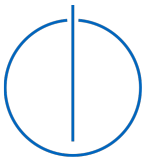
DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Improving the Usage of MRI Reconstruction: What happens if Neural Networks trained on Brain Data are used on Knee Scans?

Shizhe He



DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Understanding Domain Shift in Learned MRI Reconstruction: A Quantitative Analysis on fastMRI Knee and Neuro Sequences

Author:	Shizhe He
Supervisor:	Prof. Dr. Daniel Rueckert
Advisor:	Dr. Kerstin Hammernik
Submission Date:	11.02.2023

Abstract

We investigate the problem of domain shift in the context of state-of-the-art MRI reconstruction networks with respect to variations in training data. We provide visualization tools and support our findings with statistical analysis for the networks evaluated on the 1.5T/3T fastMRI knee/neuro data. We observe that the signal-to-noise ratio of the examined sequences plays an essential role, and we statistically prove the hypothesis that both the type and amount of training data are less important for low acceleration factors. Finally, a visualization tool facilitating the examination of the networks' performance on each individual subject of the fastMRI data is provided. In this context, we identify the substantial impact certain neural network architectures and configurations have on reconstruction quality.

Keywords. Deep Learning, fastMRI, Domain Shift, Physics-based Reconstruction, Statistical Analysis,

Contents

1 Introduction

With the rapid development of digital technologies and the increasing complexity of required algorithms, Artificial Intelligence (AI) has become a terminology and methodology frequently used in a broad spectrum of applications. Applied in medical imaging, Deep Learning (DL), a further sub-field of AI and Machine Learning (ML), has enabled and improved several methodologies in radiology beyond what was previously thought possible. For instance, in order to automate brain tumor image segmentation, DNNs were utilized to develop efficient AI solutions due to their versatility and performance [7].

1.1 Past, Present, and Future of Artificial Intelligence in Medicine

Simultaneous with the rise of DL, the usage of AI has increased in medical sciences. We are currently at the emergence of a new era of Medical Technologys (MTs) in which AI is fused into daily clinical decisions and it has become crucial for us to understand the development of AI in Medicine (AIM).

Today, AI has found its usage in a wide variety of applications in the 4P model of medicine (Predictive, Preventive, Personalized, and Participatory). Above all, the application of DL to medical imaging has shown promising results, even outperforming the diagnosis of radiologists in detecting pneumonia in chest x-rays [17]. In medical imaging, DL has shown promising results for numerous types of problem statements and widely used imaging techniques, including Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Outside of radiology, AI has also been proposed to improve health systems, patient autonomy, data analysis, and computer-assisted diagnosis.

Despite the fact that AI has already established its usage in many fields today, the application of this relatively new technology in such a sensitive field as medicine awaits many formidable roadblocks regarding regulatory and ethical approval and explainability in the near future. Although deep learning seems to be as reliable, AI solutions have shown limitations and require further clinical validation. Limitations include the susceptibility to security breaches as well as the requirement of human surveillance: As is often said, AI will not replace radiologist but rather support them in their diagnosis.

1.2 Deep Learning in Magnetic Resonance Imaging Reconstruction

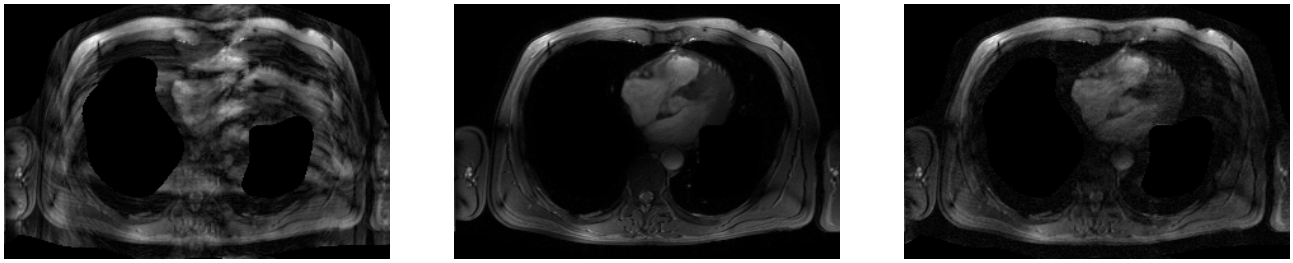
One of the most prominent use cases of DL is to compensate for the long acquisition duration of MRIs. While research on CT mainly focuses on decreasing the ionizing radiation, the long acquisition time is a major concern for MRI research. The quality of MR acquisitions highly depends on the patient's ability to remain still as movements during the scan negatively impact the image. Therefore, the acquisition duration can prove to be difficult for many patients including children and claustrophobic

patients. Other drawbacks, including high costs and lacking patient comfort, are further incentives for reducing the MRI acquisition time. However, despite efforts in accelerating MR acquisition without affecting the quality of the output, the theoretical nature of this imaging approach limits the number of frequency samples able to be recorded during a short period of time.

In the MRI reconstruction problem to facilitate the acceleration of MRI, the goal is to find a function to retrieve the reconstruction x$\in \mathbb{C}^{N_x}$ from the retrieved undersampled k-space MR signal s$\in \mathbb{C}^{N_y}$ corrupted by noise resulting from the shortened imaging duration $\epsilon \in \mathbb{C}^{N_y}$ following

$$s = Ex + \epsilon \tag{1}$$

incorporating an linear encoding operator $E : \mathbb{C}^{N_x} \to \mathbb{C}^{N_y}$ [15].



(a) Undersampled MRI (b) Fully-sampled target MRI (c) Reconstruction using SENSE

Figure 1: Comparison of One Frame in 2D+t Cardiac MRI: Undersampled input with acceleration factor R=8, fully-sampled target, and SENSE reconstruction

1.3 Research Objectives and Outline

Without a doubt, deep learning is a powerful methodology accompanied by an extensive list of potential improvements and breakthroughs in the field of medical imaging. The focus of this paper is devoted to common aspects in static and dynamic learned MRI reconstruction as well as its challenges to introduce multifaceted conclusions and suggestions for usage.

The objective is to shed light on the impact of domain shift in learned MRI reconstruction. While prior research on MRI reconstruction has focused on the implementation of deep learning algorithms and their evaluation on image quality [1, 4, 10, 16, 20], the topic of domain shift has rarely ever been studied. However, due to a deep learning model's natural dependency on the given data, this domain adaptation has proved to have a substantial impact on the performance of the deep learning algorithms in most other deep learning applications [11, 14]. This challenge is especially significant for academic and potential clinical applications of MRI reconstruction. The questions that then naturally arise are "How much of an impact do different data configurations of domain shift have on different networks?" and "Which networks are least prone to varying degrees domain shift in clinical applications?". For this, we provide visualization tools and statistically investigate the impact of domain shift in the context of state-of-the-art MRI reconstruction networks with respect to variations in training data.

Finally, based on the outcomes of our multifaceted analysis, generally applicable conclusions and recommendations for future research, application, and development are drawn.

2 Methods

2.1 Theoretical Foundations of DL-based Reconstruction

First, the theoretical foundations applicable on both DL-based reconstruction of static and dynamic imaging has to be outlined. There are many ML-based approaches to retrieving the image $x \in \mathbb{C}^{N_x}$ from the inverse problem (1) including image denoising, physics-based reconstruction, and direct mapping. Accordingly, the task is also defined differently with varying network designs and input data [5].

