Aus dem Institut für Radiologie
der Medizinischen Fakultät Charité - Universitätsmedizin Berlin

DISSERTATION

Deep Learning-based Methods for Image Reconstruction in Cardiac CT and Cardiac Cine MRI

Deep Learning-basierte Methoden zur Bildrekonstruktion in der Herz-CT und Herzfunktions-MRT

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité - Universitätsmedizin Berlin

von

Andreas Kofler
aus Meran, Italien

Datum der Promotion: 17.09.2021

Inhaltsverzeichnis

1 Abstrakt

Abstrakt (Deutsch)

Ziel: Nichtinvasive Bildgebungsverfahren wie die Computertomographie (CT) und die Magnetresonanztomographie (MRT) sind wesentliche Werkzeuge für die Diagnose von Herzerkrankungen wie z.B. koronare Herzkrankheiten oder Herzfunktionsstörungen. Die dazugehörigen Rekonstruktionsprobleme sind aus verschiedenen Gründen schlecht gestellt. Bei niedrigdosierten CT-Scans sind die Daten verrauscht, während diese bei der kardialen MRT unvollständig sind. Um diagnostische Bilder zu erhalten, werden Regularisierungsmethoden angewandt. In dieser Arbeit entwickeln und untersuchen wir verschiedene auf neuronalen Netzen (NN) basierende Methoden zur Bildrekonstruktion in der Herz-CT und Herz-Cine-MRT.

Methoden: Wir verwendeten verschiedene NN-basierte Methoden zur Rekonstruktion von niedrig dosierten CT- und unterabgetasteten MR-Bildern. Zuerst führten wir eine Parameterstudie mit iterativen NN durch. Basierend auf den Ergebnissen und Beobachtungen haben wir uns für die Entwicklung eines NN-basierten Ansatzes mit dem Namen XT,YT-Ansatz zur Reduktion von Unterabtastungsartefakten für die 2D-Radial-Herz-Cine-MRT entschieden. Der Ansatz basiert auf einem NN, das auf räumlich-zeitlichen *xt*- und *yt*-Schichten trainiert wird, die aus den MR-Bildern extrahiert werden können. Der XT,YT-Ansatz wurde dann in einer verallgemeinerten iterativen Rekonstruktionsmethode mit NN-Priors angewandt, die wir für 2D Cine-MRT und 3D Niedrigdosis-CT evaluiert haben.

Ergebnisse: Die vorgestellte XT,YT-Methode erzielte ähnliche oder bessere Ergebnisse als andere NN-basierte Methoden und schnitt besser ab als einige andere iterative Methoden. Das Training des NN im räumlich-zeitlichen Bereich hat mehrere Vorteile. Erstens ist es geeignet für das Training mit beschränkten Datensätzen. Zweitens bietet es die Möglichkeit, die Anzahl der Trainingsparameter zu reduzieren und somit eine Überanpassung des NN zu verhindern. Drittens ist das NN stabil bezüglich Rotation in der *xy*-Ebene. Viertens wird die räumlich-zeitliche Korrelation effizient genutzt, selbst wenn nur 2D Faltungsschichten verwendet werden. Das vorgestellte allgemeine Rekonstruktionsschema mit NN-Priors übertraf für niedrigdosierte 3D CT und unterabgetastete 2D Herz Cine MR zwei andere auf Totalvariationminimierung und gelernten Dictionaries basierende iterative Rekonstruktionsmethoden.

Schlussfolgerung: Obwohl iterative NN den Stand der Technik für Bildrekonstruktionsprobleme darstellen, ist ihre Anwendbarkeit derzeit auf relativ kleine Probleme beschränkt. Iterative Rekonstruktionsmethoden anhand NN-basierter Priors übertreffen empirisch Standardmethoden und haben das Potenzial, die Strahlendosisbelastung in der CT zu reduzieren und den Messprozess in der MRT zu beschleunigen.

Abstract (English)

Objective: Non-invasive medical imaging techniques such as computed tomography (CT) and magnetic resonance imaging (MRI) are nowadays essential tools for the assessment of cardiac diseases, e.g. coronary artery disease or cardiac dysfunction. The image reconstruction problems in these imaging modalities can be ill-posed for different reasons. For example, in low-dose CT, the measured data is noisy, while in accelerated cardiac MRI, undersampling in k-space leads to incomplete data. Thus, regularization methods must be applied to obtain images suitable for diagnostic purposes. In this thesis, we develop, investigate and evaluate different Neural Networks (NNs)-based methods for image reconstruction in cardiac CT and cardiac cine MRI.

Methods: We addressed the reconstruction of low-dose CT and accelerated MR-images using different NNs-based methods. We first performed an ablation study using iterative networks. Then, based on the obtained results and observations, we opted to develop a NNs-based approach, named XT,YT-approach, tailored to the reduction of undersampling artefacts for 2D radial cardiac cine MRI. The approach is based on a NN which is trained on the xt-and yt-spatio-temporal slices which can be extracted from the cine MR images. The XT,YT-approach was then applied to a generalized iterative image reconstruction framework using NN-image priors which we evaluated for 2D radial cine MRI and 3D low-dose CT.

Results: The presented XT,YT-method achieved competitive or better results compared to other NNs-based methods and outperformed several other iterative reconstruction methods. Training the NN in spatio-temporal domain has several advantages. First, it is suitable for training on limited datasets. Second, it offers the possibility to highly reduce the number of trainable parameters and therefore prevent the NN from overfitting. Third, the NN is naturally stable with respect to rotation in the xy-plane. Fourth, spatio-temporal correlation is efficiently exploited even by only using 2D convolutional layers. The proposed generalized reconstruction scheme using NN-priors was shown to outperform two other iterative reconstruction methods based on total variation-minimization and learned dictionaries for 3D low-dose CT and 2D radial cardiac cine MRI.

Conclusion: Although iterative neural network methods constitute the state-of-the-art for image reconstruction problems, their applicability is currently still limited to relatively small problems. Iterative reconstruction methods using NN-based image priors empirically outperform standard ones and have the potential to reduce the radiation dose exposure in CT and to accelerate the measurements process in MRI.

2 Manteltext

1 Introduction

Medical imaging techniques as computed tomography (CT) or magnetic resonance imaging (MRI) have become nowadays indispensable tools in the clinical routine for the diagnose of different diseases, e.g. cardiovascular diseases. For example, coronary CT has been reported to be the most accurate non-invasive imaging technique for coronary artery disease [32]. Cardiac Cine MRI can be applied for the assessment of the cardiac function as well as left and right ventricular volumes and left ventricular mass [24].

In every non-invasive imaging technique, the objective is to obtain a visual representation of the interior of the patient. This representation is typically obtained from a set of indirect measurements which are different for each imaging modality and depend on the underlying physics. Therefore, the reconstruction of such a representation corresponds to solving an inverse problem. In CT, for example, the measurements are given by a set of X-ray projections from different angles which are measured by a detector array. In MRI, the measurements correspond to the spatial frequency information of the image. Reconstructing an image from these measurements is a classical inverse problem which can be formulated by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z} \tag{1}$$

where $\mathbf{A} : X \rightarrow Y$ denotes a discrete and possibly non-linear forward operator between finite dimensional Hilbert spaces which models the data-acquisition process, $\mathbf{x}$ denotes the unknown image we want to recover, $\mathbf{y}$ denotes the measured data and $\mathbf{z}$ is a random vector which models the noise in the acquisition process. Inverse problems in medical image reconstruction can be ill-posed for different reasons. For example, in low-dose CT, where the energy of the emitted photons is reduced to limit the patient's radiation exposure, the measured data is corrupted by Poisson-distributed noise. In cardiac cine MRI, where the measurement process needs to be accelerated in order to be able to scan the patient during a single breathhold, the measurements are incomplete. Images directly reconstructed from the measured data, i.e. $\mathbf{x}_{\text{ini}} := \mathbf{A}^{\dagger}\mathbf{y}$, where $\mathbf{A}^{\dagger}$ denotes some reconstruction operator of $\mathbf{A}$, contain severe artefacts and/or noise. Therefore, regularization techniques must be used in order to obtain diagnostic images. A possible way of regularization is to solve a relaxed version of (1), by imposing the regularization in terms of a penalty term. Then, instead of solving (1), one minimizes a functional

$$F_{D,\mathcal{R},\lambda,\mathbf{y}}(\mathbf{x}) = D(\mathbf{A}\mathbf{x}, \mathbf{y}) + \lambda\,\mathcal{R}(\mathbf{x}) \rightarrow \min_{\mathbf{x}}, \tag{2}$$

where D denotes a data-discrepancy measure which is appropriately chosen according to the considered problem, $\mathcal{R}$ is a regularization term and $\lambda > 0$ controls the contribution of the regularization.

A possible choice of hand-crafted regularization is given by the L_1-norm of the image transformed by a so-called sparsifying transform. For example, using a first-order derivative operator $\mathbf{G}$ leads to the well-known total variation (TV)-minimization problem which has been widely considered in the literature [6], [34]. Further, approaches where the sparsifying transforms are learned from data, e.g. Dictionary Learning, have also been considered and applied for the task of image reconstruction in CT and MRI, see e.g. [35], [7], [38]. While regularization approaches such as TV-minimization come with solid and well-understood mathematical theory, a drawback of hand-crafted regularizations is that the properties which are imposed on the solution can be limited and may not reflect the nature of the images one wants to obtain. Using representations such as learned dictionaries circumvents this issue by providing transforms learned on the data. However, the underlying regularizing concept is still relatively simple by assuming that a signal is sparse with respect to a dictionary.

1.1 Neural Networks for Medical Image Reconstruction

Recently, Convolutional Neural Networks (CNNs) have been considered for the regularization of inverse problems, see e.g. [17], [2], [31], [40], [25], [11], [3], [11], [4], [12], [8]. Given the nowadays available computational power and relatively large amount of data, the idea is to parameterize a regularization by some pre-defined fixed mathematical operations, e.g. convolutions, and let the regularization be fully learned from data. In this section, we discuss several different approaches to use NNs for medical image reconstruction and highlight some advantages and limitations of the respective approaches by also outlining the work of the thesis.

Neural Networks as Post-Processing Methods

Initial works in the research area of inverse problems involving the use of NNs were mainly concerned with the design and the application of CNNs as post-processing methods, see e.g. [17]. Thereby, CNNs are used to denoise/remove artefacts from the initial reconstruction obtained from the measured data, i.e. to generate an image $\mathbf{x}_{\mathrm{CNN}} = f_\Theta(\mathbf{x}_{\mathrm{ini}})$ using some previously trained CNN f_Θ. By doing so, the learned reconstruction scheme takes the form

$$\mathbf{A}_\Theta^\dagger = f_\Theta \circ \mathbf{A}^\dagger. \tag{3}$$

However, solely post-processing the initial reconstruction $\mathbf{x}_{\mathrm{ini}}$ might result in an image which lacks data-consistency because the initially measured data $\mathbf{y}$ is only used once to obtain $\mathbf{x}_{\mathrm{ini}}$. Therefore, after having obtained $\mathbf{x}_{\mathrm{CNN}}$, a natural question to ask is how well $\mathbf{x}_{\mathrm{CNN}}$ matches the data in the measurements domain by considering $D(\mathbf{A}\mathbf{x}_{\mathrm{CNN}}, \mathbf{y})$. CNNs which correct the data $\mathbf{y}$ in raw-data domain, see e.g. [12], as well as methods which aim to invert the forward operator $\mathbf{A}$, see e.g. [40], have also been proposed but are not discussed here due to space limitations.

The most widely used types of CNNs are so-called *residual networks* which have the form

$$\tilde{f}_\Theta = \mathrm{id} + f_\Theta, \tag{4}$$

where id denotes the identity mapping and f_Θ a CNN as for example the U-net [29]. The name residual network reflects the fact that the CNN $\tilde{f}_\Theta$ learns the difference (i.e. the residual) between the input and the corresponding target label.

Iterative Reconstruction with Neural Networks-based Image Priors

In order to ensure/increase data-consistency of the output of the CNN, the estimate $\mathbf{x}_{\mathrm{CNN}}$ has to be corrected to better match the measured data $\mathbf{y}$. This is typically achieved by applying the forward operator $\mathbf{A}$ to $\mathbf{x}_{\mathrm{CNN}}$ and updating $\mathbf{x}_{\mathrm{CNN}}$ in raw-data space, either by imposing strict data-consistency [33], [16], which is possible only if $\mathcal{N}(\mathbf{A}) \neq \{\mathbf{0}\}$, where $\mathcal{N}(\mathbf{A})$ denotes the null-space of $\mathbf{A}$, or by minimizing a functional of the form (2), where $\mathbf{x}_{\mathrm{CNN}}$ is integrated in the penalty term $\mathcal{R}$. For example, the functional can be chosen as

$$F_{D,\mathbf{x}_{\mathrm{CNN}},\lambda,\mathbf{y}}(\mathbf{x}) = D(\mathbf{A}\mathbf{x},\mathbf{y}) + \lambda\, ||\mathbf{x} - \mathbf{x}_{\mathrm{CNN}}||_2^2, \tag{5}$$

which means that the regularization term is given by $\mathcal{R}(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}_{\mathrm{CNN}}||_2^2$. The procedure of removing artefacts/noise and correcting the estimate can be performed only once, see e.g. [16], [22], or repeated in an iterative manner [15], [8]. If performed only once, it leads to learned reconstruction schemes of the form

$$\mathbf{A}_\Theta^\dagger = f_{\mathrm{dc}} \circ f_\Theta \circ \mathbf{A}^\dagger, \tag{6}$$

where f_{dc} is a function which ensures/increases data-consistency
The specific form of f_{dc} depends on the considered problem. For example, in MRI, if $\mathbf{A}$ is a Fourier-operator which samples data on a Cartesian grid using a single-coil and $D(\mathbf{A}\mathbf{x},\mathbf{y}) = ||\mathbf{A}\mathbf{x} - \mathbf{y}||_2^2$, then f_{dc} computes the minimizer of (5) and has a closed form solution, see e.g. [31]. In the more general case, where $\mathbf{A}$ is not an isometry or $\mathcal{N}(\mathbf{A}) \neq \{\mathbf{0}\}$, f_{dc} can be given as any iterative scheme of finite length which is used to minimize (5), see e.g. [22].
If the procedure of applying the CNN to reduce artefacts or noise and ensuring/increasing data-consistency is iterated, one obtains reconstruction schemes of the form

$$\mathbf{A}_\Theta^\dagger = \big(f_{\mathrm{dc}} \circ f_\Theta\big) \circ \ldots \circ \big(f_{\mathrm{dc}} \circ f_\Theta\big) \circ \mathbf{A}^\dagger. \tag{7}$$

Using the same mapping f_Θ during the whole reconstruction requires the CNN f_Θ to be able to perform noise/artefacts reduction at different levels of noise/artefacts, see e.g. [9]. Using different networks $f_{\Theta_i}^{(i)}$ is also possible [26], [8]. However, the overall image recovery performance largely depends on the performance of each single $f_{\Theta_i}^{(i)}$. Depending on the considered problem, the training of all CNNs $f_{\Theta_i}^{(i)}$ can be computationally demanding for large-scale problems, as each CNN requires the generation of a new training dataset which involves the application of the physical models.

Iterative Neural Networks

A particularly interesting class of NNs-based algorithms for solving ill-posed inverse problems are so-called iterative networks, see e.g. [2], [31] [3], [11], where the forward and the adjoint operators $\mathbf{A}$ and $\mathbf{A}^{\mathsf{H}}$ can be represented as network layers and are integrated in the network. This means that the network f_Θ itself defines a proper end-to-end trainable reconstruction method as an unrolled iterative scheme of finite length, i.e. $f_\Theta = \mathbf{A}_\Theta^\dagger$ with

$$f_\Theta = \big(f_{\mathrm{dc}} \circ f_{\Theta_N}^N\big) \circ \ldots \circ \big(f_{\mathrm{dc}} \circ f_{\Theta_1}^1\big) \circ \mathbf{A}^\dagger. \tag{8}$$

Note that while (7) and (8) have the same form, the key-difference between (7) and (8) consists in the implementation and in the training process. While in (7), the networks $f_{\Theta_i}^{(i)}$ are trained consecutively, see e.g. [15], the network f_Θ in (8) is trained in an end-to-end manner, which means that all $f_{\Theta_i}^{(i)}$ are jointly trained, see e.g. [31], [25], [4].
Cascaded or iterative networks have been proposed and considered for different imaging modalities and define the state-of-the-art of CNNs-based methods for medical image

reconstruction. However, the main feature of these methods, i.e. the integration of the forward and adjoint operators in the CNN architecture, at the same time limits their application to realistic large-scale problems. Since the forward and the adjoint operators are integrated in the CNN architecture, the whole object which the operators are applied to must be processed by the CNN. For large-scale problems as for example 3D low-dose CT or non-Cartesian multi-coil MR acquisition protocols this can be challenging.

Contribution of this Thesis

In the following, we discuss the main work of the thesis and put it in the context of the previously described methods. The thesis is organized as follows. First, we briefly present and discuss the approach in [21] which served as a preliminary study and provided the basis for subsequent considerations on the development of the algorithms of choice needed to tackle the problem of image reconstruction in a realistic clinical setting. Then, we present a method for post-processing 2D radial cine MR images containing undersampling artefacts and discuss the main findings and results published in [20]. We finally present a generalized CNNs-based regularization method which can be used to solve arbitrary large-scale medical image reconstruction problems [22] and conclude the work with a short discussion and summary of the thesis.

2 Methods

2.1 A U-Nets Cascade for Sparse View Computed Tomography

This section is based on the following publication [21]:

- ♦ Andreas Kofler, Markus Haltmeier, Christoph Kolbitsch, Marc Kachelrieß, and Marc Dewey. *A U-nets Cascade for Sparse View Computed Tomography.* In International Workshop on Machine Learning for Medical Image Reconstruction, p. 91–99. Springer, 2018.

Problem Formulation and Proposed Network Architecture

In the following, we consider the problem of image reconstruction in sparse-view computed tomography. The inverse problem is given by

$$\mathbf{R}_I \mathbf{x} = \mathbf{y}_I, \tag{9}$$

where $\mathbf{x}$ denotes the unknown image to recover. The data-acquisition is modeled by $\mathbf{R}_I = \mathbf{S}_I \circ \mathbf{R}$, where the operator $\mathbf{R}$ defines the forward model given by a discrete Radon transform and $\mathbf{S}_I$ defines a binary mask which masks the measurements vector $\mathbf{y}$ at angular projections indexed by the indices in $I \subset J$, where $J = \{1, \ldots, d\}$ denotes the full set of projections. In [21], we have proposed a reconstruction algorithm based on iterative networks of the form (8), where the CNNs are given by U-nets.

A U-nets cascade u_Θ of length N with trainable parameters in the set Θ is given by

$$u_\Theta := f_{\mathrm{dc}}^N \circ u_{\Theta_N}^N \circ \ldots \circ f_{\mathrm{dc}}^1 \circ u_{\Theta_1}^1, \tag{10}$$

where the k-th data-consistency layer f_{dc}^k is given by

$$f_{\mathrm{dc}}^k(\mathbf{x}_{\mathrm{CNN}}^k, \mathbf{y}_I, \lambda_k) := \mathbf{R}^\dagger \big(\mathbf{\Lambda}_k \mathbf{R} \mathbf{x}_{\mathrm{CNN}}^k + \frac{\lambda_k}{1 + \lambda_k} \mathbf{y}_I \big), \tag{11}$$

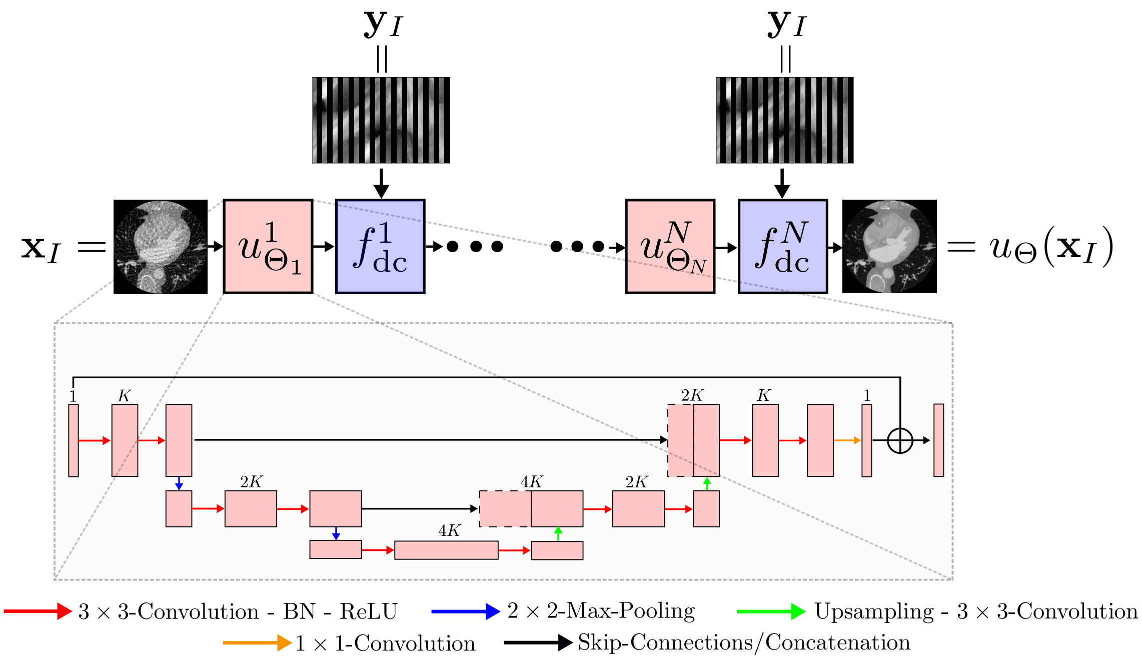


FIGURE 2.1 A U-nets cascade. The cascade u_Θ takes the initial reconstruction $\mathbf{x}_I$ as an input and alternates between applying a U-net and performing a data-consistency step. The first U-net is highlighted in order to show the different componenets in more detail.

and $\mathbf{R}^\dagger$ denotes the filtered back-projection. The parameter λ_k is treated as trainable parameter and can therefore vary for each k. Here, $\mathbf{\Lambda}_k = \mathrm{diag}(a_{1,k}, \ldots, a_{n,k})$ is a diagonal matrix of size $d \times d$ with entries $a_{i,k} = 1$ if $i \notin I$ and $a_{i,k} = 1/(1 + \lambda_k)$ otherwise.
The CNN can be trained end-to-end to minimize the L_2-error between its estimate $u_\Theta(\mathbf{x}_I)$ and the corresponding label. An example of a U-nets cascade is shown in Figure 2.1.

Experimental Set-Up and Dataset

In [21], we tested the proposed iterative CNN on a dataset of retrospectively generated sparse-view cardiac CT images of 52 patients taken from the study in [28]. The input images were obtained by performing a scan with a full set of $N_\theta = 512$ projections using a 2D parallel-beam scanner geometry. From the 512 projections, only 32 were used to obtain the initial reconstructions using the filtered back-projection (FBP), i.e. $\mathbf{x}_I = \mathbf{R}_I^\dagger$.
The implementation of the Radon-transform $\mathbf{R}_I$ and its FBP $\mathbf{R}_I^\dagger$ was done using the publicly available library `ODL` [1]. Training was carried out on 40 patients, while 6 patients were used for validation and the remaining 6 were used for testing.
We performed two different experiments. First, we analysed the effect of replacing of blocks of fully convolutional layers, as the ones used in [31], with U-nets [17]. Second, we performed an ablation study to investigate the effect of the length of the cascade on the achieved reconstruction by keeping the number of trainable parameters approximately the same for each cascade. To achieve this, we parametrized a U-nets Cascade by the following hyper-parameters:

- U - the number of U-nets used in the U-nets cascade,
- E - the number of encoding stages of each U-net,
- C - the number of convolutional layers per stage of each U-net,
- K - the number of initially applied filters of each U-net,

- F - the factor by which the number of used filters is increased after the max-pooling layers of each U-net.

This allowed us to construct cascades of different lengths in terms of number of alternations of U-nets and data-consistency layers but having approximately equally expressive network architectures in terms of trainable parameters.
For each experiment, we trained each model for 20 epochs by minimizing the L_2-norm between the CNN prediction and the corresponding target image. The performance of the networks was evaluated in terms of achieved peak signal-to-noise ratio (PSNR), normalized root mean squared error (NRMSE), structural similarity index measure (SSIM) [36] and Haar wavelet-based perceptual similarity index measure (HPSI) [27].

Results

Figure 2.2 shows results obtained with four different cascades of different lengths for $N = 1, 2, 3, 4$. For $N = 1$ the cascade only consists of one single U-net with no data-consistency layer. while for $N > 1$, the alternation is repeated N times. Interestingly, it can be seen that visually, for $N \geq 2$, the results look more appealing as the level of residual noise is lower than for $N = 1$. In particular, as it is highlighted by the yellow arrows, fine diagnostic image details as the right coronary artery are better visible in the cascades for $N \geq 2$ and in general, the edges seem to be sharper. However, the differences between $N = 2, 3, 4$ are negligible and barely visible. In terms of quantitative results, no method clearly surpasses the other, see Table 2.1. In terms of PSNR, NRMSE and SSIM, increasing N leads to poorer results, while HPSI stays approximately constant.

