathbf{E}_{I_t}$ *in* (52) *samples the image* $\mathbf{x}_t$ *on a non-Cartesian grid, the separate application of* $\mathbf{A}^{\mathsf{H}}$ *and* $\mathbf{A}$ *is relatively slow as the application of* $\mathbf{E}$ *and* $\mathbf{E}^{\mathsf{H}}$ *involves interpolation operations. The application of the composite operator* $\mathbf{E}^{\mathsf{H}}\mathbf{E}$*, in contrast, can be highly accelerated as it can be efficiently approximated by the Toeplitz-kernel trick [37], see e.g. [71, 93] for applications. Thereby, under some conditions (see [99] for more details), the composition of the forward and the adjoint* $\mathbf{E}^{\mathsf{H}}\mathbf{E}$ *can be represented by* $\mathbf{E}^{\mathsf{H}}\mathbf{E} = \mathbf{F}^{\mathsf{H}}\mathbf{W}\mathbf{F}$*, where* $\mathbf{W}$ *are Toeplitz-kernels which can be estimated depending on*

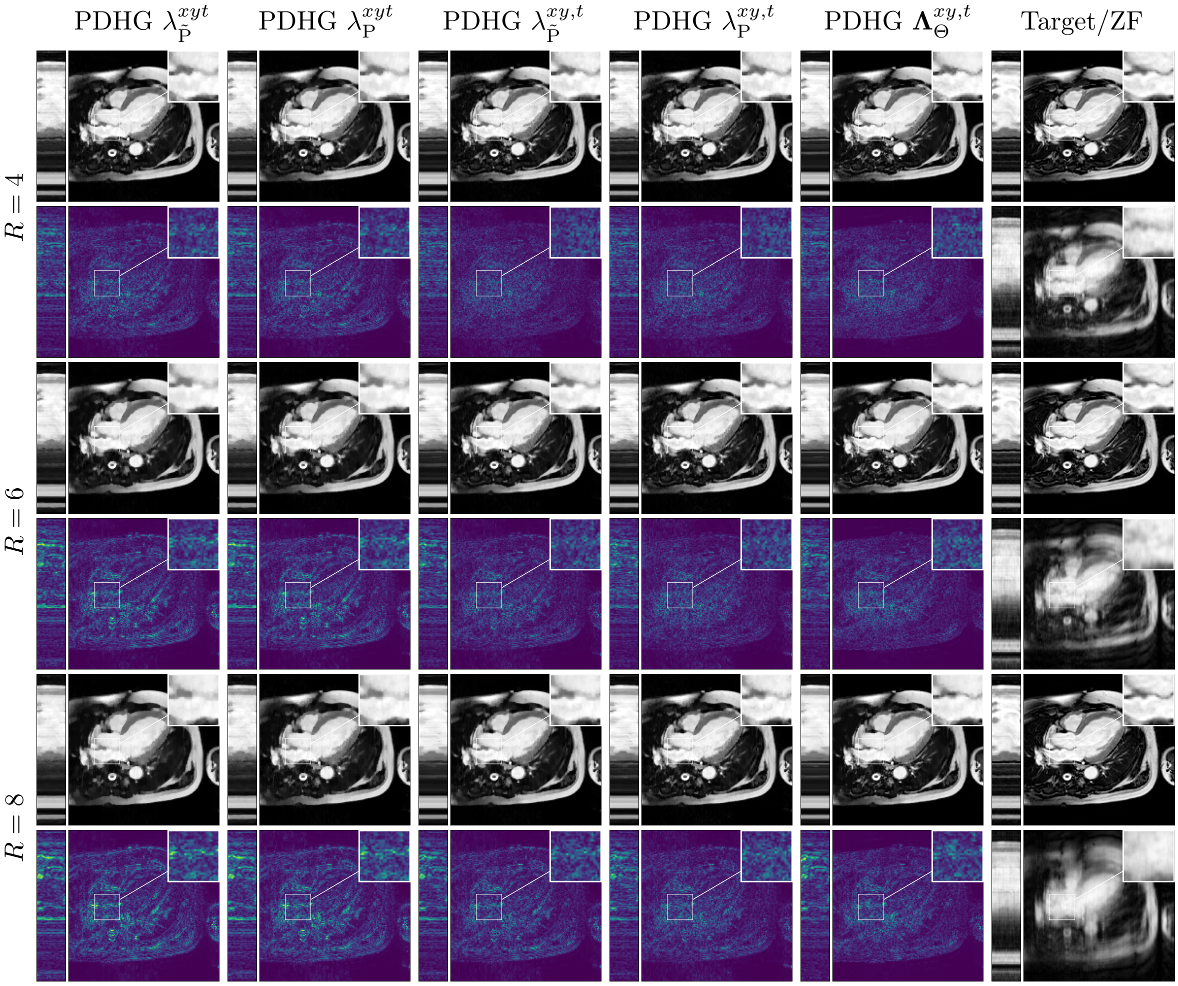
PDHG $\lambda_{\tilde{\mathrm{P}}}^{xyt}$
PDHG $\lambda_{\mathrm{P}}^{xyt}$
PDHG $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$
PDHG $\lambda_{\mathrm{P}}^{xy,t}$
PDHG $\boldsymbol{\Lambda}_{\Theta}^{xy,t}$
Target/ZF
$R = 4$
$R = 6$
$R = 8$

Figure 3: An example of images reconstructed with the primal-dual scheme in Algorithm 2 for different choices of regularization parameters and acceleration factors $R = 4, 6, 8$. Single scalar regularization parameter $\lambda_{\mathrm{P}}^{xyt}$ and $\lambda_{\tilde{\mathrm{P}}}^{xyt}$, two scalar regularization parameters for differently weighted spatial and temporal components, $\lambda_{\mathrm{P}}^{xy,t}$ and $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$, the proposed spatially and temporal dependent parameter-map $\boldsymbol{\Lambda}_{\Theta}^{xy,t}$ obtained with the network $\mathcal{N}_{\Theta}^{T}$. The last column shows the target image and the zero-filled reconstruction. Note again that the results for $\lambda_{\mathrm{P}}^{xyt}$ and $\lambda_{\mathrm{P}}^{xy,t}$ were obtained performing a grid-search for $\lambda^{xyt} > 0$ and $\lambda^{xy}, \lambda^{t} > 0$, assuming the ground truth image to be known. Therefore, the results for $\lambda_{\mathrm{P}}^{xyt}$ and $\lambda_{\mathrm{P}}^{xy,t}$ cannot be obtained in practice and merely serve for illustrating that the proposed $\boldsymbol{\Lambda}_{\Theta}^{xy,t}$ yields competitive results.

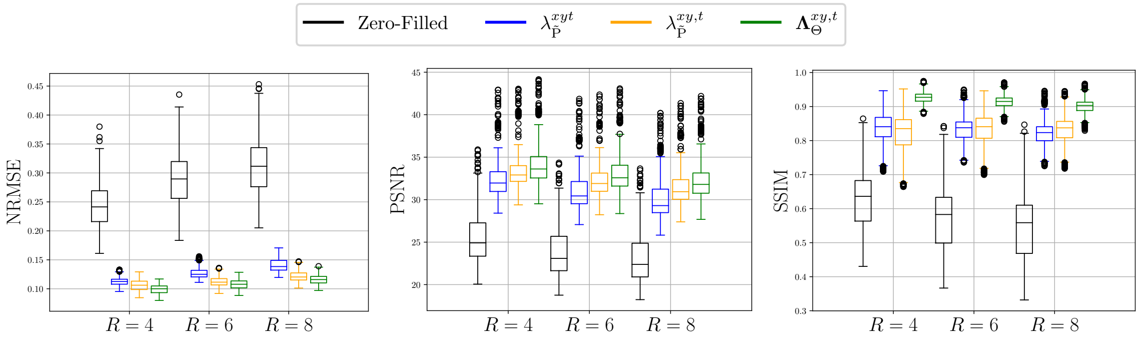

Figure 4: Box-plots summarizing the reconstruction results in terms of PSNR, NRMSE and SSIM obtained with the PDHG for a dynamic cardiac MR image reconstruction problem for different choices of the regularization parameter. Zero-filled reconstruction (black), single scalar regularization parameter ($\lambda_{\tilde{\mathrm{P}}}^{xyt}$, blue), two scalar regularization parameters; one for the spatial x- and y-direction, one for the temporal t-direction ($\lambda_{\tilde{\mathrm{P}}}^{xy,t}$, blue) and the proposed spatially and temporal dependent parameter-map $\mathbf{\Lambda}_{\Theta}^{xy,t}$ obtained with a CNN (NET_{Θ}, green).

the sampling trajectories and where $\mathbf{F}^{\mathsf{H}}$ *and* $\mathbf{F}$ *denote efficient implementations of the FFT on a Cartesian grid. Therefore, choosing a different reconstruction algorithm that requires the application of* $\mathbf{A}^{\mathsf{H}}\mathbf{A}$ *rather than* $\mathbf{A}^{\mathsf{H}}$ *and* $\mathbf{A}$ *separately, e.g. [38, 100], may be a viable option for non-Cartesian MRI, especially given the fact that NNs are typically trained on GPUs.*

5.3 Quantitative MRI Reconstruction

Here, we apply the proposed method to estimate voxel-wise regularization parameter-maps to be used in a quantitative brain MRI reconstruction problem. Similar to the previous case study, the problem consists of different decoupled 2D problems. However, the third temporal dimension contains information about the changing magnetization and thus over time, the contrast of the images changes. Moreover, the speed at which the contrast changes is voxel-depending. This suggests that, different from the previously shown dynamic MRI example, the dynamic component of the estimated regularization parameter-maps should also change over time and thus regularize each time point of the images differently.

5.3.1 Problem Formulation

Formally, the data-acquisition process for quantitative MRI reconstruction problems is given by

$$\mathbf{z} = \mathbf{A}\, q(\mathbf{u}) + \mathbf{e}, \tag{54}$$

where the operator $\mathbf{A}$ takes the exact form as in Subsection 5.2. However, instead of acquiring the k-space data of a sequence of qualitative 2D images $\mathbf{x} = [\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_Q}]^{\mathsf{T}}$ with similar image contrast, the operator $\mathbf{A}$ collects the k-space data of the qualitative images defined by

$$\mathbf{x}_{t_i} = q_{t_i}(\mathbf{u}), \tag{55}$$

where $q_{t_i} : \mathbb{R}^{un} \to \mathbb{C}^n$ relates the vector containing the u quantitative parameters $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_u]^{\mathsf{T}}$ to a qualitative image by a non-linear signal-model q_{t_i} (e.g. solution of Bloch equations evaluated at

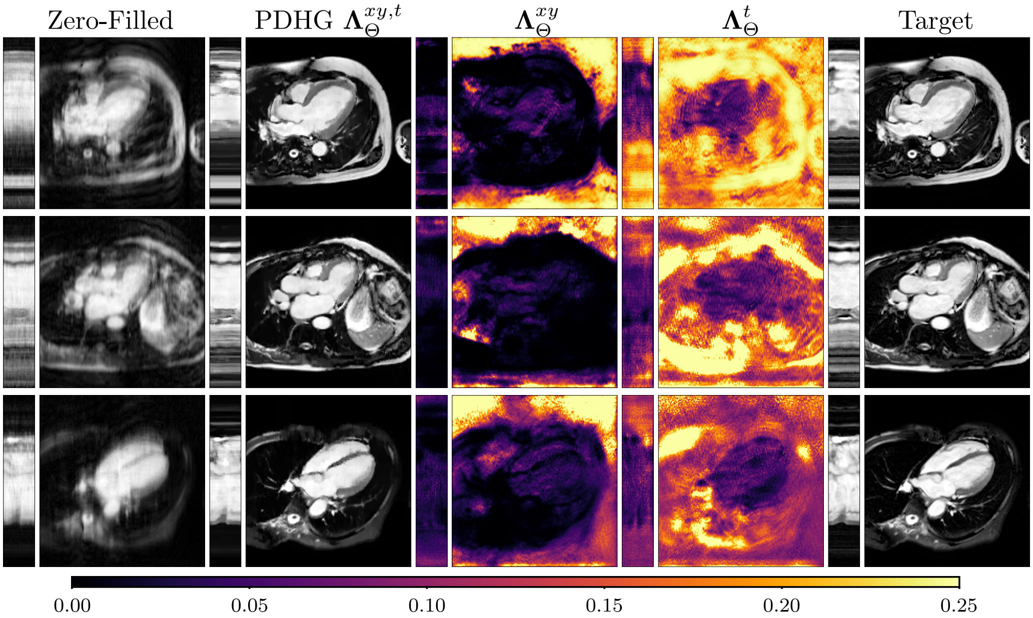
Zero-Filled
PDHG $\mathbf{\Lambda}_\Theta^{xy,t}$
$\mathbf{\Lambda}_\Theta^{xy}$
$\mathbf{\Lambda}_\Theta^{t}$
Target
0.00
0.05
0.10
0.15
0.20
0.25

Figure 5: Different examples of reconstruction results and regularization parameter-maps for an acceleration factor of $R = 6$. From left to right for each row: zero-filled reconstruction, PDHG-reconstruction with $T = 4096$ obtained with the CNN-based regularization spatio-temporal parameter-map $\mathbf{\Lambda}_\Theta^{xy,t} = (\mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{t})$, spatial parameter-map $\mathbf{\Lambda}_\Theta^{xy}$, temporal parameter-map $\mathbf{\Lambda}_\Theta^{t}$ and target ground truth image. Both $\mathbf{\Lambda}_\Theta^{xy}$ and $\mathbf{\Lambda}_\Theta^{t}$ are displayed on the scale $[0, 0.25]$.

t_i). In the following, we will consider the inversion recovery signal model for T_1-mapping given by

$$\begin{aligned} q_{t_i} &: \mathbb{R}^{3n} \to \mathbb{C}^n \\ [\mathbf{T}_1, \mathrm{Re}(\mathbf{M}_0), \mathrm{Im}(\mathbf{M}_0)]^\mathsf{T} &\mapsto q_{t_i}\left(\mathbf{T}_1, \mathrm{Re}(\mathbf{M}_0), \mathrm{Im}(\mathbf{M}_0)\right) = \mathbf{M}_0(1 - 2e^{-t_i/\mathbf{T}_1}), \end{aligned} \tag{56}$$

where the vector $\mathbf{T}_1$ denotes the longitudinal relaxation times for all pixels and $\mathrm{Re}(\mathbf{M}_0)$ and $\mathrm{Im}(\mathbf{M}_0)$ denote real and imaginary parts of the equilibrium magnetization, respectively [40].

