VORTEX: PHYSICS-DRIVEN DATA AUGMENTATIONS FOR CONSISTENCY TRAINING FOR ROBUST ACCELERATED MRI RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks have enabled improved image quality and fast inference times for various inverse problems, including accelerated magnetic resonance imaging (MRI) reconstruction. However, such models require large amounts of fully-sampled ground truth data, which are difficult to curate and are sensitive to distribution drifts. In this work, we propose applying *physics-driven* data augmentations for consistency training that leverage our domain knowledge of the forward MRI data acquisition process and MRI physics for improved data efficiency and robustness to clinically-relevant distribution drifts. Our approach, termed VORTEX, (1) demonstrates strong improvements over supervised baselines with and without augmentation in robustness to signal-to-noise ratio change and motion corruption in data-limited regimes; (2) considerably outperforms state-of-the-art data augmentation techniques that are purely image-based on both in-distribution and out-of-distribution data; and (3) enables composing heterogeneous image-based and physics-driven augmentations.

1 INTRODUCTION

Magnetic resonance imaging (MRI) is a powerful medical imaging modality that enables noninvasive visualization of anatomy and is a cornerstone for disease diagnostics. However, acquiring clinical MRI data typically requires long scan durations (30+ minutes). To reduce these durations, MRI data acquisition can be accelerated by undersampling the requisite spatial frequency measurements, referred to as *k-space* measurements. Reconstructing these undersampled images without aliasing artifacts from k-space measurements that are subsampled below the Nyquist rate – the minimum sampling rate that fully describes a given signal – is an ill-posed problem in the Hadamard sense (Hadamard). To address this challenge, previous methods utilized underlying image priors to constrain the optimization – most notably enforcing sparsity in a transformation domain, in a process called compressed sensing (Lustig et al., 2008). Nevertheless, these methods suffer from long reconstruction times and can require parameter-specific tuning (Lustig et al., 2007; Akasaka et al., 2016).

Deep learning (DL) based accelerated MRI reconstruction methods have recently enabled higher acceleration factors compared to traditional methods with fast reconstruction times and improved image quality (Hammernik et al., 2018; Sandino et al., 2020a). However, these approaches rely on large amounts of paired undersampled and fully-sampled reference data for training, which is often costly or simply impossible to acquire in many imaging applications. Methods that achieve state-of-the-art reconstruction performance still use large fully-sampled (supervised) datasets, with only a handful of methods exploring approaches such as leveraging prospectively undersampled (unsupervised) data (Chaudhari et al., 2021) or using image-based data augmentation schemes (Fabian et al., 2021) to mitigate data paucity. Perhaps more concerning is that even some of the best DL-based MR reconstruction methods are highly sensitive to clinically-relevant distribution drifts such as scanner-induced drifts, patient-induced artifacts, anatomical changes, and forward model changes (Darestani et al., 2021). Sensitivity to distribution drifts remains largely unexplored, with only a few studies that have proposed solutions for simple forward model alterations such as undersampling mask change at inference time (Gilton et al., 2021). Addressing sensitivities to clinically-relevant distribution shifts is necessary to deploy DL reconstruction models clinically with confidence.

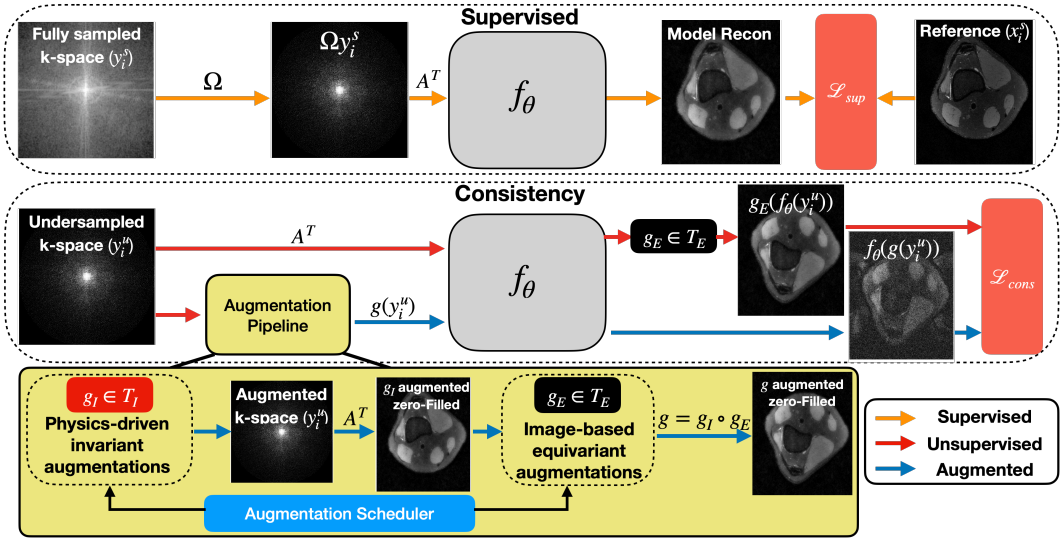

Figure 1: VORTEX uses supervised and consistency paths for robust accelerated MRI reconstruction.

In this work, we demonstrate that leveraging domain knowledge of the forward MRI data acquisition process and MRI physics through *physics-driven, acquisition-based* data augmentations for consistency training enables building data-efficient networks that are robust to clinically-relevant distribution drifts such as signal-to-noise ratio (SNR) and motion artifacts. Our proposal builds on the consistency framework Noise2Recon that conducts joint reconstruction for supervised scans and denoising for unsupervised scans (Desai et al., 2021a) by replacing the original consistency denoising objective with a data augmentation pipeline. Specifically, we propose a semi-supervised consistency training framework (described in Figure 1), termed VORTEX, that uses a data augmentation pipeline to enforce invariance to *physics-driven* data augmentations of noise and motion, and equivariance to image-based data augmentations of flipping, scaling, rotation, translation, and shearing. VORTEX allows for curriculum learning based on the difficulty of physics-driven augmentations, and composing heterogeneous augmentations. This leads to robustness to different families of perturbations at inference time without decreasing the reconstruction performance on non-perturbed, in-distribution data. We show that VORTEX outperforms the state-of-the-art data augmentation scheme, MRAugment, (Fabian et al., 2021) which solely relies on image-based data augmentations, on both in-distribution data and simulated out-of-distribution (OOD) data. Additionally while MRAugment is constrained to image augmentations to preserve noise statistics in the training data, we demonstrate that VORTEX can relax this constraint and operate on a broad family of augmentations, including acquisition-based augmentations, which inherently change the noise statistics of the training data. Our contributions in this work include the following:

- We propose VORTEX, a semi-supervised consistency training framework for accelerated MRI reconstruction that enables composing image-based data augmentations with *physics-driven* data augmentations, which leverage our knowledge of both MRI physics and the forward model of the MRI data acquisition process. We show that VORTEX improves data-efficiency and robustness.
- We demonstrate strong improvements over supervised baselines in robustness to clinically-relevant distribution drifts including scanner-induced SNR change and patient-induced motion artifacts. Notably, we obtain +10.6 structural similarity (SSIM) and +5.3 complex PSNR (cPSNR) improvement over supervised baselines on heavily motion-corrupted scans in label-scarce regimes.
- We improve over state-of-the-art data augmentation techniques for MRI reconstruction that are purely image-based (Fabian et al., 2021). We achieve +6.1 SSIM and +0.2 cPSNR improvements on in-distribution data, +12.5 SSIM and +7.8 cPSNR improvement on motion-corrupted data, and +8.9 SSIM and +2.5 cPSNR improvement on noise-corrupted data.
- We conduct ablations comparing pixel-based and latent space consistency during training and designing curricula for data augmentation difficulty.

Our code and all experimental configurations are publicly available at (blinded).

2 Related Work

Supervised accelerated MRI reconstruction methods map zero-filled images obtained from under-sampled measurements to fully-sampled ground truth images using fully-convolutional networks (e.g. U-Net (Ronneberger et al., 2015)) or unrolled networks modeling iterative proximal-update optimization methods (Adler & Öktem, 2018; Aggarwal et al., 2019; Sandino et al., 2020a). Such approaches rely on a large corpus of fully-sampled scans. Although lagging in performance with supervised approaches, prior proposals (Chaudhari et al., 2021) have leveraged unsupervised data including using generative adversarial networks (Lei et al., 2021; Cole et al., 2020), self-supervised learning (Yaman et al., 2020), and dictionary learning (Lahiri et al., 2021). Fabian et al. (2021) proposed an image-based data augmentation scheme to reduce dependence on supervised training data. Several methods have explored building neural networks robust to distribution drifts for image classification (Taori et al., 2020; Recht et al., 2019; Goel et al., 2020) and natural language processing tasks (Miller et al., 2020; Gunel et al., 2021). For accelerated MRI reconstruction, Darestani et al. (2021) recently demonstrated trained deep neural networks, un-trained networks (Darestani & Heckel, 2021), and traditional iterative approaches are sensitive to adversarial perturbations and distribution drifts. However, unlike what we propose with consistency training, the authors did not explore how to build robust neural networks to the discussed distribution drifts.

Consistency training was first proposed to include a form of denoising objective where the model is trained to be invariant to noisy input examples (Miyato et al., 2019; Sajjadi et al., 2016; Clark et al., 2018) or hidden representations (Bachman et al., 2014; Laine & Aila, 2017). These methods primarily differed in the type of noise injection applied, including additive Gaussian noise, dropout noise, and adversarial noise. Desai et al. (2021a) extended these methods to a consistency training framework that performs joint MRI image reconstruction and denoising, where noise is applied to undersampled k-space as additive complex-valued Gaussian noise. Compared to denoising-based consistency, Xie et al. (2020) showed that using semantic-preserving data augmentation consistency (RandAugment (Cubuk et al., 2020) for image tasks and back-translation for language tasks (Edunov et al., 2018)) led to significant performance boosts. Chen et al. (2021) proposed an adversarial data augmentation model that consists a set of photometric and geometric image transformations which gets jointly optimized with a segmentation network during training and evaluated on cardiac and prostate segmentation tasks. Motion correction for MRI is an active research area, as scans corrupted by patient motion affect the diagnostic image quality and clinical outcomes (Chavhan et al., 2013; Barker, 2000). Pawar et al. (2019) proposes a supervised DL method that learns to map simulated motion-corrupted scans to clean scans as a post-processing method after reconstruction. Liu et al. (2020) extends iterative application of image denoisers as imaging priors Romano et al. (2017) for general artifact removal such as motion correction. Gan et al. (2021) extends this method by training the model in the measurement domain without supervised data. However, these methods require multiple measurements of the same object undergoing nonrigid deformation which is unrealistic in most clinical settings. Shaw et al. (2020) generates realistic patient motion artefacts and uses them as an augmentation method to train robust semantic segmentation methods.

An extended discussion on related work is available in Appendix B.

