FULL PAPER

Magnetic Resonance in Medicine 00:00–00 (2017)

# Learning a Variational Network for Reconstruction of Accelerated MRI Data

Kerstin Hammernik,[1]* Teresa Klatzer,[1] Erich Kobler,[1] Michael P. Recht,[2,3] Daniel K. Sodickson,[2,3] Thomas Pock,[1,4] and Florian Knoll[2,3]

**Purpose:** To allow fast and high-quality reconstruction of clinical accelerated multi-coil MR data by learning a variational network that combines the mathematical structure of variational models with deep learning.

**Theory and Methods:** Generalized compressed sensing reconstruction formulated as a variational model is embedded in an unrolled gradient descent scheme. All parameters of this formulation, including the prior model defined by filter kernels and activation functions as well as the data term weights, are learned during an offline training procedure. The learned model can then be applied online to previously unseen data.

**Results:** The variational network approach is evaluated on a clinical knee imaging protocol for different acceleration factors and sampling patterns using retrospectively and prospectively undersampled data. The variational network reconstructions outperform standard reconstruction algorithms, verified by quantitative error measures and a clinical reader study for regular sampling and acceleration factor 4.

**Conclusion:** Variational network reconstructions preserve the natural appearance of MR images as well as pathologies that were not included in the training data set. Due to its high computational performance, that is, reconstruction time of 193 ms on a single graphics card, and the omission of parameter tuning once the network is trained, this new approach to image reconstruction can easily be integrated into clinical workflow. **Magn Reson Med 000:000–000, 2017. © 2017 International Society for Magnetic Resonance in Medicine.**

**Key words:** variational network; deep learning; accelerated MRI; parallel imaging; compressed sensing; image reconstruction

## INTRODUCTION

Imitating human learning with deep learning (1,2) has become an enormously important area of research and development, with a high potential for far-reaching application, including in the domain of Computer Vision. Taking encouragement from early successes in image classification tasks (3), recent advances also address semantic labeling (4), optical flow (5) and image restoration (6). In medical imaging, deep learning has also been applied to areas like segmentation (7,8), q-space image processing (9), and skull stripping (10). However, in these applications, deep learning was seen as a tool for image processing and interpretation. The goal of the current work is to demonstrate that the concept of learning can also be used at the earlier stage of image formation. In particular, we focus on image reconstruction for accelerated MRI, which is commonly accomplished with frameworks like Parallel Imaging (PI) (11–13) or Compressed Sensing (CS) (14–16). CS in particular relies on three conditions to obtain images from k-space data sampled below the Nyquist rate (17,18).

The first CS condition requires a data acquisition protocol for undersampling such that artifacts become incoherent in a certain transform domain (14,15). In MRI, we usually achieve incoherence by random (16) or non-Cartesian sampling trajectories (19). The second requirement for CS is that the image to be reconstructed must have a sparse representation in a certain transform domain. Common choices are the Wavelet transform (16,20) or total variation (TV) (19,21–23). In these transform domains, the $l_1$ norm is commonly applied to obtain approximate sparsity. The third CS condition requires a non-linear reconstruction algorithm that balances sparsity in the transform domain against consistency with the acquired undersampled k-space data.

Despite the high promise of CS approaches, most routine clinical MRI examinations are still based on Cartesian sequences. Especially in the case of 2D sequences, it can be challenging to fulfill the criteria for incoherence required by CS (24). One other obstacle to incorporation of CS into some clinical examinations is the fact that the sparsifying transforms employed in CS applications to date may be too simple to capture the complex image content associated with biological tissues. This can lead to reconstructions that appear blocky and unnatural, which reduces acceptance by clinical radiologists. A further drawback, not only for CS but for advanced image acquisition and reconstruction methods in general, is the long image reconstruction time typically required for iterative solution of non-linear optimization problems. A final challenge concerns the selection and tuning of hyper-parameters for CS approaches. A poor choice of hyper-parameters leads either to over-regularization, that is, excessively smooth or

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria.

[2]Center for Biomedical Imaging, Department of Radiology, NYU School of Medicine, New York, New York, USA.

[3]Center for Advanced Imaging Innovation and Research (CAI$^2$R), NYU School of Medicine, New York, New York, USA.

[4]Center for Vision, Automation & Control, AIT Austrian Institute of Technology GmbH, Vienna, Austria.

Grant sponsor: FWF START Project BIVISION; Grant number: Y729; Grant sponsor: ERC starting grant "HOMOVIS"; Grant number: 640156; Grant sponsor: NIH; Grant number: P41 EB017183; Grant sponsor: NIH; Grant number: R01 EB000447.

*Correspondence to: Kerstin Hammernik, M.Sc., Institute of Computer Graphics and Vision, Graz University of Technology, Inffeldgasse 16/II, 8010 Graz, Austria, E-mail: hammernik@icg.tugraz.at

Preliminary data for this article were presented at the 24th Annual Meeting of ISMRM, Singapore, 2016.

Received 30 March 2017; revised 19 September 2017; accepted 27 September 2017

DOI 10.1002/mrm.26977

Published online 00 Month 2017 in Wiley Online Library (wileyonlinelibrary.com).

unnatural-looking images, or else to images that still show residual undersampling artifacts. The goal of our current work is to demonstrate that, using learning approaches, we can achieve accelerated and high-quality MR image reconstructions from undersampled data which do not fulfill the usual CS conditions, which we address with both quantitative error measures and a clinical reader study.

With current iterative image reconstruction approaches, we treat every single exam and resulting image reconstruction task as a new optimization problem. We do not use information about the expected appearance of the anatomy, or the known structure of undersampling artifacts, explicitly in these optimization problems, which stands in stark contrast to how human radiologists read images. Radiologists are trained throughout their careers to look for certain reproducible patterns, and they obtain remarkable skills to "read through" known image artifacts (24). Translating this conceptual idea of human learning to deep learning allows us to shift the key effort of optimization from the online reconstruction stage to an up-front offline training task. In other words, rather than solving an inverse problem to compute, for each new data set, a suitable transform between raw data and images, we propose to *learn* the key parameters of that inverse transform in advance, so that it can be applied to all new data as a simple flow-through operation.

In this work, we introduce an efficient trainable formulation for accelerated PI-based MRI reconstruction that we term a *variational network* (*VN*). The VN embeds a generalized CS concept, formulated as a variational model, within a deep learning approach. Our VN is designed to learn a complete reconstruction procedure for complex-valued multichannel MR data, including all free parameters which would otherwise have to be set empirically. We train the VN on a complete retrospectively undersampled clinical protocol for musculoskeletal imaging, and evaluate performance for different acceleration factors, and for both regular and pseudo-random Cartesian 2D sampling. Using both retrospectively and prospectively undersampled clinical patient data, we investigate the applicability of our proposed VN approach for clinical routine examination, including improved image quality and preservation of unique pathologies that are not included in the training data set.

## THEORY

### From Linear Reconstruction to a Variational Network

In MRI reconstruction, we naturally deal with complex numbers. Here, we introduce a mapping to real-valued numbers that we will use throughout our manuscript. We define complex images $\tilde{\mathbf{u}}$ of size $n_x \times n_y = N$ as equivalent real images $\mathbf{u}$ as follows:

$$\tilde{\mathbf{u}} = \mathbf{u}_{\mathrm{re}} + j\mathbf{u}_{\mathrm{im}} \in \mathbb{C}^N \Longleftrightarrow \mathbf{u} = (\mathbf{u}_{\mathrm{re}}, \mathbf{u}_{\mathrm{im}}) \in \mathbb{R}^{2N}.$$

We consider the ill-posed linear inverse problem of finding a reconstructed image $\mathbf{u} \in \mathbb{R}^{2N}$ that satisfies the following system of equations

$$\mathbf{A}\mathbf{u} = \hat{\mathbf{f}}, \qquad [1]$$

where $\hat{\mathbf{f}} \in \mathbb{R}^{2NQ}$ is the given undersampled k-space data, where missing data are padded by zeros. The linear forward sampling operator $\mathbf{A}$ implements point-wise multiplications with $Q$ coil sensitivity maps, Fourier transforms, and undersampling according to a selected sampling pattern. Originally, the operator $\mathbf{A}$ is defined by the mapping $\mathbb{C}^N \mapsto \mathbb{C}^{NQ}$, but embedding it in our real-valued problem changes the mapping to $\mathbb{R}^{2N} \mapsto \mathbb{R}^{2NQ}$. Since the system in Equation [1] is ill-posed, we cannot solve for $\mathbf{u}$ explicitly. Therefore, a natural idea is to compute $\mathbf{u}$ by minimizing the least squares error

$$\min_{\mathbf{u}} \frac{1}{2} \|\mathbf{A}\mathbf{u} - \hat{\mathbf{f}}\|_2^2. \qquad [2]$$

In practice we do not have access to the true $\hat{\mathbf{f}}$ but only to a noisy variant $\mathbf{f}$ satisfying

$$\|\hat{\mathbf{f}} - \mathbf{f}\|_2 \leq \delta$$

where $\delta$ is the noise level. The idea is to perform a gradient descent on the least squares problem Equation [2] that leads to an iterative algorithm, which is known as the Landweber method (25). It is given by choosing some initial $\mathbf{u}^0$ and performing the iterations with step sizes $\alpha^t$

$$\mathbf{u}^{t+1} = \mathbf{u}^t - \alpha^t \mathbf{A}^*(\mathbf{A}\mathbf{u}^t - \mathbf{f}), \quad t \geq 0 \qquad [3]$$

where $\mathbf{A}^*$ is the adjoint linear sampling operator. To prevent over-fitting to the noisy data $\mathbf{f}$, it is beneficial to stop the Landweber iterative algorithm early (26), that is, after a finite number of iterations $T$.

Instead of early stopping, we can also extend the least squares problem by an additional regularization term $\mathcal{R}(\mathbf{u})$ to prevent over-fitting. The associated (variational) minimization problem is given by

$$\min_{\mathbf{u}} \left\{ \mathcal{R}(\mathbf{u}) + \frac{\lambda}{2} \|\mathbf{A}\mathbf{u} - \mathbf{f}\|_2^2 \right\}.$$

The minimizer of the regularized problem depends on the trade-off between the regularization term and the least squares data fidelity term controlled by $\lambda > 0$. One of the most influential regularization terms in the context of images is the TV semi-norm (21), which is defined as

$$\begin{aligned} \mathcal{R}(\mathbf{u}) &= \|(\mathbf{D}\mathbf{u}_{\mathrm{re}}, \mathbf{D}\mathbf{u}_{\mathrm{im}})\|_{2,1} \\ &= \sum_{l=1}^{N} \sqrt{|\mathbf{D}\mathbf{u}_{\mathrm{re}}|_{l,1}^2 + |\mathbf{D}\mathbf{u}_{\mathrm{im}}|_{l,1}^2 + |\mathbf{D}\mathbf{u}_{\mathrm{re}}|_{l,2}^2 + |\mathbf{D}\mathbf{u}_{\mathrm{im}}|_{l,2}^2}, \end{aligned}$$

where $\mathbf{D} : \mathbb{R}^N \mapsto \mathbb{R}^{N \times 2}$ is a finite differences approximation of the image gradient, see for example (27). The main advantage of TV is that it allows for sharp discontinuities (edges) in the solution while being a convex functional enabling efficient and global optimization. From a sparsity point of view, TV induces sparsity in the image edges and hence, favors piecewise constant solutions. However, it is also clear that the piecewise-constant approximation is not a suitable criterion to describe the complex structure of MR images and a more general regularizer is needed.

A generalization of the TV is the Fields of Experts model (28)

$$\mathcal{R}(\mathbf{u}) = \sum_{i=1}^{N_k} \langle \Phi_i(\mathbf{K}_i\mathbf{u}), \mathbf{1}\rangle. \quad [4]$$

Here, the regularization term is extended to $N_k$ terms and $\mathbf{1}$ denotes a vector of ones. The linear operator $\mathbf{K} = (\mathbf{K}_{\text{re}}, \mathbf{K}_{\text{im}}) : \mathbb{R}^{2N} \mapsto \mathbb{R}^N$ models convolutions with filter kernels $k \in \mathbb{R}^{s\times s\times 2}$ of size $s$, which is expressed as

$$\mathbf{K}\mathbf{u} = \mathbf{K}_{\text{re}}\mathbf{u}_{\text{re}} + \mathbf{K}_{\text{im}}\mathbf{u}_{\text{im}},$$
$$\mathbf{u} \in \mathbb{R}^{2N} \Longleftrightarrow u * k = u_{\text{re}} * k_{\text{re}} + u_{\text{im}} * k_{\text{im}}, \quad u \in \mathbb{R}^{n_x\times n_y\times 2}.$$

The non-linear potential functions $\Phi(\mathbf{z}) = (\phi(z_1), \ldots, \phi(z_N))^\top : \mathbb{R}^N \mapsto \mathbb{R}^N$ are composed by scalar functions $\phi$. In the Fields of Experts model (28), both convolution kernels and parametrization of the non-linear potential functions, such as student-t functions, are learned from data.