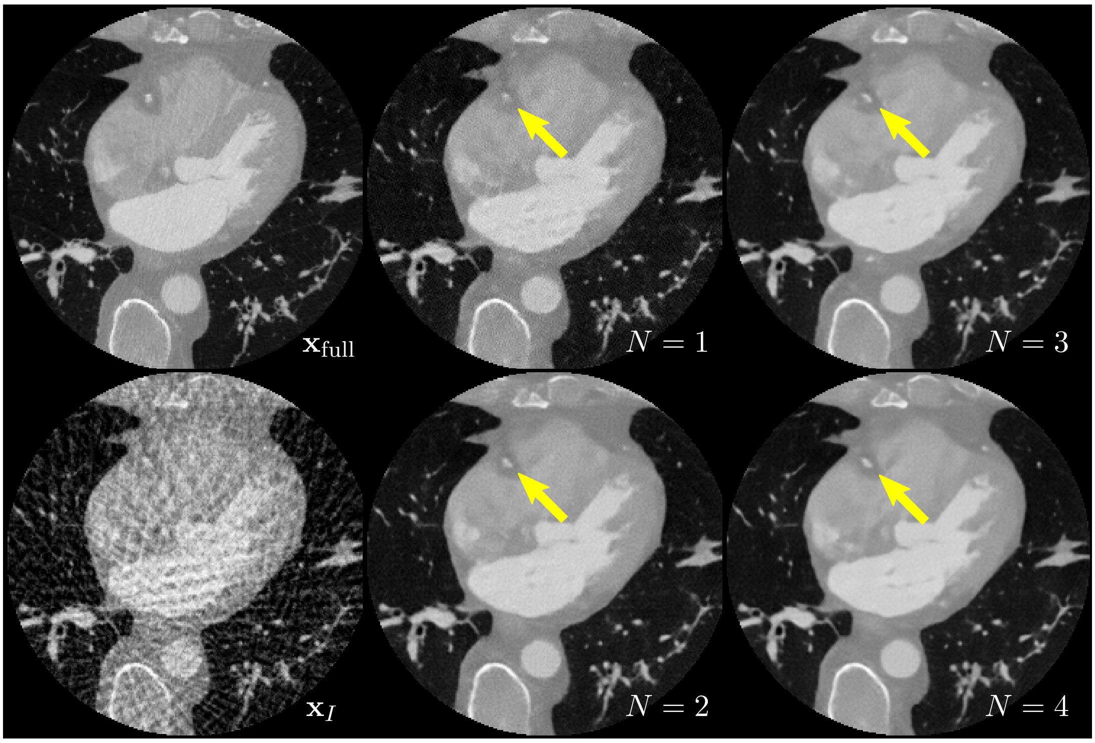


Figure 2.2 Results for the single U-net with no data-consistency layer ($N = 1$) and our proposed U-nets cascades with $2 \leq N \leq 4$. The U-nets cascades yield images with sharper edges and visually better preserve diagnostic image details. The yellow arrows point at the right coronary artery. Figure adapted from [21].

TABLE 2.1 Variation of the length of the U-nets cascade. The measures are averaged over the test set. Table adapted from [21].

Model	U1 E3 C2 K64 F2	U2 E3 C4 K32 F2	U3 E3 C3 K64	U4 E3 C2 K32 F2
PSNR	30.1920	31.1432	30.8460	30.3836
SSIM	0.9532	0.8905	0.8686	0.8559
HPSI	0.7304	0.7659	0.7732	0.7729
NRMSE	0.1832	0.1531	0.1621	0.1732
n_{params}	1 957 251	1 941 379	1 999 107	1 960 707

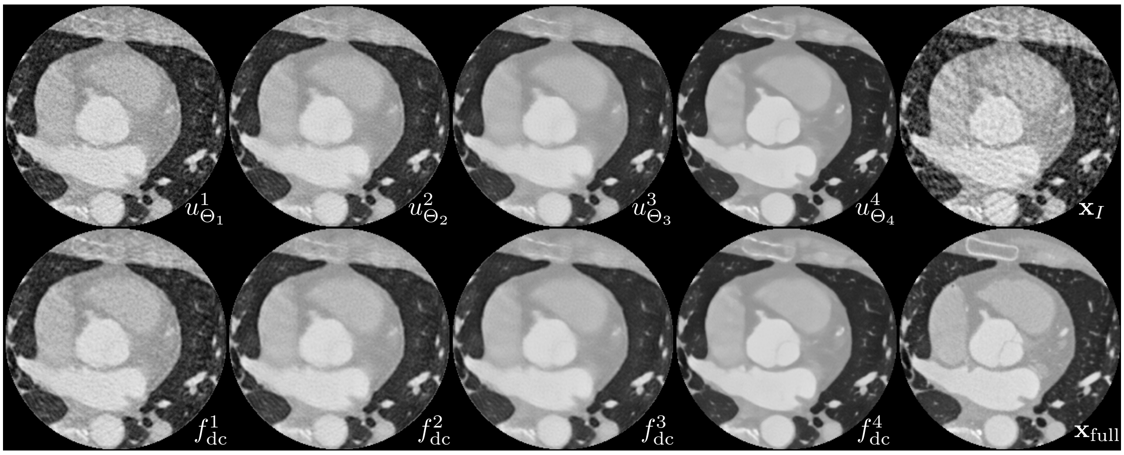


FIGURE 2.3 Output of a U-nets cascade of length $N = 4$ at the different stages. The output of the i-th U-net is denoted by $u^i_{\Theta_i}$, the output of the i-th data-consistency layer is denoted by f^i_{dc}. $\mathbf{x}_I$ and $\mathbf{x}_{\text{full}}$ show the input of the cascade and the ground truth, respectively. While the first three cascades seem to reduce the artefacts by smoothing, the last U-net re-enhances image contrast and sharpens edges.

Figure 2.3 shows the intermediate results obtained with a cascade of length $N = 4$. As we can see, the first three U-nets seem all to perform the task of removing the artefacts from the input mainly by smoothing. Interestingly, the fourth U-net is the one which re-enhances some of the edges and makes fine diagnostic details visible again. These observations indicate that some redundancy is present in the U-nets cascade. Further, the small difference between the cascades of different lengths suggests that for large-scale problems, where cascaded CNNs cannot be applied due to hardware constraints, using one single CNN to generate an image prior to be used in a subsequent iterative reconstruction might suffice.

2.2 Spatio-Temporal Artefacts Reduction in Accelerated 2D Radial Cine MRI

This section is based on the following publication [20]:

- Andreas Kofler, Marc 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.* In IEEE Transactions on Medical Imaging, 39(3), p. 703–717, 2020.

Based on the observations stated in Section 2.1, we now focus on the design of a Deep Learning-based method for obtaining an image prior which can be used within the regularization of a subsequent iterative reconstruction.

Problem Formulation and Motivation

We consider the problem

$$\mathbf{E}_I \mathbf{x} + \mathbf{e} = \mathbf{y}_I, \tag{12}$$

where $\mathbf{x} \in \mathbb{C}^{N_x \times N_y \times N_t}$ denotes the unknown 2D cine MR image, $\mathbf{E}_I$ denotes a Fourier-encoding operator which for each time point $k_t \in \{1, \ldots, N_t\}$ samples the k-space data along radial trajectories indexed by a set I_t with $\cup_{t=1}^{N_t} I_t = I \subset J$, where J denotes the "full" set of radial lines needed to satisfy the Nyquist limit. The vector $\mathbf{y}_I \in \mathbb{C}^{m_{\mathrm{rad}}}$ denotes the measured k-space data and $\mathbf{e}$ denotes a random noise vector. The radial trajectories are chosen according to the golden-angle method [37].
As already outlined, the idea is to tackle the image reconstruction problem for realistic large-scale problems by means of an iterative reconstruction using a CNN-generated image prior as regularization. Thereby, the image prior should be as close as possible to the (unknown) ground truth image for which the measured data $\mathbf{y}_I$ is given. Put in other words, the network f_Θ should be robust and reliably provide a "good" image prior $\mathbf{x}_{\mathrm{CNN}}$ to be further used in the reconstruction. Therefore, increasing the number of trainable parameters of the CNN to a maximum before experiencing overfitting seems to be a viable strategy for obtaining such a prior. In fact, increasing the model's capacity as well as applying more sophisticated approaches, e.g. using generative adversarial networks, is common practice for the design of CNN architectures as post-processing methods. However, in medical imaging applications, it is rarely the case to have access to a large number of training samples. For example, in cardiac cine MRI, obtaining ground truth data is challenging due to the fact that patients are scanned during a single breathhold which is difficult for patients with limited breathing capabilities. Training a CNN on a dataset only consisting of healthy volunteers is a possible option but does not give insights in the applicability of the method for diagnostic purposes.

Intuition of the Proposed Method

In order to be independent of a large amount of data for training, we adopt an approach which is based on a change of perspective on the data. Given problem (12), the object of interest $\mathbf{x}$ is a sequence of $2D$ images which vary over time and show the cardiac movement. Therefore, the nature of the problem offers the possibility to exploit the high correlation among adjacent image frames and has been considered in different methods in the literature, for example using spatio-temporal total variation-based constraints [6], regularization methods based on spatio-temporal dictionary learning [7], [35] or spatio-temporal CNNs [31], [30], [25], [14].

The main idea behind our method is to decompose an undersampled 2D cine MR image with artefacts into spatio-temporal slices and to train a CNN to map them to their corresponding ground-truth spatio-temporal slices. The immediate advantage is to have direct access to a large number of training samples since for each 2D cine MR image one can extract N_x yt-slices and N_y xt-slices. Figure 2.4 shows different CNN-based approaches for mapping an initial reconstruction $\mathbf{x}_I$ to its corresponding ground-truth image. Figure 2.4 A, Figure 2.4 B and Figure 2.4 C show the approaches presented in [17], [30] and [14], respectively, while Figure 2.4 D illustrates our proposed XT,YT approach [20].

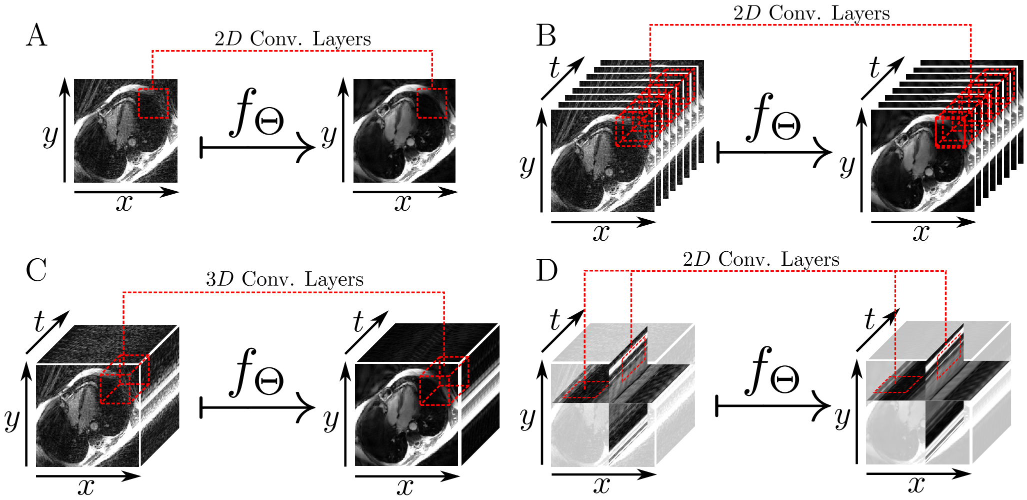


FIGURE 2.4 Different $2D$ and $3D$ CNNs-based approaches for post-processing images with undersampling artefacts. A: 2D CNN using a frame by frame approach [17], B: 2D CNN using an image sequence to image-sequence approach where the cardiac phases are aligned as channels [30], C: $3D$ CNN using an image sequence to image sequence mapping with three-dimensional convolutional kernels [14], D: our proposed method using 2D spatio-temporal slices [20]. Figure taken from [20].

The intuition behind choosing the just described decomposition of a 2D cine MR image is to facilitate the learning of a mapping between an undersampled image and a ground truth image. The set of spatio-temporal xt- and yt-slices consists of images containing mostly horizontal lines except for regions where the cardiac movement is visible. Furthermore, the content of these spatio-temporal slices is rather independent of the specific subject one is considering. Therefore, intuitively speaking, samples across different subjects are more indistinguishable from the xt- and yt-perspective compared to the xy-perspective and a small number of subjects is expected to already contain the needed information to allow for the proper training of a CNN.

To show that the set of 2D spatio-temporal slices has indeed a simpler structure than the 2D image frames, we analysed our dataset using Persistent Homology analysis, a tool for assessing the topological complexity of datasets. The results of the analysis confirmed that the manifold of the ground-truth spatio-temporal slices in xt- and yt-direction has a lower topological complexity than the manifold of the ground truth images in the xy-plane. Further, for both manifolds in the xt, yt- and xy-domain, the ground-truth images have a lower topological complexity than their corresponding residual manifolds. For the interested reader, we refer to [20] for more details on the persistent homology analysis.

Based on these results, the CNN used in the XT,YT approach is constructed in such a way that the network learns to map the spatio-temporal slices of the initial NUFFT-reconstructions to their corresponding ground-truth spatio-temporal slices. For the XT,YT approach, we adopted a U-net as in [17] which has the usual form

$$\tilde{u}_\Theta = \mathrm{id} + u_\Theta. \tag{13}$$

Since (13) is a residual network, only u_Θ contains trainable parameters. Therefore, pushing $\tilde{u}_\Theta$ to learn the ground-truth image manifold is achieved by using the residuals $\mathbf{r}_I := \mathbf{x}_I - \mathbf{x}_\mathrm{f}$ as target labels instead of the ground truth images $\mathbf{x}_\mathrm{f}$. By doing so, the CNN learns the manifold of the ground-truth images up to a change of sign. If not otherwise stated, the used U-net was the one parameterized by $E3\,C4\,K64$ following the notation introduced in Subsection 2.1. For more details, please see [20].

Experimental Set-up

For the following experiments, we used a dataset of different 2D cine MR images of $n = 19$ subjects (4 patients + 15 healthy volunteers). The images were obtained with a bSSFP sequence on a 1.5 T MR scanner during a 10 s breathhold. For all healthy volunteers and for two patients, $N_z = 12$ different orientations of 2D cine MR images were used, while for two patients, only $N_z = 6$ slices were available due to limited breathing capabilities. Therefore, the dataset $\mathcal{D}$ consisted of 216 2D cine MR images. The images which served as target ground-truth images were obtained by reconstructing the k-space data acquired along $N_\theta = 3400$ radial lines using kt-SENSE [18]. Each cine image has a shape of $N_x \times N_y \times N_t = 320 \times 320 \times 30$ with an in-plane resolution of 2 mm and a slice thickness of 8 mm. The input images for the CNN were given by the direct reconstruction from the measured data using a non-uniform inverse Fourier transform (NUFFT). Note that sampling the k-space data along 3400 radial spokes already corresponds to an acceleration factor of approximately ~ 3 which is needed to perform the scan during a single breathhold. Therefore, only acquiring 1130 radial trajectories corresponds to an acceleration factor of approximately ~ 9 and reduces the required scan time to 3-4 seconds. We split the dataset $\mathcal{D}$ in 12/3/4 subjects used for training, validation and testing. All experiments were performed using a 4-fold cross validation. This means the reported performance of the networks corresponds to the average performance over the different folds. For one fold, we used only patients' images as test data in order to have the possibility to investigate clinically relevant features. All the images shown in this section are images of patients.

In [20], we performed the following experiments:

1. *Training with limited data*: To show that the proposed XT,YT is applicable when only limited training data is available, we trained the network on different datasets where we restricted the number of subjects n whose images were included in the training dataset. We trained on $n = 1, 2, 4, 8, 12$ subjects and tested the network on the remaining 4 subjects and compared our proposed approach to the spatially trained U-net [17].

2. *Comparison with other CNNs*: We compared our proposed XT,YT method to the spatially trained 2D U-net [17], the spatio-temporal 2D U-net [30] and the computationally heavier 3D U-net [14] which we abbreviate by XY, XY,T and XYT, respectively.

3. *Comparison with other iterative reconstruction methods:* In this experiment, we compared our method to kt-FOCUSS [18], kt-SENSE [10], a TV-minimization approach

[6] and a dictionary learning + TV-based approach [35], [7].

4. *Comparison with state-of-the-art cascaded networks:* For the sake of completeness, we further compared our method to the two cascaded networks [25] and [31].

5. *Deep vs more shallow CNNs*: Here, we showed that the change of perspective on the data allows our XT,YT method to highly reduce the number of trainable parameters needed by the CNN. The performance of deeper U-nets is compared to the one of more shallow CNNs.

6. *Rotation equivariance*: Here, we showed that our method naturally achieves the property of being stable with respect to image rotations we and compared it to XY. For the experiment, we tested the pre-trained CNNs on images of the training set which were rotated by the angles ±66°, ±33°, ±90° and 180°. By doing so, we were able to evaluate the effect of the rotations on the performance of the CNNs.

The XT,YT CNN was trained by minimizing the L_2-error between input and the target label using stochastic gradient descent with a linearly decreasing learning rate from 10^{-6} to 10^{-8} and a mini-batch of 44. All CNNs were trained by performing $5 \cdot 10^4$ back-propagations.
Again, the performance of the methods was evaluated in terms of average PSNR, NRMSE, SIMM and HPSI achieved on the test set.

Main Results

1) Training with limited amount of data:
In this experiment, the XY and the proposed XT,YT method were compared in terms of achieved performance when trained on a restricted dataset. First, while observing the training and the validation error-curves in Figure 2.5, we see that successfully training with XY is only possible for $n = 8, 12$, as for lower n overfitting occurs immediately. In contrast, using our XT,YT method, even for $n = 1$, no overfitting is visible. Also, the training and validation error-curves are comparable among all n which shows that the network is properly trainable even on one single subject. Further, a small validation error is already achieved at early stages of training which suggests that removing artefacts from the *xt*- and *yt*-perspective is a particularly easily learnable task for the CNN. Note that the CNNs used in the experiments for XY and XT,YT have exactly the same number of trainable parameters but the training and validation error trends clearly differ from each other. This also suggests that the mapping to be learned by the CNN in the XT,YT perspective is particularly simple and is consistent with the results of the persistent homology analysis [20]. Figure 2.6 shows results obtained with our proposed XT,YT approach (A-D), the 2D frame-to-frame approach [17] trained on 12 subjects (E) and the ground truth *kt*-SENSE reconstruction with $N_\theta = 3400$ radial lines (F). As we can see, the obtained results are comparable for all n. This suggests that using the proposed XT,YT approach, the network is already able to properly generalize when being trained on only one single subject. Further, we see that the network using the XT,YT approach already visually outperforms the spatially trained U-net [17] even when trained on only one subject.
Table 2.2 lists the average measures achieved by the network on the testset when trained on a dataset only including n subjects. We see that even in terms of quantitative measures, the achieved performance is comparable among all experiments with different n.
Note that the curves shown in 2.5 as well as the results obtained in Figure 2.6 and Table 2.2 correspond the case that the CNNs are trained to learn the ground-truth image manifolds

as described above. We also performed the experiment for the case of residual learning and obtained similar results. However, for the sake of conciseness, we omit the results and their discussion here and refer the interested reader to [20].

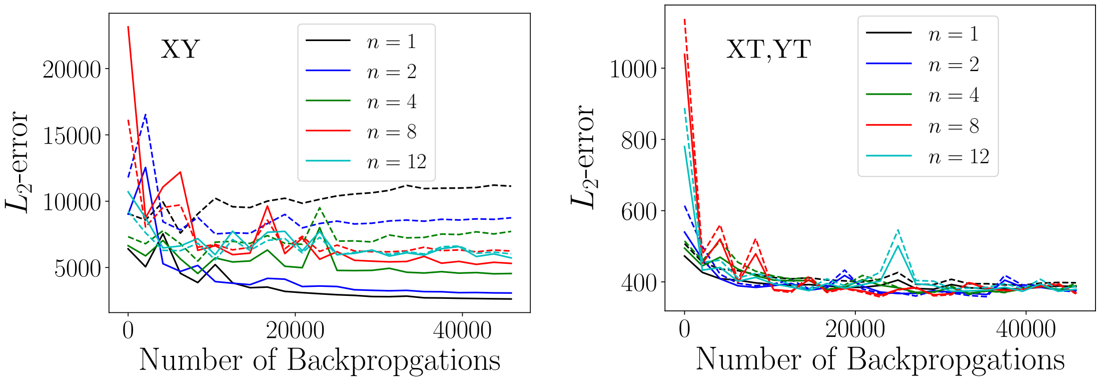


FIGURE 2.5 Loss behavior during training of the XY- and our proposed XT,YT-approach. Observing the training losses (solid lines) and validation losses (dashed lines) for different n, we see that overfitting is visible for XY for $n = 1, 2, 4$. For our XT,YT method, almost no gap is visible between the training and the validation error for all n which indicates that the network is able to properly generalize even when trained on only one subject. Figure adapted from [20].

TABLE 2.2 Estimated images and their corresponding point-wise errors when the number of subjects whose images were included in the training set is varied. Table taken from [20].

	$n = 1$	$n = 2$	$n = 4$	$n = 8$	$n = 12$
PSNR	37.25	37.79	37.66	37.84	37.83
SSIM	0.93	0.93	0.93	0.93	0.93
HPSI	0.99	0.99	0.99	0.99	0.99
NRMSE	0.11	0.11	0.11	0.11	0.11

2) Comparison with other CNNs:

Here, we compare our proposed XT,YT approach to other CNNs-based post-processing methods. More precisely, we implemented the approaches illustrated in Figure 2.4. Figure 2.7 shows results obtained with the methods XY [17] , XY,T [30], XYT [14] and our proposed XT,YT method [20]. We see that the method which delivered the poorest results is the spatio-temporal U-net XY,T, followed by the spatially trained U-net XY. The reason for the poor performance of XY,T most probably lies in the fact that the network is forced to learn the residual as proposed in the original work [30] for which the considered input images were zero-filled reconstructions opposed to our NUFFT-reconstructions. Forcing XY,T to learn the ground-truth images turned out to properly reduce the artefacts but also smoothing the cardiac movement. The method XY achieved decent results, however it does not exploit any temporal correlation of adjacent frames and therefore resulted in a less accurate reduction of the artefacts. The 3D approach XYT and our approach XT,YT performed comparably well. However, our method has considerably fewer parameters. Further, due to memory limits, for XYT, one has to decompose the initial reconstruction into patches, apply the 3D U-net to the patches and properly re-assemble the processed

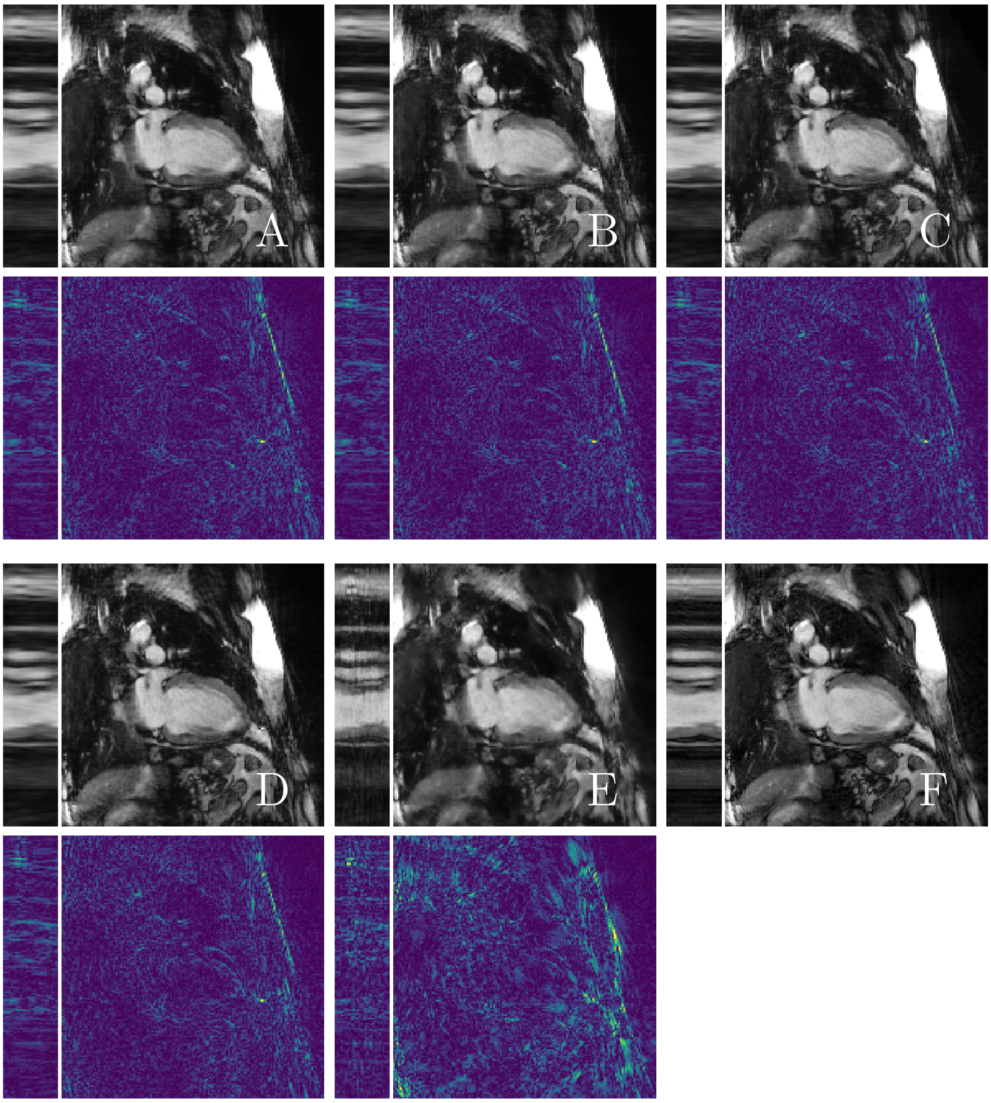


Figure 2.6 Results obtained by XT,YT and XY when the number of subjects in the training data is varied and the corresponding point-wise error-images. Note that no data-augmentation was used. Proposed XT,YT method for $n = 1$ (A), $n = 2$ (B), $n = 8$ (C), $n = 12$ (D), the spatial U-net for $n = 12$ (E) and the *kt*-SENSE reconstruction with 3400 radial lines (F). The point-wise error images were magnified by a factor of ×3. Figure adapted from [20].

patches to obtain the estimated output. Using our XT,YT approach, the initial reconstruction can be processed in approximately 1.2 s even if the CNN has to process $N_x + N_y$ spatio-temporal slices. Table 2.3 shows the average performance of the just described CNNs in terms of PSNR, NRMSE, SSIM and HPSI which also quantitatively reflects the visually deduced performances. Note that our method XT,YT yielded results comparable to the ones obtained with XYT even when trained on only $n = 1$ subject.

