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

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 under-sampling operator with $M \ll N$ which sets non-measured values in k-space to zero. Thereby, $I \subset J = \{1, \ldots, N\}$ corresponds to the set of indices of the measured Fourier coefficients. Since $M \ll N$, the problem in (2) is under-determined 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.
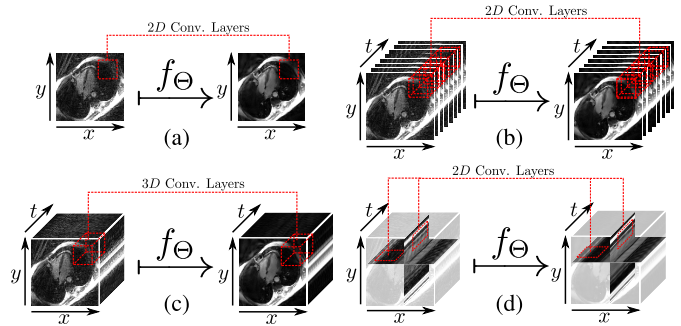

Fig. 1. Different $2D$ and $3D$ Deep Learning-based approaches for under-sampling artefacts reduction. $2D$ network for frame-wise mapping (a), $2D$ network for image sequence-wise mapping with cardiac phases aligned as channels (b), $3D$ network for image sequence-wise mapping with three-dimensional 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
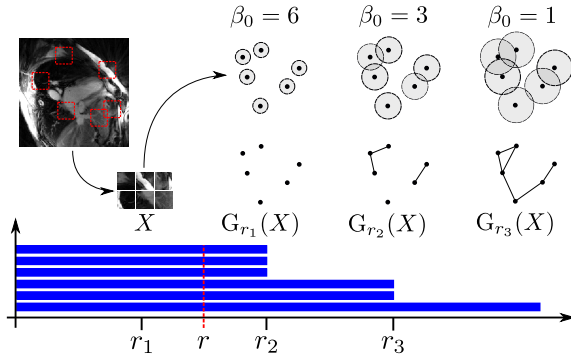
Fig. 2. Procedure of the persistent homology analysis. The image shows an example for six randomly extracted patches of an image in the spatial domain and its corresponding barcode.

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 $G_r(X)$ with vertices $V_r(X) = X$ and edges

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

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

$$H_0^r = \bigoplus_{i \in \Pi(G_r(X))} \mathbb{F}_2$$

where $\mathbb{F}_2$ is the field with two elements. For $0 \leq r < r'$ we have a map $\Pi(G_r(X)) \to \Pi(G_{r'}(X))$ which induces a map $H_0^r \to H_0^{r'}$. 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 $G_r(X)$ 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 $G_r(X)$ which is a measure of complexity for $G_r(X)$, 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
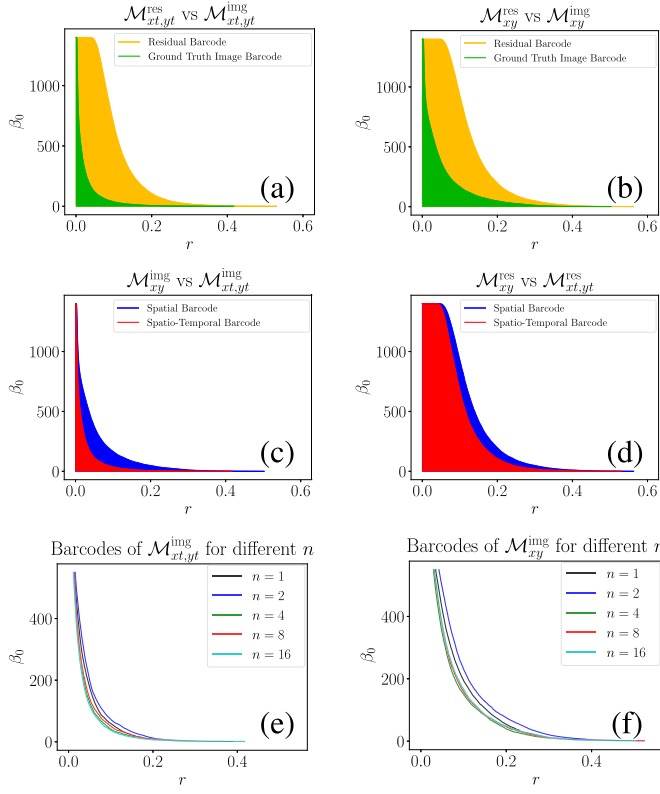
Fig. 3. The number of connected components β_0 of $G_r(X)$ for different datasets X at different r. Pairwise comparison of the persistent barcodes for $\mathcal{M}^{\mathrm{res}}_{xt,yt}$ and $\mathcal{M}^{\mathrm{img}}_{xt,yt}$ (a), for $\mathcal{M}^{\mathrm{res}}_{xy}$ and $\mathcal{M}^{\mathrm{img}}_{xy}$ (b), for $\mathcal{M}^{\mathrm{img}}_{xy}$ and $\mathcal{M}^{\mathrm{img}}_{xt,yt}$ (c) and for $\mathcal{M}^{\mathrm{res}}_{xy}$ and $\mathcal{M}^{\mathrm{res}}_{xt,yt}$ (d). Persistent codes of $\mathcal{M}^{\mathrm{img}}_{xy}$ and $\mathcal{M}^{\mathrm{img}}_{xt,yt}$ 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],
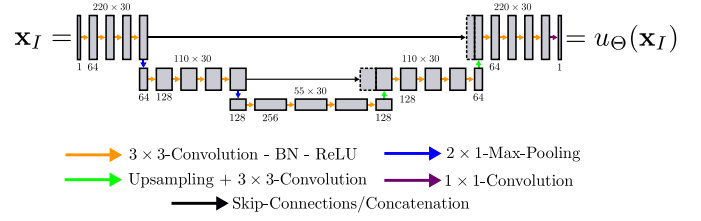