Plugging the Fields of Experts model Equation [4] into the Landweber iterative algorithm Equation [3] yields

$$\mathbf{u}^{t+1} = \mathbf{u}^t - \alpha^t\left(\sum_{i=1}^{N_k} (\mathbf{K}_i)^\top \Phi_i'(\mathbf{K}_i\mathbf{u}^t) + \lambda\mathbf{A}^*(\mathbf{A}\mathbf{u}^t - \mathbf{f})\right) \quad [5]$$

where $\Phi_i'(\mathbf{z}) = \text{diag}\big(\phi_i'(z_1), \ldots, \phi_i'(z_N)\big)$ are the activation functions defined by the first derivative of potential functions $\Phi_i$. Observe that the application of the transpose operation $(\mathbf{K}_i)^\top$ can be implemented as a convolution with filter kernels $k_i$ rotated by 180°. Chen et al. (6) introduce a trainable reaction-diffusion approach that performs early stopping on the gradient scheme Equation [5] and allows the parameters, such as filters, activation functions and data term weights, to vary in every gradient descent step $t$. All parameters of the approach are learned from data. This approach has been successfully applied to a number of image processing tasks including image denoising (6), JPEG deblocking (6), demosaicing (29) and image inpainting (30). For MRI reconstruction, we rewrite the trainable gradient descent scheme with time-varying parameters $\mathbf{K}_i^t$, $\Phi_i^{t\prime}$, $\lambda^t$ as

$$\mathbf{u}^{t+1} = \mathbf{u}^t - \sum_{i=1}^{N_k} (\mathbf{K}_i^t)^\top \Phi_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t) - \lambda^t\mathbf{A}^*(\mathbf{A}\mathbf{u}^t - \mathbf{f}), \quad 0 \le t \le T-1. \quad [6]$$

Additionally, we omit the step size $\alpha^t$ in Equation [5] because it is implicitly contained in the activation functions and data term weights.

By unfolding the iterations of Equation [6], we obtain the *variational network* (*VN*) structure as depicted in Figure 1. Essentially, one iteration of an iterative reconstruction can be related to one step in the network. In our VN approach, we directly use the measured raw data as input. Coil sensitivity maps are pre-computed from the fully sampled k-space center. A zero filled solution is computed from the undersampled k-space data by applying the adjoint operator $\mathbf{A}^*$. The measured raw data and sensitivity maps, together with the zero filled initializations, are fed into the VN as illustrated in Supporting Figure S1. The sensitivity maps are used in the operators $\mathbf{A}, \mathbf{A}^*$, which perform sensitivity-weighted image combination and can also implement other processing steps such as the removal of readout oversampling. While both raw data and operators $\mathbf{A}, \mathbf{A}^*$ are required in every iteration of the VN to implement the gradient of the data term, the gradient of the regularization is only applied in the image domain as depicted in Figure 1.

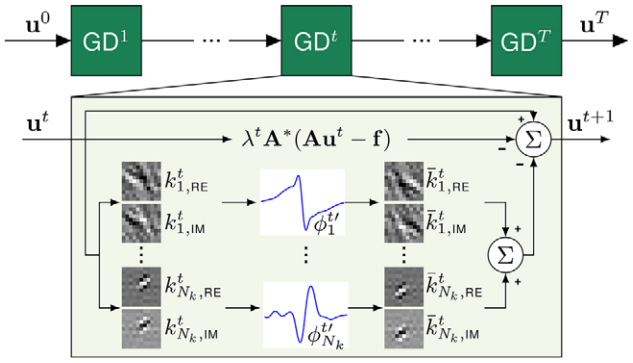


FIG. 1. Structure of the variational network (VN). The VN consists of $T$ gradient descent steps. To obtain a reconstruction, we feed the undersampled k-space data, coil sensitivity maps and the zero filling solution to the VN. Here, a sample gradient step is depicted in detail. As we are dealing with complex-valued images, we learn separate filters $k_i^t$ for the real and complex plane. The non-linear activation function $\phi_i^{t\prime}$ combines the filter responses of these two feature planes. During a training procedure, the filter kernels, activation functions and data term weights $\lambda^t$ are learned.

## METHODS

### Variational Network Parameters

The VN defined by Equation [6] and illustrated in Figure 1 contains a number of parameters: Filter kernels $k_i^t$, activation functions $\Phi_i^{t\prime}$, and data term weights $\lambda^t$. We first consider the filter kernels which requires us to introduce a vectorized version $\mathbf{k}_i^t \in \mathbb{R}^{2s^2}$ of the filter kernel $k_i^t$. We constrain the filters to be zero-mean which is defined as $\xi_{\text{re}}^\top\mathbf{k}_i^t = 0$, $\xi_{\text{im}}^\top\mathbf{k}_i^t = 0$, where $\xi_{\text{re}}^\top\mathbf{k}_i^t, \xi_{\text{im}}^\top\mathbf{k}_i^t$ estimate the individual means of the filter kernel on the real and imaginary plane, respectively. Additionally, the whole kernel is constrained to lie on the unit-sphere, that is, $\|\mathbf{k}_i^t\|_2 = 1$, to avoid a scaling problem of the activation functions. To learn the activation functions, we require a suitable function parametrization. A standard choice to smoothly approximate any functions are Gaussian radial basis functions (RBFs). We define the scalar activation functions $\phi_i^{t\prime}$ as a weighted combination of $N_w$ RBFs with nodes $\mu$ and standard deviation $\sigma = \frac{2I_{\max}}{N_w - 1}$

$$\phi_i^{t\prime}(z) = \sum_{j=1}^{N_w} w_{ij}^t \exp\left(-\frac{(z-\mu_j)^2}{2\sigma^2}\right).$$

The nodes are distributed in an equidistant way in $[-I_{\max}, I_{\max}]$ which allows us to achieve the same resolution over the whole defined range. Note here that $\mu, \sigma$ depend on the maximum estimated filter response $I_{\max}$. The final parameters that we consider are the data term weights $\lambda^t$, which are constrained to be non-negative ($\lambda^t > 0$). During

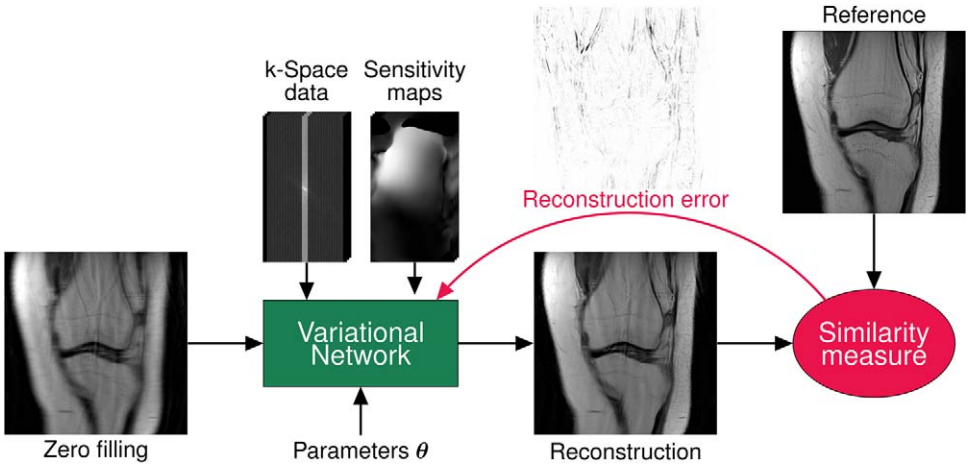


FIG. 2. Variational network training procedure: we aim at learning a set of parameters of the VN during an offline training procedure. For this purpose, we compare the current reconstruction of the VN to an artifact-free reference using a similarity measure. This gives us the reconstruction error which is propagated back to the VN to compute a new set of parameters.

training, all constraints on the parameters are realized based on projected gradient methods.

### Variational Network Training

During the offline training procedure illustrated in Figure 2, the goal is to find an optimal parameter set $\theta = \{\theta^0, \ldots, \theta^{T-1}\}$, $\theta^t = \{w_{ij}^t, \mathbf{k}_i^t, \lambda^t\}$ for our proposed VN in Eq. [6]. To set up the training procedure, we minimize a loss function over a set of images $S$ with respect to the parameters $\theta$. The loss function defines the similarity between the reconstructed image $\mathbf{u}^T$ and a clean, artifact-free reference image $\mathbf{g}$. A common choice for the loss function is the mean-squared error (MSE)

$$\mathcal{L}(\theta) = \min_{\theta} \frac{1}{2S} \sum_{s=1}^{S} \left\| \mathbf{u}_s^T(\theta) - \mathbf{g}_s \right\|_2^2.$$

As we are dealing with complex numbers in MRI reconstruction and we typically assess magnitude images, we define the MSE loss of ($\epsilon$-smoothed) absolute values

$$\mathcal{L}(\theta) = \min_{\theta} \frac{1}{2S} \sum_{s=1}^{S} \left\| |\mathbf{u}_s^T(\theta)|_\epsilon - |\mathbf{g}_s|_\epsilon \right\|_2^2, \; |\mathbf{x}|_\epsilon = \sqrt{\mathbf{x}_{\text{re}}^2 + \mathbf{x}_{\text{im}}^2 + \epsilon},$$

where $|\cdot|_\epsilon$ is understood in a point-wise manner. To solve this highly non-convex training problem, we use the Inertial Incremental Proximal Gradient (IIPG) optimizer which is related to the Inertial Proximal Alternating Linearized Minimization (IPALM) algorithm (31). For algorithmic details on IIPG refer to Appendix A and (32). First-order optimizers require both the loss function value and the gradient with respect to the parameters $\theta$. This gradient can be computed by simple back-propagation (33), that is, applying the chain rule

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta^t} = \frac{\partial \mathbf{u}^{t+1}}{\partial \theta^t} \cdot \frac{\partial \mathbf{u}^{t+2}}{\partial \mathbf{u}^{t+1}} \cdots \frac{\partial \mathbf{u}^{T}}{\partial \mathbf{u}^{T-1}} \cdot \frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^{T}}.$$

The derivation of the gradients for the parameters is provided in Appendix B. After training, the parameters $\theta$ are fixed and we can reconstruct previously unseen k-space data efficiently by forward-propagating the k-space data through the VN.

### Data Acquisition

A major goal of our work was to explore the generalization potential of a learning-based approach for MRI reconstruction. For this purpose, we used a standard clinical knee protocol for data acquisition with a representative patient population that differed in terms of anatomy, pathology, gender, age, and body mass index. The protocol consisted of five 2D turbo spin echo (TSE) sequences that differed in terms of contrast, orientation, matrix size, and signal-to-noise ratio (SNR). For each sequence, we scanned 20 patients on a clinical 3T system (Siemens Magnetom Skyra) using an off-the-shelf 15-element knee coil. All data were acquired without acceleration, and undersampling was performed retrospectively as needed. In addition, we acquired prospectively accelerated data for one case. The number of acquired slices was chosen individually for each clinical patient exam. The study was approved by our institutional review board. Sequence parameters were as follows:

**Coronal proton-density (PD)**: TR = 2750 ms, TE = 27 ms, turbo factor/echo train length TF = 4, matrix size 320 × 288, in-plane resolution 0.49 × 0.44 mm$^2$, slice thickness 3 mm, 35–42 slices, 5 female/15 male, age 15–76, BMI 20.46–32.94

**Coronal fat-saturated PD**: TR = 2870 ms, TE = 33 ms, TF = 4, matrix size 320 × 288, in-plane resolution 0.49 × 0.44 mm$^2$, slice thickness 3 mm, 33–44 slices, 10 female/10 male, age 30–80, BMI 19.76–33.87

**Axial fat-saturated $T_2$**: TR = 4000 ms, TE = 65 ms, TF = 9, matrix size 320 × 256, in-plane resolution 0.55 × 0.44 mm$^2$, slice thickness 3 mm, 33–41 slices, 10 female/10 male, age 20–70, BMI 19.20–35.69

**Sagittal fat-saturated $T_2$**: TR = 4300 ms, TE = 50 ms, TF = 11, matrix size 320 × 256, in-plane resolution 0.55 × 0.44 mm$^2$, slice thickness 3 mm, 31–40 slices, 11 female/9 male, age 12–73, BMI 18.16–37.31

**Sagittal PD**: TR = 2800 ms, TE = 27 ms, TF = 4, matrix size 384 × 307, in-plane resolution 0.46 × 0.36 mm$^2$, slice thickness 3 mm, 31–38 slices, 11 female/9 male, age 15–94, BMI 18.69–35.15

Coil sensitivity maps were precomputed from a data block of size 24 × 24 at the center of k-space using ESPIRiT (34). For both training and quantitative evaluation, each network reconstruction was compared against a gold standard reference image. We defined this gold standard as the coil-sensitivity combined fully sampled reconstruction. The fully sampled raw data were retrospectively undersampled for both training and testing.

