Received: 12 December 2017 | Revised: 20 April 2018 | Accepted: 20 April 2018

DOI: 10.1002/mrm.27355

FULL PAPER

Magnetic Resonance in Medicine

Assessment of the generalization of learned image reconstruction and the potential for transfer learning

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

[1] Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, New York, New York

[2] Center for Advanced Imaging Innovation and Research (CAI2R), New York University School of Medicine, New York, New York

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

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

Correspondence
Florian Knoll, Center for Biomedical Imaging (Department of Radiology) and Center for Advanced Imaging Innovation and Research (CAI2R), New York University School of Medicine, 550 1st Avenue, New York, NY 10016.
Email: florian.knoll@nyumc.org

Funding information
National Institutes of Health, Grant/Award number NIH P41 EB017183; Austrian Science Fund (START project BIVISION Y729); European Research Council (Horizon 2020 program); and ERC starting grant "HOMOVIS," Grant/Award number 640156

Purpose: Although deep learning has shown great promise for MR image reconstruction, an open question regarding the success of this approach is the robustness in the case of deviations between training and test data. The goal of this study is to assess the influence of image contrast, SNR, and image content on the generalization of learned image reconstruction, and to demonstrate the potential for transfer learning.

Methods: Reconstructions were trained from undersampled data using data sets with varying SNR, sampling pattern, image contrast, and synthetic data generated from a public image database. The performance of the trained reconstructions was evaluated on 10 in vivo patient knee MRI acquisitions from 2 different pulse sequences that were not used during training. Transfer learning was evaluated by fine-tuning baseline trainings from synthetic data with a small subset of in vivo MR training data.

Results: Deviations in SNR between training and testing led to substantial decreases in reconstruction image quality, whereas image contrast was less relevant. Trainings from heterogeneous training data generalized well toward the test data with a range of acquisition parameters. Trainings from synthetic, non-MR image data showed residual aliasing artifacts, which could be removed by transfer learning–inspired fine-tuning.

Conclusion: This study presents insights into the generalization ability of learned image reconstruction with respect to deviations in the acquisition settings between training and testing. It also provides an outlook for the potential of transfer learning to fine-tune trainings to a particular target application using only a small number of training cases.

KEYWORDS
accelerated imaging, deep learning, iterative image reconstruction, machine learning, transfer learning, variational network

1 | INTRODUCTION

The use of deep learning[1] for medical image reconstruction is a new and emerging field. The first early-stage developments were reported in 2016. Wang et al proposed to augment a conventional compressed-sensing reconstruction with a regularizer that is based on a convolutional neural network.[2] Kwon et al proposed to learn a parallel imaging reconstruction without explicit use of coil sensitivity maps that operates entirely within image space.[3,4] We proposed a learning approach based on the framework of variational optimization with the goal of learning the complete reconstruction procedure, which maps from multichannel k-space raw data to image space, and the associated numerical procedure.[5,6] Since then, a substantial increase of developments has occurred around the world. At the 2017 annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM), work was shown that used learning for image reconstruction for angiography,[7] multicontrast MRI,[8] cardiac imaging,[9] MR fingerprinting,[10] manifold learning,[11] partial Fourier imaging,[12] projection reconstruction,[13] and compressed sensing using residual learning.[14] Our own recent developments, which were also presented at ISMRM 2017, included a preliminary investigation of the influence of sampling patterns on the training procedure,[15] the influence of different loss functions that are used in the training,[16] and a first clinical reader study with the goal of evaluating the diagnostic content of accelerated images that were reconstructed using a variational network.[17]

One of the biggest open questions regarding the success of these technologies in practice is generalization. To what degree can the test data deviate from the data that were used during training? This is important for several reasons. First, one of the key strengths of MRI is the flexibility during data acquisition. Due to the range of available MR systems and protocols, images from different institutions commonly vary with respect to acquisition parameters. A learned reconstruction procedure that works only for a specific set of imaging parameters would therefore be only of limited practical use, as it would require retraining for every new setup. Second, collecting large data sets for training is usually expensive in medical imaging. In some cases, it is even impossible, such as in the case of time-resolved imaging, in which a high spatial and temporal resolution ground truth cannot be obtained. The necessity to collect separate training data for all protocol versions of a particular sequence would put substantial restrictions on clinical translation of these new technologies.

The main goal of this study is to assess the influence of image contrast, SNR, sampling pattern, and image content on the generalization of a learned image reconstruction. These design parameters were chosen for investigation because the goal of learning a reconstruction for accelerated data is the separation of aliasing artifacts and true image content. These parameters have a strong influence on the structure of the aliasing artifacts, and consequently, the conditioning of the reconstruction problem.

The additional goal, which is also related to the question of generalization and the issue of limited training data, is to investigate the potential for transfer learning[18] for image reconstruction using our proposed variational network architecture.[6] This particular topic was recently investigated for MR image reconstruction of brain data (Dar and Cukur, *arXiv*, 2017) with a deep CNN architecture recently proposed in Ref 9. In the context of image processing and computer vision, the general hypothesis behind transfer learning is that low-level image features, such as edges and simple geometrical structures, are independent of the actual image content of the target application. As a consequence, they can be learned from arbitrary data sets in which large amounts of training data are available. These pretrained models then serve as a baseline, which is then fine-tuned to the target domain using less training data than would be required when training from scratch. This concept is appealing for MR image reconstruction because nonmedical image data are easily available,[19,20] which can be used to simulate synthetic k-space data. In contrast, large amounts of true measurement training data are often challenging to obtain.

2 | METHODS

We used a combination of true measurement k-space data from clinical patients, additionally processed k-space data, and completely synthetic data for the experiments in this study. For in vivo data acquisition, 40 consecutive patients referred for diagnostic knee MRI to evaluate for internal derangement were enrolled in the study, which was approved by the internal review board. Fully sampled raw data were acquired on a clinical 3T system (Magnetom Skyra, Siemens, Erlangen, Germany) with a standard 15-channel knee coil. We acquired data with the conventional 2D turbo spin-echo protocol that is used clinically at our institution. Coronal proton density–weighted (PD_w) sequences with and without fat suppression (FS) were acquired. Technologists were instructed to keep the following sequence parameters constant during the study:

- PD_w: TR = 2750 ms, TE = 27ms, echo-train length = 4, matrix size = 320 × 288, in-plane resolution = 0.49 × 0.44 mm^2, slice thickness = 3 mm; and
- PD_w FS: TR = 2870 ms, TE = 33 ms, echo-train length = 4, matrix size = 320 × 288, in-plane resolution = 0.49 × 0.44 mm^2, slice thickness = 3 mm.

The number of acquired slices varied depending on the size of the patient. Twenty cases were acquired for both the PD_w (5 female, 15 male; age 15-76; body mass index 20-33) and the PD_w FS (10 female, 10 male; age 30-80; body mass index 20-34) sequence. The data were split equally into 2 categories. The first 10 acquisitions were used for training and the remaining half was used for validation. A selection of slices reconstructed by an inverse Fourier transform followed by a sum-of-squares combination of the individual coil elements is shown in Supporting Information Figure S1. These data show strong similarities in terms of the actual image content but have fundamentally different contrast and SNR. The noise level σ_{est} of the 2 sequences was estimated from an off-center slice that showed only background u_σ. The estimation was performed by averaging the SD from the real and imaginary channels of the uncombined multichannel data, and then averaging over all N_k training data cases n: $\sigma_{est}(u_\sigma)=\frac{1}{N_k}\sum_{n=1}^{N_k}(std([Re(u_\sigma)])+std([Im(u_\sigma)]))$. This resulted in an estimated noise level of $\sigma_{est}=10^{-5}$, which was identical for both the fat-suppressed and the non-fat-suppressed sequence. The signal level μ_{est} was then estimated by calculating the l2 norm of the complex multichannel k-space data f, averaged over the central $N_{sl}=20$ slices of all training data cases: $\mu_{est}(f)=\frac{1}{N_k}\sum_{n=1}^{N_k}(\frac{1}{N_{sl}}\sum_{sl=1}^{N_{sl}}\frac{\|f\|_2}{length(f)})$. The central 20 slices were selected to ensure that no slices that contained no signal because of being outside the imaged anatomy were included in this analysis. In this definition, f is organized as a single stacked column vector of the data from all receive coils. The estimated $SNR=\frac{\mu_{est}}{\sigma_{est}}$ of the PD_w data was approximately 80, and that of the PD_w FS data was approximately 20. A third data set was then generated by adding additional complex Gaussian noise to the PD_w data such that the SNR corresponded to the PD_w FS data. These data sets allowed us to study the generalization influences of SNR and image contrast independently from each other.