3) and 4) Comparison with iterative reconstruction methods and *Comparison with state-of-the-art cascaded networks:*
We further compared our proposed method to *kt*-FOCUSS [18], *kt*-SENSE [10], a spatio-temporal TV-minimization approach [6], a dictionary learning + TV-based regularization method [7] and two cascaded networks [31], [25]. Our proposed approach outperformed

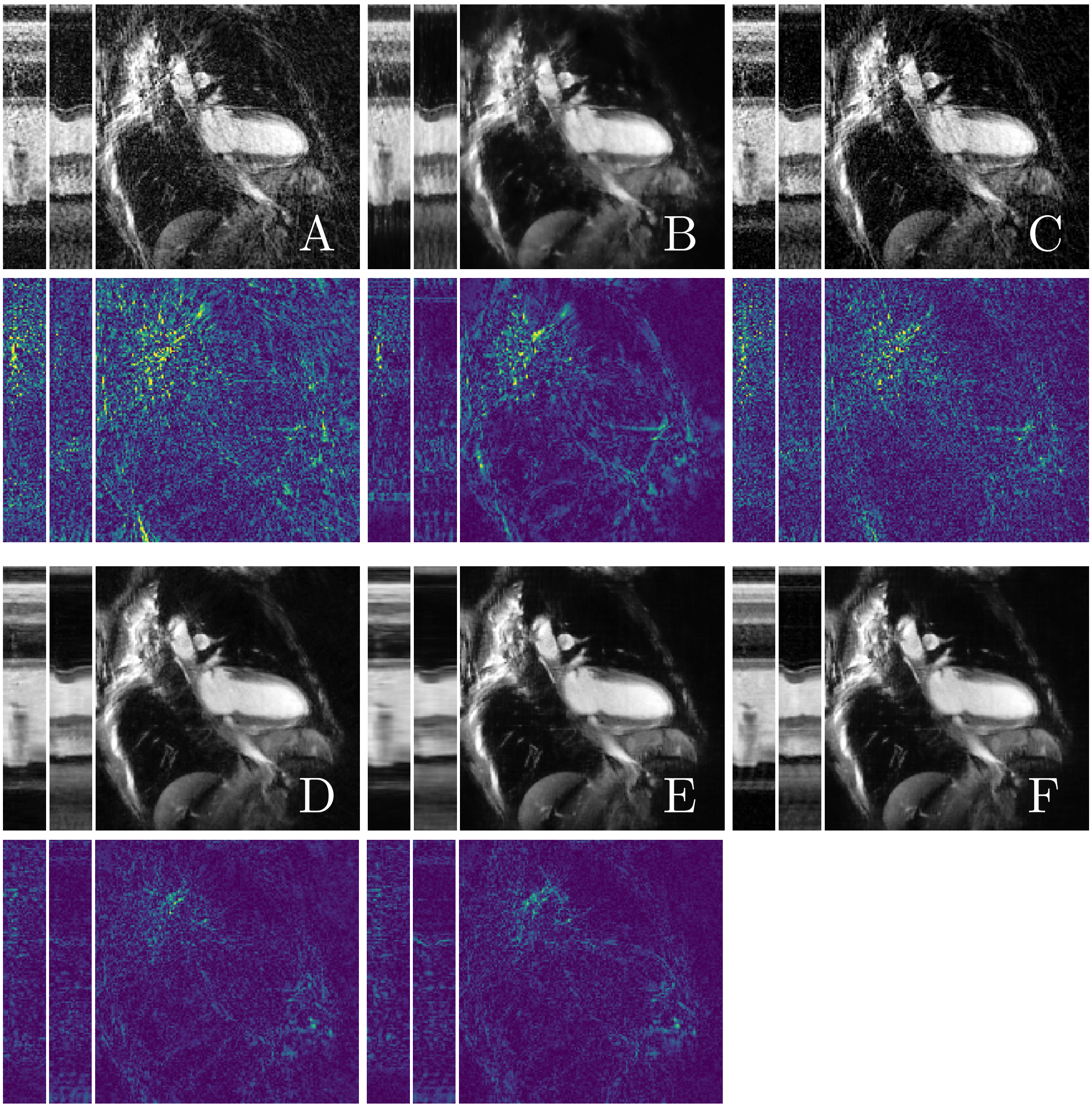


Figure 2.7 Results and point-wise error-images for the comparison with different Deep Learning-based post-processing methods. A: NUFFT reconstruction with $N_\theta = 1130$ radial lines, B: 2D spatial U-net [17], C: 2D spatio-temporal U-net [30], D: 3D spatio-temporal U-net [14], E: proposed approach 2D XT,YT spatio-temporal U-net [20], F: ground truth *kt*-SENSE reconstruction. The point-wise error images were magnified by a factor of ×3. Figure adapted from [20].

Table 2.3 Comparison of different Deep Learning-based post-processing approaches. Table taken from [20].

NN Model	**XY**	**XY,T**	**XYT**	**XT,YT**
PSNR	34.82	33.53	37.83	37.93
SSIM	0.91	0.87	0.94	0.93
HPSI	0.99	0.98	0.99	0.99
NRMSE	0.14	0.17	0.11	0.10

all non-CNN-based methods and achieved competitive results compared to the cascaded CNNs, see [20].

5) Deep vs more shallow CNNs:
Since the main idea of the work was to exploit the lower topological complexity of the spatio-temporal slices which can be extracted from the cine images, we investigated if simpler CNNs (in terms of fewer layers and a lower number of trainable parameters) were able to reduce the undersampling artefacts comparably well. For this purpose, we trained different CNNs according to our XT,YT approach. More precisely, by following the parametrization of a U-net as described in Subsection 2.1, we further trained and tested the networks E1 C8 K64 E4 C4 K64 and E5 C2 K64.

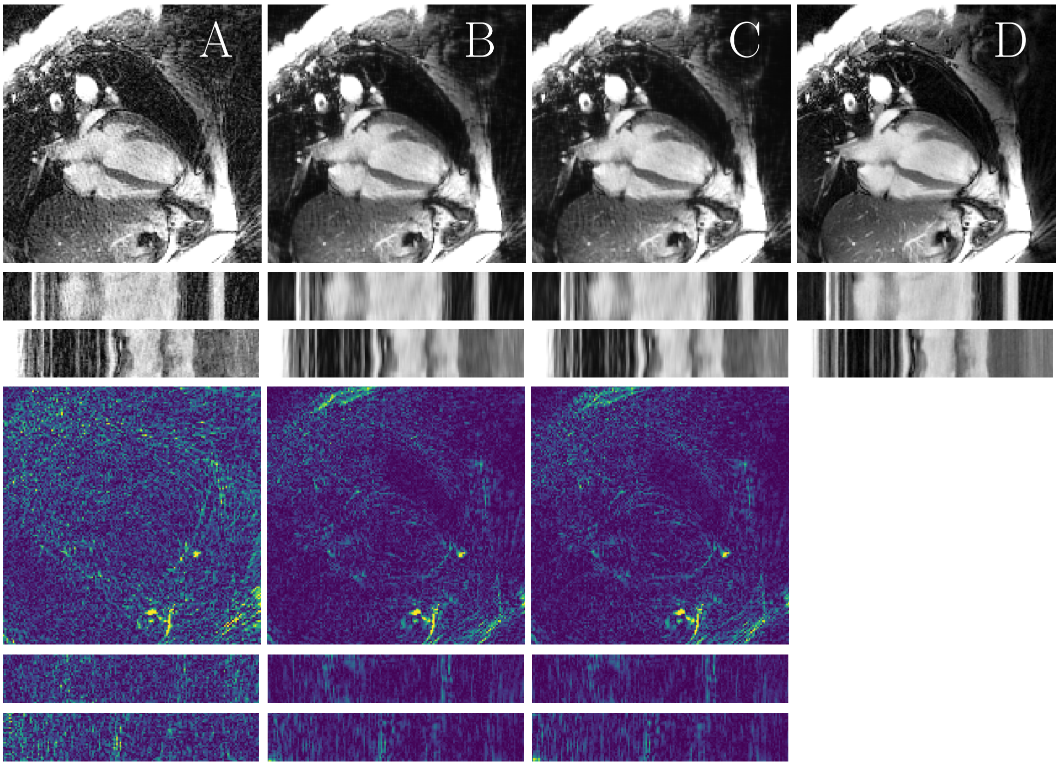


FIGURE 2.8 Results and point-wise error-images obtained with different CNNs of different depth following our proposed XT,YT aproach. A: initial NUFFT-reconstruction, B: E1 C8 K64, C: E5 C2 K64 and D: *kt*-SENSE reconstruction with $N_\theta = 3400$ radial lines. We see that the network E1 C8 K64, which only consists of 8 convolutional layers, performed equally well compared to E5 C2 K64 which consists of 18 convolutional layers. Figure adapted from [20].

Figure 2.8 shows an example of results obtained with the just mentioned networks and the comparision with the ground truth image given by the *kt*-SENSE reconstruction using $N_\theta = 3400$ radial lines. As we can see, the three networks visually performed comparably well. In particular, the network E1 C8 K64 only consists of eight convolutional layers but still achieved similar results as E5 C2 K64 which in contrast has 18 convolutional layers. The network E1 C8 K64 contains $n_{\text{params}} = 223\,492$ trainable parameters, while for E5 C2 K64, $n_{\text{params}} = 25\,087\,168$. This suggests that the change to the XT,YT perspective indeed leads to a simplification of the artefacts-reduction problem and facilitates the learning of the CNN. By being able to use more shallow CNNs and still achieve good performance, the phenomenon of overfitting can be efficiently prevented.

6) Rotation equivariance: Finally, our XT,YT approach was shown to be naturally robust with respect to rotation. The results are omitted here due to space limits. We refer the interested reader to [20].

2.3 Neural Networks-based Regularization of Large-Scale Problems in Medical Image Reconstruction

The following content is based on the following publication [22]:

- Andreas Kofler, Markus Haltmeier, Tobias Schaeffter, Marc Kachelrieß, Marc Dewey, Christian Wald, and Christoph Kolbitsch. *Neural Networks-Based Regularization for Large-Scale Medical Image Reconstruction.* In Physics in Medicine & Biology, 65(13):135003, 2020.

Here, we consider ill-posed large-scale inverse problems of the form (1) which we aim to solve via a Tikhonov-regularization by minimizing (5). In the following, we assume to have access to a properly trained CNN. The CNN is used to obtain an appropriate estimate of the unknown ground-truth image and is used as an image-prior when solving (5) with an iterative method.

Proposed Approach

Given an initial reconstruction $\mathbf{x}_{\mathrm{ini}} = \mathbf{A}^{\dagger}\mathbf{y}$, for large-scale problems, we obtain the image prior $\mathbf{x}_{\mathrm{CNN}}$ using a composite-function which decomposes the initial reconstruction $\mathbf{x}_{\mathrm{ini}}$ into patches/slices, processes them with the CNN and recomposes the patches to obtain the image prior, i.e. $\mathbf{x}_{\mathrm{CNN}} \coloneqq f_{\Theta}(\mathbf{x}_{\mathrm{ini}})$. The function f_{Θ} is needed because of the large scale of the considered problems where the complete object cannot be processed at once. Consider an image/volume $\mathbf{x}$ and its decomposition in $N_{\mathbf{p},\mathbf{s}}$ (in general overlapping) patches

$$\mathbf{x} = \mathbf{W}_{\mathbf{p},\mathbf{s}} \sum_{j=1}^{N_{\mathbf{p},\mathbf{s}}} (\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^{\mathsf{T}} \, \mathbf{R}_j^{\mathbf{p},\mathbf{s}} \, \mathbf{x}, \tag{14}$$

where the operators $\mathbf{R}_j^{\mathbf{p},\mathbf{s}}$ and $(\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^{\mathsf{T}}$ extract and re-position the patches at the original position, respectively. The diagonal operator $\mathbf{W}_{\mathbf{p},\mathbf{s}}$ accounts for proper weighting of overlapping regions. The tuples $\mathbf{p}$ and $\mathbf{s}$ define the size of the used patches and strides in each dimension and therefore implicitly determine the number of patches $N_{\mathbf{p},\mathbf{s}}$ which are extracted from a single image or volume.
Using the introduced notation, the CNN-image prior is obtained using f_{Θ} by

$$\mathbf{x}_{\mathrm{CNN}} \coloneqq f_{\Theta}(\mathbf{x}_{\mathrm{ini}}) = \mathbf{W}_{\mathbf{p},\mathbf{s}} \sum_{j}^{N_{\mathbf{p},\mathbf{s}}} (\mathbf{R}_j^{\mathbf{p},\mathbf{s}})^{\mathsf{T}} (u_{\Theta}(\mathbf{R}_j^{\mathbf{p},\mathbf{s}}(\mathbf{x}_{\mathrm{ini}}))). \tag{15}$$

Since $\mathbf{x}_{\mathrm{CNN}}$ is fixed, the subsequent minimization of (5) is independent of the CNN and can be achieved by means of any classical iterative scheme. The choice of the iterative scheme depends on the specific application and on the data-discrepancy term used in (5). For example, in MRI, a typical choice for the data-discrepancy measure D in (5) is the L_2-norm and therefore, the pre-conditioned conjugate gradient (PCG) method is a suitable iterative method. In low-dose CT, the Kullback Leibler-divergence is usually used as data-discrepancy measure as it corresponds to the log-likelihood for Poisson-distributed noise in the measured data and therefore, a simple Landweber iteration can be used.

Finally, for ill-posed large-scale image reconstruction problems, one can use a three-steps reconstruction scheme of the form described in Algorithm 1.

Algorithm 1 Proposed CNNs-based large-scale image reconstruction algorithm.

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

In [22], we tested the applicability of the proposed three-steps reconstruction scheme for 2D radial cine MRI and 3D low-dose CT and compared the results to the well-known TV-minimization and dictionary learning-based approaches [6], [7], [35].

Experimental Set-Up for 2D Radial Cine MRI

Here, we tested the reconstruction scheme in Algorithm 1 for 2D radial cine MRI. The used image dataset is the same as in Section 2.2. For these experiments, we retrospectively generated undersampled radial k-space data with a radial encoding operator $\mathbf{E}_I$ using $N_\theta = 1130$ spokes. From the generated k-space data, we reconstructed initial reconstructions using the adjoint NUFFT-operator $\mathbf{E}_I^{\mathsf{H}}$. The NUFFT-reconstructions and the ground truth images were used as training dataset for a CNN which served to generate an image prior for a subsequent iterative reconstruction. We trained the CNN according to our proposed XT,YT-approach [20]. Since in [20], the CNN was trained only on magnitude images as a post-processing method, we extended the XT,YT approach to be applicable on complex-valued images. For this, we trained one real-valued CNN to map the real and imaginary parts of the xt- and yt-spatio-temporal slices of the NUFFT-reconstruction $\mathbf{x}_I$ to their corresponding ground-truth spatio-temporal slices. Therefore, the CNN image prior was obtained by

$$\begin{aligned}\mathbf{x}_{\mathrm{CNN}} &= f_\Theta(\mathbf{x}_I) \qquad (16)\\ &= \frac{1}{2}\Big[\sum_j (\mathbf{R}_j^{xt})^{\mathsf{T}}\big(u_\Theta(\mathbf{R}_j^{xt}(\mathrm{Re}\,\mathbf{x}_I))\big) + (\mathbf{R}_j^{yt})^{\mathsf{T}}\big(u_\Theta(\mathbf{R}_j^{yt}(\mathrm{Re}\,\mathbf{x}_I))\big)\\ &\quad + \mathrm{i}\,\Big((\mathbf{R}_j^{xt})^{\mathsf{T}}\big(u_\Theta(\mathbf{R}_j^{xt}(\mathrm{Im}\,\mathbf{x}_I))\big)\Big) + \mathrm{i}\,\Big((\mathbf{R}_j^{yt})^{\mathsf{T}}\big(u_\Theta(\mathbf{R}_j^{yt}(\mathrm{Im}\,\mathbf{x}_I))\big)\Big)\Big],\end{aligned}$$

where the operators $\mathbf{R}_j^{xt}$ and $\mathbf{R}_j^{yt}$ extract the j-th spatio-temporal slices in xt- and yt-direction, respectively. The operators $(\mathbf{R}_j^{xt})^{\mathsf{T}}$ and $(\mathbf{R}_j^{yt})^{\mathsf{T}}$ denote their corresponding adjoint operators which reposition the slices at their original positions. Note that our XT,YT method discussed in Section 2.2 represents a special case of the decomposition (15). Using the XT,YT approach, the initial reconstruction $\mathbf{x}_{\mathrm{ini}}$ is decomposed in its disjoint xt- and yt-spatio-temporal slices. As in [20], we split the available data in 12/3/4 subjects for training, validation and testing and performed a 4-fold cross-validation. The CNN u_Θ was trained for 12 epochs using ADAM [19] with a learning rate of 10^{-5}.
Then, using the same notation as in the previous sections, functional (5) is given by

$$F_{\mathbf{x}_{\mathrm{CNN}},\lambda,\mathbf{y}_I}(\mathbf{x}) = ||\mathbf{E}\mathbf{x} - \mathbf{y}_I|| + \lambda\, ||\mathbf{x}_{\mathrm{CNN}} - \mathbf{x}||_2^2. \qquad (17)$$

It can be easily seen that minimizing (17) corresponds to solving the system $\mathbf{H}\mathbf{x} = \mathbf{b}$, where

$$\begin{aligned} \mathbf{H} &= \mathbf{E}_I^{\mathsf{H}}\mathbf{E}_I + \lambda\,\mathbf{I}, \\ \mathbf{b} &= \mathbf{x}_I + \lambda\,\mathbf{x}_{\mathrm{CNN}}. \end{aligned} \tag{18}$$

Since $\mathbf{H}$ is symmetric, we used the PCG method for minimizing (17). Note that because of strong convexity, (17) has a unique solution which can be approximated using any iterative scheme. For the experiments, we set $\lambda = 0.1$ and performed $n_{\mathrm{PCG}} = 16$ iterations to approximate the solution of (17).

Results for 2D Radial Cine MRI

Figure 2.9 shows the intermediate reconstructions obtained with the proposed three-steps reconstruction scheme as well as their point-wise errors. Figure 2.9 A shows the initial NUFFT-reconstruction directly obtained from the measured data. Figure 2.9 B shows the NUFFT-reconstruction after the processing with the previously trained CNN. In Figure 2.9 C, the final result of the proposed scheme can be seen which was obtained after the iterative reconstruction using the previously generated CNN-image prior. Figure 2.9 D shows the ground truth images obtained with *kt*-SENSE using $N_\theta = 3400$ radial spokes. We see that the CNN successfully removed a large portion of undersampling artefacts. However, applying the CNN also smoothed out temporal information as pointed out by the yellow arrow in Figure 2.9 B. Minimizing (17) yielded a solution with increased data-consistency. Interestingly, the previously smoothed out image details are visible again, see Figure 2.9 C and D.

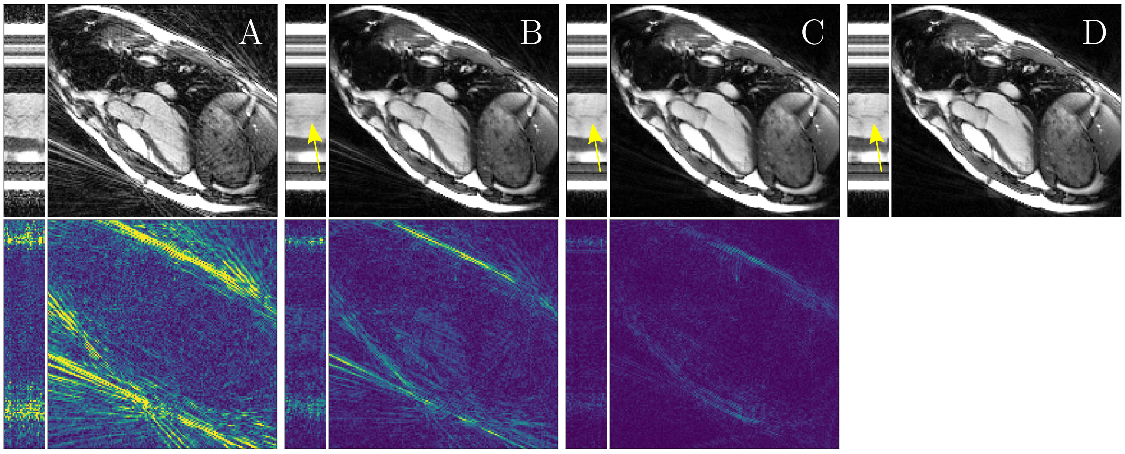


FIGURE 2.9 Intermediate results and point-wise error-images of our three-steps approach. A: Initial NUFFT-reconstruction $\mathbf{x}_I$, B: artefact-corrected image $\mathbf{x}_{\mathrm{CNN}}$ using the XT,YT method, C: CNNs-based regularized solution $\mathbf{x}_{\mathrm{REC}}$, D: ground truth image reconstruction with *kt*-SENSE and $N_\theta = 3400$ radial lines. The point-wise error images were magnified by a factor of ×3. The yellow arrows show details which were smoothed out in the CNN-prior $\mathbf{x}_{\mathrm{CNN}}$ but are visible again in the final reconstruction $\mathbf{x}_{\mathrm{REC}}$. Figure adapted from [22].

Table 2.4 shows the performance of our reconstruction scheme compared to the TV-minimization method and the dictionary learning-based method (DIC) in terms of PSNR, NRMSE, SSIM and HPSI. The measures were obtained as averages over the four different folds. We see that our proposed CNN-based regularized iterative reconstruction consistently yielded the best results with respect to all reported measures. Only applying the CNN yielded results which are superior to TV but not to DIC. However, after the subsequent iterations, the final result outperforms DIC with respect to all measures. In addition, obtaining the CNN-image prior using the pre-trained CNN is faster than DIC by several orders of magnitude since we do not need to solve the sparse-coding problem for each image patch. For more details, we refer to [22].

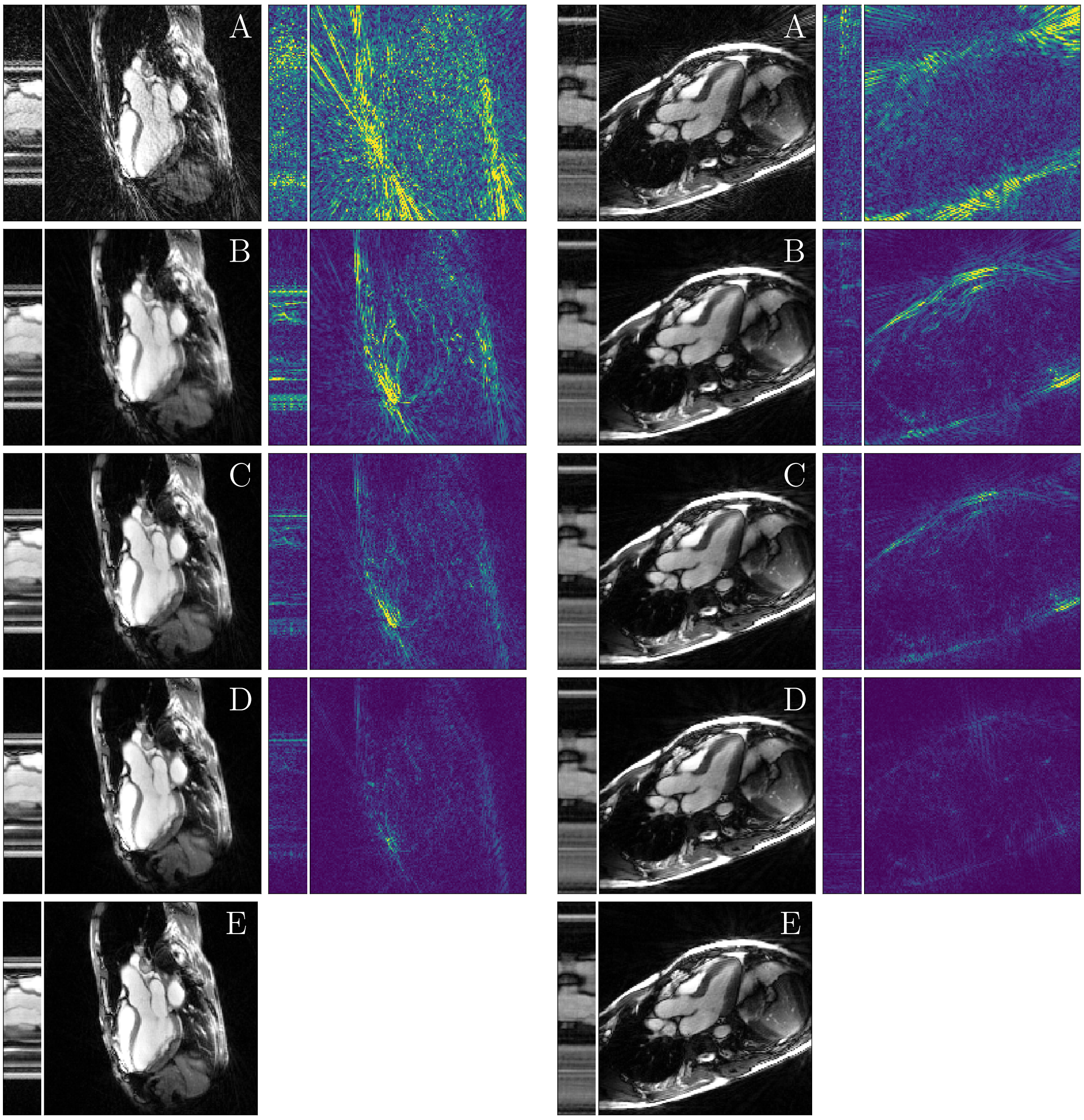


FIGURE 2.10 Results and point-wise error images for the comparison with other reconstruction methods. A: Initial NUFFT-reconstruction $\mathbf{x}_I$ using $N_\theta = 1130$ radial trajectories, B: TV-minimization solution, C: dictionary learning-based solution DIC, D: CNN-regularized solution $\mathbf{x}_{\mathrm{REC}}$, E: ground truth image obtained by kt-SENSE using $N_\theta = 3400$ radial lines. The point-wise error images were magnified by a factor of ×5. The point-wise error is the lowest for the reconstruction $\mathbf{x}_{\mathrm{REC}}$. Figure taken from [22].

Experimental Set-Up for 3D Low-Dose CT

Here, we tested the proposed reconstruction scheme for 3D low-dose CT. We used a dataset of $n = 16$ patients from the randomized DISCHARGE trial [23]. We cropped each volume

Table 2.4 Quantitative measures for the 2D radial cine MRI example. The measures are obtained as averages over the four different folds. Table taken from [22].

