Robustness of Deep Learning for Accelerated MRI: Benefits of Diverse Training Data

Kang Lin [1] **Reinhard Heckel** [1]

Abstract

Deep learning based methods for image reconstruction are state-of-the-art for a variety of imaging tasks. However, neural networks often perform worse if the training data differs significantly from the data they are applied to. For example, a model trained for accelerated magnetic resonance imaging (MRI) on one scanner performs worse on another scanner. In this work, we investigate the impact of the training data on a model's performance and robustness for accelerated MRI. We find that models trained on the combination of various data distributions, such as those obtained from different MRI scanners and anatomies, exhibit robustness equal or superior to models trained on the best single distribution for a specific target distribution. Thus training on such diverse data tends to improve robustness. Furthermore, training on such a diverse dataset does not compromise in-distribution performance, i.e., a model trained on diverse data yields in-distribution performance at least as good as models trained on the more narrow individual distributions. Our results suggest that training a model for imaging on a variety of distributions tends to yield a more effective and robust model than maintaining separate models for individual distributions.

1. Introduction

Deep learning models trained end-to-end for image reconstruction are fast and accurate and outperform traditional methods for a variety of imaging tasks ranging from denoising over super-resolution to accelerated magnetic resonance imaging (MRI) (Jin et al., 2017; Dong et al., 2014; Muckley et al., 2021).

Imaging accuracy is typically measured as in-distribution performance: A model trained on data from one source is applied to data from the same source. However, in practice a neural network for imaging is typically applied to slightly different data than it is trained on. For example, a neural network for accelerated magnetic resonance imaging trained on data from one hospital is applied in a different hospital.

Neural networks for imaging often perform significantly worse under such distribution-shifts. For accelerated MRI, a model trained on knees performs worse on brains when compared to the same model trained on brains. Similar performance loss occurs for other natural distribution-shifts (Knoll et al., 2019; Johnson et al., 2021; Darestani et al., 2021).

To date, much of research in deep learning for imaging has focused on developing better models and algorithms to improve in-distribution performance. Nevertheless, recent literature on computer vision models, in particular multi-modal models, suggest that a model's robustness is largely impacted by the training data, and a key ingredient for robust models are large and diverse training sets (Fang et al., 2022; Nguyen et al., 2022; Gadre et al., 2023).

In this work, we take a step towards a better understanding of the training data for learning robust deep networks for imaging, in particular for accelerated MRI.

First, we investigate whether deep networks for accelerated MRI compromise performance on individual distributions when trained on more than one distribution. We find for various pairs of distributions (different anatomies, image contrasts, and magnetic fields), that training a single model on two distributions yields the same performance as training two individual models.

Second, we demonstrate for a variety of distribution-shifts (anatomy, image contrast, and magnetic field shift) that the robustness of models, regardless of its architecture, is largely determined by the training set. A diverse set enhances robustness towards distribution-shifts. We further show that robustness improvements highly correlate with similarity between train and test set.

Third, we consider a distribution-shift from healthy to non-

[1]Department of Computer Engineering, Technical University of Munich, Munich, Germany. Correspondence to: Kang Lin <ka.lin@tum.de>, Reinhard Heckel <reinhard.heckel@tum.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Ground truth fastMRI brain $\mathcal{D}_P$

Figure 1: An illustrative (randomly chosen) example to demonstrate benefits of training on a large and diverse dataset: Shown are reconstructions from two VarNets (Sriram et al., 2020), one trained on fastMRI brain, the largest single dataset of brain images for accelerated MRI, and one trained on a diverse collection of datasets $\mathcal{D}_P$. Both models are evaluated out-of-distribution on an image from the CC-359-sagittal (Souza et al., 2018) dataset. The model trained on fastMRI brain shows severe artifact whereas the model trained $\mathcal{D}_P$ provides better details and fewer artifacts.

healthy subjects and find that models trained on a diverse set of healthy subjects can reconstruct images with pathologies as accurately as models trained on images containing pathologies.

Fourth, we empirically find for several distribution-shifts that what we call 'distributional overfitting' occurs: When training for long, in-distribution performance continues to improve slightly while out-of-distribution performance sharply drops. A related observation was made by Wortsman et al. (2022) for fine-tuning of CLIP models. Therefore, early stopping can be helpful for training a robust model as it can yield a model with almost optimal in-distribution performance without losing robustness.

Taken together, those four findings suggest that training a single model on a diverse set of data distributions and incorporating early stopping yields a robust model. We test this hypothesis by training a model on a **large and diverse pool of data** significantly larger than the fastMRI dataset (Zbontar et al., 2019). The resulting model, as shown in Figure 1, is significantly more robust than a model trained on the fastMRI dataset, the single largest dataset for accelerated MRI, without compromising performance on fastMRI data.

Related work. Several works have shown that deep learning models for accelerated MRI are sensitive to distribution-shifts. Johnson et al. (2021) found the models submitted to the 2019 fastMRI challenge (Knoll et al., 2020) to be sensitive to distribution-shifts. Furthermore, Darestani et al. (2021) show that reconstruction methods for MRI, regardless of whether they are trained or only tuned on data, all exhibit similar performance loss under distribution-shifts. Contrary to our work, both works do not propose robustness enhancing strategies, such as training on a diverse dataset.

Moreover, there are several works that characterise the severity of specific distribution-shifts and propose transfer learning as a mitigation strategy (Knoll et al., 2019; Huang et al., 2022; Dar et al., 2020). Those works fine-tune on data from the test distribution, whereas we study a setup without access to data from the test distribution.

A potential solution to enhance robustness in accelerated MRI is test-time training to narrow the performance gap on out-of-distribution data (Darestani et al., 2022), albeit at high computational costs. Liu et al. (2021) propose a special model architecture for improving performance of training on multiple anatomies. Ouyang et al. (2023) proposes an approach that modifies natural images for training MRI reconstruction models. In ultrasound imaging, Khun Jush et al. (2023) demonstrate that diversifying simulated training data can improve robustness on real-world data.

More broadly, several influential papers have shown that machine learning methods for problems ranging from image classification to natural language processing perform worse under distribution-shifts (Recht et al., 2019; Miller et al., 2020; Taori et al., 2020; Hendrycks et al., 2021).

Shifting to computer vision, OpenAI's CLIP model (Radford et al., 2021) is robust under distribution-shifts. Fang et al. (2022) finds that the key contributor to CLIP's robustness is the diversity of the training set. However, Nguyen et al. (2022) show that blindly combining data sources can weaken robustness compared to training on the best individual data source.

These studies underscore the pivotal role of dataset design, particularly data diversity, for a model's performance and robustness. In light of concerns regarding the robustness of deep learning in medical imaging, we explore the impact of data diversity on models trained for accelerated MRI.

While increasing the training set size generally improves performance, often following a power law (Kaplan et al., 2020; Zhai et al., 2022; Klug & Heckel, 2023), this work focuses on out-of-distribution improvements through diversity, rather than in-distribution improvements through dataset size. It's worth noting that we specifically address out-of-distribution robustness, while other notions exist, such as worst-case robustness (Antun et al., 2020; Ducotterd et al., 2022; Krainovic et al., 2023).

2. Setup and Background

We consider multi-coil accelerated MRI, where the goal is to reconstruct a complex-valued image $\mathbf{x} \in \mathbb{C}^N$ from measurements of electromagnetic signals obtained through C receiver coils according to

$$\mathbf{y}_i = \mathbf{MFS}_i\mathbf{x} + \mathbf{z}_i \in \mathbb{C}^m, \quad i = 1, \ldots, C. \tag{1}$$

Table 1: Fully-sampled k-space datasets used here. The percentages are the proportions of the data within a dataset. Scans containing multiple echoes or averages are separated as such and counted as separate volumes.

Dataset	Anatomy	View	Image contrast	Vendor	Magnet	Coils	Vol./Subj.	Slices
fastMRI knee (Zbontar et al., 2019)	knee	coronal	PD (50%), PDFS (50%)	Siemens	1.5T (45%), 3T (55%)	15	1.2k/1.2k	42k
fastMRI brain (Zbontar et al., 2019)	brain	axial	T1 (11%), T1POST (21%), T2 (60%), FLAIR (8%)	Siemens	1.5T (43%), 3T (67%)	4-20	6.4k/6.4k	100k
fastMRI prostate (Tibrewala et al., 2023)	prostate	axial	T2	Siemens	3T	10-30	312/312	9.5k
M4Raw (Lyu et al., 2023)	brain	axial	T1 (37%), T2 (37%), FLAIR (26%)	XGY	0.3T	4	1.4k/183	25k
SKM-TEA, 3D (Desai et al., 2021)	knee	sagittal	qDESS	GE	3T	8, 16	310/155	50k
Stanford 3D (Epperson, 2013)	knee	axial	PDFS	GE	3T	8	19/19	6k
Stanford 3D (Epperson, 2013)	knee	coronal	PDFS	GE	3T	8	19/19	6k
Stanford 3D (Epperson, 2013)	knee	sagittal	PDFS	GE	3T	8	19/19	4.8k
7T database, 3D (Caan, 2022)	brain	axial	MP2RAGE-ME	Philips	7T	32	385/77	112k
7T database, 3D (Caan, 2022)	brain	coronal	MP2RAGE-ME	Philips	7T	32	385/77	112k
7T database, 3D (Caan, 2022)	brain	sagittal	MP2RAGE-ME	Philips	7T	32	385/77	91k
CC-359, 3D (Souza et al., 2018)	brain	axial	GRE	GE	3T	12	67/67	17k
CC-359, 3D (Souza et al., 2018)	brain	coronal	GRE	GE	3T	12	67/67	14k
CC-359, 3D (Souza et al., 2018)	brain	sagittal	GRE	GE	3T	12	67/67	11k
Stanford 2D (Cheng, 2018)	various	various	various	GE	3T	3-32	89/89	2k
NYU data (Hammernik et al., 2018)	knee	various	PD (40%), PDFS (20%), T2FS(40%)	Siemens	3T	15	100/20	3.5k
M4Raw GRE (Lyu et al., 2023)	brain	axial	GRE	XGY	0.3T	4	366/183	6.6k

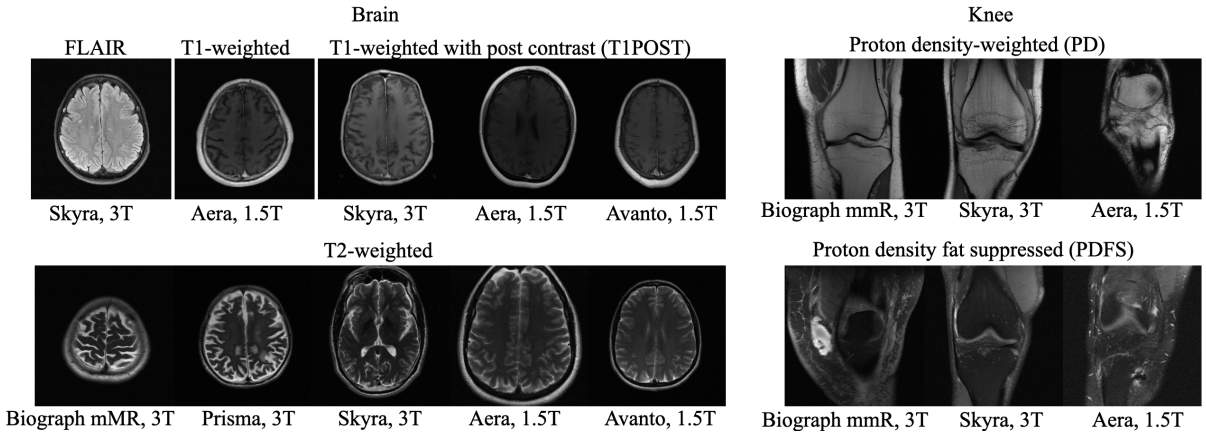


Figure 2: Example images for a selection of distributions from the fastMRI dataset (Zbontar et al., 2019) we consider here. Axial view brain images are on the left, coronal view knee images are on the right. The caption above an image describes the image contrast, and the caption below is the name of the MRI scanner used.

Here, $\mathbf{S}_i$ is the sensitivity map of the i-th coil, $\mathbf{F}$ is the 2D discrete Fourier transform, $\mathbf{M}$ is an undersampling mask, and $\mathbf{z}_i$ models additive white Gaussian noise. The measurements $\mathbf{y}_i$ are often called k-space measurements.