Note that in quantitative MR imaging, one is ultimately interested in the quantities contained in the vector $\mathbf{u}$. However, often, qualitative images are first reconstructed (using some regularization method) as an intermediate step, from which then the vector $\mathbf{u}$ is estimated in a second step using non-linear regression methods, see for example [92]. We can formulate the image reconstruction problem by

$$\begin{cases} \min\limits_{\mathbf{u}} \dfrac{1}{2} \displaystyle\sum_{i=1}^{N_t} \|\mathbf{x}_{t_i} - q_{t_i}(\mathbf{u})\|_2^2, \\ \text{where } \ \mathbf{x} = \operatorname*{argmin}\limits_{\tilde{\mathbf{x}}} \dfrac{1}{2}\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{z}\|_2^2 + \|\mathbf{\Lambda}_\Theta \nabla \tilde{\mathbf{x}}\|_1. \end{cases} \tag{57}$$

First, we train the proposed NN to estimate appropriate pixel-dependent regularization parameter-maps $\mathbf{\Lambda}_\Theta^{xy,t}$ to solve the TV-minimization problem and obtain $\mathbf{x}$, and in a second step, perform a pixel-wise regression to obtain the vector $\mathbf{u}$.

5.3.2 Results

For a detailed description of the experimental set-up we refer again to the Appendix and in particular to the Section A.2. Figure 6 shows examples of the quantitative (magnitude) images $\mathbf{u}$ of three of the 112 simulated inversion recovery measurements in the test dataset. We also show the regularization parameter-maps for regularization along the spatial directions and along the inversion-time direction generated by the network. The mean PSNR and SSIM of our proposed method is consistently higher for all considered acceleration factors, even compared to PDHG with regularization strength along spatial and inversion-time direction chosen by grid-search with access to the ground truth images (shown in Figure 8 and Table 2). Qualitatively, the reconstruction was reduced by the proposed method without resulting in noticeably increased blur. The regularization parameter-maps show strong spatial regularization within areas of homogeneous tissue and increased temporal regularization at the transitions between different tissues.

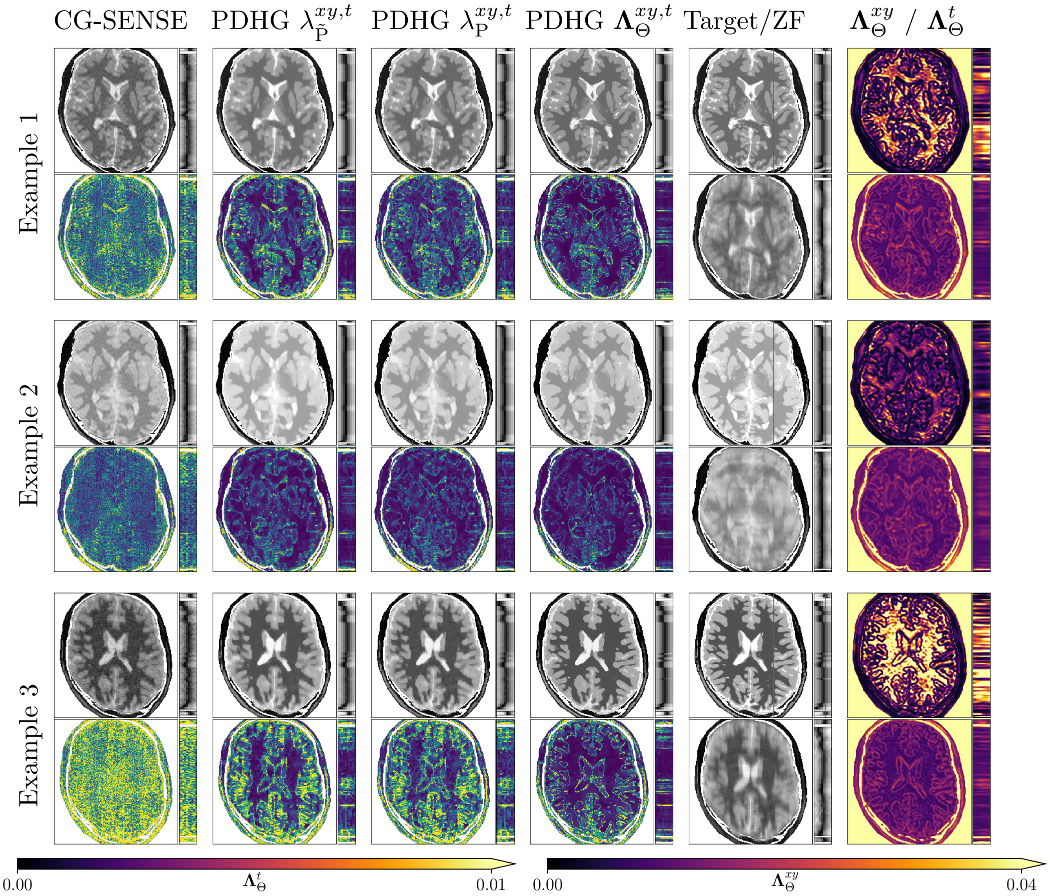
CG-SENSE
PDHG $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$
PDHG $\lambda_{\mathrm{P}}^{xy,t}$
PDHG $\boldsymbol{\Lambda}_{\Theta}^{xy,t}$
Target/ZF
$\boldsymbol{\Lambda}_{\Theta}^{xy}$ / $\boldsymbol{\Lambda}_{\Theta}^{t}$
Example 1
Example 2
Example 3
0.00
$\boldsymbol{\Lambda}_{\Theta}^{t}$
0.01
0.00
$\boldsymbol{\Lambda}_{\Theta}^{xy}$
0.04

Figure 6: Three exemplary sets of reconstructed, qualitative magnitude images (all acceleration factor 6, $\sigma = 0.10$ / 0.15 / 0.24) and absolute error compared to the ground truth (bottom rows). The last column shows spatial (top) and temporal (bottom) regularization strengths maps for the PDHG reconstruction generated by the CNN, the second to last column shows the synthetic ground truth as well as the zero-filled reconstruction (bottom rows).

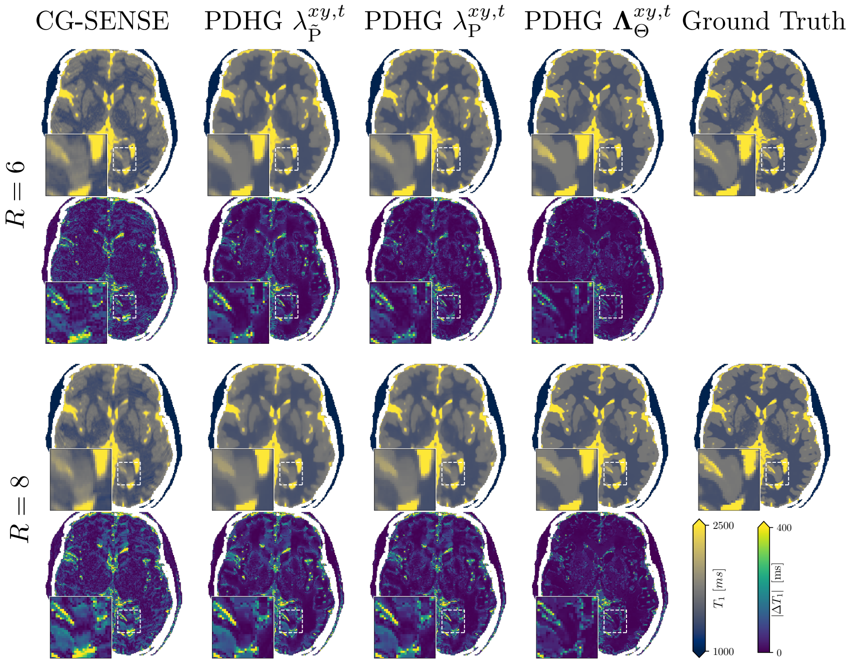
CG-SENSE
PDHG $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$
PDHG $\lambda_{\mathrm{P}}^{xy,t}$
PDHG $\mathbf{\Lambda}_{\Theta}^{xy,t}$
Ground Truth
$R = 6$
$R = 8$
2500
T_1 [ms]
1000
400
$|\Delta T_1|$ [ms]
0

Figure 7: Resulting quantitative T_1 parameter-maps after performing the regression on the reconstructed magnitude images and absolute errors compared to the ground truth (bottom row). Shown is *Example 2* of Figure 6, at acceleration 6 and 8.

The resulting T_1 parameter-maps after performing the regression on the reconstructed images are shown in Figure 7. Again, our proposed method results in the lowest RMS deviation from the ground truth images (Table 2).

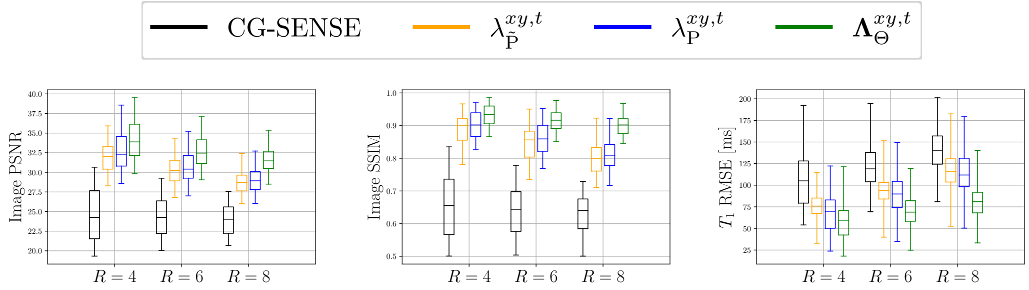

Figure 8: Quantification of the quality of the reconstructed magnitude images in terms of PSNR and SSIM, and RMSE of the T_1 values after performing the signal regression. We compare our proposed approach with a CG-SENSE reconstruction [81], and the TV-based reconstructions with parameters $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$ and $\lambda_{\mathrm{P}}^{xy,t}$.

5.4 Dynamic Image Denoising

Here, we apply the proposed method to estimate voxel-wise dependent regularization parameter-maps to be used in a dynamic image denoising problem. An important difference to the previously considered cardiac MRI example is that, while in the latter, a clear inherent distinction between the black background and the object of interest is possible, for the next videos to be considered, this is not the case. The samples might show scenes with static camera position and only moving objects or scenes in which also the camera-position changes over time.

5.4.1 Problem Formulation

The real-valued noisy video samples are denoted by $\mathbf{x} \in \mathbb{R}^n$ with $n = n_x \cdot n_y \cdot n_t$. The forward operator for the dynamic denoising problem is simply given by an $n \times n$ identity operator, i.e. $\mathbf{A} = \mathbf{I}_n$.

5.4.2 Results

We refer to the Section A.3 for description of the experimental set-up. We compare the 2D time frames of the video samples from the test dataset to the denoised frames, regularized by $\lambda_{\tilde{\mathrm{P}}}^{xyt}$, $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$ and by the spatio-temporal parameter-map $\mathbf{\Lambda}_{\Theta}^{xy,t}$. The metrics were calculated frame-wise for all samples at three different noise levels, characterized by the standard deviation of the Gaussian distribution. From the box-plots in Figure 10 we see that the PDHG reconstructions using the proposed spatio-temporal regularization parameter-map yield superior reconstructions compared to $\lambda_{\tilde{\mathrm{P}}}^{xyt}$ and $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$ with respect to all measures. Table 3 quantitatively summarizes the results. In Figure 9, we compare two samples from the test dataset with a static camera view in the first row and a dynamic camera view in the third row. The vertical red lines in Figure 9 indicate the x-location of the yt-excerpt shown to the left of each image. The second and the fourth row show the pointwise absolute errors of the respective images. For both samples, the lowest error is achieved by the $\mathbf{\Lambda}_{\Theta}^{xy,t}$ parameter-map. The spatial and the temporal components of the obtained regularization parameter-maps $\mathbf{\Lambda}_{\Theta}^{xy,t}$ are visualized in Figure 11. Here, the noisy samples, the results obtained with PDGH using $\mathbf{\Lambda}_{\Theta}$, the spatially and temporally dependent $\mathbf{\Lambda}_{\Theta}^{xy,t}$ parameter-maps and the ground truth-images are depicted. By comparing the static and dynamic case, we see that the trained CNN is able to differentiate between the two inherently different cases. Thereby, for the video sample with the static camera position, where the background remains constant over time and only objects are changing position, the CNN imposes an overall higher temporal regularization. For the video sample where the camera position also changes over time, the CNN is able to predict the less prominent potential to exploit the temporal gradient-sparsity and thus assigns relatively low

5.5 Low-Dose Computerized Tomography

In the last section, we show an application of our proposed method to a static 2D low-dose CT reconstruction problem. Because of the different noise statistics and as a result, a different fidelity term, see (58) below, the problem requires the use of a reconstruction algorithm different than PDHG which shows that our proposed method can be used in conjunction with any iterative scheme. We mention however that since this fidelity term is not strongly convex, the consistency results of Section 4 cannot be applied in this case. We leave the corresponding extension of these results to the CT case for future work.

