Using Deep Feature Distances for Evaluating MR Image Reconstruction Quality

Philip M. Adamson[1] **Arjun Desai**[1] **Jeffrey Dominic**[1] **Christian Bluethgen**[2]
Jeff P. Wood[4] **Ali B. Syed**[2] **Robert D. Boutin**[2] **Kathryn J. Stevens**[2]
Shreyas Vasanawala[2] **John M. Pauly**[1] **Akshay S. Chaudhari**[2,3] **Beliz Gunel**[1]
Departments of [1]Electrical Engineering, [2]Radiology, and [3]Biomedical Data Science, Stanford University [4]Austin Radiological Association
{padamson, arjundd, jdomini, bluethgen, alibsyed, boutin, kate.stevens, vasanawala, pauly, akshaysc, bgunel}@stanford.edu

Abstract

Evaluation of MR reconstruction methods is challenged by the need for image quality (IQ) metrics which correlate strongly with radiologist-perceived IQ. We explore Deep Feature Distances (DFDs) as MR reconstruction IQ metrics, whereby distances between ground truth and reconstructed MR images are computed in a lower-dimensional feature space encoded by a CNN. In addition to comparing DFDs to two commonly used pixel-based MR IQ metrics in PSNR and SSIM via correlations to radiologist reader scores of MR image reconstructions, we explore the impact of domain shifts between the DFD encoder training data and the evaluated MR images. In particular, we assess two state-of-the-art but "out-of-domain" DFDs with encoders trained on natural images, an in-domain DFD trained on MR images alone, and propose two domain-adjacent DFDs trained on large medical imaging datasets (not limited to MR data). IQ metric performance is assessed via their correlations to 5 expert radiologist reader scores of MR image reconstructions. We make three striking observations: 1) all DFDs out-perform traditional IQ metrics, 2) DFDs performance approaches that of radiologist inter-reader variability, and, 3) surprisingly, out-of-domain DFDs perform comparably as an MR reconstruction IQ metric to in-domain and domain-adjacent DFDs. These results make it evident that DFDs should be used alongside traditional IQ metrics in evaluating MR reconstruction IQ, and suggest that general vision encoders are able to assess visual IQ across image domains.

1 Introduction

Accelerated MR reconstruction is a common inverse in which the goal is to recover an aliasing-free image from a set of under sampled frequency-domain measurements. Compressed sensing [1] and Deep Learning (DL) [2, 3, 4, 5, 6] approaches have show great promise, but systematic evaluation is challenging because diagnostic information is not easily quantified at the pixel-level. Metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are routinely used to assess the image quality (IQ) of MR reconstructions [7], but have been shown to correlate poorly with radiologist review [8, 9, 10]. Thus, there is a pressing need to develop IQ metrics that better correlate with radiologist perceived diagnostic quality and downstream clinical utility.

The discordance between IQ metrics and human-perceived IQ is a well-known challenge in the natural image computer vision community. Recently, Zhang et. al. showed that deep feature distances (DFDs) correlate strongly with human perceived IQ, proposing the DFD Learned Perceptual Image

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Patch Similarity (LPIPS) [11]. Generally, we define a DFD (δ) between the fully-sampled ground truth image $x \in \mathbb{R}^{CxHxW}$ and the reconstructed image $\hat{x}$ as

$$\delta_l(x, \hat{x}) = G(\phi_D^{(l)}(x), \phi_D^{(l)}(\hat{x})) \tag{1}$$

where G is a distance measure, and $\phi_D^{(l)}$ maps from the image space $\mathbb{R}^{CxHxW}$ to feature space $\mathbb{R}^{C_l x H_l x W_l}$ via a CNN encoder trained on a dataset D with features extracted from convolutional layer l. LPIPS uses an L2-norm for G, VGG-16 [12] or AlexNet [13] trained on ImageNet [14] for ϕ_D, and additionally learns a linear combination of DFDs extracted from five different convolutional layers l based on perceptual judgement scores.

While LPIPS and other DFDs outperform traditional IQ metrics in terms of correlation with human perceptual judgments in large scale computer vision studies [15], these studies evaluate DFDs on natural images belonging to the same domain as the images used for encoder training D. One study has shown that LPIPS out-performs traditional IQ metrics in an MR image-based reader study, but does not explore if these out-of-domain DFDs are optimal for MR images [16]. It further evaluates MR images with artificial corruptions, rather than clinically-feasible accelerated MR reconstruction corruptions. Building on prior work from the in-domain self-supervised feature distance (SSFD) trained in a self-supervised manner on MR images [17], we assess the impact of the domain of D on the correlation of DFDs to radiologist reader scores of DL-based accelerated MR reconstructions. To this end, we compare traditional IQ metrics to two state-of-the-art out-of-domain DFDs trained on natural images, the in-domain SSFD trained on MR images, and two novel domain-adjacent DFDs trained on a large corpora of medical images (not limited to MR images).