Synthetic k-space data were generated using 200 images from the Berkeley segmentation database (BSDS).[19] Images were cropped according to the matrix size of the knee k-space data, including readout oversampling. The images were modulated with a synthetic sinusoidal phase using different randomly selected frequencies. After point-wise multiplication with randomly selected coil sensitivity maps estimated from our knee training data, the images were Fourier-transformed. Complex Gaussian noise was then added to these synthetic k-space data according to the noise levels of our knee imaging data. We generated 3 different versions of these data. The first was generated at the SNR level of the PD_w data, the second at the SNR level of the PD_w FS data, and the third with a randomly selected level of SNR for every single image using the PD_w FS data as the lower and the PD_w data as the upper bound of SNR.

K-space data were undersampled by a factor of 4, according to a regular Cartesian undersampling pattern as implemented by the scanner vendor for accelerated acquisitions using parallel imaging, and variable density pseudorandom sampling according to Ref 21. The same random sampling pattern was used in all experiments. Twenty-four reference lines at the center of k-space were used for the estimation of coil sensitivity maps in both cases, using ESPIRiT.[22]

We followed the learned image reconstruction procedure using a variational network described in Ref 5. In this approach a regularized iterative image reconstruction defined by

$$E(u)=R(u)+\frac{\lambda}{2}\|Au-f\|_2^2 \tag{1}$$

is learned from the training data. The input of the variational network is the undersampled k-space raw data f and the corresponding coil sensitivity maps (included in the forward-sampling operator A in Equation 1), and the output is a complex-valued coil-combined image u. All computations are performed by accounting for the complex valued data with the exception of the application of the regularizer on the image u, in which the image is split up into the real and imaginary part. The regularizer is defined as

$$R(u)=\sum_{i=1}^{N_k}\rho_i(k_i*u),\qquad k_i*u=k_{i,Re}*u_{Re}+k_{i,Im}*u_{Im}. \tag{2}$$

It consists of a set of N_k spatial filter kernels k for the real and imaginary component of an MR image and potential functions ρ, which are learned from the data together with the regularization parameter λ. Inserting this regularizer in an iterative image reconstruction yields the following update:

$$u_{t+1}=u_t-\sum_{i=1}^{N_k}\bar{k}_i^t*\rho_{i,t}'(k_i^t*u_t)-\lambda_t A^*(Au_t-f),\quad \lambda_t>0,\quad 0\le t\le T-1$$

where $\bar{k}_i^t$ denote the filter kernels k_i^t rotated by 180°; and $\rho_{i,t}'$ are the first derivatives of the potential functions $\rho_{i,t}$. Unfolding several iterations of this scheme leads to the variational network structure depicted in Figure 1.[6] Essentially, one gradient step GD of an iterative reconstruction can be related to one stage t in a network with a total of T stages. For reference, the used network architecture is shown in Figure 1. Ten stages were used, each consisting of 24 convolution kernels of size 11×11. The iPalm optimizer[23] and the variational network architecture were implemented and trained using Tensorflow,[24] which was extended with additional operators such as the trainable activation functions and (inverse) Fourier shift operations. The source code of our extended TensorFlow library will be provided online (https://github.com/VLOGroup/tensorflow-icg). Example training and testing code for MRI reconstruction will be provided online as well (https://github.com/VLOGroup/mri-variationalnetwork). Trainings were performed slice by slice. During both training and testing each slice u was normalized

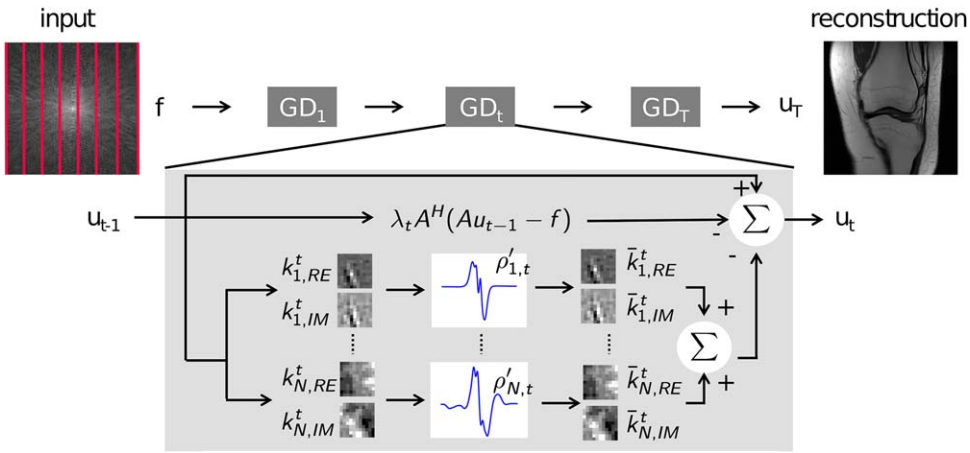

FIGURE 1 Overview of the variational network architecture used in this work

between 0 and 1 $\left(u=\frac{u_{orig}}{max(|u_{orig}|)}\right)$ for the application of the learned regularizer. The pixel-by-pixel mean-squared-error to the fully sampled reference was used as the error metric that was minimized during the training process. All trainings were performed with 150 epochs and a batch size of 5, using the iPalm algorithm.[24] One epoch was defined as a sequence of updates of the network parameters when all training examples have been used exactly once. Trainings were performed using the following training data.

First set of experiments: assessment of generalization with respect to contrast and SNR:

1. PD_w using the central 20 slices from 10 patients (total of $N = 200$ slices).
2. PD_w FS using the central 20 slices from 10 patients (total of $N = 200$ slices).
3. PD_w using the central 20 slices from 10 patients (total of $N = 200$ slices) with additional noise added to the complex multichannel k-space data such that the SNR corresponds to the SNR level of the FS sequence.
4. Joint PD_w (5 patients) and PD_w FS (5 patients) training, each using the central 20 slices (total of $N = 200$ slices).
5. Joint PD_w (5 patients) and PD_w with additional noise added (5 patients) training, each using the central 20 slices (total of $N = 200$ slices).

Second set of experiments: influence of the number of training samples and the heterogeneity of the training data:

6. Joint PD_w (10 patients) and PD_w FS (10 patients) training, each using the central 20 slices (total of $N = 400$ slices).
7. PD_w using the central 20 slices from the 5 patients used in the joint training in experiment 4 in the first set of experiments (total of $N = 100$ slices).
8. PD_w FS using the central 20 slices from the 5 patients used in the joint training in experiment 4 in the first set of experiments (total of $N = 100$ slices).

Third set of experiments: assessment of generalization with respect to the sampling pattern:

9. Training with regular sampling, testing with regular sampling.
10. Training with random sampling, testing with regular sampling.
11. Training with regular sampling, testing with random sampling.
12. Training with random sampling, testing with regular sampling.
13. Joint training with regular and random sampling, testing with regular sampling.
14. Joint training with regular and random sampling, testing with random sampling.