2.1.1 Image Denoising

In image denoising, the problem definition of image reconstruction is simplified to a regression task. Specifically, it is an image-to-image regression problem in which the NN learns to predict a continuous outcome for each pixel in the given complex-valued input image ($\mathbb{K} \in \mathbb{R}^{\mathbb{N}_1}/\mathbb{K} \in \mathbb{C}^{\mathbb{N}_1}$), generating an image in the respective number system. In mathematical terminology, the network is trained with the MRI data previously transformed to image space from k-space to learn the mapping function

$$f_\theta(x) : \mathbb{K}^{N_1} \rightarrow \mathbb{K}^{N_2} \tag{2}$$

where N_1 and N_2 denote the dimensionality of the input and output image and are therefore equal (N_1=N_2). The complex input image space is made up of a real and an imaginary number per pixel representing the pixel-wise intensity. However, the mapping function in trained image denoising, as its name suggests, is confined to the MRI image space and therefore does not incorporate the raw k-space frequency signals during denoising. In consequence, the underlying extensive physical information crucial for image reconstruction in k-space is disregarded [5].

2.1.2 A Case of Physics-based Reconstruction: Unrolled Optimization

To solve the inverse problem in Eq. (1), the regularized reconstruction problem

$$x^* \in \underset{x \in \mathbb{C}^{N_x}}{\arg\min}\ \lambda \mathcal{D}[Ex, s] + \mathcal{R}[x] \tag{3}$$

is minimized through learning. This reconstructed approximation of x incorporates a regularization term $\mathcal{R}[x]$ learned from data and the DC term $\mathcal{D}[Ex, s] = \frac{1}{2} \|Ex - s\|_2^2$ balanced by λ in a "DC layer" [6]. The unrolled optimization algorithm for the MRI reconstruction problem is defined as

$$x^{it+\frac{1}{2}} = x^{it} - f_{\theta^{it}}(x^{it}) \tag{4}$$

$$x^{it+1} = g(x^{it+\frac{1}{2}}, s, E) \tag{5}$$

by Hammernik et al [6, 4] for $0 \leq it \leq T$ with current iteration it and total number of iterations T. For the regularization network $f_\theta(x)$ before each DC layer, we employ either a U-Net [16], 5-layer CNN,

or DUNET [6] with complex-valued input and output as well as varying mathematical operations to process the complex-valued input. By including the DC layer, the optimization problem receives information regarding the physics of raw MR data and MR acquisition and therefore is able to assess the learning progress, in contrast to image denoising, based on the k-space data [6]. Further physics-based reconstruction approaches involve k-space learning and hybrid learning [5].

2.1.3 Data Consistency Layers in Unrolled Optimization Networks

DC layers guarantee k-space consistency, and hence are essential to incorporating the underlying MRI physics in the raw data in learned MRI reconstruction [6]. Therefore, in unrolled optimization, the network is able to not only consider the information and similarity in image space, but also in k-space. In all networks we inspected, either Gradient Descent (GD) or Proximal Mapping (PM) was used to model the DC term $\mathcal{D}[Ex, s]$. A GD scheme can be used [4, 6]

$$g_{gd}(x^{it+\frac{1}{2}}) = x^{it+\frac{1}{2}} - \lambda^{it} A^*(Ax^{it+\frac{1}{2}} - s) \tag{6}$$

In certain networks we inspected, as well as the PM-DUNET examined in chapter 3, DC was modelled by PM [1, 6]

$$g_{pm}(x^{it+\frac{1}{2}}) = \arg\min_{x} \in \mathbb{C}^{N_x} \frac{1}{2}\|x - x^{it+\frac{1}{2}}\|_2^2 + \frac{\lambda}{2}\|Ex - s\|_2^2. \tag{7}$$

As a result, our unrolled optimization networks consist of a fixed number of iterations that comprise either a GD or a PM modelled DC layer and a N-layer CNN as the "denoising" regularization network as shown in Fig. 2.

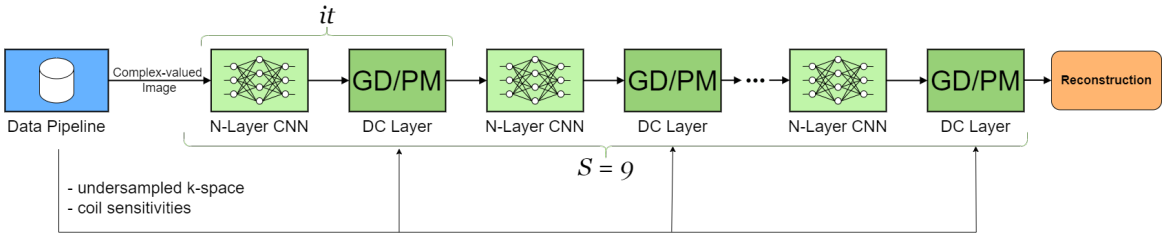


Figure 2: CNN-Net Unrolled Optimization Reconstruction DNN containing N-layer CNN as the regularization network and a PM/GD-modelled DC layer. The network variable N as well as convolutional layer configurations vary between experiments.

2.2 Experimental Setup

In this paper, we follow the experimental setup of Hammernik et al [6]. The key motivation of Hammernik et al [6] was to compare varying data consistency layers and regularization networks associated with reconstruction approaches, e.g. physics-based reconstruction, in state-of-the-art reconstructions networks. As a result, an extensive collection of network performances evaluated on a per-patient basis along with its MR acquisition parameters is generated.

knee x	x% knee data
joint x	mixed knee and neuro data, number of training subjects correspond to number of training subjects in knee x
joint x uni	mixed knee and neuro data with uniform distribution of pulse sequences, number of training subjects correspond to number of training subjects in knee x

Table 1: An Overview of all fastMRI data configurations and their compositions [6]. $x \in 25, 50, 100$. For instance, "knee 50" corresponds to "50% knee data".

2.2.1 The fastMRI Dataset

As described in Hammernik et al [6], the fastMRI multi-coil train knee and neuro dataset was used to train and evaluate the examined DNNs. The experiments were limited to the multi-coil dataset, disregarding the single-coil samples due to the emphasis on parallel imaging as explained in 1. This large-scale dataset owned by New York University and NYU Langone Health consists of fully-sampled ground-truth images and k-space data acquired by scanners of a magnetic field strength of 1.5 Tesla (T) and of 3T [20]. Specifically, all reconstruction networks were trained with varying data configurations shown in Table 1. Therefore, we will focus on the data acquisition parameters of the data configurations listed in Table 1 in the following. Further information on the MR sequence parameters are detailed in the original publication [20].

The knee and neuro datasets can be further divided into MR scanners and sequences distinctive for the respective anatomy. The knee MR samples were acquired on four different MR systems for clinical diagnostic usage, three of which had a magnetic field strength of 3T – Siemens Magnetom Skyra (496 scans), Prisma (47 scans), and Biograph mMR (124 scans) – and one with a magnetic field strength of 1.5T [20] – Siemens Magnetom Aera (505 scans). This dataset comprises data of two sequences mainly used in human joint diagnostics as shown in Fig. 3: Coronal proton-density weighted with fat-saturation – CORPDFS (588 scans) and Coronal proton-density weighted w/o fat-saturation – CORPD (584 scans). A noticeably higher overall SNR for samples acquired with the CORPD pulse sequence compared to CORPDFS data could be observed.

The neuro MR samples were acquired on five different MR systems for clinical diagnostic usage, three of which had a magnetic field strength of 3T – Siemens Magnetom Skyra (1625 scans), Prisma (602 scans), Biograph mMR (645 scans), and Tim Trio (478 scans) – and two a magnetic field strength of 1.5T [20] – Siemens Magnetom Avanto (1274 scans) and Aera (1223 scans). This dataset comprises data of four sequences mainly used used to detect structures in the central nervous system as shown in Fig. 3: Axial fluid-attenuated inversion recovery – AXFLAIR (451 scans), Axial T_1 weighted – AXT1 (667 scans), Axial T_1 weighted with contrast agent – AXT1POST (1236 scans), and Axial T_2 weighted – AXT2 (3493 scans).

All categories considered, we acknowledge the presence of imbalance between the amount of data acquired for the individual scanner models and MRI sequences, especially during model training.

Instead of repeating the MR imaging procedure with a reduced duration in order to generate prospectively undersampled k-space data matching the fully-sampled target acquisitions, masking was employed retrospectively to simulate accelerations factors of R=4 and R=8 [20].

