Nr. 66
19. December 2019

universität
innsbruck
Institut für Mathematik

Preprint-Series: Department of Mathematics - Applied Mathematics

Neural Networks-based Regularization of Large-Scale Inverse Problems in Medical Imaging

A. Kofler, M. Haltmeier, T. Schaeffter, M. Kachelriess,
M. Dewey, C. Wald, and C. Kolbitsch

APPLIEDMATHEMATICS

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

PREPRINT

Neural Networks-based Regularization of Large-Scale Inverse Problems in Medical Imaging

Andreas Kofler[1], Markus Haltmeier[2], Tobias Schaeffter[3], Marc Kachelriess[4], Marc Dewey[5], Christian Wald[1], and Christoph Kolbitsch[6]

[1] *Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany (e-mail: {andreas.kofler, christian.wald}@charite.de)*

[2] *Department of Mathematics, University of Innsbruck, Innsbruck, Austria (e-mail: markus.haltmeier@uibk.ac.at)*

[3] *Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany, King's College London, London, UK and the Department of Medical Engineering, Technical University of Berlin, Berlin, Germany (e-mail: tobias.schaeffter@ptb.de)*

[4] *Division of X-Ray Imaging and CT, German Cancer Research Center, Heidelberg, Germany (e-mail: marc.kachelriess@dkfz.de)*

[5] *Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany and the Berlin Institute of Health, Berlin, Germany (e-mail: marc.dewey@charite.de)*

[6] *Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany and King's College London, London, UK (e-mail: christoph.kolbitsch@ptb.de)*

December 19, 2019

Abstract

In this paper we present a generalized Deep Learning-based approach to solve ill-posed large-scale inverse problems occurring in medical imaging. Recently, Deep Learning methods using iterative neural networks and cascaded neural networks have been reported to achieve excellent image quality for the task of image reconstruction in different imaging modalities. However, the fact that these approaches employ the forward and adjoint operators repeatedly in the network architecture requires the network to process the whole images or volumes at once, which for some applications is computationally infeasible. In this work, we follow a different reconstruction strategy by decoupling the regularization of the solution from ensuring consistency with the measured data. The regularization is given in the form of an image prior obtained by the output of a previously trained neural network which is used in a Tikhonov regularization framework. By doing so, more complex and sophisticated network architectures can be used for the removal of the artefacts or noise than it is usually the case in iterative networks. Due to the large scale of the considered problems and the resulting computational complexity of the employed networks, the priors are obtained by processing the images or volumes as patches or slices. We evaluated the method for the cases of 3D cone-beam low dose CT and undersampled 2D radial cine MRI and compared it to a total variation-minimization-based reconstruction algorithm as well as to a method with regularization based on learned overcomplete dictionaries. The proposed method outperformed all the reported methods with respect to all chosen quantitative measures and further accelerates the regularization step in the reconstruction by several orders of magnitude.

Keywords: Deep Learning, Neural Networks, Inverse Problems, Low-Dose CT, Radial Cine MRI

1 Introduction

In inverse problems, the goal is to recover an object of interest from a set of indirect and possibly incomplete observations. In medical imaging, for example, a classical inverse problem is given by the task of reconstructing a diagnostic image from a certain number of measurements, e.g. X-ray projections in computed tomography (CT) or the spatial frequency information (k-space data) in magnetic resonance imaging (MRI). The reconstruction from the measured data can be an ill-posed inverse problem for different reasons. In low-dose CT, for example, the reconstruction from noisy data is ill-posed because of the ill-posedeness of the inversion of the Radon transform. In accelerated MRI, on the other hand, the reconstruction from incomplete data is ill-posed since the underlying problem is underdetermined and therefore no unique solution exists without integrating prior information.

In order to constrain the space of possible solutions, a typical approach is to impose specific a-priori chosen properties on the solution by adding a regularization (or penalty) term to the problem. Well known choices for the regularization are for example given by the popular total variation-minimization and sparse regularization approaches, where the solution is transformed using a sparsifying transform such as the Wavelet-transform or the Fourier-transform [19] or a finite-differences filter [5] and the L_1-norm of the latter is minimized. While the aforementioned methods use hand-crafted priors, other methods learn the regularization directly within the reconstruction of the images where the regularization is imposed patch-wise by the sparse approximation using a dictionary which is learned in an unsupervised manner during the reconstruction [33], [36]. However, these learning-based methods are usually time consuming since the regularization is adaptive and learned during an iterative reconstruction scheme. Further, in the specific dictionary learning framework, the regularization requires training of a dictionary and sparse coding of all patches of the current image estimate at each iteration. This is computationally demanding and makes the application in the clinical routine challenging.

Recently, Convolutional Neural Networks (CNNs) have been applied in the field of inverse problems, either as direct full inversion methods [38], as post processing methods [14], [27], [11], as learned iterative schemes [2], [3], or as learned regularizers [25], [17], [18], [4], [21]. When used as post-processing methods, the networks are trained to denoise or remove artefacts from images obtained by the direct reconstruction of the noisy or incomplete data. Although a wide range of different network architecture has been proposed, e.g. [11], [37], a major concern is that the estimated output of the CNN might lack data-consistency. In order to ensure the obtained image is consistent with the acquired raw data, methods have been proposed where the constructed networks define unrolled iterative schemes which employ the forward and the adjoint operators. These methods can be interpreted as learned iterative schemes and have been successfully applied to different imaging modalities [2], [3], [9], [13], [25], [17], [4], [21]. Thereby, the subnetworks containing trainable parameters can be thought of regularizers which are learned by end-to-end training of the whole network cascade. Due to the integration of the forward and the adjoint operators, iterative or cascaded networks seem to be the natural network of choice for any image reconstruction task. However, the main advantage of all these methods at the same time represents the computational bottleneck of the approaches. The fact that the forward and the adjoint operators are integrated as layers in the networks requires that the whole object of interest has to be processed at once. Since CNNs typically increase the input size by extracting several feature maps per layer, end-to-end training might be infeasible for some high-dimensional problems, including high-resolution 3D CT volumes or non-Cartesian MR acquisitions.

In order to overcome these limitations, we propose to decouple the regularization of the solution from ensuring consistency with the measured data. We present a general framework to use CNNs as learned regularizers and still ensure data-consistency of the obtained solution. In particular, we consider high-dimensional problems where either the object of interest or the measured data are high-dimensional (high-resolution 3D CT) or the evaluation of the forward or the adjoint operators is computationally expensive (dynamic 2D non-Cartesian radial MR acquisition).

This paper is organized as follows. In Section 2, we formally introduce the inverse problem of image reconstruction and motivate our proposed approach for the solution of large-scale ill-posed inverse problems. We demonstrate the feasibility of our method by applying it to 3D low-dose cone beam CT and 2D radial cine MRI in Section 3. We further compare the proposed approach to an iterative reconstruction method given by total variation-minimization (TV) and a learning-based method (DIC) using Dictionary Learning-based priors in Section 4. We then conclude the work with a discussion and conclusion in Section 5 and Section 6.

2 Iterative Image Reconstruction with CNN-Priors

In this Section, we present the proposed deep learning scheme for solving large-scale, possibly non-linear, inverse problems. For the sake of clarity, we do not focus on a functional analytical setting but consider discretized problems of the form

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \tag{1}$$

where $\mathbf{A}\colon X \to Y$ is a discrete forward operator, $\mathbf{y} \in Y$ is the measured data and $\mathbf{x} \in X$ the unknown object to be recovered, i.e. the diagnostic image. The operator $\mathbf{A}$ could for example model the measurement process in different imaging modalities such as the X-ray projection in CT or the Fourier encoding in MRI. Depending on the

PREPRINT

nature of the underlying problem one is considering, problem (1) can be ill-posed for different reasons. For example, in low-dose CT, the measurement data is inherently contaminated by noise. In cardiac MRI, k-space data is often undersampled in order to speed up the acquisition process. This leads to incomplete data and therefore to an undetermined problem with an infinite number of theoretically possible solutions. In order to constrain the space of solutions of interest, a typical approach is to impose specific a-priori chosen properties on the solution $\mathbf{x}$ by adding a regularization (or penalty) term $\mathcal{R}(\mathbf{x})$ and using Lagrange multipliers. Then, we solve the relaxed problem

$$D(\mathbf{A}\mathbf{x}, \mathbf{y}) + \lambda\,\mathcal{R}(\mathbf{x}) \to \min, \tag{2}$$

where $D(\,\cdot\,,\,\cdot\,)$ is an appropriately chosen data-discrepancy measure and $\lambda > 0$ controls the strength of the regularization. The choice of $D(\,\cdot\,,\,\cdot\,)$ depends on the considered problem. Clearly, the regularization term $\mathcal{R}(\mathbf{x})$ significantly affects the quality and the characteristics of the solution $\mathbf{x}$. Here, we propose a generalized approach for solving high-dimensional inverse problems by the following three steps: First, an initial guess of the solution is provided by a direct reconstruction from the measured data, i.e. $\mathbf{x}_\eta = \mathbf{A}^\dagger\mathbf{y}$, where $\mathbf{A}^\dagger$ denotes some reconstruction operator. Then, a CNN is used to remove the noise or the artefacts from the direct reconstruction $\mathbf{x}_\eta$ in order to obtain another intermediate reconstruction $\mathbf{x}_{\mathrm{CNN}}$ which is used as a CNN-prior in a Tikhonov functional

$$F_{\mathbf{y},\mathbf{x}_{\mathrm{CNN}},\lambda}(\mathbf{x}) := D(\mathbf{A}\mathbf{x}, \mathbf{y}) + \lambda\|\mathbf{x} - \mathbf{x}_{\mathrm{CNN}}\|_2^2 \to \min. \tag{3}$$

As a third and final step, the CNN-Tikhonov functional (3) is minimized resulting in the proposed CNN-based reconstruction.