### Experimental Setup

Our experiments differed in contrast, orientation, acceleration factor, and sampling pattern. For all our experiments, we pre-normalized the acquired k-space volumes with $n_{\mathrm{sl}}$ slices by $\frac{\sqrt{n_{\mathrm{sl}}}10000}{\|\mathbf{f}\|_2}$. We trained an individual VN for each experiment and kept the network architecture fixed for all experiments. The VN consisted of $T=10$ steps. The initial reconstruction $\mathbf{u}_0$ was defined by the zero filled solution. In each iteration $N_k=48$ real/imaginary filter pairs of size 11 × 11 were learned. For each of the $N_k$ filters, the corresponding activation function was defined by $N_w=31$ RBFs equally distributed between [−150,150]. Including the data term weight $\lambda^t$ in each step, this resulted in a total of 131,050 network parameters.

For optimization, we used the IIPG optimizer described in Appendix A. The IIPG optimizer allows handling the previously described constraints on the network parameters. We generated a training set for each of the five knee datasets. In each experiment, we used 20 image slices from 10 patients with the same contrast weighting and orientation, which amounts to 200 images, as the training set. For each patient, the central 20 slices were used for training. In fact, each single pixel of these training images provides a training example. In the case of a 320 × 320 matrix, this results in more than 20 million pixels which is orders of magnitudes larger than the number of network parameters. The training set was split into mini batches of size 10. Optimization was performed for 1000 epochs with a step size of $\eta = 10^{-3}$.

### Experiments

In the first step, we investigated whether the learning-based VN approach actually benefits from structured undersampling artifacts due to regular undersampling, or if it performs better with incoherent undersampling artifacts as are typically present in CS applications. We used a regular sampling scheme with fully-sampled k-space center consisting of 24 auto-calibration lines, identical to the vendor implementation of an accelerated TSE sequence on an MR-system. To introduce randomness, we also generated a variable-density random sampling pattern according to Lustig et al. (16). Both sampling patterns have the same fully-sampled k-space center and same number of phase encoding steps. We evaluated the acceleration factors $R \in \{3,4\}$ for two sequences which differ in contrast and SNR. The second step was to explore the generalization potential with respect to different contrasts and orientations of a clinical knee protocol. In a third step, we performed an experiment with prospectively accelerated data.

### Evaluation

We tested our algorithm on data from 10 clinical patients per sequence and reconstructed the whole imaged volume for each patient. These cases were not included in the training set, and they also contained pathology not represented in the training set. It is worth noting that the number of slices was different for each patient, depending on the individual optimization of the scan protocol by the MR technologist.

We compared our learning-based VN to the linear PI reconstruction method CG SENSE (12) and a combined PI-CS non-linear reconstruction method based on Total Generalized Variation (TGV) (22,35). Additionally, we compared our qualitative results to dictionary learning (36) and provide quantitative measures for the selected cases. However, a full comparison to dictionary learning for all cases is out of scope of this work due to the long runtime requirements (approximately 1 h per slice). The forward and adjoint operators for all three reference methods, in particular the coil sensitivity maps, were consistent with our VN approach. All hyper-parameters for CG SENSE and PI-CS TGV such as the number of iterations and regularization parameters were estimated individually by grid search for each sampling pattern, contrast and acceleration factor, such that the MSE of the reconstruction to the gold standard reconstruction was minimized. For dictionary learning, we used the standard parameters as in (36) and estimated the regularization parameter by grid search such that the MSE of the depicted slices was minimized. We assessed the reconstruction results quantitatively in terms of MSE, Normalized Root Mean Square Error (NRMSE), and Structural Similarity Index (SSIM) (37) with $\sigma = 1.5$ on the magnitude images.

In addition to the qualitative and quantitative evaluation, we performed a reader study that compared results from the proposed VN method with results from PI-CS TGV. The 50 test cases from all five sequences were independently reviewed by two fellowship trained musculoskeletal radiologists who were blinded to the MRI reconstruction method. Cases were reviewed in two different sessions, separated by 2 weeks to minimize recall bias. Each session consisted of a random selection of 25 learning and 25 TGV reconstructions. Using a 4-point ordinal scale, reconstructed images were evaluated for sharpness (1: no blurring, 2: mild blurring, 3: moderate blurring, 4: severe blurring), SNR (1: excellent, 2: good, 3: fair, 4: poor), presence of aliasing artifacts (1: none, 2: mild, 3: moderate, 4: severe) and overall image quality (1: excellent, 2: good, 3: fair, 4: poor). Comparisons in terms of image quality scores, averaged over the two readers, were made using a one-sided Wilcoxon signed-rank test. The null hypothesis that PI-CS TGV reconstruction results are equal or better than VN-based results is rejected at significance level $\alpha = 0.05$ if the resulting $P$-value of the test is lower than the significance level $\alpha$.

### Implementation Details

The VN approach as well as the reference methods were implemented in C++/CUDA with CUDNN support. We

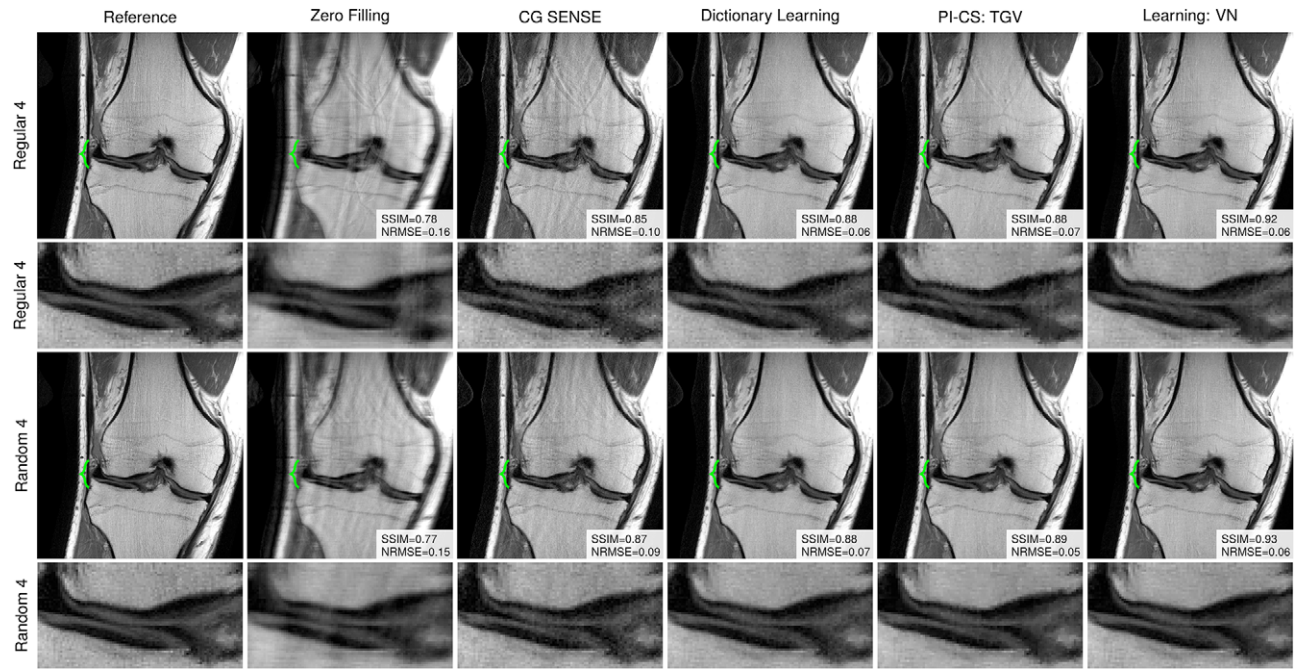


FIG. 3. Coronal PD-weighted scan with acceleration $R=4$ of a 32-year-old male. The green bracket indicates osteoarthritis. The first and third row depict reconstruction results for regular Cartesian sampling, the second and forth row depict the same for variable-density random sampling. Zoomed views show that the learned VN reconstruction appears slightly sharper than the PI-CS TGV and dictionary learning reconstruction. The dictionary learning and VN reconstruction can significantly suppress artifacts unlike CG SENSE and PI-CS TGV. Results based on random sampling show reduced residual artifacts and slightly increased sharpness in comparison to regular sampling.

provide Python and Matlab interfaces for testing. Experiments were performed on a system equipped with an Intel Xeon E5-2698 Central Processing Unit (CPU) (2.30 GHz) and a single Nvidia Tesla M40 Graphics Processing Unit (GPU). For dictionary learning, we used the Matlab implementation provided by the authors (36) and extended their formulation to be used with our multi-coil sampling operator. This requires to solve Equation [7] in their work using the conjugate gradient method which additionally increases runtime. Source code and data are available online.[1]

## RESULTS

### Retrospective Variational Network Reconstructions

Figure 3 display the impact of acceleration factor $R=4$ and sampling patterns for CG SENSE, dictionary learning, PI-CS TGV and our learned VN on coronal PD-weighted images. Additionally, we plot zero filling solutions to illustrate the amount and structure of undersampling artifacts. Difference images to the reference are visualized in Figure 4. The reconstruction results for acceleration factor $R=3$ along with the difference images are illustrated in Supporting Figures S2 and S3. Residual artifacts and noise amplification can be observed for CG SENSE, in particular for $R=4$. In case of acceleration factor $R=3$, the PI-CS image appears less noisy than CG SENSE; however, similar undersampling artifacts are present. For $R=4$ the PI-CS TGV result contains fewer undersampling artifacts than CG SENSE but small details in the image are already lost. Dictionary learning leads to improved removal of undersampling artifacts, resulting in a lower NRMSE than PI-CS TGV for this particular case. The learned VN suppresses these artifacts while still providing sharper and slightly more homogeneous images. Interestingly, dictionary learning as well as the PI-CS TGV and learned VN reconstruction with $R=3$ regular sampling perform slightly better than with variable-density random sampling in terms of intensity homogeneity and sharpness. For acceleration $R=4$, randomness improves the reconstruction results. We depict the reconstruction videos of the whole imaged volume of a 29-year-old female patient for acceleration factor $R=4$ in Supporting Video S1 for regular sampling and in Supporting Video S2 for variable-density random sampling.

Similar observations can be made for coronal PD-weighted scans with fat saturation, as depicted in Figure 5. Again, the reconstruction results for acceleration factor $R=3$ along with the difference images are illustrated in Supporting Figures S4 and S5. The main difference is that this sequence has a lower SNR compared to the non-fat-saturated version. Since additional noise reduces sparsity, the PI-CS TGV reconstructions produce an even more unnatural blocky pattern and contain substantial residual artifacts. The dictionary learning results appear blurrier at image edges and the general reconstruction quality is lowered at this level of SNR, which can best be seen in the error maps in Figure 6 and is supported by the quantitative values for this particular slice. Our learned VN is able to suppress these undersampling artifacts and shows improved image quality at this SNR level as well.

[1]https://github.com/VLOGroup/mri-variationalnetwork

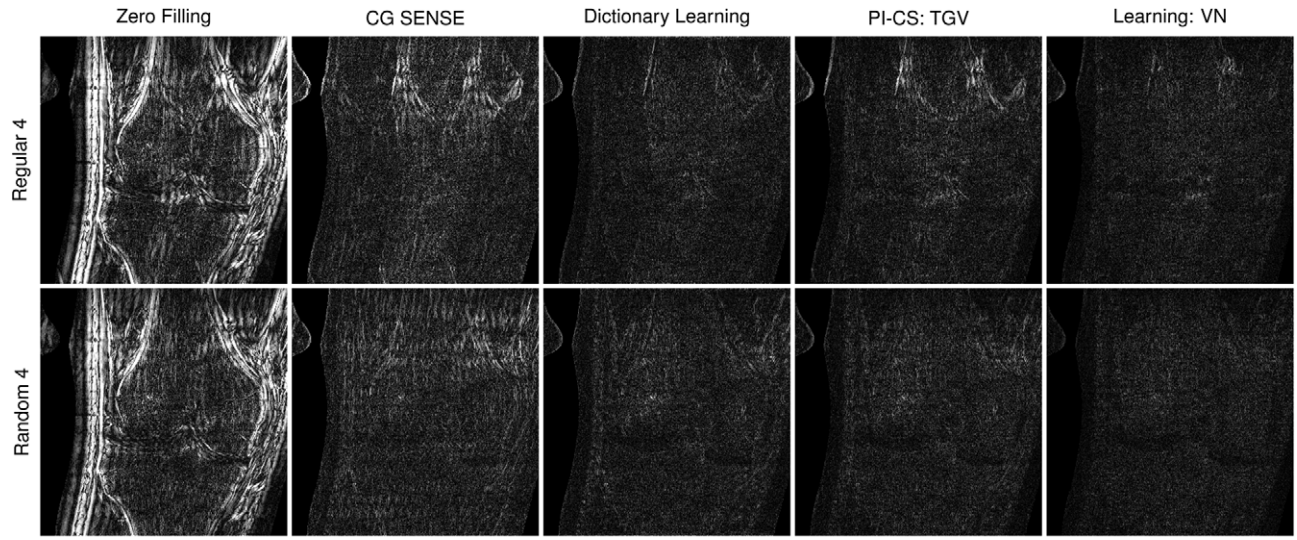


FIG. 4. Difference images to reference image for the reconstructed coronal PD-weighted scans with acceleration $R = 4$ presented in Figure 3. The undersampling artifacts can be clearly observed in the CG SENSE and zero filling results. While PI-CS TGV has a remaining undersampling artifact for regular sampling, the dictionary learning method can suppress this artifact. However, we observe larger errors at object boundaries in the dictionary learning results. The VN result has the least error compared to the reference methods.