3 Background and Preliminaries

3.1 Accelerated Multi-coil MRI Reconstruction

In MRI, measurements are acquired in the spatial frequency domain, referred to as *k-space*. In this work, we consider the case of clinically-relevant accelerated multi-coil MRI acquisition where multiple receiver coils are used to acquire spatially-localized k-space measurements modulated by corresponding *sensitivity maps*. Sensitivity maps are generally unknown and vary per patient, and thus, need to be estimated to perform reconstruction (Pruessmann et al., 1999). In accelerated MRI reconstruction, the goal is to reduce scan times by decreasing the number of samples acquired in k-space. The undersampling operation can be represented by a binary mask Ω that indexes acquired samples in k-space. The forward problem for multi-coil accelerated MRI can be written as:

$$y = \Omega \boldsymbol{F} \boldsymbol{S} x^* + \epsilon = A(x^*) + \epsilon$$

where y is the measured signal in k-space, $\boldsymbol{F}$ is the discrete Fourier transform matrix, $\boldsymbol{S}$ is the receiver coil sensitivity maps, x^* is the ground-truth signal in image-space, and ϵ is additive complex

Gaussian noise. $A = \Omega \boldsymbol{F} \boldsymbol{S}$ is the known forward operator during acquisition (see Appendix A for notation). Note that this problem is ill-posed in the Hadamard sense (Hadamard) as we have fewer measurements than variables to recover. It does not satisfy the three conditions of 1) existence of a solution, 2) uniqueness, and 3) continuous dependence on measurements to be defined as well-posed. This makes recovering the underlying image x^* impossible to recover uniquely without an assumption such as sparsity in some transformation domain as in compressed sensing. (Lustig et al., 2008).

3.2 Definitions

Equivariance. We simplify the precise definition of equivariance that requires group theory (Celledoni et al., 2021) to denote $f_\theta(t(x)) = t(f_\theta(x))$ for all $t \in T$ where T is the set of data augmentation transformations. Intuitively, if a trained model f_θ is equivariant to a transformation t, then the transformation of the input directly corresponds to the transformation of the model output.

Invariance. Similarly, we simplify the definition of invariance to $f_\theta(t(x)) = f_\theta(x)$ for all $t \in T$ where T is the the set of transformations we use for data augmentation. Intuitively, f_θ is invariant if its output does not change upon applying transformation t to the input. Details on how these definitions motivate the structure of augmentations in VORTEX are provided in Appendix C.

4 Methods

We propose VORTEX, a semi-supervised consistency training framework that integrates a generalized data augmentation pipeline for accelerated MRI reconstruction (Fig. 1). We consider the setup with dataset $\mathcal{D}$ that consists of (1) fully-sampled examples in k-space $y^{(s)}$ with corresponding supervised reference ground truth images $x^{(s)}$, and (2) prospectively undersampled examples in k-space $y^{(u)}$ without supervised references. f_θ is the learned reconstruction model with the forward model A. A pixel-wise ℓ_1 supervised loss $\mathcal{L}_{sup}$ is computed for examples with supervised references for $y^{(s)}$. Undersampled examples $y^{(u)}$ are passed through the *Augmentation Pipeline* (see §4.1 for details). We consider the case where there are considerably more unsupervised examples than supervised examples, which is often observed in clinical practice.

Let T_I be the set of transformations we want to be invariant to that consists of the physics-driven data augmentations such as additive complex Gaussian noise and motion corruption (see §4.1.1). Similarly, let set T_E denote the transformations we want to be equivariant to that includes the image-based data augmentations such as flipping, rotation, translation, scaling, and shearing (see §4.1.2). A pixel-wise l_1 consistency loss $\mathcal{L}_{cons}$ is computed between the model outputs of input undersampled examples with and without augmentation, which is formulated differently based on whether we would like to be invariant or equivariant to the given transformation. The overall training objective is the following:

$$\mathcal{L}_{\text{VORTEX}} = \sum_i \|f_\theta(y_i^s, A) - x_i^s)\|_1 + \lambda \mathcal{L}_{cons}$$

$$\text{where} \quad \mathcal{L}_{cons} = \begin{cases} \|f_\theta(y_i^u, A) - f_\theta(g(y_i^u), A)\|_1, & \text{if g} \in T_I \\ \|g(f_\theta(y_i^u, A)) - f_\theta(g(y_i^u), A)\|_1, & \text{if g} \in T_E \end{cases}$$

4.1 Generalized Data Augmentation Pipeline

Our *Augmentation Pipeline* (Fig. 1) lets us compose image-based data augmentations that resemble state-of-the-art computer vision data augmentations with the physics-driven, acquisition-based data augmentations motivated by the MRI data acquisition forward model.

4.1.1 Physics-Driven Data Augmentations

Noise. Noise in MRI scans affects signal-to-noise (SNR) ratios and is modeled as a additive complex-valued Gaussian distribution (explained in §3), occurring primarily due to thermal fluctuations in the subject and due to receiver coils, magnetic field strength, and specific imaging parameters (Macovski, 1996). In accelerated MRI, noise is propagated through an undersampling mask and the underlying signal, such that $\epsilon_i \sim \Omega \boldsymbol{F} \mathcal{N}(0, \sigma)$ for the i^{th} example, where $\mathcal{N}$ is a zero-mean complex-valued

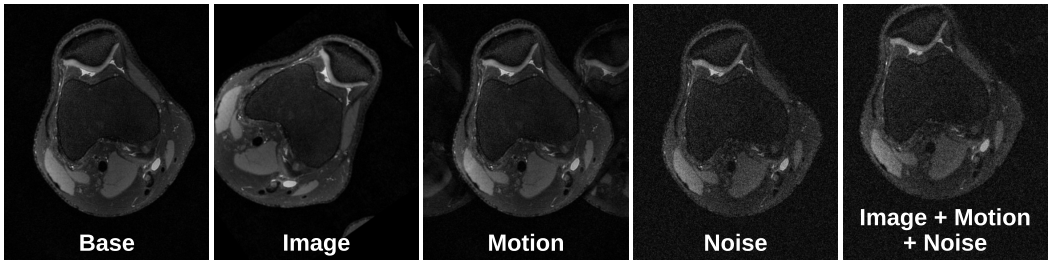

Figure 2: Sample image-based, physics-driven (motion, noise), and composed (image + physics) augmentations applied to a fully-sampled image. Motion and noise were simulated at difficulty levels of $\alpha = 0.2$ and $\sigma = 0.2$, respectively.

Gaussian distribution with standard deviation σ. Since noise is one of the most common MRI artifacts, practical MRI reconstruction methods for clinical deployment should be robust to SNR variations.

Since we precisely know the noise generating process and the MRI data acquisition forward model, we leverage that for consistency training. Specifically, we sample σ from a specified range $\mathcal{R}(\sigma) = [\sigma^{\boldsymbol{LN}}, \sigma^{\boldsymbol{HN}})$ where **LN** (light noise) and **HN** (heavy noise) are chosen based on visual inspections conditioned on clinical scans. We normalize each sampled σ with respect to the magnitude of the image so that same relative change in SNR across scans is induced. We denote the operation of adding noise to the k-space as g_N, in which case the noise-augmented unsupervised example is given by $g_N(y_i^{(u)}) = y_i^{(u)} + \epsilon_i$. We provide an example of a noise-augmented scan in Figure 2.

Motion. Patient motion during long MRI scans can degrade image quality and is an unsolved problem particularly affecting pediatric, elderly, and claustrophobic patients. While navigator sequence-based approaches that densely sample low-resolution motion states during the scan are common for motion correction, they require custom sequences that need to be carefully designed, often leading to increased acquisition time, reduced SNR, and complicated reconstruction (Zaitsev et al., 2001). A common example is rigid motion which occurs due to random patient movement and results in considerable image ghosting artifacts especially in multi-shot MR imaging (Carter, 2011). Many MRI acquisitions sample data over multiple *shots* where consecutive k-space lines are acquired in separate excitations (Anderson & Gore, 1994). Therefore, motion across every shot manifests as additional phase in k-space and as translation in image space. Thus, considering 2D-shots acquisitions, one-dimensional translational motion artifacts can be modeled using random phase errors that alter odd and even lines of k-space separately. Considering we know precisely how rigid motion can be modeled in k-space, we leverage that for consistency training. We denote the phase error due to motion for i^{th} example by $e^{-j\phi_i}$ that corresponds to a translational motion. We sample two random numbers from the uniform distribution $m_o, m_e \sim U(-1, 1)$ which is chosen from a specified range $\mathcal{R}(\alpha) = [\alpha^{\boldsymbol{LM}}, \alpha^{\boldsymbol{HM}})$ where α denotes the amplitude of the phase errors and **LM** (light motion) and **HM** (heavy motion) are chosen based on visual inspections conditioned on clinical scans. Then, for the k^{th} line in k-space, the phase error is given as in the following:

$$\phi_i^k = \begin{cases} \pi\alpha m_o, & \text{if k is odd} \\ \pi\alpha m_e, & \text{if k is even} \end{cases}$$

We denote the operation of adding motion to the k-space as g_M, in which case the motion-augmented unsupervised example is given by $g_M(y_i^{(u)}) = y_i^{(u)} e^{-j\phi_i}$ (example scan given in Fig 2).

4.1.2 Image-based Data Augmentations

In the MR reconstruction task, data augmentations need to transform the target images and their corresponding k-space and coil sensitivity measurements, in contrast to classification problems where labels stay invariant with respect to the augmentations. Moreover, unlike physics-driven augmentations that occur in k-space, image-based augmentations occur in the image domain. Since the training data initially exists as k-space measurements, we first transform it into the image domain using coil sensitivity maps. We then apply a cascade of the image-based data augmentations to both the image and the sensitivity maps. Image-based data augmentations include pixel-preserving augmentations such as flipping, translation, arbitrary and 90 degree multiple rotations, translation, as well as isotropic

and anisotropic scaling. Using the augmented image and transformed sensitivity maps, we run the forward model A to generate the corresponding undersampled k-space measurements.

Composing Augmentations. Our *Augmentation Pipeline* allows for composing different combinations of physics-driven and image-based data augmentations, with example composed augmentations shown in Figure 2. It is important to note that composing multiple physics-driven augmentations such as noise and motion corruption represents a real-world scenario as multiple artifacts can occur simultaneously during MRI acquisition. Appendix C discusses augmentation composition in detail.