Note that the regularization of the problem, i.e. obtaining the CNN-prior, is decoupled from the step of ensuring data-consistency of the solution via minimization of (3). This allows to use deeper and more sophisticated CNNs as the ones typically used in iterative networks. Given the high-dimensionality of the considered problems, network training is further carried out on sub-portions of the image samples, i.e. on patches or slices which are previously extracted from the images or volumes. This is motivated by the fact that in most medical imaging applications, one has typically access to datasets with only a relatively small number of subjects. The images or volumes of these subjects, on the other hand, are elements of a high-dimensional space. Therefore, one is concerned with the problem of having topologically sparse training data with only very few data points in the original high-dimensional image space. Working with sub-portions of the image samples increases the number of available data points and at the same time decreases its ambient dimensionality.

Let $\mathbf{x}_{\mathrm{f}}$ denote some ground truth image or volume and $\mathbf{x}_{\mathrm{est}}$ an estimate of $\mathbf{x}_{\mathrm{f}}$. Then, we can always find a decomposition of $\mathbf{x}_{\mathrm{f}}$ and $\mathbf{x}_{\mathrm{est}}$ in $N_{\mathbf{p},\mathbf{s}}$ possibly overlapping patches using the operators $\mathbf{R}_j^{\mathbf{p},\mathbf{s}}$ and $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal$, i.e.

$$\mathbf{x}_{\mathrm{f}} = \mathbf{W}_{\mathbf{p},\mathbf{s}} \sum_{j=1}^{N_{\mathbf{p},\mathbf{s}}} (\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal\, \mathbf{R}_j^{\mathbf{p},\mathbf{s}}\, \mathbf{x}_{\mathrm{f}}, \tag{4}$$

$$\mathbf{x}_{\mathrm{est}} = \mathbf{W}_{\mathbf{p},\mathbf{s}} \sum_{j=1}^{N_{\mathbf{p},\mathbf{s}}} (\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal\, \mathbf{R}_j^{\mathbf{p},\mathbf{s}}\, \mathbf{x}_{\mathrm{est}}, \tag{5}$$

where $\mathbf{R}_j^{\mathbf{p},\mathbf{s}}$ and $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal$ extract and reposition the patches at the original position, respectively and the diagonal operator $\mathbf{W}_{\mathbf{p},\mathbf{s}}$ accounts for weighting of regions containing overlaps. The tuples $\mathbf{p}$ and $\mathbf{s}$ specify the size of the patches and the strides and therefore the number $N_{\mathbf{p},\mathbf{s}}$ of patches extracted from a single image. Since the operator norm of $\mathbf{W}_{\mathbf{p},\mathbf{s}}$ is less or equal to one, by the triangle inequality, we have

$$\|\mathbf{x}_\mathrm{f} - \mathbf{x}_\mathrm{est}\|_2 \leq \sum_{j=1}^{N_{\mathbf{p},\mathbf{s}}} \|(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal \Big(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_\mathrm{f} - \mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_\mathrm{est}\Big)\|_2. \tag{6}$$

Assuming we have access to a finite set of N ground truth samples $(\mathbf{x}_{\mathrm{f},k})_{k=1}^N$ and corresponding estimates $(\mathbf{x}_{\mathrm{est},k})_{k=1}^N$, it analogously holds

$$e_N := \sum_{k=1}^{N} \|\mathbf{x}_{\mathrm{f},k} - \mathbf{x}_{\mathrm{est},k}\|_2 \leq \sum_{k=1}^{N}\sum_{j=1}^{N_{\mathbf{p},\mathbf{s}}} \|(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal \Big(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{f},k} - \mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{est},k}\Big)\|_2 =: e_{N,N_{\mathbf{p},\mathbf{s}}}. \tag{7}$$

Now assume that N is relatively small, which is usually the case for most medical imaging applications, and the considered samples $\mathbf{x}_{\mathrm{f},k}$ have a relatively large size (for example, 3D CT image volumes). At this point, we see that it might be beneficial trying to estimate $\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{f},k}^j$ with a neural network u_θ with trainable parameters θ, i.e. to find θ such that $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{est},k}^j)(\theta) \approx \mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{f},k}^j$ for all k and j, rather than trying to predict the whole sample at once, i.e. $\mathbf{x}_{\mathrm{est},k}(\theta) \approx \mathbf{x}_{\mathrm{f},k}$ for all k. By doing so, one has the advantage of having access to $N_{\mathbf{p},\mathbf{s}} \cdot N$ training samples of smaller size instead of N samples of larger size. If there exists a θ^* which minimizes $e_{N,N_{\mathbf{p},\mathbf{s}}}$ for $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{est},k})(\theta) = u_\theta(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\eta,k})$ for some input $\mathbf{x}_{\eta,k}$, then clearly, if $e_{N,N_p} \to 0$ for $\theta \to \theta^*$, also $e_N \to 0$. Using (5) we can obtain $\mathbf{x}_{\mathrm{est},k}$ for all k by estimating the single patches $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_{\mathrm{est},k})(\theta)$.

More precisely, we denote by f_θ the composite function which decomposes an image or volume into patches, applies a neural network u_θ to all patches, and reassembles the sample from them. This results in the proposed CNN-prior $\mathbf{x}_\mathrm{CNN}$ given by

$$\mathbf{x}_\mathrm{CNN} := f_\theta(\mathbf{x}_\eta) = \mathbf{W}_{\mathbf{p},\mathbf{s}} \sum_j (\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal (u_\theta(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}(\mathbf{x}_\eta))), \tag{8}$$

where $\mathbf{x}_\eta$ is the initial reconstruction obtained from the measured data $\mathbf{y}_\eta \approx \mathbf{A}\mathbf{x}$. The network u_θ is trained on a subset of pairs

$$\mathcal{D} = \Big\{ \Big(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}(\mathbf{x}_{\eta,k}), \mathbf{R}_j^{\mathbf{p},\mathbf{s}}(\mathbf{x}_{\mathrm{f},k})\Big) : (k,j) \in \mathcal{I}_{N,N_{\mathbf{p},\mathbf{s}}} \Big\}, \tag{9}$$

of all possible patches extracted from the N samples in the dataset, where $\mathcal{I}_{N,N_{\mathbf{p},\mathbf{s}}} := \{1,\ldots,N\} \times \{1,\ldots,N_{\mathbf{p},\mathbf{s}}\}$. During training, we optimize the set of parameters θ to minimize the L_2-error between the estimated output of the patches and the corresponding ground truth patch, i.e. we solve the following optimization problem

$$\mathcal{L}(\theta) = \frac{1}{N_\mathrm{train}} \sum_{(\mathbf{z}_\eta,\mathbf{z}) \in \mathcal{D}} \|u_\theta(\mathbf{z}_\eta) - \mathbf{z}\|_2^2 \to \min, \tag{10}$$

where N_train is the number of samples used for training the network u_θ. The inequality in (7) guarantees that the set of parameters θ^* found by minimizing (10) is also suitable for obtaining the prior $\mathbf{x}_{\mathrm{CNN}}$. Therefore, u_θ is powerful enough to deliver a CNN-prior to regularize the solution of (3). Figure 1 illustrates the processing of extracting a patch from a volume using the operator $\mathbf{R}_j^{\mathbf{p},\mathbf{s}}$, processing it with a neural network u_θ and repositioning it at the original position using the transposed operator $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal$. The example is shown for a 2D cine MR image sequence.

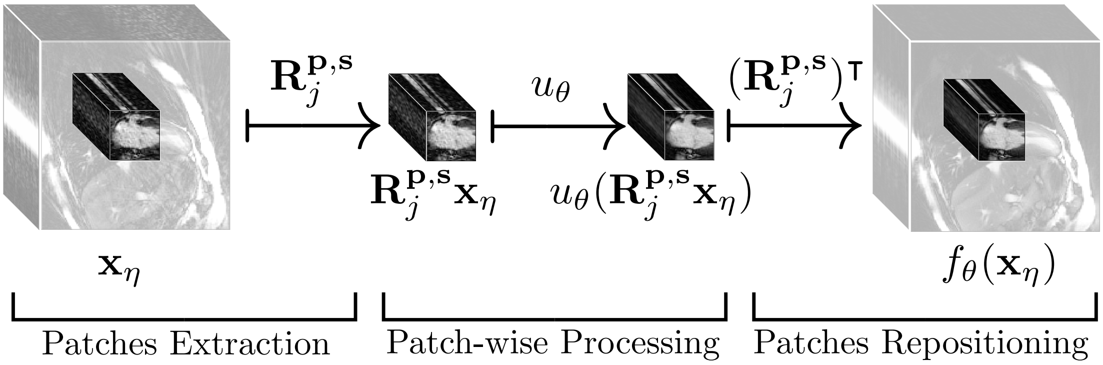
$\mathbf{R}_j^{\mathbf{p},\mathbf{s}}$
u_θ
$(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal$
$\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_\eta$
$u_\theta(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}\mathbf{x}_\eta)$
$\mathbf{x}_\eta$
$f_\theta(\mathbf{x}_\eta)$
Patches Extraction
Patch-wise Processing
Patches Repositioning

Figure 1: Workflow for obtaining a CNN-prior by patch-based processing: First, the initial reconstruction is divided into patches, then the network u_θ is applied to all patches. Reassembling all processed patches results in the CNN-prior which is then used for regularization of the inverse problem.