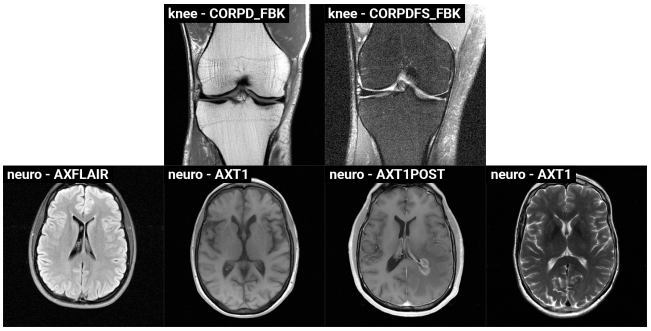


Figure 3: MR Sequences: The first row shows ground-truth samples acquired with the knee data pulse sequences. The second row shows ground-truth samples acquired with the neuro data pulse sequences.

2.2.2 Evaluated Deep Neural Networks

We employed reconstruction networks of varying architectures, including (1) three state-of-the art DL networks: U-Net [16, 20], MoDL [1], and VN [4] and (2) Down-Up Networks (DUNETs) incorporating three different data consistency layers as explained in methods, i.e. Gradient Descent (GD), Proximal Mapping (PM), and Variable Splitting (VS).

These networks along with the number of trainable parameters, reflecting on the complexity of the network, and the regularization networks are depicted in Table 2.

Network	Parameters	Regularization Network	Data Consistency Layer
U-Net	3,357,827	U-Net	None
VN	131,051	Fields-of-Experts Model	Gradient Descent
MoDL	113,155	5 layer CNN	Proximal Mapping
GD-DUNET	3,372,985	DUNET	Gradient Descent
PM-DUNET	3,372,985	DUNET	Proximal Mapping
VS-DUNET	3,372,985	DUNET	Variable Splitting

Table 2: Comparison of evaluated DNNs as described in [6].

2.2.3 Training and Validation

All networks were trained simultaneously on identical training environments for a fixed number of 60 epochs. During training on the same hardware setup, the ADAM optimizer [9] with predetermined configuration parameters along with identical loss functions was utilized. That being the case, we opted for no hyperparameter tuning prior to each experiment/training cycle whatsoever, contrary to machine learning conventions. This was specifically designed to deter possible undesirable influences from factors not investigated in this context; hence, to ensure a generalized validation environment. When training the networks on the knee dataset, the entirety of the samples was split into 80% training and 20% validation set, whereas the remarkably larger neuro dataset was split into 70% training and 30% validation set.

Furthermore, in order to examine the varying MRI reconstruction network architecture performances uniformly, both training and evaluation of each network in the initial publication [6] were performed on different data configurations of the fastMRI multi-coil validation knee and neuro dataset for the acceleration factors R=4 and R=8 and interpreted using the Structural Similarity Index Measure (SSIM). Further quality metrics such as the Mean Squared Error (MSE), the Peak Signal-to-Noise Ratio (PSNR), and the Normalized Mean Squared Error (NMSE) were computed.

2.3 Methodological Approach

In the following we will elaborate on our approach to quantitatively analyzing the different degrees of ramifications when the reconstruction networks are deployed on datasets with a different data distribution compared to the train dataset.

We differentiate between quantitative and qualitative analysis [8]. Quantitative analysis is based on objective criteria (i.e. mathematical metrics) to evaluate the quality of reconstructions and thus the performance of the reconstruction network. This way, conclusions on overall trends can be drawn from a statistical point of view, yet disregarding certain aspects such as the inequality of image similarity and medical value. Qualitative analysis for medical imaging reconstruction, on the other hand, requires human judgement from medical experts to access the medical value of the reconstruction results regarding diagnosis.

2.3.1 Statistical Analysis

In this section we present an overview on the statistical techniques and parameters we used to quantitatively inspect domain shift.

Evaluation Metrics

The adequate and accurate assessment of the received DNN output is critical in AI development. In learned MRI reconstruction, the evaluation of the reconstruction quality is commonly based on metrics on the image-level [20], not in k-space: The MSE reconstruction quality metric is calculated through the mean squared pixel-wise differences between the target and reconstructed image. In order to calculate the NMSE, the MSE is normalized by the mean squared image pixel value of the entire image [20]. The third metric utilized, the PSNR, scales the MSE metric to the maximal pixel value of the reference image (R).

The MSE, NMSE, and PSNR all are mathematically accurate measures of errors in a reconstruction result compared to the target image [19]. However, quantifying errors does not entirely correlate with perceived visual quality. In general image reconstruction, the quality of reconstruction results can often be understood as the similarity of the network output to the target data (i.e. the ground-truth k-space data) [3, 20]. This visual similarity is often determined by the similarities of the visual image structures. The usage of the mathematical SSIM is a well-established approach in identifying and assessing the differences between visible structures in order to directly evaluate them similar to a human visual perception system instead of pixel-wise comparison [19]. Possible values for the SSIM

quality metric range from -1 to 1, with 1 denoting the an identical reconstruction result. The SSIM-metric takes in the luminance l, contrast c, and structure s of the target image and reconstruction into consideration. These three relatively independent measurements are weighted and combined to the SSIM calculation

Consequently, for the sake of comprehensibility and coherence as well as usability in graphical analysis, we quantitatively analyzed domain shift in MRI reconstruction mainly based on the SSIM metric [19].

Examined Parameters

Apart from data configurations, the fastMRI knee and neuro dataset can be further broken down into several additional categories. The acquired MRI subjects were evaluated for the following three MRI acquisition variables: MRI pulse sequence, scanner model, and scanner magnetic field strength. Due to their significant impact on the SNR, the appearance, and, hence, the perception of the undersampled MRI, we will especially be focusing on these three variables of the MRI acquisition procedure in the following investigation.

Evaluation Environment

All computed metric values ("data") for each reconstructed undersampled MRI were stored along with each sample's attributes based on the sample's anatomy (neuro/knee), dataset (train/validation), and acceleration factor R (R=4/R=8). Attributes include the respective patient sample id, scanner model, pulse sequence, reconstruction network, and the data configuration the network was trained on.

Quantitative analysis of the knee and neuro data was fully completed in the high-level programming language [18] due to its versatility. We made use of its countless open-source libraries to accommodate the operations performed on the data.

When presented in graphical plots, the knee and neuro data was analyzed from different perspectives using, mostly, three types of plots: We used conventional graphical techniques such as (1) ranked lists/horizontal bar charts and (2) scatterplots to investigate (1) the performance of individual reconstructions networks/types of reconstruction networks validated on certain cohorts (determined by e.g. acceleration factor, data configuration, scanner model, MRI sequence) from a broad view and (2) the ability of individual reconstructions networks/types of reconstruction networks to e.g. generalize and minimize the impact of domain shift. Furthermore, our graphical analysis incorporated statistical graphs as box plots that display the spread, locality of the data. The visually estimated L-estimators – particularly the trimean, interquartile range, midhinge, and range revealed more in-depth trends, supporting the ensuing statistical tests. For the majority of statistical plots, the data points consisted of a pair of SSIM-values for data cohorts characterized by two contrasting variables (e.g. R=4 versus R=8). The data cohorts represented in the plots (i.e. bars or boxes) were ranked in descending order by their highest SSIM-value when it proved to be feasible and reasonable. The training dataset or validation network of every categorical data point (bar, box, dot) is identifiable through its coloring, shape, or border. For clarity, both the x-/y-axis limits and steps in all plot types were uniform for every subplots of an experiment (e.g. universal x-/y-axis limits/steps for every MRI pulse sequence when plotting the SSIM against the validation dataset configuration).

Statistical Significance

In order to identify key parameters and relationships determining the distribution of the networks' behaviours over the given data domains, we used the Mann-Whitney U Test [12, 13] with a 95% confidence interval required to reject the null hypothesis and regard the distribution as statistically significant.