Fourth set of experiments: training with synthetic data:

15. Synthetic BSDS data (total of $N = 200$ images) with (high) SNR level of PD_w, regular sampling.
16. Synthetic BSDS data (total of $N = 200$ images) with (low) SNR level of PD_w FS, regular sampling.
17. Synthetic BSDS data (total of $N = 200$ images) with variable SNR, regular sampling.
18. Synthetic BSDS data (total of $N = 200$ images) with variable SNR, random sampling.

Transfer learning experiments: fine-tuning the regular sampling variable SNR synthetic BSDS model for another 150 epochs:

19. Fine-tuning using a subset of $N = 20$ slices selected from all PD_w cases.
20. Fine-tuning using a subset of $N = 20$ slices selected from all PD_w FS cases.

In addition, trainings using only the reduced subsets that were used for fine-tuning were performed as a reference for the transfer learning experiments.

The remaining 10 knee-measurement data sets of both sequences and the data set with additional noise added were used to test the performance of the different learned image reconstruction networks. Quantitative evaluation was performed by calculating the root-mean-squared-error $RMSE(u, u_{ref}) = \frac{\|u_{ref} - u\|_2}{\|u_{ref}\|_2}$ and the structural similarity index $SSIM(u, u_{ref}) = \frac{(2\ mean(u_{ref})\ mean(u) + C_1)(2\ cov(u,u_{ref})\ + C_2)}{(mean(u_{ref})^2 + mean(u)^2 + C_1)(std(u_{ref})^2 + std(u)^2 + C_2)}$ to the fully sampled reference u_{ref} for all test slices (378 slices in the case of PD_w and 365 slices in the case of PD_w FS). In the definition of the structural similarity metric (SSIM), $C_1 = (0.01L)^2$ and $C_2 = (0.03L)^2$, with L being the dynamic range of the input images, are regularization parameters that are used to avoid instabilities in regions of the image where the local mean or SD is close to zero.[25] Finally, to obtain insight into the fine-tuning process in transfer learning, the learned network parameters before and after fine-tuning were visualized together with those from the corresponding in vivo training.

3 | RESULTS

Results of the assessment of generalization with respect to contrast and SNR are shown in Figure 2. A zoomed view to a region of interest that includes complex image texture due to bone trabeculae and fine details due to ligaments and cartilage is shown in Supporting Information Figure S2. Unsurprisingly, the best results can be achieved when applying the network to test data from the same sequence on which it was trained. When applying the network trained from high SNR PD_w data to the lower SNR FS data, a substantial level of noise is present in the reconstructed images. This leads to a reduction of SSIM from 0.89 to 0.81 for this particular slice. In contrast, applying the network trained from low SNR PD_w FS data to higher SNR PD_w data leads to slightly blurred images with some residual artifacts, leading to a SSIM reduction from 0.94 to 0.91. The behavior of the network trained from PD_w data with additional noise is comparable to the network trained from the lower SNR PD_w FS data. In particular, the SSIM is 0.88 for the PD_w FS test data in comparison to 0.89 when training with the data from the same sequence. In the case of the PD_w data test, the SSIM is identical (0.91). The PD_w test data with additional noise results in substantially lower quality in the case of the PD_w training (SSIM of 0.74) in comparison to all other trainings, including the individual training from the PD_w FS data (SSIM of 0.82). The results from the joint training using data from both contrasts are identical to using the same sequence for training and testing (SSIM of 0.89 for PD_w FS and 0.94 for PD_w). The joint training from PD_w data with and without additional noise leads to identical results for the PD_w and noisy PD_w test data. For the PD_w FS test data, SSIM is reduced slightly (0.85 in comparison to 0.88 for training with noisy PD_w). The visual impression and the SSIM values of the individual slices shown for the different experiments are confirmed by the quantitative SSIM and RMSE analysis over all cases in the test set (Supporting Information Table S1).

The influence of the number of training data points in relation to their heterogeneity is shown in Figure 3. No substantial differences in image quality can be observed between the results of these trainings and the individual trainings in Figure 2. A small improvement of 0.01 in the SSIM can be observed in the quantitative analysis (Supporting Information Table S2) for the largest training data set that includes all training samples from both sequences.

Figures 4 and 5 show the results of the assessment of generalization with respect to the sampling pattern. When the sampling pattern is consistent between training and testing, results without aliasing artifacts and preservation of fine details are obtained. Applying a network that was trained from random undersampled data to regular undersampling leads to subtle residual artifacts. With the exception of application of a network that was trained from regular undersampling to random undersampling, which shows a small SSIM drop from 0.96 to 0.95 for the PD_w test data, quantitative image metrics were identical for all other combinations of training and test data. A joint training with data from both sampling patterns leads to results that are comparable to individual trainings with no deviations in acquisition parameters. This behavior is identical to the experiments with image contrast and SNR. Quantitative SSIM and RMSE analysis over all cases in the test set are provided in Supporting Information Table S3 for this experiment.

To demonstrate that the trainings are properly converged, plots of RMSE and SSIM of the training and test sets over the training epochs are shown in Figure 6 for the experiments shown in Figures 3–5, and 6.

The results of the trainings that were performed on synthetic k-space data generated from the BSDS database are shown in Figure 7. The most substantial difference to the experiments when training with true in vivo MR data from the same anatomical area is the presence of residual aliasing artifacts in the results (indicated by red arrowheads in Figure 3). The effects of the influence of SNR can be reproduced with the synthetic data. Training data with a substantially higher SNR level leads to noise amplification. This effect is strongest in the case of no additional noise, where the SSIM drops to 0.65 for this particular slice of PD_w FS test data. Training with substantially lower SNR leads to blurring and residual artifacts. This effect is strongest when using the network trained at the noise level of the PD_w FS data for non-