All our observations are supported by the quantitative evaluation depicted in Table 1 for $R = 4$ and in Supporting Table S1 for $R = 3$. The wide range in quantitative values over the different sequences illustrates the effect of SNR on the reconstructions. The learned VN reconstructions show superior performance in terms of MSE, NRMSE and SSIM in all cases. Table 1 and Supporting Table S1 also supports the qualitative impression that there is no improvement using variable-density random sampling for $R = 3$ for PI-CS TGV and VN reconstruction. In contrast, random sampling outperforms regular sampling for $R = 4$ in all coronal cases.

We illustrate results for individual scans with regular sampling of $R = 4$ for a complete knee protocol, which

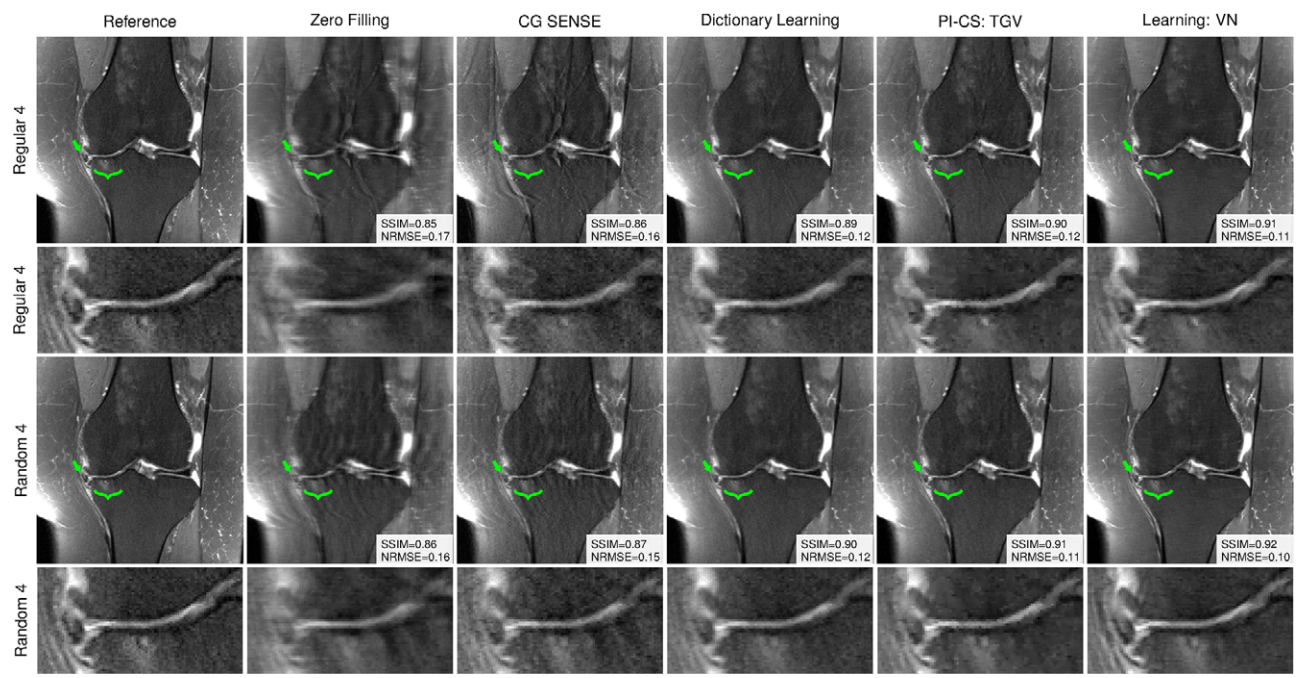


FIG. 5. Coronal fat-saturated PD-weighted scan with acceleration $R = 4$ of a 57-year-old female. The green bracket indicates broad-based, full-thickness chondral loss and a subchondral cystic change. The green arrow depicts an extruded and torn medial meniscus. The first and second row depicts reconstruction results for regular Cartesian sampling, the third and fourth row depict the same for variable-density random sampling. The zoomed views show that the learned VN reconstruction appears sharper than the PI-CS TGV and dictionary learning reconstruction. The VN reconstruction shows reduced artifacts compared to the other methods. Results based on random sampling show reduced residual artifacts and appear sharper than the results based on regular sampling.

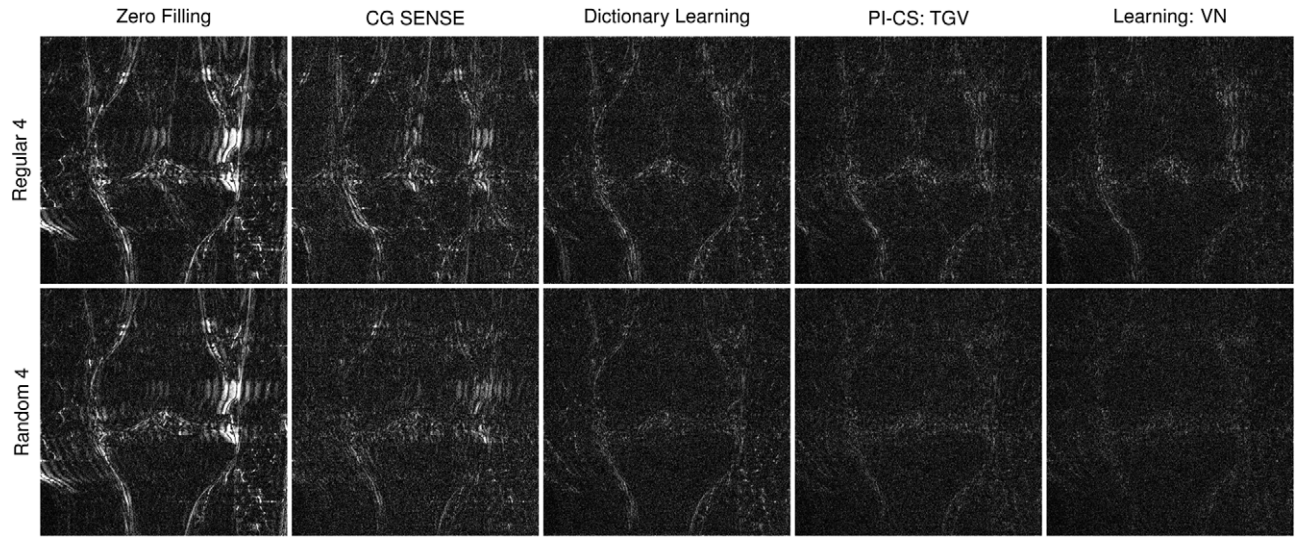


FIG. 6. Difference images to reference image for the reconstructed coronal fat-saturated PD-weighted scans with acceleration $R = 4$ presented in Figure 5. The undersampling artifacts can be clearly observed in the CG SENSE and zero filling results. Both PI-CS TGV and dictionary learning have residual undersampling artifact for regular sampling. We observe larger errors at object boundaries in the dictionary learning results. The VN result has the least error compared to the reference methods and is able to suppress the undersampling artifacts.

contains various pathologies, taken from subjects ranging in age from 15 to 57, and anatomical variants, including a pediatric case. In particular, the coronal PD-weighted scan (M32) depicted in Figure 3 shows osteoarthritis, most advanced within the lateral tibiofemoral compartment with associated marginal osteophyte formation, indicated by the green bracket. An extruded and torn medial meniscus, indicated by the green arrow, is visible in the coronal fat-saturated PD-weighted scan in Figure 5. Additionally, this patient (F57) has broad-based, full-thickness chondral loss within the medial compartment and a subchondral cystic change underlying the medial tibial plateau, as indicated by the green bracket. Further results for different orientations and contrasts are illustrated in Figure 7 for regular sampling with $R = 4$ along with the error maps in Supporting Figure S6. The sagittal PD-weighted scan illustrates a skeletally immature patient (F15) with almost completely fused tibial physes. A partial tear of the posterior cruciate ligament is visible in the sagittal fat-saturated $T_2$-weighted scan M34. A full-thickness chondral defect centered in the medial femoral trochlea (green arrow) is visible on the axial fat-saturated $T_2$-weighted scan (F45) on a background of patellofemoral osteoarthritis. A reconstruction video of all available image slices for the axial fat-saturated $T_2$-weighted case is shown in Supporting Video S3.

The presence of these particular variations, which were not included in the training data set, does not negatively affect the learned reconstruction. The reduction of residual aliasing artifacts, marked by yellow arrows, the reduced noise level, and the overall improved image quality lead to improved depiction of the pathologies when compared to the reference methods. Again, the quality improvement of the learned VN is supported by the quantitative analysis of similarity measures depicted in Table 1 and Supporting Table S1.

### Prospective Variational Network Reconstructions

The reconstruction results of prospectively undersampled data for regular sampling and acceleration $R = 4$ are depicted in Figure 8. We observe a similar behavior of the reconstruction methods as for the retrospectively undersampled data. While PI-CS TGV and dictionary learning perform reasonably well for non-fat-saturated scans, a noise pattern can be observed in certain regions for dictionary learning and blocky appearance for PI-CS TGV. Our VN reconstructions are more homogeneous and less prone to remaining artifacts.

### Reader Study

The average scores of the readers together with the *P*-values of the Wilcoxon signed-rank test are listed in Table 1. The mean values of the reader scores indicate that all VN reconstructions have equal or better scores than the PI-CS TGV reconstructions. *P*-values indicate that the null hypothesis is rejected for most of the sequences for the given significance level $\alpha$. Coronal as well as sagittal $T_2$ VN reconstructions have significantly better image quality than PI-CS TGV. The difference between the individual reconstruction methods for the sagittal PD case is not significant, which is already obvious in the negligible difference of the qualitative results and quantitative results for this sequence. No significant difference in image quality, except SNR, can be observed for the axial $T_2$-weighted scans.

### Variational Network Parameters

Examples of learned filter kernel pairs for real and imaginary feature planes are plotted along with their corresponding activation and potential functions in Figure 9. The potential functions are computed by integrating the

Table 1
Quantitative Evaluation Results in Terms of MSE, NRMSE and SSIM as well as Image Quality Reader Scores for a Clinical knee Protocol and Acceleration Factor $R = 4$ for Regular Sampling and Variable-Density Random Sampling. For the Reader Scores, we Depict the Mean Values and Standard Deviations Averaged Over Both Readers Along With the p-Value Obtained by the One-Sided Wilcoxon Signed-Rank Test. Values that Accept the Alternative Hypothesis with a Significance Level $\alpha = 0.05$ that VN Reconstructions have a Better Quality Score, are Marked as Bold.