4.2 Augmentation Scheduling

We adopt curriculum learning (Hacohen & Weinshall, 2019) for physics-driven data augmentations, where we seek to schedule the *task difficulty*. Difficulty is denoted by σ, the standard deviation of the additive zero-mean complex-valued Gaussian noise, and α, the amplitude of the phase errors for motion. Note that this is in contrast to the MRAugment scheduling strategy, which schedules the probability p of an augmentation. Concretely, for noise, we consider a time-varying range $\mathcal{R}(\sigma(t)) = [\sigma^{\boldsymbol{L}}, \sigma^{\boldsymbol{H}}(t))$, where t indexes the iteration number during training. The upper-bound $\sigma^{\boldsymbol{H}}(t)$ increases monotonically to ensure task difficulty increases during training. We consider two scheduling techniques $\beta(t)$ such that $\sigma^{\boldsymbol{H}}(t) = \sigma^{\boldsymbol{L}} + \beta(t)(\sigma^{\boldsymbol{H}} - \sigma^{\boldsymbol{L}})$: (1) **Linear:** $\beta(t) = t/M$, and (2) **Exponential:** $\beta(t) = \frac{1-e^{-t/\tau}}{1-e^{-M/\tau}}$, where M is the number of epochs until which task difficulty increases and τ is the time-constant for exponential scheduling. After M epochs, training proceeds with constant upper bound $\sigma^{\boldsymbol{H}}$. Scheduling for motion is the same where σ is replaced with α, and image-based data augmentations follow the scheduling strategy proposed in MRAugment as there is no explicit sense of difficulty for that class of data augmentations. Figure 5 included in the Appendix shows simulated $\beta(t)$ for different curricula configurations.

5 Experiments

5.1 Setup

We evaluate our method using the publicly-available mridata 3D fast-spin-echo (FSE) multi-coil knee dataset (Ong et al., 2018). 3D MRI scans were decoded into a hybrid k-space ($x \times k_y \times k_z$) using the 1D orthogonal inverse Fourier transform along the readout direction x. All methods reconstructed 2D $k_y \times k_z$ slices. Sensitivity maps were estimated for each slice using JSENSE (Ying & Sheng, 2007). 2D Poission Disc undersampling masks were used for training and evaluation. N_s training scans were randomly selected to be fully-sampled (supervised) examples while N_u scans were used to simulate undersampled-only scans. All methods used 2D U-Net network with a complex-ℓ_1 training objective both for supervised and for the consistency loss. We evaluated our reconstructions with two image quality metrics: magnitude structural similarity (SSIM) (Wang et al., 2004) and complex peak signal-to-noise ratio (cPSNR) in decibels (dB). SSIM has shown to be a more clinically-preferred metric to cPSNR for quantifying perceptual quality of MRI reconstructions (Knoll et al., 2020). Appendix D discusses the experimental setup in further detail, and Appendix F includes additional experiments across all methods on the 2D fastMRI multi-coil brain dataset (Zbontar et al., 2018).

5.2 Robustness to Clinically Relevant Distribution Drifts

Unlike many other ML domains, the source of possible distribution drifts in accelerated MRI reconstruction can be well characterized based on the known, physics-driven forward data acquisition process. This enables accurate simulations of many of these distribution drifts. Here, we simulate SNR and motion corruptions, two common and problematic artifacts, at inference time using models described in Section 4.1.1 at 16x scan acceleration. Specifically, we use $\sigma = 0.2$ for light noise and $\sigma = 0.4$ for heavy noise. Similarly, we use $\alpha = 0.2$ for light motion and $\alpha = 0.4$ for heavy motion.

In Table 1, we compare supervised baselines without and with the physics-based augmentations and Noise2Recon to VORTEX. For the supervised training with augmentation methods, augmentation is applied with probability $p = 0.2$ during training for noise, motion, and composition corresponding to $g_N(g_M(\cdot))$. For consistency-based approaches, we used $\lambda = 0.1$ for $\mathcal{L}_{cons}$ for noise, motion, and composition. Both *Aug (Motion)* and *VORTEX (Motion)* models were trained with $\mathcal{R}(\alpha) =$

Table 1: Average test results at 16x acceleration for in- and out-of-distribution data with SNR and motion perturbations. The *heavy* difficulty configuration ($\mathcal{R}(\sigma) = [0.2, 0.5)$ for noise and $\mathcal{R}(\alpha) = [0.2, 0.5)$ for motion) was used for all physics-driven augmentations during consistency training with 1:1 balanced sampling and augmentation curricula with highest validation cPSNR.

Perturbation	Metric	Supervised	Aug (Motion)	Aug (Noise)	Aug (Motion+Noise)	VORTEX (Motion)	Noise2Recon	VORTEX (Motion+Noise)
None	SSIM	0.798	0.793	0.805	0.789	0.877	**0.882**	0.869
	cPSNR (dB)	35.8	35.9	35.8	35.7	36.4	**36.4**	36.4
Motion (light)	SSIM	0.809	0.793	0.799	0.785	**0.867**	0.854	0.854
	cPSNR (dB)	33.6	35.1	34.1	35.0	**35.8**	32.8	35.4
Motion (heavy)	SSIM	0.706	0.751	0.722	0.739	**0.812**	0.731	0.803
	cPSNR (dB)	27.0	31.5	29.6	31.9	**32.3**	27.1	32.3
Noise (light)	SSIM	0.830	0.786	0.778	0.761	**0.857**	0.854	0.840
	cPSNR (dB)	33.8	33.7	34.2	34.2	34.0	**34.8**	34.8
Noise (heavy)	SSIM	0.807	0.758	0.745	0.739	0.823	**0.830**	0.812
	cPSNR (dB)	32.2	32.0	33.5	33.4	32.4	**34.0**	33.9

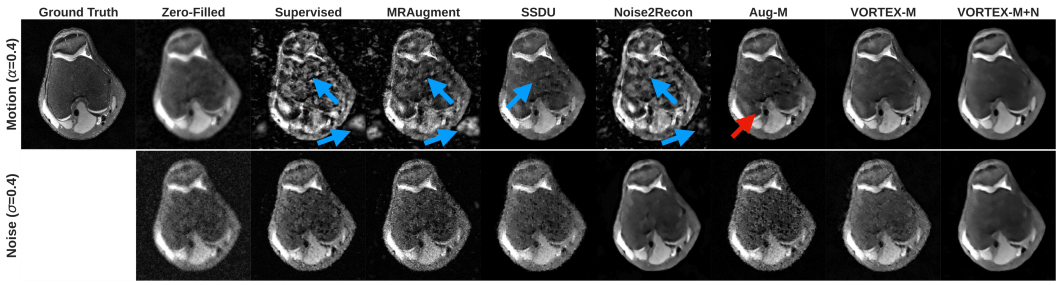

Figure 3: Example reconstructions for simulated scans with heavy motion (top) and heavy noise (bottom). M and M+N correspond to motion and motion+noise augmentations, respectively. *Supervised*, MRAugment, SSDU, and Noise2Recon amplify motion ghosting artifacts (blue arrow). Supervised training with motion augmentations (Aug-M) reduces these artifacts, but still suffers from artifacts (red arrow) and extensive blurring. VORTEX-M and VORTEX-M+N suppress these artifacts. Methods without noise augmentations (Supervised, MRAugment, SSDU, Aug-M, VORTEX-M) amplify image noise. VORTEX-M+N suppresses noise without over-blurring the image.

$[0.2, 0.5)$, and both *Aug (Noise)* and *Noise2Recon* models were trained with $\mathcal{R}(\sigma) = [0.2, 0.5)$. *Aug (Motion+Noise)* and *VORTEX (Motion+Noise)* setting also follow these ranges. We include the results in the Appendix where smaller ranges $\mathcal{R}(\alpha) = [0.1, 0.3)$ for motion and $\mathcal{R}(\sigma) = [0.1, 0.3)$ for noise were used during training. We use a balanced data sampling approach where unsupervised and supervised examples are sampled at a fixed ratio of $1:1$ during training, and all consistency training approaches used augmentation curricula with highest validation cPSNR as described in Section 4.2. Results are shown with more unsupervised slices than supervised (1600 vs 320), which is a realistic clinical scenario. We show results for different accelerations, training times and augmentation curricula in the Appendices D and E.

We demonstrate a large improvement of +8.4 SSIM and +0.6 cPSNR with respect to the supervised baseline for in-distribution data with *VORTEX (Noise)*. The vanilla supervised augmentation-based approaches (*Aug (Motion)*, *Aug (Noise)*, *Aug (Motion+Noise)*) fail to show any meaningful improvement. We observe consistent improvements over both *Supervised* and vanilla augmentation baselines for both light and heavy motion cases with an impressive improvement of +10.6 SSIM and +5.3 cPSNR with *VORTEX (Motion)* over *Supervised* for the heavy motion-corruption case at inference. Similarly, we show considerable improvements over both the *Supervised* and vanilla augmentation baselines with *VORTEX (Noise)* for both light and heavy noise cases with an improvement of +2.3 SSIM and +1.8 cPSNR for heavy noise-corruption case. We highlight that our proposed consistency-based improvements are considerably larger than values reported in DL-based accelerated MRI literature that use different architectures, loss functions, or data consistency schemes (Zbontar et al., 2018; Hammernik et al., 2021). We show example reconstructions comparing our method, supervised baseline, and standard augmentation-based approaches in Figure 3.

Table 2: Average test results for in-distribution data and out-of-distribution data with heavy motion and heavy noise perturbations. Physics augmentations are compositions of noise and motion in their *heavy* training difficulty configurations.

Perturbation	None		Motion (heavy)		Noise (heavy)	
Model	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)
Supervised	0.798 (0.038)	35.8 (0.351)	0.706 (0.048)	27.0 (0.779)	0.807 (0.015)	32.2 (0.278)
MRAugment	0.811 (0.043)	36.2 (0.533)	0.660 (0.040)	24.0 (0.954)	0.742 (0.005)	30.8 (0.293)
SSDU	0.787 (0.026)	34.9 (0.401)	0.734 (0.009)	31.9 (1.70)	0.716 (0.023)	32.5 (0.321)
Aug (Physics)	0.789 (0.045)	35.7 (0.359)	0.739 (0.010)	31.9 (2.35)	0.739 (0.051)	33.4 (0.282)
Aug (Image+Physics)	0.785 (0.050)	36.1 (0.531)	0.742 (0.022)	**32.8 (2.36)**	0.727 (0.051)	33.7 (0.435)
VORTEX (Image)	0.862 (0.030)	36.4 (0.335)	0.648 (0.080)	26.1 (0.678)	0.767 (0.016)	31.5 (0.172)
VORTEX (Physics)	**0.872 (0.033)**	**36.4 (0.296)**	**0.785 (0.019)**	31.8 (2.84)	0.817 (0.034)	**33.9 (0.227)**
VORTEX (Image+Physics)	0.861 (0.036)	36.4 (0.368)	0.777 (0.034)	31.1 (2.74)	**0.831 (0.023)**	33.3 (0.097)

Table 3: Ablation for consistency at pixel-level vs. latent space. **LM**: light motion; **HM**: heavy motion

Model	cPSNR (dB)	SSIM	cPSNR (dB) (LM)	SSIM (LM)	cPSNR (dB) (HM)	SSIM (HM)
Supervised	35.8	0.798	33.6	0.809	27.1	0.706
Pixel-Level	36.4	0.873	35.9	0.866	33.2	0.828
R_4	36.4	0.877	34.7	0.865	29.8	0.778
R_3,R_4	36.4	0.873	34.0	0.852	30.1	0.781
R_2,R_3,R_4	36.3	0.873	34.4	0.854	29.5	0.769
R_1,R_2,R_3,R_4	36.3	0.875	34.7	0.864	30.3	0.775