	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

to a fixed size of $N_x \times N_y \times N_z = 512 \times 512 \times 128$. From these volumes, we generated low-dose sinograms by simulating a low-dose CT scan acquisition using a cone beam geometry. The retrospective data generation was performed according to [2], i.e. by using the forward model

$$\mathbf{y}_\eta = \mathbf{T}\mathbf{x} + \boldsymbol{\eta} = p\exp\{-\mu\,\mathbf{R}\mathbf{x}\} + \boldsymbol{\eta}, \tag{19}$$

where μ denotes the linear attenuation coefficient of water which was chosen as $\mu = 0.02$ and p is the number of photons per voxel. The operator $\mathbf{R}$ corresponds to a discrete version of the ray-transform and $\boldsymbol{\eta}$ denotes Poisson-distributed noise which was used to contaminate the measured data in order to simulate a low-dose scanning protocol. In this experiment, the operator $\mathbf{R}$ was discretized by $N_\psi = 1000$ angles with a detector array of shape $N_{r_x} \times N_{r_y} = 320 \times 800$. The reconstruction spaces were discretized according to the pixel-spacing found in the respective DICOM files. The initial reconstruction was obtained as $\mathbf{x}_\eta = \mathbf{R}^\dagger\big(-\mu^{-1}\ln(p^{-1}\mathbf{y}_{\boldsymbol{eta}})\big)$ with $\mathbf{R}^\dagger$ being the filtered back-projection using a Ram-Lak filter. The operators $\mathbf{R}$ and $\mathbf{R}^\dagger$ were implemented using the publicly available Operator Discretization Library `ODL` [1]. For this experiment, we chose the network to be a 3D U-net as the one given in [14]. Since the volumes have a shape of $512 \times 512 \times 128$ but only volumes of $128 \times 128 \times 16$ fit the GPU with 12 GB of memory, the image prior $\mathbf{x}_{\mathrm{CNN}}$ was obtained as described in (15) where we used a patch-size of $\mathbf{p} = (128, 128, 16)$ and strides of $\mathbf{s} = (16, 6, 8)$, respectively.

We split the dataset in 12/2/2 patients for training, validation and testing and performed a seven-fold cross-validation. For each fold, we trained the 3D U-net for 115 epochs, where we used the L_2-norm between CNN prediction and target image as loss function during training.

After having obtained the CNN-image prior $\mathbf{x}_{\mathrm{CNN}}$, we considered the functional

$$F_{\mathbf{y}_\eta, \mathbf{x}_{\mathrm{CNN}}, \lambda}(\mathbf{x}) = D_{\mathrm{KL}}(\mathbf{T}\mathbf{x}, \mathbf{y}_\eta) + \lambda \|\mathbf{x} - \mathbf{x}_{\mathrm{CNN}}\|_2^2 \to \min, \tag{20}$$

where D_{KL} denotes the Kullback Leibler-divergence. Problem (20) was solved by performing $n_{\mathrm{iter}} = 4$ iterations of the Landweber method. Since the measured data $\mathbf{y}_\eta$ was contaminated by noise, the number of iterations is on purpose chosen to be relatively small in order to avoid the semi-convergence behavior of the Landweber method. The regularization parameter was set to $\lambda = 1$ for all experiments.

Results for 3D Low-Dose CT

Figure 2.11 shows the intermediate results of the proposed three-steps reconstruction, their corresponding point-wise errors and the ground truth image. In Figure 2.11 A, we see the initial FBP-reconstruction affected by the noise of the simulated low-dose scan. The volume which was patch-wise processed by the previously trained CNN is visible in Figure

2.11 B which shows that most of the noise was successfully removed. However, visually, the result is relatively smooth. Minimizing functional (20) led to the final reconstruction which can be seen in Figure 2.11 C. As expected, some noise was re-introduced in the obtained image due to the fact that the noisy data $\mathbf{y}_\eta$ was used in the Landweber iteration. However, the level of noise is relatively low and lends the final image the characteristic texture of CT images. Fine diagnostic image details as the right coronary artery which is highlighted by the yellow arrows are clearly visible in all processed images. Figure 2.11 D shows the ground truth image.

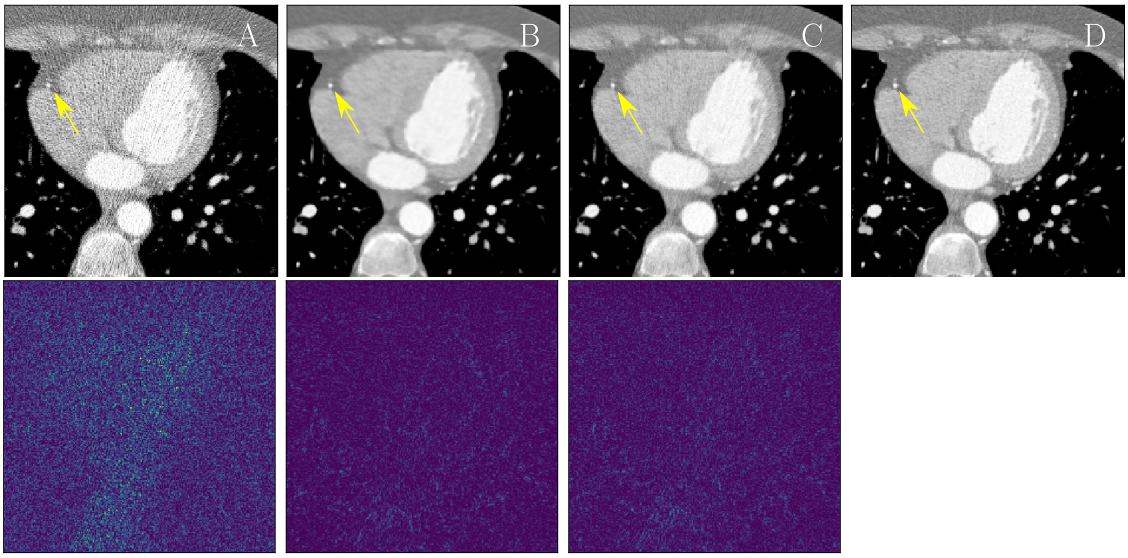


Figure 2.11 Intermediate steps of the proposed three-steps approach and corresponding error-images. A: Low-dose FBP-reconstruction $\mathbf{x}_\eta$, B: denoised image $\mathbf{x}_{\text{CNN}}$, C: CNNs-based regularized solution $\mathbf{x}_{\text{REC}}$, D: ground truth image. Diagnostic details are well visible in the prior $\mathbf{x}_{\text{CNN}}$ as welll as in $\mathbf{x}_{\text{REC}}$. The yellow arrow points at the right coronary artery. The images are windowed and displayed on the scale with $C = 0\,\text{HU}$, $W = 850\,\text{HU}$. Figure adapted from [22].

Table 2.5 Quantitative measures for the 3D low-dose CT example. The measures are obtained as averages over the seven different folds. Table take from [22].

	FBP	$\mathbf{x}_{\text{CNN}}$	$\mathbf{x}_{\text{REC}}$	**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 2.5 lists the quantitative measures obtained for the experiments averaged over the seven different folds. As can be seen, the largest improvement in our approach is given by the application of the CNN to the initial FBP-reconstruction. Minimizing the CNN-regularized functional only slightly further improved SSIM. PSNR and NRMSE decreased as expected due to the use of the noisy measurements $\mathbf{y}_\eta$. HPSI remained approximately the same.

Table 2.5 also shows the comparison of our proposed approach to TV and DIC which were clearly outperformed with respect to all reported measures. These quantitative results are also well-reflected in Figure 2.12 which shows the initial FBP-reconstruction (A), the TV-

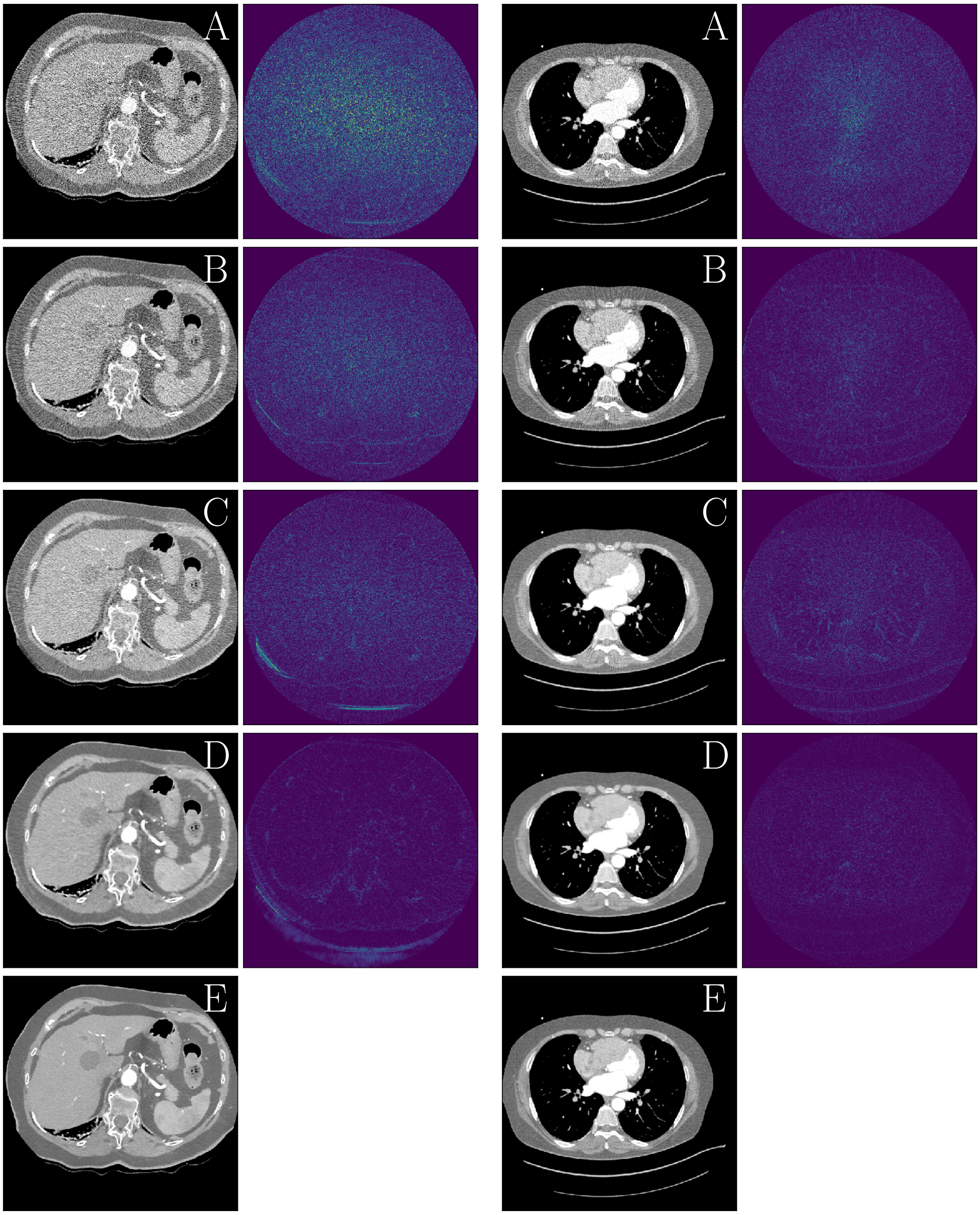


Figure 2.12 Results and point-wise error images of two patients obtained with the different reconstruction methods. A: Low-dose FBP-reconstruction $\mathbf{x}_\eta$, B: TV, C: DIC, D: proposed CNN-based reconstruction $\mathbf{x}_{\mathrm{REC}}$, E: ground-truth image. The images are windowed and displayed on the same scale with $C = 0\,\mathrm{HU}$, $W = 800\,\mathrm{HU}$. Figure taken from [22].

reconstruction (B), the DIC-reconstruction (C), the solution using our proposed scheme (D) and the ground truth image (E). From the point-wise error-images as well as from the appearance of the reconstructed images, we see the superiority of the CNN-based method compared to TV and DIC.

3 Discussion

In this thesis, we have addressed and investigated the applicability of NNs-based regularization of large-scale ill-posed inverse problems occurring in medical image reconstruction. In particular, we have posed our focus on cardiac imaging and more precisely, on 3D low-dose CT and 2D radial cine MRI, although some of the concepts, considerations and results are most probably applicable to general image reconstruction problems as well.

In Section 2.1, we have performed an ablation study for the task of image reconstruction in sparse view CT, where we tackled the reconstruction problem using a novel end-to-end trainable cascade of U-nets [21]. The U-nets cascade is an iterative network which alternates between processing the images with U-nets and data-consistency (DC) layers. We investigated the reconstruction performance of the cascaded network for different lengths of the cascade, i.e. number of alternations between CNNs and DC layers, while keeping the number of trainable parameters approximately fixed. By doing so, we were able to isolate the effect of multiple interceptions of the U-nets by DC layers. This is of interest as cascaded networks represent the state-of-the-art in medical image reconstruction but are currently only applicable to either small-scale problems (e.g. 2D scanner geometries for CT) or problems where the forward and adjoint operators $\mathbf{A}$ and $\mathbf{A}^{\mathsf{H}}$ are computationally cheap (e.g. Cartesian single-coil acquisition protocols in MRI). The results in [21] suggest that even if increasing the number of iterations seems to have a visually positive impact on the reconstruction quality, the results in terms of quantitative measures were somewhat all comparable. Further, using more sophisticated CNNs as the U-net, turned out to be favorable over using simple blocks of CNNs. Therefore, for large-scale image reconstruction problems where the construction of cascaded/iterative CNNs is computationally prohibitive, using a single CNN to obtain an image prior which is used in a subsequent iterative reconstruction seems to be a valid alternative, at least given the nowadays available hardware.

In Section 2.2, we have dedicated ourselves to the construction of a CNN for obtaining such an image prior for 2D radial cine MRI. Given the fact that the performance of a CNN is always linked to the number of available training samples and that in medical imaging applications, large datasets are rarely available, we have put our attention on the development of a robust and reliable approach which overcomes the data-availability problem. In contrast to the observable tendency in the Deep Learning research community to include more and more sophisticated blocks, e.g. perceptual losses and generative adversarial approaches [39], we adopted a different strategy based on the simplification of the problem to be solved. Our XT,YT approach uses a simple yet efficient idea. Instead of learning the mapping from the manifold of the undersampled images to the one of the ground-truth images (3D to 3D), it learns to map the corresponding 2D spatio-temporal slices which can be extracted in xt- and yt-direction from the 2D image sequences. For a single pair of undersampled and ground truth 2D cine MR images of shape $N_x \times N_y \times N_t$ (i.e. 2D + time), one immediately has access to $N_x + N_y$ training pairs. In addition, the topological complexity of the manifold of the xt- and yt-spatio temporal slices was reported to be lower than the one of images in the xy-plane and thus facilitates the learning of the mapping.

Our XT,YT approach only uses 2D convolutional layers in contrast to the 3D U-net [14] but still is able to exploit spatio-temporal correlation of adjacent time frames. Using only 2D convolutional layers further reduces the number of trainable parameters and prevents the network from overfitting. A possible limitation could be a too strong smoothing of temporal information which, however, was not observed in the experiments. As shown in

[20], our XT,YT approach outperformed classical iterative reconstruction schemes based on *kt*-FOCUSS [18], *kt*-SENSE [10], TV-minimization [6] and dictionary learning + TV [7], [35] in terms of all reported measures. Further, it achieved competitive results with state-of-the-art methods based on cascaded networks presented in [31] and [25]. In addition, as shown in Figure 2.8, using the XT,YT approach, it is possible to obtain competitive results even with relatively shallow CNNs which can prevent the network from overfitting. Using more shallow CNNs, also offers the possibility to integrate the XT,YT method as a building block in cascaded networks. Last, we showed that due to the change of perspective on the data, the method naturally achieves the property of almost being rotation-equivariant. Thus, the method is robust with respect to rotations without the need to explicitly incorporate rotation-equivariant operations in the CNN. Even though the method was presented for 2D radial cine MRI, we expect the method to be applicable to general inverse problems where temporal correlation can be exploited.

In Section 2.3, we have tackled the problem of realistic large-scale ill-posed image reconstruction problems using CNNs-based regularization within a Tikhonov regularization framework. Based on the results presented in Section 2.1 and discussed in more detail in [21], we opted for a reconstruction scheme which involves the following steps: First, the initial reconstruction is obtained by directly reconstructing the image from the measured data. As a second step, the initial reconstruction is processed by a CNN which in general operates on sub-portions of the images due to the relatively large dimensionality of the considered problems. Last, the obtained output of the CNN is used as an image prior for the formulation of a Tikhonov-regularized functional which is subsequently minimized. Since this functional only depends on the image, its minimization can be achieved by means of any iterative solver. We have applied and evaluated the proposed approach to 2D radial cine MRI and 3D low-dose CT and have compared the proposed three-steps method for the well known total variation (TV)-minimization approach and a method using dictionary learning-based regularization (DIC). The proposed method outperformed both TV and DIC in terms of all reported measures and further accelerated the regularization step by several orders of magnitude compared to DIC. A limitation of the three-steps reconstruction method compared to cascaded/iterative networks is the need to choose the regularization parameter λ which controls the strength of the contribution of the regularization. In contrast, in cascaded/iterative networks, the parameter λ can be treated as a trainable parameter and can be learned as well. While in [22] we have chosen λ empirically, there exists a large variety of methods on how to more appropriately choose λ.

Note that the applicability of cascaded/iterative networks is limited to relatively small-scale problems. Only recently, large-scale problems have been addressed as well using iterative networks, see e.g. [13] for volumetric low-dose CT reconstruction. However, the considered volume size is still relatively small being $168 \times 168 \times 168$ and using a scanner geometry which measures projections for only $N_\theta = 60$ angles. Therefore, the three-steps reconstruction approach still represents a viable option for realistic large-scale problems. An assessment of the achieved image quality with respect to clinical tasks, e.g. the assessment of the coronary artery calcium scoring in low-dose CT, is already planned as future work in collaboration with clinicians.

Neural networks-based methods for image reconstruction *empirically* seem to outperform classical methods, potentially allowing for more accurate, faster (MRI) and safer (CT) imaging protocols. However, the research area is still in its infancy and, as recently reported, these methods can be highly unstable with respect to tiny perturbations in image and sampling domain [5]. In particular, fully learned inversion methods as presented in [40] seem to be more susceptible to instability problems compared to CNNs which contain the physical models as [31]. This suggests that restricting the CNNs to only playing the role of learned regularizers in the reconstruction process is favorable. Unless a clear theoretical understanding of what neural networks *exactly do* is available, caution is advised for their integration in practice, especially in a sensible field as medical imaging.

4 Conclusion

In the last years, medical image reconstruction has experienced a paradigm shift due to the re-emergence of Neural Networks (NNs) thanks to publicly available Deep Learning software. Convolutional NNs (CNNs), in particular, have been applied in many different ways within the task of image reconstruction and inverse problems. Although end-to-end trainable iterative networks represent the state-of-the-art for medical image reconstruction, their applicability still remains limited to a class of relatively small-scale problems with the nowadays available hardware. In [21], we observed that for cascading networks with approximately the same number of trainable parameters but different numbers of interceptions of data-consistency layers and CNNs, the results are somewhat comparable. This suggests that for large-scale inverse problems in medical imaging, where the forward models and the reconstruction operators are computationally heavy to apply and the construction of cascaded networks is computationally prohibitive, generating an image prior using a CNN and subsequently ensuring/increasing data-consistency might be an efficient and viable option. In [22], we have presented a simple yet efficient generalized approach for solving large-scale image reconstruction problems in medical imaging. In contrast to iterative networks which repeatedly employ the forward and the adjoint operators within the network architecture, we only generate one single image prior using an pre-trained CNN which is then used in a subsequent generalized Tikhonov-regularization framework. By doing so, the step of ensuring/increasing data-consistency of the solution with the measured data is separated from the regularization and therefore, more powerful and sophisticated network architectures can be applied as the ones conventionally used in iterative neural networks. Further, in [20] we have developed an approach specifically designed for the task of artefacts-reduction in 2D radial cine MRI, named XT,YT approach. The XT,YT method was designed to be particularly suitable for situations where only limited training data is available. We have demonstrated that the latter can be applied as a post-processing method to generate an image prior to be used in the reconstruction scheme presented in [22]. Further, since it is computationally light, it could also be easily integrated into cascaded networks which will be subject of future work.

3 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] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 2020.

[6] 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.

[7] 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.

[8] Il Yong Chun, Xuehang Zheng, Yong Long, and Jeffrey A Fessler. Bcd-net for low-dose ct reconstruction: Acceleration, convergence, and generalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 31–40. Springer, 2019.

[9] Yong Chun and Jeffrey A Fessler. Deep bcd-net using identical encoding-decoding cnn structures for iterative image recovery. In *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2018.

[10] Li Feng, Monvadi B Srichai, Ruth P Lim, Alexis Harrison, Wilson King, Ganesh Adluru, Edward V R Dibella, Daniel K Sodickson, Ricardo Otazo, and Daniel Kim. Highly accelerated real-time cardiac cine MRI using k-t SPARSE-SENSE. *Magnetic Resonance Imaging*, aug 2012.

[11] 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.

[12] Yoseo Han, Leonard Sunwoo, and Jong Chul Ye. k-space deep learning for accelerated MRI. *IEEE Transactions on Medical Imaging*, 39(2):377–386, 2019.

[13] Andreas Hauptmann, Jonas Adler, Simon Arridge, and Ozan. Öktem. Multi-scale learned iterative reconstruction. *IEEE Transactions on Computational Imaging*, 2020, DOI: 10.1109/TCI.2020.2990299.

[14] 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.

[15] 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.

[16] Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. Deep learning for undersampled MRI reconstruction. *Physics in Medicine & Biology*, 63(13):135007, 2018.

[17] 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.

[18] Hong Jung, Kyunghyun Sung, Krishna S Nayak, Eung Yeop Kim, and Jong Chul Ye. k-t focuss: a general compressed sensing framework for high resolution dynamic MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 61(1):103–116, 2009.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Andreas Kofler, Marc 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*, 39(3):703–717, 2020.

[21] Andreas Kofler, Markus Haltmeier, Christoph Kolbitsch, Marc Kachelrieß, and Marc Dewey. A U-nets cascade for sparse view computed tomography. In *International Workshop on Machine Learning for Medical Image Reconstruction*, page 91–99. Springer, 2018.

[22] Andreas Kofler, Markus Haltmeier, Tobias Schaeffter, Marc Kachelriess, Marc Dewey, Christian Wald, and Christoph Kolbitsch. Neural networks-based regularization for large-scale medical image reconstruction. *Physics in Medicine & Biology*, 65(13):135003, 2020.

[23] Adriane E. Napp, Robert Haase, Georg M. Laule, Michael Schuetz, Matthias Rief, Henryk Dreger, Gudrun Feuchtner, Guy Friedrich, Miloslav Špaček, Vojtěch Suchánek, Klaus Fuglsang Kofoed, Thomas Engstroem, Stephen Schroeder, Tanja Drosch, Matthias Gutberlet, Michael Woinke, Pál Maurovich-Horvat, Béla Merkely, Patrick Donnelly, Peter Ball, Jonathan D. Dodd, Martin Quinn, Luca Saba, Maurizio Porcu, Marco Francone, Massimo Mancone, Andrejs Erglis, Ligita Zvaigzne, Antanas Jankauskas, Gintare Sakalyte, Tomasz Harań, Malgorzata Ilnicka-Suckiel, Nuno Bettencourt, Vasco Gama-Ribeiro, Sebastian Condrea, Imre Benedek, Nada Čemerlić

Adjić, Oto Adjić, José Rodriguez-Palomares, Bruno Garcia del Blanco, Giles Roditi, Colin Berry, Gershan Davis, Erica Thwaite, Juhani Knuuti, Mikko Pietilä, Cezary Kepka, Mariusz Kruk, Radosav Vidakovic, Aleksandar N. Neskovic, Ignacio Díez, Iñigo Lecumberri, Jacob Geleijns, Christine Kubiak, Anke Strenge-Hesse, The-Hoang Do, Felix Frömel, Iñaki Gutiérrez-Ibarluzea, Gaizka Benguria-Arrate, Hans Keiding, Christoph Katzer, Jacqueline Müller-Nordhorn, Nina Rieckmann, Mario Walther, Peter Schlattmann, Marc Dewey, and The DISCHARGE Trial Group. Computed tomography versus invasive coronary angiography: design and methods of the pragmatic randomised multicentre discharge trial. *European radiology*, 27(7):2957–2968, 2017.

[24] Valentina O Puntmann, Silvia Valbuena, Rocio Hinojar, Steffen E Petersen, John P Greenwood, Christopher M Kramer, Raymond Y Kwong, Gerry P McCann, Colin Berry, Eike Nagel, and The SCMR Clinical Trial Writing Group. Society for cardiovascular magnetic resonance (SCMR) expert consensus for CMR imaging endpoints in clinical research: part i-analytical validation and clinical qualification. *Journal of Cardiovascular Magnetic Resonance*, 20(1):67, 2018.

[25] 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.

[26] Saiprasad Ravishankar, Il Yong Chun, and Jeffrey A Fessler. Physics-driven deep training of dictionary-based algorithms for MR image reconstruction. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 1859–1863. IEEE, 2017.

[27] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018.

[28] Carlos E Rochitte, Richard T George, Marcus Y Chen, Armin Arbab-Zadeh, Marc Dewey, Julie M Miller, Hiroyuki Niinuma, Kunihiro Yoshioka, Kakuya Kitagawa, Shiro Nakamori, Roger Laham, Andrea L. Vavere, Rodrigo J. Cerci, Cerci C. Vishal, Mehra Cesar Nomura, Kofoed Klaus F, Jinzaki Masahiro, Sachio Kuribayashi, Albert De Roos, Michael Laule, Swee Yaw Tan, John Hoe, Paul Narinder, Frank J. Rybicki, Jeffery A. Brinker, Andrew E. Arai, Christopher Cox, Melvin E. Clouse, Marcelo F. Di Carli, and Joao A.C. Lima. Computed tomography angiography and perfusion to assess coronary artery stenosis causing perfusion defects by single photon emission computed tomography: the core320 study. *European heart journal*, 35(17):1120–1130, 2014.