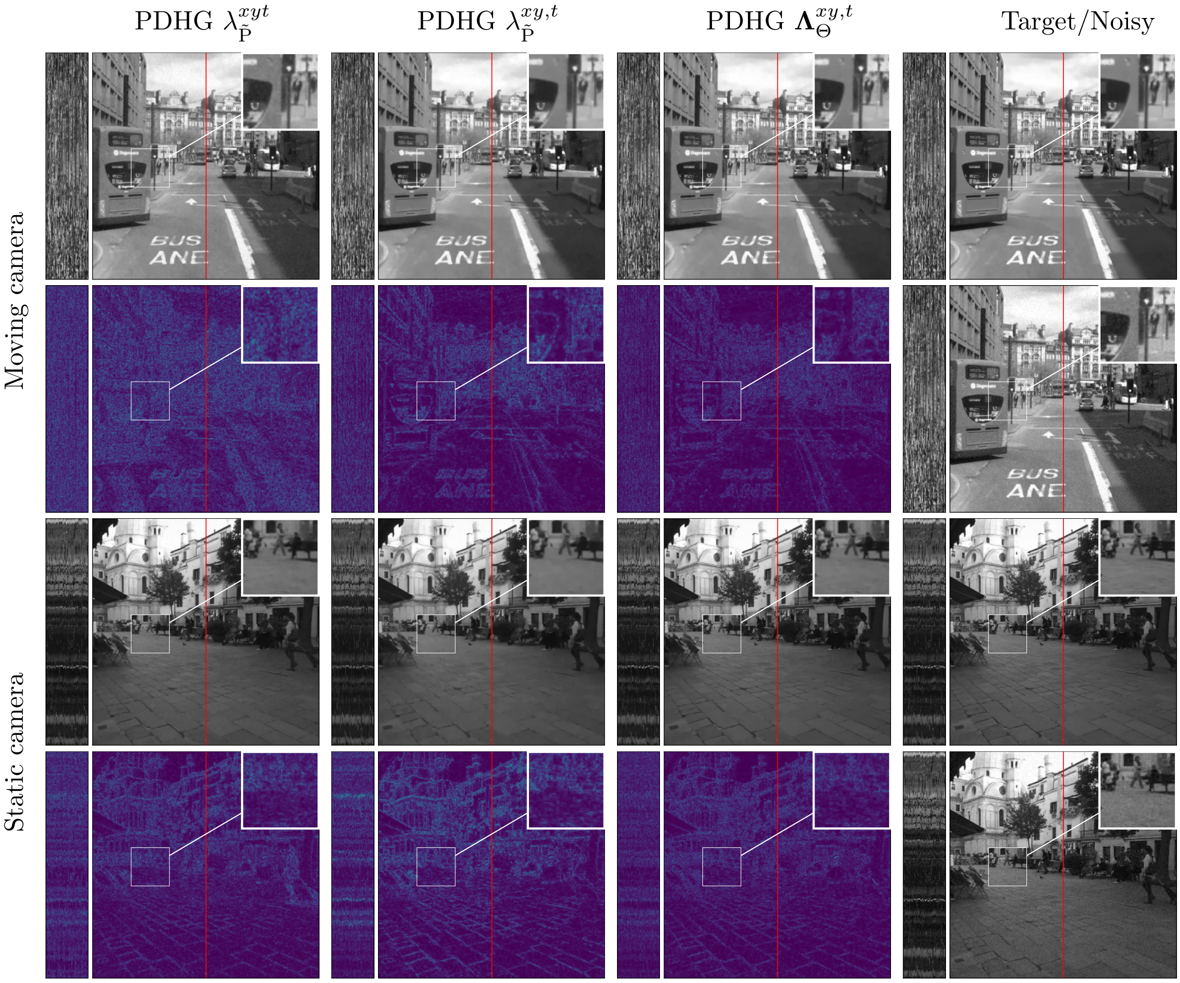
PDHG $\lambda_{\tilde{\mathrm{P}}}^{xyt}$
PDHG $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$
PDHG $\mathbf{\Lambda}_{\Theta}^{xy,t}$
Target/Noisy
Moving camera
Static camera

Figure 9: An example of dynamic denoising with the PDHG from Algorithm 2 for different choices of regularization parameters at moving (`https://motchallenge.net/vis/MOT17-14`) and static (`https://motchallenge.net/vis/MOT17-01`) camera view. Single scalar regularization parameter $\lambda_{\tilde{\mathrm{P}}}^{xyt}$, two scalar regularization parameters for differently weighted spatial and temporal components $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$, and the proposed spatially and temporally dependent parameter-map $\mathbf{\Lambda}_{\Theta}^{xy,t}$ obtained with the network $\mathcal{N}_{\Theta}^{T}$. The last column shows the target image and the noisy sample. The row underneath the denoised image shows the error map.

5.5.1 Problem Formulation

We consider the proposed NN for the low-dose Computerized Tomography (CT) setting. Here a two-dimensional parallel beam geometry is chosen and the corresponding ray transform is given by the Radon transform [82]. As forward operator, we then consider the discretized Radon transformation, which is a finite-dimensional linear map $\mathbf{A}\colon \mathbb{R}^n \to \mathbb{R}^m$, where n is the dimension of the image space and m is the product between the number of angles of the measurement and the number of the

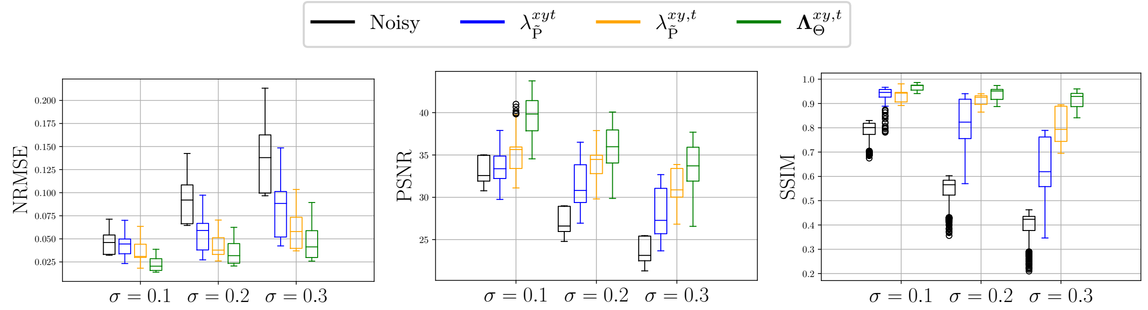

Figure 10: Box-plots summarizing the results in terms of PSNR, NRMSE and SSIM obtained with the PDHG algorithm for dynamic denoising. Single scalar regularization parameter ($\lambda_{\tilde{\mathrm{P}}}^{xyt}$, blue), two scalar regularization parameters; one for the spatial x- and y-direction, one for the temporal t-direction ($\lambda_{\tilde{\mathrm{P}}}^{xy,t}$, orange) and the proposed spatially and temporally dependent parameter-map $\mathbf{\Lambda}_{\Theta}^{xy,t}$ obtained with a CNN (NET$_\Theta$, green).

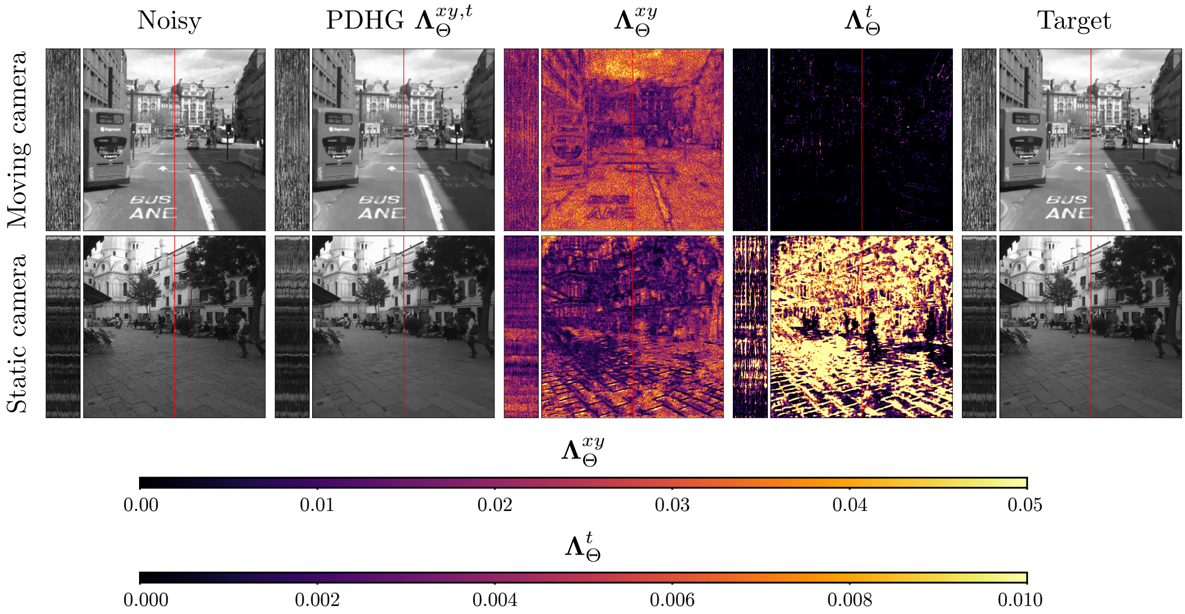
Noisy
PDHG $\mathbf{\Lambda}_{\Theta}^{xy,t}$
$\mathbf{\Lambda}_{\Theta}^{xy}$
$\mathbf{\Lambda}_{\Theta}^{t}$
Target
Moving camera
Static camera
$\mathbf{\Lambda}_{\Theta}^{xy}$
0.00
0.01
0.02
0.03
0.04
0.05
$\mathbf{\Lambda}_{\Theta}^{t}$
0.000
0.002
0.004
0.006
0.008
0.010

Figure 11: Two examples with moving (top row) and static (bottom row) camera view. The columns show the noisy sample, the denoised sample, spatial and temporal dependent parameter-maps $\mathbf{\Lambda}_{\Theta}^{xy}$ and $\mathbf{\Lambda}_{\Theta}^{t}$ and the ground truth sample. By comparing the top and the bottom row, we can see that the CNN-block NET$_\Theta$ is able to differentiate between scenes with static and dynamic camera positions as it exploits the higher temporal correlation in the first by assigning higher temporal regularization values and at the same time provides lower temporal regularization values when there is less temporal correlation to exploit due to the moving camera position.

equidistant detector bins. Then we can formulate the inverse problem as

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{e}, \text{ where } \mathbf{e} = -\mathbf{A}\mathbf{x} - \log(\tilde{\mathbf{N}}_{\mathbf{1}}/N_0) \text{ and } \tilde{\mathbf{N}}_{\mathbf{1}} \sim \mathrm{Pois}(N_0 \exp(-\mathbf{A}\mathbf{x}\mu)),$$

where μ is a normalization constant and N_0 denotes the mean photon count per detector bin without attenuation. Note that here we do not have Gaussian noise, but some noise which follows the negative log-transformation of a Poisson distribution. Therefore, the data-discrepancy in (3) is not the L^2-error, and the correct term can be derived from a Bayesian viewpoint, where the data-discrepancy corresponds to the negative log-likelihood $p_{Y|X=x}$. Using that the negative log-likelihood of a Poisson distributed random variable is given by the Kullback-Leibler divergence, the resulting data-discrepancy can be written as

$$d(\mathbf{Ax},\mathbf{z}) = \sum_{i=1}^{m} e^{-(\mathbf{Ax})_i\mu}N_0 - e^{-\mathbf{z}_i\mu}N_0\big(-(\mathbf{Ax})_i\mu + \log(N_0)\big), \tag{58}$$

see, e.g,. [7, 61] for more details. Consequently, we cannot use Algorithm 2 for reconstruction and a reformulation of Algorithm 1 for this data-discrepancy does not lead to a closed form, see Appendix A.4.

As a remedy we consider the primal-dual algorithm PD3O [106], which is a generalization of the PDHG method. The PD3O aims to minimize the sum of proper, lower semi-continuous and convex functions

$$\min_{\mathbf{x}} f(\mathbf{Kx}) + g(\mathbf{x}) + h(\mathbf{x}),$$

where $\mathbf{K}\colon V^n \to V^{\tilde{m}}$ is a bounded linear operator, h is differentiable with a Lipschitz continuous gradient and for both g and f^* the proximal operator has a analytical solution. The general scheme of PD3O is described in Algorithm 3

Algorithm 3 Unrolled PD3O algorithm (adapted from [106])

```
Input: $L = \mathrm{Lip}(\nabla h)$,  $\tau = 2/L$,  $\sigma = 1/(\tau\|\mathbf{KK}^\mathsf{T}\|)$,  initial guess $\bar{\mathbf{x}}_0$
Output: reconstructed image $\mathbf{x}_{\mathrm{TV}}$
$\mathbf{p}_0 = \bar{\mathbf{x}}_\mathbf{0}$
$\mathbf{q}_0 = \mathbf{0}$
for $k = 0, \dots, T-1$ do
    $\mathbf{q}_{k+1} = \mathrm{prox}_{\sigma f^*}(\mathbf{q}_k + \sigma\mathbf{K}\bar{x}_k)$
    $\mathbf{p}_{k+1} = \mathrm{prox}_{\tau g}(\mathbf{p}_k - \tau\nabla h(\mathbf{p}_k) - \tau\mathbf{K}^\mathsf{T}\mathbf{q}_{k+1})$
    $\bar{\mathbf{x}}_{k+1} = 2\mathbf{p}_{k+1} - \mathbf{p}_k + \tau\nabla h(\mathbf{p}_k) - \tau\nabla h(\mathbf{p}_{k+1})$
end for
$\mathbf{x}_{\mathrm{TV}} = \mathbf{x}_T$
```

Note that we recover the PDHG algorithm if we set $h = 0$. For application of PD3O to our CT case we define

$$f(\mathbf{q}) = \|\mathbf{\Lambda}\,\mathbf{q}\|_1, \quad g(\mathbf{p}) = \iota_{\{\mathbf{p}\geq 0\}}(\mathbf{p}) = \begin{cases} 0 & \text{if } \mathbf{p} \geq 0, \\ +\infty & \text{else,} \end{cases}$$

$$h(\mathbf{p}) = \sum_{i=1}^{m} e^{-\mathbf{p}_i\mu}N_0 - e^{-\mathbf{z}_i\mu}N_0\big(-\mathbf{p}_i\mu + \log(N_0)\big), \quad \mathbf{K} = \nabla.$$

The proximal operator of f^* is already given in (53), the proximal operator of g is given by

$$\mathrm{prox}_{\tau g}(\mathbf{z}) = \mathrm{ReLU}(\mathbf{z}) = \begin{cases} \mathbf{z} & \text{if } \mathbf{z} \geq 0, \\ 0 & \text{else,} \end{cases}$$

and ∇h is given by

$$\nabla h(\mathbf{p}) = \mu N_0 \mathbf{A}^{\mathsf{T}}\big(- e^{-\mathbf{A}\mathbf{p}\mu} + e^{-\mathbf{z}\mu}\big).$$

Note that ∇h is not globally Lipschitz continuous, but due to the non-negativity constraint g we only have to consider ∇h for $\mathbf{p}$ with non-negative entries. Consequently, we can find an upper bound of the Lipschitz constant of ∇h by $\mathrm{Lip}(\nabla h) \leq \|\mathbf{A}\|^2\mu^2 N_0$. The resulting scheme we use for CT reconstruction is summarized in Algorithm 4.

