Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing

Mohammad Zalbagi Darestani [1] **Jiayu Liu** [2] **Reinhard Heckel** [2 1]

Abstract

Deep learning based image reconstruction methods outperform traditional methods. However, neural networks suffer from a performance drop when applied to images from a different distribution than the training images. For example, a model trained for reconstructing knees in accelerated magnetic resonance imaging (MRI) does not reconstruct brains well, even though the same network trained on brains reconstructs brains perfectly well. Thus there is a distribution shift performance gap for a given neural network, defined as the difference in performance when training on a distribution P and training on another distribution Q, and evaluating both models on Q. In this work, we propose a domain adaptation method for deep learning based compressive sensing that relies on self-supervision during training paired with test-time training at inference. We show that for four natural distribution shifts, this method essentially closes the distribution shift performance gap for state-of-the-art architectures for accelerated MRI.

1. Introduction

Deep learning methods enable fast and accurate image reconstruction and outperform traditional methods on a variety of imaging tasks (Dong et al., 2014; Jin et al., 2017; Zhang et al., 2017; Sriram et al., 2020; Rivenson et al., 2018; Jalal et al., 2021). Performance is typically measured as in-distribution performance: A dataset is split into test and training sets, and a method trained on the training set is evaluated on the test set.

In practice, however, the train and test distributions are usually different: For example, we train a network on data from one hospital, and apply the network to data from a different hospital. Or we train on data acquired with one scanner type and acquisition mode, and apply it to a different scanner type or acquisition mode.

Deep learning imaging methods perform significantly worse under such distribution shifts. For accelerated MRI, a medical imaging technique, deep learning methods incur a significant accuracy drop when shifting from one distribution to another, as shown for three natural distribution shifts by Zalbagi Darestani et al. (2021) and for SNR changes by Knoll et al. (2019).

The part of this accuracy drop that can be overcome in principle can be measured by the *distribution shift performance gap*: Suppose P and Q denote the train and test distributions. We define the distribution shift performance gap as the reconstruction accuracy (measured by a standard metric, e.g., the SSIM score) of training on Q and testing on Q minus the reconstruction accuracy of training on P and testing on Q.

In this paper, we propose a novel domain adaptation method for deep learning based compressive sensing, and show that it overcomes the gap caused by four natural distribution shifts in accelerated MRI. Our approach consists of two parts: (1) including self-supervision during the supervised training stage of deep learning models, and (2) performing self-supervised test-time training for each new test sample at inference.

We show that our method works with two well-known network architectures: the baseline U-Net (Ronneberger et al., 2015) and the state-of-the-art end-to-end variational network (Sriram et al., 2020). We evaluate robustness under four natural distribution shifts,

[1] Department of Electrical and Computer Engineering, Rice University [2] Department of Electrical and Computer Engineering, Technical University of Munich. Correspondence to: Mohammad Zalbagi Darestani <mz35@rice.edu>, Reinhard Heckel <rh43@rice.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

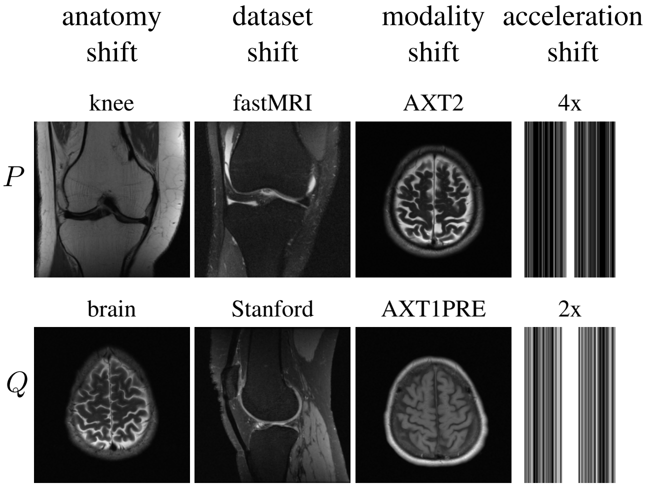

Figure 1: We study our domain adaptation method under four natural distribution shifts: anatomy, dataset, modality, and acceleration shifts. P and Q are the train and test domains.

illustrated in Figure 1: (1) anatomy shift, (2) dataset shift where the training set comes from a different hospital than the test set, (3) modality shift where the acquisition mode changes, and (4) acceleration shift where the acceleration factor changes. For the U-Net, our method closes the distribution shift performance gap by 98.6, 87.4, 96.3%, 97.4% (see Table 1) which in each case yields a significant increase in image quality.

1.1. Prior work

A series of recent influential works in image classification has shown that image classifiers often incur a significant performance drop under natural distribution shifts (Recht et al., 2019; Taori et al., 2020; Hendrycks et al., 2021; Koh et al., 2021; Miller et al., 2021). For image reconstruction, Zalbagi Darestani et al. (2021) demonstrated that reconstruction methods for accelerated MRI (even un-trained methods such as ℓ_1-minimization tuned on the train distribution) also suffer a significant accuracy drop when shifting from one distribution to another. (Knoll et al., 2019; Johnson et al., 2021) also observed a performance drop under distribution shifts in MRI. Consequently, several works, mainly in image classification, made efforts to overcome the distribution shift performance gap:

Robust optimization. Distributionally-robust optimization learns a model by minimizing a loss with a robustness notion (Duchi & Namkoong, 2019; Duchi et al., 2020; Duchi & Namkoong, 2021). Robust optimization methods yield a robustness gain on synthetic distribution shifts, and can yield a small gain on some natural shifts Koh et al. (2021), but it is unclear whether they yield significant gains on natural shifts.

Data-driven interventions. Training on larger datasets (Mahajan et al., 2018; Yalniz et al., 2019) and data augmentations (DeVries & Taylor, 2017; Geirhos et al., 2018; Zhang et al., 2018; Engstrom et al., 2019; Hendrycks et al., 2019; Yun et al., 2019; Xie et al., 2020) are popular data-driven robustness interventions. For classification, Taori et al. (2020) found training with more data to marginally improve model robustness to natural distribution shifts. Fabian et al. (2021) found a small improvement by using data augmentations, and Desai et al. (2021) recently also proposed a data augmentation scheme that yields robustness gains for accelerated MRI.

Domain adaptation. Fine-tuning-based domain adaptation pre-trains classifiers on an auxiliary distribution R, then fine-tunes on a train distribution P, and evaluates on a test distribution Q to measure the out-of-distribution (OOD) generalization performance (Sharif Razavian et al., 2014; Donahue et al., 2014; Kornblith et al., 2019). Zero-shot learning methods, pre-train classifiers on R and perform zero-shot inference on Q (Radford et al., 2021). A third group of domain adaptation methods, most closely related to ours, train classifiers on P and perform per-instance test-time training (TTT) at inference (Sun et al., 2020; Liu et al., 2021b; Wang et al., 2021). However, Miller et al. (2021) has shown that for several natural distribution shifts, zero-shot methods offer only marginal robustness improvements and the other domain adaptation methods do not offer any improvement.

Domain adaptation methods have also been proposed for improving robustness in imaging. For image denoising, Mohan et al. (2021) proposed GainTuning, which performs TTT at inference only on a few scaling factors of a neural network. We studied a version of this method tailored to MRI, and found no improvement in robustness for our problems. For accelerated MRI reconstruction, Liu et al. (2021a) proposed a method that assumes access to examples from the target domain, which we do not have here.

Finally, Yaman et al. (2021) proposed a zero-shot learn-

training scheme	setup	P: knee	P: fastMRI	P: AXT2	P: 4x	P: fastMRI
		Q: brain	Q: Stanford	Q: AXT1PRE	P: 2x	Q: adv-filt fastMRI
supervised	ctrain on Q test on Q	0.9187	0.7164	0.9026	0.9004	0.6865
	train on P test on Q	0.8521	0.6830	0.8506	0.8385	0.6861
	distribution shift performance gap	0.0666	0.0334	0.0520	0.0619	0.0004
self-supervision included	train on Q test on Q + TTT	0.9234	0.7268	0.9086	0.9192	0.6827
	train on P test on Q + TTT	0.9225	0.7226	0.9067	0.9176	0.6806
	distribution shift performance gap	0.0009	0.0042	0.0019	0.0016	0.0021
	fraction of gap closed by TTT	98.6%	87.4%	96.3%	97.4%	–

Table 1: **Self-supervision with test-time training (TTT) closes 99%, 87%, 96%, and 97% of the distribution shift performance gap for anatomy, dataset, modality, and acceleration distribution shifts.** SSIM scores are averaged over 100 samples for U-Net. The adversarially-filtered shift is an example where the distribution shift performance gap is close to zero (more than an order of magnitude smaller than for the other shifts), and thus TTT is not having an impact here (and other methods are also not expected to have an impact here).

ing method (ZS-SSL) for accelerated MRI and demonstrated that it can be used as a TTT method as well. Specifically, Yaman et al. (2021) applied ZS-SSL to a pre-trained (on one anatomy in a fully supervised manner) model and observed that it improves model robustness under anatomy shift. Our TTT approach differs from ZS-SSL in that ZS-SSL relies on creating a synthesized dataset by repeatedly splitting the given under-sampled measurement into training and validation measurements, whereas our approach is to include a self-supervised loss in pre-training and then performing TTT with respect to that self-supervised loss. See the supplement for further comparison.

2. Problem setup

We consider the problem of reconstructing an image from undersampled measurements. We focus on accelerated multi-coil magnetic resonance imaging (MRI), but our method also applies to other compressive sensing image reconstruction problems, for example to computed tomography. For such imaging problems, deep learning methods perform best. In our setup, the network is trained on one distribution (e.g., knees) and is tested on another distribution (e.g., brains).

2.1. Compressive sensing

Our goal is to reconstruct an image $\mathbf{x}^* \in \mathbb{C}^N$ from undersampled measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x}^* + \text{noise} \in \mathbb{C}^M, \tag{1}$$

where the number of measurements, M, is typically lower than the dimension of the image, N. We are given the measurement matrix $\mathbf{A}$. We focus on accelerated MRI, in which the measurements, often called k-space measurements, are obtained as

$$\mathbf{y}_i = \mathbf{M}\mathbf{F}\mathbf{S}_i\mathbf{x}^* + \text{noise} \in \mathbb{C}^{M_c}, \quad i = 1, \ldots, n_c.$$

Here, n_c denotes the number of radiofrequency coils, $\mathbf{S}_i$ is a complex-valued position-dependent coil sensitivity map, that is applied through element-wise multiplication to the image $\mathbf{x}^*$, $\mathbf{F}$ is the 2D discrete Fourier transform, and $\mathbf{M}$ is a mask (a diagonal matrix with ones and zeros on its diagonal) that implements undersampling of k-space data. The measurements $\mathbf{y}_i$ and matrices can be organized so that the measurement model has the form (1).