5.3 VORTEX VS. BASELINES

We compare VORTEX performance for in- and out-of-distribution data at 16x acceleration to supervised methods using both physics-driven and the state-of-the-art image-based MRAugment augmentations, and to the state-of-the-art self-supervised via data undersampling (SSDU) reconstruction method (Yaman et al., 2020). We describe SSDU method and our implementation in detail in the Appendix D.2. OOD simulations of SNR change and motion corruption follow the setup described in Section 5.2 where heavy motion (HM) corresponds to $\alpha = 0.4$ phase error amplitude and heavy noise (HN) corresponds to $\sigma = 0.4$ additive k-space zero-mean complex-valued Gaussian noise. *Physics* augmentations listed in Table 2 correspond to the composition of noise and motion augmentations in their *heavy* difficulty configurations during training ($\mathcal{R}(\sigma) = [0.2, 0.5)$ for noise and $\mathcal{R}(\alpha) = [0.2, 0.5)$ for motion). Consistency-weighting λ, augmentation probability p, balanced data sampling ratio, and supervised and unsupervised data amounts are identical to Section 5.2. We isolate the benefits of consistency training with VORTEX from the utility of the data augmentations (Aug) themselves by separately comparing *Aug (Physics)* and *Aug (Image + Physics)*. Note that MRAugment corresponds to *Aug (Image)*, which is based on our own implementation and the originally reported hyperparameters. See D.3.1 for hyperparameter details.

VORTEX (Physics) demonstrated substantial improvements of +7.4 SSIM and +0.6 cPSNR over the *Supervised* baseline, +6.1 SSIM and +0.2 cPSNR over *MRAugment*, and +8.5 SSIM and +1.5 cPSNR over *SSDU* for in-distribution data. As *VORTEX (Image)* also considerably improves over *Supervised* and *MRAugment*, a dominant mechanism of the benefits is attributed to the consistency training even for the in-distribution setting. For both heavy motion and heavy noise settings, including physics augmentations is vital for robust performance as *MRAugment*, *SSDU*, and *VORTEX (Image)* perform worse, even compared the *Supervised* baseline. For heavy motion, we observe an improvement of +7.9 SSIM and +4.8 cPSNR over the *Supervised* and +12.5 SSIM and +7.8 cPSNR over *MRAugment* with *VORTEX (Physics)*. Similarly, for heavy noise, we show an improvement of +2.4 SSIM and +1.1 cPSNR over the Supervised baseline and +8.9 SSIM and +2.5 cPSNR over *MRAugment* with *VORTEX (Image + Physics)*. We note that SSIM is clinically-preferred to cPSNR for quantifying MRI perceptual quality (Zbontar et al., 2018).

The substantial performance gain with VORTEX in both in-distribution and OOD settings suggests that the consistency training framework is amenable to both image-based and physics-driven, acquisition-based augmentations. While conventional supervised training requires that all augmentations preserve the noise statistics of the training data, consistency training can relax this constraint and allow for the use of acquisition-based augmentations (see Appendix B.1).

5.4 ABLATIONS

We perform ablations to understand two design questions for key components in our framework: (1) Can consistency be enforced at different points in the network; (2) How should example difficulty be specified during training. All methods use the default configurations specified in Appendix D.3. To evaluate each piece thoroughly, we consider augmentation and VORTEX approaches trained with heavy motion perturbations. Additional ablation findings are detailed in Appendix E.

Latent Space vs Pixel-level Consistency. We compare enforcing consistency at the pixel-level output image versus learned latent representations at varying U-Net resolution levels. Let R_k be k^{th} resolution level at which consistency is enforced, where $k \in \{1, 2, 3, 4\}$ since our U-Net architecture had 4 pooling layers (see Appendix E for details). We find that latent space consistency performed similarly across all resolution levels, outperforming the *Supervised* baseline on both in- and out-of-distribution data (Table 3). For in-distribution data, latent space consistency at any resolution level performed on par with pixel-level consistency. However, for OOD data, it performed considerably worse than pixel-level consistency, by at least 3.2 cPSNR and 4.7 SSIM under heavy motion. Although not common in the consistency training literature, we find that pixel-level consistency was a better technique for capturing the semantics of global distribution shifts such as motion for accelerated MRI reconstruction, which might occur at the pixel-level.

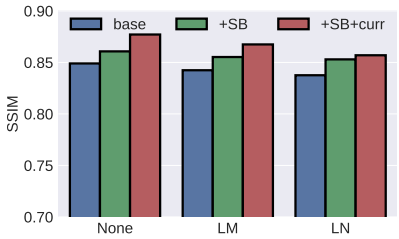

Figure 4: Ablation for balanced sampling (*SB*) and augmentation curricula (*curr*) in *VORTEX (Motion)*.

Augmentation Scheduling. We seek to quantify the utility of scheduling augmentation difficulty in VORTEX's consistency branch (see §4.2). We evaluate linear and exponential scheduling functions with different warm up schedules – 10%, 50%, and 100% of the training period. We show that curricula methods outperformed non-curricula methods for both in-distribution and OOD evaluation (Table 6 in the Appendix). However, no one curricula configuration outperformed others, which may indicate that all curricula methods are feasible ways to schedule augmentations. Curriculum learning is also compatible with the balanced sampling protocol proposed by Desai et al. (2021a), where supervised and unsupervised examples are sampled at a fixed ratio during training. Incorporating balanced sampling (SB) into training led to an increase in SSIM for both in-distribution and OOD light motion and light noise evaluation configurations (Fig.4). Increase in SSIM may indicate that curricula can help the network gradually learn useful representations without a mode collapse into the trivial solution (i.e. image blurring), which is common for pixel-level losses.

6 CONCLUSION

We propose a semi-supervised consistency training framework VORTEX for accelerated MRI reconstruction that uses a generalized data augmentation pipeline for improved data-efficiency and robustness to clinically relevant distribution drifts. VORTEX enforces invariance to *physics-driven* data augmentations of noise and motion; enforces equivariance to image-based data augmentations of flipping, scaling, rotation, translation, and shearing; enables composing data augmentations of different types; and allows for curriculum learning based on the difficulty of physics-driven augmentations. We demonstrate strong improvements compared to the fully-supervised augmentation baselines, and the state-of-the-art data augmentation scheme MRAugment, on both in-distribution and OOD data. Our framework is model-agnostic and could be used with any other MRI reconstruction models or even for other image-to-image tasks with appropriate data augmentations. In future work, we plan to extend our VORTEX physics-driven, acquisition-based augmentations to additional OOD MRI artifacts and non-Cartesian undersampling patterns to work towards building robust DL-based MR reconstruction models that can be safely deployed in clinics.

7 ETHICS STATEMENT

We, the authors, confirm that we are responsible in case of violation of rights, etc. and ensure the reproducibility of this study. All data was acquired from healthy subjects who were volunteering in a

research study with informed consent and under Institutional Review Board Approval. The data used in this study is publicly available at http://mridata.org/ and at http://fastmri.org/.

REFERENCES

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37(6):1322–1332, 2018.

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, Feb 2019. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2018.2865356.

Thai Akasaka, Koji Fujimoto, Takayuki Yamamoto, Tomohisa Okada, Yasutaka Fushumi, Akira Yamamoto, Toshiyuki Tanaka, and Kaori Togashi. Optimization of regularization parameters in compressed sensing of magnetic resonance angiography: can statistical image metrics mimic radiologists' perception? *PloS one*, 11(1):e0146548, 2016.

A. W. Anderson and J. C. Gore. Analysis and correction of motion artifacts in diffusion weighted imaging. *Magn Reson Med*, 32(3):379–387, Sep 1994.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *NeurIPS*, 2014.

GJ Barker. Technical issues for the study of the optic nerve with mri. *Journal of the neurological sciences*, 172:S13–S16, 2000.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533. PMLR, 2019.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Robert Bridson. Fast poisson disk sampling in arbitrary dimensions. *SIGGRAPH sketches*, 10(1), 2007.

James Wesley Carter. Mri: The basics, 3rd ed. *American Journal of Roentgenology*, 197(2):W361–W361, 2011. doi: 10.2214/AJR.11.6487. URL https://doi.org/10.2214/AJR.11.6487.

Elena Celledoni, Matthias Joachim Ehrhardt, Christian Etmann, Brynjulf Owren, Carola-Bibiane Schonlieb, and Ferdia Sherry. Equivariant neural networks for inverse problems. *Inverse Problems*, 37, 2021.

Akshay S. Chaudhari, Christopher M. Sandino, Elizabeth K. Cole, David B. Larson, Garry E. Gold, Shreyas S. Vasanawala, Matthew P. Lungren, Brian A. Hargreaves, and Curtis P. Langlotz. Prospective deployment of deep learning in mri: A framework for important considerations, challenges, and recommendations for best practices. *Journal of Magnetic Resonance Imaging*, 54 (2):357–371, Aug 2021. ISSN 1053-1807, 1522-2586. doi: 10.1002/jmri.27331.

Govind B Chavhan, Paul S Babyn, and Shreyas S Vasanawala. Abdominal mr imaging in children: motion compensation, sequence optimization, and protocol organization. *Radiographics*, 33(3): 703–719, 2013.

Chen Chen, Chen Qin, Cheng Ouyang, Shuo Wang, Huaqi Qiu, Liang Chen, Giacomo Tarroni, Wenjia Bai, and Daniel Rueckert. Enhancing mr image segmentation with realistic adversarial data augmentation. *ArXiv*, abs/2108.03429, 2021.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. Semi-supervised sequence modeling with cross-view training. In *EMNLP*, 2018.

Elizabeth K. Cole, John M. Pauly, Shreyas S. Vasanawala, and Frank Ong. Unsupervised mri reconstruction with generative adversarial networks, 2020.

Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3008–3017, 2020.

Mohammad Zalbagi Darestani and Reinhard Heckel. Accelerated mri with un-trained neural networks. *IEEE Transactions on Computational Imaging*, 7:724–733, 2021.

Mohammad Zalbagi Darestani, A. Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *ICML*, 2021.

Arjun D Desai, Batu M Ozturkler, Christopher M Sandino, Shreyas Vasanawala, Brian A Hargreaves, Christopher M Re, John M Pauly, and Akshay S Chaudhari. Noise2recon: A semi-supervised framework for joint mri reconstruction and denoising. *arXiv preprint arXiv:2110.00075*, 2021a.

Arjun D Desai, Andrew M Schmidt, Elka B Rubin, Christopher Michael Sandino, Marianne Susan Black, Valentina Mazzoli, Kathryn J Stevens, Robert Boutin, Christopher Re, Garry E Gold, et al. Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021b.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, 2018.