Algorithm 4 Unrolled PD3O algorithm for general linear operator $\mathbf{A}$ with $h(\,\cdot\,) = d(\mathbf{A}\cdot\,, \mathbf{z})$, d as in (58) and *fixed* regularization parameter-map $\mathbf{\Lambda}$ (adapted from [106])

```
Input: $L = \mathrm{Lip}(\nabla h)$,  $\tau = 2/L$,  $\sigma = 1/(\tau\|\nabla\|)$,  initial guess $\bar{\mathbf{x}}_{\mathbf{0}}$
Output: reconstructed image $\mathbf{x}_{\mathrm{TV}}$
$\mathbf{p}_0 = \bar{\mathbf{x}}_{\mathbf{0}}$
$\mathbf{q}_0 = \mathbf{0}$
for $k = 0, \ldots, T-1$ do
    $\mathbf{q}_{k+1} = \mathrm{clip}_{\mathbf{\Lambda}}(\mathbf{q}_k + \sigma\nabla\bar{\mathbf{x}}_k)$
    $\mathbf{p}_{k+1} = \mathrm{ReLU}(\mathbf{p}_k - \tau\mu N_0 \mathbf{A}^{\mathsf{T}}(e^{-\mathbf{z}\mu} - e^{-\mathbf{A}\mathbf{p}_k\mu}) - \tau\nabla^{\mathsf{T}}\mathbf{q}_{k+1})$
    $\bar{\mathbf{x}}_{k+1} = 2\mathbf{p}_{k+1} - \mathbf{p}_k + \tau\mu N_0 \mathbf{A}^{\mathsf{T}}(e^{-\mathbf{A}\mathbf{p}_{k+1}\mu} - e^{-\mathbf{A}\mathbf{p}_k\mu})$
end for
$\mathbf{x}_{\mathrm{TV}} = \bar{\mathbf{x}}_T$
```

5.5.2 Results

We refer to the Section A.5 for details on the experimental set-up. In Figure 12 we compare the PD3O reconstructions (top) and their corresponding errors with respect to the ground truth (bottom) using different regularization parameter choices $\lambda^{xy}_{\tilde{P}}$, λ^{xy}_{P} and $\mathbf{\Lambda}_\Theta$ for PD3O. Obviously, using the estimated parameter-map $\mathbf{\Lambda}_\Theta$ leads to a significant improvement of the reconstruction. In particular, sharp edges are retained, while using a constant regularizing parameter results in a significant blur. This can be also seen in Table 4, where we compare the NRMSE, PSNR, SSIM and blur and evaluated on the first 100 test images of the LoDoBaP dataset. These results are visualized in Figure 13 using box-plots. Note that the FBP seems to better than PD3O-$\lambda^{xy}_{\tilde{P}}$ in terms of the blur effect, but this can be explained by the fact that FBP reconstructions admit a lot of high-frequency artefacts leading to a small blur effect.

Further PD3O-$\mathbf{\Lambda}_\Theta$ reconstructions with their corresponding estimated parameter-maps $\mathbf{\Lambda}_\Theta$ are shown in Figure 14. Note that the parameter-maps are given in a logarithmic scale. As expected, the regularization is strong in constant areas and less strong on edges or finer details in order to reduce a smoothing in these regions.

5.6 Choosing the Number of Iterations T

Since our proposed method to obtain the regularization parameter-map $\mathbf{\Lambda}_\Theta$ is based on unrolling an iterative algorithm as PDHG or PD3O using a fixed number of iterations T, questions about how to choose T during training as well as at inference time are relevant. Recall from Section 4 that $\mathbf{x}_T := S^T(\mathbf{z}, \mathbf{\Lambda})$ and $\mathbf{x}^* := S^*(\mathbf{z}, \mathbf{\Lambda})$ denote the T-th iterate and the exact solution of problem (4), respectively. For addressing questions about the choice of T at training and testing time, here, we emphasize the dependence of the solutions $\mathbf{x}_T$ and $\mathbf{x}^*$ on the set of parameters Θ, by writing

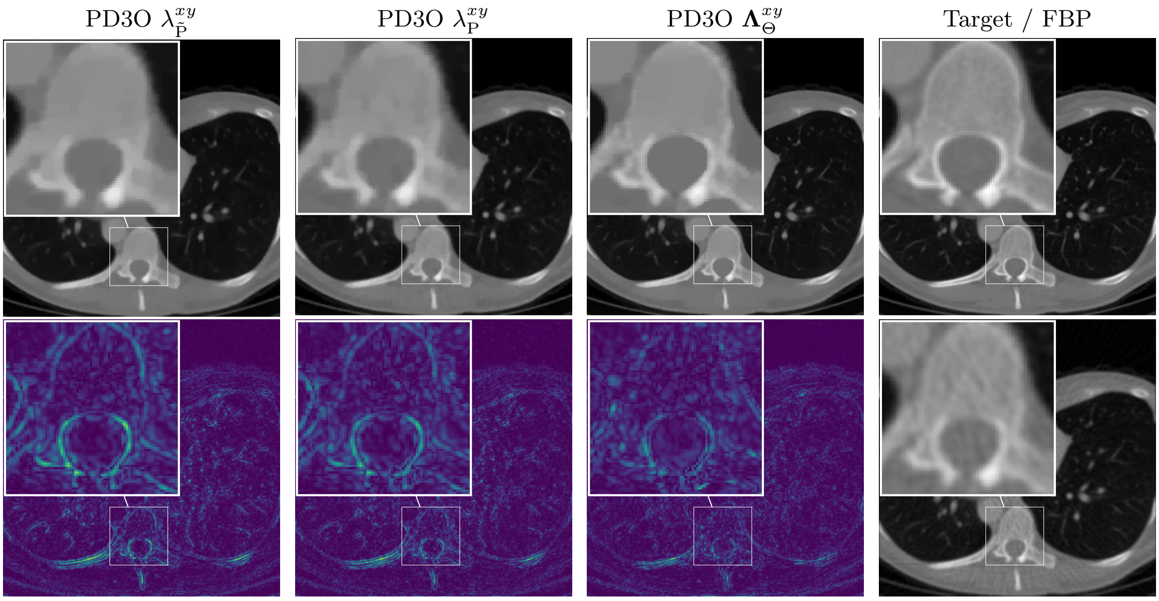
PD3O $\lambda^{xy}_{\tilde{\mathrm{P}}}$
PD3O $\lambda^{xy}_{\mathrm{P}}$
PD3O $\mathbf{\Lambda}^{xy}_{\Theta}$
Target / FBP

Figure 12: Reconstruction of the ground truth CT image using different choices of regularization parameters. The last column shows the ground truth and the FBP reconstruction. *Top*: full image. *Bottom*: difference to the ground truth.

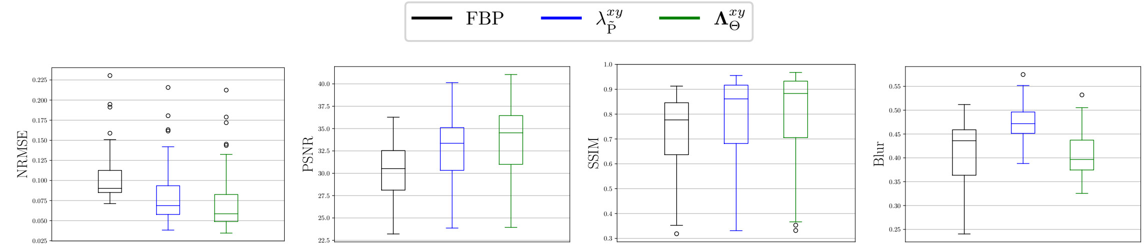

Figure 13: Box-plots summarizing the reconstruction results in terms of PSNR, NRMSE, SSIM and blur effect obtained with the PD3O algorithm for a CT reconstruction problem for different choices of the regularization parameter. Filtered back-projection (black), single scalar regularization parameter ($\lambda^{xy}_{\tilde{\mathrm{P}}}$, blue) and the proposed parameter-map $\mathbf{\Lambda}_{\Theta}$ obtained with a CNN (NET$_{\Theta}$, green).

$\mathbf{x}_T(\Theta) := S^T(\mathbf{z}, \mathbf{\Lambda}_{\Theta})$ and $\mathbf{x}^*(\Theta) := S^*(\mathbf{z}, \mathbf{\Lambda}_{\Theta})$ and by denoting Θ_T as the set of trainable parameters which is obtained by training the network which unrolls using T iterations of PDHG or PD3O.

We illustrate the following considerations relying on results obtained for the dynamic cardiac MRI application shown in Subsection 5.2 but point out that these could be derived from the other applications examples as well.

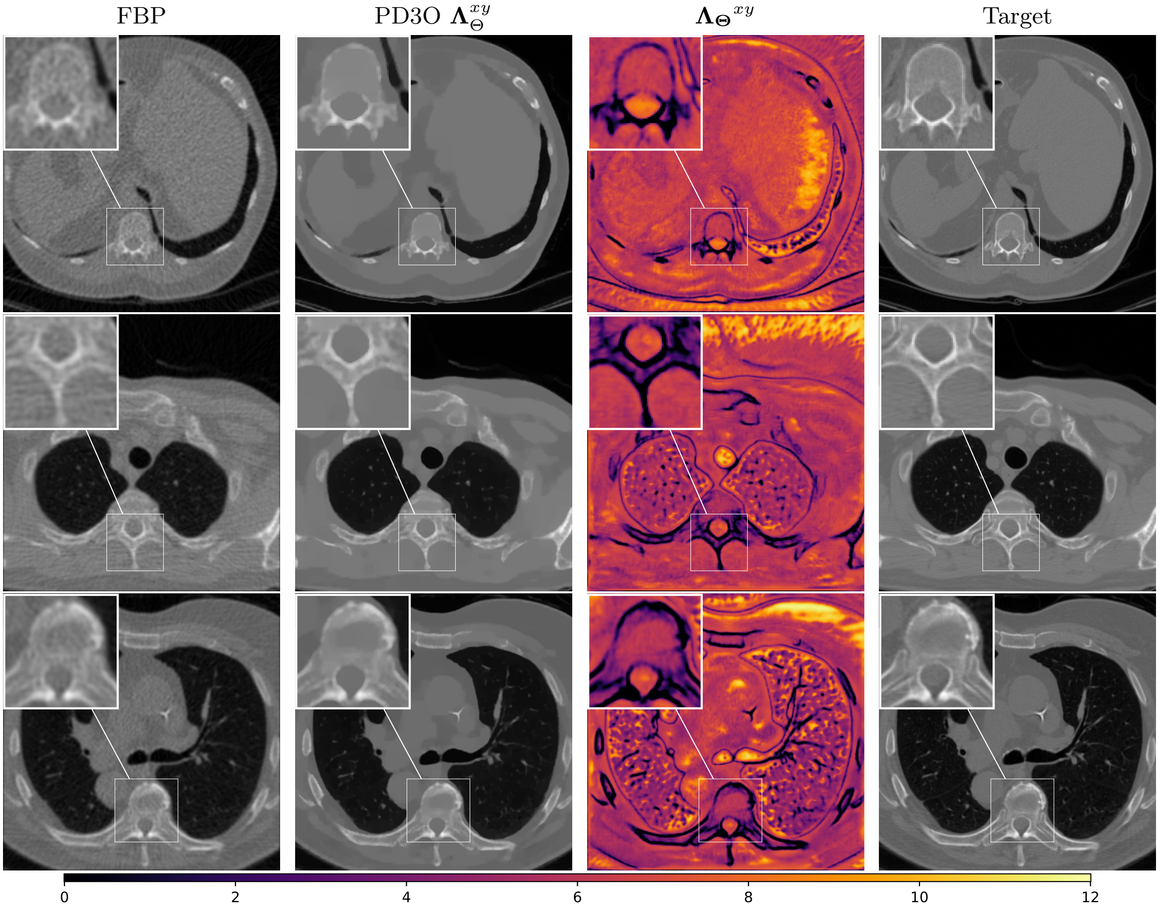
FBP
PD3O $\mathbf{\Lambda}_\Theta^{xy}$
$\mathbf{\Lambda_\Theta}^{xy}$
Target
0
2
4
6
8
10
12

Figure 14: Different reconstructions obtained with PD3O employing the regularization parameter-maps obtained with the proposed CNN. From left to right: initial FBP-reconstruction, PD3O-reconstruction, spatial regularization parameter-map and ground truth image. As can be seen, the network attributes higher regularization parameters to image content with smooth structures while it yields lower regularization parameters at the edges to prevent smoothing.

5.6.1 Choosing T at Training Time

Clearly, the obvious choice is to set T as high as possible during training. The reason is to hope to be able to have $\mathbf{x}_T(\Theta) \approx \mathbf{x}^*(\Theta)$ and therefore to optimize Θ such that its optimal when the reconstruction algorithm given by $\mathcal{N}_\Theta^T$ is run until convergence. However, choosing a high T increases training times as well as hardware requirements. Thus, one could on purpose choose or be forced to choose to use a lower T for training and hope that the sub-network NET_Θ is flexible enough to compensate for that.

Figure 15 shows the validation error during training of an unrolled PDHG for the dynamic MRI example for different T. As can be seen, for smaller T, the NNs' ability to accurately reconstruct the images is clearly reduced. Further, Figure 16 shows an example of different regularization parameter-maps which were obtained by training using a different number of iterations T. It shows that indeed, the obtained regularization parameter-maps vary depending on the number of iterations chosen for training, although they seem to share local features. Clearly, when T is set too low, the network tends to yield higher regularization parameter-maps to try to compensate for the low number of iterations. However, from Figure 15 one cannot infer whether the limited reconstruction accuracy is attributable to the too low number of iterations, a resulting sub-optimal Θ_T or combinations thereof.

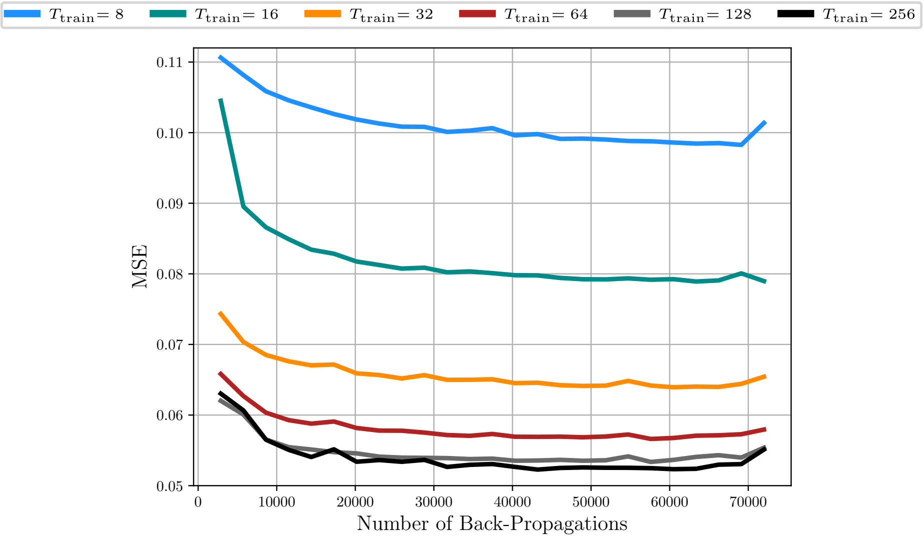

Figure 15: The validation error during training of the proposed method for the previously shown dynamic cardiac MRI example for different number of iterations T_{train} which were used for unrolling PDHG. We see that the rate of convergence for obtaining the set of parameters Θ_{train} is comparable, but using higher T_{train} results in more accurate reconstructions.