The MRI datasets we work with (see Section 5) consist of pairs of measurements and corresponding reference image $\{(\mathbf{x}_j, \mathbf{y}_j)\}$. The datasets are constructed from fully-sampled MRI data (i.e., taken with an identity mask $\mathbf{M} = \mathbf{I}$). The reference images are obtained by reconstructing the coil images from each full coil measurement as $\mathbf{x}_i = \mathbf{F}^{-1}\mathbf{y}_i$ and then combining them via the root-sum-of-squares (RSS) algorithm to a single image: $\mathbf{x} = \sqrt{\sum_{i=1}^{n_c} |\mathbf{x}_i|^2}$. Here, $|\cdot|$ and $\sqrt{\cdot}$ denote element-wise absolute value and squared root operations. The under-sampled k-space measurements (for acceleration) are obtained by applying a standard 1D random mask (random vertical lines in the frequency domain), which is the default in the fastMRI challenge. We consider 4x acceleration throughout the paper, the acceleration factor considered in the fastMRI challenge (Knoll et al., 2020; Muckley et al., 2021).

2.2. Image reconstruction with neural networks

We study our domain adaptation method for two neural networks, a standard baseline method (U-net) and the state-of-the-art reconstruction method (VarNet).

U-Net (Ronneberger et al., 2015) is a convolutional network which for MRI is trained end-to-end to map a least-squares reconstruction obtained from a measurement $\mathbf{y}$ to a clean image $\mathbf{x}$ by minimizing the loss $\mathcal{L}(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \left\|\mathbf{x} - f_{\boldsymbol{\theta}}(\mathbf{A}^{\dagger}\mathbf{y})\right\|_1$ over a training dataset $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ (Jin et al., 2017). Here, $f_{\boldsymbol{\theta}}$ is a U-Net with parameters $\boldsymbol{\theta}$.

VarNet is a variational network that gives state-of-the-art performance (Sriram et al., 2020). Similar to U-Net, VarNet is trained to map an under-sampled measurement to a clean image, but contrary to U-net, it contains data consistency blocks.

2.3. Problem statement: Overcoming the distribution shift performance gap

Neural networks for MRI are often trained on the fastMRI training set (Zbontar et al., 2020) and evaluated on the fastMRI test set. This measures in-distribution performance since the sets are constructed by collecting data and splitting it in train and test sets.

This evaluation mode is not reflective of performance in practice, where we typically train on data acquired in one setup (anatomy, scanner type, acquisition mode, etc.) and apply the network in a different setup on data from another anatomy, scanner, or acquisition mode. All those changes introduce distribution shifts.

Under anatomy shifts (training on knee and testing on brain) and dataset shifts (training on NYU data (Zbontar et al., 2020) and testing on Stanford data (Epperson et al., 2013)), different methods lose a similar and significant amount in image quality Zalbagi Darestani et al. (2021). This loss comprises two parts: one is due to variations in difficulty of the datasets; e.g., images with finer details are harder to reconstruct and result in lower scores. This part cannot be overcome by better algorithms or even by having access to the test distribution. The second part is due to a missfit of algorithm and test distribution, this can in principle be overcome if we had access to the test distribution. We call this second gap the **distribution shift performance gap**.

The goal of this work is to close the distribution shift performance gap without having access to the test distribution. We assume that we can train a network on a training distribution P, but at inference, we are only given a measurement $\mathbf{y}$ from a test distribution Q, without any other information about the test distribution.

3. The distribution shift performance gap for four natural distribution shifts

In this section, we introduce the four natural distribution shifts we consider and measure the corresponding distribution shift performance gap. We also include a fifth distribution shift for which the distribution shift performance gap is close to zero, even though we observe a performance drop.

We then show that the models we consider have the ability to close the performance gap, by training a network on data from both distributions. This show that the networks we consider can perform well on both distributions simultaneously. Of course, in practice we cannot train on the test distribution, since we do not have access to examples from the test distribution.

3.1. Natural distribution shifts considered

We consider four natural distribution shifts illustrated in Figure 1, and one artificial distribution shift.

Anatomy shift. We consider an anatomy shift from training on fastMRI knee images to testing on fastMRI brain images.

Dataset shift. We consider a dataset shift from training on fastMRI knee images (collected at NYU) to testing on Stanford knee images (Epperson et al., 2013). The main differences are: (1) The Stanford data is constructed via volumetric 3D recording (fastMRI scans are 2D), (2) The Stanford set contains samples with a lower frequency resolution than fastMRI, and (3) the slice thickness is 5 times smaller in the Stanford set.

Modality shift. A modality shift occurs when the acquisition mode of training images is different than the one for test images. A modality shift is subtle, in that it occurs within an anatomy (say brain) and only the contrast of the images changes. We consider a modality shift from AXT2 to AXT1PRE images (see Figure 1 for an example, and the supplement for the change in pixel intensity distribution).

Acceleration shift. We consider an acceleration shift from training on 4x accelerated measurements to testing on 2x accelerated measurements.

For each of these natural distribution shifts, we compute the distribution shift performance gap for a given model as the gap in SSIM (an image comparison metric) when the training domain changes. Table 1 shows the distribution shift performance gap for U-Net given each of the distribution shifts explained above.

We note that there are distribution shifts for which the distribution shift performance gap is essentially zero, and thus no robustness intervention can be helpful. An example is an **adversarially-filtered shift** which occurs when the target distribution contains hard-to-reconstruct samples from the original domain. Specifically, models trained on the fastMRI dataset achieve a significantly lower score on fastMRI-A, which contains hard-to-reconstruct samples (Zalbagi Darestani et al., 2021). However, models achieve a low score on fastMRI-A simply because those samples contain finer details. This can be seen in Table 1 where the best performance obtainable by models is 0.6865 and when models are trained on fastMRI instead of fastMRI-A, they still achieve that performance on fastMRI-A. This example shows that a performance degradation on the target domain does not necessarily translate into a distribution shift performance gap as defined above.

3.2. Networks can close the distribution shift performance gap

Before describing our domain adaptation method, we investigate whether models are capable of achieving close-to-zero performance gap. To this end, we assume access to data from distribution Q and train a U-Net on the mixture distribution of P and Q, for each distribution shift introduced in the previous section, and evaluate the models on the two distributions. The results are shown in Figure 2, which depicts performance as a function of the mixture coefficient, which indicates the proportion of data from distribution Q.

The distribution shift performance gap is captured by the difference in the vertical direction between points with mixture coefficient 1.0 and 0.0. For the natural distribution shifts we study, there is a significant distribution shift performance gap, while for adversarially-filtered shift, the gap is relatively small.

We observe an approximately vertical section intersected with an approximately horizontal section at a relatively sharp angle. This shows that, when more data from the target distribution Q is added to the training set, performance on Q increases while performance on P does not degrade. As a consequence, the performance gap is roughly closed when the model is trained on the mixture distribution with the mixture coefficient at the intersection point.

4. Method: Incorporating self-supervised training and then performing test-time training at inference

In this section, we describe our domain adaptation method for compressive sensing which incorporates a self-supervised loss into the training of a deep learning model, and performs test-time training (TTT) during inference. Let $f_{\boldsymbol{\theta}}$ be a neural network mapping a coarse reconstruction from a measurement $\mathbf{y}$ to a clean image (e.g., a U-Net or VarNet). The training and inference stages are as follows.

Training stage: Given a training set consisting of (ground-truth-image, measurement) pairs $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, we learn a model by minimizing the loss function $\mathcal{L}(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) =$

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\left\|\mathbf{x}_i - f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y}_i)\right\|_1}{\left\|\mathbf{x}_i\right\|_1}}_{\mathcal{L}_{\text{sup}}} + \underbrace{\frac{\left\|\mathbf{y}_i - \mathbf{A} f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y}_i)\right\|_1}{\left\|\mathbf{y}_i\right\|_1}}_{\mathcal{L}_{\text{self}}}. \tag{2}$$

Here, $\mathcal{L}_{\text{sup}}$ is the supervised part of the loss function which ensures that the output of the network is close to the ground truth image. For our setup, the ℓ_1-norm works well as a loss but other choices, such as the SSIM loss, also work.

The un-supervised loss we propose is simply enforcing data-consistency, and might seem like an odd choice, given that existing TTT schemes from the classification literature typically choose an auxiliary task, and not a fitting term as a self-supervised loss. We also experimented with auxiliary reconstruction tasks, such as a self-supervised denoising task (see appendix), without success. We further discuss the motivation for the un-supervised loss below.

Inference stage: At inference, we are given a (typically out-of-distribution) measurement $\mathbf{y}$, and we estimate an image as follows. We optimize the network parameters $\boldsymbol{\theta}$ with respect to the loss $\mathcal{L}_{\text{self}}$ for the given under-sampled test measurement. We refer to this step as TTT. To prevent TTT from overfitting to the measurement, we early-stop the optimization based on a self-validation loss computed over a fraction of the

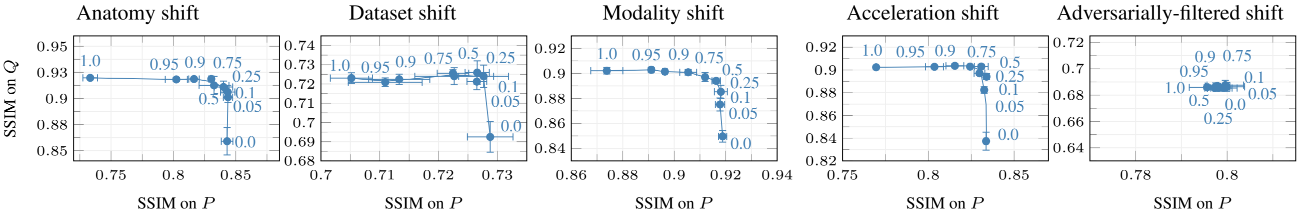

Figure 2: Performance of U-Net trained on the mixture distributions of source and target domains: Numbers are the mixture coefficients; for example 0.25 means that 25% of the training data comes from distribution Q. The performance gap is significant for the four natural distribution shifts, but essentially non-existent for the adversarially-filtered shift. The model can perform well on both distributions P and Q simultaneously irrespective of the distribution shifts. Error bars denote 95% confidence intervals.

under-sampled measurement. Specifically, let $\mathbf{y}$ be the under-sampled test measurement with M frequency-domain measurements. We split $\mathbf{y}$ into a $\mathbf{y}_{\text{train}}$ and $\mathbf{y}_{\text{val}}$, which contains a random fraction $q = M/k$ of all measurements in $\mathbf{y}$. We then perform TTT on $\mathbf{y}_{\text{train}}$ and monitor the error between $\mathbf{y}_{\text{val}}$ and the predicted frequencies by the network for self-validation to stop TTT early when the self-validation error starts to rise.