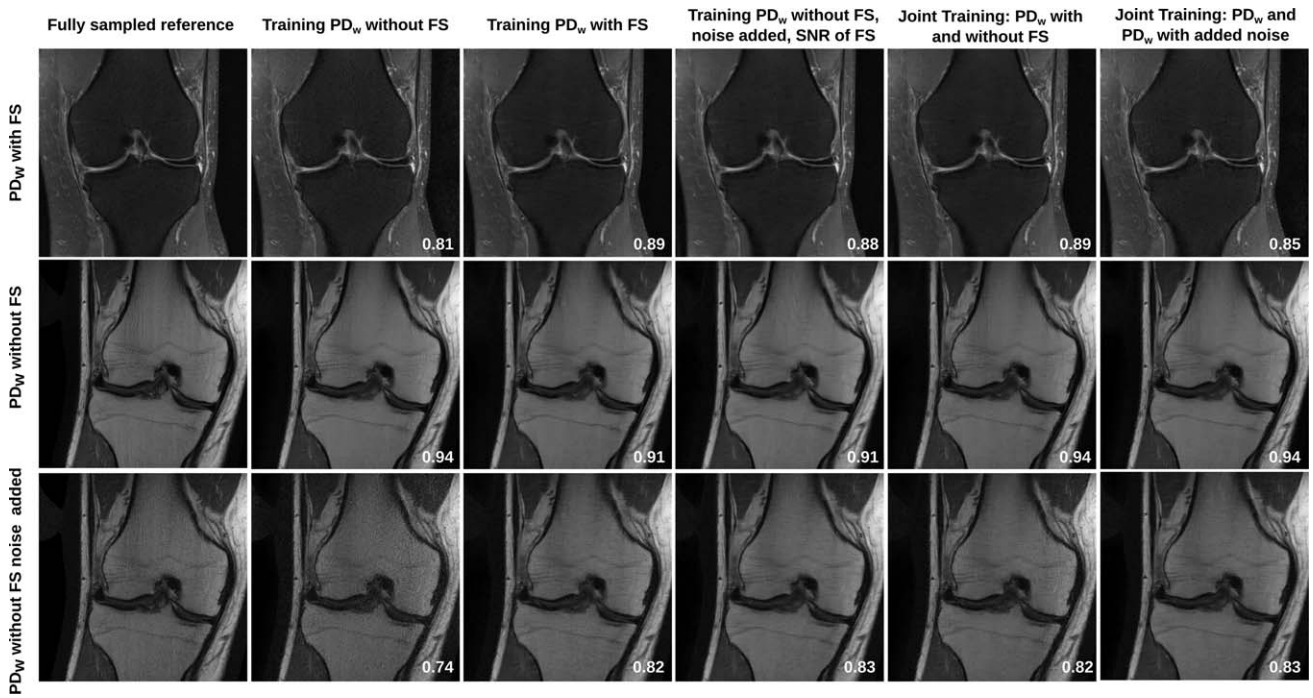

FIGURE 2 Assessment of generalization with respect to contrast and SNR. The structural similarity metric (SSIM) to the fully sampled reference is shown for the corresponding slices. When applying a network from high SNR proton density–weighted (PD_w) data to lower SNR PD_w fat-suppression (FS) data, a substantial level of noise is present in the reconstructions. Applying the low-SNR PD_w FS network to higher SNR PD_w data leads to slightly blurred images with residual aliasing artifacts. The behavior of the network trained from PD_w data with additional noise was comparable to the network trained from the lower SNR PD_w FS data. The PD_w test data with additional noise results in substantially lower quality in the case of the PD_w training in comparison to all other trainings. The results from the joint training using data from both contrasts are identical to using the same sequence for training and testing. The joint training from PD_w data with and without additional noise leads to identical results for the PD_w and noisy PD_w test data and slightly reduced SSIM for the PD_w FS test data. A zoomed view of these results is included in the Supporting Information

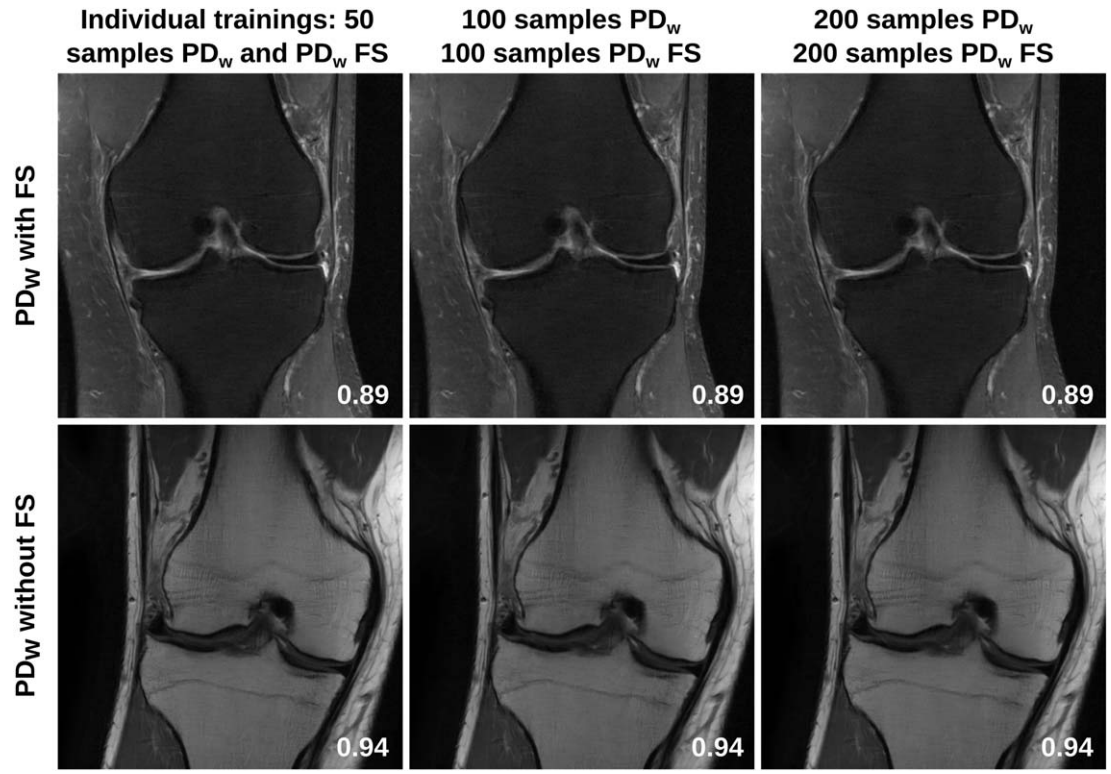

FIGURE 3 Influence of the number of training data points in relation to their heterogeneity. No substantial differences in image quality can be observed between the results of these trainings and the individual trainings in Figure 2

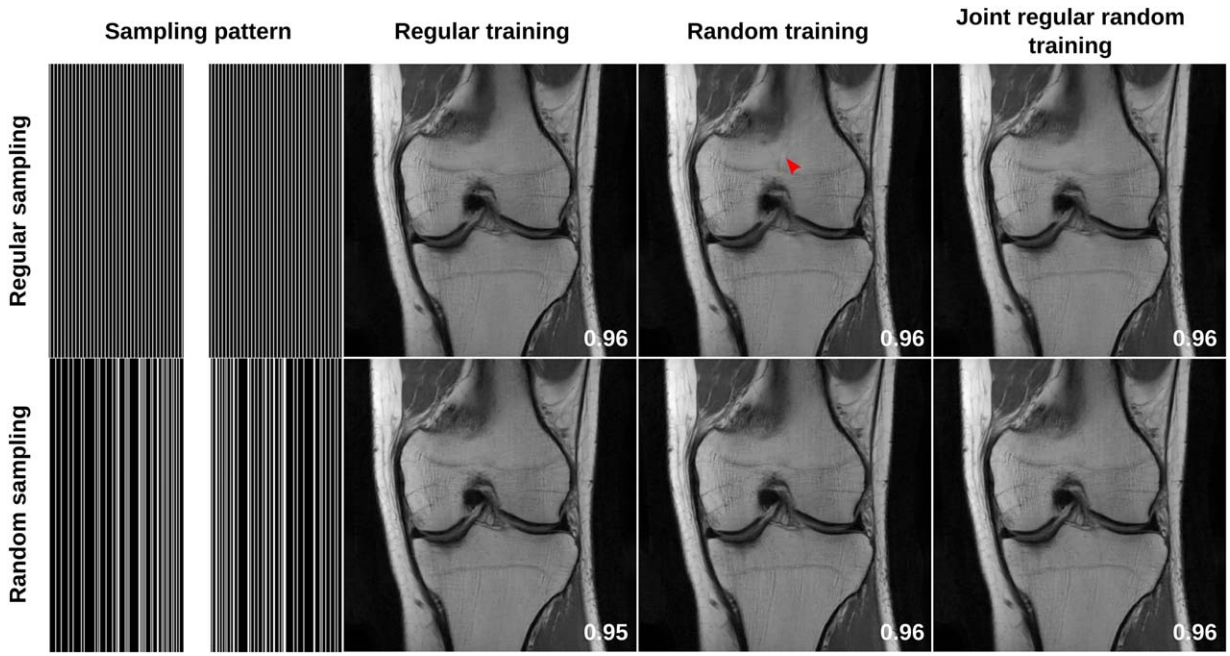

FIGURE 4 Assessment of generalization with respect to the sampling pattern for the higher SNR non-FS data. When the sampling pattern is consistent between training and testing, results without aliasing artifacts and preservation of fine details are obtained. Applying a network that was trained from regular undersampling to random undersampling leads to subtle oversmoothing, which is also reflected in a slight drop of the SSIM from 0.96 to 0.95. Applying a network that was trained from random undersampled data to regular undersampling leads to residual artifacts. Identical to the experiments with image contrast and SNR, a joint training with data from both sampling patterns leads to results that are comparable to individual trainings with no deviations in acquisition parameters

FS test cases. Again, training with a range of SNR values leads to results that are comparable to training with data that are consistent with the test data in terms of SNR. The corresponding quantitative analysis is provided in Supporting Information Table S4.