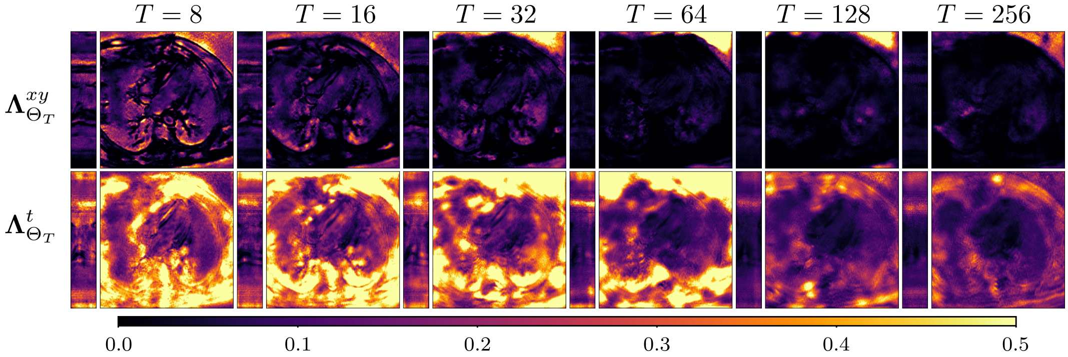
$T = 8$
$T = 16$
$T = 32$
$T = 64$
$T = 128$
$T = 256$
$\mathbf{\Lambda}^{xy}_{\Theta_T}$
$\mathbf{\Lambda}^{t}_{\Theta_T}$
0.0
0.1
0.2
0.3
0.4
0.5

Figure 16: An example of different spatio-temporal regularization parameter-maps $\mathbf{\Lambda}^{xy,t}_{\Theta_T}$ for the dynamic cardiac MRI example obtained by training the unrolled network $\mathcal{N}^T_\Theta$ using different numbers of iterations T. All images are shown on the scale $[0, 0.5]$. Although for different T the regularization parameter-maps have a similar structure - for example, the cardiac region is consistently regularized less strongly along time - for lower T, the values tend to be higher in general. Intuitively speaking, the network estimates higher regularization values in order to compensate for the lower T.

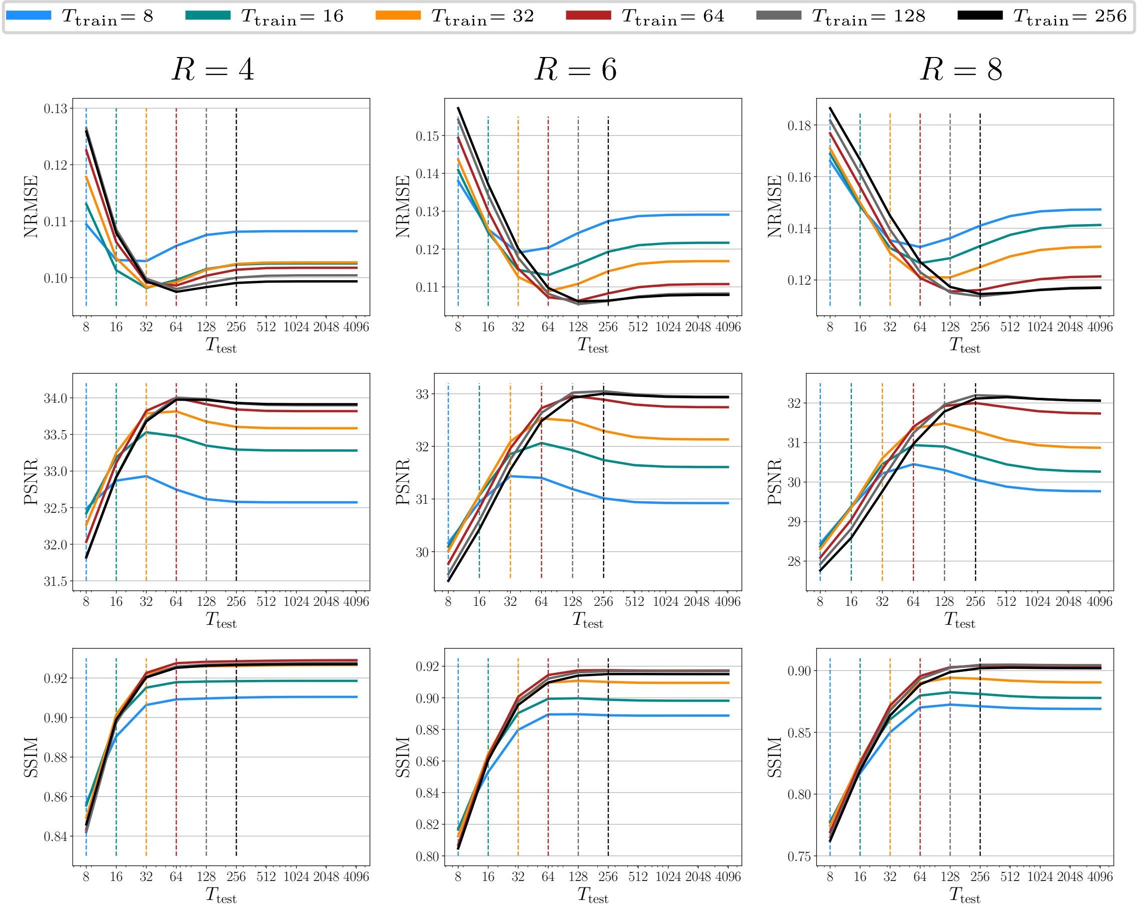

Figure 17: Average NRMSE, PSNR and SSIM for different PDHG-networks $\mathcal{N}_{\Theta_{T_{\text{train}}}}^{T_{\text{test}}}$ trained and tested with different combinations of T_{train} and T_{test} shown for the accelerated dynamic cardiac MRI example for acceleration factors $R = 4, 6, 8$. The color of the dashed lines encodes the number of iterations which were used for training, and consequently, the number of iterations used at test time for which one could expect the corresponding network to yield the best measure. We see that, however, this does not consistently hold, especially for lower T_{train}, where choosing $T_{\text{test}} > T_{\text{train}}$ quite consistently improves the results up to some point with respect to all measures.

5.6.2 Choosing T at Test Time

Recall that at test time, the proposed method generates an input-dependent regularization parameter-map $\mathbf{\Lambda}_{\Theta_T}$ which is inherently dependent on the number of iterations T the network was trained with. With $\mathbf{\Lambda}_{\Theta_T}$, we can then formulate the reconstruction problem (21). Conceptually, it might be desirable to exactly solve problem (21), i.e., to run the network $\mathcal{N}_{\Theta_T}^{T'}$ until convergence by setting T' high enough, i.e. conceptually let $T' \to \infty$. By the triangle inequality, we have

$$\|\mathbf{x}^*(\Theta_T) - \mathbf{x}_{\text{true}}\|_2 \leq \|\mathbf{x}^*(\Theta_T) - \mathbf{x}_{T'}(\Theta_T)\|_2 + \|\mathbf{x}_{T'}(\Theta_T) - \mathbf{x}_{\text{true}}\|_2, \tag{59}$$

where $\|\mathbf{x}^*(\Theta_T)-\mathbf{x}_{T'}(\Theta_T)\|_2 \to 0$ as $T' \to \infty$ due to the convergence of the algorithm that the network $\mathcal{N}_{\Theta_T}^{T'}$ implicitly defines. Note however, that the contribution of the second term is not necessarily the smallest for $T' = T$ (see also Figure 17, especially for lower T_{test}), since Θ_T was obtained by training only Θ (see (29)) and not Θ *and* T' jointly. This means that, in general, there could exist a configuration which further improves the results, i.e. $\|\mathbf{x}_T(\Theta_T) - \mathbf{x}_{\text{true}}\|_2 \geq \|\mathbf{x}_{T'}(\Theta_T) - \mathbf{x}_{\text{true}}\|_2$ for some $T' \neq T$. This is clearly visible from Figure 17, which shows the average NRMSE, PSNR and SSIM for different combinations of T_{train} and T_{test}. Interestingly, it reveals that setting $T_{\text{train}} = T_{\text{test}}$ is indeed not consistently the best choice. Especially for lower T_{train}, setting $T_{\text{test}} > T_{\text{train}}$ introduces further regularization and yields more accurate reconstructions. For higher T_{train}, in contrast, we see that setting $T_{\text{test}} > T_{\text{train}}$ possibly introduces reconstruction errors which, however, can be entirely attributed to the regularization inherently imposed by the TV, i.e., coming from the term $\|\mathbf{x}^*(\Theta_T) - \mathbf{x}_{\text{true}}\|_2$.
This means that, in general, one can view our proposed method in two different flavours. From one point of view, with the proposed method, we can generate a regularization parameter-map $\mathbf{\Lambda}_{\Theta_T}$ which we then use to formulate the reconstruction problem (21) and which we then aim to subsequently solve *exactly*, i.e., the corresponding algorithm defined by $\mathcal{N}_{\Theta_T}^{T'}$ is run until convergence by letting $T' \to \infty$. Thereby, at test time, we implicitly accept the inherent model-error which is made by choosing the TV-minimization as the underlying regularization method. All the results shown in the paper were generated following this strategy. From a second, more applied perspective, we can view the proposed approach which yields regularization parameter-maps which are also tailored to the specific number of iterations the network was trained with. Therefore, assuming one was able to use a high-enough T_{train} for training, at test time, one simply uses $T_{\text{test}} = T_{\text{train}}$ or, if T_{train} had to be strongly compromised during training (for example, due to limited GPU-memory) one can manually adjust an appropriate T_{test} on a validation set to compensate for the effects seen in Figure 17.

6 Conclusion

We have presented a data-driven approach to automatically select data/patient-adaptive spatial/spatio-temporal dependent regularization parameter-maps for the variational regularization approach focusing on TV-minimization. This constitutes a simple yet efficient and elegant way to combine variational methods with the versatility of deep learning-based approaches, yielding an interpretable reconstruction algorithm which inherits all theoretical properties of the scheme the network implicitly defines. We showed consistency results of the proposed unrolled scheme and we applied the proposed method to a dynamic MRI reconstruction problem, a quantitative MRI reconstruction problem, a dynamic image denoising problem and a low-dose CT reconstruction problem. In the following, we discuss possible future research directions and we also comment on the limitations of our approach.

We can immediately identify several different components worth further investigations. First of all, for a fixed problem formulation and choice of regularization method (i.e. the TV-minimization considered in this work) there exist several different reconstruction algorithms, all with their theoretical and practical advantages and limitations, see e.g. [21, 38, 48, 100]. It might be interesting to investigate whether our approach yields similar regularization maps regardless of the chosen reconstruction method and if not, to what extent they differ in. Second, in this work, we have considered the TV-minimization as the regularization method of choice. However, also TV minimization-based methods are known to have limitations, e.g. in producing staircasing effects. We hypothesize that the proposed method could as well be expanded to TGV-based methods [14] to overcome these limitations. In addition, the parameter-map learning can be applied when a combination of regularizers is considered. For example, similar to the dynamic MRI and denoising case studies, the proposed method can be used for Hyperspectral X-ray CT, where the spatial and spectral domains are regularized differently, see e.g., [103, 104]. Further, other regularization methods as for example Wavelet-based methods [23, 34] could be considered as well, where instead of employing the finite differences operator ∇, a Wavelet-operator $\mathbf{\Psi}$ would be the sparsity-transform of choice. Thereby, the multi-scale decomposition of the U-Net which we have used in our work also naturally fits the problem and could be utilized to estimate different parameter-maps for each different level of the Wavelet-decomposition. Third, although we have used a plain U-Net [84] for the estimation of the regularization parameter-maps, there exist nowadays more sophisticated network architectures, e.g. transformers [64, 69], which could be potentially adopted as well. From the theoretical perspective, future work can include an extension of the consistency results to stationary points instead for minimizers only as well as an extension to the non-strongly convex fidelity terms in order to cover the CT case as well. It would be also interesting to theoretically investigate in what degree CNN-produced artefacts in the parameter-maps can affect or create artefacts to the corresponding reconstructions [18, 44, 51]. This would further increase the interpretability of our approach. From a practical point of view it might be interesting to compare the proposed approach to other data-driven methods which use NNs to learn the regularizers [3,6,41,54,87]. Note however, that although our proposed approach employs as well NNs, they are (on purpose) restricted to play the role of the estimation of the regularization parameter-maps. Thus, despite the observed improvements for the demonstrated applications, it is expected that NNs-based methods which learn the regularizers will outperform the proposed approach.

The main limitations of the proposed approach are the ones which are common for every unrolled NN: the large GPU-memory consumption to store intermediate results and their corresponding gradients while training on the one hand and the possibly long training times which are attributed to the need to repeatedly apply the forward and the adjoint operator during training on the other hand. While in this work, we simply made use of `PyTorch`'s automatic differentiation package to compute the gradients necessary for training, one could rely on more sophisticated methods to do so. However, as we have mentioned, simply using the adjoint equation in the spirit of bilevel optimization [16, 27, 36, 43] in order to compute gradients would be computationally demanding so alternative routes should be sought.

Also, note that in this work we have merely used retrospectively simulated data to be able to report quantitative measures among the different choices of regularization parameters. For example, in real-world MR imaging situations the coil sensitivity maps in the operator $\mathbf{C}$ in (51) are not known a-priori but must be first estimated with appropriate methods, e.g. [95], before the proposed method can be applied. In addition, we mention that for the MR applications, the discretization of the forward operator used for the retrospective simulation and the discretization for the forward and the adjoint operators used for the reconstructions are the same. The application of the approach to real measured scanner data as well as its validation with respect to pre-defined clinical tasks might give further insights about the practical relevance of the method.

As we have seen from Figure 15, to be able to learn the regularization parameter-map with a CNN as proposed, one must be able to use a certain number of iterations T for the unrolled NN to ensure that the output image of the reconstruction network has sufficiently converged to the solution of problem (3). How large this number needs to be depends on the considered application as well as the convergence rate of the unrolled algorithm which is used for the reconstruction. To reduce training times, one could employ accelerated versions of the algorithm to be unrolled for the reconstruction, see e.g. [58] or alternatively one could use a stochastic version of PDHG with adaptive step-sizes [19].