Zalan Fabian, Reinhard Heckel, and M. Soltanolkotabi. Data augmentation for deep learning based accelerated mri reconstruction with limited data. In *ICML*, 2021.

Weijie Gan, Yu Sun, Cihat Eldeniz, Jiaming Liu, Hongyu An, and Ulugbek S. Kamilov. Modir: Motion-compensated training for deep image reconstruction without ground truth, 2021.

Davis Gilton, Greg Ongie, and R. Willett. Model adaptation for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:661–674, 2021.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *ICLR*, 2021.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks, 2019.

Jacques S. Hadamard. Sur les problemes aux derive espartielles et leur signification physique.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Kerstin Hammernik, Jo Schlemper, Chen Qin, Jinming Duan, Ronald M Summers, and Daniel Rueckert. Systematic evaluation of iterative deep neural networks for fast parallel mri reconstruction with sensitivity-weighted coil combination. *Magnetic Resonance in Medicine*, 2021.

Kyong Hwan Jin, Ji-Yong Um, Dongwook Lee, Juyoung Lee, Sung-Hong Park, and Jong Chul Ye. Mri artifact correction using sparse+ low-rank decomposition of annihilating filter-based hankel matrix. *Magnetic resonance in medicine*, 78(1):327–340, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Florian Knoll, Tullie Murrell, Anuroop Sriram, Nafissa Yakubova, Jure Zbontar, Michael Rabbat, Aaron Defazio, Matthew J Muckley, Daniel K Sodickson, C Lawrence Zitnick, et al. Advancing machine learning for mr image reconstruction with an open competition: Overview of the 2019 fastmri challenge. *Magnetic resonance in medicine*, 84(6):3054–3070, 2020.

Anish Lahiri, Guanhua Wang, Saiprasad Ravishankar, and Jeffrey A Fessler. Blind primed supervised (blips) learning for mr image reconstruction. *IEEE Transactions on Medical Imaging*, pp. 1–1, 2021. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2021.3093770.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242, 2017.

Ke Lei, Morteza Mardani, John M. Pauly, and Shreyas S. Vasanawala. Wasserstein gans for mr imaging: From paired to unpaired training. *IEEE Transactions on Medical Imaging*, 40(1):105–115, Jan 2021. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2020.3022968.

Jiaming Liu, Yu Sun, Cihat Eldeniz, Weijie Gan, Hongyu An, and Ulugbek S. Kamilov. Rare: Image reconstruction using deep priors learned without groundtruth. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1088–1099, Oct 2020. ISSN 1941-0484. doi: 10.1109/jstsp.2020.2998402. URL http://dx.doi.org/10.1109/JSTSP.2020.2998402.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Wenmiao Lu, Kim Butts Pauly, Garry E Gold, John M Pauly, and Brian A Hargreaves. Semac: slice encoding for metal artifact correction in mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(1):66–76, 2009.

Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed sensing mri. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008. doi: 10.1109/MSP.2007.914728.

Albert Macovski. Noise in mri. *Magnetic resonance in medicine*, 36(3):494–497, 1996.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. *ArXiv*, abs/2004.14444, 2020.

Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2019.

F Ong and M Lustig. Sigpy: a python package for high performance iterative reconstruction. In *Proceedings of the ISMRM 27th Annual Meeting, Montreal, Quebec, Canada*, volume 4819, 2019.

F Ong, S Amin, S Vasanawala, and M Lustig. Mridata.org: An open archive for sharing mri raw data. In *Proc. Intl. Soc. Mag. Reson. Med*, volume 26, 2018.

Kamlesh Pawar, Zhaolin Chen, N. Jon Shah, and Gary F. Egan. Suppressing motion artefacts in mri using an inception-resnet network with motion simulation augmentation. *NMR in Biomedicine*, Dec 2019. ISSN 1099-1492. doi: 10.1002/nbm.4225. URL http://dx.doi.org/10.1002/nbm.4225.

Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. Sense: Sensitivity encoding for fast mri. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999. doi: https://doi.org/10.1002/(SICI)1522-2594(199911)42:5⟨952::AID-MRM16⟩3.0.CO;2-S.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *ICML*, abs/1902.10811, 2019.

Philip M Robson, Aaron K Grant, Ananth J Madhuranthakam, Riccardo Lattanzi, Daniel K Sodickson, and Charles A McKenzie. Comprehensive quantification of signal-to-noise ratio and g-factor for image-based and k-space-based parallel imaging reconstructions. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 60(4):895–907, 2008.

Peter B Roemer, William A Edelstein, Cecil E Hayes, Steven P Souza, and Otward M Mueller. The nmr phased array. *Magnetic resonance in medicine*, 16(2):192–225, 1990.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Mehdi S. M. Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016.

Christopher M. Sandino, Joseph Y. Cheng, Feiyu Chen, Morteza Mardani, John M. Pauly, and Shreyas S. Vasanawala. Compressed sensing: From research to clinical practice with deep neural networks: Shortening scan times for magnetic resonance imaging. *IEEE Signal Process. Mag.*, 37 (1):117–127, 2020a. doi: 10.1109/MSP.2019.2950433. URL https://doi.org/10.1109/MSP.2019.2950433.

Christopher M. Sandino, Peng Lai, Shreyas S. Vasanawala, and Joseph Y. Cheng. Accelerating cardiac cine mri using a deep learning-based espirit reconstruction. *Magnetic Resonance in Medicine*, 85: 152 – 167, 2020b.

Richard Shaw, Carole H. Sudre, Thomas Varsavsky, Sébastien Ourselin, and Manuel Jorge Cardoso. A k-space model of movement artefacts: Application to segmentation augmentation and artefact removal. *IEEE Transactions on Medical Imaging*, 39:2881–2892, 2020.

Anuroop Sriram, J. Zbontar, Tullie Murrell, Aaron Defazio, C. L. Zitnick, N. Yakubova, F. Knoll, and P. Johnson. End-to-end variational networks for accelerated mri reconstruction. *ArXiv*, abs/2004.06688, 2020.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, abs/2007.00644, 2020.

Martin Uecker, Peng Lai, Mark J Murphy, Patrick Virtue, Michael Elad, John M Pauly, Shreyas S Vasanawala, and Michael Lustig. Espirit—an eigenvalue approach to autocalibrating parallel mri: where sense meets grappa. *Magnetic resonance in medicine*, 71(3):990–1001, 2014.

Muhammad Usman, Siddique Latif, Muhammad Asim, Byoung-Dai Lee, and Junaid Qadir. Retrospective motion correction in multishot mri using generative adversarial network. *Scientific reports*, 10(1):1–11, 2020.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kamil Ugurbil, and Mehmet Akcakaya. Self-supervised physics-based deep learning mri reconstruction without fully-sampled data. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 921–925. IEEE, Apr 2020. ISBN 9781538693308. doi: 10.1109/ISBI45749.2020.9098514. URL https://ieeexplore.ieee.org/document/9098514/.

Leslie Ying and Jinhua Sheng. Joint image reconstruction and sensitivity estimation in sense (jsense). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 57(6):1196–1202, 2007.

Maxim Zaitsev, Karl Zilles, and Nadim Joni Shah. Shared k-space echo planar imaging with keyhole. *Magnetic Resonance in Medicine*, 45, 2001.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

A GLOSSARY

Table 4 provides the notation used in the paper.

Table 4: Summary of notation used in this work.

	Notation	Description
MRI forward model	x, y	Image, k-space measurements
	$y_i^{(s)}, y_i^{(u)}$	Fully-sampled (supervised) k-space, prospectively undersampled (unsupervised) k-space
	$\Omega, \boldsymbol{F}, \boldsymbol{S}$	Undersampling mask, fourier transform matrix, coil sensitivity maps
	A	The forward MRI acquisition operator
	ϵ	Additive complex-valued Gaussian noise
Augmentation transforms	T	Set of data transforms
	T_I, T_E	Set of invariant and equivariant data transforms
	g, g_E, g_I	Transform, equivariant transform, invariant transform
	G_E, G_I	Sequence of sampled invariant and equivariant data transforms
	$\mathcal{N}(0,\sigma)$	Complex gaussian distribution with zero-mean, variance σ^2
	α	Motion-induced phase error amplitude
	ϕ_i^k	Phase error for k$^{\text{th}}$ phase encode line in example i
	$\mathcal{R}(\cdot)$	Range
	$\beta(t)$	Difficulty scale
	LM, HM	Light motion (α=0.2), heavy motion (α=0.4)
	LN, HN	Light noise (σ=0.2), heavy noise (σ=0.4)
Model components and losses	$\mathcal{L}_{sup}$, $\mathcal{L}_{cons}$	Supervised, consistency loss
	λ	Consistency loss weight
	R_i	U-Net resolution level i

B EXTENDED RELATED WORK

In this section, we summarize the key differences between VORTEX and prior work in augmentations (i.e. MRAugment) and in consistency training (i.e. Noise2Recon). Specifically, we highlight two advantages of VORTEX:

1. **Image-based *and* Acquisition-based Augmentations.** VORTEX can relax the assumption that augmentations must preserve the noise statistics of the data (Fabian et al., 2021). This allows VORTEX to leverage both image-based and acquisition-based augmentations, which do not preserve the noise statistics of the data.
2. **Regularization Beyond Noise.** VORTEX can leverage physics-driven augmentations beyond the standard denoising regularization used in prior work in both consistency (Desai et al., 2021a) and pre-training (Romano et al., 2017). Thus, it may be feasible to extend VORTEX to other relevant clinical artifacts while maintaining the regularization properties of the well-studied denoising task.

B.1 VORTEX VS MRAUGMENT

MRAugment proposes a framework for applying image-based augmentations on fully-supervised training data. This approach showed improved performance in data-limited settings, which may suggest the family of image-based augmentations are helpful in reducing model overfitting. It also suggests scheduling the likelihood of applying an augmentation can be helpful for reducing the number of augmented examples in early stages of training.

Image vs Acquisition Augmentations. MRAugment focuses on the use of image-based augmentations for supervised training. In VORTEX, both image-based and MRI acquisition-based augmentations are used for semi-supervised consistency training to 1) reduce dependence on supervised training data and 2) increase robustness to physics-driven perturbations that are frequently observed during MRI acquisition.

Relaxing Assumption of Preserved Noise Statistics. MRAugment notes that the family of image-based augmentations were selected to ensure that noise statistics of the training data were preserved. However, this constraint excludes acquisition-based augmentations, particularly noise and motion,

which are needed to build robustness to noise and motion artifacts in MRI. However, these acquisition-based augmentations inherently change the effective noise floor (and thus SNR) of the scan, and thus violate this constraint. We empirically validate this claim in supervised settings, where acquisition-based augmentations perform worse than standard supervised training in in-distribution settings (Table 1). This tradeoff between in-distribution performance and OOD robustness would preclude the application of acquisition-based augmentations in practice.