The results from the transfer learning experiments are shown in Figure 8. For data of both sequences, results from fine-tuned networks outperform both baseline trainings using only synthetic data, as well as reference trainings using the same subset of knee MRI data that were used during fine-

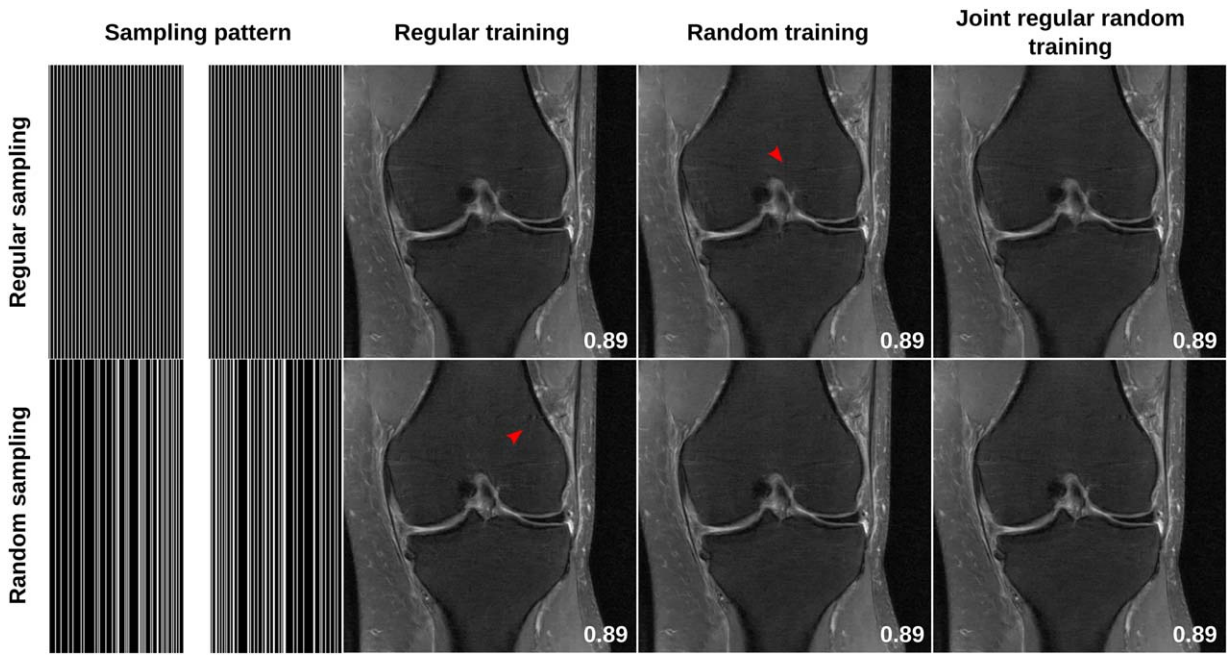

FIGURE 5 Assessment of generalization with respect to the sampling pattern for the lower SNR FS data shows the same behavior as the experiments with non-FS data (Figure 4). However, residual aliasing artifacts are subtler because the image corruption is primarily dominated by noise amplification in this lower SNR case and aliasing artifacts are buried under the noise level

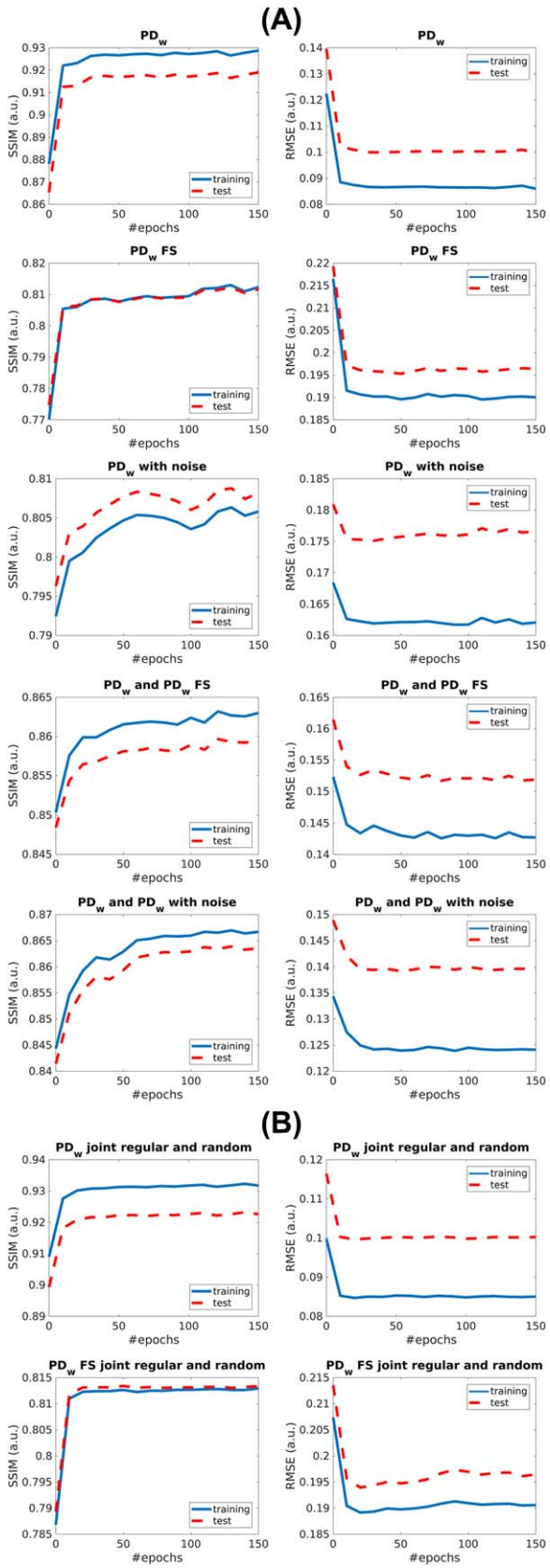

FIGURE 6 A, Plots of RMSE and SSIM of the training and test sets over the training epochs for SNR and contrast generalization experiments in Figure 3. B, Experiments with changes in the sampling pattern from Figures 5 and 6

tuning in terms of removal of residual artifacts. Quantitative SSIM and RMSE analysis are given in Supporting Information Table S5. In particular, with the exception of SSIM for training and testing with non-FS PD_W data (0.96 versus 0.95 for the corresponding transfer learning experiment), results for transfer learning lead to the same SSIM values as the corresponding trainings with in vivo MRI data from the same sequence. Figure 9 shows plots of RMSE and SSIM for the transfer learning experiments. The first 150 epochs are baseline training, and the training error is shown for synthetic data. Epochs 151 to 300 are fine-tuning and the training error is shown for the corresponding subset of in vivo data. The test error is shown for the in vivo test cases that were used for all experiments in this study. A substantial jump in the training error can be observed at the transition between baseline training and fine-tuning. This is to be expected, because the data set is used to obtain the error metric changes at this point. The effect is subtler for the test error, but the baseline training reached a plateau and then improved further during the fine-tuning period.