In the non-parametric Mann-Whitney U Test, also referred to as the Wilcoxon Rank Sum Test, we analyze two network performance data groups to investigate a possible statistical significance in the network performance measured in SSIM when trained on two different data configurations (dataset 1 and 2), the "base" configuration that is tested against (dataset 1) being the same dataset the network is evaluated on [13]. The significance test examines whether the two given dataset distribution are correlated. The presence of statistical significance for two fastMRI training datasets would point towards lacking network generalization for the examined network when trained on training dataset 2, the dataset unequal to the validation dataset. Accordingly, the amount of training datasets the analyzed network proves to be statistically significant on could be an indicator for the overall ability of the network to generalize, and, hence, to perform under the impact of domain shift.

3 Results

This chapter investigates the impact of domain shift in the context of state-of-the-art MRI reconstruction networks with respect to variations in training data. In particular, we thoroughly examine a wide range of factors and imaging parameters of the fastMRI knee and neuro dataset that contribute to a change in the networks' performances based on statistical analysis.

3.1 General Network Performance

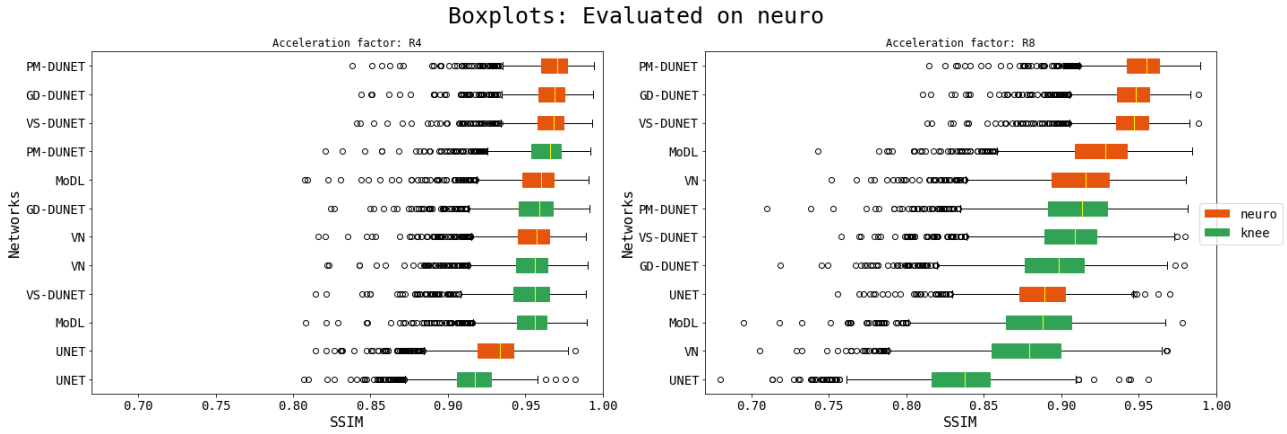
In addition to the ranked bar charts for networks evaluated on knee and neuro data examined by Hammernik et al. [6], we enhance the analysis on the reconstruction networks over the entire knee/neuro dataset with ranked box plots with descending median SSIM-values in Fig. 4, giving us a more in-depth understanding of the capabilities of different reconstruction network designs. We observe that for R = 4, the best iteration of U-Net (neuro: 0.8845/knee: 0.9170) achieves lower SSIM values than all other networks (e.g. MoDL: 0.9160/VN : 0.9248).

The L-estimators visible in the box plots, in particular the difference between the upper/lower whisker ΔW and the IQ range, shed new light on this surprising phenomenon. For the configuration neuro evaluation dataset and R=8, we report a ΔW of 0.1170/IQR of 0.0295 for U-Net [16] trained on neuro data compared to the least dispersed non-DUNET trained on neuro data (MoDL [1]) with a ΔW of 0.1262/IQR of 0.0340. For the configuration knee evaluation dataset and R=8, we report a ΔW of 0.1638/IQR of 0.0440 for U-Net [16] trained on neuro data compared to the least dispersed non-DUNET trained on neuro data (VN [4]) with a ΔW of 0.1890/IQR of 0.0542. Table 3 presents a detailed overview on the exact values of the metrics calculated in the experiment for networks evaluated on the **neuro** fastMRI dataset with **R = 8** (Fig. 4(a) R=8).

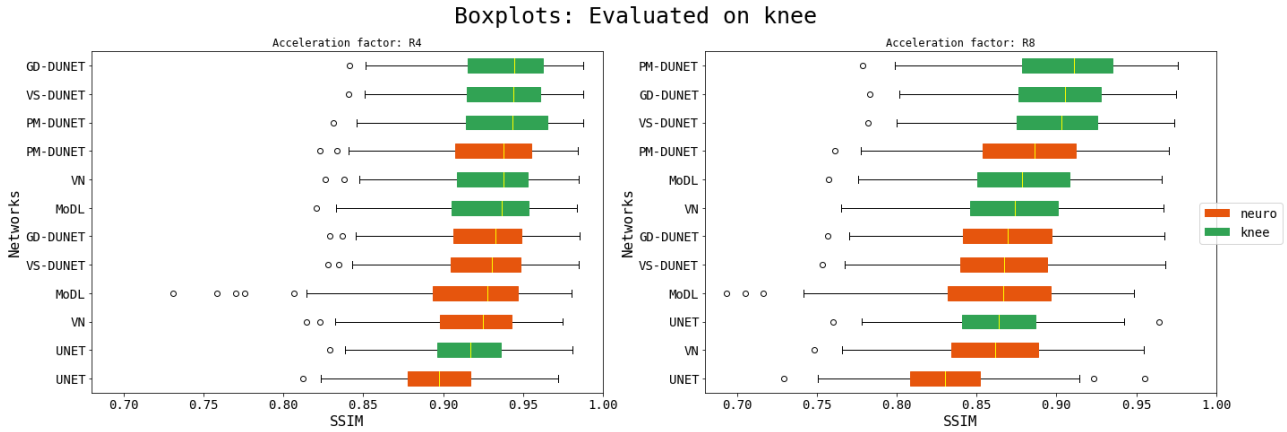
3.2 Network Generalization

The scatter plots in Fig. 5 compare the correlation of SSIM values separately on the fastMRI knee and neuro validation dataset, for R=4 and R=8, for networks that were trained separately on knee and neuro data. The positioning of SSIM values on the linear function y=x, i.e., SSIM-knee = SSIM-neuro, would represent the perfect model generalization. The scatter plots show larger discrepancies between the yellow and white outlined markers for R=8 than R=4, representing a weaker, less linear association between the results when trained on the neurological and the brain dataset than for R=4.

Furthermore, to examine the impact of scanner models on network generalization, we plotted ranked horizontal bar charts for VN [4] trained only on neuro or knee data. The performance of the reconstruction network VN evaluated on knee and neuro data at R=4/R=8 categorized into the respective scanner models is illustrated in Fig. 6(a) (knee validation dataset, R=4), Fig. 6(b) (knee



(a) Validation dataset: neuro



(b) Validation dataset: knee

Figure 4: Ranked box plots for fastMRI neuro/knee training and validation datasets at R=4 and R=8. All physics-based reconstruction networks (VN, MoDL, DUNETs), when trained on identical datasets, outperform the image denoising network U-Net regarding the upper whisker- and median SSIM-value in all experiments without exception. Preliminary results reveal varying degrees of generalization for different networks.

validation dataset, R=8), Fig. 6(c) (neuro validation dataset, R=4), and Fig. 6(d) (neuro validation dataset, R=8).