In this work, we consider 4-fold accelerated (i.e., $m = N/4$) multi-coil 2D MRI reconstruction with Cartesian undersampling. The central k-space region is fully sampled including 8% of all k-space lines, and the remaining lines are sampled equidistantly with a random offset from the start. We choose 4-fold acceleration as going beyond 4-fold acceleration, radiologists tend to reject the reconstructions by neural networks and other methods as not sufficiently good (Muckley et al., 2021; Radmanesh et al., 2022). Equidistant sampling is chosen due to the ease of implementation on existing machines (Zbontar et al., 2019).

Class of reconstruction methods. We focus on deep learning models trained end-to-end for accelerated MRI because they give state-of-the-art performance in accuracy and speed (Hammernik et al., 2018; Aggarwal et al., 2019; Sriram et al., 2020; Fabian et al., 2022). There are different methods to image reconstruction with neural networks including un-trained neural networks (Ulyanov et al., 2020; Heckel & Hand, 2019; Darestani & Heckel, 2021) and methods based on generative neural networks (Bora et al., 2017; Jalal et al., 2021; Zach et al., 2023).

A neural network $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ mapping measurements $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}$ to an image is commonly trained to reconstruct an image from the measurements $\mathbf{y}$ by minimizing the supervised loss $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \text{loss}(f_{\boldsymbol{\theta}}(\mathbf{y}_i), \mathbf{x}_i)$ over a training set consisting of target images and cor-

responding measurements $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$. This dataset is typically constructed from fully-sampled k-space data (i.e., where the undersampling mask M is identity). From the fully-sampled data, a target image $\mathbf{x}$ is estimated, and retrospectively undersampled measurements $\mathbf{y}$ are generated by applying the undersampling mask to the fully-sampled data.

Several choices of network architectures work well. A standard baseline is a U-net (Ronneberger et al., 2015) trained to reconstruct the image from a coarse least-squares reconstruction of the measurements (Zbontar et al., 2019). A vision transformer (Dosovitskiy et al., 2021) for image reconstruction applied in the same fashion as the U-net also works well (Lin & Heckel, 2022). The best-performing models are unrolled networks such as the variational network (Hammernik et al., 2018) and a deep cascade of convolutional neural networks (Schlemper et al., 2018). The unrolled networks often use either the U-net as backbone, like the end-to-end VarNet (Sriram et al., 2020), or a transformer based architecture (Fabian et al., 2022).

We expect our results in this paper to be model agnostic, and show that this is indeed the case for the U-net, ViT, and end-to-end VarNet.

Datasets. We consider the fully-sampled MRI dataset with varying attributes listed in Table 1. The datasets include the largest publicly available fully-sampled MRI datasets, and contain altogether around **500k slices**.

Many of our experiments are based on splits of the fastMRI dataset (Zbontar et al., 2019), the most commonly used dataset for MRI reconstruction research. Figure 2 depicts samples from the fastMRI dataset and shows that MRI data vary in appearance across different anatomies and image contrasts (T1, T2, etc.). The image distribution also varies across vendors and magnetic field strengths of scanners, as the strength of the magnet impacts the signal-to-noise ratio (SNR), with stronger magnets leading to higher SNRs.

The fastMRI dataset stands out for its diversity and size, making it particularly well-suited for exploring how different data distributions can affect the performance of deep learning models for accelerated MRI. In our experiments in Section 3, 4, 5, and 6 we split the fastMRI dataset according to different attributes of the data. In Section 7, we showcase the generalizability of our findings on a diverse collection of 17 different datasets.

3. Training a Single Model or Separate Models on Different Distributions

We start with studying whether training a model on data from a diverse set of distributions compromises the performance on the individual distributions. In its simplest instance, the question is whether a model for image reconstruction trained on data from both distributions P and Q performs as well on distributions P and Q as a model trained on P and applied on P and a model trained on Q and applied on Q.

In general, this depends on the distributions P and Q, and on the estimator. For example, consider a simple toy denoising problem, where the data from distribution P is generated as $\mathbf{y} = \mathbf{x} + \mathbf{e}$, with $\mathbf{x}$ is drawn i.i.d. from the unit sphere of a subspace, and $\mathbf{e}$ is drawn from a zero-mean Gaussian with co-variance matrix $\sigma_P \mathbf{I}$. Data for distribution Q is generated equally, but the noise is drawn from a zero-mean distribution with different noise variance, i.e., $\mathbf{e} \sim \mathcal{N}(0, \sigma_Q^2 \mathbf{I})$ with $\sigma_P^2 \neq \sigma_Q^2$. Then the optimal linear estimator learned from data drawn from both distribution P and Q is sub-optimal for both distributions P and Q. However, there exists a non-linear estimator that is as good as the optimal linear estimator on distribution P and distribution Q.

In addition, conventional approaches to MRI such as ℓ_1-regularized least-squares need to be tuned individually on different distributions to achieve best performance, as discussed in Appendix A.

Thus it is unclear whether it is preferable to train a neural network for MRI on diverse data from many distributions or to train several networks and use them for each individual distribution. For example, is it better to train a network specific for knees and another one for brains or to train a single network on knees and brains together? Here, we find that training a network on several distributions simultaneously does not compromise performance on the individual distribution relative to training one model for each distribution.

Experiments for training a joint or separate models. We consider two distributions P and Q, and train U-nets (Ronneberger et al., 2015), ViTs (Dosovitskiy et al., 2021) and end-to-end VarNets (Sriram et al., 2020) on data $\mathcal{D}_P$ from distributions P and on data D_Q from distribution Q separately. We also train the same models on data from P and Q, i.e., $\mathcal{D}_P \cup \mathcal{D}_Q$. We then evaluate on separate test sets from distribution P and Q. We consider the end-to-end VarNet because it is a state-of-the-art model for accelerated MRI, and consider the U-net and ViT as popular baseline models. This diverse selection of architectures (unrolled, convolutional, transformer) aims to demonstrate that our qualitative results are independent of the specific architectural choice. We consider the following choices for $\mathcal{D}_P$ and $\mathcal{D}_Q$, which are subsets of the fastMRI dataset specified in Figure 2:

- **Anatomies.** P are knees scans collected with 6 different combinations of image contrasts and scanners and Q are the brain scans collected with 10 different combinations of image contrasts and scanners.

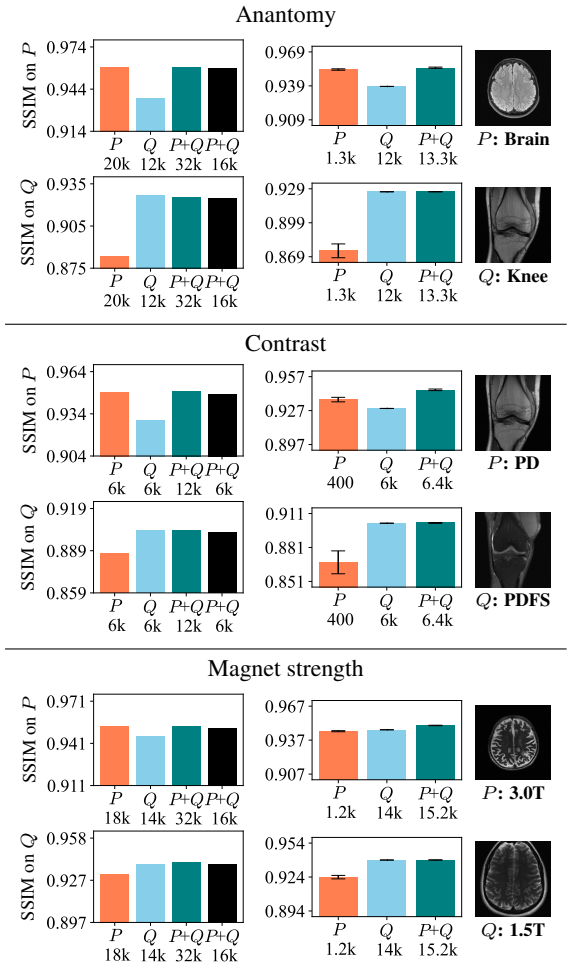


Figure 3: The **orange** and **blue** bars are models trained only on data from P ($\mathcal{D}_P$) and Q ($\mathcal{D}_Q$), respectively, and the **teal** bars are the models trained on both sets $\mathcal{D}_P \cup \mathcal{D}_Q$. As a reference point, the **black** bars are the performance of models trained on random samples of $\mathcal{D}_P \cup \mathcal{D}_Q$ of **half the size**. Below each bar is the number of training images. **Left:** High-data regime where increasing the dataset further gives only minor improvements. For all distributions, the joint model trained on P and Q performs as well on P and Q as the models trained individually for each of those distributions. **Right:** Results for skewed datasets. We report the mean $\pm$ two standard deviations of five runs with different training data from P and model initialization. We note for training sets exceeding 3k images, there is next to no variation (see Figure 12), therefore we only have error bars for this experiment which includes training runs on small datasets.

- **Contrasts.** We select P as PD-weighted knee images from 3 different scanners and Q are PDFS-weighted knee images from the same 3 scanners.
- **Magnetic field.** Here, we pick P to contain all 3.0T scanners and Q to contain all 1.5T scanners regardless of anatomy or image contrast.

Figure 3 (left) shows for U-net that the models trained on both P and Q achieve essentially the same performance on both P and Q as the individual models. The model trained on both $P + Q$ uses more examples than the model trained on P and Q individually. To rule out the possibility that the joint model is only as good as the individual models since it is trained on more examples, we also trained a model on $P + Q$ with half the number of examples (obtained by randomly subsampling). Again, the model performs essentially equally well as the other models. We refer to Appendix B.1 and B.2 for details regarding the setup.

Results for VarNet and ViT are qualitatively the same as the results in Figure 3 for U-net (see Appendix B.3), and indicate that our findings are architecture-independent.

Thus, separating datasets into data from individual distributions and training individual models does not yield benefits, unlike for ℓ_1-regularized least squares or the toy-subspace example.

Experiments for training a joint or separate models on skewed data. Next, we consider skewed data, i.e., the training set $\mathcal{D}_P$ is by a factor of about 10 smaller than the training set $\mathcal{D}_Q$. The choices for distributions P and Q are as in the previous experiment. Figure 3 (right) shows that even for data skewed by a factor 10, the performance on distributions P and Q of models (U-net) trained on both distributions is comparable to the models trained on the individual distributions.

4. Data Diversity Enhances Robustness Towards Distribution-Shifts

We now study how training on diverse data affects the out-of-distribution performance of a model. Nguyen et al. (2022) note that for image recognition two outcomes can be expected from training a model jointly on two distributions P_1 and P_2, and evaluating on another distribution Q. Assuming that training a model on P_1 gives better performance on Q than training a model on P_2, the model trained on P_1 and P_2 could perform (i) at least as well as a model trained on P_1 or (ii) performs worse than the model trained on P_1 but better than a model trained on P_2. Here, we find that for accelerated MRI, training a model on diverse data improves a model's out-of-distribution performance, i.e., case (i).

Measuring robustness. Our goal is to measure the expected robustness gain by training models on diverse data, and we would like this measure to be independent of the model itself. Hence, we measure robustness with the notion of 'effective robustness' by Taori et al. (2020). We evalu-

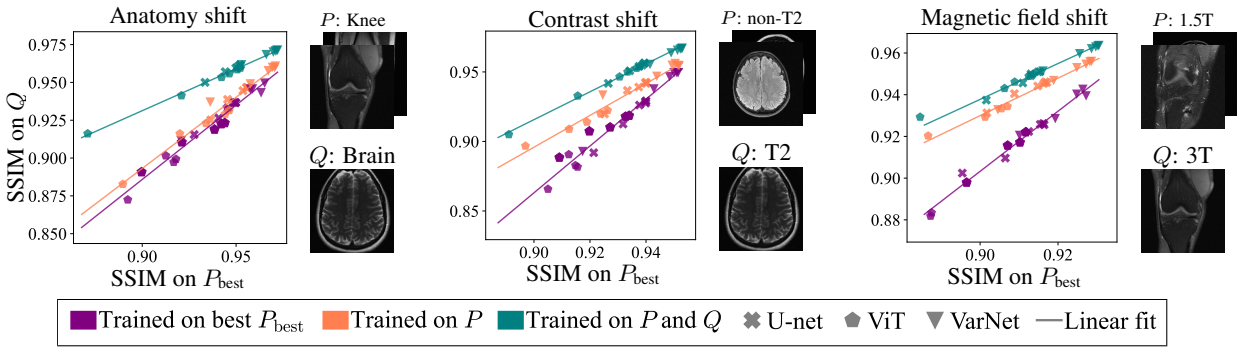


Figure 4: For a distribution-shift from distributions $P = \{P_1, \ldots, P_m\}$ to distribution Q, we compare robustness of models trained on P (**orange**) to baselines trained on the best single distribution P_{best} (**violet**). As additional reference, we also report models trained on both P and Q to imitate ideally robust models (**teal**). For the three distribution-shifts shown, training on the more diverse dataset P is beneficial compared to training on P_{best} alone.

ate models on a standard in-distribution test set (i.e., data from the same source that generated the training data) and on an out-of-distribution test set. We then plot the out-of-distribution performance for different models as a function of the in-distribution performance, see Figure 4.