4 | DISCUSSION

The results from this study demonstrate that a deviation of SNR between training and test data leads to a substantial reduction of image quality when using a trained variational network for image reconstruction. This can be related to the influence of 2 design parameters in image reconstruction, which are usually tuned by hand in a conventional image reconstruction approach: the number of iterations in an iterative reconstruction[26] and the regularization parameter in compressed sensing.[21] These parameters determine the trade-off between resolution, g-factor-based noise amplification, and residual aliasing artifacts. In a machine learning approach, the parameter that balances the data consistency term and the regularization term, as well as the step size of the numerical algorithm, is learned from the training data. Interestingly, reconstructed test case images showed the same behavior when an SNR deviation occurred due to a change of the pulse sequence that was used between training and testing, and when data from the same pulse sequence was retrospectively corrupted by additional noise. In particular, lower SNR PD_w FS test data showed comparable image quality for trainings from PD_w FS data and PD_w data with additionally added noise. The PD_w test data with additionally added noise showed substantially higher image quality for trainings from PD_w FS data, with different contrast but matched SNR, than for trainings from the PD_w data, with matched contrast but different SNR. This demonstrates that although SNR is a critical parameter that has to be consistent between training and testing, image contrast is a less critical factor. This particular behavior can only be interpreted for

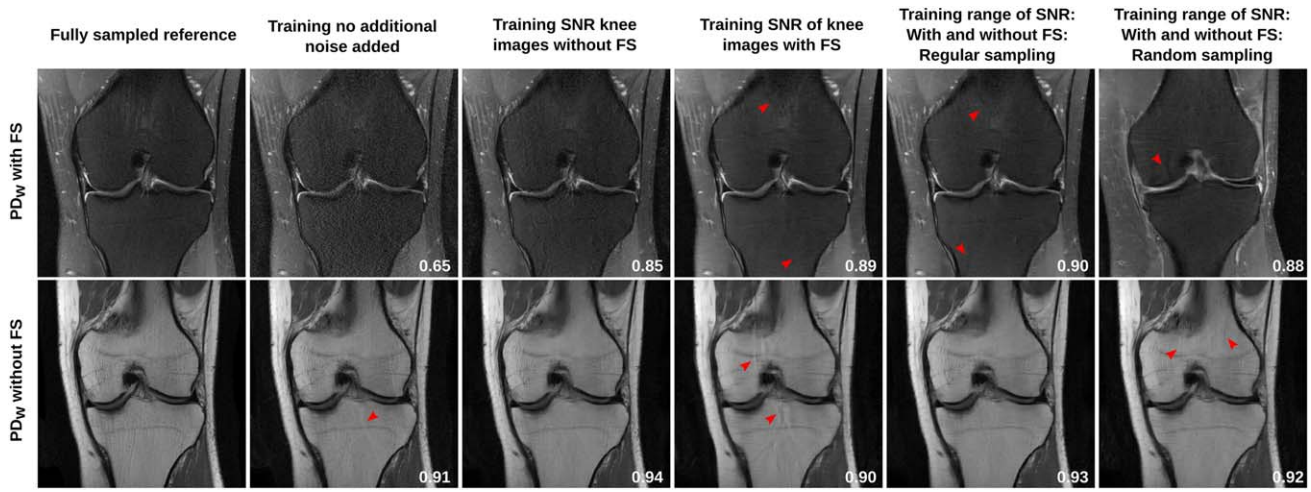

FIGURE 7 Trainings from synthetic, regularly undersampled, k-space data generated from the Berkeley segmentation database (BSDS) database. The SSIM to the fully sampled reference is shown for the corresponding slices. Reconstructions show a larger degree of residual aliasing artifacts (red arrowheads) in comparison to trainings with in vivo knee data. The effects of the influence of SNR can be reproduced with the synthetic data. Experiments with deviating SNR levels between training and test data again lead to either noise amplification or blurring and residual artifacts. Again, training with a range of SNR values leads to results that are comparable to training with data that are consistent with the test data in terms of SNR. This particular training was also performed and tested with random sampling, with comparable behavior

our particular network architecture, implementation of the training procedure and difficulty of the reconstruction problem defined by the acceleration factor, SNR, and the quality of the particular multichannel receive coil.

As expected, because the structure of the aliasing artifacts is influenced by the used sampling pattern, deviations between training and testing influence the quality of the reconstructions. However, it is important to note that our approach does not learn a particular aliasing pattern by heart. Elimination of artifacts is performed locally, by spatial convolution of learned filter kernels in image space. This explains why the negative effects on the reconstruction quality are relatively benign. However, it cannot be expected that these results generalize to more substantial changes in the

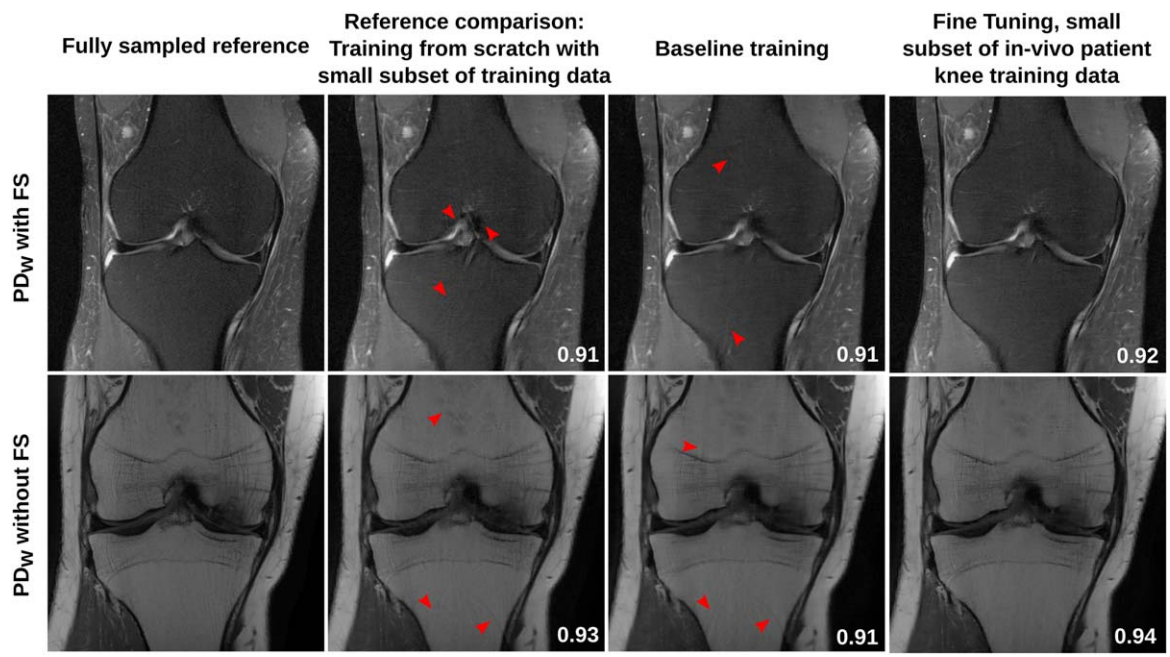

FIGURE 8 Transfer learning experiments. For both PD_w and PD_w FS data, results from fine-tuned networks outperform baseline trainings using only synthetic data, as well as reference trainings using the same subset of knee MRI data that were used during fine-tuning in terms of removal of residual artifacts. This indicates the possibility of fine-tuning networks that were pretrained from generic data with only a very small number of training cases for a particular target application