Motivation for the self-supervised loss: Our self-supervised loss is low if the network generates an image that is consistent with the measurement. This loss might seem an unusual choice, given that TTT methods from the literature usually use a loss based on an auxiliary task. However, we observe footprints of this loss in works on cycle consistency losses in image translation (Zhu et al., 2017) and image classification (Hoffman et al., 2018). For image-to-image translation, Zhu et al. (2017) proposed cycle consistency to learn a consistent mapping from one image domain X to another image domain Y. Oh et al. (2020) used a cycle consistency loss to train a GAN for image reconstruction.

Our self-supervised loss uses the network as an image model to enforce consistency between the output and the under-sampled measurement. Therefore, we expect such self-supervision to work well for architectures that are *good* image models. Both U-Net and VarNet are Convolutional Neural Networks (CNNs), and CNNs are good image priors even without any training Ulyanov et al. (2018). Un-trained CNNs are such good image priors so that they can perform image reconstruction without any training (Ulyanov et al., 2018; Heckel & Hand, 2019; Arora et al., 2020; Heckel & Soltanolkotabi, 2020; Wang et al., 2020; Bostan et al., 2020; Zalbagi Darestani & Heckel, 2021).

5. Experiments

In this section, we show that self-supervised training with test-time training (TTT) closes the distribution shift performance gap for the four natural distribution shifts considered.

For each experiment, we have an original distribution P and a target distribution Q. Performance is measured in terms of the structural similarity index measure (SSIM). The distribution shift performance gap when training is fully supervised is defined as $\text{gap}_{\text{before}} = \text{SSIM(train on } Q \text{ test on } Q) - \text{SSIM(train on } P \text{ test on } Q)$, and is defined as $\text{gap}_{\text{after}} = \text{SSIM(train on } Q \text{ test on } Q + \text{TTT}) - \text{SSIM(train on } P \text{ test on } Q + \text{TTT})$ when we include self-supervision during training and then apply TTT. If the robustness intervention is successful, then $\text{gap}_{\text{after}}$ is significantly smaller than $\text{gap}_{\text{before}}$.

Throughout the experiments, we work with a U-Net with 8 layers and 64 channels as the width factor and a VarNet with 12 cascades and 18 channels as the width factor. For training, we run the Adam optimizer (Kingma & Ba, 2015) with learning rate set to 1e-5 for U-Net (and 1e-4 for VarNet).

The train and test datasets we used for each experiment are described below for each distribution shift. We observed that a training set of 300-400 slices gives a good performance on the validation set. For larger training sets, we observed no significant performance improvement (less than 0.5% SSIM score).

Anatomy shift: Here, the distribution P are knee images, and Q are brain images. The knee training set consists of 376 PD knee slices of the fastMRI training dataset and the brain training set consists of 310

		anatomy shift		**dataset shift**		**modality shift**		**acceleration shift**	
		P:knee	**Q:brain**	**P:fastMRI**	**Q:Stanford**	**P:AXT2**	**Q:AXT1PRE**	**P:4x**	**Q:2x**
training scheme	setup	U-Net	VarNet	U-Net	VarNet	U-Net	VarNet	U-Net	VarNet
		SSIM	SSIM	SSIM	SSIM	SSIM	SSIM	SSIM	SSIM
self-supervision included	train on P test on Q + TTT	0.9225	0.9315	0.7226	0.7247	0.9067	0.9084	0.9176	0.9203
	train on Q test on Q + TTT	**0.9234**	0.9322	**0.7268**	0.7295	**0.9086**	**0.9110**	**0.9192**	0.9207
supervised	train on Q test on Q	0.9187	**0.9344**	0.7164	**0.7442**	0.9026	0.9105	0.9004	**0.9224**
	train on P test on Q	0.8521	0.8717	0.6830	0.7062	0.8506	0.8744	0.8385	0.8744
	train on P test on Q + TTT	0.8648	0.8533	0.6130	0.6026	0.8695	0.8102	0.8090	0.8324
fraction of gap closed by TTT		98.6%	98.8%	87.4%	87.3%	96.3%	92.7%	97.4%	99.1%

Table 2: **Including self-supervision while training deep learning models combined with TTT improves model robustness to natural anatomy, dataset, modality, and acceleration shifts.** SSIM scores are averaged over 100 slices from the test set of Q. Note that TTT is only effective when self-supervision is included during training, and as shown in the second to the last row, it offers marginal improvement when applied to a model that is trained in a supervised fashion.

AXT2 brain slices of the fastMRI training dataset. We train U-Net and VarNet on both training sets separately, and test them on a brain test set consisting of 100 AXT2 brain slices of the fastMRI validation dataset, this yields the numbers for supervised training in Table 2. Next, we performed the same experiment, but this time we included self-supervision into the training, and we applied test-time-training as described above.

Dataset shift: Here, the distribution P is a subset of the fastMRI dataset, and Q is the Stanford dataset. The version of the fastMRI training set consists of 366 PDFS knee slices, and the Stanford training set consists of 308 PDFS knee slices. We then repeat the same experiment explained above for anatomy shift.

Modality shift: Here, the P and Q distributions are the AXT2 and AXT1PRE slices of the fastMRI brain dataset. The AXT2 training set consists of 310 slices, and the AXT1PRE consist of 320 slices. We then repeat the same experiment explained above.

Acceleration shift: Here, the distribution P are 4x accelerated knee measurements, and Q are 2x accelerated knee measurements. The 4x training set consists of 376 PD knee slices of the fastMRI training dataset and the 2x training set consists of the same 376 PD knee slices but accelerated 2 times instead of 4.

Table 2 contains the results of the experiments. Figure 3 contains example reconstructions, more examples are in the appendix in Figures 11, 12, 13, and 14.

5.1. Discussion on the results

We draw the following conclusions from the results.

TTT essentially closes the distribution shift performance gap. For the four considered natural distribution shifts, our method closes the distribution shift performance gap by 87-99%. Thus, our method closes the gap for practical purposes for those four distribution shifts. This is also reflected in a significant increase in perceptual quality (see Figure 3 and the appendix).

TTT also slightly improves in-distribution performance, which is not surprising as it tunes the parameters on each instance individually as opposed to finding the parameters that work best on average on all slices. That is why we measure the performance gap with and without TTT included.

Increased computation cost. Our approach offers significant reduction in the distribution shift performance gap. However, TTT comes at the cost of more computations at the inference as shown in Table 3.

5.2. Ablation studies

Including self-supervision in training is critical for performance. TTT on a model trained only on the supervised loss (without the self-supervised loss during training) gives only minimal (for anatomy shift) to no (for dataset shift) performance improvement.

Early-stopping TTT is critical for performance. TTT is performed on the under-sampled test measurement and overfits to it without our early-stopping mechanism, which chooses the early-stopping time based on a hold-out set obtained from the under-sampled measurement. Figure 5 in the supplement illustrates that early stopping is critical for TTT to perform well.

		anatomy shift **P:knee Q:brain**		**dataset shift** **P:fastMRI Q:Stanford**		**modality shift** **P:AXT2 Q:AXT1PRE**		**acceleration shift** **P:4x Q:2x**	
training scheme	setup	U-Net runtime (mins/slice)	VarNet runtime (mins/slice)	U-Net runtime (mins/slice)	VarNet runtime (mins/slice)	U-Net runtime (mins/slice)	VarNet runtime (mins/slice)	U-Net runtime (mins/slice)	VarNet runtime (mins/slice)
self-supervision included	train on P test on Q + TTT	10.1	12.9	1.2	1.5	1.6	6.5	3.6	5.2
	train on Q test on Q + TTT	3.3	4.2	0.4	0.7	0.7	2.6	2.6	5.1

Table 3: **Reducing the robustness gap comes at a noticeable computational cost for TTT.** Runtimes are averaged over 100 slices from the test set of Q and are reported for a single GPU.

Variants of TTT. TTT in image classification is typically quite different than the version of TTT proposed here, in that a network with two heads is trained, one part on a supervised loss and one part at a self-supervised loss on an auxiliary task, such as predicting rotations of an image. At inference, TTT is performed on the self-supervised loss (Sun et al., 2020). We experimented with the analogous idea for image reconstruction: We took a U-Net with two decoders and a joint encoder, and trained the network on a self-supervised denoising problem and the supervised compressive sensing problem. At inference, we performed TTT on the denoising problem. This approach fails to improve model robustness (see Figure 6 in the supplement).

We also experimented with another variant of TTT that also does not work well. We exploited the idea behind the CycleGAN (Zhu et al., 2017) to build CycleU-Net. CylcleU-Net comprises two U-Nets in tandem and is trained based on two forward passes: (1) a supervised pass mapping the ground-truth image to itself, and (2) a self-supervised pass mapping the low-quality image to itself (by switching the places of the two U-Nets). This type of training enables TTT w.r.t. the self-supervised pass at inference. Table 5 in the supplement shows that only 29% of the gap is closed for anatomy shift when this approach is employed.

6. Test-time training can provably adapt to a distribution shift

In this section, we discuss an example illustrating that test-time training (TTT) with an appropriate loss can optimally adapt to a particular distribution shift.

Consider the problem of denoising a signal that lies in a subspace. The training distribution draws a signal from an unknown d-dimensional subspace and corrupts it by Gaussian noise with noise variance σ^2:

$$P\colon \mathbf{y} = \mathbf{x} + \mathbf{z}, \mathbf{x} = \mathbf{U}\mathbf{c}, \mathbf{c} \sim \mathcal{N}(0, \mathbf{I}), \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}\sigma^2).$$

The test distribution draws a signal from the same subspace, but corrupts it with noise that has a different noise variance:

$$Q\colon \mathbf{y} = \mathbf{x} + \mathbf{z}, \mathbf{x} = \mathbf{U}\mathbf{c}, \mathbf{c} \sim \mathcal{N}(0, \mathbf{I}), \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}\varsigma^2).$$

Here, $\mathbf{U} \in \mathbb{R}^{n \times d}$ is an orthonormal basis for the signal subspace. We consider a linear denoiser of the form $f(\mathbf{y}) = \alpha \mathbf{V}\mathbf{V}^T\mathbf{y}$, where $\mathbf{V} \in \mathbb{R}^{n \times d}$ is an orthonormal basis and $\alpha \in [0, 1]$ is a scalar. We measure performance in terms of the population risk

$$R_Q(\alpha, \mathbf{V}) = \mathbb{E}_Q \left[\left\| \mathbf{x} - \alpha \mathbf{V}\mathbf{V}^T\mathbf{y} \right\|_2^2 \right].$$

The population risk is minimized by the estimator $\alpha = \frac{1}{1+\varsigma^2}$ and $\mathbf{V} = \mathbf{U}$. So the optimal estimator for the unknown signal $\mathbf{x}$ is $\hat{\mathbf{x}} = \frac{1}{1+\varsigma^2}\mathbf{U}\mathbf{U}^T\mathbf{y}$ and it projects the observation onto the subspace and then shrinks with a coefficient that is dependent on the noise variance, the larger the noise variance, the more the estimator shrinks.