Fig. 4. The U-net with three encoding stages and four convolutional layers per stage, no residual connection and batch-normalization (BN). In the case we train on the spatial domain, max-pooling is performed in both spatial dimensions, whereas in our proposed approach max-pooling is solely performed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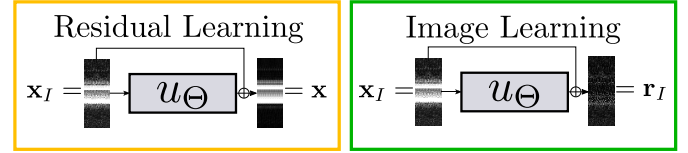

Fig. 5. Residual and Image Learning: For a NN $\tilde{u}_\Theta$ with residual connection, learning the residuals is achieved by using the ground truth images x as labels (left). Learning the ground truth images x is achieved by using the residuals r_I as labels (right).

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 $\mathbf{r}_{\mathrm{cnn}} = \tilde{u}_\Theta(\mathbf{x}_I) = \mathbf{x}_I + u_\Theta(\mathbf{x}_I) = \mathbf{x}_I - \mathbf{x}_{\mathrm{cnn}}$, then the estimates of the images are obtained by $\mathbf{x}_I - \mathbf{r}_{\mathrm{cnn}} = \mathbf{x}_I - (\mathbf{x}_I - \mathbf{x}_{\mathrm{cnn}}) = \mathbf{x}_{\mathrm{cnn}} \approx \mathbf{x}$.

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

C. Loss Function

Dependent on what we want the network to learn, we train the network architecture to minimize different loss functions. Let $\mathcal{D}^{\mathrm{res}}_{xy}, \mathcal{D}^{\mathrm{img}}_{xy}$ and $\mathcal{D}^{\mathrm{res}}_{xt,yt}, \mathcal{D}^{\mathrm{img}}_{xt,yt}$ 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