It can be seen that the in- and out-of-distribution performance of models trained on data from one distribution, (e.g., in-distribution data **violet**) is well described by a linear fit. If a robustness intervention only moves a model along the line, it doesn't increase out-of-distribution performance beyond what's expected for a given in-distribution performance and therefore does not yield effective robustness. Thus, a dataset yields more effective robustness if models trained on it lie above the **violet** line, since such models have higher out-of-distribution performance than what's expected for a fixed in-distribution performance.

Experiment. We are given data from two distributions P and Q, where distribution P can be split up into m sub-distributions $P_1, \ldots, P_m$. We consider the following choices for the two distributions, all based on the knee and brain fastMRI datasets illustrated in Figure 2:

- **Anatomy shift:** $P_1, \ldots, P_6$ is knee data collected with all 6 different combinations of image contrasts and scanners, and Q are the different brain datasets collected with 8 different combinations of image contrasts (FLAIR, T1, T1POST, T2) and scanners (Skyra, Prisma, Aera, Biograph mMR).
- **Contrast shift:** $P_1, \ldots, P_5$ are all FLAIR, T1POST, or T1 brain images and Q are T2 brain data.
- **Magnetic field shift:** $P_1, \ldots, P_7$ are brain and knee data collected with 1.5T scanners (Aera, Avanto) and Q are brain and knee data collected with 3T scanners (Skyra, Prisma, Biograph mMR).

For each of the distributions $P_1, \ldots, P_m$ we construct a training set with 2048 images and a test set with 128 images.

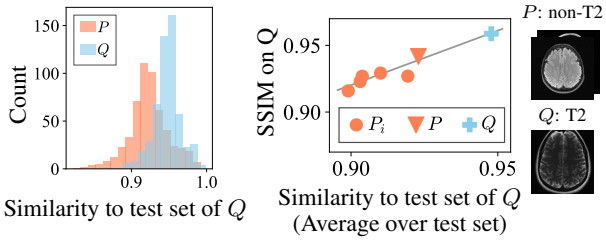


Figure 5: We compute the cosine similarity in CLIP feature-space between a test sample and the nearest neighbor from the training set and relate this similarity to model performance. **Left:** Histograms of nearest neighbor similarity for each test sample of Q to the training set from distributions P or Q. **Right:** Strong correlation between nearest neighbor similarity and performance. Compared to datasets from distributions P_i, a more diverse dataset from distribution $P = \{P_1, \ldots, P_m\}$ increases both similarity to the out-of-distribution test set and model (U-net) performance.

We then train U-nets on each of the distributions $P_1, \ldots, P_m$ separately and select from these distributions the distribution P_{best} that maximizes the performance of the U-net on a test set from the distribution Q.

Now, we train a variety of different model architectures including the U-net, end-to-end VarNet (Sriram et al., 2020), and vision transformer (ViT) for image reconstruction (Lin & Heckel, 2022) on data from the distribution P_{best}, data from the distribution P (which contains P_{best}), and data from the distribution P and Q. We also sample different models by early stopping and by decreasing the training set size by four. We plot the performance of the models evaluated on the distribution Q as a function of their performance evaluated on the distribution P_{best}. The configurations of

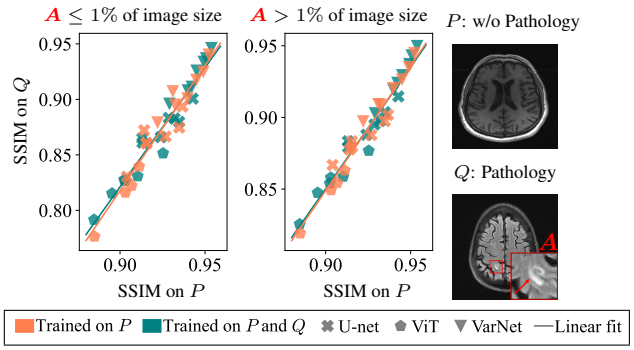


Figure 6: Models trained on images without pathologies can reconstruct pathologies as well as models trained on images with pathologies. SSIM is calculated for the pathology region (A) for small (**left**) and large (**right**) pathologies.

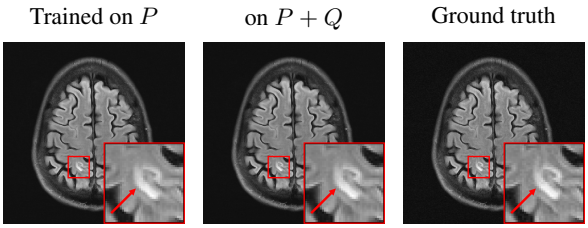


Figure 7: Reconstructions by a VarNet trained on images without pathologies (P) and with pathologies ($P + Q$) of an image containing a small-sized pathology.

our models are the same as in Section 3.

Figure 4 shows that the models **trained on P** are outperformed by models **trained on P and Q** when evaluated on Q, as expected, since a model **trained on P and Q** is an ideal robust baseline (as it contains data from Q). The difference of the **trained on P and Q**-line and the **trained on P_{best}**-line is a measure of the severity of the distribution-shift, as it indicates the loss in performance when a model trained on P_{best} is evaluated on Q. Comparing the difference between the line for the models **trained on P** and the line for models **trained on P_{best}** shows that effective robustness is improved by training on a diverse dataset, even when compared to distribution P_{best} which is the most beneficial distribution for performance on Q. Moreover, we find similar results for distribution-shifts in the acceleration factor and in the number of receiver coils as detailed in Appendix C.

Robustness improvements are related to increasing similarity of train and test distributions. A plausible explanation for the observed performance gains through training on a more diverse dataset is that a diverse dataset increases the similarity between samples in the training and target test set, and thus the target test set is less out-of-distribution. To test this hypothesis, we use CLIP-similarity, similar to how Mayilvahanan et al. (2024) used CLIP-similarity to study CLIP's robustness in image classification.

Specifically, for each test sample we find the nearest neighbor to a training sample in terms of the cosine similarity of CLIP features, and then take the mean of the histogram of similarities as the similarity of the test set to a training set. The left plot in Figure 5, illustrates this measure for a contrast shift. The histograms for P and Q show the distribution of the similarity-scores between the test set and a training set. As expected, the training set from the same distribution (Q) shows higher similarity than a training set from a different distribution (P). The right plot extends this analysis to sub-distributions P_i within P, showing that a more diverse training set P is more similar to the test set, and most importantly this metric correlates well with reconstruction performance. More details are in Appendix D.

5. Reconstruction of Pathologies Using Data From Healthy Subjects

In this section, we investigate the distribution-shift from healthy to non-healthy subjects by measuring how well models reconstruct images containing a pathology if no pathologies are contained in the training set. We find that models trained on fastMRI data without pathologies reconstruct fastMRI data with pathologies as accurately as the same models trained on fastMRI data with pathologies.

Experiment. We rely on the fastMRI+ annotations (Zhao et al., 2022) to partition the fastMRI brain dataset into sets of images with and without pathologies. The annotations cover various pathologies in the fastMRI dataset. We extract a set of volumes without pathologies by selecting all scans with the fastMRI+ label "Normal for age", and select images with pathologies by taking all images with slice-level annotations of a pathology. The training set contains 4.5k images without pathologies (P) and 2.5k images with pathologies (Q). We train U-nets, ViTs, and VarNets on P and on $P+Q$, and sample different models by varying the training set size by factors of 2, 4 and 8, and by early stopping. While the training set from distribution P does not contain images with pathologies, P is a diverse distribution containing data from different scanners and image contrasts.

Figure 6 shows the models' performance on Q relative to their performance on P. Reconstructions are evaluated only on the region containing the pathology, where we distinguish between small pathologies ($\leq 1\%$ of the total image size) and large pathologies ($> 1\%$ of the total image size) to see potential dependencies on the size of the pathology.

We see that the models **trained on P** show essentially the same performance (SSIM) as models **trained on $P + Q$** regardless of pathology size. The results indicate that mod-

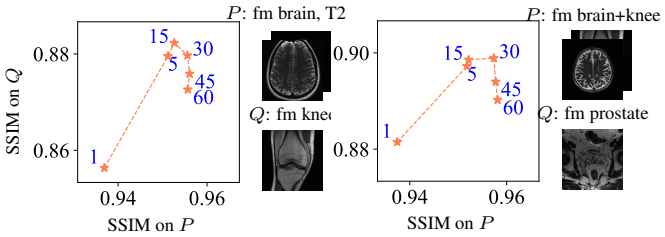


Figure 8: For a distribution-shift from distribution P to distribution Q, the in- and out-of-distribution performance is plotted as a function of training epochs (1 to 60). At the beginning of training, out-of-distribution performance increases together with in-distribution performance. Then, out-of-distribution performance starts to drop while in-distribution performance continues to marginally increase.

els trained on images without pathologies can reconstruct pathologies as accurately as models trained on images with pathologies. This is further illustrated in Figure 7, where we show reconstructions given by the VarNet of images with a pathology: The model recovers the pathology well even though no pathologies are in the training set. We provide additional results, reconstruction examples, and discussion in Appendix E, including a more nuanced evaluation of the SSIM values for VarNet (Figure 14).

6. Distributional Overfitting

We observed that when training for long, while in-distribution performance continues to improve slightly, out-of-distribution performance can deteriorate. We refer to this as distributional overfitting. Unlike conventional overfitting, where a model's in-distribution performance declines after prolonged training, distributional overfitting involves a decline in out-of-distribution performance while in-distribution performance continues to improve (slightly). A similar observation has been made in the context of fine-tuning CLIP models (Wortsman et al., 2022).

Figure 8 illustrates distributional overfitting on two distribution-shifts. Each plot depicts the in and out-of-distribution (P and Q) performance of an U-net as a function of trained epochs. For example, in the left plot P is fastMRI T2-weighted brain data and Q is fastMRI knee data. We observe as training progresses, initially, the model's in-distribution and out-of-distribution performance both improve. However, after epoch 15, out-of-distribution performance deteriorates, despite marginal improvements in in-distribution performance. In Appendix F, we show that distributional overfitting also occurs for VarNet and ViT, and when using other optimizers.

This finding indicates that early stopping before conventional overfitting sets in, can help to improve model robustness with minimal impact on in-distribution performance.

7. Robust Models for Accelerated MRI

The results from the previous sections based on the fastMRI dataset suggest that training a single model on a diverse dataset consisting of several data distributions is beneficial to out-of-distribution performance without sacrificing in-distribution performance on individual distributions.

We now move beyond the fastMRI dataset and demonstrate that this finding continues to hold on a large collection of datasets, enabling significant out-of-distribution performance improvements. We train a single large model for 4-fold accelerated 2D MRI on a **diverse collection of 13 datasets** including the fastMRI brain and knee datasets, and evaluate on 4 out-of-distribution datasets (the descriptions of the sets are in Table 1). The resulting model, when compared to models trained only on the fastMRI dataset, shows significant robustness improvements while maintaining its performance on the fastMRI dataset.

Experiment. We train an U-net, ViT, and an end-to-end VarNet on the collection of the first 13 datasets listed in Table 1. We denote this collection of datasets by $\mathcal{D}_P$. For the fastMRI knee and brain datasets, we exclude the fastMRI knee validation set and the fastMRI brain test set from the training set, as we use those for testing. The total number of training slices after the data preparation is 413k. For each model family, we also train a model on fastMRI knee, and one on fastMRI brain as baselines. To mitigate the risk of distributional overfitting, we early stop training when the improvement on the fastMRI knee dataset becomes marginal. Further experimental details are in Appendix G.1 and G.2.

The fastMRI knee validation and fastMRI brain test set are used to measure in-distribution performance. We measure out-of-distribution performance on CC-359 sagittal view (Souza et al., 2018), Stanford 2D (Cheng, 2018), M4Raw GRE (Lyu et al., 2023), and NYU data (Hammernik et al., 2018). These datasets constitute a distribution-shift relative to the training data with respect to vendors, anatomic views, anatomies, time-frame of data collection, anatomical views, MRI sequences, contrasts and combinations thereof and therefore enable a broad robustness evaluation. As a further reference point we also train models on the out-of-distribution datasets to quantify the robustness gap.