We train the method on the supervised loss

$$\mathcal{L}_P(\alpha, \mathbf{V}) = R_P(\alpha, \mathbf{V}). \tag{3}$$

For simplicity, we choose as the supervised loss the population loss, because we are not interested in finite-sample effects. This corresponds to a setup where we have abundant training data on the distribution P.

A minimizer of the loss is $\alpha = \frac{1}{1+\sigma^2}$, $\mathbf{V} = \mathbf{U}$, thus minimizing on the loss gives an estimator that is optimal on the distribution P. However, the estimator is sub-optimal on the distribution Q.

We next obtain an observation $\mathbf{y}$ from the distribution Q, and our goal is to estimate the corresponding signal $\mathbf{x}$. Towards this end, we first perform TTT on the distribution Q, i.e., we minimize the self-supervised loss

$$L_{SS}(\alpha, \mathbf{U}, \mathbf{y}) = \left\| \mathbf{y} - \alpha \mathbf{U}\mathbf{U}^T\mathbf{y} \right\|_2^2 + \frac{2\alpha d}{n-d} \left\| (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{y} \right\|_2^2$$

over the scale parameter α. Note that we already

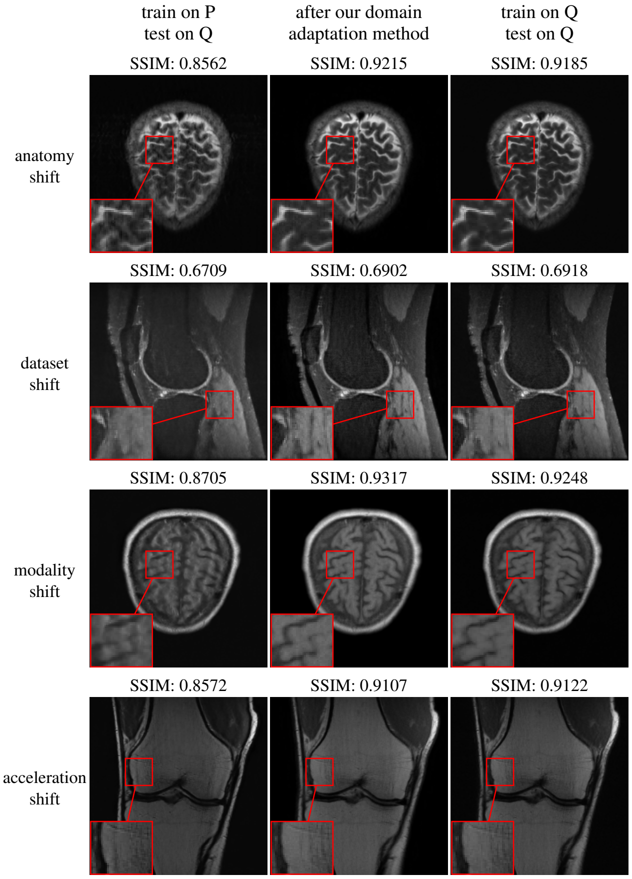

Figure 3: Our domain adaptation method significantly improves high perceptual quality when closing the distribution shift performance gap. All images are U-Net reconstructions.

obtained $\mathbf{V} = \mathbf{U}$ from the training on the distribution P. For d and n large, the self-supervised loss $L_{SS}(\alpha, \mathbf{U}, \mathbf{y})$ concentrates around the expectation, which can be shown to be $(1-\alpha)^2 d + \alpha^2 d\varsigma^2$. Minimizing this over α gives $\alpha = \frac{1}{1+\varsigma^2}$. Thus, for this example, TTT yields the optimal estimator for the signal $\mathbf{x}$ under the distribution Q.

This example demonstrates that in theory, TTT can yield an optimal estimator under a distribution shift. It also illustrates that for TTT to work, the main task and the self-supervised learning task have to be related, and the choice of the self-supervised loss has to take this relation into account. Also note that this particular self-supervised loss will not work for a different distribution shift, it is tailored to the anticipated shift in noise variance. This highlights that different distribution shifts might require different self-supervised losses for TTT to work.

7. Conclusion

Distribution shifts are a key limiting factor in deep learning based imaging. In this paper, we proposed a novel domain adaptation method based on self-supervised training and test-time training (TTT) that reduces the distribution shift performance gap for four natural distribution shift by 87-99%, and thus gives significantly better image quality at inference.

It is perhaps surprising that TTT works so well for natural distribution shifts, in particular since a variety of domain adaptation methods for classification only give a marginal improvement (if at all) for natural distribution shifts Miller et al. (2021). However, image reconstruction problems are much more amenable for domain adaptation methods, since at test time, we are given an entire measurement of an image, which contains lots of information about the image and a potential distribution shift. We exploit this information through TTT and to perform early stopping during TTT. Contrary, in an image classification setup, we are not given any information about the label and thus it might be much harder to adapt at test time.

Many important questions remain: Due to the TTT, our method has significantly higher computational cost at inference than using a plain neural network without TTT. Reducing this computational cost is desirable. Moreover, our method is specific to compressive sensing problems, and thus developing TTT methods that are also applicable to inverse problems beyond compressive sensing is an important future direction. Finally, it is important to gain a better theoretical understanding of the mechanisms making TTT work.

Reproducibility

Our repository at https://github.com/MLI-lab/ttt_for_deep_learning_cs contains the code to reproduce all results in this paper.

Acknowledgment

M. Zalbagi Darestani and R. Heckel are (partially) supported by NSF award IIS-1816986, and R. Heckel acknowledges support by the Institute of Advanced Studies at the Technical University of Munich, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 456465471, 464123524. J. Liu is supported by the 6G Future Lab Bavaria.

References

Arora, S., Roeloffs, V., and Lustig, M. Untrained modified deep decoder for joint denoising parallel imaging reconstruction. In *International Society for Magnetic Resonance in Medicine Annual Meeting*, 2020.

Bostan, E., Heckel, R., Chen, M., Kellman, M., and Waller, L. Deep phase decoder: Self-calibrating phase microscopy with an untrained deep neural network. In *Optica*, pp. 559–562, 2020.

Desai, A. D., Gunel, B., Ozturkler, B. M., Beg, H., Vasanawala, S., Hargreaves, B. A., Ré, C., Pauly, J. M., and Chaudhari, A. S. Vortex: Physics-driven data augmentations for consistency training for robust accelerated MRI reconstruction. In *arXiv preprint: 2111.02549 [eess.IV]*, 2021.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. In *arXiv preprint: 1708.04552 [cs.CV]*, 2017.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, pp. 647–655, 2014.

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 (ECCV)*, pp. 184–199, 2014.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

Duchi, J. and Namkoong, H. Variance-based regularization with convex objectives. In *The Journal of Machine Learning Research*, volume 20, pp. 2450–2504, 2019.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. In *arXiv preprint: 2007.13982 [cs.LG]*, 2020.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. In *The Annals of Statistics*, volume 49, pp. 1378–1406, 2021.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, pp. 1802–1811, 2019.

Epperson, K., Sawyer, A., Lustig, M., Alley, M., and Uecker, M. Creation of fully sampled MR data repository for compressed sensing of the knee. In *Proceedings of the 22nd Annual Meeting for Section for Magnetic Resonance Technologists*, 2013.

Fabian, Z., Heckel, R., and Soltanolkotabi, M. Data augmentation for deep learning based accelerated MRI reconstruction with limited data. In *International Conference on Machine Learning (ICML)*, 2021.

Feng, C.-M., Yan, Y., Chen, G., Fu, H., Xu, Y., and Shao, L. Accelerated multi-modal MR imaging with transformers. In *arXiv preprint:2106.14248 [cs, eess]*, 2021a.

Feng, C.-M., Yan, Y., Fu, H., Chen, L., and Xu, Y. Task transformer network for joint MRI reconstruction and super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 307–317, 2021b.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2018.

Heckel, R. and Hand, P. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Heckel, R. and Soltanolkotabi, M. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. In *International Conference on Machine Learning (ICML)*, 2020.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021.

Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A., and Darrell, T. CyCADA: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 1989–1998, 2018.

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 (NeurIPS)*, volume 34, 2021.

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. Deep convolutional neural network for inverse problems in imaging. In *IEEE Transactions on Image Processing*, pp. 4509–4522, 2017.

Johnson, P. M., Jeong, G., Hammernik, K., Schlemper, J., Qin, C., Duan, J., Rueckert, D., Lee, J., Pezzotti, N., Weerdt, E. D., et al. 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 *International Workshop on Machine Learning for Medical Image Reconstruction*, pp. 25–34, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

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. In *Magnetic Resonance in Medicine*, volume 81, pp. 116–128, 2019.

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

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Korkmaz, Y., Yurt, M., Dar, S. U. H., Özbey, M., and Cukur, T. Deep MRI reconstruction with generative vision transformers. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pp. 54–64, 2021.

Korkmaz, Y., Dar, S. U., Yurt, M., Özbey, M., and Cukur, T. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. In *IEEE Transactions on Medical Imaging*, 2022.

Kornblith, S., Shlens, J., and Le, Q. V. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2661–2671, 2019.

Lin, K. and Heckel, R. Vision transformers enable fast and robust accelerated MRI. In *Medical Imaging with Deep Learning (MIDL)*, 2021.

Liu, X., Wang, J., Liu, F., and Zhou, S. K. Universal undersampled MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021a.

Liu, Y., Kothari, P., Delft, B. V., Bellot-Gurlet, B., Mordan, T., and Alahi, A. TTT++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021b.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European

Conference on Computer Vision (ECCV), pp. 181–196, 2018.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning (ICML)*, pp. 7721–7735, 2021.

Mohan, S., Vincent, J., Manzorro, R., Crozier, P., Fernandez-Granda, C., and Simoncelli, E. Adaptive denoising via GainTuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Muckley, M. J., Riemenschneider, B., Radmanesh, A., Kim, S., Jeong, G., Ko, J., Jun, Y., Shin, H., Hwang, D., Mostapha, M., et al. State-of-the-art machine learning MRI reconstruction in 2020: Results of the second fastMRI challenge. In *IEEE Transactions on Medical Imaging*, 2021.

Oh, G., Sim, B., Chung, H., Sunwoo, L., and Ye, J. C. Unpaired deep learning for accelerated MRI using optimal transport driven CycleGAN. In *IEEE Transactions on Computational Imaging*, volume 6, pp. 1285–1296, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

Rivenson, Y., Zhang, Y., Günaydın, H., Teng, D., and Ozcan, A. Phase recovery and holographic image reconstruction using deep learning in neural networks. In *Light: Science & Applications*, volume 7, pp. 17141–17150, 2018.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 806–813, 2014.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 64–73, 2020.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 9229–9248, 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 (NeurIPS)*, 2020.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9446–9454, 2018.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Wang, F., Bian, Y., Wang, H., Lyu, M., Pedrini, G., Osten, W., Barbastathis, G., and Situ, G. Phase imaging with an untrained neural network. In *Light: Science & Applications*, pp. 1–7, 2020.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., and Le, Q. V. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 819–828, 2020.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification. In *arXiv preprint: 1905.00546 [cs.CV]*, 2019.