However, with VORTEX, not only is this tradeoff mitigated but the performance in both in-distribution and OOD settings is significantly improved (Table 9). This improved performance empirically demonstrates that the assumption that augmentations must preserve noise statistics can be relaxed in the VORTEX framework. Thus, both image-based and acquisition-based augmentations can be leveraged simultaneously, which leads to improvements in performance over either family of augmentations alone (Table 9).

Precomputing Coil Sensitivity Maps. Integrating coil sensitivity maps is standard clinical practice to help constrain the optimization problem for MRI image reconstruction (Sandino et al., 2020b; Robson et al., 2008; Roemer et al., 1990). MRAugment utilizes the end-to-end VarNet, which *learns* to jointly estimate coil sensitivities and reconstruct images (Sriram et al., 2020). Thus, the augmentation pipeline in MRAugment does not need to explicitly account for the effect of image-based transformations on sensitivity maps. It also has the added benefit of optimzing sensitivity map estimtion with respect to augmented data. In practice, precomputing coil sensitivities is feasible and routine with sensitivity map estimation methods such as ESPIRiT (Uecker et al., 2014) and JSENSE (Ying & Sheng, 2007). Additionally, precomputed maps are important in multi-coil datasets where the number of coils are not constant across different scans, which is critical when patients with heterogenous anatomies are being imaged (Desai et al., 2021b).

VORTEX utilizes precomputed sensitivity maps estimated from auto-calibration regions in each scan. Because image-based augmentations are designed to emulate shifts in the imaging target, they also impact the coil geometry and sensitivity maps that are estimated. In contrast to the MRAugment sensitivity map formulation, which assumes sensitivity maps are fixed, VORTEX integrates physics-based modeling to appropriately warp sensitivity maps based on image-based augmentations. Given some equivariant image-based transform g_E, the augmented image for coil i ($\tilde{x_i}$) can be defined as

$$\tilde{x_i} = g_E(\boldsymbol{S}_i)g_E(x)$$

Scheduling Augmentation Difficulty. MRAugment and VORTEX also different in the mechanism of how augmentations are scheduled. MRAugment proposes an augmentation scheduling method that schedules the probability of applying an augmentation. Thus, training can occur predominantly on collected data in earlier stages of training and augmentations can help reduce overfitting at later.

VORTEX is designed to build robustness to OOD perturbations, where the *extent* (and, more generally, difficulty) of these perturbations will be unknown at test time. In this framework, augmentations must not only regularization method for improved performance on in-distribution data, but rather appropriately model a separate distribution of data with respect to which the model can be trained. Thus, the model must learn to jointly optimize for both in-distribution (default training data) and OOD (perturbation-corrupt data) examples simultaneously. Intuitively, we need to design an augmentation scheduling scheme that will allow the model to gradually learn to generalize to higher extents (more difficult) perturbations over time while still ensuring examples from both distributions are sampled for joint optimization. To ensure that augmentations are always applied but at different extents, we propose a curriculum learning strategy for scheduling the *difficulty* of the augmentation.

B.2 VORTEX VS NOISE2RECON

Noise2Recon proposes a semi-supervised consistency based framework for joint denoising and reconstruction. This approach showed improved performance in label-limited settings, where the training dataset consists of both supervised and unsupervised data. VORTEX 1) extends this consistency training paradigm to a broader family of acquisition-based perturbations, 2) exhaustively studies how this framework can be leveraged for *both* image and acquisition-based augmentations, and 3) proposes a curriculum learning strategy to gradually increase reconstruction difficulty.

Robustness to Motion. Noise2Recon proposes a novel consistency framework for semi-supervised MRI reconstruction but solely focuses on applications to noise artifacts. While denoising is a well-known regularizer for inverse problems (Romano et al., 2017; Batson & Royer, 2019), many other acquisition-related artifacts in MRI are commonplace. In VORTEX, we explore the utility of motion augmentations as 1) a regularizer to improve robustness in label-limited settings and 2) a method to increase robustness to OOD motion artifacts. We demonstrate that motion artifact removal is as effective of a regularizer as denoising (Tables 1 and 2).

Composing Augmentations for Multi-Artifact Correction. Existing MRI artifact correction/removal methods, including Noise2Recon, separately handle reconstruction and artifact removal tasks, are limited to correcting for a single artifact or require multiple unique workflows to correct for different artifacts (Usman et al., 2020; Lu et al., 2009; Jin et al., 2017). However, in practice, effects of multiple acquisition-related artifacts can be compounded even in accelerated MRI. Thus a unified framework for removing these artifacts is desirable. VORTEX establishes a framework for both image-based and acquisition-based augmentations that can be utilized to jointly reconstruction and remove multiple artifacts with a single approach.

Curriculum Learning for Augmentations. VORTEX extends basic consistency training to include a scheduling protocol for increasing the difficulty of augmentations over the training cycle. Results demonstrates that designing curricula for augmentations in the consistency framework can lead to considerable performance improvements in OOD settings without losing performance among in-distribution scans (Table 6). Such curricula can be helpful for joint optimization of both artifactual and artifact-free images, particularly when example difficulty is extensive Bengio et al. (2009).

B.3 Summary of Technical Contributions

In this work, we characterize the interface between physics-based MRI acquisition-motivated and image-based augmentations to 1) reduce data dependency and 2) increase robustness to clinically-relevant distribution shifts that are pervasive during MRI acquisition. We extend the semi-supervised consistency framework in Noise2Recon to handle both acquisition and image based perturbations in a way that is motivated by the physics-driven forward model of MRI acquisition. To ensure that we are inclusive of a broader family of acquisition-based perturbations than was available in Noise2Recon, we propose extending the semi-supervised consistency framework proposed to handle motion, a common artifact in MRI. We exhaustively study the interaction between physics/acquisition based and image based augmentations in both fully supervised training with augmentations and semi-supervised training with the proposed consistency.

C Equivariant and Invariant Transforms

We provide an extended discussion of the choice of and interaction between equivariant and invariant transforms.

Choosing Equivariance or Invariance. It is important to note that, practically, specifying which transforms the network should be equivariant or invariant to is a design choice and often task-dependent. In the case of MRI, image-based augmentations proposed in MRAugment are meant to simulate differences in patient positioning and spatial scan parameters (e.g. field-of-view, nominal resolution). The differences are typically prescribed at scan time (i.e. scan parameters) or are correctable prior to the scan. In contrast, motion and noise are perturbations that occur *during acquisition*, and therefore cannot be corrected a priori. Thus, building networks that are invariant to these perturbations are critical. Based on this paradigm of transforms in MRI, spatial image transforms are classified as equivariant transforms while the physics-based transforms we propose are classified as invariant transforms.

Composing Transforms (Extended). §4.1 introduces the intuition for equivariant and invariant transformations. In this section, we formalize transforms from these families are composed.

Let $g_1, \ldots, g_K$ be an ordered sequence of unique transforms sampled from a set of transforms T. Let $\bar{G}_E, \bar{G}_I$ be the sequence of sampled equivariant and invariant transforms, respectively. Thus,

$\bar{G}_E = (g_i \text{ if } g_i \in T_E \; \forall i = 1, \ldots, K)$ (similarly for $\bar{G}_I$). Let G_E and G_I be the compositions of each transform in $\bar{G}_E, \bar{G}_I$, respectively. Thus, $G_E = \bar{G}_{E_{|\bar{G}_E|}} \circ \cdots \circ \bar{G}_{E_1}$ (similarly for G_I).

As a design choice, we select all physics-driven, acquisition-related transforms to be in the family of invariant transforms. This choice is made to ensure reconstructions are invariant to plausible acquisition-related perturbations. Thus, the family of physics-driven transforms are synonymous with the family of invariant transforms for our purposes.

Because signal from physics-driven perturbations (noise and motion) is sampled at acquisition, these perturbations are applied after undersampling in the supervised augmentation methods, where fully-sampled data is available.

D Experimental Details

All code and experimental/data configurations are available at (blinded).

D.1 Dataset

In this section, we provide details for the two datasets used in this study: the mridata 3D FSE knee dataset and the fastMRI multi-coil brain dataset.

D.1.1 mridata 3D FSE Knee Dataset

Dataset Splits. The mridata 3D FSE knee dataset consists of 6080 fully-Cartesian-sampled knee slices (19 scans) from healthy participants. The dataset was randomly partitioned into 4480 slices (14 scans) for training, 640 slices (2 scans) for validation, and 960 slices (3 scans) for testing.

Simulating Data-Limited and Label-Limited Settings. In this study, we evaluate all methods in the data-limited and label-limited regimes, where supervised examples are scarce compared to unsupervised (undersampled) examples. To simulate this scenario, a subset of training scans are retrospectively undersampled using fixed undersampling masks, resulting in unsupervised training examples. To limit the total (supervised and unsupervised) amount of available training data, we train with only 6 of the 14 training scans, where 1 scan is supervised and 5 scans are unsupervised.

K-space Hybridization and Sensitivity Maps. 3D FSE scans were acquired in 3D, resulting in Fourier encoded signal along all dimensions ($k_x \times k_y \times k_z$). Because the readout dimension k_x is fully-sampled in these scans, scans were decoded along the k_x dimension, resulting in a hybridized k-space as mentioned in §5.1. All sensitivity maps were estimated with JSENSE as implemented in SigPy (Ong & Lustig, 2019), with a kernel width of 8 and a 20×20 center k-space auto-calibration region.

Mask Generation. Scans for training and evaluation were undersampled using 2D Poisson Disc undersampling, a compressed sensing-motivated pattern for 3D Cartesian imaging. Given an acceleration rate R, undersampling masks were generated in the $k_y \times k_z$ dimensions for all scans such that the number of pixels sampled would be approximately $\frac{|k_y||k_z|}{R}$. To maintain consistency with generated sensitivity maps, a 20×20 center k-space auto-calibration region was used when constructing undersampling masks for all examples. To simulate prospectively undersampled acquisitions, scans were retrospectively undersampled with a fixed 2D Poisson Disc undersampling pattern (Bridson, 2007). Following Cartesian undersampling convention, all $k_y \times k_z$ slices for a single scan are undersampled with the identical 2D Poisson Disc mask. This procedure was used for both simulating prospectively undersampled scans during training (i.e. unsupervised examples) and evaluation. All undersampling masks are generated with an unique, fixed random seed for each scan to ensure reproducibility.

D.1.2 fastMRI Brain Multi-coil Dataset

Dataset Splits. The distributed validation split of the fastMRI 2D brain multi-coil dataset was divided into 757 scans for training, 207 scans for validation, and 414 scans for testing. To control for confounding variables when comparing performance between reconstruction methods, all data

splits were filtered to include only T2-weighted scans acquired at a 3T field strength, resulting in 266, 70, and 137 scans for training, validation, and testing, respectively. Data-limited and label-limited training settings were simulated by limiting training data to 18 supervised and 36 unsupervised scans and validation data to 50 scans.

Sensitivity Maps. Like for mridata, sensitivity maps were estimated using JSENSE with a kernel width of 8 and calibration region of 12×12. This calibration region corresponds to the 4% auto-calibration region used for 8x undersampling.