Acknowledgments

The authors acknowledge the support of the German Research Foundation (DFG) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689) as this work was initiated during the Hackathon event "Maths Meets Image", Berlin, March 2022, which was part of the MATH+ Thematic Einstein Semester on "Mathematics of Imaging in Real-World Challenges". This work is further supported via the MATH+ project EF3-7.

This work was funded by the UK EPSRC grants the "Computational Collaborative Project in Synergistic Reconstruction for Biomedical Imaging" (CCP SyneRBI) EP/T026693/1; "A Reconstruction Toolkit for Multichannel CT" EP/P02226X/1 and "Collaborative Computational Project in tomographic imaging" (CCPi) EP/M022498/1 and EP/T026677/1. This work made use of computational support by CoSeC, the Computational Science Centre for Research Communities, through CCP SyneRBI and CCPi.

This work is part of the Metrology for Artificial Intelligence for Medicine (M4AIM) project that is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) in the framework of the QI-Digital initiative.

References

[1] J. Adler, H. Kohr, and O. Öktem. Operator discretization library (ODL). Zenodo, 2017. (Cited on page 46).

[2] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. Inverse Problems, 33(12):124007, 2017. https://doi.org/10.1088/1361-6420/aa9581. (Cited on page 5).

[3] J. Adler and O. Öktem. Learned primal-dual reconstruction. IEEE Transactions on Medical Imaging, 37(6):1322–1332, 2018. https://doi.org/10.1109%2Ftmi.2018.2799231. (Cited on page 34).

[4] J. Adler and O. Öktem. Learned primal-dual reconstruction. IEEE Transactions on Medical Imaging, 37(6):1322–1332, 2018. https://doi.org/10.1109%2Ftmi.2018.2799231. (Cited on page 5).

[5] B. M. Afkham, J. Chung, and M. Chung. Learning regularization parameters of inverse problems via deep neural networks. Inverse Problems, 37(10):105017, 2021. https://doi.org/10.1088/1361-6420/ac245d. (Cited on pages 4, 5, and 11).

[6] H. K. Aggarwal, M. P. Mani, and M. Jacob. Modl: Model-based deep learning architecture for inverse problems. IEEE Transactions on Medical Imaging, 38(2):394–405, 2018. https://doi.org/10.1109%2Ftmi.2018.2865356. (Cited on page 34).

[7] F. Altekrüger, A. Denker, P. Hagemann, J. Hertrich, P. Maass, and G. Steidl. PatchNR: Learning from small data by patch normalizing flow regularization. arXiv preprint arXiv:2205.12021, 2022. https://doi.org/10.48550/arXiv.2205.12021. (Cited on page 27).

[8] S. Armato, G. McLennan, M. McNitt-Gray, C. Meyer, A. Reeves, L. Bidaut, B. Zhao, B. Croft, and L. Clarke. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics, 38(2):915–931, 2011. https://doi.org/10.1118%2F1.3469350. (Cited on page 46).

[9] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. Acta Numerica, 28:1–174, 2019. https://doi.org/10.1017/S0962492919000059. (Cited on page 5).

[10] B. Aubert-Broche, M. Griffin, G. B. Pike, A. C. Evans, and D. L. Collins. Twenty new digital brain phantoms for creation of validation image data bases. IEEE Transactions on Medical Imaging, 25(11):1410–1416, 2006. https://doi.org/10.1109%2Ftmi.2006.883453. (Cited on page 44).

[11] M. Benning, C. B. Schönlieb, T. Valkonen, and V. Vlacic. Explorations on anisotropic regularisation of dynamic inverse problems by bilevel optimisation. arXiv preprint arXiv:1602.01278, 2016. https://doi.org/10.48550/arXiv.1602.01278. (Cited on page 7).

[12] M. Bergounioux, E. Papoutsellis, S. Stute, and C. Tauber. Infimal convolution spatiotemporal PET reconstruction using total variation based priors. HAL preprint, 2018. https://hal.archives-ouvertes.fr/hal-01694064. (Cited on page 8).

[13] K. T. Block, M. Uecker, and J. Frahm. Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint. Magnetic Resonance in Medicine, 57(6):1086–1098, 2007. https://doi.org/10.1002/mrm.21236. (Cited on page 2).

[14] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. SIAM Journal on Imaging Sciences, 3(3):492–526, 2010. http://dx.doi.org/10.1137/090769521. (Cited on pages 3 and 33).

[15] M. Burger and S. Osher. A Guide to the TV Zoo, pages 1–70. Springer International Publishing, 2013. https://doi.org/10.1007/978-3-319-01712-9_1. (Cited on page 2).

[16] L. Calatroni, C. Chung, J. C. De Los Reyes, C. B. Schönlieb, and T. Valkonen. Bilevel approaches for learning of variational imaging models. In RADON book Series on Computational and Applied Mathematics, vol. 18. Berlin, Boston: De Gruyter, 2017. https://www.degruyter.com/view/product/458544. (Cited on pages 3 and 34).

[17] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari. Tikhonov regularization and the l-curve for large discrete ill-posed problems. Journal of Computational and Applied Mathematics, 123(1):423–446, 2000. https://doi.org/10.1016/S0377-0427(00)00414-3. (Cited on page 3).

[18] V. Caselles, A. Chambolle, and M. Novaga. The discontinuity set of solutions of the TV denoising problem and some extensions. Multiscale Modeling & Simulation, 6(3):879–894, 2007. http://dx.doi.org/10.1137/070683003. (Cited on pages 5 and 34).

[19] A. Chambolle, C. Delplancke, M. J. Ehrhardt, C.-B. Schönlieb, and J. Tang. Stochastic primal dual hybrid gradient algorithm with adaptive step-sizes. arXiv preprint arXiv:2301.02511, 2023. https://doi.org/10.48550/arXiv.2301.04764. (Cited on page 35).

[20] A. Chambolle and P. L. Lions. Image recovery via total variation minimization and related problems. Numerische Mathematik, 76:167–188, 1997. http://dx.doi.org/10.1007/s002110050258. (Cited on page 2).

[21] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011. https://doi.org/10.1007%2Fs10851-010-0251-1. (Cited on pages 2, 5, 8, 9, and 33).

[22] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. Acta Numerica, 25:161–319, 2016. https://doi.org/10.1017/S096249291600009X. (Cited on page 2).

[23] S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. IEEE Transactions on Image Processing, 9(9):1532–1546, 2000. https://doi.org/10.1109%2F83.862633. (Cited on pages 2 and 34).

[24] L. Chung, J. C. De los Reyers, and C. B. Schönlieb. Learning optimal spatially-dependent regularization parameters in total variation image denoising. Inverse Problems, 33:074005, 2017. https://doi.org/10.1088/1361-6420/33/7/074005. (Cited on page 3).

[25] M. J. Colbrook, V. Antun, and A. C. Hansen. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. Proceedings of the National Academy of Sciences, 119(12):e2107151119, 2022. https://doi.org/10.1073/pnas.2107151119. (Cited on page 5).

[26] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In Human Vision and Electronic Imaging XII, volume 6492, pages 196–206. SPIE, 2007. https://doi.org/10.1117%2F12.702790. (Cited on page 16).

[27] C. Crockett and J. A. Fessler. Bilevel methods for image reconstruction. Foundations and Trends in Signal Processing, 15(2-3):121–289, 2022. http://dx.doi.org/10.1561/2000000111. (Cited on page 34).

[28] G. Dal Maso. Introduction to Γ-convergence. Birkhäuser, 1993. https://doi.org/10.1007%2F978-1-4612-0327-8. (Cited on page 15).

[29] J. C. De los Reyes and C. B. Schönlieb. Image denoising: learning the noise model via nonsmooth PDE-constrained optimization. Inverse Problems and Imaging, 7(4):1183–1214, 2013. http://dx.doi.org/10.3934/ipi.2013.7.1183. (Cited on page 3).

[30] J. C. De Los Reyes, C. B. Schönlieb, and T. Valkonen. Bilevel parameter learning for higher-order total variation regularisation models. Journal of Mathematical Imaging and Vision, 57(1):1–25, 2017. https://doi.org/10.1007/s10851-016-0662-8. (Cited on page 3).

[31] J. C. De los Reyes and D. Villacís. Bilevel Optimization Methods in Imaging, pages 1–34. Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-03009-4_66-1. (Cited on page 3).

[32] J. C. De los Reyes and D. Villacís. Optimality conditions for bilevel imaging learning problems with total variation regularization. SIAM Journal on Imaging Sciences, 15(4):1646–1689, 2022. https://doi.org/10.1137/21M143412X. (Cited on page 13).

[33] M. D'Elia, J. C. De Los Reyes, and A. Miniguano-Trujillo. Bilevel parameter learning for nonlocal image denoising models. Journal of Mathematical Imaging and Vision, 63:753–775, 2021. https://doi.org/10.1007/s10851-021-01026-2. (Cited on page 3).

[34] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3):425–455, 1994. https://doi.org/10.1093%2Fbiomet%2F81.3.425. (Cited on pages 2 and 34).

[35] M. Duff, N. D. F. Campbell, and M. J. Ehrhardt. Regularising inverse problems with generative machine learning models. arXiv preprint arXiv:2107.11191, 2021. https://doi.org/10.48550/ARXIV.2107.11191. (Cited on page 5).

[36] M. J. Ehrhardt and L. Roberts. Analyzing inexact hypergradients for bilevel learning. arXiv preprint arXiv:2301.04764, 2023. https://doi.org/10.48550/arXiv.1402.0026. (Cited on page 34).

[37] H. G. Feichtinger and T. Strohmer. Efficient numerical methods in non-uniform sampling theory. Numerische Mathematik, 69(4):423–440, 1995. https://doi.org/10.1007%2Fs002110050101. (Cited on page 18).

[38] T. Goldstein and S. Osher. The split Bregman method for l1-regularized problems. SIAM journal on imaging sciences, 2(2):323–343, 2009. https://doi.org/10.1137%2F080725891. (Cited on pages 20 and 33).

[39] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21(2):215–223, 1979. https://doi.org/10.1080/00401706.1979.10489751. (Cited on page 3).

[40] A. Haase. Snapshot flash MRI. applications to T1, T2, and chemical-shift imaging. Magnetic Resonance in Medicine, 13(1):77–89, 1990. https://doi.org/10.1002%2Fmrm.1910130109. (Cited on page 21).

[41] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated MRI data. Magnetic Resonance in Medicine, 79(6):3055–3071, 2018. https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.26977. (Cited on pages 5 and 34).

[42] C. He, C. Hu, W. Zhang, and B. Shi. A fast adaptive parameter estimation for total variation image restoration. IEEE Transactions on Image Processing, 23(12):4954–4967, 2014. https://doi.org/10.1109/TIP.2014.2360133. (Cited on page 3).

[43] M. Hintermüller and K. Papafitsoros. Generating structured nonsmooth priors and associated primal-dual methods. In Ron Kimmel and Xue-Cheng Tai, editors, Processing, Analyzing and Learning of Images, Shapes, and Forms: Part 2, volume 20 of Handbook of Numerical Analysis, pages 437–502. 2019. https://doi.org/10.1016/bs.hna.2019.08.001. (Cited on pages 3 and 34).

[44] M. Hintermüller, K. Papafitsoros, and C. N. Rautenberg. Analytical aspects of spatially adapted total variation regularisation. Journal of Mathematical Analysis and Applications, 454(2):891–935, 2017. https://doi.org/10.1016/j.jmaa.2017.05.025. (Cited on pages 5 and 34).

[45] M. Hintermüller, K. Papafitsoros, C. N. Rautenberg, and H. Sun. Dualization and automatic distributed parameter selection of total generalized variation via bilevel optimization. Numerical Functional Analysis and Optimization, 43(8):887–932, 2022. https://doi.org/10.1080/01630563.2022.2069812. (Cited on page 3).

[46] M. Hintermüller and C. N. Rautenberg. Optimal selection of the regularization function in a weighted total variation model. Part I: Modelling and theory. Journal of Mathematical Imaging and Vision, 59(3):498–514, 2017. https://doi.org/10.1007/s10851-017-0744-2. (Cited on page 3).

[47] M. Hintermüller, C. N. Rautenberg, T. Wu, and A. Langer. Optimal selection of the regularization function in a weighted total variation model. Part II: Algorithm, its analysis and numerical tests. Journal of Mathematical Imaging and Vision, 59(3):515–533, 2017. https://doi.org/10.1007/s10851-017-0736-2. (Cited on page 3).

[48] M. Hintermüller and G. Stadler. An infeasible primal-dual algorithm for total bounded variation–based inf-convolution-type image restoration. SIAM Journal on Scientific Computing, 28(1):1–23, 2006. http://dx.doi.org/10.1137/040613263. (Cited on pages 2 and 33).

[49] M. Holler and K. Kunisch. On infimal convolution of TV-type functionals and applications to video and image reconstruction. SIAM Journal on Imaging Sciences, 7(4):2258–2300, 2014. https://doi.org/10.1137/130948793. (Cited on page 7).

[50] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, and J. K. Seo. Deep learning for undersampled MRI reconstruction. Physics in Medicine & Biology, 63(13):135007, 2018. https://doi.org/https://doi.org/10.1088/1361-6560/aac71a. (Cited on page 5).

[51] K. Jalalzai. Discontinuities of the minimizers of the weighted or anisotropic total variation for image reconstruction. arXiv preprint 1402.0026, 2014. https://doi.org/10.48550/arXiv.1402.0026. (Cited on pages 5 and 34).

[52] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. https://doi.org/10.48550/arXiv.1412.6980. (Cited on pages 44, 45, and 46).

[53] A. Kofler, M. Haltmeier, T. Schaeffter, M. Kachelrieß, M. Dewey, C. Wald, and C. Kolbitsch. Neural networks-based regularization for large-scale medical image reconstruction. Physics in Medicine & Biology, 65(13):135003, 2020. https://doi.org/10.1088/1361-6560/ab990e. (Cited on page 5).

[54] A. Kofler, M. Haltmeier, T. Schaeffter, and C. Kolbitsch. An end-to-end-trainable iterative network architecture for accelerated radial multi-coil 2D cine MR image reconstruction. Medical Physics, 48(5):2412–2425, 2021. https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14809. (Cited on pages 5 and 34).

[55] C. Kolbitsch, C. Prieto, and T. Schaeffter. Cardiac functional assessment without electrocardiogram using physiological self-navigation. Magnetic Resonance in Medicine, 71(3):942–954, 2013. https://doi.org/10.1002%2Fmrm.24735. (Cited on page 44).