Yaman, B., Hosseini, S. A. H., and Akcakaya, M. Zero-shot physics-guided deep learning for subject-specific MRI reconstruction. In *NeurIPS Workshop on Deep Learning and Inverse Problems*, 2021.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019.

Zalbagi Darestani, M. and Heckel, R. Accelerated MRI with un-trained neural networks. In *IEEE Transactions on Computational Imaging*, volume 7, pp. 724–733, 2021.

Zalbagi Darestani, M., Chaudhari, A. S., and Heckel, R. Measuring robustness in deep learning based compressive sensing. In *International Conference on Machine Learning (ICML)*, 2021.

Zbontar, J., Knoll, F., Sriram, A., Muckley, M. J., Bruno, M., Defazio, A., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., et al. fastMRI: An open dataset and benchmarks for accelerated MRI. In *Radiology: Artificial Intelligence*, 2020.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. In *IEEE Transactions on Image Processing*, volume 26, pp. 3142–3155, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

A. Intensity distribution changes under modality shift

One of the natural distribution shifts we study in our work is a modality shift, where the acquisition mode of train and test domains differ. This difference results in a shift in intensity distributions that is illustrated in Figure 4 for the shift from T2 to T1PRE brain images that we consider.

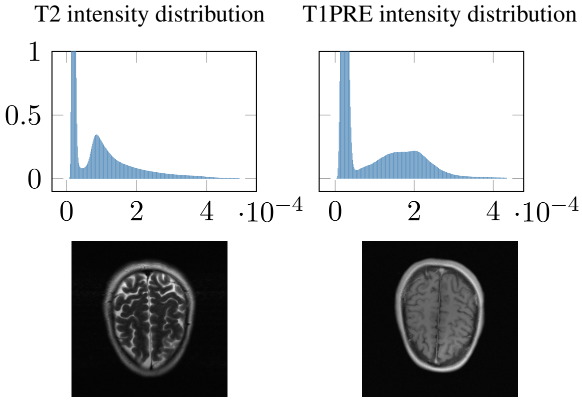

Figure 4: Modality shifts occur within an anatomy where the contrast changes via modality change, and thus the intensity distribution of the images changes.

B. Comparison to ZS-SSL

We mentioned in the introduction that an alternative test-time training (TTT) approach compared to our work is to use the method ZS-SSL (Yaman et al., 2021), which was originally introduced as a method for performing reconstruction for a single instance. Given a single under-sampled measurement $\mathbf{y}$ of an image $\mathbf{x}$, ZS-SSL creates a dataset $\{(\mathbf{y}_1, \mathbf{y}'_1, \mathbf{y}''_1), \ldots, (\mathbf{y}_K, \mathbf{y}'_K, \mathbf{y}''_K)\}$ by randomly sampling pixels from $\mathbf{y}$. Each triplet is a random partition of the under-sampled k-space and the proportions are determined by two hyper-parameters p and p'. In each triplet, $\mathbf{y}_i$ is fed to the network as input, $\mathbf{y}'_i$ is used in the TTT loss based on the network output, and $\mathbf{y}''_i$ is used for self-validation to determine when to stop training. Unlike our method that relies on no hyper-parameters, ZS-SSL has three hyper-parameters to tune which are K, the number of splits for the synthesized dataset, and p and p' which determine how to split the under-sampled measurement to three partitions

Our TTT approach incorporates a consistency-based self-supervised loss during the supervised pre-training stage, and then TTT is performed w.r.t. that self-supervised loss. Thus, the main advantage of ZS-SSL over our method is that it does not impose any constraints on the pre-training scheme. Performance-wise, Table 4 compares our method to ZS-SSL for anatomy shift. Both methods are applied to unrolled networks (our approach is applied to VarNet and ZS-SSL is applied to a similar unrolled network explained in (Yaman et al., 2021)). Table 4 shows that both achieve on-par performance in terms of closing the distribution shift performance gap. However, our approach is computationally cheaper than ZS-SSL which is expected since our method does not rely on dataset synthesis for TTT.

C. Early stopping of test-time training

We discussed in the min body that early stopping is a critical component of our test-time training (TTT) approach. Figure 5 shows that the self-validation error in the middle panel (which is computed using a fraction of under-sampled test measurement) yields a good early-stopping time, in that the validation loss is inversely correlated with the true accuracy, as expected.

	anatomy shift			
	P:knee Q:brain			
	Our approach		ZS-SSL	
setup	SSIM	TTT runtime (mins/slice)	SSIM	TTT runtime (mins/slice)
train on P test on Q + TTT	0.9358	12.9	0.9343	59.4
train on Q test on Q + TTT	0.9375	4.2	**0.9365**	33.5
train on Q test on Q	**0.9396**	-	0.9331	-
train on P test on Q	0.8802	-	0.9029	-
fraction of gap closed by TTT	97.1%		93.1%	

Table 4: **Both ZS-SSL and our test-time training (TTT) approach are highly effective for overcoming natural anatomy shift**. SSIM scores are averaged over 30 validation slices of the fastMRI brain dataset. For both methods, TTT is applied with a similar learning rate as the one used during pre-training.

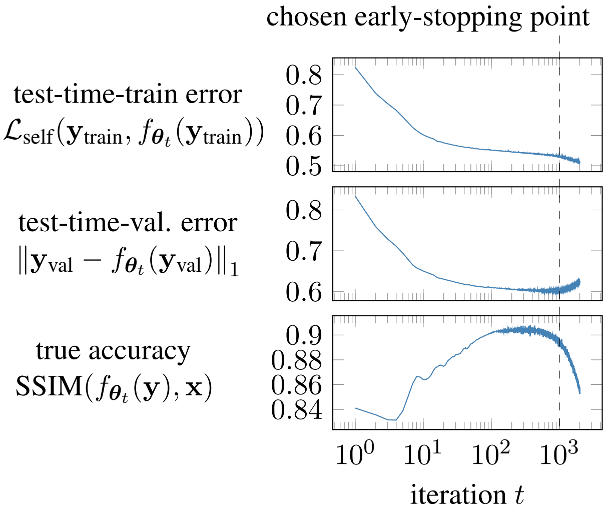

Figure 5: Test-time-training error, validation error, and error with respect to the original (unknown) image $\mathbf{x}$. The given measurement $\mathbf{y}$ is split into a part for test-time-training ($\mathbf{y}_{\text{train}}$) and into a part for validation ($\mathbf{y}_{\text{val}}$), which in turn is used to determine the early-stopping time. Comparing the validation error with the true accuracy shows that i) the validation error is a good proxy for the true accuracy, and ii) early stopping is critical for the performance of test-time training.

D. Variants of TTT

As mentioned in the main body, we also experimented with two variants of TTT, which fail to improve robustness. Those variants are discussed here.

Variant 1. The first variant is very similar to how TTT is typically performed in image classification. Let $E_{\boldsymbol{\beta}}$ be the encoder of the U-Net and let $D_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{\vartheta}}$ be two decoders. The encoder with the first decoder is trained to solve the main task which is the supervised compressed sensing task, and the encoder with the second decoder is trained to solve an auxiliary task which we take as a self-supervised denoising task. Specifically, we train the method on the loss:

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n} \left(\underbrace{\frac{\left\|\mathbf{x}_i - E_{\boldsymbol{\beta}}(D_{\boldsymbol{\theta}}(\mathbf{A}^\dagger\mathbf{y}_i))\right\|_1}{\|\mathbf{x}_i\|_1}}_{\mathcal{L}_{\text{sup}}} + \underbrace{\frac{\left\|\mathbf{A}^\dagger\mathbf{y}_i - \mathbf{A}E_{\boldsymbol{\beta}}(D_{\boldsymbol{\vartheta}}(\mathbf{A}^\dagger\mathbf{y}_i + \mathbf{z}_i))\right\|_1}{\|\mathbf{A}^\dagger\mathbf{y}_i\|_1}}_{\mathcal{L}_{\text{self}}} \right), \tag{4}$$

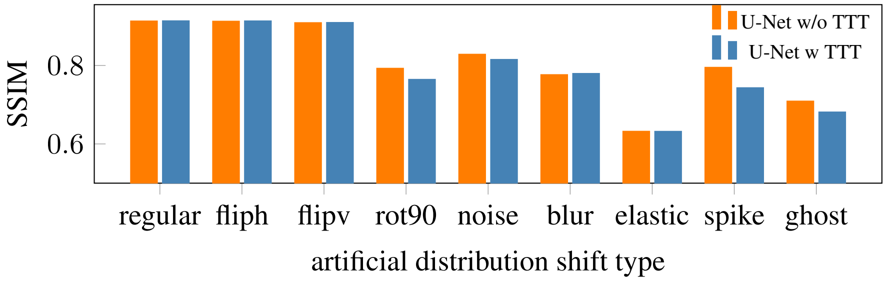

Figure 6: **A variant of test-time training (TTT) with a two-head U-Net inspired from image classification does not improve model robustness.** Including denoising as an auxiliary task during supervised training of U-Net even impairs model robustness for some artificial distribution shifts. Evaluation scores when each transformation on the x axis is applied to the brain validation set.

where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$ is random Gaussian noise that we generated for the self-supervised loss, and $\mathbf{A}^\dagger \mathbf{y}$ is the input of the U-Net (i.e., the least-squares reconstruction from measurement $\mathbf{y}$). Similar to our domain adaptation method, at the inference, we optimize the self-supervised loss $\mathcal{L}_{\text{self}}(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \mathbf{y})$ with respect to the weights $\boldsymbol{\beta}, \boldsymbol{\vartheta}$ for a given undersampled measurement $\mathbf{y}$. The stopping iteration was set to 10 heuristically (as opposed to our domain adaptation method where we used a self-validation-based automatic early stopping).

Figure 6 shows the results when training U-Net with loss function (4) in comparison to a U-Net trained in a supervised manner. Because we observed no benefit from this type of TTT under a natural anatomical shift (i.e., 0% of the gap is closed), in Figure 6, we illustrate the result for artificial distribution shifts. We specifically evaluate this variant on the brain validation set of fastMRI under (1) no transformation (regular), (2) horizontal flipping (fliph), (3) vertical flipping (flipv), (4) 90-degree rotation (rot90), (5) Gaussian noise (noisy), (6) blur artifacts (blur), (7) elastic transformation (elastic), (8) spike artifacts (spike), and (9) ghosting artifacts (ghost). Interestingly, this variant of TTT is not even helpful with these artificial shifts according to SSIM scores reported in Figure 6.