[29] 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.

[30] Christopher M. Sandino, Neerav Dixit, Joseph Y. Cheng, and Shreyas S. Vasanawala. Deep convolutional neural networks for accelerated dynamic magnetic resonance imaging. *Proceedings of 31st Conference of Neural Information Processing Systems (NIPS), Medical Imaigng meets NIPS Workshop.*, 2017. Online. Available at www.doc.ic.ac.uk/bglocker/public/mednips2017/mednips_2017_paper_19.pdf.

[31] 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.

[32] Georg M Schuetz, Niki Maria Zacharopoulou, Peter Schlattmann, and Marc Dewey. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. *Annals of internal medicine*, 152(3):167–177, 2010.

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin. Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1310–1321, 2017.

[40] 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):487–492, 2018.

4 Eidesstattliche Versicherung

Ich, Andreas Kofler, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema *Deep Learning-based Methods for Image Reconstruction Cardiac CT and Cardiac Cine MRI - Deep Learning-basierte Methoden zur Bildrekonstruktion in der Herz-CT und Herzfunktions-MRT* selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe. Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Erstbetreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.org) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.

Datum Unterschrift

5 Ausführliche Anteilserklärung an den erfolgten Publikationen

Publikation:

♦ Andreas Kofler, Marc 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.* In IEEE Transactions on Medical Imaging, 39(3), p. 703–717, 2020.

Beitrag im Einzelnen:

- Thema, Idee, Konzeption (Kofler, Kolbitsch, Dewey)
- Bereitstellung der MR Daten (Koblitsch)
- Datenaufbereitung (Kofler, Kolbitsch, Wald)
- Entwicklung der Hypothesen zur Persistent Homology Analysis (Kofler)
- Persistent Homology Analysis (Wald)
- Methoden- und Algorithmenentwicklung: Alle auf Neuronalen Netzen basierenden Post-Processing Methoden (Kofler)
- Methoden- und Algorithmenentwicklung: Implementierung der Methode Dictionary Learning + Total Variation sowie Anpassung der Netzwerkkaskade DnCn3DDS (Kofler)
- Anpassung der Netzwerkkaskade CRNN (Wald)
- Bereitstellung der Vergleichsmethoden *kt*-FOCUSS, *kt*-SENSE sowie TVT (Kolbitsch)
- Quantitative und statistische Auswertung der Ergebnisse - sämtliche Tabellen (Kofler)
- Erstellung aller Bilder (Kofler) mit Ausnahme jener zur Persistent Homology Analysis (Wald)
- Qualitative Auswertung der Ergebnisse (Kofler und Kolbitsch)
- Diskussion der Methode und Ergebnisse (Kofler)
- Erster Manuskriptentwurf (Kofler)
- Überarbeitung und weitere Strukturierung (Kofler, Wald, Kolbitsch, Schaeffter)
- Korrekturlesen (Dewey, Schaeffter, Wald, Kolbitsch)
- Präsentation auf Konferenzen: ISMRM 2019, ICIAM 2019, AIP 2019 (Kofler)

Unterschrift, Datum und Stempel des betreuenden Hochschullehrers

---

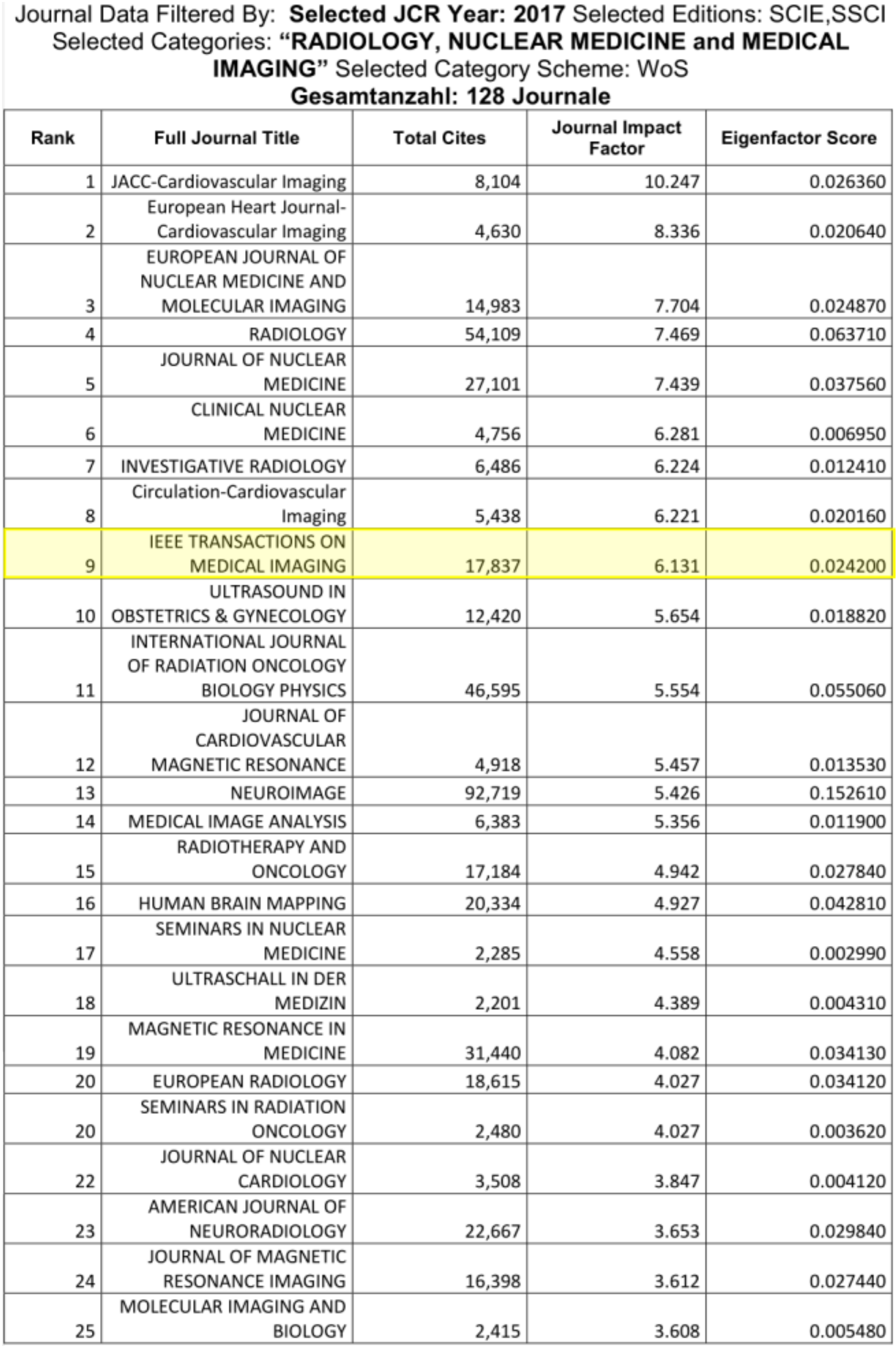
Unterschrift des Doktoranden

6 Originalpublikation

- Andreas Kofler, Marc 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.* In IEEE Transactions on Medical Imaging, 39(3), p. 703–717, 2020.

https://doi.org/10.1109/TMI.2019.2930318

7 Auszug aus der Journal Summary List



Journal Data Filtered By: **Selected JCR Year: 2017** Selected Editions: SCIE,SSCI Selected Categories: **"RADIOLOGY, NUCLEAR MEDICINE and MEDICAL IMAGING"** Selected Category Scheme: WoS

Gesamtanzahl: 128 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	JACC-Cardiovascular Imaging	8,104	10.247	0.026360
2	European Heart Journal-Cardiovascular Imaging	4,630	8.336	0.020640
3	EUROPEAN JOURNAL OF NUCLEAR MEDICINE AND MOLECULAR IMAGING	14,983	7.704	0.024870
4	RADIOLOGY	54,109	7.469	0.063710
5	JOURNAL OF NUCLEAR MEDICINE	27,101	7.439	0.037560
6	CLINICAL NUCLEAR MEDICINE	4,756	6.281	0.006950
7	INVESTIGATIVE RADIOLOGY	6,486	6.224	0.012410
8	Circulation-Cardiovascular Imaging	5,438	6.221	0.020160
9	IEEE TRANSACTIONS ON MEDICAL IMAGING	17,837	6.131	0.024200
10	ULTRASOUND IN OBSTETRICS & GYNECOLOGY	12,420	5.654	0.018820
11	INTERNATIONAL JOURNAL OF RADIATION ONCOLOGY BIOLOGY PHYSICS	46,595	5.554	0.055060
12	JOURNAL OF CARDIOVASCULAR MAGNETIC RESONANCE	4,918	5.457	0.013530
13	NEUROIMAGE	92,719	5.426	0.152610
14	MEDICAL IMAGE ANALYSIS	6,383	5.356	0.011900
15	RADIOTHERAPY AND ONCOLOGY	17,184	4.942	0.027840
16	HUMAN BRAIN MAPPING	20,334	4.927	0.042810
17	SEMINARS IN NUCLEAR MEDICINE	2,285	4.558	0.002990
18	ULTRASCHALL IN DER MEDIZIN	2,201	4.389	0.004310
19	MAGNETIC RESONANCE IN MEDICINE	31,440	4.082	0.034130
20	EUROPEAN RADIOLOGY	18,615	4.027	0.034120
20	SEMINARS IN RADIATION ONCOLOGY	2,480	4.027	0.003620
22	JOURNAL OF NUCLEAR CARDIOLOGY	3,508	3.847	0.004120
23	AMERICAN JOURNAL OF NEURORADIOLOGY	22,667	3.653	0.029840
24	JOURNAL OF MAGNETIC RESONANCE IMAGING	16,398	3.612	0.027440
25	MOLECULAR IMAGING AND BIOLOGY	2,415	3.608	0.005480

Spatio-Temporal Deep Learning-Based Undersampling Artefact Reduction for 2D Radial Cine MRI With Limited Training Data

Andreas Kofler, Marc Dewey, Tobias Schaeffter, Christian Wald, and Christoph Kolbitsch

***Abstract*—In this work we reduce undersampling artefacts in two-dimensional (2D) golden-angle radial cine cardiac MRI by applying a modified version of the U-net. The network is trained on 2D spatio-temporal slices which are previously extracted from the image sequences. We compare our approach to two 2D and a 3D deep learning-based post processing methods, three iterative reconstruction methods and two recently proposed methods for dynamic cardiac MRI based on 2D and 3D cascaded networks. Our method outperforms the 2D spatially trained U-net and the 2D spatio-temporal U-net. Compared to the 3D spatio-temporal U-net, our method delivers comparable results, but requiring shorter training times and less training data. Compared to the compressed sensing-based methods *kt*-FOCUSS and a total variation regularized reconstruction approach, our method improves image quality with respect to all reported metrics. Further, it achieves competitive results when compared to the iterative reconstruction method based on adaptive regularization with dictionary learning and total variation and when compared to the methods based on cascaded networks, while only requiring a small fraction of the computational and training time. A persistent homology analysis demonstrates that the data manifold of the spatio-temporal domain has a lower complexity than the one of the spatial domain and therefore, the learning of a projection-like mapping is facilitated. Even when trained on only one single subject without data-augmentation, our approach yields results which are similar to the ones obtained on a large training dataset. This makes the method particularly suitable for training a network on limited training data. Finally, in contrast to the spatial 2D U-net, our proposed method is shown to be naturally robust with respect to image rotation in image space and almost achieves rotation-equivariance where neither data-augmentation nor a particular network design are required.**

***Index Terms*—Deep learning, neural networks, dynamic MRI, image processing, compressed sensing, persistent homology analysis.**

I. Introduction

MAGNETIC Resonance Imaging (MRI) is a widely used non-invasive imaging modality in clinical practice. Especially for cardiac applications, MRI does not only provide anatomical imaging with excellent soft tissue contrast but also allows for functional assessment by using 2D cine MRI. Such images show the heart anatomy for different phases of the cardiac cycle providing valuable information of the heart function [1], [2].

However, MRI suffers from long data-acquisition which determines the achievable spatial and temporal resolution. In order to shorten scan times, ensure sufficiently large spatial coverage and high spatial and temporal resolution, a wide range of undersampling and reconstruction techniques have been proposed, ranging from Parallel Imaging to Compressed Sensing (CS) and Dictionary Learning [3], [4]. Cine MRI provides a temporal sequence of images and therefore offers the possibility to exploit the temporal correlation of adjacent frames in order to reduce undersampling artefacts. The movement of the heart during the cardiac cycle is mainly smooth and continuous. Ensuring that undersampling artefacts along time are incoherent and using a sparsifying transform along time such as Fourier transform [3], Principal Component Analysis [5], [6], Wavelet transform [7] or a transform learned from data [8], [9] combined with a L_1-norm minimization approach can strongly reduce undersampling artefacts. The main challenges of these techniques are to ensure that the

Manuscript received February 26, 2019; revised May 28, 2019; accepted July 12, 2019. Date of publication August 9, 2019; date of current version February 28, 2020. The work of A. Kofler was supported by the German Research Foundation (DFG), BIOQIC, under Project GRK 2260. The work of M. Dewey was supported in part by the German Research Foundation (DFG), BIOQIC, under Project GRK 2260 and in part by the Digital Health Accelerator (DHA) of the Berlin Institute of Health. The work of C. Wald was supported by the DHA of the Berlin Institute of Health. *(Corresponding author: Andreas Kofler.)*

A. Kofler and C. Wald are with the Department of Radiology, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany (e-mail: andreas.kofler@charite.de; christian.wald@charite.de).

M. Dewey is with the Department of Radiology, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany, and also with the Berlin Institute of Health, 10178 Berlin, Germany (e-mail: marc.dewey@charite.de).

T. Schaeffter and C. Kolbitsch are with the Physikalisch-Technische Bundesanstalt (PTB), 10587 Berlin, Germany, and also with King's College London, London WC2R 2LS, U.K. (e-mail: tobias.schaeffter@ptb.de; christoph.kolbitsch@ptb.de).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2930318

sparsifying transform really leads to a sparse signal and long reconstruction times due to the iterative reconstruction.

Recently, Neural Networks (NNs) have been applied to inverse problems as image reconstruction in MRI [10], [11], [12], [13] and computed tomography (CT) [10], [14], [15]. Autoencoders, and in particular the U-net [16], a convolutional NN (CNN) which was first introduced for biomedical image segmentation, and different derivations of it [17], [18], have been widely used for removing undersampling artefacts in different medical imaging modalities.

In initial works, the images were most commonly reconstructed or processed frame by frame, see e.g. [10]. In the case of dynamic MRI, however, the temporal correlation of $2D$ MRI sequences can be exploited by aligning frames along the channel axis. Thus, $2D$ CNNs can be trained to map whole undersampled image sequences to their corresponding fully sampled image sequences [19], [20]. Further, also CNNs employing $3D$-convolutions were shown to be trainable on entire image sequences, either as post-processing methods [21] or as unrolled iterative reconstruction schemes [19]. However, in general, due to the resulting high dimensionality of the considered problem, either a large dataset or the application of data-augmentation techniques are indispensable to obtain satisfactory results, see e.g. [19], [21].

Nowadays it is common practice to learn the filters of the convolutional layers by considering the images in the spatial domain. In this work, we propose to apply CNNs to two-dimensional slices extracted from the spatio-temporal dimension in order to remove undersampling artefacts from a $2D$ cine MR scan obtained with a $2D$ Golden radial sampling scheme [22]. A persistent homology analysis shows that the manifold of the spatio-temporal slices has a lower topological complexity than the manifold of the two-dimensional spatial image frames and suggests that the learning process of the network can therefore be facilitated. We compare our proposed approach to a $2D$ U-net trained frame-by-frame [10], a $2D$ U-net trained image sequence-wise [20] and a $3D$ U-net [21] in terms of image quality, amount of required training data and stability with respect to rotation of the images. The latter is important because $2D$ cine MRI is commonly obtained in oblique planes which are adapted to the patients anatomy. Our spatio-temporal approach method is also compared to three CS-based approaches for image reconstruction of cine MRI: kt-FOCUSS [23], a total variation minimization-based method [4] and a Dictionary Learning- and total variation-based reconstruction method [9]. Further, we compare our method to two methods for cine MRI based on cascaded networks [19], [24].

The paper is organized as follows. In Section II, we shortly discuss how NNs have been integrated within the problem of image reconstruction in MRI so far. Section III introduces our proposed method by discussing an a priori performed persistent homology analysis of the data which is needed to derive the approach as well as the network's architecture. We then show results of In-Vivo experiments and compare our method to other Deep Learning- and CS-based methods in Section IV and finish with a discussion and conclusion in Section V.

II. Problem Formulation

In dynamic MRI, the image reconstruction problem is given by finding a solution of the inverse problem

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

where $\mathbf{x} \in \mathbb{C}^N$ denotes the complex-valued image sequence with $N = N_x N_y N_t$, $\mathbf{F}$ denotes the Fourier encoding matrix and $\mathbf{y}$ corresponds to the measured data in k-space. As the data-acquisition process in MRI is slow, undersampling schemes are applied to fasten the measurement process. Therefore, the inverse problem one encounters in applications is of the form

$$\mathbf{F}_I\mathbf{x} = \mathbf{y}_I, \tag{2}$$

where $\mathbf{F}_I = \mathbf{S}_I\mathbf{F}$ and $\mathbf{S}_I \in \mathbb{C}^{M\times N}$ denotes a binary undersampling operator with $M \ll N$ which sets non-measured values in k-space to zero. Thereby, $I \subset J = \{1, \dots, N\}$ corresponds to the set of indices of the measured Fourier coefficients. Since $M \ll N$, the problem in (2) is underdetermined and therefore ill-posed. Hence, a direct solution is not possible and usually regularization approaches have to be applied in order to constrain the sought solution. Two widely used regularization techniques are based on Dictionary Learning [8], [9] and total-variation (TV) minimization [4], [25]. However, since the methods employ the regularization within an iterative reconstruction, solving the problem in (2) is time consuming and NNs have been considered as a valid and powerful alternative, see e.g. [10]–[12], [19], [21].

Most commonly, the networks are trained by considering the images in the spatial domain. By doing so, the network learns to distinguish between diagnostic content of the image and the artefacts by exploiting the natural correlation of neighbouring pixel values in spatial domain. Given a dynamic process, one can further make use of the correlation of temporal slices amongst each other. In [20], the work of [10] is extended in the sense that a U-net is trained to directly map whole $2D$ image sequences of undersampled image reconstructions to $2D$ image sequences of ground truth images. In [19], the temporal dimension of the sequence is taken into account in the same manner, where furthermore, a weighted data-sharing and a data-consistency approach further improve the quality of the reconstruction. For the $2D$ networks, frames corresponding to different cardiac phases are aligned along the channel axis. As shown in [19] and [21], CNNs employing $3D$ convolutional layers can also be applied for the task of removing undersampling artefacts in dynamic sequences. Note that, for a network employing $2D$ convolutional layers and assuming the channel's dimension to be the one along which feature maps are combined by linear combination, aligning temporal frames along the channel's axis only slightly increases the computational complexity of the CNN. In this case, the filters size only increases for the first and the last convolutional layers. Employing $3D$ convolutional layers, in contrast, adds further non-negligible computational cost as well as hardware requirements, increases training time, the number of trainable parameters and therefore the number of samples required to successfully train a network without experiencing overfitting.

In the aforementioned methods, the resulting number of available training samples reduces to the number of $2D$ image sequences. Since NNs are well known to require a large number of training samples and as the collection of proper data can be challenging, using these approaches, one usually has to heavily rely on the use of data-augmentation techniques, see e.g. [19], have access to a large dataset [21] or both in order to obtain a good representation of the data manifold. However, data-augmentation might also be non-trivial, time consuming or not possible to be performed on the fly. In the case of image reconstruction, the dataset is obtained by a prior data-acquisition process. In a simulation-based framework, one can for example apply arbitrary transformations to a ground truth image, e.g. elastic transformations, and then simulate the data-acquisition process. Also, using different undersampling masks to obtain zero-filled reconstructions can further enrich the data, see for example [19], [20]. However, assuming a fixed dataset of pairs of undersampled image reconstructions and ground truth images, transformations would have to be applied to each pair, possibly altering the structure of the undersampling artefacts in the input images.

The same holds true for including rotated versions of training pairs into the dataset. As CNNs are not necessarily rotation-invariant or rotation-equivariant, these properties are usually achieved by properly augmenting the dataset [26]. In contrast, other approaches explicitly incorporate mathematical operations in the design of the network architectures and therewith attempt to reach rotation-invariance or -equivariance [27], [28]. High quality images in cardiac MRI are usually reconstructed by applying iterative methods. Thus, obtaining realistic versions of images rotated by a non-trivial rotation, i.e. by a rotation of $\theta \notin \{\frac{k\pi}{2} : k \in \{0, 1, 2, 3\}\}$, is computationally demanding, as the k-space data has to be rotated and the iterative reconstruction has to be performed on the rotated data. Therefore, rotation-equivariance, in this case, can either be achieved by means of the network architecture design or by a possibly time consuming data-augmentation process.

III. Proposed Approach

In medical imaging, the number of available training samples is usually very small compared to the underlying mathematical dimension of the data, i.e. the number of pixels of an image. Therefore, we are particularly interested in the question of whether or not it is possible to train a CNN on a highly limited dataset by making best use of the given data. We propose to train a CNN employing $2D$ convolutional layers on $2D$ spatio-temporal slices which can be extracted from the cine image sequences over the cardiac cycle. Once the network is trained, the image sequences can be reconstructed by properly reassembling the spatio-temporal slices. Later, we demonstrate that with our proposed approach, already a small number of $2D$ cine MRI datasets suffices to successfully train a network. Furthermore, robustness with respect to rotation in the spatial domain is achieved in a natural way by the change of perspective on the given dataset and our method is therefore almost rotation-equivariant.

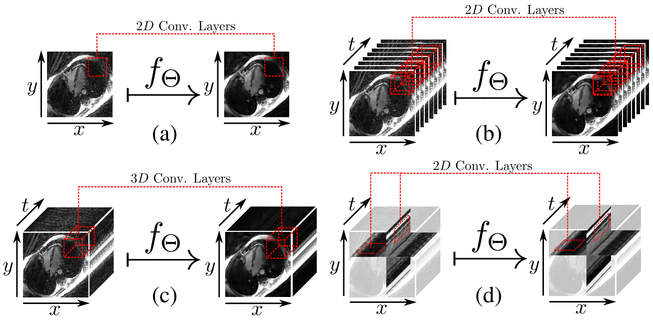


Fig. 1. Different $2D$ and $3D$ Deep Learning-based approaches for undersampling artefacts reduction. $2D$ network for frame-wise mapping (a), $2D$ network for image sequence-wise mapping with cardiac phases aligned as channels (b), $3D$ network for image sequence-wise mapping with three-dimensional convolutional kernels (c), $2D$ network for our proposed approach on two-dimensional spatio-temporal slices (d).

TABLE I
Different Deep Learning-Based Approaches With Their Corresponding Number of Available Training Samples

Approach	Conv. Layers	Available Training Samples
Frame-wise	$2D$	$n \cdot N_z \cdot N_t$
Sequence-wise	$2D$	$n \cdot N_z$
Sequence-wise	$3D$	$n \cdot N_z$
Proposed	$2D$	$n \cdot (N_x + N_y) \cdot N_z$

Consider a dataset of $2D$ cine MR images $\mathcal{D}$ of n subjects, each with N_z slices of size $N_x \times N_y$ and N_t cardiac phases. Figure 1 shows different possible Deep Learning-based methods for removing undersampling artefacts in dynamic MRI sequences. In the first case, the artefacts are removed by training a network f_Θ to map frames to frames, see Figure 1 (a). Given the temporal correlation of adjacent frames, one could also align temporal frames along the channel's axis and apply a network which is trained to map whole image sequences to image sequences, see Figure 1 (b). The same approach can be extended to map image sequences to image sequences but with the network employing three-dimensional convolutional filters, see Figure 1 (c). Our approach exploits spatio-temporal correlation but employs $2D$ convolutional filters which are trained on the spatio-temporal slices of the image sequences, see Figure 1 (d). Table I lists the number of immediately available training samples, i.e. without data augmentation, for the different approaches. Note that with our proposed approach, the number is by far the highest.

A. Persistent Homology Analysis

As a trained denoising autoencoder can geometrically be interpreted to perform a projection-like mapping onto a manifold [29], the study of topological features of the manifold of the input and output images might be of interest for the design of the network architecture, [14], [30]. Persistent homology is a mathematical tool that can be used for analysing datasets $X \subset \mathbb{R}^n$ [31]. For a two-classes classification problem, singular homology has been used as a complexity measure of the positively labelled submanifold of the input space and a

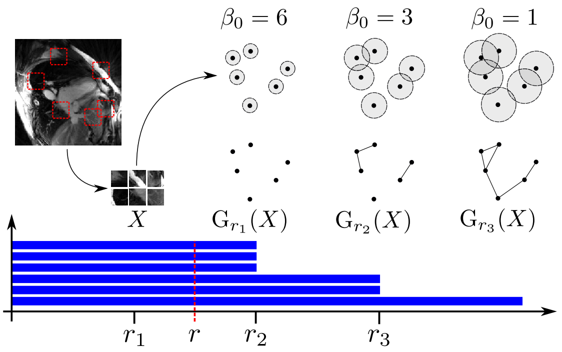


Fig. 2. Procedure of the persistent homology analysis. The image shows an example for six randomly extracted patches of an image in the spatial domain and its corresponding barcode.

relation between this complexity and the depth of the networks was proven in [32]. This and experimental evidence using persistent homology [14], [30], motivates that it might be beneficial to investigate the persistent homology of datasets since it might explain the superiority of specific approaches to others. For a concise introduction to persistent homology see [33], Chapter 1. In general, persistent homology H_* assigns a family of persistence modules $\{H_i(X) : i \in \mathbb{N}\}$ over some field F to a set $X \subset \mathbb{R}^n$, see [33], Chapter 2. We will only use H_0 which has a much simpler interpretation as follows, see Figure 2. Let $X \subset \mathbb{R}^n$ be a finite set and let $r \geq 0$. Then, we can define a graph $\mathrm{G}_r(X)$ with vertices $\mathrm{V}_r(X) = X$ and edges

$$\mathrm{E}_r(X) = \left\{(x, y) \in X^2 : x \neq y \text{ and } \|x - y\|_2 \leq r\right\}.$$

This graph is the Rips complex restricted to simplices of dimension at most 1 [31], Chapter 1.3. Let $\Pi(\mathrm{G}_r(X))$ be the set of connected components of $\mathrm{G}_r(X)$. Then, we can define

$$H_0^r = \bigoplus\nolimits_{i \in \Pi(\mathrm{G}_r(X))} \mathbb{F}_2$$

where $\mathbb{F}_2$ is the field with two elements. For $0 \leq r < r'$ we have a map $\Pi(\mathrm{G}_r(\mathrm{X})) \rightarrow \Pi(\mathrm{G}_{r'}(\mathrm{X}))$ which induces a map $H_0^r \rightarrow H_0^{r'}$. The family of these maps is called the 0-th persistent homology of X. A good visualization of persistent homology is the persistent barcode, see Figure 2. For a real number $r > 0$, the number of connected components of $\mathrm{G}_r(X)$ is equal the number of intersections of the vertical line at $x = r$ with the barcodes, see Figure 2. This is also the 0-th Betti number β_0 of $\mathrm{G}_r(X)$ which is a measure of complexity for $\mathrm{G}_r(X)$, see [31], Chapter 2.3. Hence, the faster the persistent barcode of a dataset X decreases, the less complex the dataset is.