Figure 9 shows for all architectures considered, the model trained on the collection of datasets $\mathcal{D}_P$ significantly outperforms the models trained on fastMRI data when evaluated on out-of-distribution data, without compromising performance on fastMRI data. For example, on the CC-359 sagittal view dataset, the VarNet trained on $\mathcal{D}_P$ almost closes the distribution-shift performance gap (i.e., the gap to the

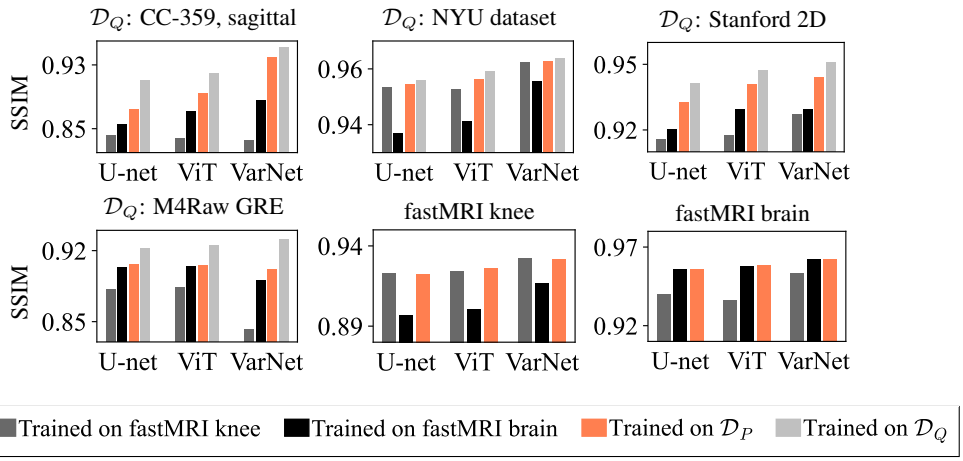


Figure 9: Training on a diverse collection of datasets improves robustness under distribution-shifts. A model trained on the diverse set of datasets $\mathcal{D}_P$ can significantly outperform models trained on fastMRI data when evaluated on out-of-distribution data $\mathcal{D}_Q$, while maintaining the same performance on fastMRI data.

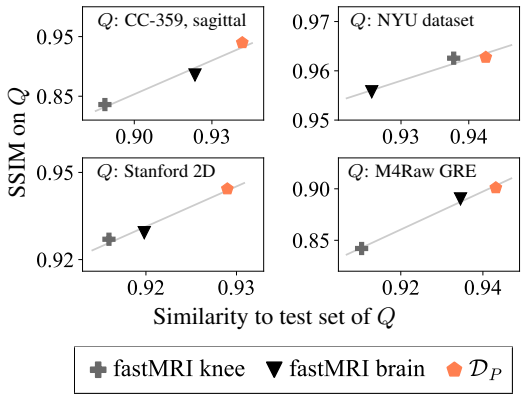


Figure 10: Nearest-neighbor CLIP similarity between a training set and out-of-distribution (Q) test set correlate well with the performance of VarNet.

model trained on the out-of-distribution data). We refer to Figure 1 and Figure 20 (in the Appendix) for reconstruction examples.

Robustness improvements are related to increasing similarity of train and test distributions. In Figure 10, we compute the CLIP similarity between the training sets and the out-of-distribution test sets and observe a strong correlation between similarity and performance. This supports the idea that diverse datasets enhance similarity to out-of-distribution data, leading to improved performance.

We further validate our findings in Appendix G.3, where we evaluate model performance using the deep feature metrics LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022), because these metrics were found to align well with radiologist evaluations (Adamson et al., 2023; Kastryulin et al., 2023). The qualitative results are consistent with the SSIM results in this section: As shown in Table 3, the models trained on the large collection of datasets yield better scores on the out-of-distribution datasets than the fastMRI baselines while achieving similar scores on the fastMRI datasets. In addition, we provide an error analysis by quantifying the prominence of artifacts in the reconstructions (Table 4) and find that models trained on the diverse dataset $\mathcal{D}_P$ produce less pronounced artifacts compared to the fastMRI baseline.

The results in this section reinforce our earlier findings, affirming that large and diverse MRI training sets can significantly enhance robustness without compromising in-distribution performance.

8. Conclusion and Limitations

While our research shows that diverse training sets significantly enhances out-of-distribution robustness for deep learning models for accelerated 2D MRI, training a model on a diverse dataset often doesn't close the distribution-shift performance gap, i.e., the gap between the model and the same idealized model trained on the out-of-distribution data (see Figure 4 and 9). Nevertheless, as datasets grow in size and diversity, training networks on larger and even more diverse data might progressively narrow the distribution-shift performance gap (as suggested in Figure 5 and 10). However, in practice it might be difficult or expensive to collect diverse and large datasets.

Besides demonstrating the effect of diverse training data, our work shows that care must be taken when training models for long as this can yield to a less robust model due to distributional overfitting. This finding also emphasizes the importance of evaluating on out-of-distribution data.

Acknowledgements

We thank Stefan Ruschke for helpful discussions.

This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy as part of the project "6G Future Lab Bavaria". This work also received support by the Institute of Advanced Studies at the Technical University of Munich and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 456465471, 464123524.

Impact Statement

The findings of this work highlight the importance of dataset design and in particular data diversity for deep learning based accelerated MRI. As this work encourages the utilization of large and diverse datasets for training robust deep learning models, ethical considerations around data privacy, consent, and security become crucial. While this work relies on public datasets intended for research purposes, practitioners should prioritize the ethical handling of patient data and ensure compliance with data protection regulations. Our code is available at https://github.com/MLI-lab/mri_data_diversity.

References

Adamson, P. M., Desai, A. D., Dominic, J., Bluethgen, C., Wood, J. P., Syed, A. B., Boutin, R. D., Stevens, K. J., Vasanawala, S., Pauly, J. M., Chaudhari, A. S., and Gunel, B. Using Deep Feature Distances for Evaluating MR Image Reconstruction Quality. In *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023.

Aggarwal, H. K., Mani, M. P., and Jacob, M. MoDL: Model-Based Deep Learning Architecture for Inverse Problems. *IEEE Transactions on Medical Imaging*, 2019.

Antun, V., Renna, F., Poon, C., Adcock, B., and Hansen, A. C. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 2020.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed Sensing using Generative Models. In *International Conference on Machine Learning*, 2017.

Caan, M. Quantitative motion-corrected 7T sub-millimeter raw MRI database of the adult lifespan. https://doi.org/10.34894/IHZGQM, 2022.

Cheng, J. Y. Stanford 2D FSE. http://mridata.org/list?project=Stanford 2D FSE, 2018.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Dar, S. U. H., Özbey, M., Çatlı, A. B., and Çukur, T. A Transfer-Learning Approach for Accelerated MRI Using Deep Neural Networks. *Magnetic Resonance in Medicine*, 2020.

Darestani, M. Z. and Heckel, R. Accelerated MRI With Un-trained Neural Networks. *IEEE Transactions on Computational Imaging*, 2021.

Darestani, M. Z., Chaudhari, A. S., and Heckel, R. Measuring Robustness in Deep Learning Based Compressive Sensing. In *International Conference on Machine Learning*, 2021.

Darestani, M. Z., Liu, J., and Heckel, R. Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing. In *International Conference on Machine Learning*, 2022.

Desai, A. D., Schmidt, A. M., Rubin, E. B., Sandino, C. M., Black, M. S., Mazzoli, V., Stevens, K. J., Boutin, R., Re, C., Gold, G. E., Hargreaves, B., and Chaudhari, A. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Dong, C., Loy, C. C., He, K., and Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In *European Conference on Computer Vision*, 2014.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Ducotterd, S., Goujon, A., Bohra, P., Perdios, D., Neumayer, S., and Unser, M. Improving Lipschitz-Constrained Neural Networks by Learning Activation Functions. arXiv:2210.16222, 2022.

Epperson, K. Creation of fully sampled mr data repository for compressed sensing of the knee. In *SMRT Annual Meeting*, 2013.

Fabian, Z., Tinaz, B., and Soltanolkotabi, M. HUMUS-Net: Hybrid Unrolled Multi-scale Network Architecture for

Accelerated MRI Reconstruction. In *Advances in Neural Information Processing Systems*, 2022.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). In *International Conference on Machine Learning*, 2022.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P. W. W., Saukh, O., Ratner, A. J., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., and Schmidt, L. DataComp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 2023.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 2002.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., and Knoll, F. Learning a Variational Network for Reconstruction of Accelerated MRI Data. *Magnetic resonance in medicine*, 2018.

Heckel, R. and Hand, P. Deep Decoder: Concise Image Representations from Untrained Non-convolutional Networks. In *International Conference on Learning Representations*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *IEEE International Conference on Computer Vision*, 2021.

Huang, J., Wang, S., Zhou, G., Hu, W., and Yu, G. Evaluation on the generalization of a learned convolutional neural network for MRI reconstruction. *Magnetic Resonance Imaging*, 2022.

Ishida, M., Doi, K., Loo, L. N., Metz, C. E., and Lehr, J. L. Digital image processing: Effect on detectability of simulated low-contrast radiographic patterns. *Radiology*, 1984.

Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A. G., and Tamir, J. Robust Compressed Sensing MRI with Deep Generative Priors. In *Advances in Neural Information Processing Systems*, 2021.

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 2017.

Johnson, P. M., Jeong, G., Hammernik, K., Schlemper, J., Qin, C., Duan, J., Rueckert, D., Lee, J., Pezzotti, N., De Weerdt, E., Yousefi, S., Elmahdy, M. S., Van Gemert, J. H. F., Schülke, C., Doneva, M., Nielsen, T., Kastryulin, S., Lelieveldt, B. P. F., Van Osch, M. J. P., Staring, M., Chen, E. Z., Wang, P., Chen, X., Chen, T., Patel, V. M., Sun, S., Shin, H., Jun, Y., Eo, T., Kim, S., Kim, T., Hwang, D., Putzky, P., Karkalousos, D., Teuwen, J., Miriakov, N., Bakker, B., Caan, M., Welling, M., Muckley, M. J., and Knoll, F. Evaluation of the Robustness of Learned MR Image Reconstruction to Systematic Deviations Between Training and Test Data for the Models from the fastMRI Challenge. In *Machine Learning for Medical Image Reconstruction*, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models. arXiv:2001.08361, 2020.

Kastryulin, S., Zakirov, J., Pezzotti, N., and Dylov, D. V. Image Quality Assessment for Magnetic Resonance Imaging. *IEEE Access*, 2023.

Khun Jush, F., Biele, M., Dueppenbecker, P. M., and Maier, A. Deep Learning for Ultrasound Speed-of-Sound Reconstruction: Impacts of Training Data Diversity on Stability and Robustness. *Machine Learning for Biomedical Imaging*, 2023.

Klug, T. and Heckel, R. Scaling Laws For Deep Learning Based Image Reconstruction. In *International Conference on Learning Representations*, 2023.

Knoll, F., Hammernik, K., Kobler, E., Pock, T., Recht, M. P., and Sodickson, D. K. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic Resonance in Medicine*, 2019.

Knoll, F., Murrell, T., Sriram, A., Yakubova, N., Zbontar, J., Rabbat, M., Defazio, A., Muckley, M. J., Sodickson, D. K., Zitnick, C. L., and Recht, M. P. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. *Magnetic Resonance in Medicine*, 2020.

Krainovic, A., Soltanolkotabi, M., and Heckel, R. Learning Provably Robust Estimators for Inverse Problems via Jittering. In *Neural Information Processing Systems*, 2023.

Lin, K. and Heckel, R. Vision Transformers Enable Fast and Robust Accelerated MRI. In *International Conference on Medical Imaging with Deep Learning*, 2022.

Liu, X., Wang, J., Liu, F., and Zhou, S. K. Universal Undersampled MRI Reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, 2021.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 2007.

Lyu, M., Mei, L., Huang, S., Liu, S., Li, Y., Yang, K., Liu, Y., Dong, Y., Dong, L., and Wu, E. X. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 2023.

Mayilvahanan, P., Wiedemer, T., Rusak, E., Bethge, M., and Brendel, W. Does CLIP's generalization performance mainly stem from high train-test similarity? In *International Conference on Learning Representations*, 2024.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The Effect of Natural Distribution Shift on Question Answering Models. In *International Conference on Machine Learning*, 2020.

Muckley, M. J., Riemenschneider, B., Radmanesh, A., Kim, S., Jeong, G., Ko, J., Jun, Y., Shin, H., Hwang, D., Mostapha, M., Arberet, S., Nickel, D., Ramzi, Z., Ciuciu, P., Starck, J.-L., Teuwen, J., Karkalousos, D., Zhang, C., Sriram, A., Huang, Z., Yakubova, N., Lui, Y. W., and Knoll, F. Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 2021.