Finally, the optimality condition for problem (3) is solved with an iterative method which is typically dependent on the application. The solution of (3) is then the final CNN-based reconstruction. Algorithm 1 summarizes the complete reconstruction scheme. Note that the strategy for minimizing (3) depends on the specific application. In the case of an inverse problem with noisy measurements, (3) is only minimized approximately. For example, for the case of low-dose CT, early stopping of the Landweber iteration is already considered to be a regularization method which is applied due to the semi-convergence property of the Landweber iteration [29].

Algorithm 1 CNNs-based Regularized Reconstruction

```
  Data: trained network $u_\theta$, function $f_\theta$, noisy or incomplete measured data $\mathbf{y}_\eta \approx \mathbf{A}\mathbf{x}$,
  regularization parameter $\lambda > 0$
  Output: reconstruction $\mathbf{x}_{\mathrm{REC}}$
  1) $\mathbf{x}_\eta \leftarrow \mathbf{A}^\dagger \mathbf{y}_\eta$
  2) $\mathbf{x}_{\mathrm{CNN}} \leftarrow f_\theta(\mathbf{x}_\eta)$
  3) $\mathbf{x}_{\mathrm{REC}} \leftarrow \arg\min_{\mathbf{x}} \; D(\mathbf{A}\mathbf{x}, \mathbf{y}_\eta) + \lambda \|\mathbf{x} - \mathbf{x}_{\mathrm{CNN}}\|_2^2$
  Return $\mathbf{x}_{\mathrm{REC}}$
```

3 Experiments

In the following, we evaluated our proposed method on two different examples of large-scale inverse problems given by 3D low-dose CT and 2D undersampled radial cine MRI. We compared our proposed method to the well-known TV-minimization-based and dictionary learning-based approaches presented in [5], [33] and [31], [36], which we abbreviate by TV and DIC, respectively. Further details about the comparison methods are discussed later in the paper.

3.1 2D Radial Cine MRI

Here we applied our method to image reconstruction in undersampled 2D radial cine MRI. Typically, MRI is performed using multiple receiver coils and therefore, the inverse problem is given by

$$\mathbf{E}_I \mathbf{x} = \mathbf{y}_I, \tag{11}$$

where $\mathbf{x} \in \mathbb{C}^N$ with $N = N_x \cdot N_y \cdot N_t$ is an unknown complex-valued image sequence. The encoding operator $\mathbf{E}_I$ is given by $\mathbf{E}_I = \mathbf{S} \circ \mathbf{E} \circ \mathbf{C}$ where

$$\begin{aligned} \mathbf{C} &= [\mathbf{C}_1, \ldots, \mathbf{C}_{n_c}]^\intercal, && (12) \\ \mathbf{E} &= \operatorname{diag}(\mathbf{F}, \ldots, \mathbf{F}), && (13) \\ \mathbf{S} &= \operatorname{diag}(\mathbf{S}_I, \ldots, \mathbf{S}_I). \end{aligned}$$

Here, $\mathbf{C}_i$ denotes the i-th coil sensitivity map, n_c is the number of coil-sensitivity maps, $\mathbf{F}$ the 2D frame-wise operator and $\mathbf{S}_I$ with $I \subset J = \{1, \ldots, N_{\text{rad}}\}$, $|I| := m \leq N_{\text{rad}}$, a binary mask which models the undersampling process of the N_{rad} Fourier coefficients sampled on a radial grid. The vector $\mathbf{y}_I \in \mathbb{C}^M$ with $M = m \cdot n_c$ corresponds to the measured data. Here, we sampled the k-space data along radial trajectories chosen according to the golden-angle method [35]. Note that problem (11) is mainly ill-posed not due to the presence of noise in the acquisition, but because the data acquisition is accelerated and hence only a fraction of the required measurements is acquired.

If we assume a radial data-acquisition grid, problem (11) is a large-scale inverse problem mainly because of two reasons. First, the measurement vector $\mathbf{y}_I$ corresponds to n_c copies of the Fourier encoded image data multiplied by the corresponding coil sensitivity map. Second, the adjoint operator $\mathbf{E}_I^H$ consists of two computationally demanding steps. The radially acquired k-space data is first properly re-weighted and interpolated to a Cartesian grid, for example by using Kaiser-Bessel functions [22]. Then, a 2D inverse Fourier operation is applied to the image of each cardiac phase and the final image sequence is obtained by weighting the images from each estimated coil-sensitivity map and combining them to a single image sequence. We refer to the reconstruction obtained by $\mathbf{x}_I = \mathbf{E}_I^H \mathbf{y}_I$ as the non-uniform fast Fourier-transform (NUFFT) reconstruction. Therefore, in radial multi-coil MRI, the measured k-space data is high-dimensional and the application of the encoding operators $\mathbf{E}_I$ and $\mathbf{E}_I^H$ is further more computationally demanding than sampling on a Cartesian grid, see e.g [28]. This makes the construction of cascaded networks which also process the k-space data [10] or by repeatedly employing the forward and adjoint operators [25], [21] computationally challenging. Therefore, decoupling the regularization given by the CNNs from the data-consistency step is necessary in this case.

We solve a regularized version of problem (11) by considering

$$F_{\mathbf{y}_I, \mathbf{x}_{\text{CNN}}, \lambda}(\mathbf{x}) = \|\mathbf{E}_I \mathbf{x} - \mathbf{y}_I\|_2^2 + \lambda \|\mathbf{x} - \mathbf{x}_{\text{CNN}}\|_2^2 \to \min, \tag{14}$$

where $\mathbf{x}_{\mathrm{CNN}}$ is obtained a-priori by using an already trained network. For this example, for obtaining the CNN-prior $\mathbf{x}_{\mathrm{CNN}}$, we adopted the XT,YT approach presented in [16], where a modified version of the 2D U-net is used to process spatio-temporal slices which can be extracted from the image sequence. Since the XT,YT method was previously introduced to only process real-valued data (i.e. the magnitude images), we followed a similar strategy by processing the real and imaginary parts of the image sequences separately but using the same real-valued network u_θ. This further increases the amount of training data by a factor of two.

More precisely, let $\mathbf{R}_j^{xt}$ and $\mathbf{R}_j^{yt}$ denote the operators which extract the j-th two-dimensional spatio-temporal slices in xt- and yt-direction from a 3D volume. Assuming $N_x = N_y$, we denote by $\mathbf{R}_j^{xt,yt}$ the composition $\mathbf{R}_j^{xt,yt} = (\mathbf{R}_j^{xt} + \mathbf{R}_j^{yt})$ and by $(\mathbf{R}_j^{xt,yt})^\intercal$ its transposed operation which repositions the spatio-temporal slices at their original position. By u_θ we denote a 2D U-net as the one described in [16] which is trained on spatio-temporal slices, i.e. on a dataset of pairs which consist of the spatio-temporal slices in xt- and yt-direction of both the real and imaginary parts of the complex-valued images. The network u_θ was trained to minimize the L_2-error between the ground truth image and the estimated output of the CNN. Our dataset consists of radially acquired 2D cine MR images from $n = 19$ subjects (15 healthy volunteers and 4 patients with known cardiac dysfunction) with 30 images covering the cardiac cycle. The ground truth images were obtained by kt-SENSE reconstruction using $N_\theta = 3400$ radial lines. We retrospectively generated the radial k-space data $\mathbf{y}_I$ by sampling the k-space data along $N_\theta = 1130$ radial spokes using $n_c = 12$ coils. Note that sampling $N_\theta = 3400$ already corresponds to an acceleration factor of approximately ~ 3 and therefore, $N_\theta = 1130$ corresponds to an accelerated data-acquisition by an approximate factor of ~ 9. The forward and the adjoint operators $\mathbf{E}_I$ and $\mathbf{E}_I^H$ were implemented using the `ODL` library [1]. The CNN-regularized (complex-valued) image sequence $\mathbf{x}_{\mathrm{CNN}}$ was obtained by

$$\mathbf{x}_{\mathrm{CNN}} = f_\theta(\mathbf{x}_I) = \frac{1}{4}\sum_j (\mathbf{R}_j^{xt,yt})^\intercal \big(u_\theta(\mathbf{R}_j^{xt,yt}(\mathrm{Re}\,\mathbf{x}_I))\big) + \mathrm{i}\Big((\mathbf{R}_j^{xt,yt})^\intercal \big(u_\theta(\mathbf{R}_j^{xt,yt}(\mathrm{Im}\,\mathbf{x}_I))\big)\Big)$$

Given $\mathbf{x}_{\mathrm{CNN}}$, functional (14) was minimized by setting the derivative with respect to $\mathbf{x}$ to zero and applying the pre-conditioned conjugate gradient (PCG) method to iteratively solve the resulting system. Since $\frac{1}{4}\sum_j (\mathbf{R}_j^{xt,yt})^\intercal \mathbf{R}_j^{xt,yt} = \mathbf{I}$, where $\mathbf{I}$ is the identity-operator, PCG was used to solve the system $\mathbf{H}\mathbf{x} = \mathbf{b}$ with

$$\begin{aligned} \mathbf{H} &= \mathbf{E}_I^H \mathbf{E}_I + \lambda\,\mathbf{I}, \\ \mathbf{b} &= \mathbf{x}_I + \lambda\,\mathbf{x}_{\mathrm{CNN}}. \end{aligned} \tag{15}$$

Since the XT,YT method gives access to a large number of training samples, training the network u_Θ for 12 epochs was sufficient. The CNN was trained by minimizing the L_2-norm of the error between labels and output by using the Adam optimizer [15]. We split our dataset in 12/3/4 subjects for training, validation and testing and performed a 4-fold cross-validation. For the experiment, we performed $n_{\mathrm{iter}} = 16$ subsequent iterations of PCG and we empirically chose $\lambda = 0.1$. The obtained results can be found in Subsection 4.1.