[56] K. Kunisch and M. Hintermüller. Total bounded variation regularization as a bilaterally constrained optimization problem. SIAM Journal on Applied Mathematics, 64(4):1311–1333, 2004. https://doi.org/10.1137/S0036139903422784. (Cited on page 2).

[57] K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. SIAM Journal on Imaging Sciences, 6(2):938–983, 2013. http://dx.doi.org/10.1137/120882706. (Cited on page 3).

[58] H. T. V. Le, N. Pustelnik, and M. Foare. The faster proximal algorithm, the better unfolded deep learning architecture? The study case of image denoising. In 2022 30th European Signal Processing Conference (EUSIPCO), pages 947–951, 2022. https://doi.org/10.23919/EUSIPCO55093.2022.9909592. (Cited on page 35).

[59] Y. Le Montagner, E. D. Angelini, and J. C. Olivo-Marin. An unbiased risk estimator for image denoising in the presence of mixed Poisson–Gaussian noise. IEEE Transactions on Image Processing, 23(3):1255–1268, 2014. https://doi.org/10.1109%2Ftip.2014.2300821. (Cited on page 3).

[60] J. Leuschner, M. Schmidt, D. O. Baguer, and P. Maass. LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction. Scientific Data, 8(109), 2021. https://doi.org/10.1038%2Fs41597-021-00893-z. (Cited on page 46).

[61] J. Leuschner, M. Schmidt, P. S. Ganguly, V. Andriiashen, S. B. Coban, A. Denker, D. Bauer, A. Hadjifaradji, K. J. Batenburg, P. Maass, and M. van Eijnatten. Quantitative comparison of deep learning-based image reconstruction methods for low-dose and sparse-angle CT applications. Journal of Imaging, 7(3), 2021. https://www.mdpi.com/2313-433X/7/3/44. (Cited on page 27).

[62] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. NETT: solving inverse problems with deep neural networks. Inverse Problems, 36(6):065005, 2020. https://doi.org/10.1088/1361-6420/ab6d57. (Cited on page 5).

[63] Z. Li, Y. K. Dewaraja, and J. A. Fessler. Training end-to-end unrolled iterative neural networks for SPECT image reconstruction. IEEE Transactions on Radiation and Plasma Medical Sciences, 2023. https://doi.org/10.1109/TRPMS.2023.3240934. (Cited on page 5).

[64] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. SwinIR: Image restoration using swin transformer. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, 2021. https://doi.org/10.1109%2Ficcvw54120.2021.00210. (Cited on page 34).

[65] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In International Conference on Machine Learning, pages 6305–6315. PMLR, 2020. (Cited on page 11).

[66] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A general descent aggregation framework for gradient-based bi-level optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. https://doi.org/10.1109%2Ftpami.2022.3140249. (Cited on page 11).

[67] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2018. https://doi.org/10.48550/arXiv.1711.05101. (Cited on page 45).

[68] H. Lu and J. Yang. On the infimal sub-differential size of primal-dual hybrid gradient method. arXiv preprint arXiv:2206.12061, 2022. https://doi.org/10.48550/arXiv.2206.12061. (Cited on page 13).

[69] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng. Transformer for single image super-resolution. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2022. https://doi.org/10.1109%2Fcvprw56347.2022.00061. (Cited on page 34).

[70] S. Lunz, O. Öktem, and C. B. Schönlieb. Adversarial regularizers in inverse problems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. https://proceedings.neurips.cc/paper/2018/file/d903e9608cfbf08910611e4346a0ba44-Paper.pdf. (Cited on page 5).

[71] M. Mani, M. Jacob, V. Magnotta, and J. Zhong. Fast iterative algorithm for the reconstruction of multishot non-cartesian diffusion data. Magnetic Resonance in Medicine, 74(4):1086–1094, 2015. https://doi.org/10.1002%2Fmrm.25486. (Cited on page 18).

[72] M. T. McCann, K. Hwan Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. IEEE Signal Processing Magazine, 34(6):85–95, 2017. https://doi.org/10.1109/MSP.2017.2739299. (Cited on page 5).

[73] T. Meinhardt, M. Moller, C. Hazirbas, and D. Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 1799–1808, Los Alamitos, CA, USA, 2017. IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.198. (Cited on page 5).

[74] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016. https://doi.org/10.48550/arXiv.1603.00831. (Cited on page 45).

[75] 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. https://doi.10.1109/MSP.2020.3016905. (Cited on pages 4 and 5).

[76] M. J. Muckley, R. Stern, T. Murrell, and F. Knoll. TorchKbNufft: A high-level, hardware-agnostic non-uniform fast Fourier transform. In ISMRM Workshop on Data Sampling & Image Reconstruction, 2020. Source code available at https://github.com/mmuckley/torchkbnufft. (Cited on page 17).

[77] R. R. D. Nekhili, X. Descombes, and L. Calatroni. A hybrid approach combining cnns and variational modelling for blind image denoising. HAL preprint, 2022. https://hal.archives-ouvertes.fr/hal-03596605. (Cited on page 4).

[78] P. Ochs, R. Ranftl, T. Brox, and T. Pock. Bilevel optimization with nonsmooth lower level problems. In International Conference on Scale Space and Variational Methods in Computer Vision, pages 654–665. Springer, 2015. https://doi.org/10.1007%2F978-3-319-18461-6_52. (Cited on page 11).

[79] V. Pagliari, K. Papafitsoros, B. Raita, and A. Vikelis. Bilevel training schemes in imaging for total-variation-type functionals with convex integrands. SIAM Journal on Imaging Sciences, 15(4):1690–1728, 2022. https://doi.org/10.1137/21M1467328. (Cited on page 3).

[80] E. Papoutsellis, E. Ametova, C. Delplancke, G. Fardell, J. S. Jørgensen, E. Pasca, M. Turner, R. Warr, W. R. B. Lionheart, and P. J. Withers. Core imaging library - Part II: multichannel reconstruction for dynamic and spectral tomography. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379(2204):20200193, 2021. https://doi.org/10.1098/rsta.2020.0193. (Cited on page 8).

[81] K. P. Pruessmann, M. Weiger, P. Börnert, and P. Boesiger. Advances in sensitivity encoding with arbitrary k-space trajectories. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 46(4):638–651, 2001. https://doi.org/10.1016%2Fj.mri.2007.01.003. (Cited on pages 23 and 45).

[82] J. Radon. On the determination of functions from their integral values along certain manifolds. IEEE Transactions on Medical Imaging, 5(4):170–176, 1986. https://doi.org/10.1109%2Ftmi.1986.4307775. (Cited on page 25).

[83] Y. Romano, M. Elad, and P. Milanfar. The little engine that could: Regularization by denoising (red). SIAM Journal on Imaging Sciences, 10(4):1804–1844, 2017. https://doi.org/10.1137/16M1102884. (Cited on page 5).

[84] 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. https://doi.org/10.1007%2F978-3-662-54345-0_3. (Cited on pages 10 and 34).

[85] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1-4):259–268, 1992. http://dx.doi.org/10.1016/0167-2789(92)90242-F. (Cited on page 2).

[86] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. Variational methods in imaging. Springer, New York, 2009. https://doi.org/10.1007/978-0-387-69277-7. (Cited on page 2).

[87] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE Transactions on Medical Imaging, 37(2):491–503, 2018. https://doi.org/10.1109%2Ftmi.2017.2760978. (Cited on pages 5 and 34).

[88] M. Schloegl, M. Holler, A. Schwarzl, K. Bredies, and R. Stollberger. Infimal convolution of total generalized variation functionals for dynamic MRI. Magnetic Resonance in Medicine, 78(1):142–155, 2017. https://doi.org/10.1002/mrm.26352. (Cited on page 7).

[89] F. Sherry, M. Benning, J. C. De los Reyes, M. J. Graves, G. Maierhofer, G. Williams, C. B. Schönlieb, and M. J. Ehrhardt. Learning the sampling pattern for MRI. IEEE Transactions on Medical Imaging, 39(12):4310–4321, 2020. https://doi.org/10.1109/TMI.2020.3017353. (Cited on page 3).

[90] E. Y. Sidky, J. H. Jørgensen, and X. Pan. Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle–Pock algorithm. Physics in Medicine & Biology, 57(10):3065, 2012. https://doi.org/10.1088%2F0031-9155%2F57%2F10%2F3065. (Cited on page 18).

[91] E. Y. Sidky and X. Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Physics in Medicine & Biology, 53(17):4777, 2008. https://doi.org/10.1088/0031-9155/53/17/021. (Cited on page 2).

[92] J. I. Tamir, F. Ong, S. Anand, E. Karasan, K. Wang, and M. Lustig. Computational MRI with physics-based constraints: Application to multicontrast and quantitative imaging. IEEE signal processing magazine, 37(1):94–104, 2020. https://doi.org/10.1109%2Fmsp.2019.2940062. (Cited on page 21).

[93] J. I. Tamir, M. Uecker, W. Chen, P. Lai, M. T. Alley, S. S. Vasanawala, and M. Lustig. T2 shuffling: sharp, multicontrast, volumetric fast spin-echo imaging. Magnetic Resonance in Medicine, 77(1):180–195, 2017. https://doi.org/10.1002%2Fmrm.26102. (Cited on page 18).

[94] A. Tikhonov. Solution of incorrectly formulated problems and the regularization method. In Soviet Mathematics Doklady, volume 5, pages 1035–1038, 1963. (Cited on page 2).

[95] M. Uecker, P. Lai, M. J. Murphy, P. Virtue, M. Elad, J. M. Pauly, S. S. Vasanawala, and M. Lustig. ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. Magnetic resonance in medicine, 71(3):990–1001, 2014. (Cited on page 34).

[96] S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with low complexity priors. Information and Inference: A Journal of the IMA, 4(3):230–287, 2015. https://doi.org/10.1093/imaiai/iav005. (Cited on page 12).

[97] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili. Robust sparse analysis regularization. IEEE Transactions on information theory, 59(4):2001–2016, 2012. https://doi.org/10.1109/TIT.2012.2233859. (Cited on page 12).

[98] F. von Knobelsdorff-Brenkenhoff, G. Pilz, and J. Schulz-Menger. Representation of cardiovascular magnetic resonance in the AHA/ACC guidelines. Journal of Cardiovascular Magnetic Resonance, 19(1):1–21, 2017. https://doi.org/10.1186%2Fs12968-017-0385-z. (Cited on page 16).

[99] F. T. A. W. Wajer and K. P. Pruessmann. Major speedup of reconstruction for sensitivity encoding with arbitrary trajectories. In Proc. Intl. Soc. Mag. Res. Med, page 767, 2001. http://cds.ismrm.org/ismrm-2001/PDF3/0767.pdf. (Cited on page 18).

[100] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. SIAM Journal on Imaging Sciences, 1(3):248–272, 2008. https://doi.org/10.1137%2F080724265. (Cited on pages 2, 20, and 33).

[101] Y. Wang and L. Ying. Compressed sensing dynamic cardiac cine MRI using learned spatiotemporal dictionary. IEEE Transactions on Biomedical Engineering, 61(4):1109–1120, 2013. https://doi.org/10.1109%2Ftbme.2013.2294939. (Cited on page 8).

[102] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on image processing, 13(4):600–612, 2004. https://doi.org/10.1109%2Ftip.2003.819861. (Cited on page 16).

[103] R. Warr, E. Ametova, R. J. Cernik, G. Fardell, S. Handschuh, J. S. Jørgensen, E. Papoutsellis, E. Pasca, and P. J. Withers. Enhanced hyperspectral tomography for bioimaging by spatiospectral reconstruction. Scientific Reports, 11(1), 2021. https://doi.org/10.1038/s41598-021-00146-4. (Cited on page 34).

[104] R. Warr, S. Handschuh, M. Glösmann, R. J. Cernik, and P. J. Withers. Quantifying multiple stain distributions in bioimaging by hyperspectral X-ray tomography. Scientific Reports, 12(1), 2022. https://doi.org/10.1038/s41598-022-23592-0. (Cited on page 34).

[105] C. Wu and X. Tai. Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. SIAM Journal on Imaging Sciences, 3(3):300–339, 2010. https://doi.org/10.1137/090767558. (Cited on page 2).

[106] M. Yan. A new primal–dual algorithm for minimizing the sum of three functions with a linear operator. Journal of Scientific Computing, 76:1698–1717, 2018. https://doi.org/10.1007%2Fs10915-018-0680-3. (Cited on pages 9, 27, and 28).

[107] H. Zhang, M. Yan, and W. Yin. One condition for solution uniqueness and robustness of both l1-synthesis and l1-analysis minimizations. Advances in Computational Mathematics, 42(6):1381–1399, 2016. https://doi.org/10.1007/s10444-016-9467-y. (Cited on page 12).

A Experimental Set-Ups

A.1 Experimental Set-Up: Dynamic Cardiac MR Image Reconstruction

We used a set of 216 complex-valued cardiac cine MR images of the study [55] which we split in portion of 144/36/36 for training, validation and testing. The images have shape $n_x \times n_y \times n_t = 160 \times 160 \times 30$ and a resolution of 2×2 mm^2 with a slice thickness of 8 mm^2. The number of receiver coils is $n_c = 12$. We retrospectively simulated complex-valued k-space data according to (1) using the model in (51) as the forward operator simulating acceleration factors of $R = 4, 6, 8$ with complex-valued Gaussian noise with standard deviation $\sigma = 0.15, 0.30, 0.45$.
As described in Section 3.1, we constructed u_Θ such that it yields two different parameter-maps. One for the spatial x- and y-directions and one for the temporal direction, i.e., $\mathbf{\Lambda}_\Theta := (\mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{xy}, \mathbf{\Lambda}_\Theta^{t})$. The CNN u_Θ here corresponds to a 3D U-Net with two input-channels (for the real and the imaginary part of the image, respectively), three encoding stages, two convolutional layers per stage and an initial number of eight filters which are applied to the input image. As in Figure 2, the last layer consists of a $1 \times 1 \times 1$ convolution with two output channels (the first for the parameter-map for the x- and y-directions, the second for the parameter-map for the t-direction) and the softplus activation function ϕ. Note that the gradients of the real and the imaginary parts of the images share the same regularization parameter-map. The scaling factor t in (27) was set to $t = 0.1$. The overall number of trainable parameters of u_Θ is 97 290. To reduce training times, the network was trained on patches of shape $n_x' \times n_y' \times n_t' = 160 \times 160 \times 16$. The network's number of overall iterations was set to $T = 256$ during training, while at test time, we used $T = 4096$ iterations. The reason for the different number of iterations at training and test time is discussed later in Subsection 5.6. The parameters σ, τ and θ were trained as well and constrained to be in the intervals $(0, 1/L), (0, 1/L)$ and $(0, 1)$, respectively, by using a sigmoid activation-function. Despite of the training, we mention that no noteworthy changes were visible after training, i.e. $\sigma \approx \tau \approx 1/L$. Not training σ, τ and θ also led to similar results as the ones shown later. As training routine, we used the Adam optimizer [52] with initial learning rate of 10^{-4} to minimize the mean squared error (MSE) between the reconstructed image and the target image. We trained all networks for 200 epochs while evaluating the network 25 times over the entire training and validation datasets. We then used the model configuration for which the MSE on the validation set was the lowest.