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP. In *Advances in Neural Information Processing Systems*, 2022.

Ouyang, C., Schlemper, J., Biffi, C., Seegoolam, G., Caballero, J., Price, A. N., Hajnal, J. V., and Rueckert, D. Generalizing Deep Learning MRI Reconstruction across Different Domains. arXiv:1902.10815, 2023.

Pruessmann, K. P., Weiger, M., Scheidegger, M. B., and Boesiger, P. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 1999.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021.

Radmanesh, A., Muckley, M. J., Murrell, T., Lindsey, E., Sriram, A., Knoll, F., Sodickson, D. K., and Lui, Y. W. Exploring the Acceleration Limits of Deep Learning Variational Network-based Two-dimensional Brain MRI. *Radiology. Artificial Intelligence*, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., and Rueckert, D. A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 2018.

Souza, R., Lucena, O., Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rittner, L., Frayne, R., and Lotufo, R. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 2018.

Sriram, A., Zbontar, J., Murrell, T., Defazio, A., Zitnick, C. L., Yakubova, N., Knoll, F., and Johnson, P. End-to-End Variational Networks for Accelerated MRI Reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, 2020.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring Robustness to Natural Distribution Shifts in Image Classification. In *Advances in Neural Information Processing Systems*, 2020.

Tibrewala, R., Dutt, T., Tong, A., Ginocchio, L., Keerthivasan, M. B., Baete, S. H., Chopra, S., Lui, Y. W., Sodickson, D. K., Chandarana, H., and Johnson, P. M. FastMRI Prostate: A Publicly Available, Biparametric MRI Dataset to Advance Machine Learning for Prostate Cancer Imaging. arXiv:2304.09254, 2023.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep Image Prior. *International Journal of Computer Vision*, 2020.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zach, M., Knoll, F., and Pock, T. Stable Deep MRI Reconstruction using Generative Priors. *IEEE Transactions on Medical Imaging*, 2023.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Yakubova, N., Pinkerton, J., Wang, D., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. arXiv:1811.08839, 2019.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Zhao, R., Yaman, B., Zhang, Y., Stewart, R., Dixon, A., Knoll, F., Huang, Z., Lui, Y. W., Hansen, M. S., and Lungren, M. P. fastMRI+, Clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 2022.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., and E, W. Towards Theoretically Understanding Why Sgd Generalizes Better Than Adam in Deep Learning. In *Advances in Neural Information Processing Systems*, 2020.

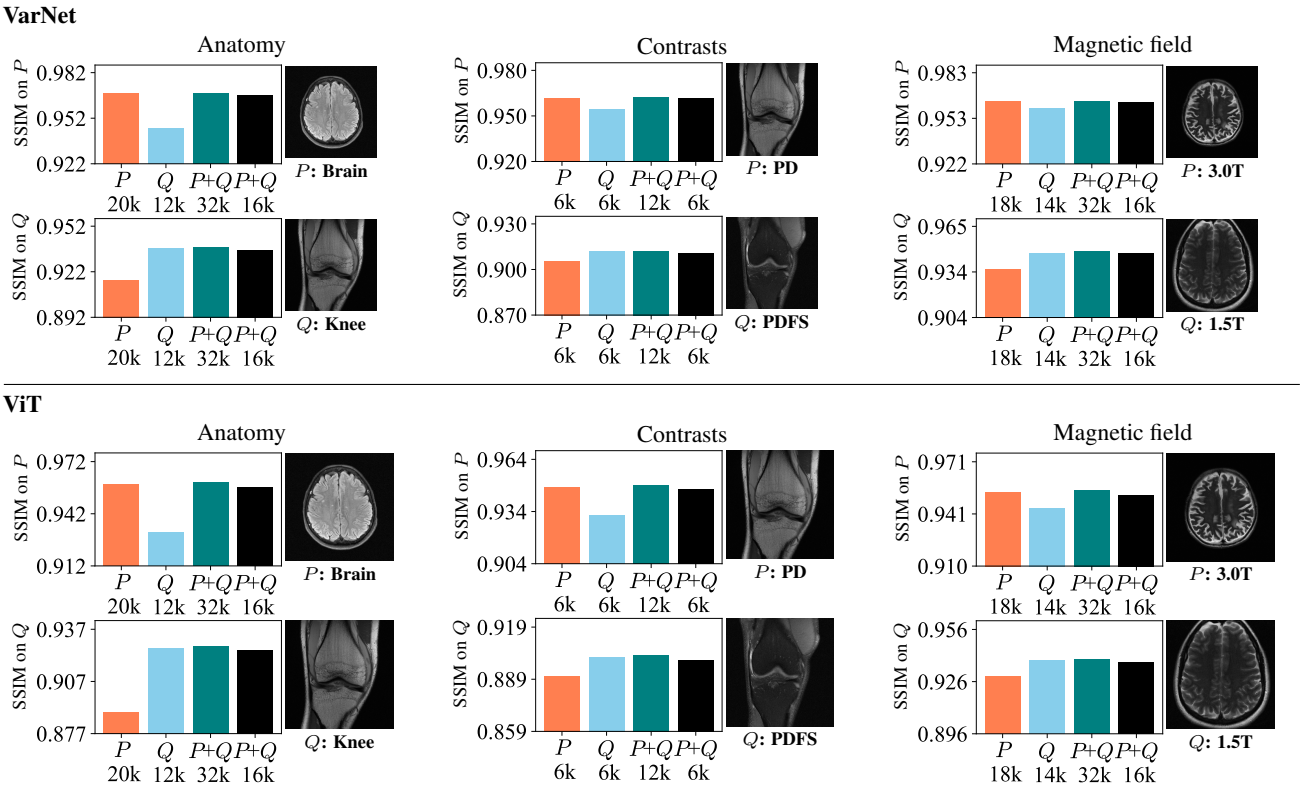


Figure 11: The **orange** and **blue** bars are the VarNet (**Top**) and ViT (**Bottom**) trained exclusively on data from P ($\mathcal{D}_P$) and Q ($\mathcal{D}_Q$), respectively, and the **teal** bars are the models trained on both sets $\mathcal{D}_P \cup \mathcal{D}_Q$. As a reference point, the **black** bars are the performance of models trained on random samples of $\mathcal{D}_P \cup \mathcal{D}_Q$ of **half the size**. The number below each bar is the total number of training images. It can be seen that we are in the high-data regime where increasing the dataset further gives minor improvements. For all distributions, the joint model trained on P and Q performs as well on P and Q as the models trained individually for each of those distributions.

A. ℓ_1-Regularized Least-Squares Requires Different Hyperparameters on Different Distributions

The traditional approach for accelerated MRI is ℓ_1-regularized least-squares (Lustig et al., 2007). While ℓ_1-regularized least-squares is not considered data-driven, the regularization hyperparameter it typically chosen in a data-driven manner. For different distributions like different anatomies or contrasts, the regularization parameter takes on different values and thus the method needs to be tuned separately for different distributions. This can be seen for example from Table 4 of Zbontar et al. (2019).

To demonstrate this, we performed wavelet-based ℓ_1-regularized least-squares on the single-coil knee version of the fastMRI dataset (Zbontar et al., 2019) using 100 images from distribution P: PD Knee Skyra, 3.0T and 100 from distribution Q: PDFS Knee Aera, 1.5T. Using a regularization weight $\lambda = 0.01$ on distribution P gives a SSIM of 0.792, while $\lambda = 0.001$ yields subpar SSIM of 0.788. Contrary, on distribution Q, $\lambda = 0.01$ only yields 0.602, while $\lambda = 0.001$ yields SSIM 0.609. Thus, using the same model (i.e., the same regularization parameter for both distributions) is suboptimal for ℓ_1-regularized. least squares. We used BART `https://mrirecon.github.io/bart/` for running ℓ_1-regularized least squares.

B. Experimental Details and Additional Results for Section 3

B.1. Data Preparation

For each of the distributions in Figure 2, we randomly sample volumes from the fastMRI training set for training so that the total number of slices is around 2048, and we randomly sample from the validation set for testing so that the number of test slices is around 128. Training sets of combination of distributions are then constructed by aggregating the training data from

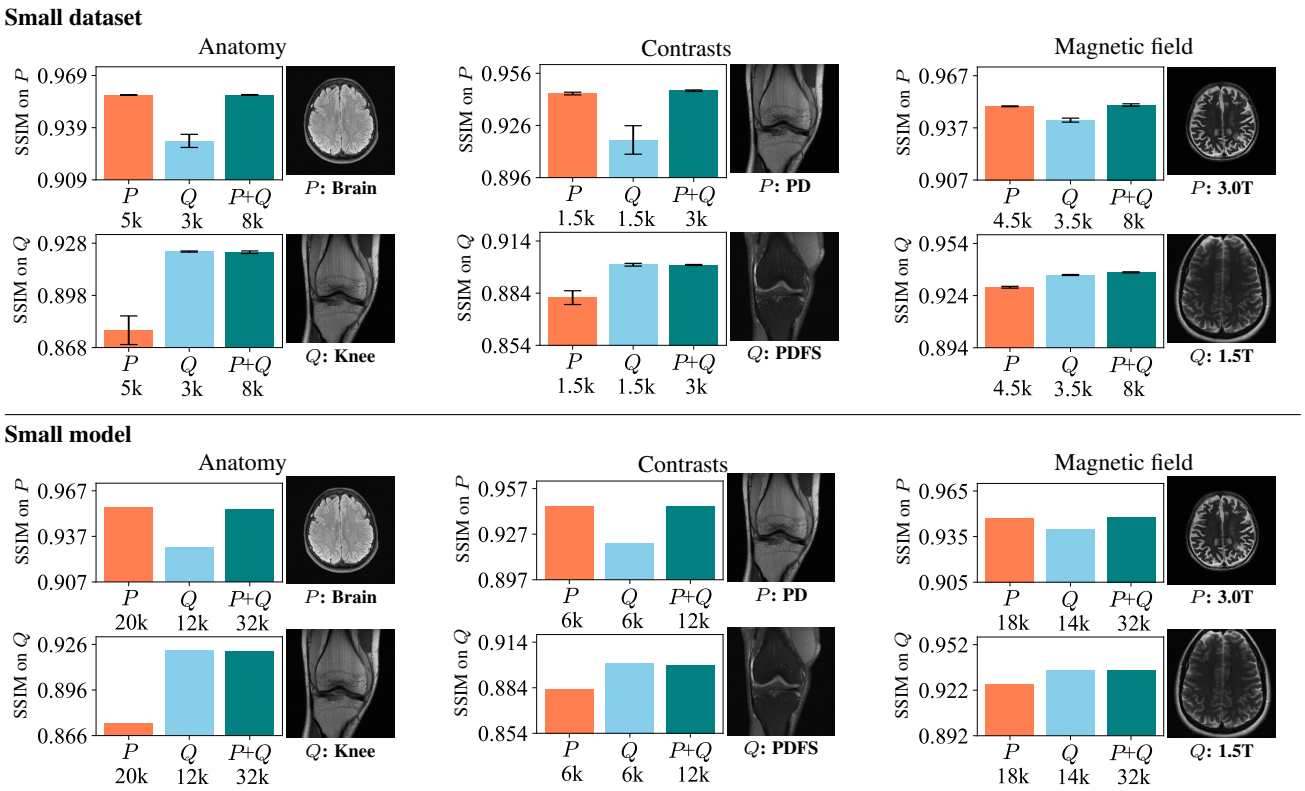


Figure 12: Also for smaller sized datasets (**Top**) or small models (U-net with 757k parameters, **Bottom**), a single model performs at least as good as separate models. We report the mean $\pm$ two standard deviations from five runs with the U-net, each with a different random seed for sampling training data and model initialization. Note that when training on datasets with more than 3k images, there is next to no variation.

the individual distributions. For example, if we consider distribution P to contain all six knee distributions from Figure 2, then the corresponding training set has $6 \cdot 2048$ training images. Likewise, if Q is for example all T2-weighted brain images the corresponding training set has $5 \cdot 2048$ training images.

B.2. Models, Training, and Evaluation

Our configuration of the end-to-end VarNet (Sriram et al., 2020) contains 8 cascades, each containing an U-net with 4 pooling layers and 12 channels in the first pooling layer. The sensitivity-map U-net of the VarNet has 4 pooling layers and 9 channels in the first pooling layer. The code for the model is taken from fastMRI's GitHub repository.