| | | Regular | | | Random | | | | Reader scores regular | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data set | Method | MSE | NRMSE | SSIM in % | MSE | NRMSE | SSIM in % | Criterion | PI-CS TGV | Learning | *P*-value |
| Coronal PD | Zero filling | 19.41 ± 4.43 | 0.17 ± 0.02 | 79.00 ± 2.36 | 15.83 ± 3.68 | 0.16 ± 0.02 | 80.64 ± 2.41 | Artifact | 3.60 ± 0.57 | 1.65 ± 0.07 | **0.0010** |
| | CG SENSE | 5.20 ± 0.97 | 0.16 ± 0.03 | 84.01 ± 2.21 | 4.26 ± 0.98 | 0.15 ± 0.03 | 85.57 ± 2.29 | Sharpness/Blur | 2.90 ± 0.14 | 2.15 ± 0.07 | **0.0234** |
| | PI-CS TGV | 2.35 ± 0.40 | 0.09 ± 0.02 | 89.80 ± 1.75 | 1.91 ± 0.45 | 0.09 ± 0.02 | 90.36 ± 1.79 | SNR | 2.60 ± 0.28 | 1.45 ± 0.21 | **0.0078** |
| | Learning | **1.64 ± 0.28** | **0.08 ± 0.02** | **92.14 ± 1.68** | **1.37 ± 0.32** | **0.08 ± 0.02** | **92.86 ± 1.63** | Overall image quality | 3.30 ± 0.14 | 2.05 ± 0.21 | **0.0010** |
| Coronal | Zero filling | 20.71 ± 4.07 | 0.23 ± 0.03 | 73.96 ± 3.04 | 17.69 ± 3.30 | 0.22 ± 0.03 | 75.10 ± 3.17 | Artifact | 3.95 ± 0.07 | 2.90 ± 0.42 | **0.0020** |
| fat-sat. PD | CG SENSE | 14.55 ± 1.62 | 0.25 ± 0.05 | 73.06 ± 4.62 | 11.79 ± 1.39 | 0.24 ± 0.04 | 74.78 ± 4.55 | Sharpness/Blur | 3.95 ± 0.07 | 3.15 ± 0.64 | **0.0020** |
| | PI-CS TGV | 7.73 ± 1.14 | 0.19 ± 0.04 | 79.19 ± 4.14 | 7.07 ± 1.07 | 0.18 ± 0.03 | 79.69 ± 4.09 | SNR | 3.75 ± 0.21 | 2.90 ± 0.71 | **0.0049** |
| | Learning | **6.49 ± 0.80** | **0.17 ± 0.03** | **81.97 ± 3.60** | **5.81 ± 0.85** | **0.17 ± 0.03** | **82.47 ± 3.67** | Overall image quality | 3.95 ± 0.07 | 3.20 ± 0.57 | **0.0020** |
| Sagittal | Zero filling | 16.66 ± 3.14 | 0.19 ± 0.03 | 85.71 ± 2.62 | 17.35 ± 3.21 | 0.19 ± 0.03 | 84.91 ± 2.59 | Artifact | 2.90 ± 0.14 | 2.80 ± 0.28 | 0.3750 |
| fat-sat. $T_2$ | CG SENSE | 6.27 ± 1.62 | 0.15 ± 0.04 | 87.86 ± 3.08 | 9.55 ± 2.11 | 0.18 ± 0.04 | 85.06 ± 3.11 | Sharpness/Blur | 3.40 ± 0.14 | 2.75 ± 0.21 | **0.0156** |
| | PI-CS TGV | 3.39 ± 0.82 | **0.11 ± 0.03** | 91.84 ± 2.81 | 4.76 ± 0.95 | 0.13 ± 0.03 | 90.29 ± 2.70 | SNR | 3.20 ± 0.28 | 2.50 ± 0.28 | **0.0234** |
| | Learning | **2.99 ± 0.68** | **0.11 ± 0.03** | **92.83 ± 2.40** | **3.92 ± 0.81** | **0.12 ± 0.03** | **91.85 ± 2.35** | Overall image quality | 3.30 ± 0.28 | 2.75 ± 0.07 | **0.0078** |
| Sagittal PD | Zero filling | 5.17 ± 0.75 | 0.11 ± 0.01 | 87.53 ± 1.95 | 3.32 ± 0.51 | 0.09 ± 0.01 | 89.49 ± 1.80 | Artifact | 2.10 ± 0.14 | 2.00 ± 0.14 | 0.4063 |
| | CG SENSE | 0.86 ± 0.15 | 0.06 ± 0.02 | 92.74 ± 1.46 | 1.03 ± 0.16 | 0.07 ± 0.02 | 92.37 ± 1.48 | Sharpness/Blur | 2.10 ± 0.14 | 2.10 ± 0.14 | 0.6875 |
| | PI-CS TGV | 0.49 ± 0.09 | 0.05 ± 0.01 | 96.22 ± 1.17 | 0.64 ± 0.11 | **0.05 ± 0.01** | 95.47 ± 1.24 | SNR | 1.60 ± 0.00 | 1.50 ± 0.28 | 0.3828 |
| | Learning | **0.44 ± 0.07** | **0.04 ± 0.01** | **96.64 ± 1.16** | **0.52 ± 0.09** | **0.05 ± 0.01** | **96.07 ± 1.17** | Overall image quality | 2.20 ± 0.14 | 2.05 ± 0.07 | 0.2656 |
| Axial fat-sat. $T_2$ | Zero filling | 44.57 ± 9.95 | 0.27 ± 0.02 | 78.52 ± 1.92 | 48.03 ± 11.13 | 0.28 ± 0.02 | 77.80 ± 1.98 | Artifact | 3.15 ± 0.07 | 3.10 ± 0.57 | 0.5000 |
| | CG SENSE | 23.75 ± 4.56 | 0.24 ± 0.03 | 80.30 ± 3.20 | 31.98 ± 4.88 | 0.27 ± 0.02 | 78.87 ± 2.43 | Sharpness/Blur | 3.05 ± 0.07 | 2.95 ± 0.49 | 0.3750 |
| | PI-CS TGV | 13.65 ± 3.78 | 0.18 ± 0.03 | 85.51 ± 3.25 | 15.30 ± 2.57 | 0.19 ± 0.02 | 84.93 ± 2.60 | SNR | 3.10 ± 0.14 | 2.75 ± 0.49 | **0.0313** |
| | Learning | **10.63 ± 2.48** | **0.16 ± 0.02** | **88.46 ± 2.43** | **12.06 ± 2.13** | **0.17 ± 0.02** | **87.74 ± 2.30** | Overall image quality | 3.20 ± 0.14 | 3.05 ± 0.49 | 0.2266 |

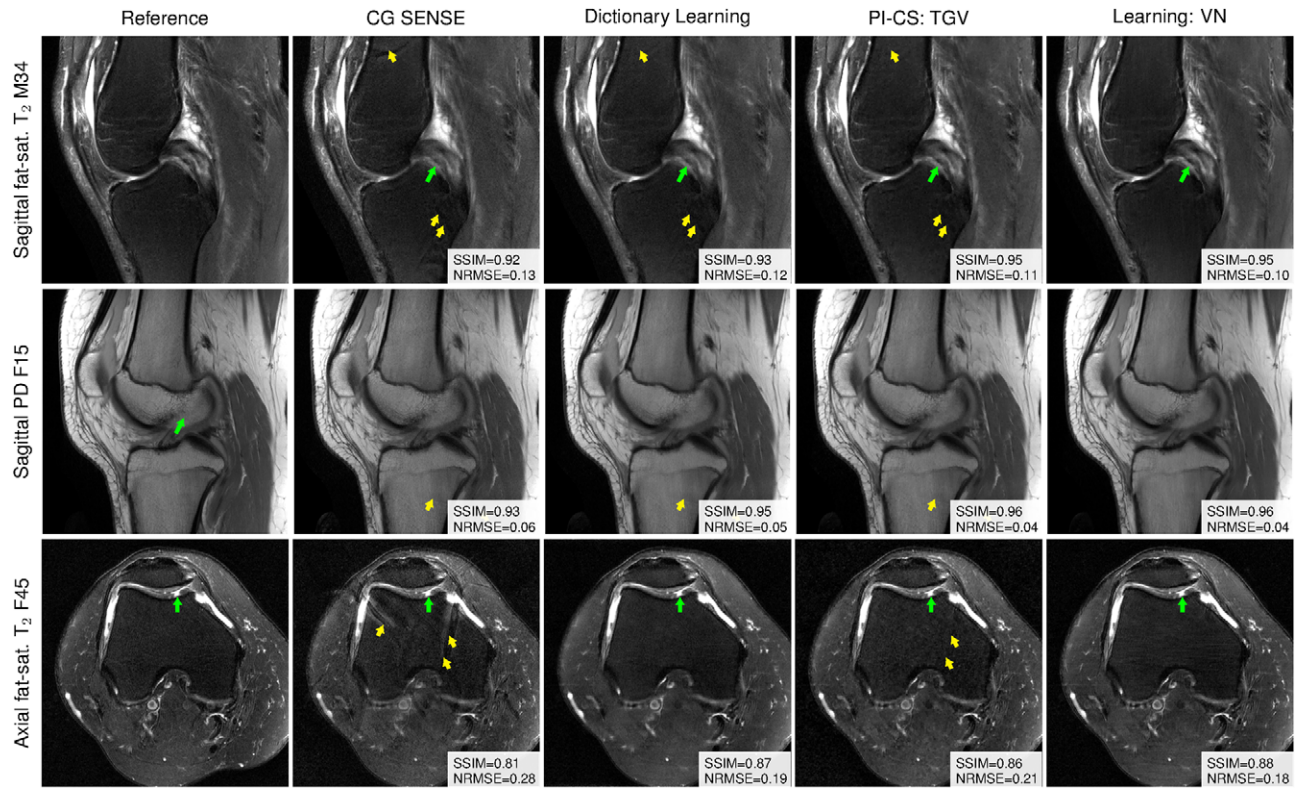


FIG. 7. Reconstruction results for sagittal fat-saturated $T_2$-weighted, sagittal PD-weighted and axial fat-saturated $T_2$-weighted sequences of a complete knee protocol for acceleration factor $R = 4$ with regular undersampling. Each sequence here is illustrated with results from a different patient, identified by gender and age (e.g., M50 indicates a 50-year-old male). Pathological cases and a pediatric case are shown for both male and female patients of various ages. Green arrows and brackets indicate pathologies. Yellow arrows show residual artifacts that are visible in the different reconstructions, but not in the learned VN reconstructions.

learned activation functions, and they can be linked directly to the norms that are used in the regularization terms of traditional CS algorithms. We observe that some potential functions are very close to the convex $l_1$ norm used in CS (e.g., the function in the 2nd column), but we can also observe substantial deviations. We can identify functions with student-t characteristics and concave functions. Some of the learned filter pairs have the same structure in both the real and imaginary plane while some of them seem to be inverted in the real and imaginary part.

## DISCUSSION

While deep learning has resulted in clear breakthroughs in Computer Vision, the application of deep learning to medical image reconstruction is just beginning (38). Initial results for our deep learning image reconstruction approach presented in detail here were first presented at the Annual Meeting of the International Society for Magnetic Resonance in Medicine in May of 2016 (39). Early attempts to use machine learning for MRI reconstruction were based on dictionary learning (36,40,41). The key difference to our VN approach is that they learn a reconstruction online as a combination of dictionary elements directly from undersampled data, hence, no reference data is required. Although the learned dictionary might be reused, a new optimization problem has to be performed for every new reconstruction, which is computationally demanding. While dictionary learning methods act on patches, which need to be properly combined, and do not involve non-linearities in the combination of dictionary elements, our proposed VN approach directly reconstructs the whole images and learns non-linearities, which are important to enhance or suppress certain filter responses. Wang et al. (42) showed first results using convolutional neural network (CNN) architecture to define a relationship between zero filled solution and high-quality images based on pseudo-random sampling. The learned network can then be used as regularization in a non-linear reconstruction algorithm. Yang et al. (43) introduced a network architecture that is based on unrolling the Alternating Direction Method of Multipliers algorithm. They proposed to learn all parameters including image transforms and shrinkage functions for CS-based MRI. Han et al. (44) learned destreaking on CT images and then fine-tuned the learning on MR data to remove streaking from radially undersampled k-space data. All three approaches used single-coil data, and it remains unclear how they deal with the complex domain of MR images. Kwon et al. (45) introduced a neural network architecture to estimate the unfolding of multi-coil Cartesian undersampled data. Similar to a classic SENSE reconstruction (12), unfolding is performed line-by-line. This restricts the applicability to a fixed matrix size and a particular 1D undersampling pattern. Most recently, Lee et al. (46) used residual learning to train two CNNs

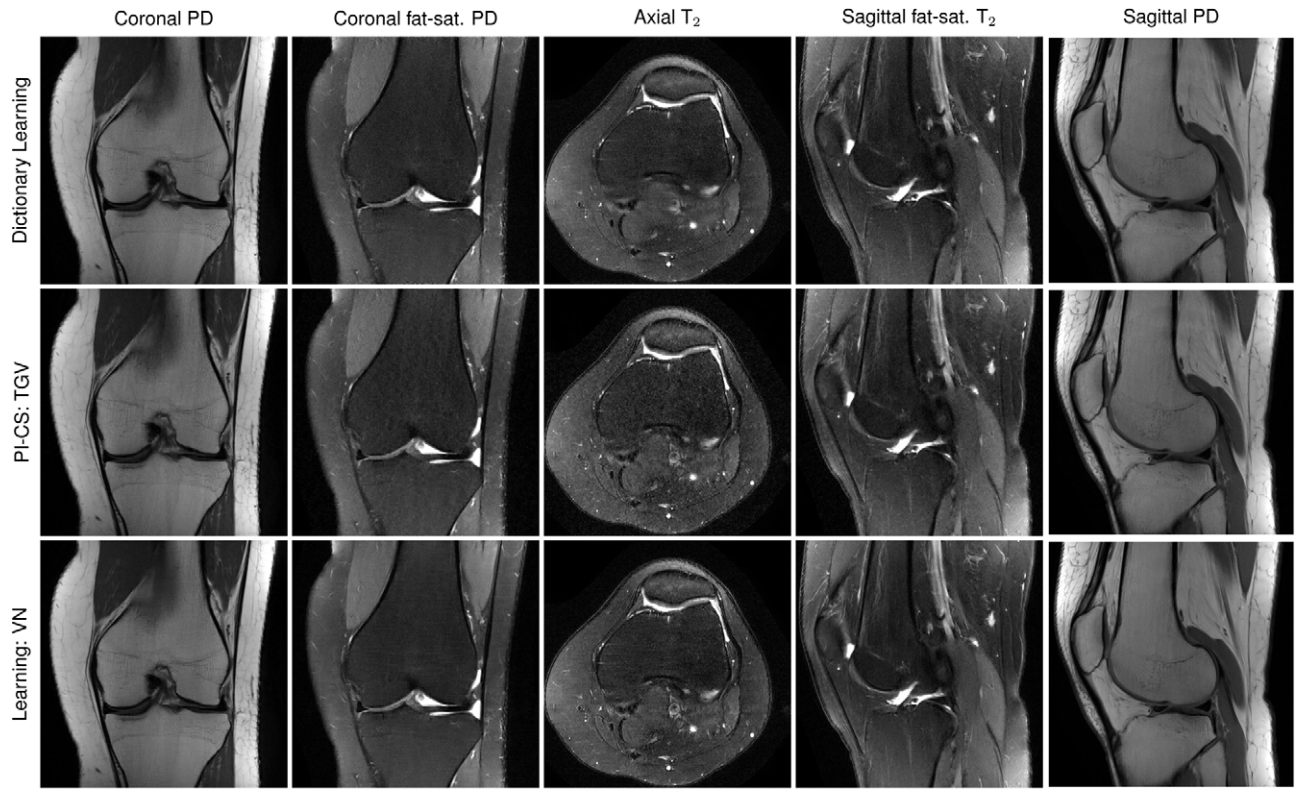


FIG. 8. Reconstruction results of prospectively undersampled data for regular sampling $R = 4$. We show reconstruction results for dictionary learning, PI-CS TGV and our VN for a whole knee protocol of a 27-year-old female volunteer. We observe a similar behavior as for the retrospectively undersampled data. Dictionary learning and PI-CS TGV perform reasonably well for non-fat-saturated scans. While the fat-saturated scans appear artificial with a PI-CS TGV reconstruction, we observe a noise pattern in the dictionary learning results, most prominent in the sagittal fat-saturated $T_2$-weighted scan. Dictionary learning appears slightly blurrier, which is best seen in the axial slice. The VN reconstructions have less undersampling artifacts and an improved SNR.

to estimate the magnitude and phase images of Cartesian undersampled data.