Variant 2. The second variant is a method we propose based on CycleGAN (Zhu et al., 2017), dubbed CycleU-Net, and works as follows. Suppose we put two U-Nets $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\beta}}$ in tandem to form a larger model. We then train the resulting model by making two forward passes for each input pair $(\mathbf{y}_i, \mathbf{x}_i)$ at each epoch: (1) a self-supervised pass as $g_{\boldsymbol{\beta}}(f_\theta(\mathbf{A}^\dagger \mathbf{y}_i))$, and (2) a supervised forward pass as $f_{\boldsymbol{\theta}}(g_{\boldsymbol{\beta}}(\mathbf{x}_i))$. This is illustrated in Figure 7.

By defining those two forward passes, we can build the training loss function as follows:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \Bigg(\left\| \mathbf{x}_i - f_{\boldsymbol{\theta}}(g_{\boldsymbol{\beta}}(\mathbf{x}_i)) \right\|_1 + \left\| \mathbf{A}^\dagger \mathbf{y}_i - g_{\boldsymbol{\beta}}(f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y}_i)) \right\|_1 + \left\| \mathbf{x}_i - f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y}_i) \right\|_1 + \left\| \mathbf{A}^\dagger \mathbf{y}_i - g_{\boldsymbol{\beta}}(\mathbf{x}_i) \right\|_1 \Bigg).$$

Here, the first two terms enforce input-output equality for each of the two forward passes. The third term ensures that $f_{\boldsymbol{\theta}}$ learns a mapping from the under-sampled to the ground-truth domain (likewise, the fourth term ensures that $g_{\boldsymbol{\beta}}$ learns a mapping from the ground-truth to the under-sampled domain). Note that without the last two terms, there is no guarantee that $f_{\boldsymbol{\theta}}$ reconstructs the ground-truth image from the under-sampled measurement.

At inference, we perform TTT w.r.t $\left\| \mathbf{A}^\dagger \mathbf{y}_i - g_{\boldsymbol{\beta}}(f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y}_i)) \right\|_1$ (both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are optimized) which is fully self-supervised, then detach $f_{\boldsymbol{\theta}}$ from the architecture and use it for reconstruction as $f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y})$.

Table 5 shows the performance of this approach for anatomy shift (the training and test data are the same as

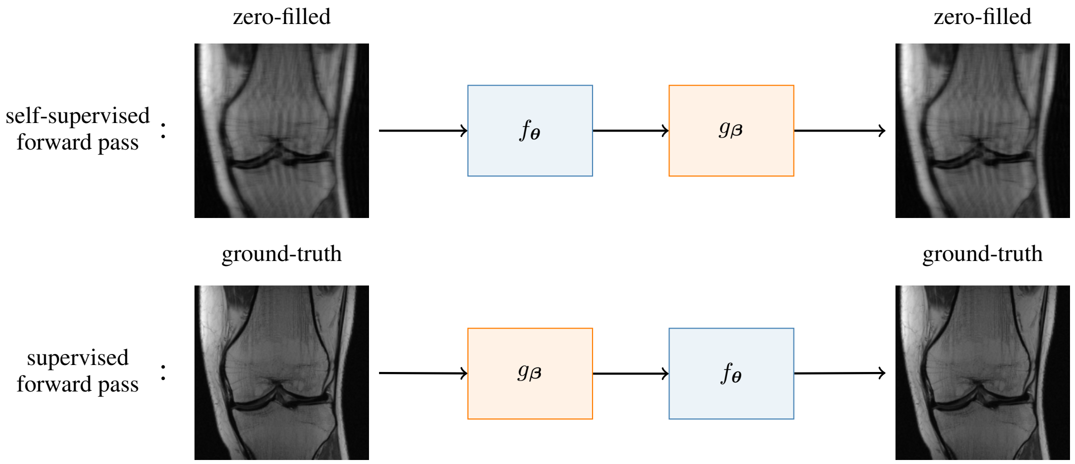

Figure 7: **Forward passes in CycleU-Net.** f_ϑ and g_β are two U-Nets with the same number of parameters. Training based on the two forward passes enables test-time training (TTT) w.r.t. the self-supervised pass. Only f_θ is needed for inference after TTT.

setup	P: knee Q: brain
train on Q test on Q	0.9187
train on P test on Q	0.8521
distribution shift performance gap	0.0666
train on Q test on Q + TTT	0.9212
train on P test on Q + TTT	0.8742
distribution shift performance gap	0.0470
fraction of gap closed by TTT	29.4%

Table 5: **A variant of test-time training (TTT) with a CycleU-Net inspired from CycleGAN improves model robustness only slightly.** The first three rows are SSIM scores for U-Net when trained in a supervised manner. The second three rows are SSIM scores for CycleU-Net when TTT is applied at the inference. This variant closes the gap by 29.4% but is outperformed by our original method which closes the gap by 98.6% which is discussed in the main body.

Section 5). As shown, the fraction of gap closed by performing TTT on CycleU-Net is 29.4% which demonstrates that this approach is not effective in closing the gap.

E. Relation to imaging with un-trained neural networks

Our domain adaptation method consists of training a network with supervised and self-supervised loss and at inference, training again on the self-supervised loss with early stopping.

The inference step is very similar to how an un-trained neural network is used for image recovery. To see this, reconstruction of a signal from an observation with an un-trained network works as follows. Let f_θ be a convolutional network that is initialized randomly, and optimized on the loss

$$\mathcal{L}(\boldsymbol{\theta}) = \left\|\mathbf{y} - \mathbf{A} f_{\boldsymbol{\theta}}(\mathbf{z})\right\|_1,$$

with gradient descent and early-stopping the iterates for regularization. Here, $\mathbf{z}$ is an input that is typically random, and has observed to be relatively irrelevant. Ulyanov et al. (2018) demonstrated that this method works well for denoising and super-resolution, and Heckel & Hand (2019); Arora et al. (2020); Zalbagi Darestani &

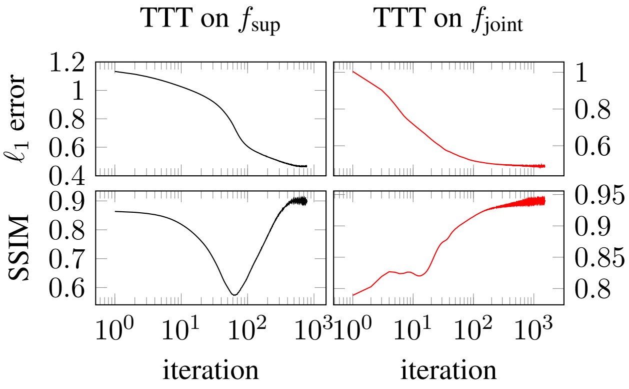

Figure 8: Test-time training (TTT) improves the reconstruction accuracy of a pre-trained model when self-supervision is included during training (right column), but treats a pre-trained model as an un-trained network when pre-training is fully supervised (left column). The first row shows the self-validation error which is used to stop TTT early. The second row shows SSIM w.r.t. the ground truth image during TTT.

Heckel (2021) have shown that un-trained convolutional networks work well for denoising and accelerated MRI. Reconstruction with an un-trained network is very similar to our TTT step, with the difference that the un-trained networks start from a random initialization of the weights (which in this case, for the DIP-based model introduced in (Zalbagi Darestani & Heckel, 2021), they achieve 0.9046 SSIM on brain images for the anatomy shift setup we consider in Table 2).

We could also just initialize the weights of an un-trained network by pre-training the network on a dataset, and we might expect that this improves performance. Viewed from that angle, our method might look like an un-trained network in disguise.

In this section, we argue that our method is not an un-trained network in disguise by demonstrating that TTT improves over the image prior learned by pre-training. We perform the following experiment. We consider the anatomy shift and train a U-Net f_{sup} in a supervised manner and another U-net f_{joint} using our joint loss function (2) that incorporates self-supervision. We then perform TTT by applying f_{sup} and f_{joint} to a brain test sample.

Figure 8 depicts the results. The first row shows the self-validation error whose increase determines the early-stopping point. The second row depicts the SSIM with respect to the unknown ground truth image during TTT. Observe that when TTT is applied to f_{joint}, SSIM improves during TTT. Thus TTT improves over the learned prior from knees to achieve a good reconstruction accuracy on the brain test sample.

Contrary, when TTT is applied to f_{sup}, SSIM first decreases dramatically and then starts to rise again. This suggests that TTT applied to a fully self-supervised loss ignores the learned prior and uses the model as an un-trained network.

F. Proof of claims in Section 6: Test-time training can provably adapt to a distribution shift

Proposition F.1. *The supervised loss*

$$R_P(\alpha, \mathbf{V}) = \mathbb{E}_P \left[\left\| \mathbf{x} - \alpha \mathbf{V}\mathbf{V}^T \mathbf{y} \right\|_2^2 \right]$$

is minimized by $\alpha = 1/(1+\sigma^2)$ *and* $\mathbf{V} = \mathbf{U}$.

Proof. Let $\mathbf{W} = \alpha \mathbf{V}\mathbf{V}^T$ and consider the following related convex optimization problem

$$\min_{\mathbf{W}} \mathbb{E}_P \left[\|\mathbf{x} - \mathbf{W}\mathbf{y}\|_2^2 \right].$$

Note that here we optimizer over a larger space, as we do not constrain $\mathbf{W}$ to be symmetric. We show that a solution to the optimization problem is $\mathbf{W} = \frac{1}{1+\sigma^2}\mathbf{U}\mathbf{U}^T$, therefore $(\alpha, \mathbf{U})$ is a minimizer of $R_P(\alpha, \mathbf{V})$, which proves the claim.