The U-nets used in the experiments have 4 pooling layers and 32 channels in the first pooling layer. The implementation of the model is taken from the fastMRI GitHub repository. Our configuration of the vision transformer (Dosovitskiy et al., 2021) for image reconstruction is the ViT-S configuration from Lin & Heckel (2022), and the code is taken from the paper's GitHub repository. As input data for the U-net and ViT, we first fill missing k-space values with zeros, then apply 2D-IFFT, followed by a root-sum-of-squares (RSS) reconstructions to combine all the coil images into one single image, and lastly normalize it to zero-mean and unit-variance. The mean and variance are added and multiplied back to the model output, respectively. This is a standard prepossessing step, see for example fastMRI's GitHub repository. The models are trained end-to-end with the objective to maximize SSIM between output and ground-truth.

For any model and any choice of distributions P or Q, the models are trained to maximize SSIM between model output and RSS target for a total of 60 epochs and we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The mini-batch size is set to 1. We use linear learning-rate warm-up until a learning-rate of 1e-3 is reached and linearly decay the learning rate to 4e-5. The warm up period amounts to 1% of the total number of gradient steps. Gradients are clipped at a global ℓ_2-norm of 1. During training, we randomly sample a different undersampling mask for each mini-batch independently.

Table 2: A model trained on data with various acceleration factors (2, 4, 8, 16-fold) performs comparable as models trained individually for each acceleration factor. Additionally, the model trained on various acceleration factors can enhance performance for unseen acceleration factors (3-fold). Models were trained and evaluated on the fastMRI PDFS knee subset.

Train \ Test	2-fold	4-fold	8-fold	16-fold	3-fold
2-fold	0.945	—	—	—	0.906
4-fold	—	0.903	—	—	0.899
8-fold	—	—	0.867	—	0.834
16-fold	—	—	—	0.828	0.758
All of above	0.944	0.902	0.866	0.829	0.912
3-fold	—	—	—	—	0.921

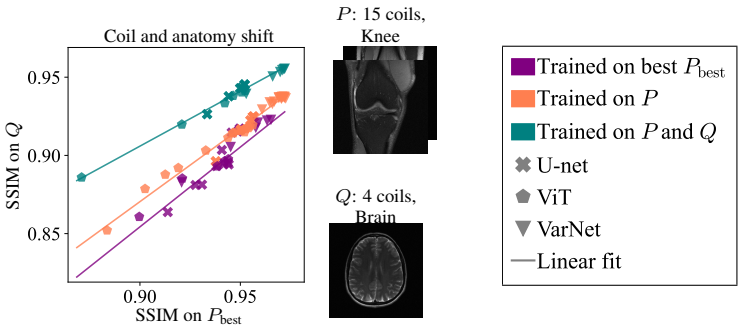


Figure 13: Training on a diverse dataset (P) increases effective robustness under distribution-shifts related to the number of coils.

During evaluation, for each volume we generated an undersampling mask randomly, and this mask is then used for all slices within the volume.

The maximal learning-rate for each model is tuned based on a grid search on the values {1.3e-3, 1e-3, 7e-4, 4e-4} and training on a random subset (2k slices) of the fastMRI dataset. We found negligible differences between learning rates {1.3e-3, 1e-3, 7e-4} and therefore keep the learning rate to 1e-3 for simplicity. We also performed the same grid search on fastMRI subsets for PD-weighted knee and PDFS knee scans and made the same observations.

B.3. Results for VarNet and ViT

In the main body we presented results for the U-net, here we present results for the VarNet and ViT. In Figure 11, we see that, like for the U-net discussed in the main body, for the VarNet and ViT training on two distributions gives the same performance as separate models trained on the individual distributions. Moreover, Figure 12 shows the same experiment for the U-net on smaller datasets and smaller models, where the same observation can be made.

C. Distribution-Shifts Induced by Changes of the Forward Model

In the main body, we considered distribution-shifts primarily related to the images, such as different contrast and anatomies. In this section, we consider two distribution-shifts that are induced by changes of the forward model, i.e., the relation of measurements and object to be imaged. We consider shifts in the acceleration factor and a distribution-shift related to the number of coils.

C.1. Results for Training With Multiple Acceleration Factors

In the main body, we presented reults for a single accelerated factor of 4.

We now train a (U-net) model simultaneously on data with 2-fold, 4-fold, 8-fold, and 16-fold acceleration factors, and analyze how the performance compares to U-net models trained for each acceleration separately. We train only on the

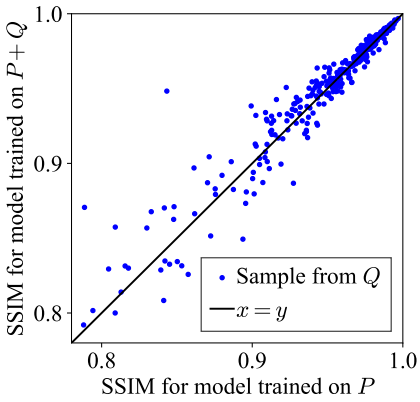


Figure 14: Reconstruction performance for each small pathology reconstructed with a VarNet trained only on data without pathologies (i.e., P) relative to the performance of a VarNet trained on data with and without pathologies (P and Q). SSIM is measured only within the region containing the pathology. The majority of pathologies are reconstructed similarly well by both models, however in the regime where SSIM is low some images are reconstructed better by the P-model and others are reconstructed better by the $P + Q$-model.

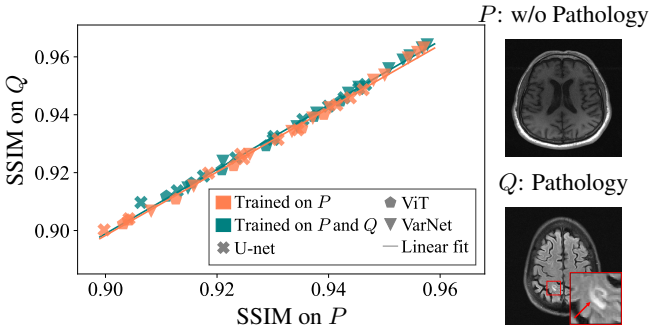


Figure 15: Models trained only on images without pathologies and models trained on images with pathologies have similar global SSIM. Different models are sampled by varying the training set size by factors of 2, 4 and 8, and by early stopping.

fastMRI PDFS knee scans (see Figure 2). We report the performance in SSIM for each model and for each acceleration factor in Table 2. We also evaluate each model on 3-fold acceleration to see how combining different accelerations affects robustness towards a distribution-shifts related to the acceleration factor.

In Table 2, we see that the model trained on all four acceleration factors simultaneously (5th row) yields similar performance to the models trained individually on each acceleration factor (the differences in performance are within 0.001 SSIM, which is negligible).

For the out-of-distribution setup (i.e., evaluation on 3-fold acceleration), we observe that the model trained on all accelerations performs by 0.006 SSIM better relative to the best separately trained model (2-fold acceleration).

Taken together, these two observations suggest that training on combinations of different acceleration factors can increase the effective robustness of a model towards distribution-shifts related to changes to the acceleration factor slightly.

C.2. Distribution-Shifts Related to Number of Receiver Coils

We now consider a shift in the number of receiver coils. For distribution P, we select all knees scans collected with the 6 different combinations of image contrasts and scanners (see Figure 2). All knee scans are collected with 15 coils. For distribution Q, we select the brain scans from the scanner Avanto since measurements from this scanner are collected with 4 coils.

For this distribution-shift we noticed that models, in particular VarNet which estimates sensitivity maps, struggles to accurately predict the mean value of the images resulting in a noticeable drops in SSIM. However, this degradation was

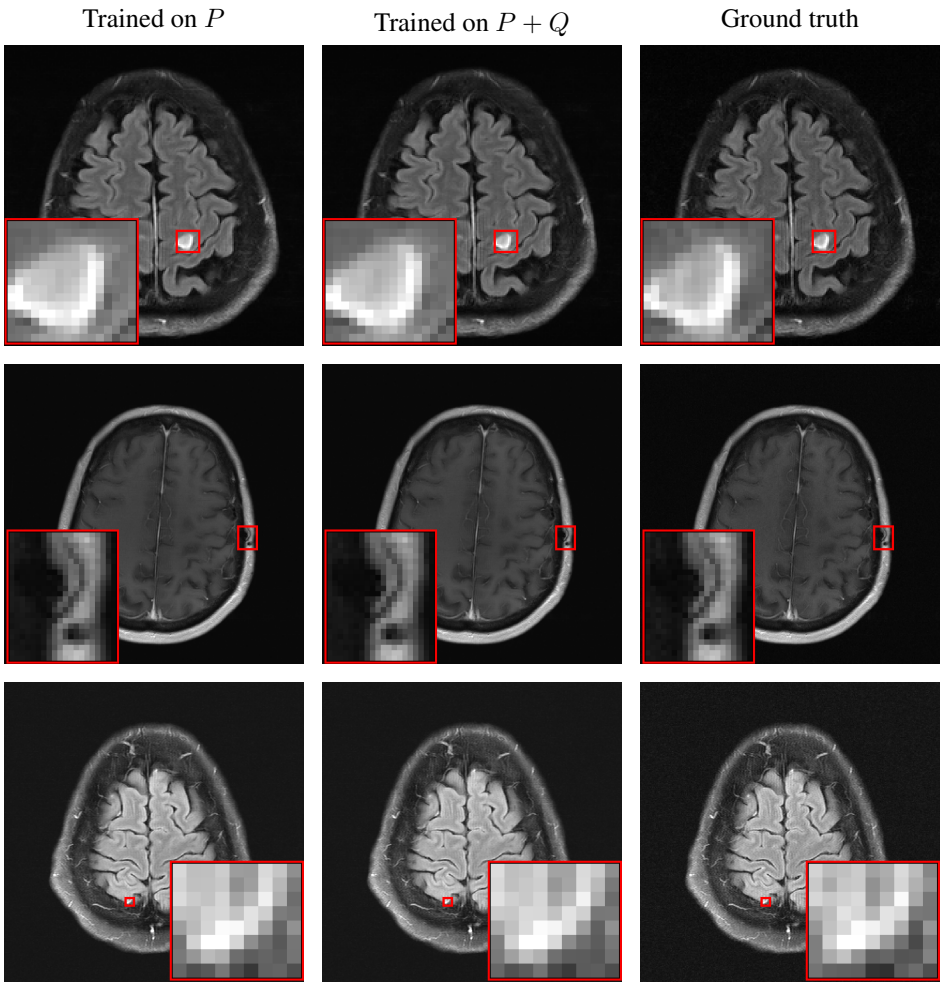


Figure 16: Random selection of reconstructions of small pathologies, given by the VarNet when trained on images without pathologies (P), and on images without and with pathologies ($P + Q$).

hardly noticeable when looking at the reconstructions. Given that radiologists routinely adjust the brightness and contrast of MRI images during inspection through a process known as windowing (Ishida et al., 1984), we normalize the model output and target to have the same mean and variance during evaluation. The results are depicted in Figure 13, where we see that training on a diverse set P increases effective robustness on this distribution-shift.

D. Measuring Similarity Between Two Datasets With CLIP Similarity

A possible explanation for performance improvements on out-of-distribution data is that a more diverse training set is more likely to contain data that is similar to out-of-distribution data and thus diverse data is 'less' out-of-distribution. Training on data that is 'less' out of distribution is in turn expected to increase performance.

There are several measures for the the similarity of two image dataset. We adapt the method from Mayilvahanan et al. (2024). Mayilvahanan et al. (2024) demonstrate that similarity between training set and test set plays a crucial role in explaining robustness of CLIP (Radford et al., 2021) models for image recognition. The method utilizes a pre-trained CLIP model to compute features of images in CLIP's image embedding space. Then given a training set and test set, this measure computes the nearest neighbor in the training set for each sample in the test set based on the cosine similarity of the CLIP image features resulting in a distribution of similarity scores—one for each sample in the test set. We then compute the mean of the histograms of those nearest neighbors.

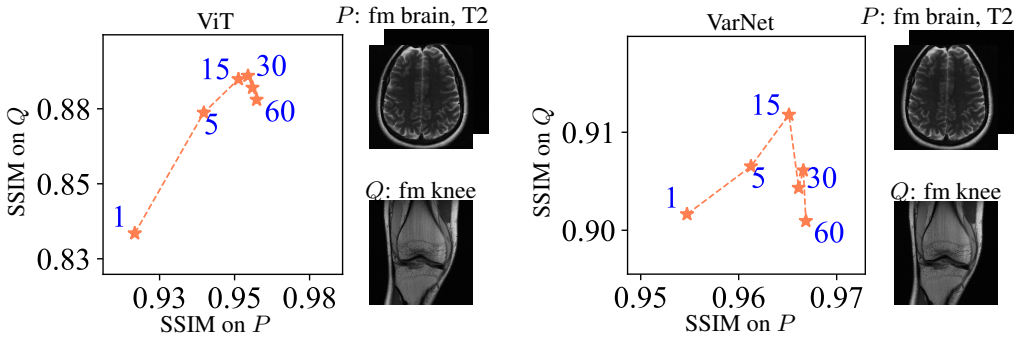


Figure 17: Distributional overfitting for ViT and VarNet. Models are trained on data from distribution P, and evaluated at different training epochs (1, 5, 15, 30, 45, 60) on P and Q.