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

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

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

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

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

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 $160 \times 160 \times 30$ 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 data-balance reasons, we crop the image sequences using a cut-off 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
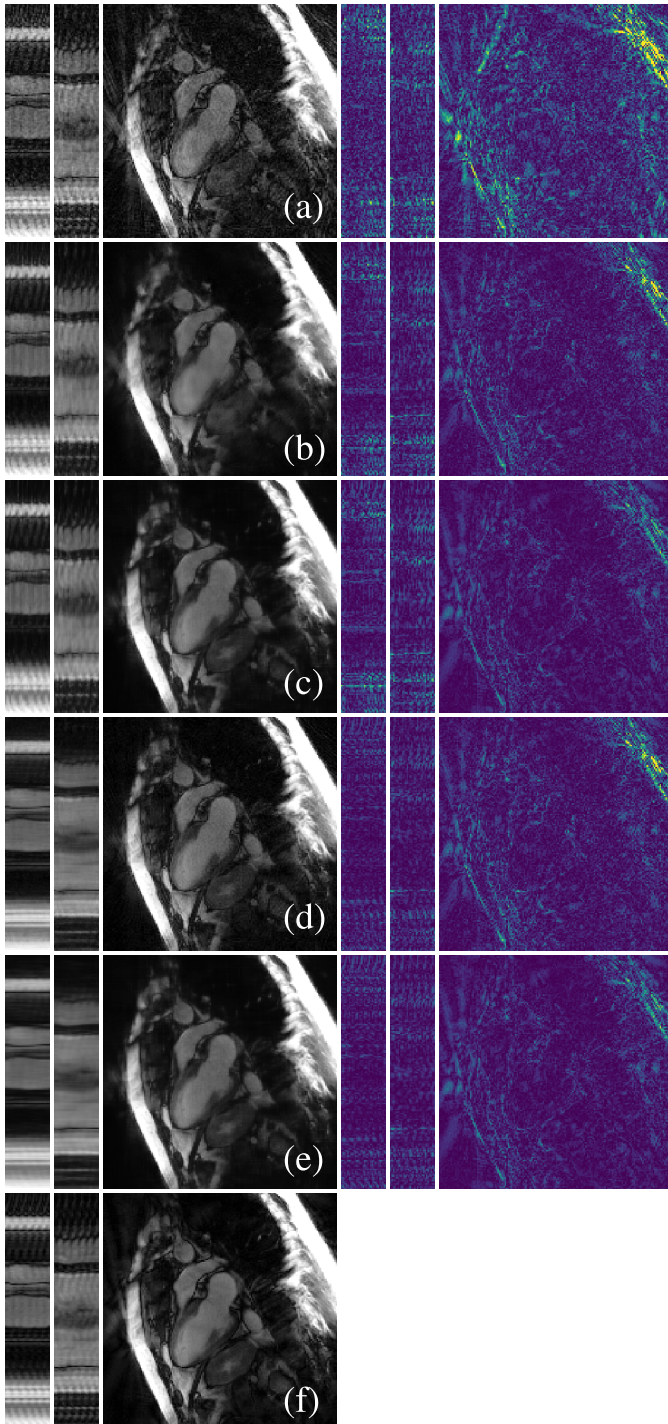
Fig. 6. Comparison of different training approaches for U-nets with residual connection. NUFFT reconstruction with $N_\theta = 1130$ radial lines (a), spatially trained U-nets u_{xy}^{res} (b) and u_{xy}^{img} (c), proposed spatio-temporal approaches $u_{xt,yt}^{\text{res}}$ (d) and $u_{xt,yt}^{\text{img}}$ (e), ground truth (f). The point-wise error images are magnified by a factor of $\times 3$. All images are displayed on the same scale.

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

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.