We confine our third experiment to a single network, the PM-DUNET [6] as the best performing network in most scenarios. We explicitly accounted for the performance of PM-DUNET for different scanner models, and thus specifications. The boxplots in Fig. 7 show the performance of PM-DUNET trained only on knee and neuro data [6], evaluated for scanner models at 1.5T and 3T on the knee and neuro data at R=4 and R=8. Statistical differences ($p<0.001$) between knee and neuro training data are found within scanner models of neuro validation data, indicating that the number of training subjects plays a vital role to span a larger solution manifold. For knee validation data, we only

Network	Anatomy	Lower Wh.	Median	Upper Wh.	ΔW
PM-DUNET	neuro	0.9114	0.9548	0.9900	0.0785
GD-DUNET	neuro	0.9053	0.9481	0.9832	0.0779
VS-DUNET	neuro	0.9052	0.9471	0.9827	0.0775
MoDL	neuro	0.8585	0.9285	0.9846	0.1262
VN	neuro	0.8384	0.9156	0.9806	0.1422
PM-DUNET	knee	0.8350	0.9132	0.9817	0.1466
VS-DUNET	knee	0.8386	0.9086	0.9731	0.1346
GD-DUNET	knee	0.8198	0.8983	0.9685	0.1487
U-Net	neuro	0.8295	0.8890	0.9464	**0.1170**
MoDL	knee	0.8010	0.8882	0.9671	0.1661
VN	knee	0.7888	0.8791	0.9650	0.1762
U-Net	knee	0.7611	0.8375	0.9094	**0.1483**

Table 3: L-Estimators of box plot for reconstruction networks for fastMRI knee and neuro training datasets evaluated on the neuro validation dataset at R = 8. Both UNETs show surprisingly low ΔW compared to reconstruction networks for the same training dataset. Abbreviation Wh. denotes Whisker.

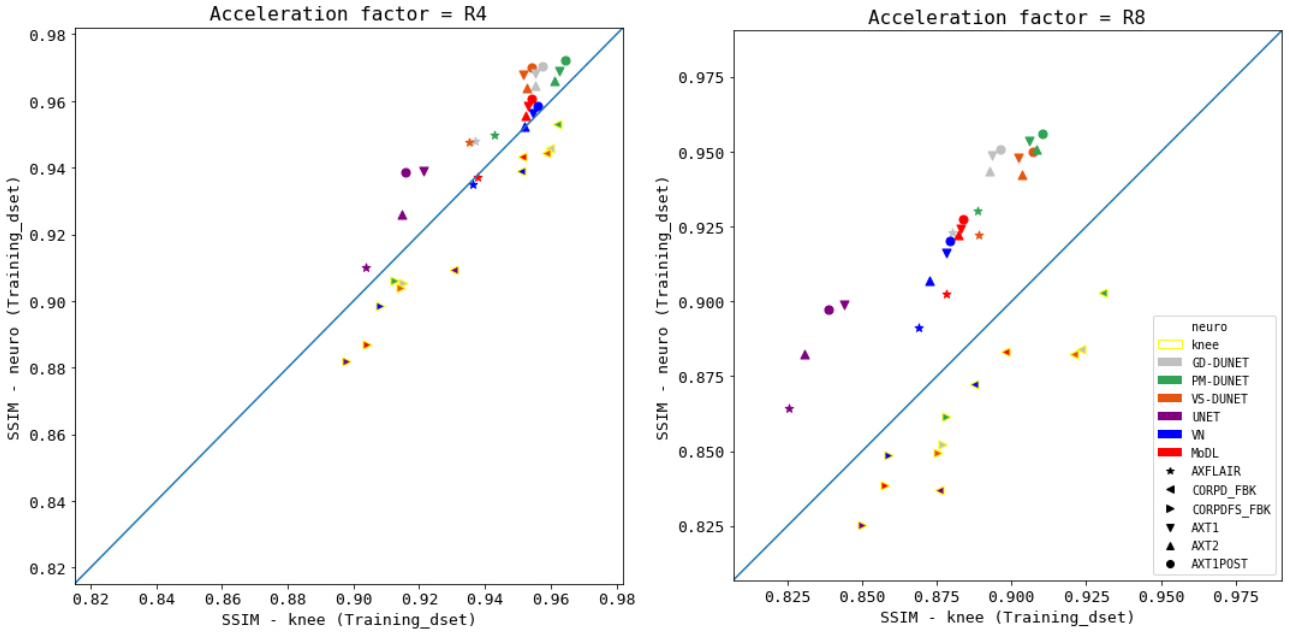
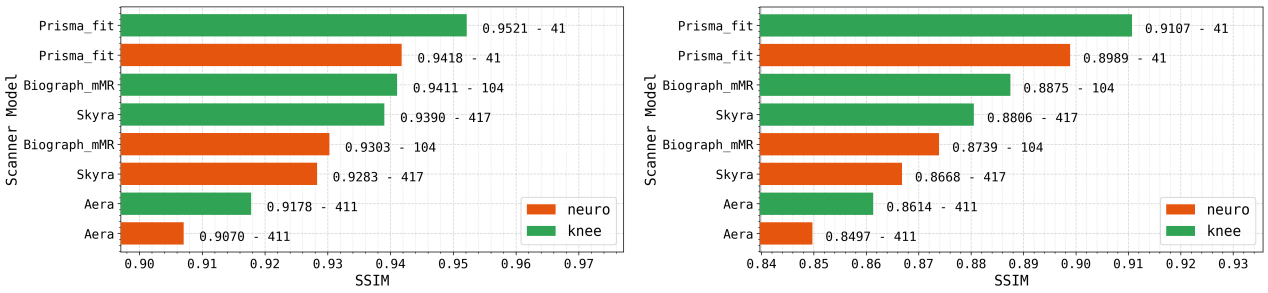


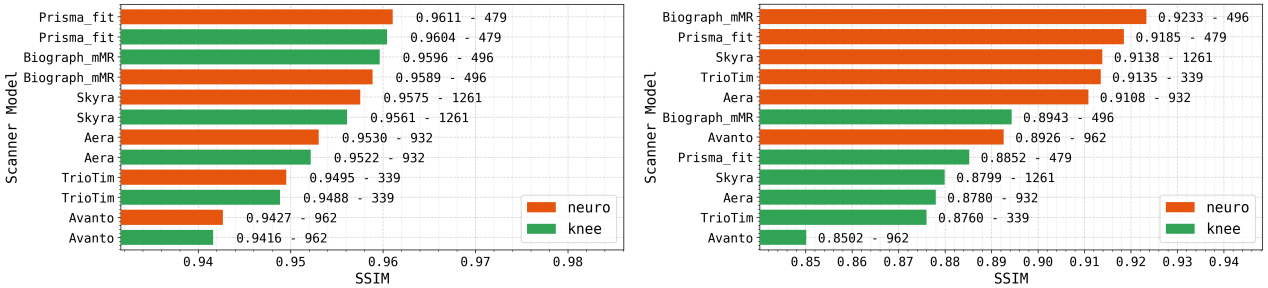
Figure 5: Scatter plots for variations in training data, for all acquisition types, for all examined networks at R=4 and R=8. Distribution of data points along the blue line represents the ideal scenario, i.e., best generalization. Data points with a yellow border were tested on knee data, without border on neuro data.

observe statistical differences for Skyra at R=4 ($p<0.01$), Skyra at R=8 ($p<0.001$), and Aera at R=8 ($p<0.01$). The plots reveal a remarkably smaller interquartile range for for both R=4 and R=8 for the neuro validation dataset.