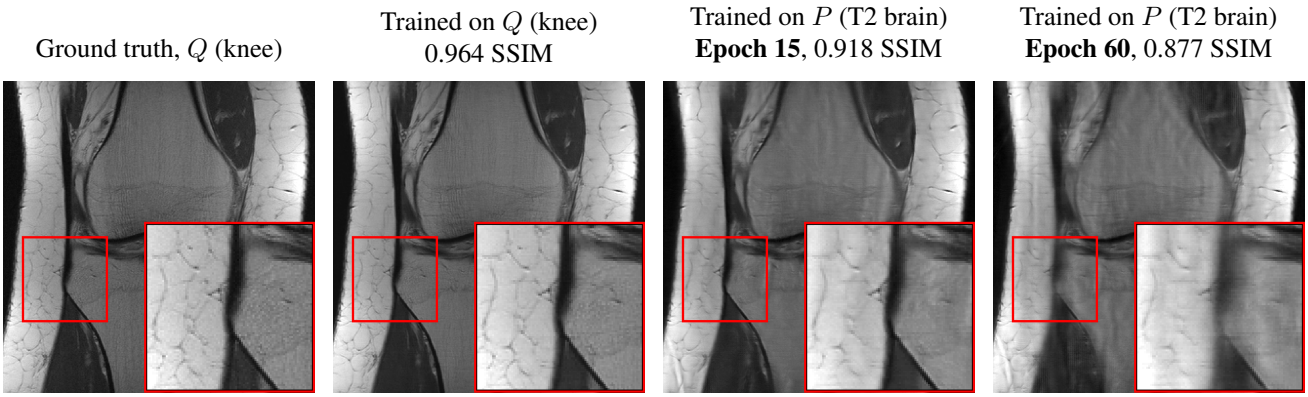


Figure 18: Reconstruction example showing how distributional overfitting affects the VarNet from Figure 17. The VarNet trained for 60 epochs on distribution P (fastMRI T2-weighted brain) produces more blurriness and artifacts in the reconstruction of a sample from distribution Q (fastMRI knee) compared to the same VarNet trained for 15 epochs. This is despite the fact, that the VarNet trained for 60 epochs performs better in-distribution (P), as seen in Figure 17.

For this study, we rely on the pre-trained CLIP models by Cherti et al. (2023) and use the ViT-B/16 (Dosovitskiy et al., 2021) pre-trained on DataComp-1B (Gadre et al., 2023), a large and diverse high-quality dataset that includes medical images. From the root-sum-of-square ground truth images, we first randomly extract smaller image patches of size 80×80 and discard those that mainly contain background and noise. CLIP features are then computed for the remaining image patches.

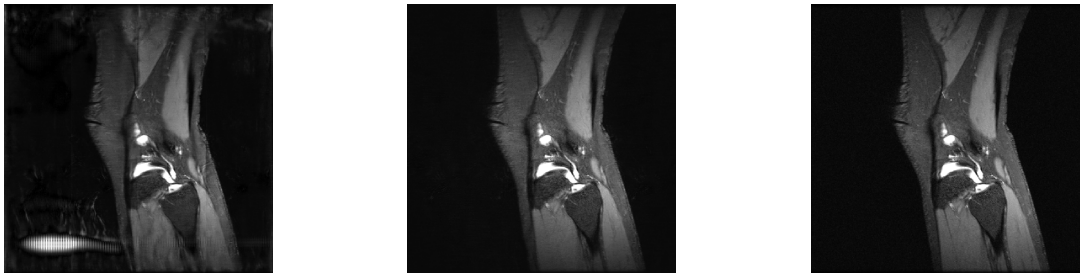
E. Additional Results for Section 5

Figure 14 presents the reconstruction performance evaluated for individual images in the test set, focusing on small pathologies. The evaluation specifically targeted the pathology regions. Results are provided for VarNet trained solely on images without pathologies (P) and VarNet trained on images with and without pathologies ($P + Q$). Both models exhibit similar mean SSIM values for test images without pathology (approximately 0.957 SSIM) and also similar SSIM values for test images with small pathologies (approximately 0.948 SSIM).

Both models perform well for the majority of samples, indicated by high SSIM scores. In the low-SSIM regime where the SSIM is low for both models, some samples are better reconstructed by the model trained on $P + Q$ and some are better reconstructed by the model trained solely on P.

In Figure 15, we show reconstruction performance when SSIM is calculated across the entire image and not just for the pathology region as in the main body. It can be seen that even when evaluated globally, models trained on data without pathologies perform as well as models trained on data with and without pathologies.

In Figure 16 we provide a selection of reconstruction examples for images with small pathologies, obtained by a VarNet



Before size adjustment | After size adjustment | Ground Truth

Figure 19: Shown are two reconstructions by a VarNet trained on fastMRI knee dataset and applied to a sample of the Stanford 3D sagittal view dataset. A mismatch in image size between training and test time can lead to artifacts (**left**). Adjusting the size of the input k-space mitigates these artifacts (**middle**).

trained on data without pathologies and a VarNet trained on data with and without pathologies. It can be seen that the reconstructions by the two models are essentially indistinguishable.

Our intuition on why models can generalize well for this particular distribution-shift is as follows: First, compared to the distributions shifts in Section 4, which include shifts between anatomies or image contrasts inducing strong differences in structure and content (see Figure 2), the pathology distribution-shift is less drastic as many image characteristics stay the same, i.e, same anatomy, scanners, image modalities. Second, we believe that models for image reconstruction learn local priors rather than global ones. Pathologies consist of local patterns which are also present in other areas of an image without pathologies.

F. Additional Results on Distributional Overfitting

In Section 6, we discussed distributional overfitting for the U-net; here we demonstrate that distributional overfitting happens equally for the ViT and VarNet. Figure 17 demonstrates for a distribution-shift from fastMRI T2-weighted brain to fastMRI knee that ViT and VarNet also suffer from distributional overfitting. Figure 18 further illustrates distributional overfitting for the VarNet on a random reconstruction example.

For image classification, studies have shown that using SGD instead of Adam can improve generalizability of neural networks (Zhou et al., 2020). To see whether distributional overfitting could be mitigated by using SGD instead of Adam, we trained U-nets with SGD on fastMRI brain data, the same data as in Section 4 for non-T2 weighted images. We trained the models for a maximum of 100 epochs, either with zero momentum or with 0.9 momentum, and maximal learning rates of 0.1, 0.01 and 0.001. Other hyperparameters are the same as in Appendix B.2. We observed that distributional overfitting was not mitigated by using SGD when evaluating on fastMRI knee data, and we also did not observe better in-distribution nor out-of-distribution performance (on T2-weighted brain and fastMRI knee) compared to using Adam.

G. Experiment Details and Additional Results for Section 7

We now discuss the experimental details for the results in section 7 on training a robust model for accelerated MRI on diverse data and provide additional analysis.

G.1. Preparation of Datasets

We convert all the dataset listed in Table 1 to follow the convention of the fastMRI knee and brain datasets, where the anatomies in images are vertically flipped, targets are RSS reconstructions, and the k-space is oriented such that the horizontal axis corresponds to the phase-encoding direction and the vertical axis corresponds to the read-out direction.

If predefined train and test splits are not already provided with a dataset, we randomly select 85% of the volumes as training set and the remaining volumes as test set. If a dataset has a designated validation set that is separate from the test set, then we include the validation set in the training set. For 3D MRI volumes, we synthesize 2D k-spaces by taking the 1D IFFT in the 3D k-space along either x, y or z dimension to create 2D volumes of different anatomical views (axial, sagittal and

Table 3: Performance of models from Section 7 when evaluated on deep feature metrics LPIPS/DISTS (lower is better). For any architecture, the model trained on the diverse collection of datasets $\mathcal{D}_P$ performs better on out-of-distribution data than the models trained on fastMRI knee/brain while showing similar performance on the in-distribution fastMRI data.

Model	Train \ Test	fastMRI knee	fastMRI brain	CC-359 sag.	NYU	Stanford 2D	M4Raw GRE	Mean
U-net	fastMRI knee	0.230/0.132	0.249/0.146	0.294/0.185	0.226/0.128	0.260/0.163	0.276/0.183	0.256/0.156
	fastMRI brain	0.299/0.163	0.224/0.130	0.281/0.168	0.270/0.150	0.283/0.168	0.253/0.160	0.268/0.157
	$\mathcal{D}_P$	0.230/0.132	0.224/0.132	0.247/0.149	0.221/0.128	0.257/0.154	0.249/0.160	0.238/0.142
	Test distribution	0.230/0.132	0.224/0.130	0.215/0.136	0.214/0.123	0.246/0.143	0.239/0.153	0.228/0.136
ViT	fastMRI knee	0.225/0.132	0.246/0.146	0.295/0.185	0.219/0.128	0.258/0.163	0.265/0.183	0.251/0.156
	fastMRI brain	0.280/0.163	0.220/0.130	0.264/0.168	0.254/0.150	0.258/0.168	0.253/0.160	0.255/0.157
	$\mathcal{D}_P$	0.223/0.132	0.221/0.132	0.228/0.149	0.219/0.128	0.243/0.154	0.252/0.160	0.231/0.142
	Test distribution	0.225/0.132	0.220/0.130	0.206/0.136	0.207/0.123	0.236/0.143	0.237/0.153	0.222/0.136
VarNet	fastMRI knee	0.209/0.122	0.226/0.132	0.306/0.189	0.201/0.106	0.244/0.143	0.344/0.226	0.255/0.153
	fastMRI brain	0.246/0.139	0.206/0.120	0.251/0.152	0.223/0.118	0.241/0.141	0.273/0.174	0.240/0.140
	$\mathcal{D}_P$	0.210/0.122	0.206/0.119	0.173/0.117	0.198/0.095	0.228/0.131	0.269/0.179	0.214/0.127
	Test distribution	0.209/0.122	0.206/0.120	0.161/0.105	0.193/0.114	0.226/0.131	0.219/0.147	0.202/0.123

coronal). However, for the SKM-TEA dataset, we only consider the sagittal view. Depending on the dataset, the first and last 15-70 slices of the synthesized 2D volumes are omitted as we mostly observe pure noise measurements:

- CC-359, sagittal view: First 15 and last 15 slices are omitted.
- CC-359, axial view: First 50 slices are omitted.
- CC-359, coronal view: First 25 and 15 slices are omitted.
- Stanford 3D, axial view: First 5 and last 5 slices are omitted.
- Stanford 3D, coronal view: First 40 and last 40 slices are omitted.
- Stanford 3D, sagittal view: First 30 and last 30 slices are omitted.
- 7T database, axial view: First 70 and last 70 slices are omitted.
- 7T database, coronal view: First 30 and last 30 slices are omitted.
- 7T database, sagittal view: First 30 and last 30 slices are omitted.

For the other datasets that are not mentioned above, all slices are used. Moreover, each of the volumes of the SKM-TEA dataset contains originally two echos due to the use of the qDESS sequence. We separate the two echos and count them as separate volumes.

fastMRI prostate T2. Originally, each volume of the fastMRI prostate T2 dataset contains three averages (Tibrewala et al., 2023): two averages sampling the odd k-space lines and one average sampling the even k-space lines. Then, for each average the authors estimate the missing k-space lines with GRAPPA (Griswold et al., 2002) and perform SENSE (Pruessmann et al., 1999) reconstruction. The final ground truth image is then obtained by taking the mean across the three averages (see code in the paper's GitHub repository). However, we convert the data as follows: we take the raw k-space and average the two averages corresponding to the odd k-space lines and then fill the missing even k-space lines with the average corresponding to the even k-space lines. This k-space serves as our k-space data. We then take this k-space and apply a 2D-IFFT and finally perform a RSS reconstruction and use this image as ground truth.