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

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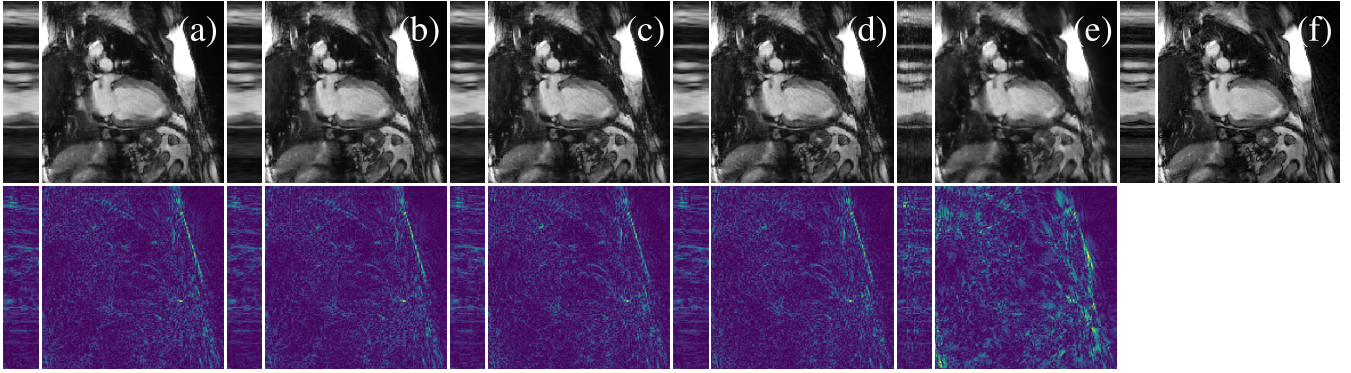
Fig. 7. Results on the test set for $N_\theta = 1130$ radial lines when the number of subjects whose spatio-temporal slices are extracted was varied. Note that no data-augmentation was used. Proposed method for $n = 1$ (a), $n = 2$ (b), $n = 8$ (c), $n = 12$ (d), the spatial U-net for $n = 12$ (e) and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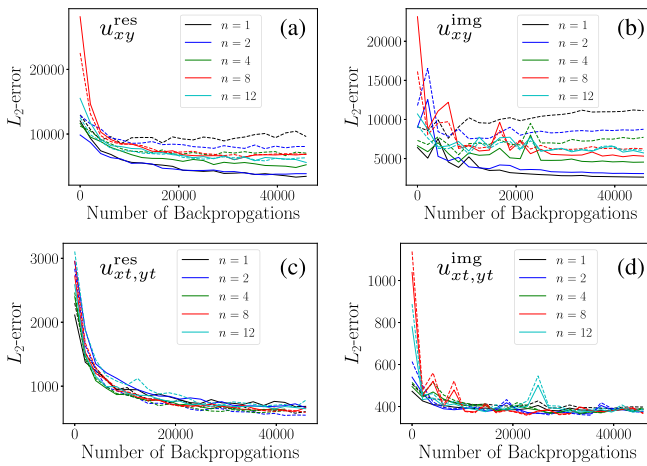

Fig. 8. Loss behaviour during training with $N_\theta = 1130$ for different number of volunteers n contained in the dataset. Training loss (solid) and validation loss (dashed) for the spatial and spatio-temporal U-nets. Spatial residual learning (a), spatial image learning (b), spatio-temporal residual 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 $\mathcal{D}_{xy}^{\psi_\theta}$ and $\mathcal{D}_{xt,yt}^{\psi_\theta}$ 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
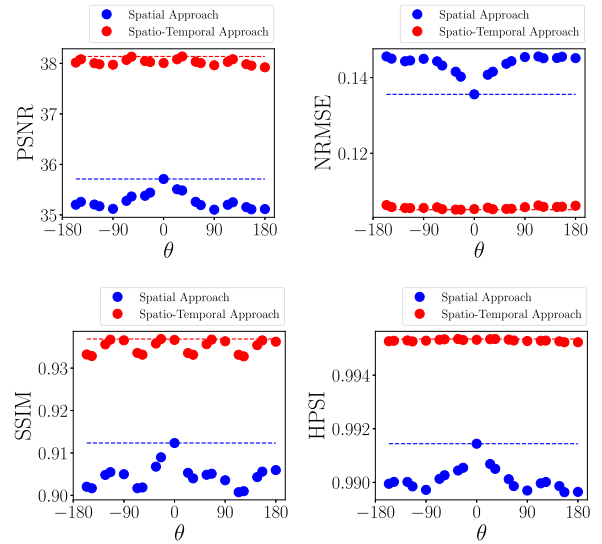

Fig. 9. Performance of the networks when tested on rotated copies of the images contained in the training set. While the network trained in the spatio-temporal domain is robust with respect to rotation, the network trained on images in the spatial domain loses generalization power when tested on rotated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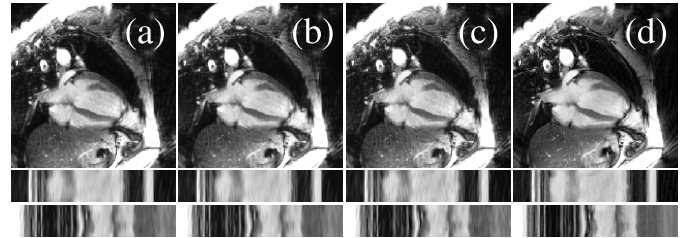

Fig. 10. Results obtained with different CNNs following our proposed approach $u_{xt,yt}^{img}$. E1 C8 (a), E4 C4 (b) and E5 C2 (c), kt-SENSE reconstruction with $N_\theta = 3400$ radial lines (d). Our approach therefore offers the possibility to further reduce the network complexity as well as training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

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.