3.2 3D Low-Dose Computed Tomography

The current generation of CT scanners performs the data-acquisition by emitting X-rays along trajectories in the form of a cone-beam for each angular position of the scanner. Therefore, for each angle ϕ of the rotation, one obtains an X-ray image which is measured by the detector array and thus, the complete sinogram data can be identified with a 3D array of shape $(N_\phi, N_{r_x}, N_{r_y})$. Thereby, N_ϕ corresponds to the number of angles the rotation of the scanner is discretized by and N_{r_x} and N_{r_y} denote the number of elements of the detector array. The values of these parameters vary from scanner to scanner but are in the order of $N_\phi \approx 1000$ for a full rotation of the scanner and $N_{r_x} \times N_{r_y} \approx 320 \times 800$ for a 320-row detector array, which is for example used for cardiac CT scans [8]. The volumes obtained from the reconstructions are typically given by an in-plane number of pixels of $N_x \times N_y = 512 \times 512$ and varying number of slices N_z, dependent on the specific application. For this example, we consider a similar set-up as in [2]. The non-linear problem is given by

$$\mathbf{y}_\eta = \mathbf{T}\mathbf{x} + \eta = p \exp\{-\mu\, \mathbf{R}\mathbf{x}\} + \eta, \tag{16}$$

where p denotes the average number of photons per pixel, μ is the linear attenuation coefficient of water, $\mathbf{R}$ corresponds to the discretized version of a ray-transform with cone-beam geometry and the vector η denotes the Poisson-distributed noise in the measurements. Following our approach, we are interested in solving the following problem:

$$F_{\mathbf{y}_\eta, \mathbf{x}_{\text{CNN}}, \lambda}(\mathbf{x}) = D_{\text{KL}}(\mathbf{T}\mathbf{x}, \mathbf{y}_\eta) + \lambda \|\mathbf{x} - \mathbf{x}_{\text{CNN}}\|_2^2 \rightarrow \min, \tag{17}$$

where D_{KL} denotes the Kullback-Leibler divergence which corresponds to the log-likelihood function for Poisson-distributed noise. According to the previously introduced notation, the prior $\mathbf{x}_{\text{CNN}}$ is given by $\mathbf{x}_{\text{CNN}} = f_\theta(\mathbf{x}_\eta)$, where f_θ denotes a CNN-based processing method with trainable parameters θ and $\mathbf{x}_\eta = \mathbf{R}^\dagger(-\mu^{-1}\ln(p^{-1}\mathbf{y}_\eta))$ with $\mathbf{R}^\dagger$ being the filtered back-projection (FBP) reconstruction. Since our object of interest $\mathbf{x}$ is a volume, it is intuitive to choose a NN which involves 3D convolutions in order to learn the filters by exploiting the spatial correlation of adjacent voxels in x-, y- and z-direction. In this particular case, u_θ denotes a 3D U-net similar to the one presented in [12]. Due to the large dimensionality of the volumes $\mathbf{x}$, the network u_θ cannot be applied to the whole volume. Instead, following our approach, the volume was divided into patches to which the network u_θ is applied. Therefore, the output $\mathbf{x}_{\text{CNN}}$ was obtained as described in (8), where u_θ operates on 3D patches given by the vector $\mathbf{p} = (128, 128, 16)$, which denotes the maximal size of 3D patches which we were able to process by a 3D U-net. The strides used for the extraction and the reassembling of the volumes used in (8) is empirically chosen to be $\mathbf{s} = (16, 16, 8)$.

Training of the network u_θ was performed on a dataset of pairs according to (9), where we retrospectively generated the measurements $\mathbf{y}_\eta$ by simulating a low-dose scan on the ground truth volumes. For the experiment, we used 16 CT volumes from the randomized DISCHARGE trial [20] which we cropped to a fixed size of $512 \times 512 \times 128$. The simulation of the low-dose scan was performed as described in [2] by setting $p = 10\,000$ and $\mu = 0.02$. The operator $\mathbf{R}$ is assumed to perform $N_\phi = 1000$ projections which are measured by a detector array of shape $N_{r_x} \times N_{r_y} = 320 \times 800$. For the implementation of the operators, we used the `ODL` library [1]. The source-to-axis and source-to-detector distances were chosen according to the DICOM files. Since the dataset is relatively small, we performed a 7-fold cross-validation where for each fold we split the dataset in 12 patients for training, 2 for validation and 2 for testing. The number of training samples N_{train} results from the number of patches times the number of volumes contained in the training set. We trained the network u_θ for 115 epochs by minimizing the L_2-norm of the error between labels and outputs. For training, we used the Adam optimizer [15]. With the described configuration of $\mathbf{p}$ and $\mathbf{s}$, the resulting number of patches to be processed in order to obtain the prior $\mathbf{x}_{\mathrm{CNN}}$ is therefore given by $N_{\mathbf{p},\mathbf{s}} = 9\,375$. In this example, the solution $\mathbf{x}_{\mathrm{REC}}$ to problem (17) was then obtained by performing $n_{\mathrm{iter}} = 4$ iterations of Landweber's method where we further used the filtered-back projection $\mathbf{R}^\dagger$ as a left-preconditioner to accelerate the convergence of the scheme. For the derivation of the gradient of (17) with respect to $\mathbf{x}$, we refer to [2]. The regularization parameter was empirically set to $\lambda = 1$. The results can be found in Subsection 4.2.

3.3 Reference Methods

Here we discuss the methods of comparison in more detail and report the times needed to process and reconstruct the images or volumes. The data-discrepancy term $D(\,\cdot\,,\,\cdot\,)$ was again chosen according to the considered examples as previously discussed. The TV-minimization approach used for comparison is given by solving

$$\arg\min\nolimits_{\mathbf{x}} D(\mathbf{A}\mathbf{x}, \mathbf{y}) + \lambda \|\mathbf{G}\mathbf{x}\|_1, \tag{18}$$

where $\mathbf{G}$ denotes the discretized version of the isotropic first order finite differences filter in all three dimensions. The solution of problem (18) was obtained by introducing an auxiliary variable $\mathbf{z}$ and alternating between solving for $\mathbf{x}$ and $\mathbf{z}$. For the solution of one of the sub-problems, an iterative shrinkage method was used, see [7] for more details. The second resulting sub-problem was solved by iteratively solving a system of linear equations, either by Landweber for the CT example or by PCG for the MRI example, as mentioned before.

The dictionary learning-based method used for comparison is given by the solution of the problem

$$\arg\min\nolimits_{\mathbf{x}} D(\mathbf{A}\mathbf{x} - \mathbf{y}) + \lambda \|\mathbf{x} - \mathbf{x}_{\mathrm{DIC}}\|_2^2, \tag{19}$$

where, in contrast to our proposed method, $\mathbf{x}_{\mathrm{DIC}}$ was obtained by the patch-wise sparse approximation of the initial image estimate using an already trained dictionary $\mathbf{D}$. Therefore, using a similar notation as in (8), the prior $\mathbf{x}_{\mathrm{DIC}}$ is given by

$$\mathbf{x}_{\mathrm{DIC}} = \mathbf{W}_{\mathbf{p},\mathbf{s}} \sum_j (\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^\intercal\, \mathbf{D}\gamma_j,, \tag{20}$$

where the dictionary $\mathbf{D}$ was previously trained by 15 iterations of the iterative thresholding and K residual means algorithm (ITKRM) [26] on a set of ground truth images which were given by the high-dose images for the CT example and the kt-SENSE reconstructions from $N_\theta = 3400$ radial lines for the MRI example. Note that for each fold, for training the dictionary $\mathbf{D}$, we only used the data which we included in the training set for our method. This means we trained a total of seven dictionaries for the CT example and four dictionaries for the MRI example. For each iteration of ITKRM, we randomly selected a subject to extract $10\,000$ 3D training patches. The corresponding sparse codes γ_j were then obtained by solving

$$\min_{\{\gamma_j\}_j} \sum_j \big(\|\mathbf{R}_{j,(\mathbf{p},\mathbf{s})}\mathbf{x}_0 - \mathbf{D}\gamma_j\|_2^2 + \|\gamma_j\|_0\big), \tag{21}$$

which is a sparse coding problem and was solved using orthogonal matching pursuit (OMP) [32]. Thereby, the image $\mathbf{x}_0$ corresponds to either the FBP-reconstruction $\mathbf{x}_\eta$ for the CT example or to the NUFFT-reconstruction $\mathbf{x}_I$ for the MRI example. In both cases, we used patches of shape given by $\mathbf{p} = (4, 4, 4)$ and strides given by $\mathbf{s} = (2, 2, 2)$. The number of atoms K and the sparsity levels were set to $K = 4 \cdot d$, with $d = 4 \cdot 4 \cdot 4$ and $S = 16$. Note that, in contrast to [36] and [34], [6], the dictionary and the sparse codes were not learned during the reconstruction, as the sparse coding step of all patches would be too time consuming for very large-scale inverse problems, such as the CT example. Instead, the dictionary and the sparse codes were used to generate the prior $\mathbf{x}_{\mathrm{DIC}}$ which makes the method also more similar and comparable to ours.