In this work, we present the first learning-based MRI reconstruction approach for clinical multi-coil data. Our VN architecture combines useful properties of two successful fields: variational methods and deep learning. We formulate image reconstruction as a variational model and embed this model in a gradient descent scheme, which forms the specific VN structure. The VN was first introduced as a trainable reaction-diffusion

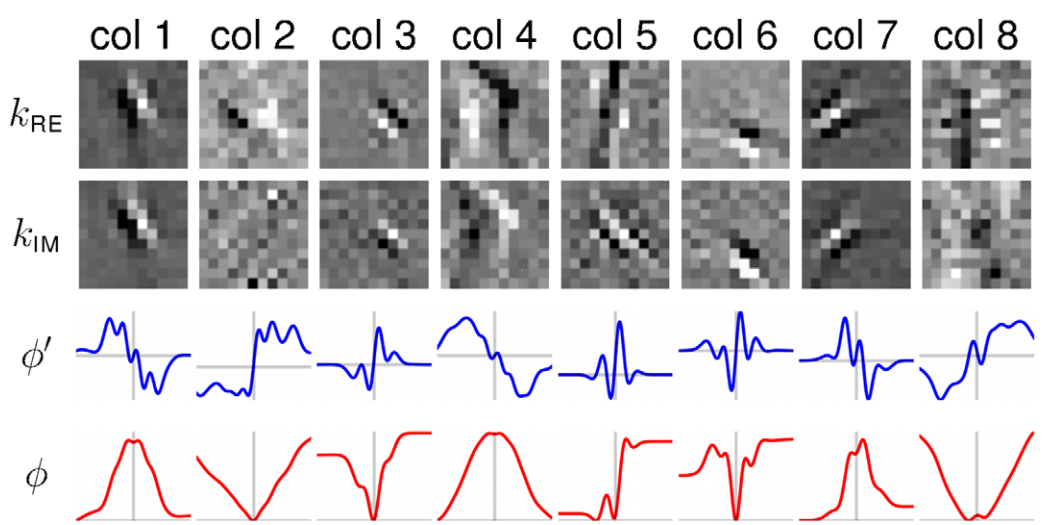


FIG. 9. Examples of learned parameters of the VN. Filter kernels for the real $k_{re}$ and imaginary $k_{im}$ plane as well as their corresponding activation $\phi'$ and potential function $\phi$ are shown. The potential function $\phi$ was obtained by integrating the activation function $\phi'$ including an additional integration constant.

model (6) with application to classic image processing tasks (6,29,30). All these tasks are similar in the sense that the data are corrupted by unstructured noise in the image domain. MR image reconstruction presents several substantial differences: complex-valued multi-coil data are acquired in the Fourier domain and transformed into the image domain. This involves the use of coil sensitivity maps and causes distinct artifacts related to the sampling pattern. For our MR image reconstruction task, the optimal design of the VN, such as the number of stages, the number of filters per stage and the kernel size, is currently an open question. Our particular design choice is based on preliminary experiments (39) and, in line with the experiments presented here, delivered consistent results for a wide range of experimental conditions. We also found that the performance of our VN was stable when varying the design of the architecture. In practice, the design of the network is essentially a trade-off between model complexity and training efficiency. For example, the number of RBFs that are used to model the activation functions in a smoothed function approximation defines the flexibility to approximate arbitrary functions in an accurate way. In our experimental setup as well as in the latest studies on image processing tasks (32), we reduced the number of RBFs compared to the initial work (6) by a half without a loss in performance but with reduced training time.

Our VN structure allows us to visualize the learned parameters, which is non-trivial for classical CNNs (47). In general, the filters in both the real and imaginary part represent different (higher-order) derivative filters of various scales and orientations, similar to Gabor filters (48,49). Handcrafted Gabor filters have been successfully used in image processing (50), and learning-based approaches (3) report similar filters. It has also been shown that these types of filters have a strong relation to the human perceptual system (51).

Some of the learned potential functions in Figure 9 are very close to the convex $l_1$ norm used in CS (e.g., the function in the 2nd column), but we can also observe substantial deviations. We can identify functions with student-t characteristics also used in (28). Indeed, non-convex functions of student-t type introduce more sparsity than, for example, the convex $l_1$-norm and are reported to fit the statistics of natural images better than the $l_1$-norm (52). Potential functions like those in columns 1, 4, and 7 have been associated with image sharpening in the literature (53).

Designing filters and functions is not a trivial task. Using learning-based approaches provides a way to tune these parameters such that they are adapted to specific types of image features and artifact properties. The strength of our algorithm are the trainable activation functions which stands in contrast to other deep learning approaches that use fixed activation functions such as Rectified Linear Units or sigmoid functions. Hence, instead of adding more and more layers and creating deeper networks, we introduce more structure and flexibility in the individual layers, which might help to reduce the overall complexity of the network. As shown in (32) for image denoising and non-blind deblurring, fixing the activation functions to less flexible, for example, convex, functions might also lead to a decrease in performance for our application.

Compared to convex L1 minimization where we can understand the characteristics and artifacts of handcrafted filters and potential functions, learning-based methods are often considered to be black-boxes, which are difficult to interpret. While we cannot claim insight into the properties of the model and the resulting images to the same degree of a simpler model like TV, one of the key strengths of our proposed VN is the motivation by a generalized, trainable variational model. To gain an understanding of what the VN learns, we first inspect the intermediate outputs of the gradient descent steps of our VN (see Supporting Video S4). We observe successive low-pass and high-pass filtering, and note that the prevalence of undersampling artifacts decreases after each single iteration. A continuous improvement over the iterations does not occur because our training is designed such that the result after the last gradient step is optimal in terms of the error metric chosen for evaluation. Although it would be possible to train the VN for progressive improvement, this would reduce the flexibility of the algorithm for adjusting the learned parameters during the training procedure.

In any iterative CS approach, every reconstruction is handled as an individual optimization problem. This is a fundamental difference to our proposed data-driven VN. In our VN approach, we perform the computationally expensive optimization as an offline pre-computation step to learn a set of parameters for a small fixed number of iterations. In our experiments, one training took approximately 4 days on a single graphics card. Once the VN is trained, the application to new data is extremely efficient, because no new optimization problem has to be solved and no additional parameters have to be selected. In our experiments, the VN reconstruction took only 193 ms for one slice. In comparison, the reconstruction time for zero filling was 11 ms, for CG SENSE with 6 iterations 75 ms and for PI-CS TGV with 1000 primal–dual iterations (22) 11.73 s on average. Thus, the online VN reconstruction using the learned parameters for the fixed number of iterations does not affect the hard time constraints during a patient exam.

Our VN is individually trained for different sampling patterns, reflected in the forward and adjoint operators. We do not learn a global mapping between undersampled k-space and reconstruction, but how to enhance local structures, while ensuring consistency to the acquired k-space data. First results towards learning a general regularizer, that could be applied for any sampling pattern, were recently presented at the annual meeting of ISMRM in 2017 (54): We showed that a network trained for regular sampling patterns can be used for reconstruction of randomly sampled data, but a network trained for randomly sampled data is not capable of removing coherent undersampling artifacts, which indicates that the dependency of sampling patterns is required to train the regularizer. However, the systematic performance evaluation for a wide range of sampling patterns is beyond the scope of this particular manuscript,

and will be the target of future work. We will not only explore joint training of various sampling patterns, acceleration factors and sequences, but also the application of VN reconstruction to non-Cartesian sampling, dynamic and multi-parametric data.

The reconstruction quality of all methods does not only rely on the sampling pattern, but also on other parameters. Larger filter sizes, such as the $11 \times 11$ filters used in our VN architecture, provide the possibility to capture more efficiently the characteristic backfolding artifacts of Cartesian undersampled data, which are spread over several pixels. This stands in contrast to models like TV or TGV that are based on gradient filters in a small neighborhood (e.g., only forward differences in the $x$ and $y$ direction are considered). To suppress artifacts with PI-CS TGV, the regularization parameters must be chosen in such a way that the remaining image appears over-smoothed, and fine details are lost. Even though the piecewise-affine prior model of TGV is more complex than the piecewise-constant prior model of TV, the images appear artificial, especially if MR images with low SNR are reconstructed. Dictionary learning involves also larger filter kernels and works reasonably well for data with high SNR, reconstructions of low SNR data contain lots of noisy regions and blurry edges.

The image quality reader study confirms our quantitative and qualitative observations for regular sampling of $R=4$. In general, the image quality of the fat-saturated sequences was rated lower than for the non-fat-saturated sequences for both VN and PI-CS TGV. The difference between the two types of sequences is the baseline SNR, which is much lower for the fat-saturated sequences. It is well known that in all CS-based methods, the best performance can be achieved in the case of a high baseline SNR and incoherent artifacts. The presented experiments demonstrate that if the corruption of the reconstructed images is dominated by noise, performance of both CS and VN reconstruction drops. If the baseline SNR drops to a level where the noise has a higher impact than aliasing artifacts, the VN concentrates on denoising instead of undersampling artifact removal. In addition, some of our results show residual artifacts, most prominent in the axial sequences. The source of these artifacts is residual aliasing and Gibbs' ringing. These residual artifacts are present in all our reconstructions and not unique for our VN.

While radiologists learn throughout their career to distinguish certain patterns in images such as artifacts, we have to reflect the quality of learning in our presented approach by not only choosing the right architecture but also a suitable similarity measure. As demonstrated by our evaluation, quantitative scores are not always on par with image quality readings by radiologists. The used MSE for training compares pixel-wise differences and is likely not optimal for representing similarity to artifact-free reference reconstructions. Future investigations will also involve the choice of different error metrics or the investigation of generative adversarial networks (55) for training.

## CONCLUSION

Inspired by variational models and deep learning, we present a new approach, termed VN, for efficient reconstruction of complex multi-coil MR data. We learn the whole reconstruction procedure and all associated model parameters in an offline training step on clinical patient data sets. The VN-based reconstructions preserve important features not presented in the training data. Our proposed learning-based VN reconstruction approach outperforms traditional reconstructions for a wide range of pathologies and offers high reconstruction speed, which is substantial for integration into clinical workflow.

## ACKNOWLEDGMENTS

We acknowledge grant support from the Austrian Science Fund (FWF) under the START project BIVISION, No. Y729, the European Research Council under the Horizon 2020 program, ERC starting grant "HOMOVIS", No. 640156, and from the US National Institutes of Health (NIH P41 EB017183, NIH R01 EB000447), as well as hardware support from Nvidia corporation. We would like to thank Dr. Tobias Block for his support with the Yarra Framework, Dr. Elisabeth Garwood for helping us with clinical evaluation, and Ms. Mary Bruno for assistance with the data acquisition. We thank Dr. Elisabeth Garwood and Dr. Gina Ciavarra for serving as readers in our clinical reader study.