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
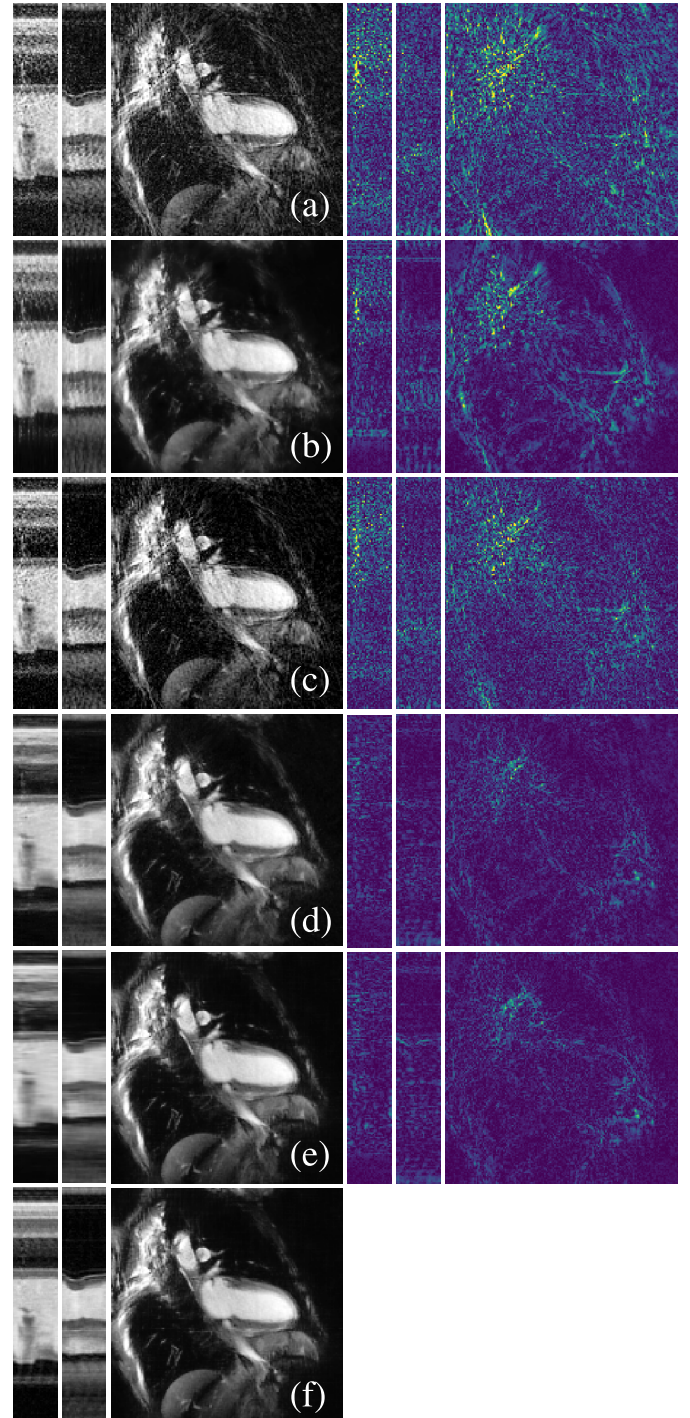

Fig. 11. Comparison with different Deep Learning-based post-processing methods. NUFFT reconstruction with $N_\theta = 1130$ radial lines (a), u_{xy}^{img} (b), $u_{xy,t}$ (c), u_{xyt} (d), proposed approach u_{xy}^{img} (e), ground truth *kt*-SENSE reconstruction (f). The point-wise error images are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