A.2 Experimental Set-Up: Quantitative MRI Reconstruction

We used the BrainWeb [10] dataset of 20 segmented healthy human heads as a basis to generate a quantitative MRI dataset with known ground truth. The subjects were split 17/1/2 for training, validation, and testing. We considered axial slices, rescaled to 192×192 pixels. In each axial slice, we sampled the magnetization and the longitudinal relaxation rate $R_1 = 1/T_1$ for each tissue class from uniform distributions around anatomically plausible values. The phase of the complex magnetization M_0 was modulated to approximate residual phases in the acquisition model. We used low amplitude time-independent random polynomials in both spatial dimensions (third order, maximum amplitude 0.2), as well as random polynomials with coefficients varying at inversion time points within one simulated measurement (second order, maximum amplitude 0.1). Following the signal model (56), we generated images for the inversion times 0.05 s, 0.1 s, 0.2 s, 0.35 s, 0.5 s, 1.0 s, 1.5 s, 2 s, 3 s, 4 s and transformed them into (undersampled) Cartesian k-space. The number of simulated receiver coils was 8. Here, we further modulated the phase of the simulated sensitivity map for each coil with random 2D polynomials (second order, maximum amplitude 2π). The acceleration factor R was chosen from 4, 6, and 8 for comparisons. In each case, complex Gaussian noise with σ randomly chosen from [0.04, 0.4] was added in k-space. The proposed unrolled network was used to reconstruct the complex-valued (qualitative) images at different inversion time points.

Similar to Section 5.2, we choose a simple 3D U-Net with two downsampling steps, two 3D convolution layers for each encoder and decoder block with LeakyReLu activation, and 8 initial filters, resulting in only 97402 parameters. We used a scaled softplus activation, $\beta\phi(x/\beta)$ with $\beta = 5$, for the final activation to keep the predicted regularization strength positive. We initialized the bias of the final convolution layer with -1 (empirically chosen) to stabilize training by starting at a low regularization. We trained the network with AdamW [67] (weight decay 10^{-4}), cosine annealing learning rate schedule with linear warmup over one epoch with a maximum learning rate of 10^{-2}, and a batch size of 4. The number of iterations of the unrolled PDHG is set to $T = 32$ during warmup and $T = 128$ for the rest of the training. Again, σ, τ and θ were trainable parameters. Optimization was done by minimizing the MSE between the ground truth images and the obtained images after masking out non-brain regions. To find λ_{P} and $\lambda_{\tilde{\mathrm{P}}}$, grid searches, similar to those in the dynamic MRI case, were performed with a fixed number of $T = 256$ iterations. For evaluation, we used $T = 256$ iterations of PDHG. We calculated the PSNR and SSIM of the reconstructed images. As a comparison method, we also performed standard iterative MRI reconstruction (CG with early stopping) without any TV regularization [81]. We determined the optimal number of iterations based on the MSE to the ground truth images. Finally, we performed a pixelwise regression on the reconstructed images $\mathbf{x}$ using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, minimizing $\| |q_{t_i}(\mathbf{u})| - |\mathbf{x}| \|_2^2$, to obtain T_1-maps and calculated the RMSE.

A.3 Experimental Set-Up: Dynamic Image Denoising

For training and testing, we used video samples from the benchmark dataset for multiple object tracking [74], containing both dynamic and static camera scenes. For training and validation, we scaled the resolution of the video samples by 0.5 in each direction and extracted patches of size $n_x \times n_y \times n_t = 192 \times 192 \times 32$. During the training process, we used 1751 patches for training and 1000 patches from different video samples for validation. We tested the trained model on scaled resolution but the full spatial dimension with 100 time points per test sample. Gaussian noise with a random standard deviation in the range of $\sigma = 0.1, 0.2, 0.3$ was added to the samples during training. For simplicity and increased training speed, we use a grey-scaled version of the video samples. Because the grey-scaled image data is real-valued, the CNN u_Θ was constructed as described in Section A.1, but with only one output-channel per output-dimension. For this example, we use the same CNN-block u_Θ as in Figure 2. For comparison, we also trained λ^{xyt} which holds a single value for both, the spatial and temporal dimension, and $\lambda^{xy,t}$ which holds two different values for the spatial and temporal dimension. During training we used $T = 128$ network iterations, for testing we increased the number of iterations to $T = 1024$. We minimized the mean squared error (MSE) between denoised and ground truth patches using the Adam optimizer [52] with an initial learning rate of 10^{-4}. All the training was performed for 100 epochs, where validation was performed every second epoch.

A.4 Proximal operator in the log-Poisson case

We shortly describe, why we do not get a closed form for the data-discrepancy term in the case of log-Poisson noise. This can be seen by the proximal operator of f^* in line 3 of Algorithm 1. In our case the convex functional f would be given by $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{z})$, i.e.,

$$f(\mathbf{x}) = \sum_{i=1}^{m} e^{-\mathbf{x}_i \mu} N_0 - \tilde{\mathbf{z}} N_0 \big(- \mathbf{x}_i \mu + \log(N_0) \big), \tag{60}$$

where we set $\tilde{\mathbf{z}} = e^{-\mathbf{z}_i \mu}$ for simplicity. Then the convex conjugate is given by

$$f^*(\mathbf{p}) = \max_{\mathbf{x}} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{p}_i - f(\mathbf{x}) = \max_{\mathbf{x}} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{p}_i - e^{-\mathbf{x}_i \mu} N_0 - \tilde{\mathbf{z}} N_0 \mathbf{x}_i \mu, \tag{61}$$

where we used in the second equality that $\tilde{\mathbf{z}} \log(N_0)$ is independent of $\mathbf{x}$. Differentiation with respect to $\mathbf{x}$ shows that for a maximizer $\hat{\mathbf{x}}$ it holds

$$\hat{\mathbf{x}}_i = -\frac{1}{\mu} \log\big(\mathbf{z} - \frac{\mathbf{z}_i}{\mu N_0}\big).$$

Inserting this in (61) yields the convex conjugate of f

$$\begin{aligned} f^*(\mathbf{p}) &= \sum_{i=1}^{m} -\frac{1}{\mu} \log\big(\mathbf{z} - \frac{\mathbf{p}_i}{\mu N_0}\big)\mathbf{p}_i - \big(\mathbf{z} - \frac{\mathbf{p}_i}{\mu N_0}\big) + \mathbf{z} \log\big(\mathbf{z} - \frac{\mathbf{p}_i}{\mu N_0}\big) \\ &= \sum_{i=1}^{m} \big(\mathbf{z} - \frac{\mathbf{p}_i}{\mu}\big) \log\big(\mathbf{z} - \frac{\mathbf{p}_i}{\mu N_0}\big) - \big(\mathbf{z} - \frac{\mathbf{p}_i}{\mu N_0}\big). \end{aligned}$$

Then for this f^* the proximal operator does not have a simple closed form.

A.5 Experimental Set-Up: Low-Dose Computerized Tomography

We use the LoDoPaB dataset [60][1] for low-dose CT imaging. It is based on scans of the Lung Image Database Consortium and Image Database Resource Initiative [8] which serve as ground truth images, while the measurements are simulated. The dataset contains 35820 training images, 3522 validation images and 3553 test images. Here the ground truth images have a resolution of 362 × 362 on a domain of 26cm × 26cm. We only use the first 300 training images and the first 10 validation images. For the forward operator we consider a normalization constant $\mu = 81.35858$, the mean photon count per detector bin $N_0 = 4096$ as well as 513 equidistant detector bins and 1000 equidistant angles between 0 and π. A detailed description of the data generation process is given in [60]. Following the naming convention of Figure 2, the network u_θ is a 2D U-Net[2], where the number of channels at different scales are 32, 32, 64, 64 and 128 resulting in 610673 trainable parameters. For training we use Adam optimizer [52] with a learning rate of 10^{-4}, a batch size of 1 and train for 50 epochs. Then we used the model configuration for which the MSE on the validation set was lowest. The number of iterations of PD3O is set to $T = 512$ resulting in a training time of around 24 hours on a single NVIDIA GeForce RTX 2080 super GPU with 8 GB GPU memory. At test time, we use $T = 1024$ iterations for reconstruction. The forward and the adjoint operator as well as the FBP were implemented using the publicly available library `ODL` [1].

[1] available at `https://zenodo.org/record/3384092#.Ylglz3VBwgM`

[2] available at `https://jleuschn.github.io/docs.dival/_modules/dival/reconstructors/networks/unet.html`

B Additional Tables

		PDHG - $\lambda_{\tilde{\mathrm{P}}}^{xyt}$			**PDHG -** $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$			**PDHG -** $\mathbf{\Lambda}_{\Theta}^{xy,t}$		
$R = 4$	**SSIM**	0.836	±	0.048	0.824	±	0.059	**0.927**	±	0.016
	PSNR	32.35	±	2.21	33.19	±	2.09	**33.91**	±	2.23
	NRMSE	0.113	±	0.007	0.106	±	0.010	**0.099**	±	0.008
	Blur	0.369	±	0.018	**0.353**	±	0.017	0.359	±	0.019
$R = 6$	**SSIM**	0.833	±	0.038	0.834	±	0.048	**0.915**	±	0.018
	PSNR	30.98	±	2.28	32.28	±	2.10	**32.94**	±	2.21
	NRMSE	0.127	±	0.009	0.112	±	0.008	**0.108**	±	0.008
	Blur	0.391	±	0.021	0.369	±	0.018	**0.365**	±	0.019
$R = 8$	**SSIM**	0.822	±	0.035	0.832	±	0.042	**0.902**	±	0.021
	PSNR	29.95	±	2.32	31.37	±	2.14	**32.10**	±	2.18
	NRMSE	0.141	±	0.011	0.122	±	0.009	**0.117**	±	0.008
	Blur	0.408	±	0.025	0.382	±	0.020	**0.371**	±	0.020

Table 1: Dynamic Cardiac MR Image Reconstruction: Mean and standard deviation of the measures SSIM, PSNR and NRMSE and Blur obtained over the test set. The TV-reconstruction using the proposed spatio-temporal parameter-maps $\mathbf{\Lambda}_{\Theta}^{xy,t}$ improves the results especially in terms of SSIM and PSNR.

		CG-SENSE			**PDHG -** $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$			**PDHG -** $\lambda_{\mathrm{P}}^{xy,t}$			**PDHG -** $\mathbf{\Lambda}_{\Theta}^{xy,t}$		
$R = 4$	**PSNR**	24.62	±	3.45	31.89	±	1.70	32.85	±	2.48	**34.23**	±	2.50
	SSIM	0.654	±	0.095	0.884	±	0.044	0.902	±	0.042	**0.930**	±	0.031
	RMSE [ms]	107	±	34	76	±	15	68	±	21	**58**	±	18
$R = 6$	**PSNR**	24.25	±	2.49	30.19	±	1.58	30.62	±	1.80	**32.62**	±	1.79
	SSIM	0.639	±	0.077	0.843	±	0.048	0.859	±	0.044	**0.914**	±	0.027
	RMSE [ms]	122	±	26	94	±	19	91	±	22	**70**	±	16
$R = 8$	**PSNR**	23.87	±	1.90	28.70	±	1.42	28.93	±	1.53	**31.61**	±	1.48
	SSIM	0.623	±	0.063	0.799	±	0.043	0.810	±	0.043	**0.897**	±	0.026
	RMSE [ms]	140	±	25	117	±	24	114	±	25	**82**	±	18

Table 2: Quantitative MRI Reconstruction: Mean and standard deviation of the measures PSNR and SSIM of the qualitative images and RMSE of the T_1 parameter-maps over the test set.

		PDHG - $\lambda_{\tilde{\mathrm{P}}}^{xyt}$			**PDHG -** $\lambda_{\tilde{\mathrm{P}}}^{xy,t}$			**PDHG -** $\mathbf{\Lambda}_{\Theta}^{xy,t}$		
$\sigma = 0.1$	**SSIM**	0.941	±	0.017	0.934	±	0.021	**0.968**	±	0.011
	PSNR	33.26	±	1.53	34.65	±	1.75	**39.30**	±	2.06
	NRMSE	0.044	±	0.009	0.038	±	0.011	**0.022**	±	0.007
$\sigma = 0.2$	**SSIM**	0.837	±	0.075	0.914	±	0.021	**0.940**	±	0.021
	PSNR	31.32	±	2.25	33.52	±	1.63	**35.537**	±	2.24
	NRMSE	0.056	±	0.014	0.043	±	0.011	**0.035**	±	0.011
$\sigma = 0.3$	**SSIM**	0.649	±	0.105	0.814	±	0.067	**0.915**	±	0.028
	PSNR	28.09	±	2.54	31.04	±	2.06	**33.36**	±	2.33
	NRMSE	0.082	±	0.024	0.058	±	0.016	**0.045**	±	0.014

Table 3: Dynamic Image Denoising: Mean and standard deviation of the measures SSIM, PSNR and NRMSE and Blur obtained over the test set for the dynamic image denoising example. The TV-reconstruction using the proposed spatio-temporal parameter-maps $\mathbf{\Lambda}_{\Theta}^{xy,t}$ improves the results especially in terms of SSIM and PSNR.

	FBP			**PD3O - $\lambda^{xy}_{\tilde{\mathrm{P}}}$**			**PD3O - $\boldsymbol{\Lambda}^{xy}_{\Theta}$**		
PSNR	30.37	±	2.95	32.87	±	3.59	**33.90**	±	3.94
SSIM	0.739	±	0.141	0.796	±	0.152	**0.809**	±	0.157
NRMSE	0.101	±	0.028	0.079	±	0.032	**0.071**	±	0.033
Blur Effect	0.412	±	0.067	0.472	±	0.038	**0.407**	±	0.043
Training time	-			5 h			24 h		
Runtime	0.03 s			5.08 s			5.08 s		

Table 4: Low-Dose Computerized Tomography: Mean and standard deviation of the measures SSIM, PSNR, NRMSE and Blur effect obtained over the CT test set. The best value is marked in bold. The TV-reconstruction using the proposed parameter-maps $\boldsymbol{\Lambda}^{xy}_{\Theta}$ improve the results in every quality measure.