(a) VN trained on knee at R=4, Validation dataset: neuro

(b) VN trained on knee at R=8, Validation dataset: neuro



(c) VN trained on neuro at R=4, Validation dataset: neuro

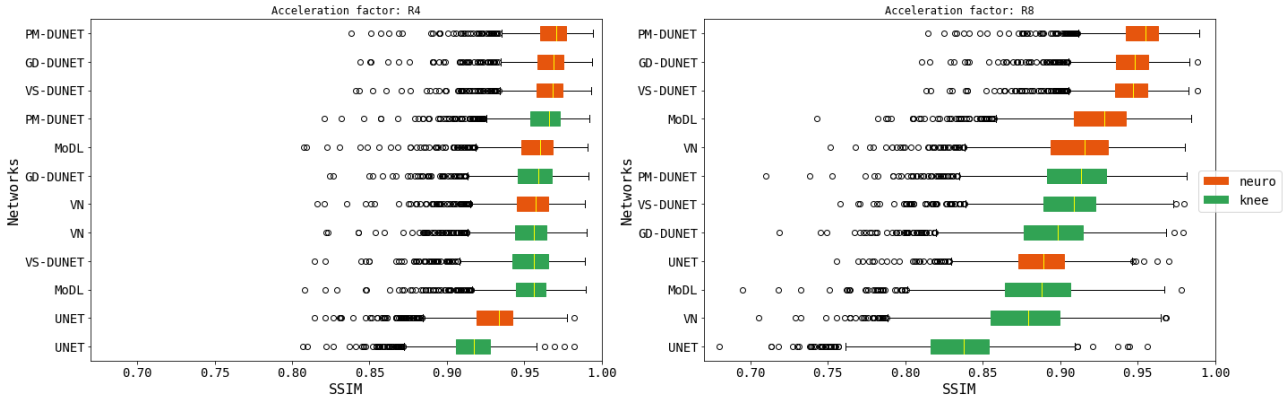
(d) VN trained on neuro at R=8, Validation dataset: neuro

Figure 6: Ranked Bar Charts for fastMRI neuro/knee training and validation datasets, evaluated for Scanner Models using reconstruction network VN [4] at R=4 and R=8. No clear correlation between the number of training samples for each individual scanner model and the network performance can be identified. When evaluated on data acquired on 1.5T scanners (Avanto, Aera) with lower SNR, reconstructions of VN tend to have lower SSIM than for 3T data.

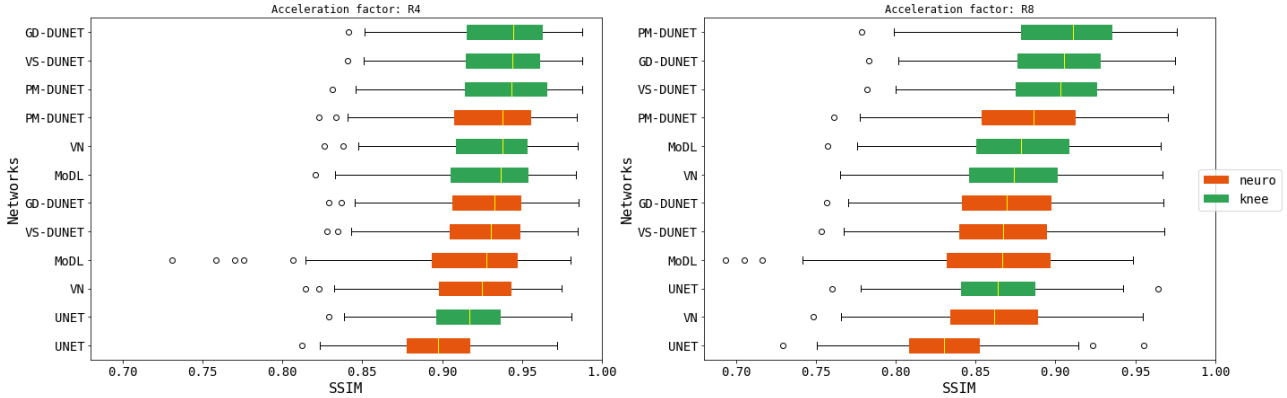
The performance of PM-DUNET trained on all data configurations, evaluated for the knee MRI sequences CORPD-FBK/CORPDFS-FBK at R=4 (Fig. 8, (a)) and R=8 (Fig. 8, (b)) is illustrated in Fig. 8. Based on these box plots, we study the reconstruction results of PM-DUNET evaluated individually on the sequences CORPD-FBK and CORPDFS-FBK of the fastMRI knee validation set at R=4 and R=8. Significant differences ($p<0.05$) between training with knee 100 against all other anatomies are marked with red stars. It is important to note that CORPDFS-FBK measurements have lower SNR compared to CORPD-FBK.

3.3 Subject-to-Network Performance Visualization Tool

Fig. 9 depicts the SSIM values for each individual subject of the fastMRI knee and neuro validation set, reconstructed with the six state-of-the-art networks. Specifically, the ranked scatterplot shows the selection of a specific reconstruction sample computed by GD-DUNET, a DUNET using Data Consistency (DC) layers modelled by GD [6]. Moreover, the reconstruction samples can be further categorized into scanner designs, MR sequences, field strength, and other variables in order to dissect the performance distributions. The categories, being the reconstruction networks in Fig. 9, are



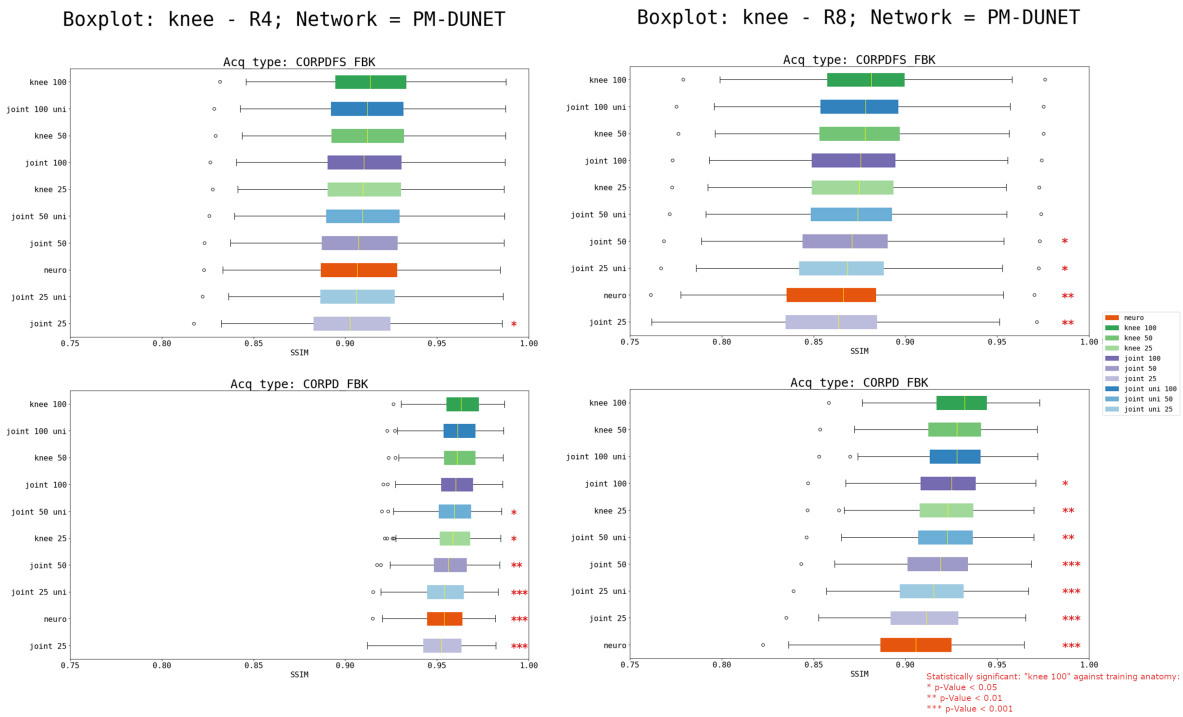
(a) Evaluation of PM-DUNET on neuro data



(b) Evaluation of PM-DUNET on knee data

Figure 7: Comparison of PM-DUNET when trained on knee (green bars) and neuro (orange bars) data, evaluated for scanner models at R=4 and R=8. In (a), statistical differences ($p<0.001$) are found within all scanner models for neuro data. Furthermore, we observe substantially worse reconstruction quality for 1.5T Avanto, and large outliers and standard deviations for 1.5T Aera. In (b), statistical differences are found within scanner models in knee data for Skyra at R=4 ($p<0.01$), Skyra at R=8 ($p<0.001$) and Aera at R=8 ($p<0.01$).

visualized as ranked lists with descending maximum reconstruction quality measured in SSIM. This visualization allows us to examine which subjects were reconstructed best/worst for the individual networks, and identify outliers.