## APPENDIX A

### INERTIAL INCREMENTAL PROXIMAL GRADIENT ALGORITHM (IIPG)

For network training, we consider following optimization problem:

$$
\begin{aligned}
\mathcal{L}(\theta) &= \min_{\theta} \frac{1}{2S} \sum_{s=1}^{S} \left\| |\mathbf{u}_s^T(\theta)|_\epsilon - |\mathbf{g}_s|_\epsilon \right\|_2^2 \\
\theta &= \{\theta^0, \ldots, \theta^{T-1}\}, \quad \theta^t = \{w_{ij}^t, \mathbf{k}_i^t, \lambda^t\} \\
\mathbf{u}_s^{t+1} &= \mathbf{u}_s^t - \sum_{i=1}^{N_k} (\mathbf{K}_i^t)^\top \Phi_i'^t(\mathbf{K}_i^t \mathbf{u}_s^t) - \lambda^t \mathbf{A}^*(\mathbf{A}\mathbf{u}_s^t - \mathbf{f}_s), \\
& 0 \le t \le T-1 \\
\text{s.t.} \quad \theta \in \mathcal{C} &= \left\{ \lambda^t \ge 0,\, \xi_{\text{re}}^\top \mathbf{k}_i^t = 0,\, \xi_{\text{im}}^\top \mathbf{k}_i^t = 0,\, \left\|\mathbf{k}_i^t\right\|_2 = 1 \right\}.
\end{aligned}
$$

To solve this highly non-convex training problem, we use the Inertial Incremental Proximal Gradient (IIPG) optimizer. This IIPG variant of projected gradient descent is related to the Inertial Proximal Alternating Linearized Minimization (IPALM) algorithm (31). The whole sequence generated by IPALM is guaranteed to converge to a stationary point in the non-convex non-stochastic case under certain constraints on the step size and inertial parameters. The analysis for the stochastic version is left to future research. In the IIPG Algorithm 1, the parameter updates are calculated in a stochastic way on a single mini batch. First, we perform over-relaxation where we set an over-relaxation constant $\beta_e$ dependent on the current epoch $e$ to achieve moderate

acceleration. Second, we compute the gradient with respect to the parameters on the current mini batch which yields a new parameter update $\tilde{\theta}^{m+1}$ for the current iteration $m$. To realize additional constraints on the parameters, we finally perform the projections

$$(\lambda^{m+1}, \mathbf{k}^{m+1}) = \text{proj}_{\mathcal{C}}^{\eta}\left(\tilde{\lambda}^{m+1}, \tilde{\mathbf{k}}^{m+1}\right).$$

As the constraints do not depend on each other, we can consider the projections independently. To realize the non-negativity constraint on the data term weights $\lambda^{m+1}$, the parameter update $\tilde{\lambda}^{m+1}$ is clamped at zero

$$\lambda^{m+1} = \max(0, \tilde{\lambda}^{m+1}).$$

For the projection onto the filter kernel constraints, we first subtract the means ${\xi_{\text{re}}}^{\top}\tilde{\mathbf{k}}^{m+1}$, ${\xi_{\text{im}}}^{\top}\tilde{\mathbf{k}}^{m+1}$ from the current kernel parameter estimates and then project the kernel onto the unit-sphere

$$\begin{aligned}\tilde{\mathbf{k}}_{\xi}^{m+1} &= (\tilde{\mathbf{k}}_{\xi,\text{re}}^{m+1}, \tilde{\mathbf{k}}_{\xi,\text{im}}^{m+1}) \\ &= (\tilde{\mathbf{k}}_{\text{re}}^{m+1} - {\xi_{\text{re}}}^{\top}\tilde{\mathbf{k}}^{m+1}, \tilde{\mathbf{k}}_{\text{im}}^{m+1} - {\xi_{\text{im}}}^{\top}\tilde{\mathbf{k}}^{m+1}) \\ \mathbf{k}^{m+1} &= \frac{\tilde{\mathbf{k}}_{\xi}^{m+1}}{\|\tilde{\mathbf{k}}_{\xi}^{m+1}\|_2}.\end{aligned}$$

**Algorithm 1**: Inertial Incremental Proximal Gradient (IIPG) Algorithm

**Input**: Split training set $\mathcal{S}$ into $N_B$ mini batches $\mathcal{B}$ s.t. $\mathcal{S} = \bigcup_{b=1}^{N_B} \mathcal{B}_b$;
**Choose**: Step size $\eta$, number of epochs $N_E$, initial parameters $\boldsymbol{\theta}^0$;
Iteration $m \leftarrow 1$, $\boldsymbol{\theta}^1 \leftarrow \boldsymbol{\theta}^0$;
**for** $e \leftarrow 1$ **to** $N_E$ **do**
  // Over-relaxation constant
  $\beta_e = \frac{e-1}{e+2}$;
  **for** $b \leftarrow 1$ **to** $N_B$ **do**
    // Over-relaxation
    $\hat{\boldsymbol{\theta}}^{m+1} = \boldsymbol{\theta}^m + \beta_e(\boldsymbol{\theta}^m - \boldsymbol{\theta}^{m-1})$;
    // Compute gradient on current mini batch $\mathcal{B}_b$
    $\mathbf{g}^{m+1} = \frac{\partial \mathcal{L}(\hat{\boldsymbol{\theta}}^{m+1})}{\partial}$;
    // Compute gradient step
    $\tilde{\boldsymbol{\theta}}^{m+1} = \hat{\boldsymbol{\theta}}^{m+1} - \eta\mathbf{g}^{m+1}$;
    // Compute projections
    $\boldsymbol{\theta}^{m+1} = \text{proj}_{\mathcal{C}}^{\eta}(\tilde{\boldsymbol{\theta}}^{m+1})$;
    $m \leftarrow m+1$;
  **end**
**end**

## APPENDIX B

## GRADIENT DERIVATION OF NETWORK PARAMETERS

In every gradient step $t$, we seek the derivatives with respect to the parameters $\theta^t = \{w_{ij}, \mathbf{k}_i^t, \lambda^t\}$ of the loss function

$$\mathcal{L}(\theta) = \min_{\theta} \frac{1}{2S}\sum_{s=1}^{S} \| |\mathbf{u}_s^T(\theta)|_{\epsilon} - |\mathbf{g}_s|_{\epsilon} \|_2^2, \; |\mathbf{x}|_{\epsilon} = \sqrt{\mathbf{x}_{\text{re}}^2 + \mathbf{x}_{\text{im}}^2 + \epsilon}$$

where $|\cdot|_{\epsilon}$ is understood in a point-wise manner. For simplicity, we drop the dependency of $\mathbf{u}^T$ on the parameters $\theta$ and the subscript $s$ and show the calculations only for a single training example. The gradient steps are given as

$$\begin{aligned}\mathbf{u}^{t+1} &= \mathbf{u}^t - \sum_{i=1}^{N_k} (\mathbf{K}_i^t)^{\top}\Phi_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t) - \lambda^t\mathbf{A}^*(\mathbf{A}\mathbf{u}^t - \mathbf{f}), \\ &0 \le t \le T-1.\end{aligned}$$

The derivatives with respect to the parameters $\theta^t$ are obtained by back-propagation (33)

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta^t} = \frac{\partial \mathbf{u}^{t+1}}{\partial \theta^t} \cdot \underbrace{\frac{\partial \mathbf{u}^{t+2}}{\partial \mathbf{u}^{t+1}} \cdots \frac{\partial \mathbf{u}^{T}}{\partial \mathbf{u}^{T-1}} \cdot \frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^T}}_{\mathbf{e}^{t+1}}.$$

The reconstruction error of the $t$-th gradient step is given by $\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^{t+1}} = \mathbf{e}^{t+1}$.

### Derivative of the Loss Function

First, we require the gradient of the loss function $\mathcal{L}$ with respect to the reconstruction $\mathbf{u}^T$ defined as $\mathbf{e}^T$. It is computed as

$$\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^T} = \mathbf{e}^T \iff e_l^T = \frac{u_l^T}{|u_l^T|_{\epsilon}}\left(|u_l^T|_{\epsilon} - |g_l|_{\epsilon}\right), \quad l = 1, \ldots, N.$$

### Derivative of the Data Term Weights $\lambda^t$

The derivative of the reconstruction $\mathbf{u}^t$ wrt. to $\lambda^t \in \mathbb{R}$ for the $t$-th gradient step is expressed as:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \lambda^t} = \frac{\partial \mathbf{u}^{t+1}}{\partial \lambda^t}\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^{t+1}} = \langle -(\mathbf{A}^*(\mathbf{A}\mathbf{u}^t - \mathbf{f})), \mathbf{e}^{t+1}\rangle.$$

### Derivative of the Activation Functions $\Phi_i^{t\prime}$

A single activation function $\Phi_i^{t\prime}(\mathbf{z}) = (\phi_i^{t\prime}(z_1), \ldots, \phi_i^{t\prime}(z_N)) : \mathbb{R}^N \mapsto \mathbb{R}^N$ is defined by a weighted combination of $N_w$ Gaussian radial basis functions:

$$\begin{aligned}\phi_i^{t\prime}(z_l) &= \sum_{j=1}^{N_w} w_{ij}^t \exp\left(-\frac{(z_l - \mu_j)^2}{2\sigma^2}\right), \quad l = 1, \ldots, N, \\ &w_{ij}^t \in \mathbb{R}.\end{aligned}$$

This can be rewritten in a matrix-vector notation:

$$\Phi_i^{t\prime}(z) = \begin{pmatrix} \phi_i^{t\prime}(z_1) \\ \vdots \\ \phi_i^{t\prime}(z_N) \end{pmatrix} = \begin{bmatrix} \exp\left(-\frac{(z_1-\mu_1)^2}{2\sigma^2}\right) & \cdots & \exp\left(-\frac{(z_1-\mu_{N_w})^2}{2\sigma^2}\right) \\ \vdots & \ddots & \vdots \\ \exp\left(-\frac{(z_N-\mu_1)^2}{2\sigma^2}\right) & \cdots & \exp\left(-\frac{(z_N-\mu_{N_w})^2}{2\sigma^2}\right) \end{bmatrix} \begin{pmatrix} w_{i1}^t \\ \vdots \\ w_{iN_w}^t \end{pmatrix} = \mathbf{M}_i^t(\mathbf{z})\mathbf{w}_i^t.$$

During training, we learn the weights $\mathbf{w}_i^t \in \mathbb{R}^{N_w}$ and express its gradient as:

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{w}_i^t} &= \frac{\partial \mathbf{u}^{t+1}}{\partial \mathbf{w}_i^t}\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^{t+1}} = -\frac{\partial}{\partial \mathbf{w}_i^t}\left\{(\mathbf{K}_i^t)^\top \mathbf{M}_i^t(\mathbf{K}_i^t\mathbf{u}^t)\mathbf{w}_i^t\right\}\mathbf{e}^{t+1} \\ &= -\left(\mathbf{M}_i^t(\mathbf{K}_i^t\mathbf{u}^t)\right)^\top \mathbf{K}_i^t\mathbf{e}^{t+1}. \end{aligned}$$

### Derivative of the Intermediate Reconstructions $\mathbf{u}^t$

Further gradients with respect to the reconstructions from intermediate steps are given as:

$$\frac{\partial \mathbf{u}^{t+1}}{\partial \mathbf{u}^t} = I - \sum_{i=1}^{N_k} (\mathbf{K}_i^t)^\top \operatorname{diag}\left(\Phi_i^{t\prime\prime}(\mathbf{K}_i^t\mathbf{u}^t)\right)\mathbf{K}_i^t - \lambda^t \mathbf{A}^*\mathbf{A}$$

where $I$ denotes the identity matrix. This also requires the second derivative of the potential functions $\Phi_i^{t\prime\prime}(\mathbf{z})$, which is expressed as:

$$\Phi_i^{t\prime\prime}(\mathbf{z}) = \begin{bmatrix} -\frac{(z_1-\mu_1)}{\sigma^2}\exp\left(-\frac{(z_1-\mu_1)^2}{2\sigma^2}\right) & \cdots & -\frac{(z_1-\mu_{N_w})}{\sigma^2}\exp\left(-\frac{(z_1-\mu_{N_w})^2}{2\sigma^2}\right) \\ \vdots & \ddots & \vdots \\ -\frac{(z_N-\mu_1)}{\sigma^2}\exp\left(-\frac{(z_N-\mu_1)^2}{2\sigma^2}\right) & \cdots & -\frac{(z_N-\mu_{N_w})}{\sigma^2}\exp\left(-\frac{(z_N-\mu_{N_w})^2}{2\sigma^2}\right) \end{bmatrix} \mathbf{w}_i^t$$