By $\mathbf{x}_I, \mathbf{x}$ and $\mathbf{r}_I := \mathbf{x}_I - \mathbf{x}$ we denote the vector representations of direct reconstruction from undersampled radially acquired data using a non-uniform fast Fourier transform approach (NUFFT), the ground truth reconstruction and the residual, respectively. Since our network reduces artefacts arising from the NUFFT reconstruction as a post-processing step similar to denoising, we operate on the real-valued magnitude images. However, the method can also be applied to complex-valued images by treating real- and imaginary part separately. Note that, in order to keep notation as simple as possible, by abuse of notation, we do not explicitly distinguish between a spatio-temporal slice and a $2D$ frame, but the meaning of the symbols should easily emerge from the context. Therefore, in the spatio-temporal training scenario, $\mathbf{x}_I$ denotes a spatio-temporal slice extracted from an undersampled NUFFT reconstruction, $\mathbf{x}$ its corresponding artefact-free spatio-temporal slice and $\mathbf{r}_I$ its spatio-temporal residual. In the spatial training scenario, $\mathbf{x}_I$, $\mathbf{x}$ and $\mathbf{r}_I$ denote $2D$ frames. In the following, we compare the complexity of the manifolds given by the set of the ground truth images and their residuals in the spatial as well as in the spatio-temporal domain and denote them by $\mathcal{M}_{xy}^{\text{img}}$, $\mathcal{M}_{xy}^{\text{res}}$ and $\mathcal{M}_{xt,yt}^{\text{img}}$, $\mathcal{M}_{xt,yt}^{\text{res}}$. Note that, in contrast to [14], we find ourselves in the situation where spatio-temporal slices and spatial images do not have the same mathematical dimension, and therefore, to be able to compare the manifolds, we restrict our considerations to image patches of the same shape. We performed a persistent homology analysis of the manifold to be learned by using GUDHI [34], [35]. We randomly selected 1400 patches of size 18×18, obtaining a set $X \subset \mathbb{R}^{18^2}$ for which we computed its persistent homology. To be able to compare the persistent barcodes at the same scale, we normalized the patches by the maximal pairwise L_2-distance of points in X. The persistent homology analysis was performed for all patches extracted from the spatio-temporal slices and from spatial image frames by repeating the experiment ten times and averaging the obtained number of connected components for each $r \geq 0$ over the experiments. The corresponding barcode diagrams in Figure 3 (a) and (b) clearly show that in the spatio-temporal domain as well as in the spatial domain, the residual manifolds are more complex than the manifolds of the ground truth images, i.e. the connected components merge at larger scales r. Figure 3 (c) also shows that for the ground truth images, the spatial manifold is more complex than the spatio-temporal manifold which is intuitively clear, as the spatial-temporal slices exhibit the temporal correlation of the sequence. This suggests that a network should achieve the best performance when trained to learn the *ground truth spatio-temporal manifold*. Furthermore, we see that in the case of the spatio-temporal domain, the topological complexity tends to be independent of the number of subjects whose patches are extracted to perform the analysis, see Figure 3 (c) and (d). In contrast, in the spatial domain, a higher number of subjects used to extract the patches slightly reduces the topological complexity of the data. Therefore, we conclude that a small number of $2D$ image sequences may already contain a good representation of all possible two-dimensional spatio-temporal slices and thus, the number of $2D$ image sequences needed to successfully train a network in the spatio-temporal domain should be lower than for training the network in the spatial domain.

B. Network Architecture

In the following, we always refer to Θ as the set of trainable parameters of a network and denote a U-net by u_Θ. Figure 4 shows the single components of a U-net without residual connection, similar as originally proposed in [16]. The network consists of five stages, where each stage is a

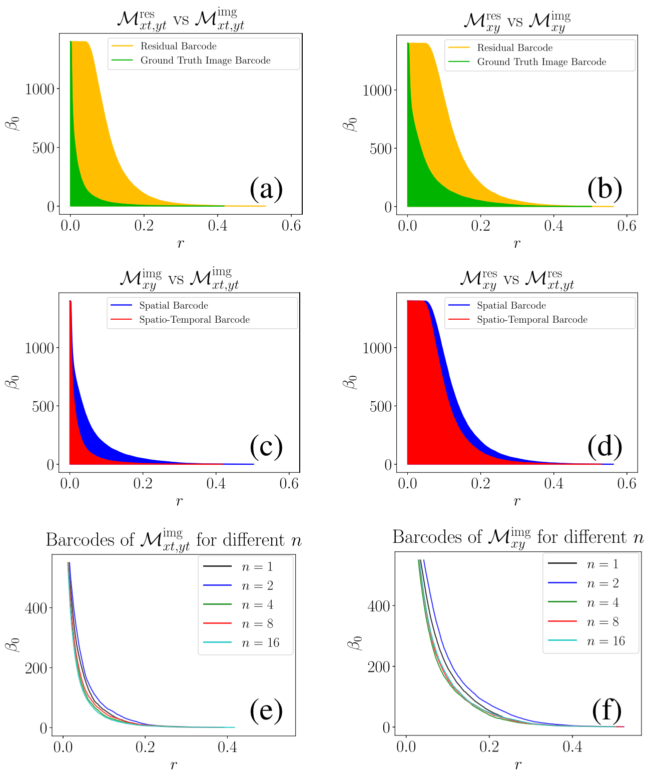


Fig. 3. The number of connected components β_0 of $\mathrm{G}_r(X)$ for different datasets X at different r. Pairwise comparison of the persistent barcodes for $\mathcal{M}^{\text{res}}_{xt,yt}$ and $\mathcal{M}^{\text{img}}_{xt,yt}$ (a), for $\mathcal{M}^{\text{res}}_{xy}$ and $\mathcal{M}^{\text{img}}_{xy}$ (b), for $\mathcal{M}^{\text{img}}_{xy}$ and $\mathcal{M}^{\text{img}}_{xt,yt}$ (c) and for $\mathcal{M}^{\text{res}}_{xy}$ and $\mathcal{M}^{\text{res}}_{xt,yt}$ (d). Persistent codes of $\mathcal{M}^{\text{img}}_{xy}$ and $\mathcal{M}^{\text{img}}_{xt,yt}$ for different n, (e) and (f). For the sake of visibility, in (e) and (f), only the endpoints of the bars are displayed.

block of four convolutional layers with $2D$ filters of shape 3×3, followed by batch-normalization [36] and a component-wise ReLU as activation function. The stages are intercepted by 2×1-max-pooling layers in the encoding phase and by bilinear interpolation layers followed by 3×3 convolutional layers with no activation function in the decoding phase. The initial number of feature maps extracted from the first convolutional layer is set to 64 and is doubled in each block in the encoding phase. The network's output is given by a 1×1-convolutional layer which corresponds to a linear combination of the last extracted feature maps. The replacement of the original 2×2-max-pooling by a contraction solely along the spatial dimension empirically turned out to deliver superior results. The black arrows in Figure 4 denote concatenations between the last and the first layer of the corresponding encoding and decoding phases.

Recall from Figure 3 in Section III-A that the manifolds of the ground truth images have a lower topological complexity compared to the manifolds of their corresponding residuals. Therefore, according to [14] and [30], one should train the network to learn the features of the artefact-free images. Note that, if the U-net employs a residual connection as in [10], the output is of the form $\tilde{u}_\Theta(\mathbf{x}_I) = \mathbf{x}_I + u_\Theta(\mathbf{x}_I)$. If $\mathbf{x}$ is used as a label, $\tilde{u}_\Theta$ is trained to learn the residual up to a change of sign, as u_Θ is the only part of the network containing trainable parameters. Therefore, being consistent with [14], [30], [37],

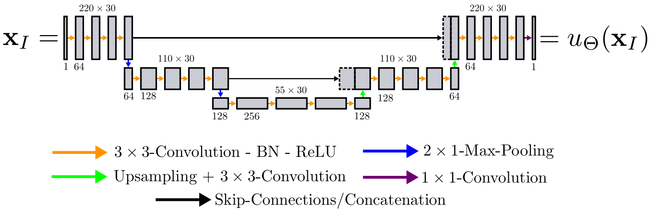

$\mathbf{x}_I =$
$= u_\Theta(\mathbf{x}_I)$
220 × 30
220 × 30
110 × 30
110 × 30
55 × 30
1
64
64
128
128
256
128
128
64
64
1
3 × 3-Convolution - BN - ReLU
2 × 1-Max-Pooling
Upsampling + 3 × 3-Convolution
1 × 1-Convolution
Skip-Connections/Concatenation

Fig. 4. The U-net with three encoding stages and four convolutional layers per stage, no residual connection and batch-normalization (BN). In the case we train on the spatial domain, max-pooling is performed in both spatial dimensions, whereas in our proposed approach max-pooling is solely performed along the spatial dimension without contracting the data along the temporal dimension.

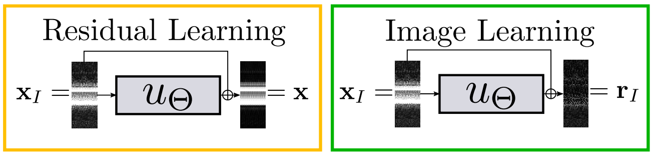


Fig. 5. Residual and Image Learning: For a NN $\tilde{u}_\Theta$ with residual connection, learning the residuals is achieved by using the ground truth images x as labels (left). Learning the ground truth images x is achieved by using the residuals r_I as labels (right).

in order to exploit the simpler topological complexity of the ground truth images and still be able to benefit from the residual connection as in [10], we propose to train a U-net with residual connection to estimate the image residuals $\mathbf{r}_I$ of the spatio-temporal slices. More precisely, if by $\tilde{u}_\Theta$ we denote a U-net with residual connection which is trained to map $\mathbf{x}_I$ to the ground truth residuals $\mathbf{r}_I$, and $\mathbf{r}_{\text{cnn}} = \tilde{u}_\Theta(\mathbf{x}_I) = \mathbf{x}_I + u_\Theta(\mathbf{x}_I) = \mathbf{x}_I - \mathbf{x}_{\text{cnn}}$, then the estimates of the images are obtained by $\mathbf{x}_I - \mathbf{r}_{\text{cnn}} = \mathbf{x}_I - (\mathbf{x}_I - \mathbf{x}_{\text{cnn}}) = \mathbf{x}_{\text{cnn}} \approx \mathbf{x}$.

Figure 5 shows different approaches for training a U-net to remove undersampling artefacts by training on spatio-temporal slices. Note that, using $\mathbf{x}$ as labels for training a U-net *with* residual connection and using the residuals $\mathbf{r}_I$ as labels for training a U-net *without* residual connection is equivalent in the sense that the trainable parameters are fitted to learn the residuals $\mathbf{r}_I$. On the other hand, if we want the network to learn the artefact-free images, we can either use the $\mathbf{x}$ as labels and *not employ* a residual connection or use the residuals $\mathbf{r}_I$ as labels and *employ* a residual connection. This holds for training the network on two-dimensional frames as well as on two-dimensional spatio-temporal slices.

By u^{res}_{xy} and u^{img}_{xy} we denote spatial U-net models when trained to learn the spatial residual manifold $\mathcal{M}^{\text{res}}_{xy}$ and the spatial ground truth image manifold $\mathcal{M}^{\text{img}}_{xy}$, respectively. Analogously, we identify $u^{\text{res}}_{xt,yt}$ and $u^{\text{img}}_{xt,yt}$ as spatio-temporally trained U-nets trained to learn the spatio-temporal manifolds $\mathcal{M}^{\text{res}}_{xt,yt}$ and $\mathcal{M}^{\text{img}}_{xt,yt}$, respectively.

C. Loss Function

Dependent on what we want the network to learn, we train the network architecture to minimize different loss functions. Let $\mathcal{D}^{\text{res}}_{xy}, \mathcal{D}^{\text{img}}_{xy}$ and $\mathcal{D}^{\text{res}}_{xt,yt}, \mathcal{D}^{\text{img}}_{xt,yt}$ denote the set of available training samples, i.e. the pairs $(\mathbf{x}_I, \mathbf{r}_I)$ or $(\mathbf{x}_I, \mathbf{x})$, depending on the domain the data is considered in and on which labels are used for training. By N_{xy} and $N_{xt,yt}$ we denote their corresponding cardinality. Recall that we use the U-net $\tilde{u}_\Theta$ to estimate the residual $\mathbf{r}_I = \mathbf{x}_I - \mathbf{x}$ and therefore, the image estimate is given by $\mathbf{x}_{\text{cnn}} = \mathbf{x}_I - \tilde{u}_\Theta(\mathbf{x}_I)$. Therefore, in order to define the loss function for a network with residual connection to learn the ground truth images, we use the residuals as labels and vice versa. The models u_{xy}^{res} and u_{xy}^{img} are trained by minimizing the L_2-errors between the predicted $2D$ frames and their corresponding labels which are given by

$$\mathcal{L}_{xy}^{\text{res}}(\Theta) = \frac{1}{N_{xy}} \sum_{(\mathbf{x}_I,\mathbf{x})\in\mathcal{D}_{xy}^{\text{img}}} \|\tilde{u}_\Theta(\mathbf{x}_I) - \mathbf{x}\|_2^2,$$
$$\mathcal{L}_{xy}^{\text{img}}(\Theta) = \frac{1}{N_{xy}} \sum_{(\mathbf{x}_I,\mathbf{r}_I)\in\mathcal{D}_{xy}^{\text{res}}} \|\tilde{u}_\Theta(\mathbf{x}_I) - \mathbf{r}_I\|_2^2, \qquad (3)$$

respectively. In the spatio-temporal case, the models $u_{xt,yt}^{\text{res}}$ and $u_{xt,yt}^{\text{img}}$ are analogously trained by minimizing the loss functions

$$\mathcal{L}_{xt,yt}^{\text{res}}(\Theta) = \frac{1}{N_{xt,yt}} \sum_{(\mathbf{x}_I,\mathbf{x})\in\mathcal{D}_{xt,yt}^{\text{img}}} \|\tilde{u}_\Theta(\mathbf{x}_I) - \mathbf{x}\|_2^2,$$
$$\mathcal{L}_{xt,yt}^{\text{img}}(\Theta) = \frac{1}{N_{xt,yt}} \sum_{(\mathbf{x}_I,\mathbf{r}_I)\in\mathcal{D}_{xt,yt}^{\text{res}}} \|\tilde{u}_\Theta(\mathbf{x}_I) - \mathbf{r}_I\|_2^2. \qquad (4)$$

IV. In-Vivo Experiments

A. Data Acquisition

In the following experiments we evaluate the proposed approach on $2D$ Golden radial cine MRI images of 19 subjects (15 healthy volunteers + 4 patients) obtained with a bSSFP sequence on a 1.5T MR scanner (Achieva, Philips Healthcare, Best, The Netherlands) during a 10 s breathhold (TR/TE = 3.0/1.5 ms, FA 60°). The spatial dimensions are $N_x \times N_y = 320 \times 320$ with an in plane resolution of 2 mm and 8 mm slice thickness. The number of cardiac phases which were reconstructed based on ECG signal is $N_t = 30$. Coil sensitivity information was used to combine the image data of each coil after NUFFT-reconstruction. No further normalization was applied to the image data. The reference images used as ground truth images in the data were reconstructed with kt-SENSE [3] using $N_\theta = 3400$ spokes, which already corresponds to an undersampling factor of ~ 3 in each cine image. In addition, dynamic images with $N_\theta = 1130$ (3.4 s scan time) were reconstructed using standard gridding (NUFFT), leading to an undersampling factor of ~ 9. For each of the 15 healthy volunteers and two patients, $N_z = 12$ slices were acquired while for two patients, only $N_z = 6$ slices were obtained due to limited breathhold capabilities. Note that, in contrast to the healthy volunteers, the patients data contains images where the heart movement dysfunction can be diagnosed provided that the temporal information is enough accurate.

B. Evaluation Metrics

The performance of our method was evaluated in terms of peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [38] and Haar-Wavelet based perceptual similarity index measure (HPSI) [39] as similarity measures and normalized root mean squared error (NRMSE) as error-measure. Note that HPSI has been reported to achieve higher correlation with human opinion scores on different benchmark databases than SSIM [39]. The quantitative measures are reported for the two-dimensional frames as well as for the two-dimensional spatio-temporal slices after the image sequences were cropped to 160 × 160 × 30 in order to compute the statistics over the regions of interest of the images.

C. Training Set-Up

Due to our relatively small dataset, all the following experiments were performed in a four-fold cross-validation setting. We split our dataset in portions of 12/3/4 subjects for training/validation/test data, where for one of these configurations, the test data corresponds to the image data coming from patients with heart movement dysfunction. Obviously, the resulting number of training samples in the spatio-temporal domain is much higher than in the spatial case and therefore, for a fair comparison of the methods, we train the networks by keeping the number of backpropagations fixed. Dependent on the perspective on the dataset, this results in a different number of epochs the networks are trained for. For databalance reasons, we crop the image sequences using a cutoff of 50 pixels in x- and y direction. Therefore, the spatial dimensions per frame reduce to 220 × 220. Due to the relatively small number of temporal frames and the large receptive field of the U-net, we also conducted experiments evaluating the performance of the networks trained on spatio-temporal slices by mirroring the boundaries. However, as we did not experience any increase or decrease of performance in explicitly handling the boundary conditions, we conducted all experiments on spatio-temporal slices of shape 220 × 30. The convolutional layers use zero-padding in order to maintain the spatial shape of the samples constant over each stage. Given a U-net as displayed in Figure 4, we are able to use a mini-batch size of 44 when training in the spatio-temporal domain. Thus, we set the mini-batch size in the spatial training case to 6 in order to have a constant number of pixels which the networks are fed with per forward pass, i.e. $44 \cdot 220 \cdot 30 = 290\,400 = 6 \cdot 220 \cdot 220$. The networks are trained for $5 \cdot 10^4$ backpropagations by stochastic gradient descent (SGD) using a learning rate which was gradually decreased from 10^{-5} to 10^{-7} and from 10^{-6} to 10^{-8} for the training in the spatio-temporal domain and in the spatial domain, respectively. The learning rates were chosen in a prior parameter study on the validation set.

D. Residual Vs. Image Learning

Here we compare the performance of the spatial U-net models u_{xy}^{res} and u_{xy}^{img} and our spatio-temporal approaches $u_{xt,yt}^{\text{res}}$ and u_{xy}^{img}. The models were trained by minimizing the loss functions defined in (3) and (4), respectively. Figure 6 shows qualitative results for different possibilities of training illustrated in Figure 5. We see that in both domains, consistent with the previously shown persistent homology analysis, the networks removed the artefacts at their best when they were trained to learn the artefact-free images. From Figure 6 we also already see the superiority of our

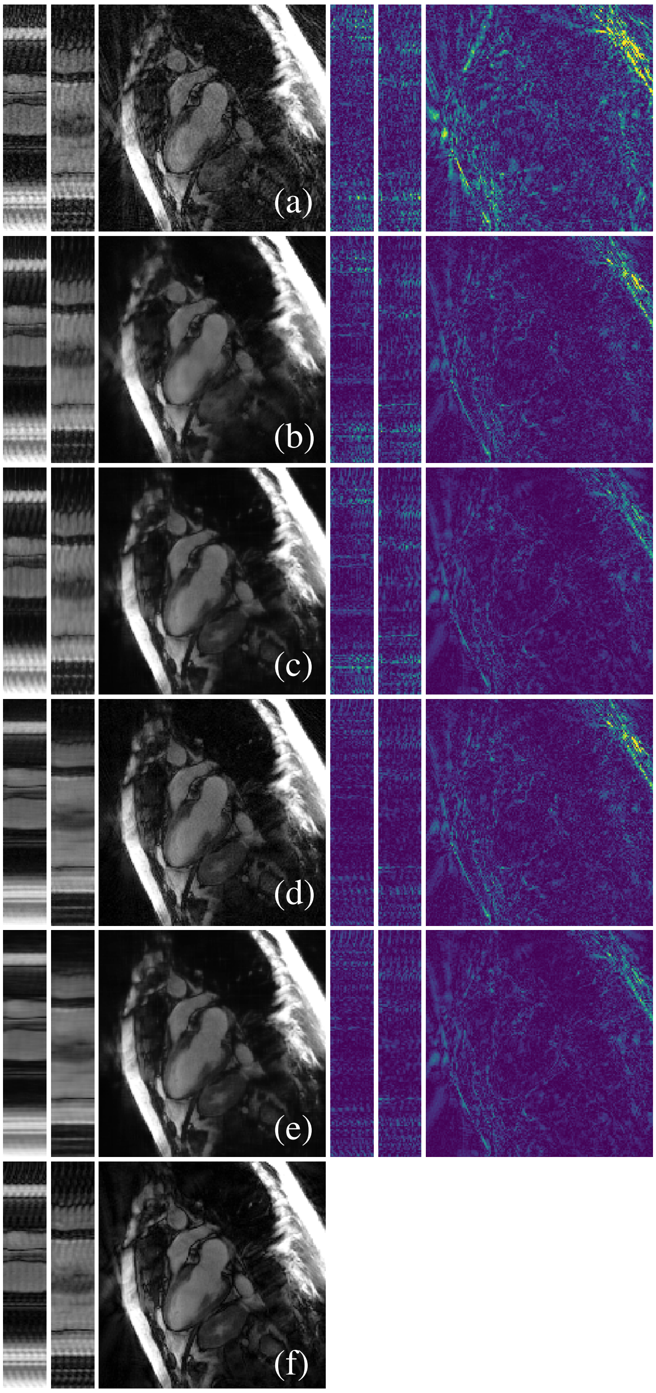


Fig. 6. Comparison of different training approaches for U-nets with residual connection. NUFFT reconstruction with $N_\theta = 1130$ radial lines (a), spatially trained U-nets u_{xy}^{res} (b) and u_{xy}^{img} (c), proposed spatio-temporal approaches $u_{xt,yt}^{\text{res}}$ (d) and $u_{xt,yt}^{\text{img}}$ (e), ground truth (f). The point-wise error images are magnified by a factor of ×3. All images are displayed on the same scale.

approach, see (d) and (e), compared to the spatially trained U-net which slightly tends to smooth out image details and less accurately removed artefacts in spatio-temporal domain, see (b) and (c). Table II shows the results obtained for the spatial U-nets u_{xy}^{res} and u_{xy}^{img} and the spatio-temporal U-nets $u_{xt,yt}^{\text{res}}$ and $u_{xt,yt}^{\text{img}}$ for $n = 12$, which confirms the heuristics given in Section III-A. Note that for the experiment, no data-augmentation was used and therefore, the results differ from the ones reported in Table IV. As a result, we conclude that for the task of removing undersampling artefacts or image denoising, the relation between the topological complexity of the residuals and the fully-sampled image reconstructions can be used to determine which labels to train the network on as well as how to design the network architecture. Since the radial acquisition is designed to be incoherent along the temporal dimension, in all our following experiments we use the U-net architecture as shown in Figure 4 where we make use of the residuals as labels and employ a residual connection as shown in Figure 5 for the case of image learning. In the next Subsection, we also see how learning the manifold $\mathcal{M}_{xt,yt}^{\text{img}}$ can reduce the training time as convergence of the training and validation errors is achieved faster.