The gradient of the objective function above is

$$\begin{aligned}\nabla_{\mathbf{W}} \mathbb{E}_P \left[\|\mathbf{x} - \mathbf{W}\mathbf{y}\|_2^2 \right] &= \nabla_{\mathbf{W}} \mathrm{tr} \left(\mathbb{E}_P \left[\mathbf{x}\mathbf{x}^T \right] - 2\mathbf{W}^T \mathbb{E}_P \left[\mathbf{x}\mathbf{y}^T \right] + \alpha^2 \mathbf{W}^T \mathbf{W} \mathbb{E}_P \left[\mathbf{y}\mathbf{y}^T \right] \right) \\ &= 2\mathbf{W}\mathbb{E}_P \left[\mathbf{y}\mathbf{y}^T \right] - 2\mathbb{E}_P \left[\mathbf{x}\mathbf{y}^T \right].\end{aligned}$$

Setting the gradient to zero, the minimizer satisfied

$$\mathbf{W} = \mathbb{E}_P \left[\mathbf{x}\mathbf{y}^T \right] \left(\mathbb{E}_P \left[\mathbf{y}\mathbf{y}^T \right] \right)^{-1}.$$

Since $\mathbf{y} = \mathbf{x} + \mathbf{z}$ and $\mathbf{x} = \mathbf{U}\mathbf{c}$, where $\mathbf{c} \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$ are independent, it holds that

$$\mathbb{E}_P \left[\mathbf{x}\mathbf{y}^T \right] = \mathbb{E}_P \left[\mathbf{x}\mathbf{x}^T \right] = \mathbf{U}\mathbf{U}^T,$$

and

$$\mathbb{E}_P \left[\mathbf{y}\mathbf{y}^T \right] = \mathbb{E}_P \left[\mathbf{x}\mathbf{x}^T + \mathbf{z}\mathbf{z}^T \right] = \mathbf{U}\mathbf{U}^T + \sigma^2 \mathbf{I}.$$

Plugging in the two expressions into the expression for $\mathbf{W}$ we obtain

$$\mathbf{W} = \frac{1}{1+\sigma^2}\mathbf{U}\mathbf{U}^T,$$

as desired. □

Proposition F.2. *For fixed* $\mathbf{U}$*, the expectation of the self-supervised loss*

$$\mathbb{E}_Q \left[L_{SS}(\alpha, \mathbf{U}, \mathbf{y}) \right] = \mathbb{E}_Q \left[\left\| \mathbf{y} - \alpha \mathbf{U}\mathbf{U}^T \mathbf{y} \right\|_2^2 + \frac{2\alpha d}{n-d} \left\| (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{y} \right\|_2^2 \right]$$

is minimized by $\alpha = 1/(1+\varsigma^2)$.

Proof. Note that

$$\begin{aligned}\mathbb{E}_Q \left[\left\| \mathbf{y} - \alpha \mathbf{U}\mathbf{U}^T \mathbf{y} \right\|_2^2 \right] &= \mathrm{tr} \left(\mathbb{E}_Q \left[\mathbf{y}\mathbf{y}^T \right] - 2\alpha \mathbf{U}^T \mathbf{U} \mathbb{E}_Q \left[\mathbf{y}\mathbf{y}^T \right] + \alpha^2 \mathbf{U}^T \mathbf{U} \mathbb{E}_Q \left[\mathbf{y}\mathbf{y}^T \right] \right) \\ &= \mathrm{tr} \left(\left(\mathbf{I} + (\alpha^2 - 2\alpha)\mathbf{U}^T \mathbf{U} \right) \mathbb{E}_Q \left[\mathbf{y}\mathbf{y}^T \right] \right) \\ &= \mathrm{tr} \left(\left(\mathbf{I} + (\alpha^2 - 2\alpha)\mathbf{U}^T \mathbf{U} \right) \left(\mathbf{U}\mathbf{U}^T + \varsigma^2 \mathbf{I} \right) \right) \\ &= \mathrm{tr} \left(\varsigma^2 \mathbf{I} + ((1-\alpha)^2 + (\alpha^2 - 2\alpha)\varsigma^2)\mathbf{U}\mathbf{U}^T \right) \\ &= \varsigma^2 n + (1-\alpha)^2 d + (\alpha^2 - 2\alpha)\varsigma^2 d.\end{aligned}$$

It follows that, for $\alpha = 1$,

$$\mathbb{E}_Q \left[\left\| \mathbf{y} - \mathbf{U}\mathbf{U}^T \mathbf{y} \right\|_2^2 \right] = \varsigma^2 n - \varsigma^2 d.$$

Hence,

$$\begin{aligned}\mathbb{E}_Q \left[L_{SS}(\alpha, \mathbf{U}, \mathbf{y}) \right] &= \varsigma^2 n + (1-\alpha)^2 d + (\alpha^2 - 2\alpha)\varsigma^2 d + \frac{2\alpha d}{n-d}(\varsigma^2 n - \varsigma^2 d) \\ &= \varsigma^2 n + (1-\alpha)^2 d + \alpha^2 \varsigma^2 d,\end{aligned}$$

whose minimum is achieved at $\alpha = 1/(1+\varsigma^2)$. To see this, take the derivative with respect to α, set it to zero, and solve for α. □

G. Test-time training for non-convolutional architectures

The two networks that we studied throughout are based on the U-Net, a convolutional neural network. There are, however, other non-convolutions neural network architectures that perform well for image reconstruction problems. In this section we explore how our test-time training (TTT) approach performs with non-convolutional networks.

We consider the Vision Transformer (ViT) (Dosovitskiy et al., 2020) that works very well for signal reconstruction problems. ViT has been tailored to accelerated MRI reconstruction as well (Feng et al., 2021a;b; Korkmaz et al., 2021; 2022), and has been shown to be computationally faster than U-Net and also slightly more robust than U-Net against anatomy shift (Lin & Heckel, 2021).

We repeated the same experiment we performed for U-Net under the anatomy shift. The results, reported in Table 6 show that TTT for a ViT is effective, but not as effective as for the U-Net (specifically, the fraction of the gap closed by TTT is only 84.5% for ViT, whereas the gap closed by TTT is 98.6% for U-Net). This is also reflected in the visual quality of the images, as illustrated in Figure 9. In Figure 9 we see that TTT for a ViT gives reconstruction artifacts.

The self-supervised loss function used in our TTT approach works better for U-Net compared to ViT, and hence our TTT is more effective for U-Net. To see this, we perform the following experiment. We train U-Net and ViT on the knee training set of fastMRI in a fully self-supervised manner, i.e., we minimize the training loss

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\left\| \mathbf{y}_i - \mathbf{A} f_{\boldsymbol{\theta}}(\mathbf{A}^{\dagger}\mathbf{y}_i) \right\|_1}{\left\| \mathbf{y}_i \right\|_1},$$

on a set of training measurements $\mathbf{y}_1, \ldots, \mathbf{y}_n$, for both U-net and ViT. During training, we monitor true accuracy, by comparing the reconstructions generated by U-Net and ViT to the (unknown) ground-truth images $\mathbf{x}_1, \ldots, \mathbf{x}_n$ associated with the measurements $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Figure 10 shows that towards convergence, there is a constant gap between the true accuracy achieved by U-Net and ViT. This demonstrates that our self-supervised loss works better with U-Net than ViT in terms of the quality of the learned prior for the images.

setup	P: knee Q: brain
train on Q test on Q	0.9041
train on P test on Q	0.8429
distribution shift performance gap	0.0612
train on Q test on Q + TTT	0.8947
train on P test on Q + TTT	0.8859
distribution shift performance gap	0.0095
fraction of gap closed by TTT	84.5%

Table 6: **For ViT, using self-supervision with test-time training (TTT) closes 84% of the distribution shift performance gap for anatomy shift.** The first three rows are SSIM scores for ViT when trained in a supervised manner. The second three rows are SSIM scores for ViT when self-supervision is included during training and then TTT is applied at the inference.

H. More reconstruction examples

The results of Table 2 demonstrate that our domain adaptation method closes the distribution shift performance gap for anatomy, dataset, modality, and acceleration shifts by about 90%. Figure 3 in the main body shows example images reconstructed with U-Net to demonstrate that the perceptual quality is improved after applying our domain adaptation method.

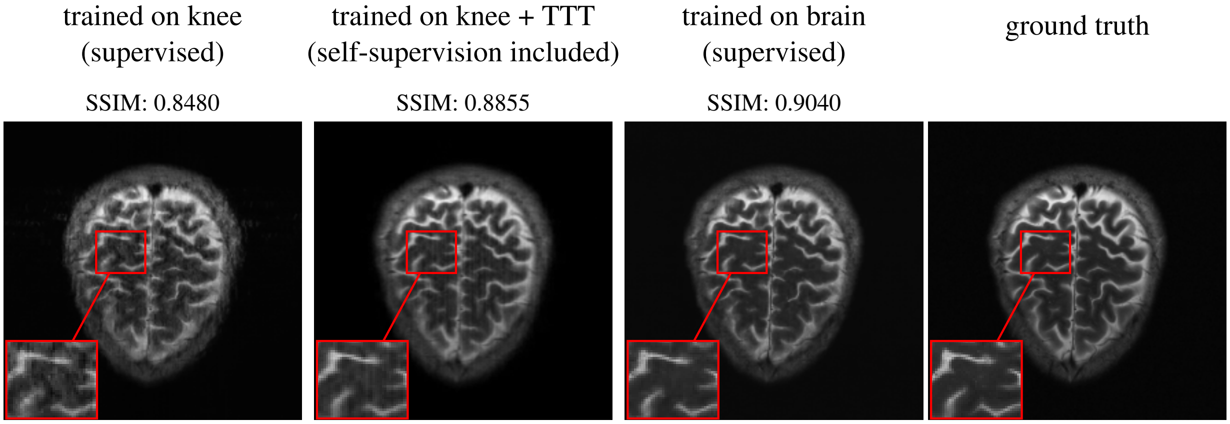

Figure 9: Including self-supervision while training DL models (a Vision Transformer in this case) combined with TTT improves model robustness to natural anatomy shifts. The sample belongs to the fastMRI brain validation dataset.

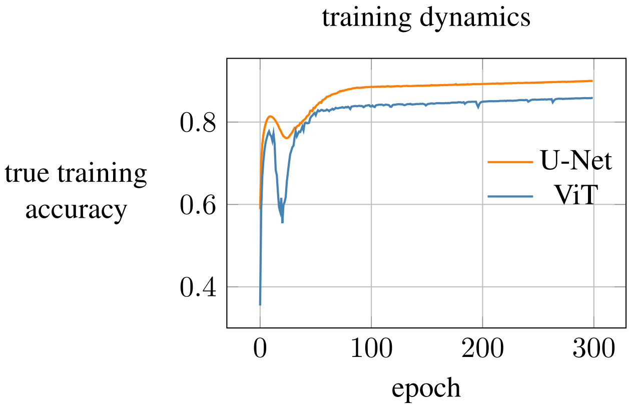

Figure 10: The inductive bias of self-supervised training of U-Net results in a higher true accuracy at convergence than ViT. True accuracy is monitored using the ground-truth data during the self-supervised training.

In this section, we provide more detailed illustrations for both U-Net and VarNet under each distribution shift. Figure 11, Figure 12, Figure 13, and Figure 14 provide reconstruction examples for anatomy, dataset, modality, and acceleration shifts before and after our domain adaptation method. As shown in the figures, the perceptual quality of the reconstructions improve noticeably with test-time training.