Mask Generation. Scans for training and evaluation were undersampled using 1D random undersampling, a compressed sensing-motivated pattern for 2D Cartesian imaging. Given an acceleration rate R, undersampling masks were generated in the k_y phase-encode dimension for all scans such that the number of pixels sampled would be approximately $\frac{|k_y|}{R}$. Training and evaluation was conducted at R=8 acceleration with a 4% auto-calibration region. Like in mridata, fixed undersampling masks were generated to simulate prospectively undersampled data and for the testing data to ensure reproduciblity.

D.2 Baselines

We compared VORTEX to state-of-the-art supervised, supervised augmentation, and self-supervised MRI reconstruction baselines. We provide an overview of these methods and their notation in the main text.

Supervised. We compared VORTEX to standard supervised training without augmentations (termed *Supervised*). In supervised training, fully-sampled scans are retrospectively undersampled. The model is trained to reconstruct the fully-sampled scan from its undersampled counterpart. Note, in supervised settings, only fully-sampled scans can be used for training. Any prospectively undersampled (unsupervised) scans cannot be leveraged in this setup.

Supervised+Augmentation (*Aug*) and MRAugment. Supervised baselines with augmentation (termed *Aug*) were trained with image and/or physics-based augmentations, which are denoted by parentheses. Image-based augmentations were applied prior to the retrospective undersampling, following the MRAugment protocol. Physics-based acquisition augmentations were applied after this undersampling to model the MRI data acquisition process. For example *Aug (Motion)* indicates a supervised method trained with motion augmentations. Image-based augmentations were identical to those used in MRAugment. As such, *Aug (Image)* is equivalent to MRAugment, and is referred to as such for readability.

SSDU. We also compared VORTEX to the state-of-the-art self-supervised learning via data undersampling (SSDU) baseline (Yaman et al., 2020). This method was originally proposed for fully unsupervised learning, in which all training scans are prospectively undersampled. We propose a trivial extension to adapt it for the semi-supervised setting. In cases of prospectively undersampled (unsupervised) data, the training protocol proposed in SSDU was used. Fully-sampled (supervised) data was retrospectively undersampled using the undersampling method and acceleration for the specified experiment. These simulated undersampled scans were used as inputs to the SSDU protocol. Because the retrospective undersampling is done dynamically (i.e. each time a supervised example is sampled), it may serve as a method of augmenting supervised scans.

Hyperparameters for all methods are provided in Appendix D.3.1.

D.3 Training Details

All training code is written in Python with PyTorch 1.6.

D.3.1 Hyperparameters

Architecture and Optimization. All models used a 2D U-Net architecture with (Ronneberger et al., 2015) with 4 pooling layers. Convolutional block at depth d consisted of two convolutional

Table 5: Data augmentation configuration for mridata 3D FSE knee dataset experiments. p is the effective probability of applying an augmentation. In MRAugment, this is equivalent to the base probability multiplied by the weighting factor. Acquisition-based augmentations were configured in separate experiments at both light and heavy settings.

Kind	Transform	Parameters	p
Image	H-Flip	N/A	0.275
	V-Flip	N/A	0.275
	$k \times 90°$ rotation	$k \in \{2\}$	0.275
	Rotation	[-180°, 180°]	0.275
	Translation	[-10%, 10%]	0.55
	Scale	[0.75, 1.25]	0.55
	Shear	[-15°, 15°]	0.55
Acquisition	Gaussian Noise	σ=[0.1,0.3] (light) σ=[0.2,0.5] (heavy)	0.2
	Motion	α=[0.1,0.3] (light) α=[0.2,0.5] (heavy)	0.2

layers with 32^d channels for $d = \{1, \ldots, 5\}$. All models were trained with the Adam optimizer with default parameters (β_1=0.9, β_2=0.999) and weight decay of 1e-4 for 200 epochs (Kingma & Ba, 2014; Loshchilov & Hutter, 2017). Training was conducted with an effective training batch size of 24 and learning rate η=1e-3. All models used VORTEX methods used 1:1 balanced sampling between supervised and unsupervised examples (Desai et al., 2021a).

***Aug* Baselines and MRAugment.** Supervised augmentation baselines were trained with image-based and acquisition-based augmentations. Image-based augmentations for each dataset followed the augmentation configuration provided in the MRAugment. With the mridata 3D FSE knee dataset, integer rotations could only be conducted at 180 degrees due to the anisotropic matrix shape of the $ky \times kz$ slice. *Aug* baselines using physics-driven acquisition-based augmentations used a maximum probability of $p = 0.2$ as recommended by Desai et al. (2021a), and use the same range of σ for noise and α for motion that are used in the corresponding VORTEX experiments. Augmentations, their parameters, and their effective probabilities used for the mridata 3D FSE knee dataset are listed in Table 5. All augmentation methods were trained with the exponential augmentation probability scheduler with $\gamma = 5$ and a scheduling period equivalent to the training length as proposed by Fabian et al. (2021).

SSDU. SSDU is sensitive to the loss function and masking extent (ρ). Thus, these hyperparameters that should be optimized for different datasets. We sweeped through loss functions k-space ℓ_1, k-space ℓ_1-ℓ_2, and image ℓ_1 and masking extent $\rho = 0.2, 0.4, 0.6$. Models with the highest validation cPSNR were selected for all SSDU experiments. For the mridata 3D FSE knee dataset, the configuration with loss function k-space ℓ_1 and $\rho = 0.4$ was used. For the fastMRI multi-coil brain dataset, the configuration with ℓ_1-ℓ_2 loss in k-space and $\rho = 0.2$ was used.

Consistency Augmentations in VORTEX. Like *Aug* baselines, VORTEX was trained with combinations of image and physics-based augmentations. We use the same parenthetical nomenclature to indicate the augmentation type used in the consistency branch (e.g. *VORTEX (Motion)* for motion consistency). The family of image augmentations used for consistency in VORTEX were identical to those used in MRAugment. Physics-based consistency augmentations were sampled from either the light ($\mathcal{R}(\cdot)$=[0.1, 0.3)) or heavy ($\mathcal{R}(\cdot)$=[0.2, 0.5)) range during training.

D.4 Evaluation

D.4.1 Evaluation Settings

We perform evaluation in both in-distribution and clinically-relevant, simulated OOD settings. In-distribution evaluation consisted of evaluation on the test set described in D.1.

Table 6: Comparison of different scheduling methods and warmup periods on the mridata knee multi-coil dataset with heavy motion augmentations. All scheduling methods outperform non-scheduled training (base). There is no advantage of a specific scheduling protocol, suggesting that some curriculum is better than none.

Perturbation	None		Motion (light)		Motion (heavy)	
Curricula	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)
None	0.861	36.4	0.855	35.8	0.819	33.2
Linear (20e)	0.866	36.4	0.862	35.8	**0.828**	33.3
Linear (100e)	0.877	36.3	**0.871**	35.8	0.822	32.6
Linear (200e)	0.869	36.4	0.865	35.8	0.817	32.7
Exp (20e, $\gamma = 5$)	0.865	**36.4**	0.857	**35.9**	0.822	**33.4**
Exp (100e, $\gamma = 5$)	0.864	36.3	0.857	35.8	0.812	33.2
Exp (200e, $\gamma = 5$)	**0.877**	36.4	0.867	35.8	0.812	32.3

Table 7: Impact of training duration on cPSNR of supervised methods without augmentations (*Supervised*), supervised methods with motion augmentations (*Aug (Motion)*), MRAugment, and VORTEX with motion consistency (*VORTEX (Motion)*). Training duration are percentages of the full training duration (200 epochs). * indicates the default training configuration. Both supervised augmentation methods and MRAugment are more sensitive to training time than Supervised or VORTEX methods. Supervised underperforms Aug, MRAugment, and VORTEX. VORTEX achieves highest performance and is insensitive to training duration relative to the other methods.

	Perturbation		
Model	None	Motion (light)	Motion (heavy)
Supervised (10%)	35.0	33.3	27.4
Supervised (25%)	35.3	32.3	27.1
Supervised (50%)	35.5	32.0	26.4
Supervised (100%)*	35.8	33.6	27.0
Supervised (200%)	36.0	33.9	27.6
Supervised (300%)	36.0	33.9	27.6
MRAugment (10%)	35.4	32.3	26.0
MRAugment (25%)	35.8	31.5	25.1
MRAugment (50%)	36.0	31.5	24.3
MRAugment (100%)	36.2	31.8	24.0
MRAugment (200%)	36.3	32.2	24.3
MRAugment (300%)	36.4	33.4	25.0
Aug (Motion) (10%)	34.8	33.9	30.8
Aug (Motion) (25%)	35.3	34.5	31.4
Aug (Motion) (50%)	35.4	34.6	31.1
Aug (Motion) (100%)*	35.9	35.1	31.5
Aug (Motion) (200%)	36.0	35.1	30.8
Aug (Motion) (300%)	36.0	35.2	32.1
VORTEX (Motion) (10%)	36.2	35.5	32.4
VORTEX (Motion) (25%)	36.3	35.7	33.1
VORTEX (Motion) (50%)	36.4	35.8	33.2
VORTEX (Motion) (100%)*	36.4	35.8	33.2
VORTEX (Motion) (200%)	36.3	35.7	33.0
VORTEX (Motion) (300%)	36.3	35.7	33.0

For OOD evalution, we considered two critical settings that have been shown to affect image quality: (1) decrease in SNR and (2) motion corruption. The extent of the distribution shift is synonymous with the difficulty level for each perturbation (σ for noise, α for motion), where larger difficulty levels indicate correspond to larger shifts. Thus, we define *low* and *heavy* noise and motion difficulty levels for evaluation – low noise σ=0.2, heavy noise σ=0.4, low motion α=0.2, heavy motion $alpha$=0.4. These values are selected based on visual inspection of clinical scans (see 4.1.1). Note, by definition ($\sigma = 0$, α=0) corresponds to the in-distribution evaluation.

Table 8: Ablation for acceleration — 12x vs 16x. Like in the 16x regime, *VORTEX (Motion)* outperformed supervised methods, and MRAugment at 12x acceleration. This may suggest that VORTEX is broadly applicable to different acceleration levels.