TABLE II
PERFORMANCE FOR THE SPATIAL AND OUR SPATIO-TEMPORAL APPROACHES DEPENDENT ON THE USED ARCHITECTURES

	u_{xy}^{res}	u_{xy}^{img}	$u_{xt,yt}^{\text{res}}$	$u_{xt,yt}^{\text{img}}$
	Statistics on $2D$ Frames			
PSNR	34.120	34.715	37.291	37.833
SSIM	0.876	0.897	0.928	0.935
HPSI	0.968	0.979	0.992	0.994
NRMSE	0.151	0.149	0.106	0.105
	Statistics on $2D$ Spatio-Temporal Slices			
PSNR	26.436	26.420	29.422	29.949
SSIM	0.735	0.735	0.795	0.804
HPSI	0.983	0.985	0.992	0.994
NRMSE	0.212	0.209	0.160	0.159

E. Training With Limited Amount of Data

Here we demonstrate the performance of our proposed approach when we restrict the number of available training samples. For this purpose, we trained the same network on different datasets where we fixed a different number of subjects n whose images we included in the training dataset. We show that with our proposed approach we are able to obtain comparable results even with a small number of subjects. Note how in the spatial training scenario, the given training data is naturally constrained by the fact that for a fixed slice, different time frames of the ground truth images exhibit a high similarity. Therefore, regardless of the fact that in the spatial domain the ground truth image manifold has a lower complexity than the residual manifold, a network which is trained to learn the ground truth images should be expected to suffer from the limited variability of the data. In contrast, due to the temporal incoherence of the undersampling pattern, this issue should be overcome when learning the residuals. In the spatio-temporal domain, the availability of the data is not an issue as we have $n\, N_z(N_x + N_y) \gg n\, N_z\, N_t$ samples. Therefore, one would expect the performance of the network to be to some extent independent of the number of subjects n the samples are extracted from. Also, according to the performed persistent homology analysis, the training of the network

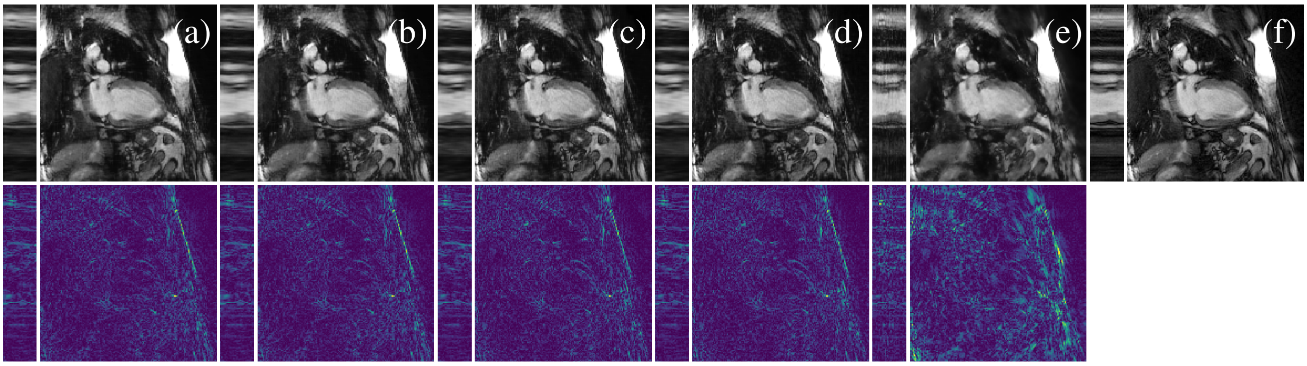


Fig. 7. Results on the test set for $N_\theta = 1130$ radial lines when the number of subjects whose spatio-temporal slices are extracted was varied. Note that no data-augmentation was used. Proposed method for $n = 1$ (a), $n = 2$ (b), $n = 8$ (c), $n = 12$ (d), the spatial U-net for $n = 12$ (e) and the *kt*-SENSE reconstruction with 3400 radial lines (f). The point-wise error images are magnified by a factor of ×3. All images are displayed on the same scale.

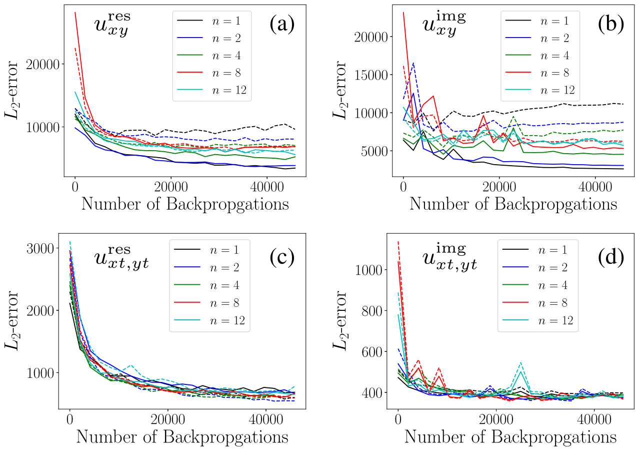


Fig. 8. Loss behaviour during training with $N_\theta = 1130$ for different number of volunteers n contained in the dataset. Training loss (solid) and validation loss (dashed) for the spatial and spatio-temporal U-nets. Spatial residual learning (a), spatial image learning (b), spatio-temporal residual learning (c), spatio-temporal image learning (d). Note that the scales differ due to the different losses and the different domains in which the networks are trained.

should be facilitated when trained to learn the manifold of the ground truth images.

Figure 8 shows the behaviour of the loss decay for the spatial approach ((a) and (b)), the spatio-temporal training approach ((c) and (d)), and in both cases, for the situation where the residuals are learned ((a) and (c)) and where the ground truth images are learned ((b) and (d)). We see that for the spatial U-net, for the residual learning and the image learning, increasing the number of subjects n leads to a decrease of the gap between training and validation error. Further, we see that the gaps are larger in the case where the ground truth images are learned which can be related to the low variability of the dataset. In both cases, for $n = 12$ the gap is small enough to assume that the networks have been properly trained and generalize well. For $n = 1$ and for $n = 1, 2, 4$, the spatially trained U-nets u_{xy}^{res} and u_{xy}^{img} poorly generalize in both training scenarios, as the networks almost immediately start to overfit the data, see (a) and (b). Spatial training of the networks without data-augmentation is possible for $n = 2, 4, 8, 13$ for the residual learning and for $n = 8, 13$ for the image learning. However, our method outperforms the spatially trained U-net as it better maintains diagnostic details in spatial and spatio-temporal domain, see Figure 7 for the case $n = 12$. For the spatio-temporal approaches, the gaps between training and validation error are smaller compared to the ones for the spatial approaches. This holds for the residual learning as well as the image learning scenario. Further, when the network is trained to learn the ground truth images, the errors converge faster than in the residual training approach, compare Figure 8 (c) and (d). Also, the convergence rate is highly independent on the number of subjects n. From these experiments, we first conclude that our proposed method is well suited for training a network on a limited number of subjects. Second, forcing the network to learn the manifold given by the ground truth images $\mathcal{M}_{xt,yt}^{\text{img}}$ facilitates the training, which leads to a faster convergence of the errors and therefore to lower training times. Figure 7 shows a slice of the output of an image in the test set which was obtained with our proposed method. For all n, the artefacts have been successfully removed. We also see that even for $n = 1$, the dataset is already rich enough in order to allow for a good depiction of cardiac contraction and expansion during the heart cycle. Table III shows the achieved average of the quantitative measures. Even if in terms of quantitative measures the network performs better the larger the training data, the differences are marginal and hardly perceivable by the human eye, see Figure 7. Therefore, we conclude that since the data has a particularly simple structure, little data is already sufficient for a successful training.

F. Rotation Equivariance

CNNs are well known to be able to achieve properties as translation-invariance and -equivariance [40]. However, they are not naturally invariant or equivariant with respect to rotation and one of the still most used methods to achieve these properties is to appropriately augment the dataset, [26], [41]. In contrast, other approaches [27], [28], [42] explicitly incorporate invariant/equivariant convolutional operations in the networks which comes at the cost of a more complex network design. As a rotation in image space, i.e. due to a rotation of the field of view in order to adapt the scan to the geometry

TABLE III
RESULTS ON THE TEST WHEN THE NUMBER OF SUBJECTS WHOSE IMAGES WERE INCLUDED IN THE TRAINING SET IS VARIED

	$n=1$	$n=2$	$n=4$	$n=8$	$n=12$
	Statistics on $2D$ Frames				
PSNR	37.245	37.785	37.659	37.845	37.833
SSIM	0.931	0.934	0.934	0.934	0.935
HPSI	0.994	0.994	0.994	0.994	0.994
NRMSE	0.109	0.106	0.107	0.105	0.105
	Statistics on $2D$ Spatio-Temporal Slices				
PSNR	29.584	29.901	29.774	29.952	29.949
SSIM	0.793	0.801	0.802	0.803	0.804
HPSI	0.993	0.994	0.994	0.994	0.994
NRMSE	0.160	0.160	0.162	0.161	0.159

of the patient's heart, might easily be encountered, we are interested in achieving rotation-equivariance, i.e. $f_\Theta(\psi(\mathbf{x}_I)) = \psi(f_\Theta(\mathbf{x}_I))$ for an already trained network f_Θ and rotation ψ in the xy-plane. For the following experiment, we generated new different test sets $\mathcal{D}^{\psi_\theta}_{xy}$ and $\mathcal{D}^{\psi_\theta}_{xt,yt}$ by applying rotations ψ_θ with rotation-angle θ and tested the networks which were previously trained on the non-rotated images on the different test sets. By doing so, we were able to isolate and measure the direct effect of the sole rotation in image space on the performance of the network.

We rotated the measured data in k-space and reconstructed the training set for different angles θ. Note that the process is time consuming since the images were reconstructed with kt-SENSE. Therefore, we only reconstructed rotated images for $\theta = \pm 66°, \pm 33°$ and for each θ we further rotated the frames by $\pm 90°$ and $180°$, obtaining an overall number of 19 rotated test sets. Figure 9 compares our approach to the $2D$ spatially trained U-net in terms of quantitative measures calculated over the $2D$ frames of the different test sets with different rotation angles. For $\theta = 0$, the measures indicate the average measure achieved on the training set. First, we see again that the spatio-temporal training approach clearly outperforms the spatial training approach in terms of all quantitative measures. Further, while rotating the $2D$ frames yields a noticeable decrease of performance of the network trained in the spatial domain, the network trained on the spatio-temporal slices performs similarly well on the different rotated test sets and is therefore almost rotation-equivariant.

G. Experiments With Shallower Networks

Even if we used the network architecture shown in Figure 4 for all experiments, the strength of the method lies in the change of perspective on the data. To demonstrate this, we applied different network architectures following our suggested approach. More precisely, we tested different types of CNNs which can be seen as special cases of the U-net. If by E and C we denote the numbers of encoding stages and convolutional layers per stage of a U-net, E3 C4 corresponds to the network displayed in Figure 4. E1 C8, on the other hand, denotes a single-scale fully CNN with eight convolutional layers and no max-pooling. Figure 10 shows results obtained with different network architectures parametrized by E and C. We see that the networks E1 C8 and E4 C4 which differ in terms of number of trainable parameters by approximately a factor of 10, achieve similar performance. This suggests that the number of trainable parameters and consequently, also training times, could further be reduced without significantly losing performance. Figure 10 shows results obtained by E1 C8 (a), E4 C4 (b) and E5 C2 (d), where the networks were trained for $3 \cdot 10^4$ backpropagations. The training of E1 C8, for example, see Figure 10 (a), amounted to only 40 minutes.

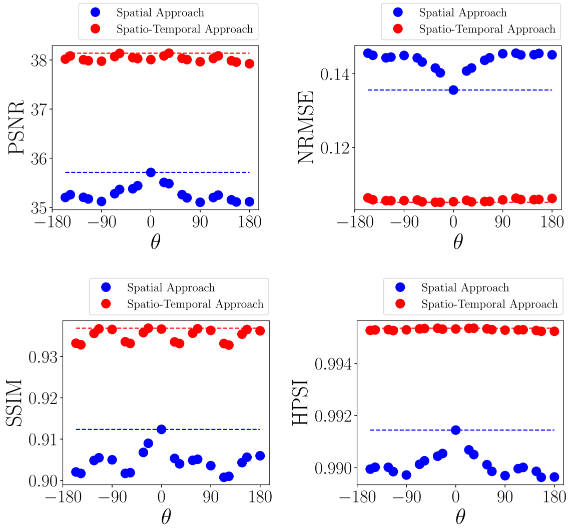


Fig. 9. Performance of the networks when tested on rotated copies of the images contained in the training set. While the network trained in the spatio-temporal domain is robust with respect to rotation, the network trained on images in the spatial domain loses generalization power when tested on rotated copies of the images it was trained on. The dashed lines correspond to the corresponding measure achieved on the training set.

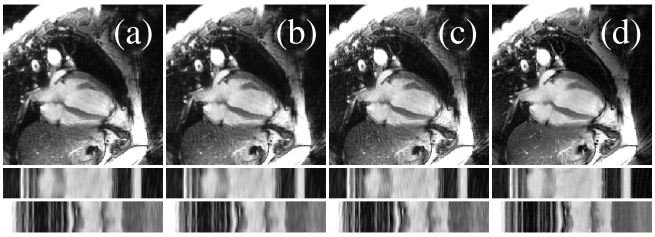


Fig. 10. Results obtained with different CNNs following our proposed aproach $u^{\text{img}}_{xt,yt}$. E1 C8 (a), E4 C4 (b) and E5 C2 (c), *kt*-SENSE reconstruction with $N_\theta = 3400$ radial lines (d). Our approach therefore offers the possibility to further reduce the network complexity as well as training times.

H. Comparison With Other Deep Learning-Based Methods

Here we compare our approach to other methods based on post-processing with deep NNs. Since we only have access to a limited dataset, for the following experiments, we made use of data-augmentation by using all our rotated images, flipping, shifting the images along the channel axis and adding random constant values to the whole image sequences. By doing so, we created a potentially infinite training set. Note that we did not include elastic deformations as a data-augmentation technique, as the data-acquisition process is not simulated and elastic deformations might alter the structure of the undersampling artefacts in the input data. The first

TABLE IV
COMPARISON OF DIFFERENT DEEP LEARNING-BASED POST-PROCESSING APPROACHES

NN Model	u_{xy}^{img}	$u_{xy,t}$	u_{xyt}	$u_{xt,yt}^{\text{img}}$
	Statistics on $2D$ Frames			
PSNR	34.817	33.526	37.827	37.930
SSIM	0.910	0.868	0.935	0.935
HPSI	0.988	0.984	0.994	0.994
NRMSE	0.141	0.172	0.105	0.104
	Statistics on $2D$ Spatio-Temporal Slices			
PSNR	26.959	25.238	29.873	30.048
SSIM	0.740	0.693	0.809	0.804
HPSI	0.990	0.991	0.994	0.994
NRMSE	0.208	0.290	0.165	0.158

method of comparison is the already discussed spatially trained U-net u_{xy}^{img}. It is trained to map frames to frames and corresponds to the method discussed in [10] and [14]. The second method of comparison is a natural extension of the first and corresponds to the $2D$ U-net approach shown in Figure 1 (b) which we refer to as $u_{xy,t}$. The net is trained to map whole image sequences to whole image sequences by aligning the cardiac phases along the channel's axis and was presented in [20]. Further, we compare our method to the $3D$ U-net approach u_{xyt} presented in [21], see Figure 1 (c). While for the $2D$ NNs, we cropped the images to 220×220 and $220 \times 220 \times 30$ in order to let the networks focus on the diagnostic content of the images, for the $3D$ U-net, the images used for training needed to be cropped to $128 \times 128 \times 20$, as the network is computationally more expensive. The shape was the one used in [21]. In order to obtain image sequences of $320 \times 320 \times 30$, the outputs of the networks were treated as patches and the image sequences were reconstructed from the patches by properly averaging over regions with overlapping patches. In contrast to the models employing $2D$ convolutional layers, which were trained using SGD, the $3D$ U-net u_{xyt} was trained in the same setting as suggested in [21] using *ADAM* [43]. Figure 11 and Table IV show and summarize the obtained results with the described networks. For more detailed information about the reassembling of the image sequences from the patches, see Section IV-K.

The spatially trained U-net u_{xy}^{img} correctly removed the undersampling artefacts in the spatial domain. However, the reduction of the artefacts is less accurate than for $u_{xt,yt}^{\text{img}}$, see Figure 11 (b) and (e). Although we report a successful training in terms of consistent decrease of training as well as validation error, the model $u_{xy,t}$ poorly removed the artefacts. Intuitively, the temporal incoherence of the radial undersampling pattern which differs from the one in [20] hinders the learning of the residual manifold and the network is therefore not suitable for our used undersampling scheme. Further, in [20], a zero-filled reconstruction is used as input of the network and therefore, the relation between the manifolds of the residuals and the ground truth images might differ as well from our case. In contrast, learning the manifold of ground truth sequences is highly facilitated by the temporal correlation of the $2D$ frames. In fact, already a network with one single convolutional layer with N_t channels and 64 filters accurately removed all the artefacts from the image sequence. However, temporal information is lost and we point out we were not able to obtain satisfactory results by the application of deeper networks. The $3D$ U-net u_{xyt} and our proposed method $u_{xt,yt}^{\text{img}}$ perform comparably well. Both correctly removed the undersampling artefacts in spatial as well in spatio-temporal domain and led to a good preservation of the heart movement. In terms of the image-error-based PSNR and NRMSE measures, our

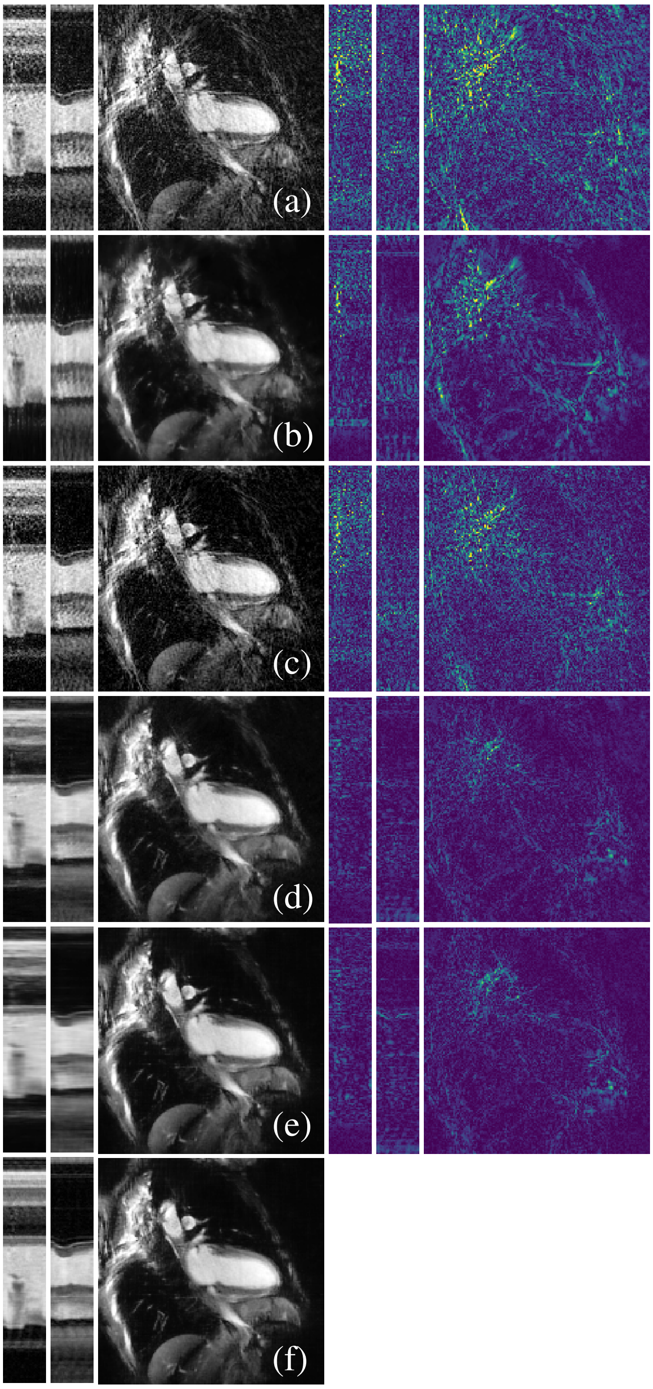


Fig. 11. Comparison with different Deep Learning-based post-processing methods. NUFFT reconstruction with $N_\theta = 1130$ radial lines (a), u_{xy}^{img} (b), $u_{xy,t}$ (c), u_{xyt} (d), proposed approach u_{xy}^{img} (e), ground truth *kt*-SENSE reconstruction (f). The point-wise error images are magnified by a factor of $\times 3$. All images are displayed on the same scale.

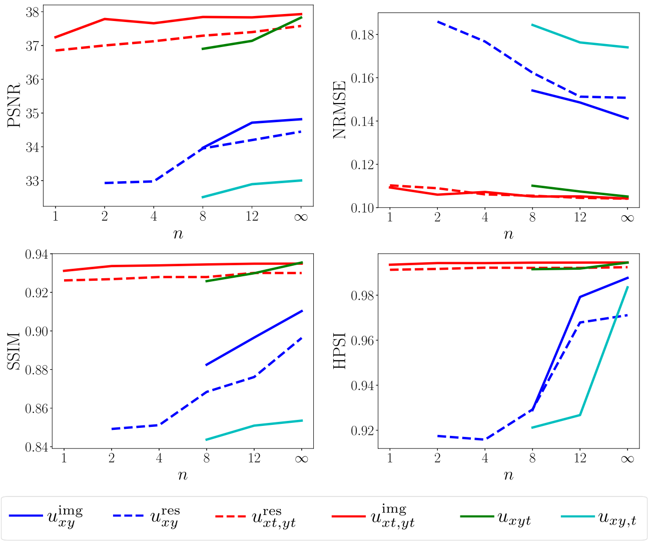


Fig. 12. Quantitative measures for all discussed Deep Learning-based post-processing methods when trained on datasets including different number of subjects *n*. Missing values for some *n* denote that the network was not properly trainable on the restricted dataset.

method performs slightly better than the $3D$ U-net u_{xyt} which, on the other hand, yields slightly better results in terms of SSIM and HPSI. However, the differences are marginal and barely visible. Further, note how our proposed method achieves similar results as the $3D$ U-net u_{xyt} even when trained on one single patient, see Table III. Figure 12 shows the statistics calculated on the $2D$ frames for all different discussed Deep Learning-based post-processing approaches where the number of subjects n contained in the training dataset was varied. The case $n = \infty$ corresponds to $n = 12$ with all previously mentioned data-augmentation techniques. Clearly, our proposed method of training on the $2D$ spatio-temporal slices is the most suitable for obtaining satisfactory results when training a network on a highly limited dataset. The models $u_{xt,yt}^{\mathrm{img}}$ and $u_{xt,yt}^{\mathrm{res}}$ are the only ones to allow the successful training of a network on data extracted from one single subject. For u_{xy}^{img} and u_{xy}^{res}, the results obtained for $n = 2$ and $n = 4$ were obtained by early stopping due to early overfitting. The models $u_{xy,t}$ and u_{xyt} are properly trainable starting from $n = 8$. The $3D$ U-net u_{xyt} and our method $u_{xt,yt}^{\mathrm{img}}$ achieve comparable performance in terms of the reported measures for $n = \infty$.

I. Comparison With State-of-the-Art Iterative Reconstruction Methods

Here, we compare our proposed approach to established state-of-the-art iterative reconstruction methods for cine cardiac MRI. Since iterative reconstruction methods are time consuming, we only reconstructed images from the patients' data which corresponds to one training/validation/testing setting of our four-fold cross-validation set-up. For comparison, images were reconstructed with kt-FOCUSS, a CS-based approach [7], an iterative reconstruction approach using spatial and temporal total variation (TV+TVT) for regularization [4] and a method employing regularization based on learned spatio-temporal dictionaries as well as spatial and total variation minimization (DL+TV) [38]. The latter method was extended by combining the approach proposed in [38] with [8] by learning the dictionaries jointly from the real and imaginary part of the image data. Further, we extended the method to be applicable to multi-coil datasets. We implemented the method using the operator discretization library (ODL) [44] for all needed operators.

TABLE V
COMPARISON WITH DIFFERENT ITERATIVE RECONSTRUCTION METHODS

Reconstruction	kt-**FOCUSS**	**TV+TVT**	**DL+TV**	$u_{xt,yt}^{\mathrm{img}}$
	Statistics on $2D$ **Frames**			
PSNR	31.231	34.794	35.154	37.572
SSIM	0.887	0.916	0.932	0.933
HPSI	0.966	0.987	0.980	0.992
NRMSE	0.213	0.140	0.132	0.102
	Statistics on $2D$ **Spatio-Temporal Slices**			
PSNR	24.493	27.092	27.942	29.554
SSIM	0.735	0.786	0.836	0.800
HPSI	0.970	0.989	0.978	0.992
NRMSE	0.257	0.203	0.171	0.158

Figure 13 shows examples of the results obtained on the patients' data for the mentioned iterative reconstruction methods and our proposed model $u_{xt,yt}^{\mathrm{img}}$. Although our method was trained on healthy volunteers, pathological heart wall motion (septal flash in Figure 13 (a)-(e) and hypo-kinetic anterior and posterior wall with strongly reduced ejection fraction in Figure 13 (f) - (j)) is clearly visible with the proposed method. Also small features, such as the chordae tendinae connecting the valves and the papillary muscles, are well preserved, see Figure 13 (i). Table V shows the obtained results with the iterative reconstruction methods as well as with our proposed network $u_{xt,yt}^{\mathrm{img}}$. We see that our method clearly outperforms the methods kt-FOCUSS and TV+TVT with respect to all reported quantitative measures. The most significant increase of performance is achieved against kt-FOCUSS, where, on the $2D$ frames, our method yields an increase of approximately 6 dB, 4.9% and 2% in terms of PSNR, SSIM and HPSI. Further, our proposed method's NRMSE is approximately half of the one of kt-FOCUSS. TV+TVT surpasses kt-FOCUSS in terms of all reported measures. Even if DL+TV surpasses TV+TVT with respect to all reported measures but HPSI, DL+TV tends to slightly smooth image details, possibly caused by a too strong regularization as well as the smoothing effect of the average of the reconstruction from patches. Further, note that the complex-valued patches were obtained by a disjoint sparse coding of the real and imaginary part of the patches as in [8]. Our method $u_{xt,yt}^{\mathrm{img}}$ outperforms DL+TV with respect to all reported measures except for SSIM on the spatio-temporal slices. Note that the reconstruction time for DL+TV is higher than for our method by several orders of magnitude, see Section IV-K.