G.2. Models, Training, and Evaluation

The U-net used has 124M parameter with 4 pooling layers and 128 channels in the first pooling layer. The maximal learning rate is set to 4e-4. The ViT has 127M parameters, where we use a patch-size of 10, an embedding dimension of 1024, 16 attention heads, and a depth of 10. The maximal learning rate is set to 2e-4. The VarNet contains 8 cascades, each containing

Table 4: Quantification of artifacts in the reconstructions of the models from Section 7. We quantify the prominence of artifacts by computing the variance of the Laplacian applied to the difference image, abs(target – reconstruction), where lower values indicate fewer artifacts. Models trained on the diverse collection of datasets $\mathcal{D}_P$ tend to have less pronounced artifacts compared to models trained on the fastMRI dataset.

Model	Train \ Test	fastMRI knee	fastMRI brain	CC-359 sag.	NYU	Stanford 2D	M4Raw GRE	Mean
U-net	fastMRI knee	257	184	389	153	357	246	264
	fastMRI brain	294	157	348	167	334	197	245
	$\mathcal{D}_P$	267	169	309	165	346	194	242
	Test distribution	257	157	239	147	267	156	204
ViT	fastMRI knee	255	189	409	158	348	239	266
	fastMRI brain	293	154	329	173	303	202	242
	$\mathcal{D}_P$	255	157	288	158	265	194	220
	Test distribution	255	154	227	148	249	156	198
VarNet	fastMRI knee	245	174	434	183	318	331	281
	fastMRI brain	271	149	292	214	323	247	249
	$\mathcal{D}_P$	245	148	151	136	271	199	192
	Test distribution	245	149	117	138	211	139	166

an U-net with 4 pooling layers and 12 channels in the first pooling layer. We use linear learning rate decay and gradients are clipped at a global ℓ_2-norm of 1. For U-net and ViT, training is set for 40 epochs but we early stopped the models at epoch 24, and we use a mini-batch size of 8. The VarNet is trained for 40 epochs and we use a mini-batch size of 4. Since slice dimensions can vary across different volumes, the images within a mini-batch are chosen randomly from the same volume without replacement. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and the models are trained to maximize SSIM between model output and RSS target. Training was carried out on two NVIDIA RTX A6000 GPUs. Training the U-net took 384 GPU hours, the ViT took 480 GPU hours, and the VarNet took 960 GPU hours.

Early stopping criterion. Observing whether distributional overfitting occurs and its significance depends on both the training dataset and the test dataset as for some combinations we might not be able to observe distributional overfitting, e.g., when training on fastMRI knee data and evaluating on fastMRI brain data. However, our results suggest a common thread: when distributional overfitting occurs, in-distribution performance improvements are marginal. Therefore, we base our early stopping criterion on this observation.

Resolution mismatch. The U-net and ViT are trained on center-cropped zero-filled reconstructions, the VarNet is trained on the entire k-space and therefore on the full-sized image. For example, the average image size of fastMRI knee dataset is 640×360. Now, if we consider for example a distributions-shift from the fastMRI knee dataset to the Stanford 3D dataset which contains images of approximately half the size, we additionally introduce an artificial distribution-shift by having a mismatch between the image size from training to evaluation.

To mitigate this artificial distribution-shift we implement the following steps during inference: given the undersampled k-space and mask, we first repeat the undersampled k-space one time in an interleaved fashion in horizontal direction and another time in vertical direction, and adjust the undersampling mask accordingly. The repeated k-space and mask serve as input to the VarNet and the output is center-cropped to the original image size. As can be seen in Figure 19, these processing steps heavily reduces artifacts of the VarNets trained on the fastMRI datasets when evaluated on the Stanford 2D, CC-359 sagittal view, and M4Raw GRE dataset.

Output normalization. Similar to Appendix C.2, we observed significant drops in SSIM on out-of-distribution evaluations due to hardly visible mismatches in terms of mean or variance between model output and target. Hence, we normalize the output of the models to have the same mean and variance as the target during evaluation. This normalization reduces the SSIM score's sensitivity to variations in brightness and contrast, enabling it to better reflect structural differences.

G.3. Additional Metrics and Error Analysis

Summarizing the reconstruction quality in a single number is difficult. While SSIM is a rather standard metric for evaluating the reconstruction quality, there are other metrics that agree better with radiologit's ratings, such as such as LPIPS (Zhang et al., 2018) or DISTS (Ding et al., 2022). LPIPS an DISTS are based on features of pretrained neural networks, and

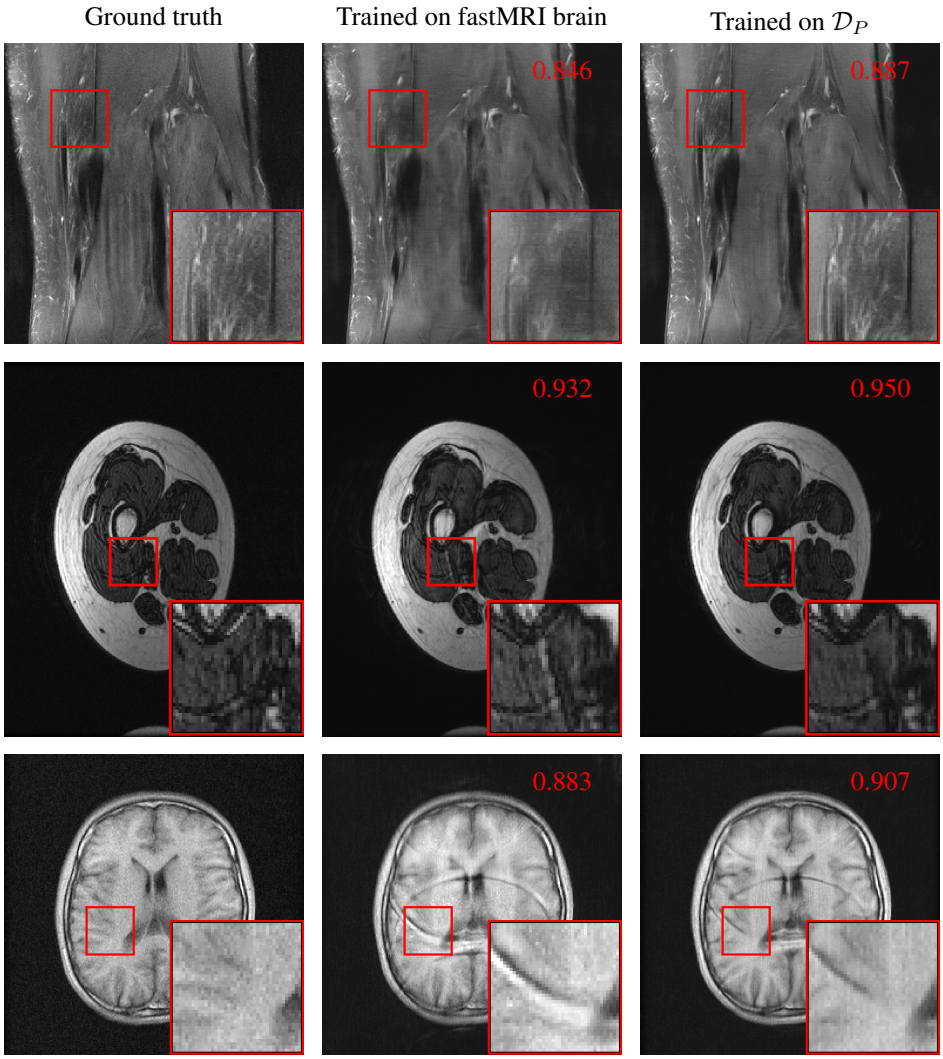


Figure 20: Out-of-distribution reconstruction examples for the NYU, Stanford 2D, and M4Raw GRE datasets. Shown are reconstructions by the VarNet trained on fastMRI brain are compared to the VarNet trained on the collection of datasets ($\mathcal{D}_P$). The numbers are the SSIM between the reconstruction and the ground truth image.

correlate relatively well with radiologist evaluations (Adamson et al., 2023; Kastryulin et al., 2023). Those metrics compute the distance between the ground-truth and the reconstruction in the feature space of pretrained neural networks.

In Table 3, we evaluate the performance of models from Section 7 on LPIPS and DISTS. We find that, consistent with the SSIM results in Section 7, the models trained on the diverse collection of datasets $\mathcal{D}_P$ perform better out-of-distribution than models trained on fastMRI knee or brain without sacrificing performance on the fastMRI datasets.

Quantification of artifacts. When inspecting the reconstructions of our models in out-of-distribution setups, we observed significant errors caused by artifacts. For example, Figure 1 (middle) depicts a reconstruction severely affected by artifacts. To quantify the prominence of artifacts, we compute the variance of the Laplacian applied to the difference image, i.e., abs(target – reconstruction), where lower values indicate fewer artifacts.

Table 4 shows the evaluation of the models from Section 7 on the prominence of artifacts. We observe that the models trained on the diverse data $\mathcal{D}_P$ tend to yield less pronounced artifacts.

In Figure 20, we provide further out-of-distribution reconstruction examples by the VarNet trained on fastMRI brain and trained on the collection of datasets $\mathcal{D}_P$ for the NYU, Stanford 2D, and M4Raw GRE datasets. A reconstruction example,

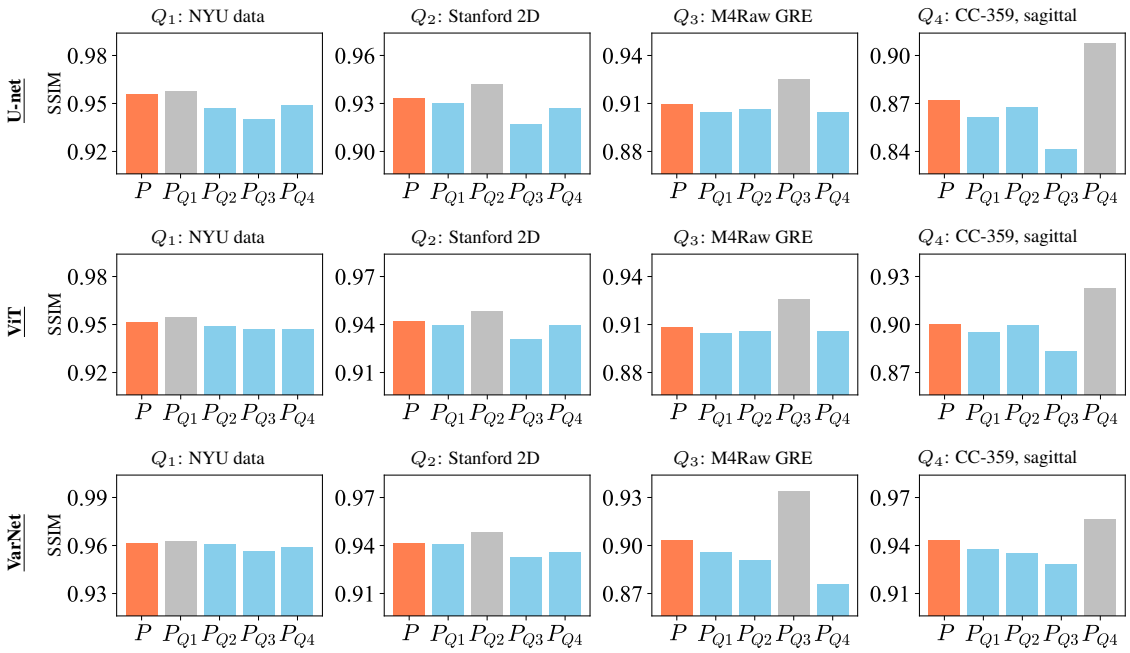


Figure 21: Fine-tuning reduces out-of-distribution robustness. P_{Qi} indicates the model obtained by fine-tuning the model trained on $\mathcal{D}_P$ (orange bar) indicated by P, on one of the distributions Q_1, Q_2, Q_3 or Q_4. It can be seen that while the models perform better on the data they are fine-tuned on (gray bar), the fine-tuned models perform worse on out-of-distribution data than the model P.

for CC-359 sagittal view is provided in Figure 1.

G.4. Finetuning Reduces Out-Of-Distribution Robustness

Our results indicate that training a model on a diverse dataset enhances its robustness towards natural distribution-shifts. In this section we demonstrate that fine-tuning an already diversely trained model on a new dataset reduces its overall robustness.

For this experiment, we take the models from Section 7 that were trained on $\mathcal{D}_P$ and fine-tune them on one of the four out-of-distribution datasets $\mathcal{D}_{Qi}$ (see last four rows in Table 1). We denote the model fine-tuned on Q_i by P_{Qi}. As depicted in Figure 21, the fine-tuned model P_{Qi} exhibits improved performance on the specific distribution Q_i it is fine-tuned on, as expected. However, the model under-performs on all other datasets in comparison to the model trained on $\mathcal{D}_P$ (i.e., prior to fine-tuning).