3.4 Quantitative Measures

For the evaluation of the reconstructions we report the normalized root mean squared error (NRMSE) and the peak signal-to-noise ratio (PSNR) as error-based measures and the structural similarity index measure (SSIM) [34] and the Haar Wavelet-based perceptual similarity index measure (HPSI) [23] as image-similarity-based measures. The reported statistics were obtained by calculating the measures of the images in the xy-plane and averaging them over the different folds.

4 Results

4.1 Results for 2D Radial Cine MRI

Figure 2 shows an example of the results obtained with our proposed method. Figure 2A shows the initial NUFFT-reconstruction $\mathbf{x}_I$ obtained from the undersampled k-space data $\mathbf{y}_I$. The CNN-prior $\mathbf{x}_{\mathrm{CNN}}$ obtained by the XT,YT network can be seen in Figure 2B and shows a strong reduction of undersampling artefacts but also blurring of small structures as indicated the yellow arrows. The CNN-prior $\mathbf{x}_{\mathrm{CNN}}$ is then used as a prior in functional (14) which is subsequently minimized in order to obtain the solution $\mathbf{x}_{\mathrm{REC}}$ which can be seen in Figure 2C. Figure 2D shows the kt-SENSE reconstruction from the complete sampling pattern using $N_\theta = 3400$ radial spokes for the acquisition. From the point-wise error images, we clearly see that the NRMSE is further reduced after performing the further iterations to minimize the CNN-prior-regularized functional. Further, fine details are recovered as can be seen from the yellow arrows in Figure 2C.

Figure 3 shows a comparison of all different reported methods. As can be seen from the point-wise error in Figure 3B, the TV-minimization [5] method was able to eliminate

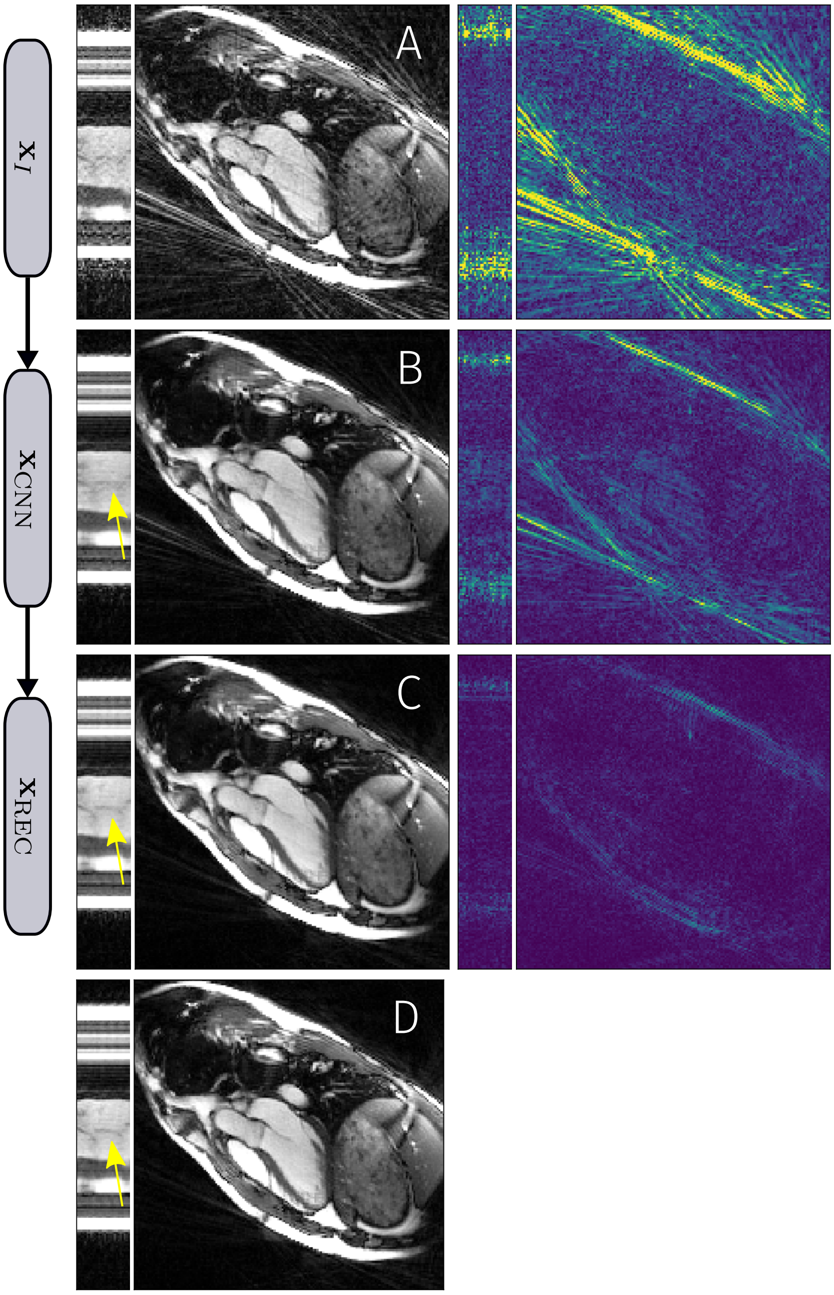

Figure 2: Results for a healthy volunteer showing two slices with different orientations. A: Initial NUFFT-reconstruction $\mathbf{x}_I$ using $N_\theta = 1130$ radial spokes, B: estimated output $\mathbf{x}_{\mathrm{CNN}}$ using the spatio-temporal 2D XT,YT U-net, C: solution of the CNNs-based regularized functional $\mathbf{x}_{\mathrm{REC}}$, D: ground truth image reconstruction with kt-SENSE and $N_\theta = 3400$ radial spokes. All images are displayed in the same scale. For better visibility, the point-wise error images are magnified by a factor of $\times 3$. The yellow arrows point at details which are smoothed out in the CNN-prior $\mathbf{x}_{\mathrm{CNN}}$ but are visible again in the final reconstruction $\mathbf{x}_{\mathrm{REC}}$.

some artefacts but less accurately compared to both learning-based methods, see Figure 3C and Figure 3D. Table 1 lists the obtained quantitative measures for all methods

am PREPRINT

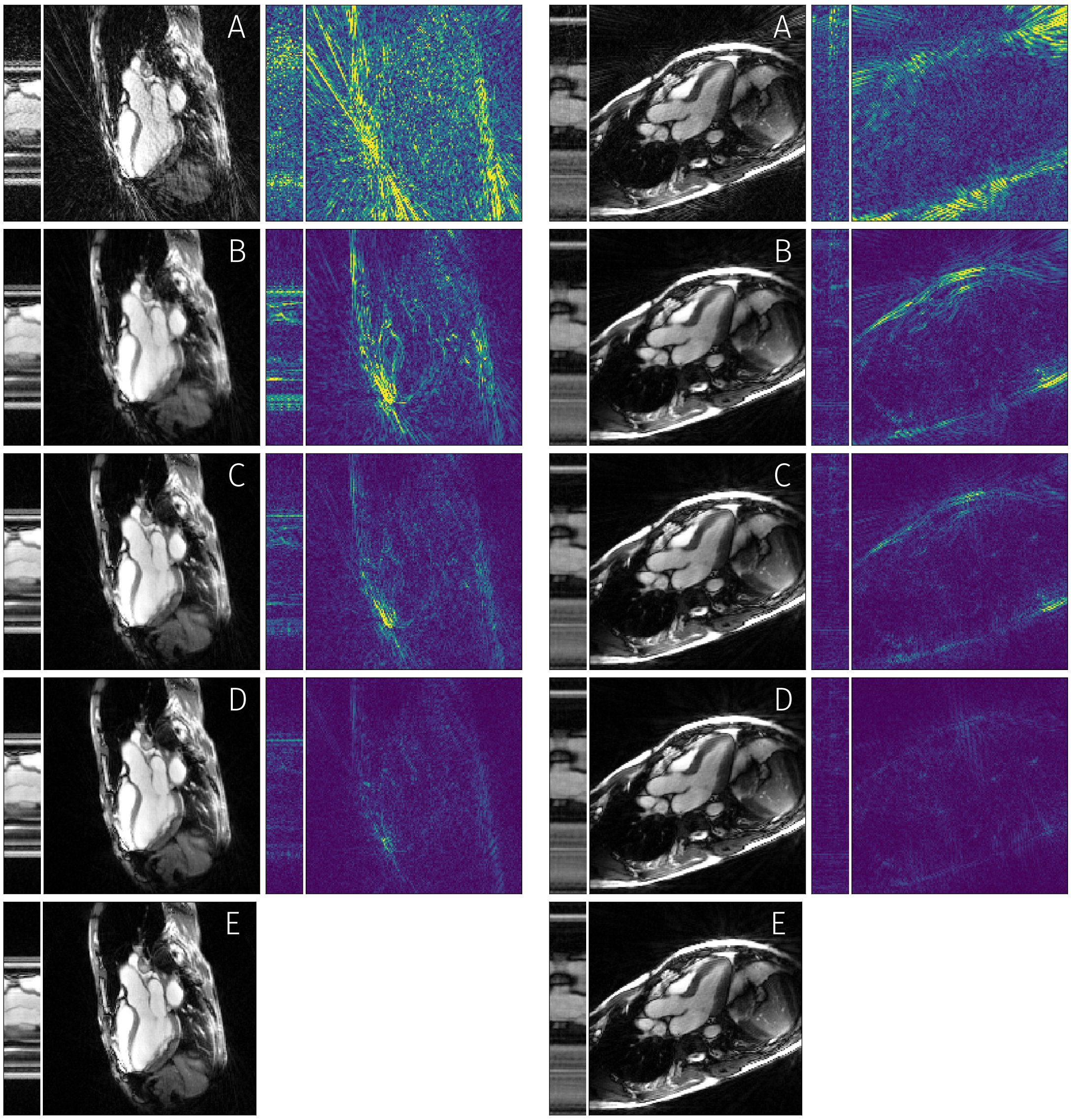

Figure 3: Results for a patient (left panel) and a healthy volunteer (right panel). A: Initial NUFFT-reconstruction $\mathbf{x}_I$ using $N_\theta = 1130$ radial spokes, B: solution of the TV-minimization approach (TV), C: dictionary learning-based regularization solution (DIC), D: CNN-regularized solution $\mathbf{x}_{\text{REC}}$, E: ground truth images obtained by kt-SENSE using $N_\theta = 3400$ radial spokes. All images are displayed in the same scale. For better visibility, the point-wise error images are magnified by a factor of $\times 5$. The point-wise error is the lowest for the reconstruction $\mathbf{x}_{\text{REC}}$.