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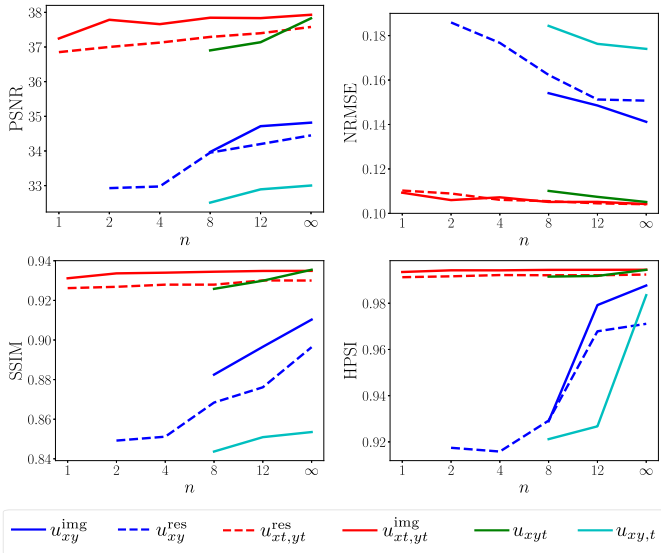
Fig. 12. Quantitative measures for all discussed Deep Learning-based post-processing methods when trained on datasets including different number of subjects *n*. Missing values for some *n* denote that the network was not properly trainable on the restricted dataset.

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

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

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.

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 $6\,\mathrm{dB}$, 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
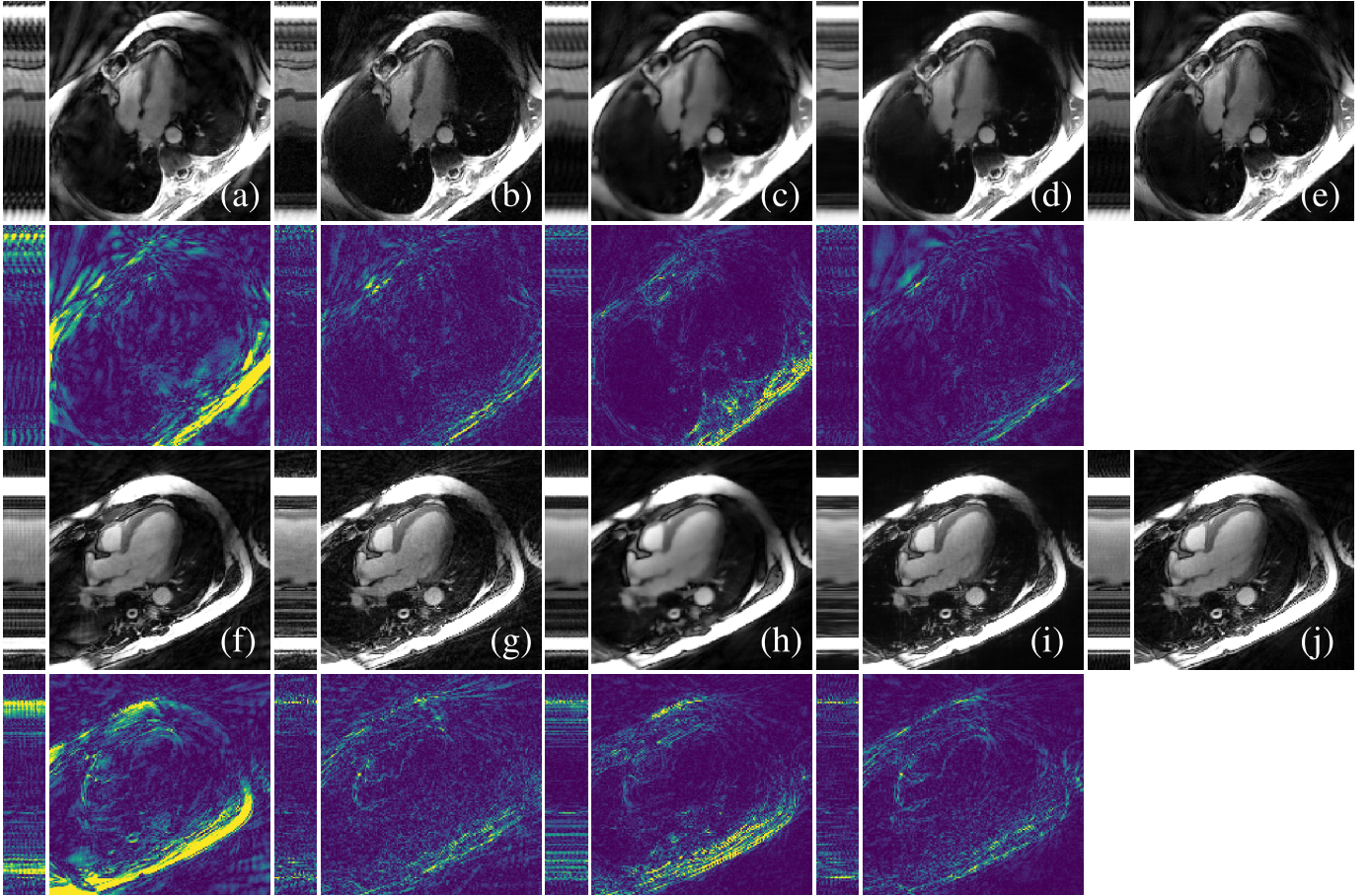
Fig. 13. Comparison with different state-of-the-art iterative reconstruction methods. *kt*-FOCUSS (a) and (f), TV+TVT (b) and (g), DL+TV (c) and (h), proposed method (d) and (i), *kt*-SENSE reconstruction with $N_\theta = 3400$ radial lines. The point-wise error images are magnified by a factor of $\times 3$. 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