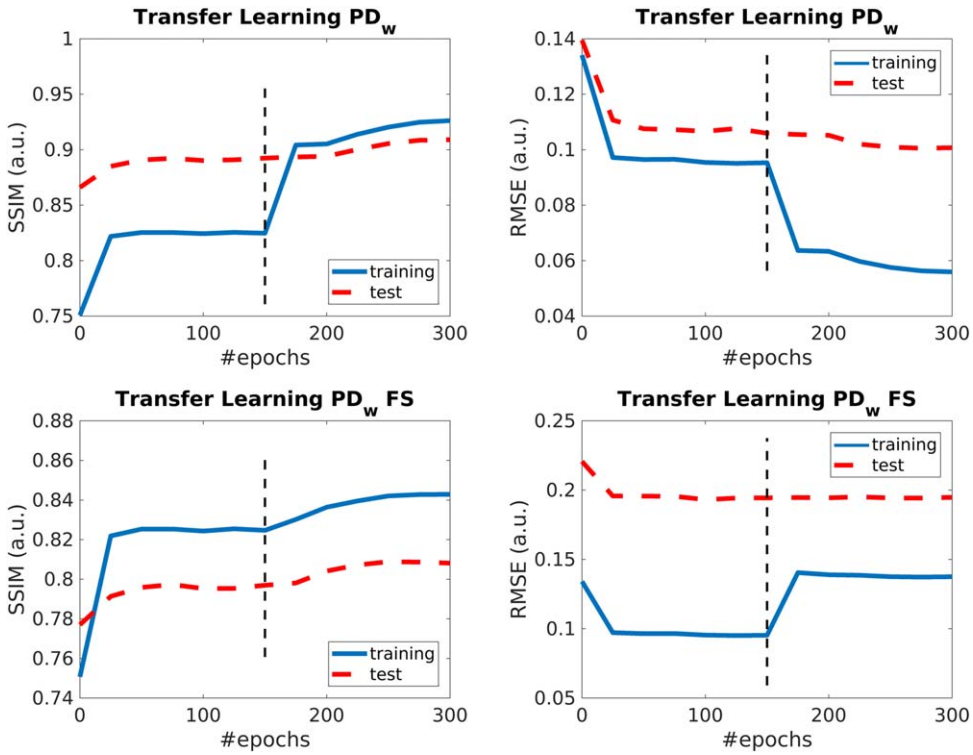

FIGURE 9 Plots of RMSE and SSIM of the training and test sets over the training epochs for the transfer learning experiments. The first 150 epochs are baseline training and the training error is shown for synthetic data. Epochs 151 to 300 are fine-tuned and the training error is shown for the corresponding subset of in vivo data. The test error is shown for the in vivo test cases that were used for all experiments in this study. The dashed line illustrates the epoch in which the training changes from baseline training to fine-tuning

trajectory (e.g., 3D pseudorandom sampling or non-Cartesian trajectories like radial or spirals, which have fundamentally different aliasing properties).

Training a reconstruction from heterogeneous data leads to the same results as training from data that were consistent between training and testing. This behavior was consistent for contrast, SNR, and multiple sampling patterns. These experiments demonstrate that a reconstruction can be learned that generalizes with respect to changes in acquisition parameters, under the condition that the corresponding heterogeneity is included in the training data. It is currently an open question to what degree an increase of heterogeneity in the training data also requires an increase of the total samples to achieve generalization. The experiments in this study did not show substantial deviations in performance when varying the number of training samples. However, the goal of these experiments was not a large-scale analysis of the influence of the number of training data sets in learned image reconstruction. The goal was to study the cases in which the exact same data sets are used in single-contrast and heterogeneous multicontrast trainings ($N = 100$ to $N = 400$ slices). An analysis of the influence of the number of training samples in the context of transfer learning for the network architecture proposed in Ref 9 is presented in an *arXiv* preprint by Dar and Cukur ("A Transfer-Learning Approach for Accelerated MRI using Deep Neural Networks," *arXiv*, 2017). The authors report improvements of SSIM from 0.93 to 0.96 when increasing the number of baseline training images by a factor of 8 and the number of fine-tuning images by a factor of 4. Although the acquisition of large data sets with sufficient heterogeneity can be challenging in practice, the results from this study indicate that data augmentation can potentially be used successfully. Given the availability of fully sampled training k-space data, experiments with different sampling patterns and acceleration factors can be performed without the need to acquire additional measurements. The influence of SNR versus image contrast further supports this strategy. Different levels of SNR can easily be achieved with data augmentation, whereas a change of image contrast would require either additional acquisitions or the use of synthetic data and numerical simulations of the MR signal of different pulse sequence and sequence parameters.

The influence of SNR that was observed in the generalization experiments with knee data from different sequences can be reproduced entirely with experiments using synthetic data. This again shows that SNR plays a more critical role than image content in the context of learned image reconstruction. However, results from trainings with synthetic data showed a substantially higher level of residual aliasing artifacts, illustrating that both the sampling pattern and the actual image content define the structure of the introduced aliasing. Our experiments were designed around the 2

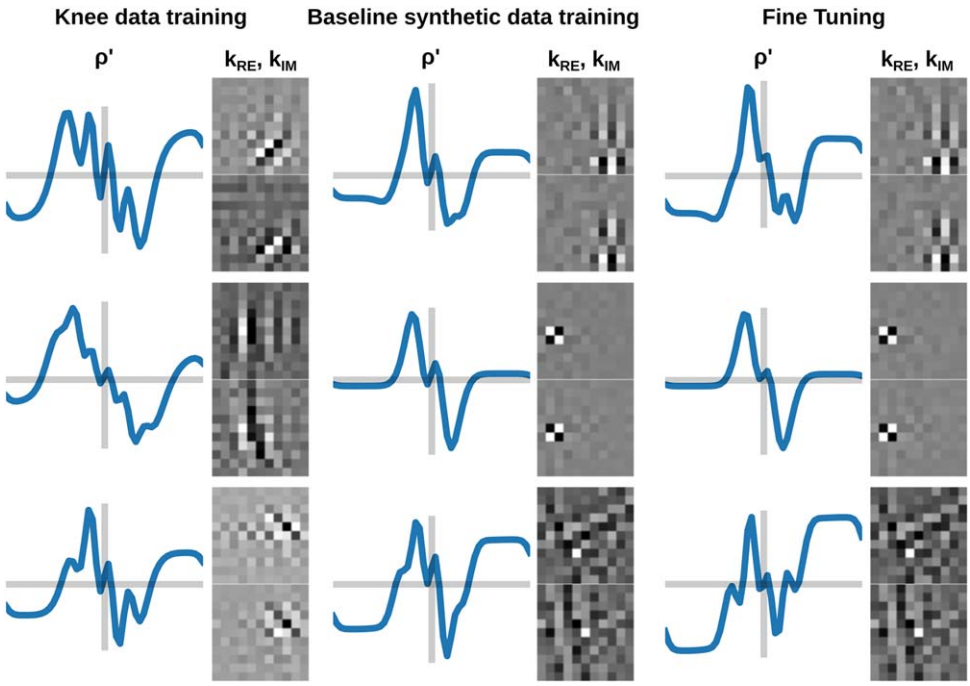

FIGURE 10 Visualization of a selection of learned nonlinear activation functions ρ' and filter kernels k_{RE} and k_{IM} for the real and imaginary plane of the regularizer. The fine-tuned kernels closely resemble the kernels from the baseline training, indicating that they are not updated substantially during the fine-tuning process. Larger updates can be observed for the learned influence functions

extreme ends of the spectrum of training and test data consistency: training from the same anatomy and scan orientation, and training from arbitrary nonmedical images. Future work should be conducted to investigate training from the same anatomical structure but different scan orientations and scans from different anatomical areas. In addition, the particular choice of using natural camera images to generate synthetic k-space data, and the particular image database, was only one of several possible experiment design choices in this study. Another topic of future research is the use of dedicated numerical simulation data that are designed specifically with the idea of training an image reconstruction procedure. This is particularly interesting for dynamic imaging applications, in which the acquisition of high spatial and temporal resolution training data is even more challenging.

Transfer learning–inspired fine-tuning with a substantially smaller size of target domain knee MR images reduced the effect of residual aliasing artifacts. The results were close to the optimal case of using MR imaging data from the same anatomical area and pulse sequence for training and testing. These results are not only in line with studies in computer vision, in which transfer learning was used to classify class labels that were not present in the original training data set,[27] but they are also comparable to transfer learning for a neural network architecture for MR reconstruction proposed in Ref 9, evaluated on brain MR data sets (Dar and Cukur, *arXiv*, 2017). Visualizing the parameters of the learned network provides some additional insight into the fine-tuning procedure. Figure 10 shows a visualization of the learned filter kernels k_{RE} and k_{IM} for the real and imaginary plane of the regularizer together with the learned nonlinear activation functions ρ′. The fine-tuned kernels closely resemble the kernels from the baseline training, indicating that they are not updated substantially during the fine-tuning process. Larger updates can be observed for the learned influence functions. They still bear closer resemblance to the baseline trainings than the corresponding in vivo trainings. However, a direct comparison between the trainings is challenging because they do not necessarily perform the same operations on the images at the same stage in the network. The reason for this is that the training process is a highly nonconvex optimization problem. The parameters of the whole network are trained simultaneously, and the error metric is the final output after the last stage. This indicates that the result of each training is one of multiple local minima that leads to approximately the same result, a property that is known from deep learning.[1]

The relation of the number of training data samples that was used for fine-tuning in comparison to the full trainings (one tenth) was chosen empirically for this study. A more systematic comparison of the influence of the relation between the number of samples for baseline training and for fine-tuning is presented by Dar and Cukur (Dar and Cukur, *arXiv*, 2017). Moreover, although the particular network architecture that was used in this study can be trained successfully from data sets in the order of several hundred samples, it is a topic for further research if additional performance benefits can be achieved by training different

architectures with a larger number of free parameters using synthetic data followed by a fine-tuning step using real MR data.