averaged over the 4 different folds. From Table 1, we see that the DIC method yielded better results than TV with respect to all reported measures. Our proposed solution $\mathbf{x}_{\text{REC}}$ further surpassed the dictionary learning-based method, by additionally increasing the PSNR and SSIM by approximately 3dB and 0.04, respectively. The difference with respect to HPSI, on the other hand, is relatively small. Our method also reduced the NRMSE by about 0.014 compared to the DIC method. In addition, from Table 1, we see that for this example, even though processing the initial NUFFT-reconstruction with a CNN improved image quality with respect to all reported measures, further iterations to minimize the CNN-prior regularized functional increased data-consistency and additionally improved the PSNR, SSIM, HPSI and NRMSE. In fact, the statistics of the CNN-prior show that only post-processing the initial NUFFT-reconstruction leads to results which are inferior to the DIC method with respect to all reported measures.

	NUFFT	$\mathbf{x}_{\mathrm{CNN}}$	$\mathbf{x}_{\mathrm{REC}}$	**TV**	**DIC**
PSNR	36.8023	42.5647	48.7752	41.6968	45.4743
NRMSE	0.1228	0.0612	0.0302	0.0693	0.0442
SSIM	0.6649	0.7876	0.952	0.8635	0.9175
HPSI	0.9679	0.9910	0.9985	0.9878	0.9959

Table 1: Quantitative measures of the intermediate steps of our proposed framework, the TV-minimization method and the dictionary learning-based method. The measures are obtained as averages over the four different folds.

4.2 Results for 3D Low-Dose CT

Figure 4 shows all the intermediate results obtained with the proposed method. Figure 4A shows the initial FBP-reconstruction which is contaminated by noise. The FBP-reconstruction was then processed using the function f_θ described in (8) to obtain the prior $\mathbf{x}_{\mathrm{CNN}}$ which can be seen in Figure 4B. From the point-wise error, we see that patch-wise post-processing with the 3D U-net removed a large portion of the noise resulting from the low-dose acquisition. Solving problem (17) increases data-consistency since we make use of the measured data $\mathbf{y}_\eta$. Note that in contrast to the previous example of undersampled radial MRI, the minimization of the functional increased data-consistency of the solution but also contaminated the solution with noise, since the measured data is noisy due to the simulated low-dose scan protocol.

Table 2 summarizes the obtained quantitative measures for all intermediate reconstructions of our approach as well as for the TV and the DIC method. In the first three columns of Table 2 we see the results obtained for all three intermediate reconstructions of our proposed scheme. The reconstruction metrics improved substantially from the FBP-reconstruction to the estimated prior $\mathbf{x}_{\mathrm{CNN}}$. The difference in terms of PSNR was almost 10 dB, while the NRMSE decreased by approximately 0.11. Further, the similarity measures SSIM and HPSI were increased by about 0.14 and 0.04, respectively. Finally, the estimated solution given by $\mathbf{x}_{\mathrm{REC}}$ which was obtained by performing $n_{\mathrm{iter}} = 4$ iterations of Landweber to minimize (17) showed a slight decrease in PSNR and NRMSE which is related to the use of the noisy-measured data. However, fine diagnostic details as the coronary arteries are still visible in the prior $\mathbf{x}_{\mathrm{CNN}}$ and in the solution $\mathbf{x}_{\mathrm{REC}}$ as indicated by the yellow arrows. SSIM slightly increased while HPSI stayed approximately the same.

Figure 5 shows a comparison of images obtained by the different reconstruction methods. In Figure 5A, we see again the FBP-reconstruction obtained from the noisy data. Figure 5B shows the result obtained by the TV-minimization method which removed some of the noise as can be taken from the point-wise error image. The result ob-

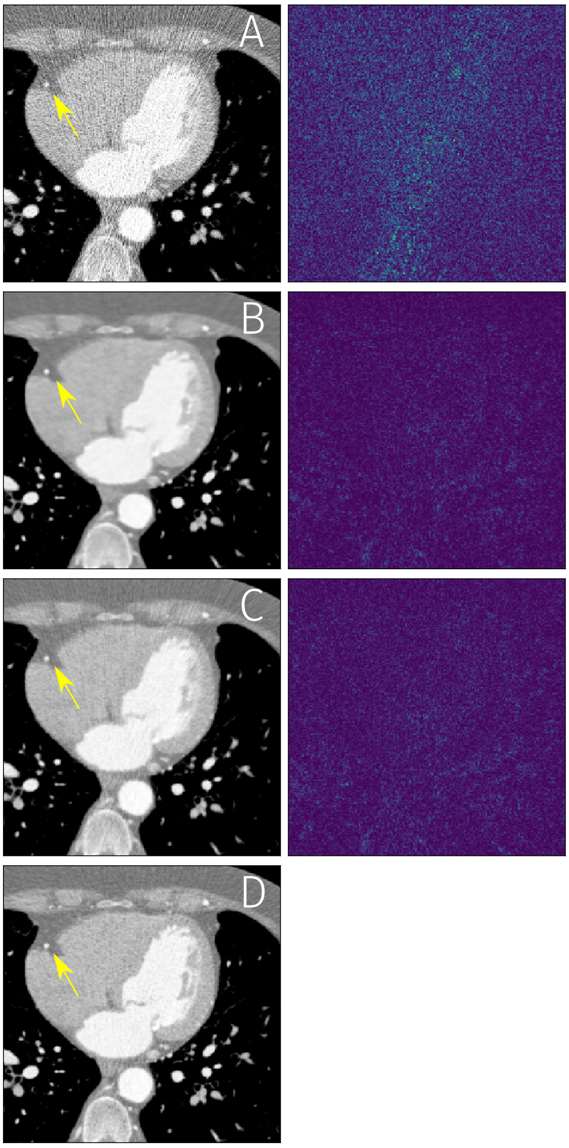

Figure 4: Axial view of image reconstructions of low-dose 3D CT data of a 55 years old female patient. A: Low-dose FBP-reconstruction $\mathbf{x}_\eta$, B: estimated output $\mathbf{x}_{\mathrm{CNN}}$ using a 3D U-net, C: solution of the CNNs-based regularized functional $\mathbf{x}_{\mathrm{REC}}$, D: ground truth image. The yellow arrow points at the right coronary artery, which is visible in the prior $\mathbf{x}_{\mathrm{CNN}}$ as well as in the final reconstruction $\mathbf{x}_{\mathrm{REC}}$. All images are windowed and displayed on the scale with $C = 0\,$HU, $W = 850\,$HU.

tained by the DIC method can be seen in Figure 5C which further reduced image noise compared to the TV method and surpasses TV with respect to the reported statistics, as can be seein in Table 2 . Finally, Figure 5D shows the solution $\mathbf{x}_{\mathrm{REC}}$ obtained with our proposed scheme and Figure 5E shows the ground truth image. The reconstruction using the CNN output as a prior further increased the PSNR, SSIM and HPSI by also reducing the NRMSE as can be taken from Table 2.