J. Comparison With State-of-the-Art Cascaded Networks

For the sake of completeness, we compare our method to the two state-of-the-art methods for $2D$ cine MRI based on cascaded networks presented in [19] and [24]. Cascaded

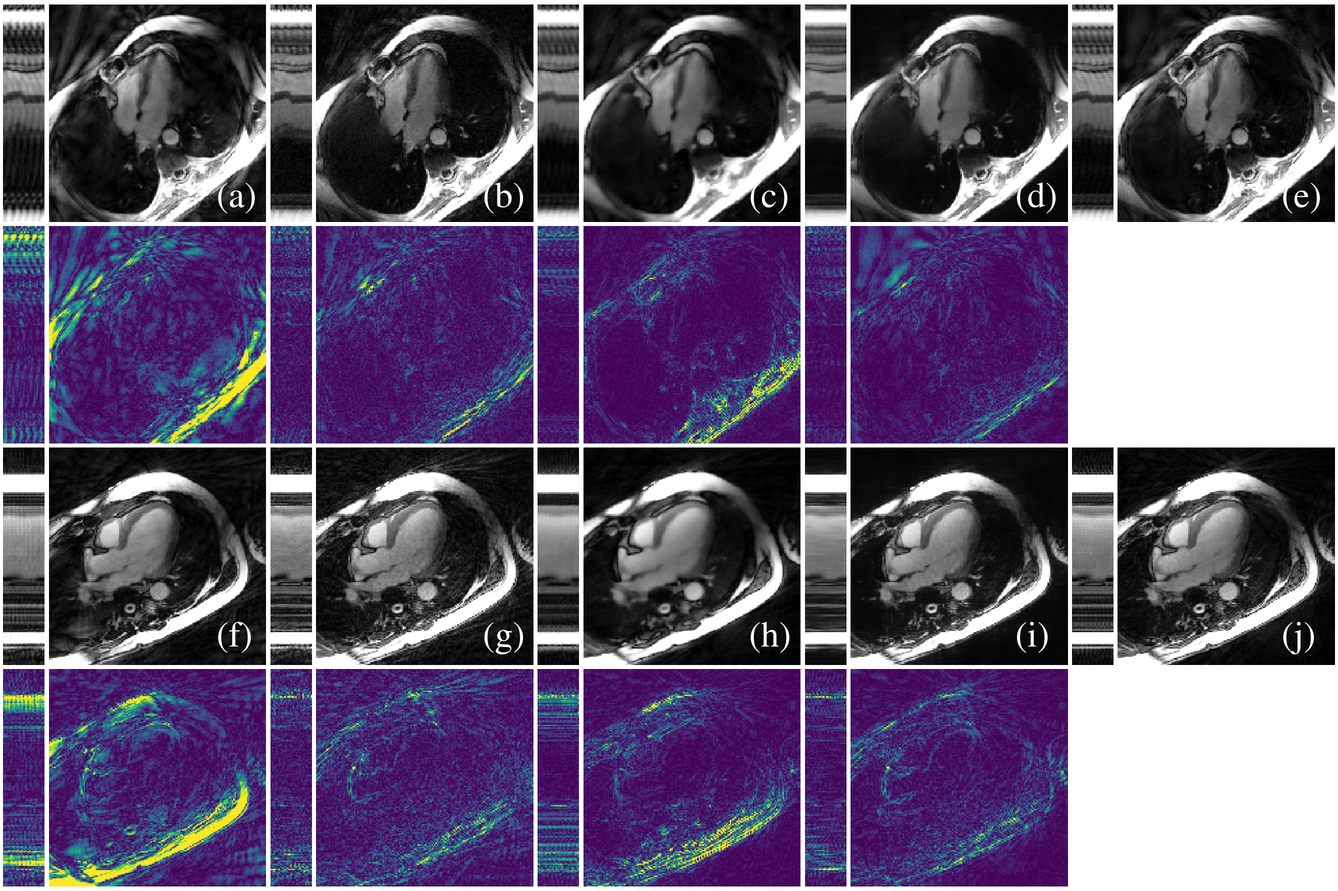


Fig. 13. Comparison with different state-of-the-art iterative reconstruction methods. *kt*-FOCUSS (a) and (f), TV+TVT (b) and (g), DL+TV (c) and (h), proposed method (d) and (i), *kt*-SENSE reconstruction with $N_\theta = 3400$ radial lines. The point-wise error images are magnified by a factor of ×3. All images are displayed on the same scale.

networks combine iterative reconstruction methods and NNs in the sense that they can be interpreted as unrolled iterative schemes where the networks play the role of regularizers learned from data [12], [45]–[47]. While the NNs remove the artefacts from the undersampled image reconstructions, the data-consistency (DC) layers ensure that the outputs provided by the single networks match the measured data in k-space domain. In [19], the used NNs are $3D$ CNNs, while in [24], the $3D$ CNNs are replaced by $2D$ recurrent CNNs. For the comparison, we used the codes available in [19] and [24]. Note that the main underlying assumption for cascaded networks is that the forward and adjoint operators can be integrated in the network architecture. For our data, the forward operator is given by a NUFFT encoding operator which measures k-space data from $n_c = 12$ coils. Since building a deep cascade of CNNs is not possible by including our operator in the DC layers, we trained the networks on the image and k-space data for each coil separately. The final image estimates were then obtained by combining the images from the single coils using coil sensitivity information. Table VI summarizes the results of the cascaded networks. The $3D$ CNN cascade approach yields slightly better image quality metrics compared to our approach, most probably due to the integration of the forward and adjoint operators in the DC layers. Note that for this experiment, the input images $\mathbf{x}_I$ were retrospectively simulated from the kt-SENSE reconstructions $\mathbf{x}$ and therefore, the statistics for our approach differ from the ones reported in Tables IV and V, where the images are reconstructed from real k-space data obtained from the scanner. Further, we report that, even if we did not observe overfitting, for the fold where the test set consists of patient data, the cascaded networks show a significant decrease in performance. This might indicate that the networks are more susceptible to possible significant differences between the training and test set data. Figure 14 shows qualitative results for the comparison of the two cascaded networks and our approach. The statistics in Table VI were obtained by averaging the results on the test set for each fold. On each

TABLE VI
COMPARISON WITH DIFFERENT CASCADED CNNS

Reconstruction	$3D$ CNN cascade	$2D$ CRNN cascade	$u^{\mathrm{img}}_{xt,yt}$
	Statistics on $2D$ Frames		
PSNR	41.831	37.945	40.376
SSIM	0.969	0.960	0.954
HPSI	0.989	0.973	0.989
NRMSE	0.068	0.103	0.079
	Statistics on $2D$ Spatio-Temporal Slices		
PSNR	33.779	30.383	32.281
SSIM	0.908	0.885	0.842
HPSI	0.988	0.970	0.985
NRMSE	0.104	0.140	0.126

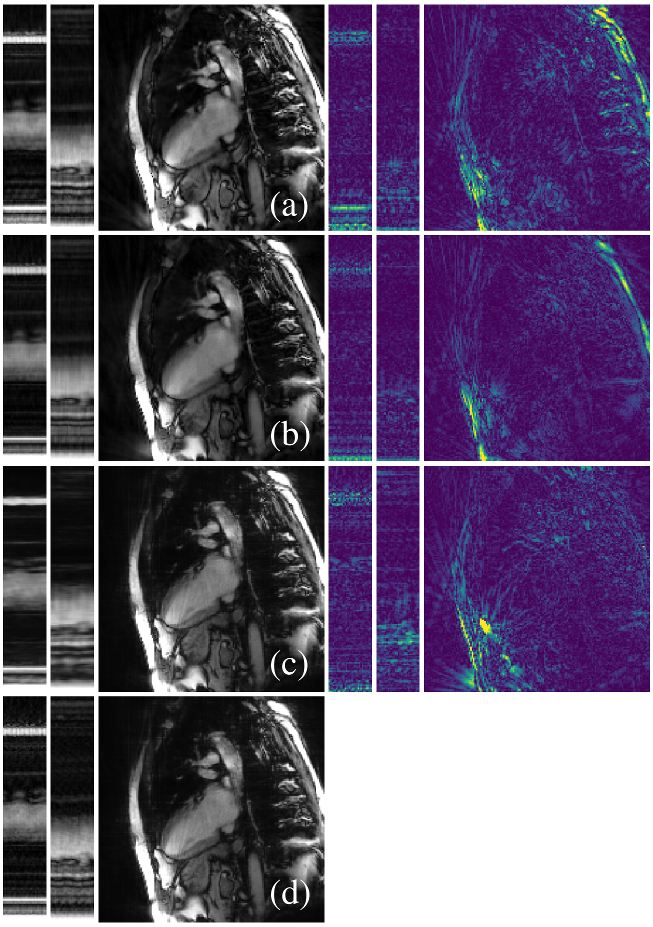


Fig. 14. Comparison with different cascaded CNNs: $2D$ CRNN Cascade (a), $3D$ CNN-Cascade (b), proposed (c) and the reference *kt*-SENSE reconstruction (d). The Figure show results for the fold where only patient's data is included in the test set. Qualitatively, all the three methods perform similarly.

test set, the measures were obtained by testing the networks for which the trainable parameters led to the smallest average error on the whole validation set. The results for the different folds can be found in the supplementary materials which are available in the multimedia tab.

K. Reconstruction Times

We report the reconstruction times needed for the reconstruction of the images with the different previously discussed methods. First, we note that the methods employing iterative reconstruction are the most demanding in terms of computational times. kt-FOCUSS, kt-SENSE and TV+TVT are in the same range, where the reconstruction times per slice vary from approximately 110 s to approximately 180 s. The DL+TV method is by far the most computationally expensive method, as the regularized inverse problem has to be solved for each coil separately. Therefore, the average overall reconstruction time per slice amounts to roughly 13 000 s, where nearly 1 500 s are needed by ITKrM [48] which replaced the computationally heavier K-SVD [49], 7 800 s by the sparse coding with orthogonal matching pursuit, 310 s for the reconstruction from the sparsely approximated patches and 2 058 s for the preconditioned conjugate gradient (PCG) method.

TABLE VII
COMPARISON OF THE RECONSTRUCTION TIMES SLICE

Method	**Reconstruction Time** $[s]$
NUFFT	5
NUFFT + $u_{xy}^{\text{res}}/u_{xy}^{\text{img}}$	$5+7$
NUFFT + $u_{xy,t}$	$5+0.64$
NUFFT + u_{xyt}	$5+5$
NUFFT + $u_{xt,yt}^{\text{res}}/u_{xt,yt}^{\text{img}}$	$5+4.4$
kt-**FOCUSS**	110
kt-**SENSE**	150
TV+TVT	180
DL+TV	13 036
$2D$ **CRNN cascade**	16.8
$3D$ **CNN cascade**	8.8

Note that we trained all the $2D$ U-nets on image sequences which were previously cropped to $220 \times 220 \times 30$. Also, due to memory limits, the shape of the image sequences which are processed by the $3D$ U-net was $128 \times 128 \times 20$. Therefore, for the methods u_{xy}, $u_{xy,t}$ and u_{xyt}, the $320 \times 320 \times 30$ image-sequences were reconstructed from patches. In particular, we used strides of size 25×25 for the spatial and spatio-temporal $2D$ U-nets and strides of $32 \times 32 \times 5$ for the $3D$ U-net, resulting in $5 \cdot 5 \cdot 30 = 750$, $5 \cdot 5 = 25$ and $7 \cdot 7 \cdot 3 = 147$ samples to be processed for the reconstruction of a single slice. For our method $u_{xt,yt}^{\text{img}}$, the strides are 50 (in x- and y direction), resulting in $3 \cdot (220+220)$ samples to be processed per slice. Processing one sample on a Titan Xp GPU takes on average 0.0093 s for u_{xy}^{img} and u_{xy}^{res}, 0.0236 s for $u_{xy,t}$, 0.0340 s for u_{xyt} and 0.0034 s for our proposed approaches $u_{xt,yt}^{\text{img}}$ and $u_{xt,yt}^{\text{res}}$. Table VII summarizes the reconstruction times for a slice of size $320 \times 320 \times 30$ for all the reported methods with the aforementioned strides. The times needed to denoise a slice obviously heavily depend on the number of patches the sequence is reconstructed from and could be easily reduced by using larger strides. For the $2D$ methods, one could also obtain the $320 \times 320 \times 30$ image sequences by directly applying the networks to the $320 \times 320 \times 30$ samples. Note that for the $3D$ U-net this not possible because of memory limits. The training times needed for the $2D$ CRNN cascade and the $3D$ CNN cascade amounted to approximately 1 day and 3 days and 14 hours while processing a single slice and all cardiac phases takes about 16.8 s and 8.8 s, respectively,. Note that the reconstruction of one slice involves the processing of the images of all $n_c = 12$ coils.

V. DISCUSSION AND CONCLUSION

In this work, we have presented a new approach for the task of undersampling artefacts reduction in $2D$ cine MRI. Even if the employed U-net is a widely used network architecture for various inverse problems, to the best of our knowledge, this is the first work in which the U-net is applied to $2D$ spatio-temporal slices. We have investigated and demonstrated several advantages of the approach compared to the training in the spatial domain. Consistent with [14], [30], [37], the performed persistent homology analysis confirms the motivation that the superiority of the proposed approach can be attributed to the simpler topological complexity of the two-dimensional spatio-temporal slices. Further, the analysis suggests that the architecture should be chosen such that the network is trained to learn the ground truth images rather than the residuals. Note that our analysis is consistent with the results presented in [10] and [14], where streaking artefacts resulting from a sparse view CT acquisition are most efficiently removed when U-net learns the residual manifold, which was shown to have a lower complexity than the one of the ground truth images [14]. This is related to the fact that the undersampling pattern in sparse view CT is regular. Conversely, in CS MRI, where the undersampling schemes, e.g. golden-angle radial undersampling, are designed to be incoherent with the assumed sparsifying basis [50], one would expect the residual manifolds to have a more complex topological structure and therefore, the network's architecture should be chosen appropriately. Further investigation of the relation between the topological complexity of the residuals and the artefact-free images in different imaging modalities and the performance of the trained networks will be investigated in the future.

Our approach allows to successfully train a U-net on highly limited data, overcoming the problem of unavailability of large datasets or the need to rely on data-augmentation. We demonstrated that our method already outperforms the spatially trained U-net when trained on one single healthy volunteer in terms of all quantitative measures. When trained on a small number of volunteers, our network is already able to accurately preserve the heart movement and delivers results which are similar to the ones obtained when training on 12 subjects. In contrast to the spatial training approach, the proposed method naturally almost achieves rotation-equivariance by the sole change of perspective on the data. The network does therefore neither require changes in the architecture, nor data-augmentation based on rotation to achieve this property. Clearly, the reason lies in how a rotation in image space results in a transformation similar to a translation in the spatio-temporal domain, and therefore, since the network consists of convolutional and max-pooling layers, it is stable with respect to rotation in image space. Even if the reconstruction of a single slice and all its cardiac phases requires the evaluation of a large number of samples, reconstruction is fast and can be achieved in approximately 4.4 s on a Titan Xp GPU.

As discussed in [17] and [18], the U-net tends to smooth out image details when trained in the spatial domain. In the proposed approach, however, image details in the spatial domain are well preserved. Our method, on the other hand, well preserves image details and further outperforms all other tested $2D$ CNNs with respect to all reported measures and achieves results comparable to the $3D$ U-net even when trained only on two subjects. Due to the small size of the data when considered in spatio-temporal domain, training times could be shortened to 3 hours compared to 6 hours for the $3D$ U-net. Further, since the spatio-temporal manifold $\mathcal{M}^{\text{img}}_{xt,yt}$ has a particularly simple structure, the reducing the artefacts reduces to a simpler task than in the spatial domain and training times could be further reduced by earlier stopping the training.

As for all Deep Learning-based post-processing methods, the main limitation of our proposed method is the possible lack of data-consistency. Even if our method is based on post-processing of the magnitude images, the method could be easily extended to process the real and imaginary part of the spatio-temporal slices separately. Therefore, handling complex-valued data does not represent a limitation and data-consistency could be enforced by for example performing several iterations of PCG for minimizing a properly chosen functional including a data-consistency and regularization term based on the output of our method, see for example [51].

We have compared our proposed method to several state-of-the-art methods for iterative reconstruction in dynamic MRI. Our method outperforms kt-FOCUSS and TV+TVT with respect to all reported measures and achieves similar results as the dictionary learning- and total variation-based method DL+TV. However, our method is faster than DL+TV by several orders of magnitude as it performs a one-step regularization based on an initial NUFFT reconstruction. The iterative reconstruction methods kt-FOCUSS, TV+TVT and DL+TV used for comparison require the tuning of several parameters which were kept fixed for all patients. Therefore, further patient-specific parameter tuning might further improve the image quality in Figure 13 (a), (b), (c), (f), (g) and (h). In particular, DL+TV makes specific parameter tuning difficult due to its prohibitive reconstruction times.

Further, we have compared our method with two state-of-the-art methods based on cascaded CNNs [19], [24] trained on retrospectively simulated data. Although the $3D$ cascaded network's performance is slightly superior to our method, note that for the cascades the input images are zero-filled reconstructions using a Cartesian mask whose support is given by the indices of the k-space coefficients which were interpolated from the radially acquired k-space data. Therefore, the input images for the cascades contain artefacts which are inherently different from the ones obtained by our NUFFT reconstruction using $n_c = 12$ coils and $N_\theta = 1130$ spokes. Also, even if our method only performs subsequent post-processing, the obtained results are qualitatively competitive with the ones obtained by the cascaded networks and we point out that our approach could also be easily extended to be integrated in cascaded networks. This will be subject of future work.

In this work, we used kt-SENSE to obtain the ground truth samples from a 10 s breathhold. Although this yielded high image quality, residual undersampling artefacts which might impair the trained U-net might still be visible. Also, kt-SENSE already makes assumptions about the temporal smoothness of the image data. Therefore, further improvement of our method might be achieved by increasing the duration of the breathhold scan to achieve higher ground truth-image quality.

References

[1] C. M. Kramer, J. Barkhausen, S. D. Flamm, R. J. Kim, and E. Nagel, "Standardized cardiovascular magnetic resonance (CMR) protocols 2013 update," *J. Cardiovascular Magn. Reson.*, vol. 15, no. 1, p. 91, Dec. 2013.

[2] F. von Knobelsdorff-Brenkenhoff, G. Pilz, and J. Schulz-Menger, "Representation of cardiovascular magnetic resonance in the AHA / ACC guidelines," *J. Cardiovascular Magn. Reson.*, vol. 19, no. 1, p. 70, Dec. 2017. [Online]. Available: http://jcmr-online.biomedcentral.com/articles/10.1186/s12968-017-0385-z

[3] J. Tsao, P. Boesiger, and K. P. Pruessmann, "$k-t$ BLAST and $k-t$ SENSE: Dynamic MRI with high frame rate exploiting spatiotemporal correlations," *Magn. Reson. Med.*, vol. 50, no. 5, pp. 1031–1042, Nov. 2003.

[4] K. T. Block, M. Uecker, and J. Frahm, "Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint," *Magn. Reson. Med.*, vol. 57, pp. 1086–1098, Jun. 2007.

[5] H. Pedersen, S. Kozerke, S. Ringgaard, K. Nehrke, and W. Y. Kim, "$k-t$ PCA: Temporally constrained $k-t$ BLAST reconstruction using principal component analysis," *Magn. Reson. Med.*, vol. 62, no. 3, pp. 706–716, Sep. 2009.

[6] A. S. Gupta and Z.-P. Liang, "Dynamic imaging by temporal modeling with principal component analysis," in *Proc. 9th Annu. Meeting Int. Soc. Magn. Reson. Med.*, Apr. 2001, pp. 21–27.

[7] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "$k-t$ FOCUSS: A general compressed sensing framework for high resolution dynamic MRI," *Magn. Reson. Med.*, vol. 61, no. 1, pp. 103–116, Jan. 2009.

[8] J. Caballero, A. N. Price, D. Rueckert, and J. V. Hajnal, "Dictionary learning and time sparsity for dynamic MR data reconstruction," *IEEE Trans. Med. Imag.*, vol. 33, no. 4, pp. 979–994, Apr. 2014.

[9] Y. Wang and L. Ying, "Compressed sensing dynamic cardiac cine MRI using learned spatiotemporal dictionary," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1109–1120, Apr. 2014.

[10] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[11] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-Net for compressive sensing MRI," in *Proc. NIPS*, 2016, pp. 10–18.

[12] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, Jun. 2018.

[13] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, p. 487, 2018.

[14] Y. S. Han, J. Yoo, and J. C. Ye, "Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis," 2016, *arXiv:1611.06391*. [Online]. Available: https://arxiv.org/abs/1611.06391

[15] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *Med. Phys.*, vol. 44, no. 10, pp. e360–e375, Oct. 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. (MICCAI)*, Munich, Germany, 2015, pp. 234–241.

[17] J. C. Ye, Y. Han, and E. Cha, "Deep convolutional framelets: A general deep learning framework for inverse problems," *SIAM J. Imag. Sci.*, vol. 11, no. 2, pp. 991–1048, 2018.

[18] Y. Han and J. C. Ye, "Framing U-Net via deep convolutional framelets: Application to sparse-view CT," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1418–1429, Jun. 2018.

[19] 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 Trans. Med. Imag.*, vol. 37, no. 2, pp. 491–503, Feb. 2017.

[20] C. M. Sandino, N. Dixit, J. Y. Cheng, and S. S. Vasanawala, "Deep convo-lutional neural networks for accelerated dynamic magnetic resonance imaging," in *Proc. Med. Imag. Meets Neural Inf. Process. Syst. Conf.*, Long Beach, CA, USA, 2017, p. 19.

[21] A. Hauptmann, S. Arridge, F. Lucka, V. Muthurangu, and J. A. Steeden, "Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease," *Magn. Reson. Med.*, vol. 81, no. 2, pp. 1143–1156, Feb. 2019.

[22] C. Kolbitsch, C. Prieto, and T. Schaeffter, "Cardiac functional assessment without electrocardiogram using physiological self-navigation," *Magn. Reson. Imag.*, vol. 71, no. 3, pp. 942–954, May 2014. doi: 10.1002/mrm.24735.

[23] H. Jung, J. Park, J. Yoo, and J. C. Ye, "Radial $k-t$ FOCUSS for high-resolution cardiac cine MRI," *Magn. Reson. Imag.*, vol. 63, no. 1, pp. 68–78, Jan. 2010. doi: 10.1002/mrm.22172.

[24] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Convolutional recurrent neural networks for dynamic MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 280–290, Jan. 2019.

[25] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM J. Imag. Sci.*, vol. 1, no. 3, pp. 248–272, Aug. 2008.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *ImageNet Classification Deep Convolutional Neural Netw.*, 2012, pp. 1097–1105.

[27] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2017, pp. 5028–5037.

[28] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "Rotation equivariant vector field networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5048–5057.

[29] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[30] W. Bae, J. Yoo, and J. C. Ye, "Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification," in *Proc. CVPR Workshops*, Jul. 2017, pp. 145–153.

[31] R. Ghrist, "Barcodes: The persistent topology of data," *Bull. Amer. Math. Soc.*, vol. 45, no. 1, pp. 61–75, 2008.

[32] M. Bianchini and F. Scarselli, "On the complexity of shallow and deep neural network classifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1553–1565, Aug. 2014.

[33] S. Y. Oudot, *Persistence Theory: From Quiver Representations to Data Analysis*, vol. 209. Providence, RI, USA: American Mathematical Society, 2015.

[34] C. Maria, "Persistent cohomology," in *GUDHI User Reference Manual*. GUDHI Editorial Board, 2015. [Online]. Available: http://gudhi.gforge.inria.fr/doc/latest/group__persistent__cohomology.html

[35] C. Maria, P. Dlotko, V. Rouvreau, and M. Glisse, "Rips complex," in *GUDHI User Reference Manual*. GUDHI Editorial Board, 2016. [Online]. Available: http://gudhi.gforge.inria.fr/doc/latest/group__rips__complex.html

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift sergey," *J. Mol. Struct.*, vol. 37, pp. 1–11, Feb. 2017.

[37] D. Lee, J. Yoo, and J. C. Ye, "Deep residual learning for compressed sensing MRI," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, Apr. 2017, pp. 15–18.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[39] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Process., Image Commun.*, vol. 61, pp. 33–43, Feb. 2018.

[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[41] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[42] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 2990–2999.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[44] J. Adler. (2013). *Odl—Operator Discretization Library*. [Online]. Available: https://github.com/odlgroup/odl

[45] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock, "Variational networks: Connecting variational methods and deep learning," in *Proc. 39th German Conf. Pattern Recognit. (GCPR)*, Basel, Switzerland, Sep. 2017, pp. 281–293.

[46] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, Jun. 2018.

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

[48] K. Schnass, "Convergence radius and sample complexity of ITKM algorithms for dictionary learning," *Appl. Comput. Harmon. Anal.*, vol. 45, no. 1, pp. 22–58, Jul. 2018.

[49] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, p. 4311, Nov. 2006.

[50] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

[51] S. Wang *et al.*, "Accelerating magnetic resonance imaging via deep learning," in *Proc. IEEE 13th Int. Symp. Biomed. Imag.*, Apr. 2016, pp. 514–517.

8 Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

9 Publikationsliste

- Andreas Kofler, Markus Haltmeier, Christoph Kolbitsch, Marc Kachelrieß, and Marc Dewey. *A U-nets Cascade for Sparse View Computed Tomography.* In International Workshop on Machine Learning for Medical Image Reconstruction, p. 91–99. Springer, 2018.

- Andreas Kofler, Marc 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.* In IEEE Transactions on Medical Imaging, 39(3), p. 703–717, 2020.

- Andreas Kofler, Markus Haltmeier, Tobias Schaeffter, Marc Kachelrieß, Marc Dewey, Christian Wald, and Christoph Kolbitsch. *Neural Networks-Based Regularization for Large-Scale Medical Image Reconstruction.* In Physics in Medicine & Biology, 65(13):135003, 2020.

- Marc Dewey, Maria Siebes, Marc Kachelrieß, Klaus F. Kofoed, Pál Maurovich-Horvat, Konstantin Nikolaou, Wenjia Bai, Andreas Kofler, Robert Manka, Sebastian Kozerke, Amedeo Chiribiri, Tobias Schaeffter, Florian Michallek, Frank Bengel, Stephan Nekolla, Paul Knaapen, Mark Lubbernik, Roxy Senior, Meng-Xing Tang, Ja J. Piek, Tim Van De Hoef, Johannes Martens, Laura Schreiber on behalf of the Quantitative Cardiac Imaging Study Group. *Clinical quantitative cardiac imaging for the assessment of myocardial ischaemia.* In Nature Reviews Cardiology, p. 1–24, 2020.

10 Danksagung

I would like to express my gratitude to all people who were in the one or the other way, directly as well as indirectly, involved in the realization of this PhD thesis and other side projects arising from it.
First, I would like to thank my colleagues Christian, Sepehr, Matteo, Eser, Steffen and Nader for all the great and fun time we spent together drinking good coffee and going for (way too long) lunch breaks to escape our sorrows.
Second, my greatest gratitude goes to Naomi for all her psychological support and her acquired capability of standing all my complaining and frustration about whatsoever.
Further, many thanks to Christoph Kolbitsch and Markus Haltmeier for all the technical and mathematical support, the fruitful discussions and the different projects we worked on during the last three years.
Thanks to Marie Pali for the collaboration on Dictionary Learning for MRI, to Marc Kachelrieß for the help and the support regarding all CT-related matters and to Marc Dewey for the support concerning the medical details.
Last but not least, many thanks to my family for the constant support over the years.