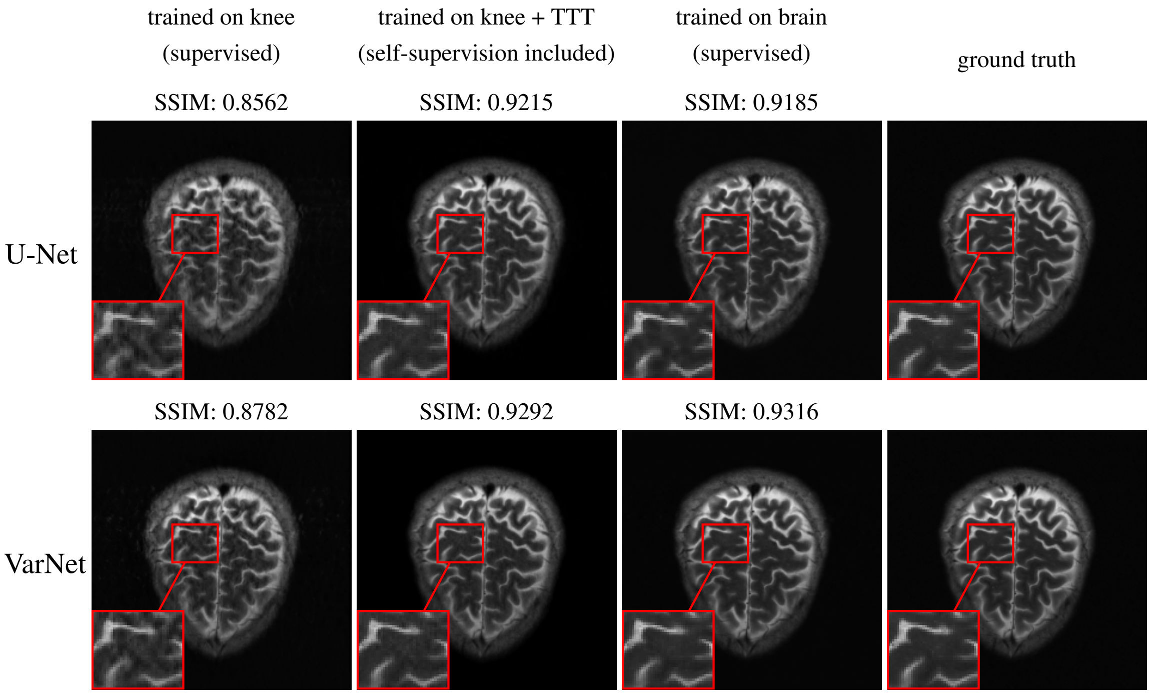

Figure 11: Including self-supervision while training DL models combined with TTT improves model robustness to natural anatomy shifts. The sample belongs to the fastMRI brain validation dataset.

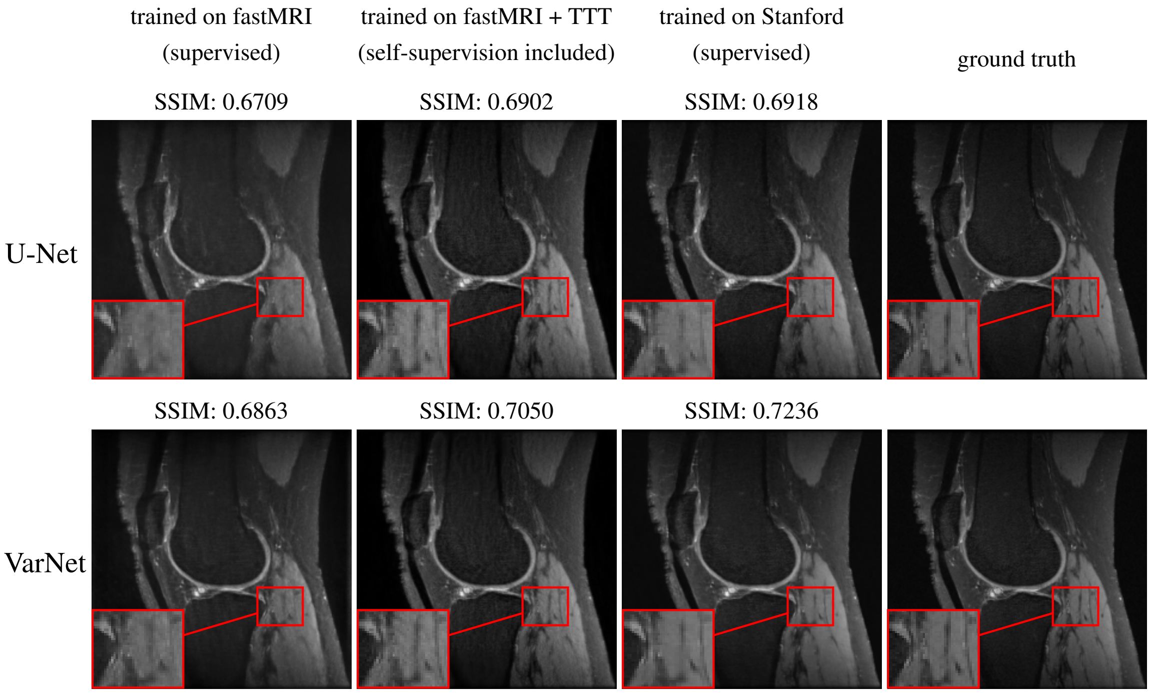

Figure 12: Including self-supervision while training DL models combined with TTT improves model robustness to natural dataset shifts. The sample belongs to the Stanford validation dataset and the pointed region reveals how each setup shines or fails at reconstruction.

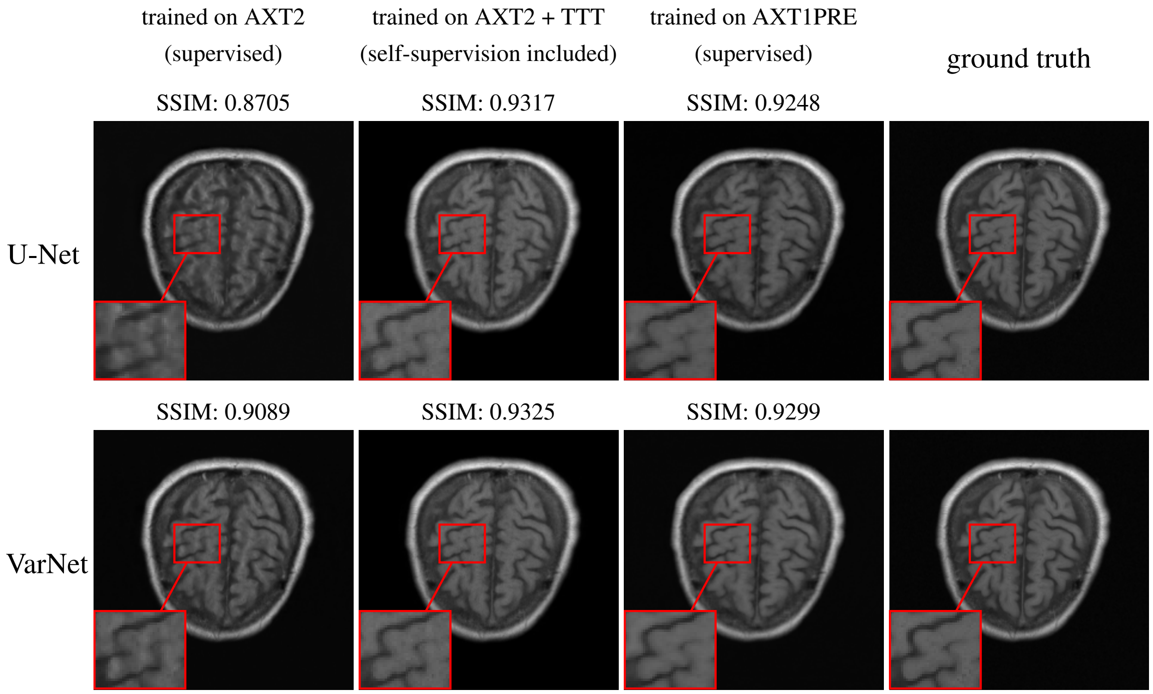

Figure 13: Including self-supervision while training DL models combined with TTT improves model robustness to natural modality shifts. The AXT1PRE sample belongs to the fastMRI brain validation dataset.

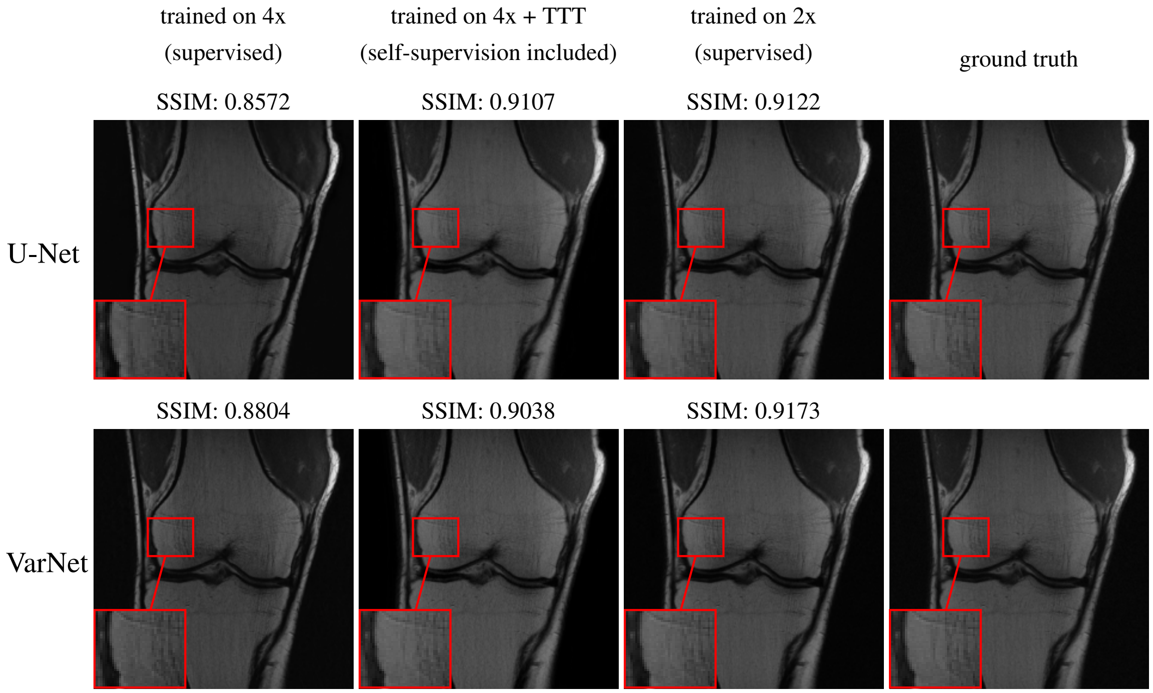

Figure 14: Including self-supervision while training DL models combined with TTT improves model robustness to natural acceleration shifts. The sample belongs to the fastMRI knee validation dataset and the pointed region reveals how each setup shines or fails at reconstruction. VarNet, unlike U-Net, does not include a region of artifact and the overall contrast of the image has changed under the shift.