5 | CONCLUSIONS

This study presents insights into the general properties and the generalization ability of a learned variational network for MR image reconstruction with respect to deviations in the acquisition settings between training and testing for a clinically representative set of test cases. Our results show that mismatches in SNR have the most severe influence. Our experiments also demonstrate that by increasing the heterogeneity of the training data set, trained networks can be obtained that generalize toward wide-range acquisition settings, including contrast, SNR, and the particular k-space sampling pattern. Finally, our study provides an outlook for the potential of transfer learning to fine-tune trainings of our variational network to a particular target application using only a small number of training cases.

ACKNOWLEDGMENTS

We acknowledge the grant support from the National Institutes of Health (NIH P41 EB017183); the Austrian Science Fund under the START project BIVISION Y729; the European Research Council under the Horizon 2020 program; the ERC starting grant "HOMOVIS" (640156); and the hardware support from Nvidia Corporation. We also thank Ms. Mary Bruno for support during data acquisition.

REFERENCES

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521:436-444.

[2] Wang S, Su A, Ying L, et al. Accelerating magnetic resonance imaging via deep learning. In Proceedings from the 13th *IEEE International Symposium on Biomedical Imaging (ISBI)*, Prague, Czech Republic, 2016. pp. 514-517.

[3] 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.

[4] Kwon K, Kim D, Park H. A parallel MR imaging method using multilayer perceptron. *Med Phys.* 2017;44:6209-6224.

[5] 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.

[6] Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn Reson Med*. 2018;79:3055-3071.

[7] Jun Y, Eo T, Kim T, Jang J, Hwang D. Deep convolutional neural network for acceleration of magnetic resonance angiography (MRA). In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 686.

[8] Gong E, Zaharchuk G, Pauly J. Improving the PI+CS reconstruction for highly undersampled multi-contrast MRI using local deep network. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 5663.

[9] Schlemper J, Caballero J, Hajnal J, Price A, Rueckert D. A deep cascade of convolutional neural networks for MR image reconstruction. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 643.

[10] Cohen O, Zhu B, Rosen M. Deep learning for fast MR fingerprinting reconstruction. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 688.

[11] Zhu B, Liu J, Rosen B, Rosen M. Neural network MR image reconstruction with AUTOMAP: automated transform by manifold approximation. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 640.

[12] Wang S, Huang N, Zhao T, Yang Y, Ying L, Liang D. 1D partial Fourier parallel MR imaging with deep convolutional neural network. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 642.

[13] Han YS, Lee D, Yoo J, Ye JC. Accelerated projection reconstruction MR imaging using deep residual learning. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 690.

[14] Lee D, Yoo J, Ye JC. Compressed sensing and parallel MRI using deep residual learning. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 641.

[15] Hammernik K, Knoll F, Sodickson D, Pock T. On the influence of sampling pattern design on deep learning-based MRI reconstruction. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 644.

[16] Hammernik K, Knoll F, Sodickson D, Pock T. L2 or not L2: impact of loss function design for deep learning MRI reconstruction. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 687.

[17] Knoll F, Hammernik A, Garwood K, et al. Accelerated knee imaging using a deep learning based reconstruction. In Proceedings of the 25th Annual Meeting of the ISMRM, Honolulu, HI, 2017. p. 645.

[18] Pratt LY. Discriminability-based transfer between neural networks. In: *Advances in Neural Information Processing Systems 5*. Burlington, MA: Morgan Kaufmann Publishers; 1993. pp. 204-211.

[19] Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings of the 8th IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 2001. pp. 416–423.

[20] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 2009. pp. 248–255.

[21] Lustig M, Donoho D, Pauly JM. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn Reson Med.* 2007;58:1182-1195.

[22] Uecker M, Lai P, Murphy MJ, et al. ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. *Magn Reson Med.* 2014;71:990-1001.

[23] Pock T, Sabach S. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM J Imaging Sci.* 2016;9:1756-1787.

[24] Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Savannah, GA, 2016. pp. 265-284.

[25] 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.

[26] Pruessmann KP, Weiger M, Boernert P, Boesiger P. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn Reson Med*. 2001;46:638-651.

[27] Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 2009. pp. 951-958.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the supporting information tab for this article.

FIGURE S1 Overview of the training data that were used in this study. A, Data from a coronal PD_w 2D turbo spin-echo sequence for knee imaging with and without FS. These two data sets show substantial differences in terms of image contrast and SNR. B, Knee data from the PD_w sequence without FS, with additional noise added to the complex multichannel k-space data, such that the SNR is comparable to the lower SNR FS data. Completely synthetic k-space data were generated from images from the BSDS

FIGURE S2 Zoomed regions of interest from the results shown in Figure 2, further demonstrating the noise amplification and blurring in cases of SNR mismatch between training and test data and the effects of joint training with heterogeneous data. The selected ROI highlights these effects in an anatomical region that includes complex image texture due to bone trabeculae and fine details due to ligaments and cartilage.

TABLE S1 Quantitative analysis of SSIM and RMSE to the fully sampled reference for the generalization experiments with respect to contrast and SNR. The mean and SDs are shown for all cases in the test set, consisting of 378 slices for PD_w and PD_w with additional noise. The PD_w FS test set consisted of 365 slices. Highest image quality values are achieved when data from the same sequence are used for both training and testing. Drops can be observed when the SNR level deviates between training and testing. Results for trainings with PD_w FS data and PD_w data with added noise are comparable. Joint trainings from multiple contrasts and SNR levels show comparable performance to using consistent data between training and testing.

TABLE S2 Quantitative analysis of SSIM and RMSE to the fully sampled reference for the experiments with different numbers of training examples for joint training. The mean and SDs are shown for all cases in the test set, consisting of 378 slices for PD_w and 365 slices for PD_w FS.

TABLE S3 Quantitative analysis of SSIM and RMSE to the fully sampled reference for the generalization experiments with respect to the sampling pattern. The mean and SDs are shown for all cases in the test set, consisting of 378 slices for PD_w and 365 slices for PD_w FS. Trainings show good generalization ability with respect to changes in the sampling pattern. In particular, for non-FS test cases, the SSIM values are identical for all combinations of sampling patterns in training and testing.

TABLE S4 Quantitative analysis of SSIM and RMSE to the fully sampled reference for the trainings with synthetic data. The mean and SDs are shown for all cases in the test set, consisting of 378 slices for PD_w and 365 slices for PD_w FS. A slight drop in image-quality values can be observed in comparison to trainings with in vivo data. The experiments show the same behavior in terms of SNR dependency.

TABLE S5 Quantitative analysis of SSIM and RMSE to the fully sampled reference for the experiments with different numbers of training examples for the transfer learning experiments. The mean and SDs are shown for all cases in the test set, consisting of 378 slices for PD_w and 365 slices for PD_w FS. Transfer learning fine-tuned results outperform both baseline trainings using only synthetic data, as well as reference trainings using the same subset of knee MRI data that were used during fine-tuning. Quantitative image quality values are close to trainings with consistent in vivo data.

How to cite this article: Knoll F, Hammernik K, Kobler E, Pock T, Recht MP, Sodickson DK. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magn. Reson. Med.* 2018;00:1–13. https://doi.org/10.1002/mrm.27355