### Derivative of the Filter Kernels $k_i^t$

To compute the derivative with respect to the filter kernels $k_i^t$ we have to introduce further relationships between our given parameters. The convolution can be defined as matrix-vector multiplication:

$$k_i^t * u^t \Longleftrightarrow \mathbf{K}_i^t\mathbf{u}^t = \mathbf{U}^t\mathbf{k}_i^t$$

where the matrix $\mathbf{U}^t : \mathbb{R}^{2s^2} \mapsto \mathbb{R}^N$ is a suitably shifted representation of the image $\mathbf{u}^t$ and $\mathbf{k}_i^t \in \mathbb{R}^{2s^2}$ is the vectorized filter kernel. The gradient step also involves rotated filter kernels $\bar{k}_i^t$ due to the transpose operation of the kernel matrix $(\mathbf{K}_i^t)^\top$. As we want to calculate the derivative with respect to $\mathbf{k}_i^t$ and not to their rotated version, we introduce a rotation matrix $\mathbf{R} : \mathbb{R}^{2s^2} \mapsto \mathbb{R}^{2s^2}$ that has the same effect as the transpose operation

$$\overline{\mathbf{k}}_i^t = \mathbf{R}\mathbf{k}_i^t.$$

The convolution can be rewritten as

$$(\mathbf{K}_i^t)^\top \Phi_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t) = \tilde{\Phi}_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)\overline{\mathbf{k}}_i^t = \tilde{\Phi}_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)\mathbf{R}\mathbf{k}_i^t$$

where $\tilde{\Phi}_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t) : \mathbb{R}^N \mapsto \mathbb{R}^{2s^2}$ is a suitable matrix representation of $\Phi_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)$. Applying the product rule yields following expression for the kernel derivative

$$\begin{aligned} \frac{\partial (\mathbf{K}_i^t)^\top \Phi_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)}{\partial \mathbf{k}_i^t} &= \frac{\partial \Phi_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)}{\partial \mathbf{k}_i^t}\mathbf{K}_i^t + \frac{\partial \mathbf{k}_i^t}{\partial \mathbf{k}_i^t}\left[\tilde{\Phi}_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)\mathbf{R}\right]^\top = \\ &(\mathbf{U}^t)^\top \operatorname{diag}\left(\Phi_i^{t\prime\prime}(\mathbf{K}_i^t\mathbf{u}^t)\right)\mathbf{K}_i^t + \mathbf{R}^\top\tilde{\Phi}_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t). \end{aligned}$$

The full derivative may be expressed as

$$\begin{aligned} &\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{k}_i^t} = \frac{\partial \mathbf{u}^{t+1}}{\partial \mathbf{k}_i^t}\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{u}^{t+1}} = \\ &-\left[(\mathbf{U}^t)^\top \operatorname{diag}\left(\Phi_i^{t\prime\prime}(\mathbf{K}_i^t\mathbf{u}^t)\right)\mathbf{K}_i^t + \mathbf{R}^\top\tilde{\Phi}_i^{t\prime}(\mathbf{K}_i^t\mathbf{u}^t)\right]\mathbf{e}^{t+1}. \end{aligned}$$

## REFERENCES


1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.
2. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (NIPS). New York: Curran Associates, Inc.; 2012. pp. 1097–1105.
4. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In International Conference on Learning Representations (ICLR), San Diega, CA, USA, 2015. p. 1–14.
5. Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, van der Smagt P, Cremers D, Brox T. FlowNet: learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015; 2758–2766.
6. Chen Y, Yu W, Pock T. On learning optimized reaction diffusion processes for effective image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015; pp. 5261–5269.

7. Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, Shen D. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage 2015;108:214–224.
8. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Isgum I, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. IEEE Trans Med Imaging 2016; 35:1252–1261.
9. Golkov V, Dosovitskiy A, Sperl JI, Menzel MI, Czisch M, Samann P, Brox T, Cremers D. q-Space deep learning: twelve-fold shorter and model-free diffusion MRI scans. IEEE Trans Med Imaging 2016;35: 1344–1351.
10. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, Biller A. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. Neuroimage 2016;129:460–469.
11. Sodickson DK, Manning WJ. Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays. Magn Reson Med 1997;38:591–603.
12. Pruessmann KP, Weiger M, Scheidegger MB, Boesiger P. SENSE: sensitivity encoding for fast MRI. Magn Reson Med 1999;42:952–962.
13. Griswold MA, Jakob PM, Heidemann RM, Nittka M, Jellus V, Wang J, Kiefer B, Haase A. Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 2002;47:1202–1210.
14. Candes EJ, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inf Theory 2006;52:489–509.
15. Donoho DL. Compressed sensing. IEEE Trans Inf Theory 2006;52: 1289–1306.
16. Lustig M, Donoho D, Pauly JM. Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn Reson Med 2007;58: 1182–1195.
17. Nyquist H. Certain topics in telegraph transmission theory. Trans Am Inst Elect Eng 1928;47:617–644.
18. Shannon CE. Communication in the presence of noise. Proc Inst Radio Eng 1949;37:10–21.
19. Block KT, Uecker M, Frahm J. Undersampled radial MRI with multiple coils: iterative image reconstruction using a total variation constraint. Magn Reson Med 2007;57:1086–1098.
20. Daubechies I. Ten lectures on wavelets. In Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.
21. Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Physica D 1992;60:259–268.
22. Knoll F, Bredies K, Pock T, Stollberger R. Second order total generalized variation (TGV) for MRI. In Proceedings of the 18th Scientific Meeting and Exhibition of ISMRM, Stockholm, Sweden, 2010. vol. 65, pp. 480–491.
23. Knoll F, Clason C, Bredies K, Uecker M, Stollberger R. Parallel imaging with nonlinear reconstruction using variational penalties. Magn Reson Med 2012;67:34–41.
24. Hollingsworth KG. Reducing acquisition time in clinical MRI by data undersampling and compressed sensing reconstruction. Phys Med Biol 2015;60:R297–R322.
25. Landweber L. An iteration formula for Fredholm integral equations of the first kind. Am J Math 1951;73:615–624.
26. Hanke M, Neubauer A, Scherzer O. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. Numer Math 1995;72:21–37.
27. Chambolle A, Pock T. An introduction to continuous optimization for imaging. Acta Numerica 2016;25:161–319.
28. Roth S, Black MJ. Fields of experts. Int J Comput Vis 2009;82:205–229.
29. Klatzer T, Hammernik K, Knöbelreiter P, Pock T. Learning joint demosaicing and denoising based on sequential energy minimization. In Proceedings of the IEEE International Conference on Computational Photography (ICCP), Evanston, IL, USA 2016. p. 1–11.
30. Yu W, Heber S, Pock T. Learning reaction-diffusion models for image inpainting. In Pattern Recognition: 37th German Conference, GCPR 2015, Aachen, Germany, October 7–10, 2015. Cham: Springer, 2015. p. 356–367.
31. Pock T, Sabach S. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J Imaging Sci 2016;9:1756–1787.
32. Kobler E, Klatzer T, Hammernik K, Pock T. Variational networks: connecting variational methods and deep learning. In Proceedings of the German Conference on Pattern Recognition (GCPR), Basel, Switzerland, 2017. p. 281–293.
33. LeCun YA, Bottou L, Orr GB, Müller KR. Efficient backprop. In Neural networks: tricks of the trade. Berlin: Springer; 2012. pp. 9–50.
34. Uecker M, Lai P, Murphy MJ, Virtue P, Elad M, Pauly JM, Vasanawala SS, Lustig M. ESPIRiT–an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. Magn Reson Med 2014;71:990–1001.
35. Bredies K, Kunisch K, Pock T. Total generalized variation. SIAM J Imaging Sci 2010;3:492–526.
36. Ravishankar S, Bresler Y. MR image reconstruction from highly undersampled k-space data by dictionary learning. IEEE Trans Med Imaging 2011;30:1028–1041.
37. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600–612.
38. Wang G. A perspective on deep imaging. IEEE Access 2016;4:8914–8924.
39. Hammernik K, Knoll F, Sodickson DK, Pock T. Learning a variational model for compressed sensing MRI reconstruction. In Proceedings of the 24th Annual Meeting of ISMRM, Singapore, 2016. p. 1088.
40. Caballero J, Price AN, Rueckert D, Hajnal JV. Dictionary learning and time sparsity for dynamic MR data reconstruction. IEEE Trans Med Imaging 2014;33:979–994.
41. Ravishankar S, Bresler Y. Data-driven learning of a union of sparsifying transforms model for blind compressed sensing. IEEE Trans Comput Imaging 2016;2:294–309.
42. Wang S, Su Z, Ying L, Peng X, Zhu S, Liang F, Feng D, Liang D. Accelerating magnetic resonance imaging via deep learning. In IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, 2016. pp. 514–517.
43. Yang Y, Sun J, Li H, Xu Z. Deep ADMM-Net for compressive sensing MRI. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, Editors. Advances in neural information processing systems (NIPS). New York: Curran Associates, Inc.; 2016. pp. 10–18.
44. Han YS, Yoo J, Ye JC. Deep learning with domain adaptation for accelerated projection reconstruction MR. arXiv:170301135 preprint, 2017.
45. Kwon K, Kim D, Seo H, Cho J, Kim B, Park HW. Learning-based reconstruction using artificial neural network for higher acceleration. In Proceedings of the 24th Annual Meeting of ISMRM, Singapore, 2016. p. 1081.
46. Lee D, Yoo J, Ye JC. Deep artifact learning for compressed sensing and parallel MRI. arXiv:170301120 preprint, 2017.
47. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference Computer Vision (ECCV), Zurich, Berlin: Springer, 2014. pp. 818–833.
48. Gabor D. Theory of communication. J Inst Electr Eng 1946, vol. 93, pp. 429–459.
49. Daugman JG. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am 1985;2:1160–1169.
50. Jain AK, Farrokhnia F. Unsupervised texture segmentation using Gabor filters. Pattern Recognit 1990;24:1167–1186.
51. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 1996; 381:607–609.
52. Huang JHJ, Mumford D. Statistics of natural images and models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Fort Collins, 1999. pp. 541–547.
53. Zhu SC, Mumford D. Prior learning and Gibbs reaction-diffusion. IEEE Trans Pattern Anal Mach Intell 1997;19:1236–1250.
54. Hammernik K, Knoll F, Sodickson D, Pock T. On the influence of sampling pattern design on deep learning-based MRI reconstruction. In Proceedings of 25th Annual Meeting of ISMRM, Honolulu, Hawaii, USA, 2017. p. 644.
55. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Adv Neural Inf Process Syst 2014;27:2672–2680.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Fig. S1.** Proposed image reconstruction pipeline: A zero filled solution is computed from the undersampled k-space data by applying the adjoint

operator A*. The adjoint operator A* involves application of coil sensitivity maps. We feed the undersampled k-space data, coil sensitivity maps and the zero filling solution to the VN to obtain a reconstruction. For simplicity, we show the magnitude images, but all the input and output data of the VN are complex-valued.

**Fig. S2.** Coronal PD-weighted scan with acceleration $R = 3$ of a 32-year-old male. The green bracket indicates osteoarthritis. The first and second row depict reconstruction results for regular Cartesian sampling, the third and fourth row depict the same for variable-density random sampling. Zoomed views show that the learned VN reconstruction appears slightly sharper than the PI-CS TGV reconstruction. Although dictionary learning can handle artifacts better than PI-CS TGV and produce a visually more appealing results, the quantitative values are slightly worse. For regular sampling, the results illustrate that the VN reconstruction can suppress undersampling artifacts better than CG SENSE and PI-CS TGV, and works on similar lines with dictionary learning. For this acceleration factor of $R = 3$, the results based on random sampling appear slightly blurrier than the results based on regular sampling.

**Fig. S3.** Difference images to reference image for the reconstructed coronal PD-weighted scans with acceleration $R = 3$ presented in Supporting Figure S2. The VN reconstructions show the least error compared to the other methods.

**Fig. S4.** Coronal fat-saturated PD-weighted scan with acceleration $R = 3$ of a 57-year-old female. The green bracket indicates broad-based, full-thickness chondral loss and a subchondral cystic change. The green arrow depicts an extruded and torn medial meniscus. The first and second row depict reconstruction results for regular Cartesian sampling, the third and fourth row depict the same for variable-density random sampling. The zoomed views show that the learned VN reconstruction appears sharper than the PI-CS TGV and dictionary learning reconstruction. For regular sampling, the results illustrate that the VN reconstruction can suppress undersampling artifacts better. Again, results based on random sampling appear slightly blurrier than the results based on regular sampling.

**Fig. S5.** Difference images to reference image for the reconstructed coronal fat-saturated PD-weighted scans with acceleration $R = 3$ presented in Supporting Figure S4. We observe large errors at boundaries for dictionary learning. The VN reconstructions show the least error compared to the other methods.

**Fig. S6.** Difference images for sagittal fat-saturated $T_2$-weighted, sagittal PD-weighted and axial fat-saturated $T_2$-weighted sequences of a complete knee protocol presented in Figure 7.

**Table S1.** Quantitative evaluation results in terms of MSE, NRMSE and SSIM for a clinical knee protocol and acceleration factor $R = 3$ for regular sampling and variable-density random sampling.

**Video S1.** Reconstruction of a complete imaged volume for a coronal PD-weighted sequence in a 50-year-old male, for regular sampling with acceleration $R = 4$.

**Video S2.** Reconstruction of a complete imaged volume for a coronal PD-weighted sequence in the same 50-year-old male patient as in Supporting Video S1, for variable-density random sampling with acceleration $R = 4$.

**Video S3.** Reconstruction of a complete imaged volume for an axial fat-saturated $T_2$-weighted sequence in a 45-year-old female patient, for regular sampling with acceleration $R = 4$.

**Video S4.** Intermediate gradient step outputs of the reconstruction algorithm for a coronal PD-weighted slice with acceleration $R = 4$. We observe alternating low-pass and high-pass filtering over the intermediate steps. The undersampling artifacts are continuously suppressed until we obtain an artifact-free image after the final step.