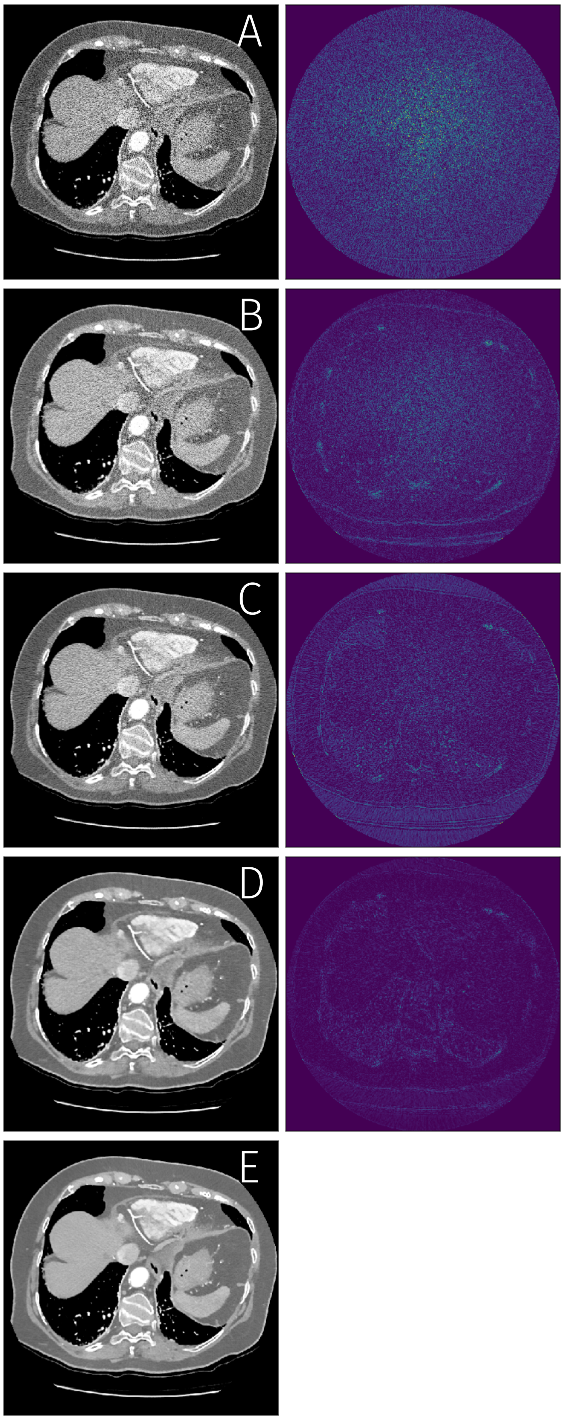

Figure 5: Axial view of image reconstructions of low-dose 3D CT data of a 76 years old female patient. A: Low-dose FBP-reconstruction $\mathbf{x}_\eta$, B: TV-minimization based reconstruction (TV), C: DIC-regularization based reconstruction (DIC), D: CNN-regularization based reconstruction $\mathbf{x}_{\mathrm{REC}}$, E: ground truth image. All images are windowed and displayed on the same scale with $C = 0\,$HU, $W = 800\,$HU.

	FBP	$\mathbf{x}_{\mathrm{CNN}}$	$\mathbf{x}_{\mathrm{REC}}$	**TV**	**DIC**
PSNR	30.0052	40.3546	39.6264	33.946	34.7807
NRMSE	0.1657	0.0498	0.0538	0.1051	0.0938
SSIM	0.425	0.5755	0.5813	0.4985	0.5465
HPSI	0.9394	0.9821	0.9819	0.9503	0.9581

Table 2: Quantitative measures of the intermediate steps of our proposed framework, the TV-minimization method and the dictionary learning-based method. The measures are obtained as averages over the seven different folds.

4.3 Reconstruction Times

Table 3 summarizes the times for the different components of the reconstructions using all different approaches for both examples. The abbreviations "SHRINK" and "LS1" stand for "shrinkage" and "linear system - one iteration" and denote the times which are needed to apply the iterative shrinkage method for the TV approach and to solve the sub-problems which are solved using iterative schemes, respectively.

		3D Low Dose CT	**2D Radial Cine MRI**
$\mathbf{A}^{\dagger}\mathbf{y}_{\eta}$		≈ 23 s (FBP)	≈ 11 s (NUFFT)
TV	SHRINK	$\ll 1$ s	$\ll 1$ s
	LS1	≈ 40 s	$\approx 1:20$m
	Total	≈ 11 m	≈ 42 m
DIC	$\mathbf{x}_{\mathrm{DIC}}$	$\approx$ **1:24 h**	$\approx$ **7 m**
	LS1	≈ 40 s	$\approx$ 1:20 m
	Total	$\approx$ 1:28 h	≈ 28 m
Proposed	$\mathbf{x}_{\mathrm{CNN}}$	$\approx$ **4 m**	$\approx$ **5 s**
	LS1	≈ 40 s	$\approx$ 1:20 m
	Total	≈ 8 m	≈ 21 m

Table 3: Reconstruction and processing times for the different methods for one 3D CT volume and a 2D cine MR image sequence.

Obviously, in terms of achieved image quality, the advantage of the DIC- and the CNN-based Tikhonov regularization are given by obtaining stronger priors which allow to use a smaller number of iterations to regularize the solution. The advantage of our proposed approach compared to the dictionary learning-based is the highly reduced time to compute the prior which is used for regularization. The reason lies in the fact that the DIC-based method requires to solve problem (21) to obtain the prior $\mathbf{x}_{\mathrm{DIC}}$, while in our method a CNN is used to obtain the prior $\mathbf{x}_{\mathrm{CNN}}$. Since problem (21) is separable, OMP is applied for each image/volume patch which is prohibitive as the number of overlapping patches in a 3D volume is in the order of $\mathcal{O}(N_x \cdot N_y \cdot N_z)$ or $\mathcal{O}(N_x \cdot N_y \cdot N_t)$, respectively. Obtaining $\mathbf{x}_{\mathrm{CNN}}$, on the other hand, does not involve the solution of any minimization problem but only requires the application of the network u_θ to the different patches. As this corresponds to matrix-vector multiplications with sparse matrices, its computational cost is lower and the calculations are further highly accelerated by performing the computations on a GPU.

5 Discussion

The proposed three-steps reconstruction scheme provides a general framework for solving large-scale inverse problems. The method is motivated by the observations stated in the ablation study [17], where the performance of cascades of CNNs with different numbers of intercepting data-consistency layers but approximately fixed number of trainable parameters was studied. First, it was noted that the replacement of simple blocks of convolutional layers by multi-scale CNNs given by U-nets had a visually positive impact on the obtained results. Further, it was empirically shown that the results obtained by cascades of U-nets of different length but with approximately the same number of trainable parameters were all visually and quantitatively comparable in terms of all reported measures. This suggests that, for large-scale problems, where the construction of cascaded networks might be infeasible, investing the same computational effort and expressive power in terms of number of trainable parameters in one single network might be similarly beneficial to intercepting several smaller sub-networks by data-consistency layers as for example in [25], [21].

Due to the large sizes of the considered objects of interest, the prior $\mathbf{x}_{\mathrm{CNN}}$ is obtained by processing patches of the images. Training the network on patches or slices of the images further has the advantage of reducing the computational overhead while naturally enlarging the available training data and therefore being able to successfully train neural networks even with datasets coming from a relatively small number of subjects. Further, as demonstrated in [16], for the case of 2D radial MRI, one can also exploit the low topological complexity of 2D spatio-temporal slices for training the network u_θ. This allows to reduce the network complexity by using 2D- instead of 3D-convolutional layers and still exploiting spatio-temporal correlations and therefore to prevent overfitting. Note that the network architectures we are considering are CNNs and, since they mainly consist of convolutional and max-pooling layers, we can expect the networks to be translation-equivariant and therefore, patch-related artefacts arising from the re-composition of the processed overlapping patches are unlikely to occur in the CNN-prior. We have tested and evaluated our method on two examples of large-scale inverse problems given by 2D undersampled radial MRI and 3D low-dose CT. For both examples, our method outperformed the TV-minimization method and the dictionary learning-based method with respect to all reported quantitative measures. For the case of 2D undersampled radial cine MRI, using the CNN-prior as a regularizer in the subsequent iterative reconstruction increased the achieved image quality with respect to all reported measures, as can be taken from Table 1. For the CT example, due to the inherent presence of noise in the measured data, the quantitative measures of the final reconstruction are only similar to the ones obtained by post-processing the FBP-reconstruction. However, performing a few iterations to minimize functional (17), increased data-consistency of the obtained solution and resulted in a slight re-enhancement of the edges and gave back the CT images their characteristic texture. Future work to qualitatively assess the achieved image quality with respect to clinically relevant features, e.g. the visibility of coronary arteries for the assessment of coronary artery disease in cardiac CT, is already planned.

Using the CNN for obtaining a learning-based prior is faster by several orders of magnitude compared to the dictionary learning-based approach. This is because obtaining the prior with a CNN reduces to a forward pass of all patches, i.e. to multiplications of vectors with sparse matrices, where instead, the sparse coding of all patches involves the solution of an optimization problem for each patch. Further, the time needed for OMP is dependent on the sparsity level and the number of atoms of the dictionary, see [30]. In our comparison, for the 2D radial MRI example, the total reconstruction times of our proposed method and the DIC-based regularization method mainly differ in the step of obtaining the priors $\mathbf{x}_{\mathrm{DIC}}$ and $\mathbf{x}_{\mathrm{CNN}}$. Note that, in contrast to [33] and [6], in our comparison, the prior $\mathbf{x}_{\mathrm{DIC}}$ was only calculated once. In the original works, however, the proposed reconstruction algorithms use an alternating direction method of multipliers (ADMM) which alternates between first training the dictionary $\mathbf{D}$ and sparse coding with OMP and then updating the image estimate. Therefore, the realistic time needed to reconstruct the 2D cine MR images according to [34] and [6] is given by the product of the seven minutes needed for one sparse approximation and the number of iterations in the ADMM algorithm and the total time used for PCG for solving the obtained linear systems. Note that for the 3D low-dose CT example, even one patch-wise sparse approximation of the whole volume already takes about one hour and therefore, applying an ADMM type of reconstruction method is computationally prohibitive. Also, note that, even if the size of the image sequences for the MRI example is smaller than the one of the 3D CT volumes, the reconstruction of the 2D cine MR images takes relatively long compared to the CT example due to the fact that we use two different iterative methods (Landweber and PCG) for two different systems with different operators. Further, the number of iterations for the CT example is on purpose smaller than for the MR example, as the measurement data is noisy and early stopping of the iteration can already be thought of as a proper regularization method, see for example [29]. Also, the operators used for the CT examples were implemented by using the operators provided by the `ODL` library and are therefore optimized for performing calculations on the GPU. On the other hand, for the MRI example, we used our own implementation of a radial encoding operator $\mathbf{E}$ which could be further improved and accelerated.