TABLE VI
COMPARISON WITH DIFFERENT CASCADED CNNS

Reconstruction	$3D$ **CNN cascade**	$2D$ **CRNN cascade**	$u_{xt,yt}^{\mathrm{img}}$
	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

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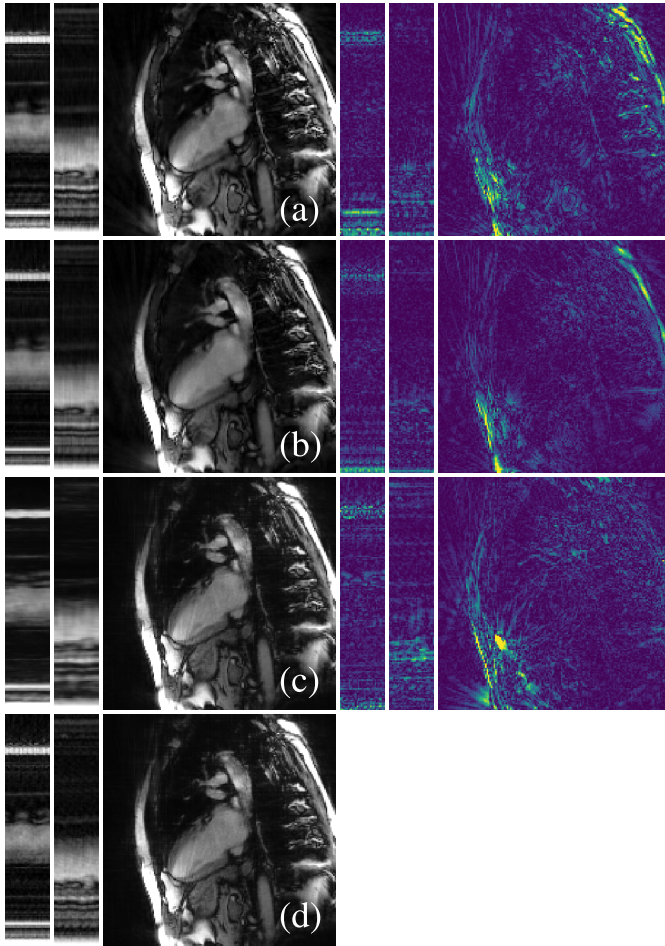
Fig. 14. Comparison with different cascaded CNNs: $2D$ CRNN Cascade (a), $3D$ CNN-Cascade (b), proposed (c) and the reference *kt*-SENSE reconstruction (d). The Figure show results for the fold where only patient's data is included in the test set. Qualitatively, all the three methods perform similarly.

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 $\mathcal{M}_{xt,yt}^{\text{img}}$ 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.