	Perturbation	None		Motion (light)		Motion (heavy)	
Aug Range	Model	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)
N/A	Supervised 12x	0.814	36.2	0.814	32.4	0.689	25.4
	Supervised 16x	0.798	35.8	0.809	33.6	0.706	27.0
N/A	MRAugment 12x	0.828	36.5	0.814	31.9	0.637	23.6
	MRAugment 16x	0.811	36.2	0.793	31.8	0.660	24.0
N/A	SSDU 12x	0.819	34.9	0.816	34.5	0.762	30.9
	SSDU 16x	0.787	34.9	0.783	34.7	0.734	31.9
light	Aug (Motion) 12x	0.811	36.1	0.807	35.3	0.765	31.3
	Aug (Motion) 16x	0.802	35.6	0.793	34.7	0.739	30.4
heavy	Aug (Motion) 12x	0.818	36.1	0.811	35.2	0.758	31.2
	Aug (Motion) 16x	0.793	35.9	0.793	35.1	0.751	31.5
light	VORTEX Motion 12x	0.881	36.8	0.875	36.1	0.815	32.1
	VORTEX Motion 16x	0.882	36.4	0.875	35.7	0.813	31.5
heavy	VORTEX Motion 12x	0.888	36.7	0.883	36.1	0.846	33.5
	VORTEX Motion 16x	0.861	36.4	0.855	35.8	0.819	33.2

Given difficulty levels for motion and noise, each scan was perturbed by a noise or phase error (motion) maps generated with a set difficulty level. These perturbations were fixed for each testing scan to ensure reproducibiilty and identical perturbations in the test set across different experiments.

In the text, we refer to different evaluation configurations as *perturbations*. *None* indicates the in-distribution setting. *LN*, *HN*, *LM*, *HM* correspond to light noise, heavy noise, light motion, and heavy motion, respectively.

D.4.2 Metric Selection

Conventional computational imaging uses magnitude metrics for quantifying image quality. However, MRI images contain both magnitude and phase information (i.e. real and imaginary components). Because phase-related errors may not be captured by magnitude metrics, we use a combination of complex and magnitude metrics – complex PSNR (cPSNR) and magnitude SSIM metrics to quantify image quality. Equation 1 defines the cPSNR formulation for complex-valued ground truth x_{ref} and prediction x_{pred}. $||\cdot||_2$ corresponds to the complex-ℓ_2 norm and $|\cdot|$ denotes the magnitude of complex-valued input. Additionally, SSIM has shown to be a better corollary for MRI reconstruction quality compared to pSNR on magnitude images (Knoll et al., 2020). Thus, we use SSIM to quantify magnitude image quality.

$$\text{cPSNR (dB)} = 20\log_{10}\frac{\max|x_{ref}|}{||x_{pred} - x_{ref}||_2} \tag{1}$$

By default, metrics were computed over the full 3D scan. An additional set of metrics were also computed per reconstructed slice (termed *slice metrics*). Because different slices have different extents of relevant anatomy, per-slice metrics can provide a more nuanced comparison of 2D slice reconstructions among different methods. Statistical comparisons were conducted using Kruskal-Wallis tests and corresponding Dunn posthoc tests with Bonferroni correction (α=0.05). All statistical analyses were performed using the SciPy library.

E Ablations

Pixel-level vs Latent Space Consistency Setup. For the k^{th} resolution level, we enforce consistency after the final convolution in the encoder, and after the transpose convolution in the decoder. For $k = 4$, consistency is enforced at the bottleneck layer, after the convolution in the encoder. To

Table 9: Slice metrics (mean [standard deviation]) on the mridata knee dataset. Asterisk (*) indicates significant performance of VORTEX over *all* baselines (p <0.05).

Perturbation	None		Motion (heavy)		Noise (heavy)	
Model	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)
Supervised	0.635 (0.133)	29.7 (3.65)	0.545 (0.117)	21.4 (2.53)	0.591 (0.139)	25.8 (2.87)
MRAugment	0.653 (0.130)	30.1 (3.48)	0.505 (0.106)	18.6 (2.34)	0.563 (0.128)	25.0 (2.83)
SSDU	0.621 (0.147)	28.9 (3.39)	0.564 (0.146)	25.9 (3.49)	0.528 (0.142)	26.7 (2.77)
Aug (Physics)	0.623 (0.144)	29.6 (3.63)	0.566 (0.136)	26.0 (3.78)	0.557 (0.144)	27.6 (3.05)
Aug (Image+Physics)	0.618 (0.136)	30.1 (3.38)	0.565 (0.134)	**26.9 (3.79)**	0.540 (0.134)	27.9 (2.81)
VORTEX (Image)	0.718 (0.125)*	**30.4 (3.41)**	0.499 (0.110)	20.6 (2.25)	0.584 (0.104)	25.8 (2.47)
VORTEX (Physics)	**0.729 (0.138)***	30.3 (3.38)	**0.628 (0.137)***	26.0 (3.76)	0.653 (0.143)*	**28.1 (2.80)**
VORTEX (Image+Physics)	0.716 (0.131)*	30.3 (3.39)	0.616 (0.130)*	25.3 (3.69)	**0.658 (0.132)***	27.5 (2.73)

Table 10: Test performance (mean [standard deviation]) on the fastMRI multi-coil brain dataset at 8x acceleration. Results are shown on both in-distribution data and different motion levels of $\alpha = 0.6$, $\alpha = 0.8$, $\alpha = 1.0$ for Supervised, SSDU, MRAugment, augmentation baselines, and VORTEX.

Perturbation	None		Motion ($\alpha = 0.6$)		Motion ($\alpha = 0.8$)		Motion ($\alpha = 1.0$)	
Model	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)	SSIM	cPSNR (dB)
Supervised	0.851 (0.041)	30.1 (1.24)	0.652 (0.167)	19.0 (4.22)	0.585 (0.185)	16.5 (4.36)	0.564 (0.172)	14.8 (4.12)
SSDU	0.856 (0.036)	27.7 (1.26)	0.712 (0.177)	19.6 (4.29)	0.628 (0.209)	17.0 (4.60)	0.600 (0.197)	15.2 (4.40)
MRAugment	**0.869 (0.033)**	29.7 (1.31)	0.653 (0.166)	18.7 (4.18)	0.586 (0.191)	16.3 (4.33)	0.565 (0.181)	14.6 (4.09)
Aug (Motion, $\mathcal{R}(\alpha) = [0.2, 0.5)$)	0.836 (0.046)	28.4 (1.31)	0.695 (0.191)	21.3 (4.43)	0.618 (0.218)	18.4 (5.08)	0.594 (0.204)	16.4 (5.19)
Aug (Motion, $\mathcal{R}(\alpha) = [0.5, 0.7)$)	0.825 (0.044)	28.0 (1.40)	0.701 (0.186)	22.6 (4.00)	0.631 (0.218)	20.1 (4.73)	0.610 (0.206)	18.3 (4.68)
VORTEX (Image)	0.858 (0.035)	**30.2 (1.30)**	0.655 (0.166)	18.9 (4.22)	0.589 (0.190)	16.4 (4.33)	0.569 (0.178)	14.7 (4.11)
VORTEX (Image+Motion, $\mathcal{R}(\alpha) = [0.5, 0.7)$)	0.838 (0.041)	29.4 (1.45)	0.700 (0.139)	22.2 (3.31)	0.641 (0.156)	19.9 (3.95)	0.621 (0.147)	18.1 (4.15)
VORTEX (Motion, $\mathcal{R}(\alpha) = [0.2, 0.5)$)	0.839 (0.044)	29.7 (1.36)	0.726 (0.154)	23.3 (3.79)	0.664 (0.187)	20.1 (5.22)	0.649 (0.176)	17.8 (5.76)
VORTEX (Motion, $\mathcal{R}(\alpha) = [0.5, 0.7)$)	0.829 (0.046)	29.2 (1.44)	**0.763 (0.085)**	**24.4 (2.83)**	**0.726 (0.104)**	**22.8 (3.31)**	**0.710 (0.100)**	**21.5 (3.25)**

control for the impact of loss weighting, we normalize λ by the number of consistency losses that are computed in latent space when consistency is enforced at multiple resolution levels R_k. We compare these approaches in the case of light and heavy motion.

Training Time. We ablate the sensitivity of the performance of supervised, augmentation, and VORTEX methods to training duration. To compute the performance at different training duration, we select best checkpoints (quantified by validation cPSNR) up to a given duration and run evaluation using these weights. As all methods were trained for 200 epochs, we compare the performance at training times of 10% (20 epochs), 25% (50 epochs), 50% (100 epochs), and 100% (200 epochs). Supervised methods were insensitive to training time, but considerably underperformed both supervised augmentation (Aug) and VORTEX (Table 7). Aug was sensitive to training time, with changes in cPSNR of >1dB. VORTEX achieved the highest performance across all metrics and evaluation setups and was relatively insensitive to training duration.

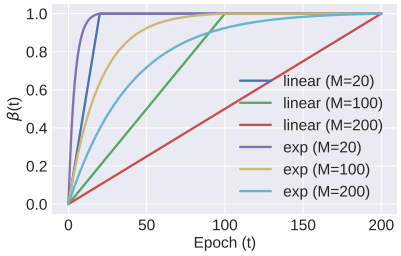

Figure 5: Simulated augmentation difficulty scheduling over training period of 200 epochs using linear and exponential (exp) schedulers defined in §4.2. Time constant for exponential scheduling $\tau = \frac{M}{\gamma}$ where γ=5.

Sensitivity to Acceleration Factors. We evaluated the performance of VORTEX at different acceleration factors in Table 8. At 12x acceleration, VORTEX trained with heavy motion recovered +6.1 SSIM and +0.8 dB cPSNR compared to the *Supervised* baseline in the in-distribution setting. At the same acceleration, VORTEX also outperformed the *Supervised* baseline by +15.7 SSIM and 8.1 dB cPSNR. The stability of VORTEX at different accelerations may indicate that VORTEX is generalizable across different acceleration extents.

F Extended Results

In this section, we provide results for the mridata dataset using slice metrics and for the fastMRI multi-coil brain dataset (Zbontar et al., 2018).

F.1 SLICE METRICS

Table 9 shows slice metrics of baselines and VORTEX on the mridata knee dataset. Among slice metrics, VORTEX also outperforms all baselines in both in-distribution and OOD settings. In particular, VORTEX significantly outperformed all baselines in SSIM in all evaluation settings ($p<0.05$). This may indicate that VORTEX has higher fidelity in recovering image structure even in OOD settings where perturbations can result in a considerable degradation in SSIM.

F.2 FASTMRI RESULTS

We compare VORTEX to Supervised, SSDU, Aug (Motion), and MRAugment baselines for in distribution and OOD motion settings of different motion levels on the fastMRI multi-coil brain dataset in Table 10 (Zbontar et al., 2018). Data preparation and experimental details follow the description in Appendix D, and all experiments are conducted at 8x acceleration. We demonstrate that VORTEX has comparable performance to baselines for in distribution, and outperforms SSDU by +5.1 SSIM and +4.8 cPSNR, and MRAugment by +11.1 SSIM and +5.7 cPSNR on motion level $\alpha = 0.6$; SSDU by +9.8 SSIM and +5.8 cPSNR, and MRAugment by +14 SSIM and +6.5 cPSNR on motion level $\alpha = 0.8$; SSDU by +11 SSIM and +6.3 cPSNR, and MRAugment by +14.5 SSIM and +6.9 cPSNR on motion level $\alpha = 1.0$. This demonstrates that the effectiveness of VORTEX for both in distribution and OOD data generalizes to 2D MRI sequences which implies broader clinical utility.