(a) R=4 (b) R=8

Figure 8: Boxplots for variations in training data, evaluated individually for CORPD-FBK and CORPDFS-FBK of the fastMRI knee validation set, for PM-DUNET at R=4 and R=8. CORPDFS-FBK is statistically less affected by domain shift compared to CORPD-FBK. The red stars mark statistical significance (p-values<0.05).

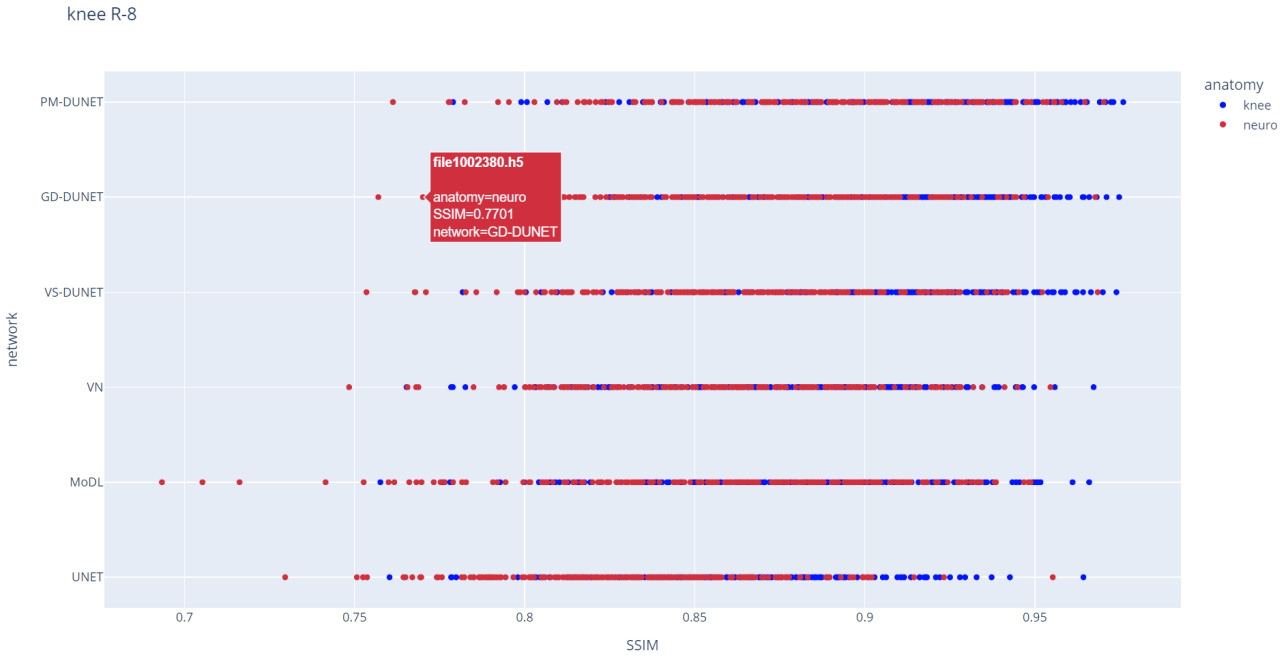


Figure 9: A snapshot illustrating the usage of our plotly-based visualization tool, examining the data samples, the training anatomy, and their respective SSIM values categorized by the evaluation network for an acceleration factor of R=8. This visualization facilitates more efficient identification and analysis of abnormalities/outliers (individual subjects).

4 Discussion

Using graphical techniques and statistical tests, we were able to identify key parameters and relationships determining the distribution of the performance of the networks over the data configurations.

4.1 General Network Performance

We verify the claims made by Hammernik et al. [6]: Incorporating the L-estimators, our point of view supports the premise that physics-based reconstruction networks (VN, MoDL, DUNETs), when trained on identical datasets, outperform the image denoising network U-Net regarding the upper whisker- and median SSIM-value in all experiments without exception.

Additionally, the results demonstrate two further findings. First, when evaluated on (neuro/knee) data acquired with R = 4, even the worst reconstruction network based on the median SSIM-value impacted by domain shift outperforms the best iteration of U-Net [16], i.e. U-Net trained on neuro and knee data. Second, we observe predominantly lower dispersion in reconstruction quality in two of our four experiments (neuro R=8/knee R=8). This observation reveals a pattern for those evaluated on data undersampled with R = 8, and hence evaluated on data with naturally lower SNRs.

Looking at the the difference between the upper/lower whisker ΔW, the U-Net [16] outperforms all reconstruction networks, including the DUNETs [6], when the analysis is confined to the dispersion rate. Together, the results from examining ΔW and the IQ-range point towards "intra-cohort" network generalization that is equally well, if not even better than state-of-the-art non-DUNET approaches when applied on low SNR data (R = 8). This observation is further verified by the lower standard deviation, another measurement of the data variance/dispersion, that can be measured for U-Net [16] when evaluated on R = 8, that is, when evaluated on data with low SNR translating to low acquisition quality.

4.2 Network Generalization

Google Developers have noted that network generalization "refers to your model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model" [2], revealing the inverse correlation between network generalization and the impact of domain shift on reconstruction networks. With our experiments, we explicitly accounted for the performance impact on individual reconstruction networks when introducing variations in data distribution between training an test data. Often, our experiments dissected the influence of specific factors such as MRI sequence, scanner model, and acceleration factor in the greatly imbalanced and heterogeneous fastMRI knee and neuro dataset [20]. For this purpose, we examined the generalization potential for the investigated reconstruction networks using box plots and scatterplots to visualize the networks' behaviour under different circumstances for scanner models, and knee/neuro sequences.

We observe that the models for R=4 generalize substantially better than for R=8, hence, the type of training data is less important for low accelerations. The best performing network in all cases is PM-DUNET [6] and the worst performing network U-Net [16].

Furthermore, we observe no clear correlation between the number of training samples for each individual scanner model and the network performance when evaluated on sampled acquired on the individual scanner models. In fact, contrary to deep learning beliefs – "The more data, the better", scanner models with the smallest percentage of total training data (Prisma fit) show the best performance for VN in certain scenarios (knee R=4, knee R=8, neuro R=4). On the contrary, scanner models that take up a large share of the knee/neuro training dataset (Aera, Aera/Avanto, Aera) show moderate to poor performance for VN. Our results also reveal that when evaluated on data acquired on 1.5T scanners with lower SNR, reconstructions tend to have lower SSIM than for 3.0T data. Consequently, we conclude that network generalization is not based on the amount of data, but is rather dependant on the quality of the data (SNR).

Also, we conclude that the large number of neuro training data generalizes well for knee data. Furthermore, we observe substantially worse reconstruction quality of neuro data at 1.5T Avanto and Aera. We suspect a potential source of this behaviour in a low SNR. We also partially confirm the assumption that the networks generalize better for R=4 than for R=8 regarding the training data configurations: The boxplots in Fig. 7 for PM-DUNET evaluated on neuro at R=8 show a clear aggregation of neuro and knee training data, regardless the scanner model; We observe a clear segregation between the worst applicable scanner model Avanto for PM-DUNET trained on neuro and the best applicable scanner model Biograph nMR for PM-DUNET trained on knee, whilst the performance on the scanner models at R=4 for the neuro and knee training datasets seem more evenly distributed.

With PM-DUNET, we observe that training with knee 100, knee 50, and joint uni 100 data yields no statistical difference for CORPD-FBK, at R=4 and R=8. Therefore, our statistical analysis supports that the type and amount of training data are critical for R=8. Low SNR, i.e, CORPDFS-FBK, data generalize better for a wide range of training configurations, while having a lower SSIM and a high standard deviation.