Clearly, one difficulty of the proposed method is the difficulty which is shared by all iterative reconstruction schemes with regularization: the need to choose the hyper-parameter λ which controls the strength of the regularization compared to the data-fidelity term can highly affect the achieved image quality, especially when the data is contaminated by noise. In cascaded networks, the parameter λ can on the other hand be learned as well during training. Further, some other hyper-parameters as the number of iterations to minimize Tikhonov functional have to be chosen as well.

The proposed method is related to the one presented in [25], [21], [17] in the sense that steps 2 and 3 in Algorithm 1 are iterated in a cascaded network which represents the different iterations. However, in [25] and [21], the encoding operator is given by a Fourier transform sampled on a Cartesian grid and therefore is an isometry. Thus, assuming a single-coil data-acquistion, given $\mathbf{x}_{\mathrm{CNN}}$, the solution of (3) has a closed-form solution which is also fast and cheap to compute since it corresponds to performing a linear combination of the acquired k-space data and the one estimated from the CNN outputs and subsequently applying the inverse Fourier transform. In the case where the operator $\mathbf{A}$ is not an isometry, one usually needs to either solve a system of linear equations in order to obtain a solution which matches the measured data or, alternatively, rely on another formulation of the functional (3) which is suitable for more general, also non-orthogonal operators [17]. However, if the operator $\mathbf{A}$ and its adjoint $\mathbf{A}^H$ are computationally demanding to apply as in the case of radial multi-coil MRI, or if the objects of interest are high-dimensional, e.g. 3D volumes in low-dose CT, the construction of cascaded or iterative networks is prohibitive with nowadays available hardware. In contrast, in the proposed approach, since the regularization is separated from the data-consistency step, large-scale problems can be tackled as well. Hence, by decoupling the regularization from further iteration of the reconstruction, one can also choose to employ more complex and sophisticated neural networks to obtain the prior $\mathbf{x}_{\mathrm{CNN}}$ as it is typically the case for cascaded or iterative networks. For example, in [25] or [2], the CNNs were given by simple blocks of fully convolutional neural networks with residual connection. In contrast, in [17], the CNNs were replaced by more sophisticated U-nets [24], [14]. However, the examples in [17], [2] or [3] all use two-dimensional CT geometries, which do not correspond to the ones used in clinical practice. Therefore, particularly for large-scale inverse problems where the construction of iterative networks is infeasible, our method represents a valid alternative to obtain accurate reconstructions.

While in this work we used a relatively simple neural network architecture given by a plain U-net as in [14], further focus could be put on the choice of the network u_θ, also by using more sophisticated approaches, e.g. improved versions of the U-net [11] or generative adversarial networks for obtaining a more accurate prior to be further used in the proposed reconstruction scheme.

6 Conclusion

In this work, we have presented a general framework for the solution of high-dimensional ill-posed inverse problems in medical imaging. The reconstruction strategy consists in decoupling the regularization of the solution from ensuring data-consistency by solving the problem in three stages. First, an initial guess of the solution is obtained by the direct reconstruction from the measured data. As a second step, the initial solution is patch-wise processed by a previously trained CNN in order to obtain a prior which is then used in a Tikhonov-regularized functional to obtain the final reconstruction in a third step. The decoupling of the steps of obtaining a CNN-prior and minimizing a Tikhonov-functional allows to tackle large-scale problems. For both shown examples of 2D undersampled radial MRI and 3D low-dose CT, the proposed method outperformed the total variation-minimization method and the dictionary learning-based approach with respect to all reported quantitative measures. Since the reconstruction scheme is a general one, we expect the proposed method to be successfully applicable to other imaging modalities as well.

Acknowledgements

A. Kofler and M. Dewey acknowledge the support of the German Research Foundation (DFG), project number GRK2260, BIOQIC. We thank Dr. V. Wieske for providing clinically relevant cases for the 3D low-dose CT experiments.

References

[1] Jonas Adler, Holger Kohr, and Ozan Oktem. Operator discretization library. *https://github. com/odlgroup/odl*, 2017.

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

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

[4] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2018.

[5] Kai Tobias Block, Martin Uecker, and Jens Frahm. Undersampled radial mri with multiple coils. iterative image reconstruction using a total variation constraint. *Magnetic Resonance in Medicine*, 57(6):1086–1098, 2007.

[6] Jose Caballero, Anthony N Price, Daniel Rueckert, and Joseph V Hajnal. Dictionary learning and time sparsity for dynamic mr data reconstruction. *IEEE Transactions on Medical Imaging*, 33(4):979–994, 2014.

[7] Antonin Chambolle. Total variation minimization and a class of binary mrf models. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152. Springer, 2005.

[8] Marc Dewey, Elke Zimmermann, Florian Deissenrieder, Michael Laule, Hans-Peter Dübel, Peter Schlattmann, Fabian Knebel, Wolfgang Rutsch, and Bernd Hamm. Noninvasive coronary angiography by 320-row computed tomography with lower radiation exposure and maintained diagnostic accuracy: comparison of results with cardiac catheterization in a head-to-head pilot investigation. *Circulation*, 120(10):867–875, 2009.

[9] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

[10] Yoseob Han, Leonard Sunwoo, and Jong Chul Ye. k-space deep learning for accelerated MRI. *IEEE Transactions on Medical Imaging*, 2019.

[11] Yoseob Han and Jong Chul Ye. Framing U-Net via deep convolutional framelets: Application to sparse-view ct. *IEEE Transactions on Medical Imaging*, 37(6):1418–1429, 2018.

[12] Andreas Hauptmann, Simon Arridge, Felix Lucka, Vivek Muthurangu, and Jennifer A Steeden. Real-time cardiovascular mr with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease. *Magnetic resonance in medicine*, 81(2):1143–1156, 2019.

[13] Andreas Hauptmann, Felix Lucka, Marta Betcke, Nam Huynh, Jonas Adler, Ben Cox, Paul Beard, Sebastien Ourselin, and Simon Arridge. Model-based learning for accelerated, limited-view 3-d photoacoustic tomography. *IEEE Transactions on Medical Imaging*, 37(6):1382–1393, 2018.

[14] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. arxiv:1412.6980 Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[16] Andreas Kofler, Mark Dewey, Tobias Schaeffter, Christian Wald, and Christoph Kolbitsch. Spatio-temporal deep learning-based undersampling artefact reduction for 2d radial cine MRI with limited training data. *IEEE Transactions on Medical Imaging*, (DOI: 10.1109/TMI.2019.2930318), 2019.

[17] Andreas Kofler, Markus Haltmeier, Christoph Kolbitsch, Marc Kachelrieß, and Marc Dewey. A U-nets cascade for sparse view computed tomography. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[18] Housen Li, Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. NETT: Solving inverse problems with deep neural networks. *arXiv:1803.00092*, 2018.

[19] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72, 2008.

[20] Adriane E Napp, Robert Haase, Michael Laule, Georg M Schuetz, Matthias Rief, Henryk Dreger, Gudrun Feuchtner, Guy Friedrich, Miloslav Špaček, Vojtěch Suchánek, et al. Computed tomography versus invasive coronary angiography: design and methods of the pragmatic randomised multicentre discharge trial. *European radiology*, 27(7):2957–2968, 2017.

[21] Chen Qin, Jo Schlemper, Jose Caballero, Anthony N Price, Joseph V Hajnal, and Daniel Rueckert. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 38(1):280–290, 2018.

[22] Volker Rasche, Roland Proksa, R Sinkus, Peter Bornert, and Holger Eggers. Resampling of data between arbitrary grids using convolution interpolation. *IEEE Transactions on Medical Imaging*, 18(5):385–392, 1999.

[23] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 2018.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[25] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018.

[26] Karin Schnass. Convergence radius and sample complexity of itkm algorithms for dictionary learning. *Applied and Computational Harmonic Analysis*, 45(1):22–58, 2018.

[27] Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. Deep null space learning for inverse problems: convergence analysis and rates. *Inverse Problems*, 35(2):025008, 2019.

[28] David S Smith, Saikat Sengupta, Seth A Smith, and E Brian Welch. Trajectory optimized NUFFT: Faster non-cartesian mri reconstruction through prior knowledge and parallel architectures. *Magnetic Resonance in Medicine*, 81(3):2064–2071, 2019.

[29] Otto Neall Strand. Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4):798–825, 1974.

[30] Bob L Sturm and Mads Græsbøll Christensen. Comparison of orthogonal matching pursuit implementations. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 220–224. IEEE, 2012.

[31] Zhen Tian, Xun Jia, Kehong Yuan, Tinsu Pan, and Steve B Jiang. Low-dose ct reconstruction via edge-preserving total variation regularization. *Physics in Medicine & Biology*, 56(18):5949, 2011.

[32] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. on information theory*, 53(12):4655–4666, 2007.

[33] Yanhua Wang and Leslie Ying. Compressed sensing dynamic cardiac cine mri using learned spatiotemporal dictionary. *IEEE Transactions on Biomedical Engineering*, 61(4):1109–1120, 2014.

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[35] Stefanie Winkelmann, Tobias Schaeffter, Thomas Koehler, Holger Eggers, and Olaf Doessel. An optimal radial profile order based on the golden ratio for time-resolved mri. *IEEE Transactions on Medical Imaging*, 26(1):68–76, 2006.

[36] Qiong Xu, Hengyong Yu, Xuanqin Mou, Lei Zhang, Jiang Hsieh, and Ge Wang. Low-dose x-ray ct reconstruction via dictionary learning. *IEEE Transactions on Medical Imaging*, 31(9):1682–1697, 2012.

[37] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging*, 37(6):1348–1357, 2018.

[38] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.