4.3 Subject-to-Network Performance Visualization Tool

During the evaluation the reconstruction networks' performances, we identified the lack of transparency regarding the SSIM-distribution of individual patient samples. Therefore, we propose an interactive tool organized as a ranked scatterplot illustrating all data samples and their respective anatomy categorized by networks and sorted by the SSIM-values (Fig. 9). This way, having the ability to inspect the undersampled MRI input, ground-truth, and learned reconstructions, our tool will aid researchers to further examine possible underlying factors (wrong labels, poor acquisition quality) to the discrepancies in reconstruction quality (based on the SSIM metric). As researchers looks into individual MR reconstruction data points using the proposed visualization tool, they would be able to efficiently conduct qualitative analysis of individual samples and variables accessible in the evaluation dataset. We had previously not been able to specifically dig into individual samples in a time-efficient manner and hence were limited to quantitative analysis. Another potential use case could be to identify the subjects that were reconstructed the best/worst for the individual networks.

5 Conclusion

Over the past few decades, the vision of AI-based methodologies in medicine have developed into one of the most popular ideas in MT literature, including a wide range of possible applications and benefits. Nevertheless, many gaps in literature and technical prerequisites still remain. In learned MRI reconstruction, a fundamental problem had previously been the absence of a large-scale, high-quality, diverse dataset of MRI samples for the research and development of DL-based approaches. The introduction of the fastMRI knee and neuro dataset [20] provided an opportunity to examine and dissect effects between scanner models, field strengths, and anatomies for domain shift. In spite of merely incorporating 2D high-field strength (1.5T, 3.0T) MRI acquisitions, the investigation of DL-based undersampled MRI reconstruction based on fastMRI data shed light on the key parameters and their respective influence on reconstruction quality in general. Similarly, conclusions for network design and selection for undersampled MRI reconstruction in general can be derived from exploring the fastMRI dataset in conjunction with ML evaluation results. For instance, a superiority of physics-based learning and the respective reconstruction neural networks could be identified in an environment experiencing domain shift. Key factors such as the ability to adapt to previously unknown data, different to the dataset used to train the network [2] identified by this paper are crucial for potential future approval and usage in real-world clinical environments.

Hence, in this work, the impact of domain shift for state-of-the-art neural networks in undersampled MRI reconstruction on the highly heterogeneous fastMRI dataset was investigated. First, the claims made by Hammernik et al [6] regarding general network performance were verified. The validation of these claims also reveal the varying ability for networks to generalize partially due to their varying ability to model the acquisition physics (physics-based reconstruction vs image denoising). This paper statistically proved that networks trained for R=4 are less prone to domain shift, hence, the type and amount of training data are less critical at low accelerations. However, different reactions for the knee and neuro data to domain shift could be observed, and the results indicate that this might be related to differences in SNR rather than differences in anatomy. However, my investigations disregard the drawbacks of mathematical image quality quantifiers. For clinical applicability, quantitative analysis of image quality is not sufficient and support from medical specialists, though with substantial overheads, is required to individually rate the reconstructed images with respect to their diagnostic value. Accordingly, the experiments also identified the lacking transparency and the inefficiency in consequence when it comes to large-scale qualitative analysis in the field of MRI reconstruction. Therefore, the design of a scalable interactive tool was implemented to simplify the inspection of network performances on individual MRI subjects. This provides a good starting point for further qualitative analysis of domain shift from a medical point of view and can be found encapsulated in a Jupyter Notebook, along with the source code, at https://github.com/h3seas0n/ismrm2022-domainshift-fastMRI.

Acknowledgments

This research project marks an important milestone in the early stages of my life, forming me and my future. This unique, invaluable time would not have been possible without many I've met along this journey.

First of all, I would like to profoundly thank my thesis advisor Dr. Kerstin Hammernik from the Chair for AI in Medicine and Healthcare at the Technical University of Munich for her patient guidance, continuous encouragement, inspiring insights and perspective throughout my research – She introduced me to the intriguing world of academic research and scientific conferences. I would also like to thank Dr. Veronika Zimmer for her valuable guidance in the world of statistical analysis.

I am also truly grateful to Prof. Daniel Rückert and the TUM School of Education for establishing this connection, and hence facilitating this life-changing research opportunity. Thank you Prof. Rückert for your openness towards me as a secondary school student and your support during the writing of the ISMRM conference abstract.

Furthermore, I would like to extend my gratitude to the organizational team of the TUMKolleg program for enabling this learning experience. Here, I am particularly grateful to Mrs. Katrin Lison for being my advisor at our school as well as Mr. Markus Stöckle and Dr. Ralf Laupitz for their efforts in finding and introducing me to a suitable research project.

I would also like to express my gratitude towards our collaborators at the Physikalisch-Technische Bundesanstalt (PTB) in Berlin, Germany for providing us with the foundational 7T data for our research in 2D+t cardiac MRI reconstruction. A specific acknowledgement must address Dr. Christoph Aigner for his expertise in 7T MRI.

Finally, special thanks go to my family and friends for their moral support, friendship, and love.

Bibliography

[1] H. K. Aggarwal, M. P. Mani, and M. Jacob. "MoDL: Model-based deep learning architecture for inverse problems." In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 394–405.

[2] G. Developers. *Generalization*. 2020. URL: https://developers.google.com/machine-learning/crash-course/generalization/video-lecture (visited on 12/26/2021).

[3] J. R. Fienup. "Invariant error metrics for image reconstruction." In: *Applied optics* 36.32 (1997), pp. 8352–8357.

[4] K. Hammernik, T. Klatzer, E. Kobler, M. Recht, D. K. Sodickson, T. Pock, and F. Knoll. "Learning a Variational Network for Reconstruction of Accelerated MRI Data." English (US). In: *Magnetic Resonance in Medicine* 79.6 (June 2018), pp. 3055–3071. ISSN: 0740-3194. DOI: 10.1002/mrm.26977.

[5] K. Hammernik, T. Küstner, and D. Rueckert. "Machine Learning For MRI Reconstruction."

[6] K. Hammernik, J. Schlemper, C. Qin, J. Duan, R. M. Summers, and D. Rueckert. "Systematic evaluation of iterative deep neural networks for fast parallel MRI reconstruction with sensitivity-weighted coil combination." In: *Magnetic Resonance in Medicine* 86.4 (2021), pp. 1859–1872. DOI: https://doi.org/10.1002/mrm.28827. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.28827.

[7] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. "Brain tumor segmentation with deep neural networks." In: *Medical image analysis* 35 (2017), pp. 18–31.

[8] Q. Huang, D. Yang, P. Wu, H. Qu, J. Yi, and D. Metaxas. *MRI Reconstruction via Cascaded Channel-wise Attention Network*. 2019. arXiv: 1810.08229 [cs.CV].

[9] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980* (2014).

[10] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson. "Assessment of the generalization of learned image reconstruction and the potential for transfer learning." In: *Magnetic resonance in medicine* 81.1 (2019), pp. 116–128.

[11] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang. "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2507–2516.

[12] H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491.

[13] N. Nachar et al. “The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution.” In: *Tutorials in quantitative Methods for Psychology* 4.1 (2008), pp. 13–20.

[14] E. H. Pooch, P. L. Ballester, and R. C. Barros. “Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification.” In: *arXiv preprint arXiv:1909.01940* (2019).

[15] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger. “SENSE: sensitivity encoding for fast MRI.” In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pp. 952–962.

[16] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer. 2015, pp. 234–241.

[17] E. J. Topol. “High-performance medicine: the convergence of human and artificial intelligence.” In: *Nature Medicine* 25.1 (Jan. 2019), pp. 44–56. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7.

[18] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image quality assessment: from error measurement to structural similarity.” In: *IEEE transactions on image processing* 13.1 (2004).

[20] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, et al. “fastMRI: An open dataset and benchmarks for accelerated MRI.” In: *arXiv preprint arXiv:1811.08839* (2018).