2 Methods

2.1 MR Image Reconstructions

We used the fastMRI multi-coil knee dataset with both sparse and fully acquired k-space data for DL-based accelerated MR reconstructions [18]. Supervised DL-reconstruction models were trained to reconstruct 2x, 4x and 6x accelerated scans using both a UNet model and an unrolled network [2]. The UNet models followed the architecture in the fastMRI challenge [18], while the unrolled models followed the fast iterative shrinkage-thresholding algorithm (FISTA) unrolled architecture [19] implemented in [2]. Reference images were computed from the fully-sampled k-space data with the JSENSE method to integrate coil sensitivities [20]. We split the dataset into training, validation, and testing splits with 27,774 slices (778 3D scans), 6,968 slices (195 scans), and 7,135 slices (199 scans) respectively. Additional architecture and training details are provided in Appendix 6.1.

2.2 Reader Study Criteria

Five radiologists rated the diagnostic quality of the center slice of 366 accelerated MR reconstructions from 61 patients, each reconstructed with the 6 models described above. The reader study was blinded such that readers did not have access to the image IDs or reconstruction methods, nor to the scores provided by other readers. The radiologists provided two scores for the image reconstructions, one each for aliasing artifacts and for the diagnostic quality of the cartilage and meniscus (the most commonly evaluated tissues on knee MRI) on the following 1-9 scale: 1- completely non-diagnostic, 3- severe corruptions, 5- diagnostically acceptable, 7- good quality, 9- perfect quality. The mean radiologist IQ score was compared against various IQ metrics from which the Spearman Rank Order Correlation Coefficient (SROCC) was computed. SROCC measures the strength of the monotonic (but not necessarily linear) relationship between the reader scores and IQ metrics.

2.3 Image Quality Metrics

PSNR and SSIM were used as baselines for evaluating IQ in the reader study using implementations described previously [21, 22].

2.3.1 Out-of-Domain DFDs

We evaluate several DFDs used in natural domain IQ assessment [15] to serve as benchmarks for the domain-specific DFDs proposed in this study. We refer to these DFDs as "out-of-domain" in the sense that natural images belong to a different image domain than the MR images on which the DFDs were computed and evaluated in the reader study. In addition to LPIPS, we use the Deep Image Structure and Texture Similarity (DISTS) index [23]. DISTS uses the same $\phi^{(l)}_{D_{tr}}$ as LPIPS, but with a distance function G inspired by the form of SSIM that assesses texture and structure similarity of the feature maps. Like LPIPS, DISTS combines DFDs from various layers l as a learned weighted sum. Additional implementation details are provided in Appendix 6.2.1.

2.3.2 In-Domain DFD

We used SSFD [17] as an example "in-domain" DFD, as its encoder $\phi^{(l)}_{D}$ is trained in a self-supervised fashion on the same MR dataset D used to train the reader study DL reconstruction models. Details on the self-supervised masking task [24, 25] and model training are given in Appendix 6.2.2.

2.3.3 Domain-Adjacent Feature Distances

We propose two additional DFDs in this study leveraging pre-existing DL models trained on large medical imaging datasets. We refer to these as "domain-adjacent" DFDs since the medical imaging datasets D are not limited to the fastMRI dataset used in the reader study reconstructions, and contain medical images from modalities beyond MR.

We leverage the RadImageNet model [26], a ResNet50 pretrained on 1.4 million labeled medical images (including 670,000 MRI images) trained in a supervised manner, to compute a RadImageNet Feature Distance (RINFD). To disentangle the model architecture ϕ from the training data D, we also compared to both a ResNet50 trained with ImageNet, and a ResNet50 with randomly initialized, untrained weights. Additional RINFD implementation details are provided in Appendix 6.2.3.

We also propose a Medical Variational Autoencoder Feature Distance (Med-VAEFD) using a medical variational autoencoder (Med-VAE) designed for neural compression trained on dataset D with 1 million radiograph and mammography images [27, 28]. Med-VAE is a convolutional VAE trained with a combination of a perceptual loss, a patch-based adversarial objective, and a penalty based on the Kullback-Leibler (KL) divergence [27, 28]. We use the VAE bottleneck latent embeddings (their compressed representations) for $\phi^{(l)}_{D_{tr}}$. Additional Med-VAEDF implementation details are provided in Appendix 6.2.3.

3 Experimental Results and Discussion

All IQ metrics versus mean reader score SROCC values are shown in Figure 1. Reader score versus IQ metric correlation plots for an example traditional IQ metric (PSNR), out-of-domain DFD (LPIPS), in-domain DFD (SSFD) and domain-adjacent DFD (RINFD) are shown in Figure 2. All DFDs substantially outperform traditional IQ metrics, and even approach or exceed human-level performance in terms of inter-reader variability (SROCC of 0.85). More details on IRV can be found in Appendix 6.4).

Interestingly, we find that out-of-domain DFDs perform comparably to in-domain and domain-adjacent DFDs. In the fixed-encoder architecture case for example, training the ResNet50 encoder with the out-of-domain ImageNet dataset (Aliasing SROCC 0.86, Cartilage/Meniscus SROCC 0.85) performs comparably to training with the domain-adjacent RadImageNet dataset (Aliasing SROCC 0.85, Cartilage/Meniscus SROCC 0.83). This is a surprising finding considering the importance of domain-specific pretraining in comparable transfer learning studies [26, 25]. One potential explanation for this finding is that transfer learning benefits from task-specific pretraining for eventually solving specific downstream tasks, while general vision encoders are strong enough feature extractors for assessing visual IQ, regardless of the image domain.

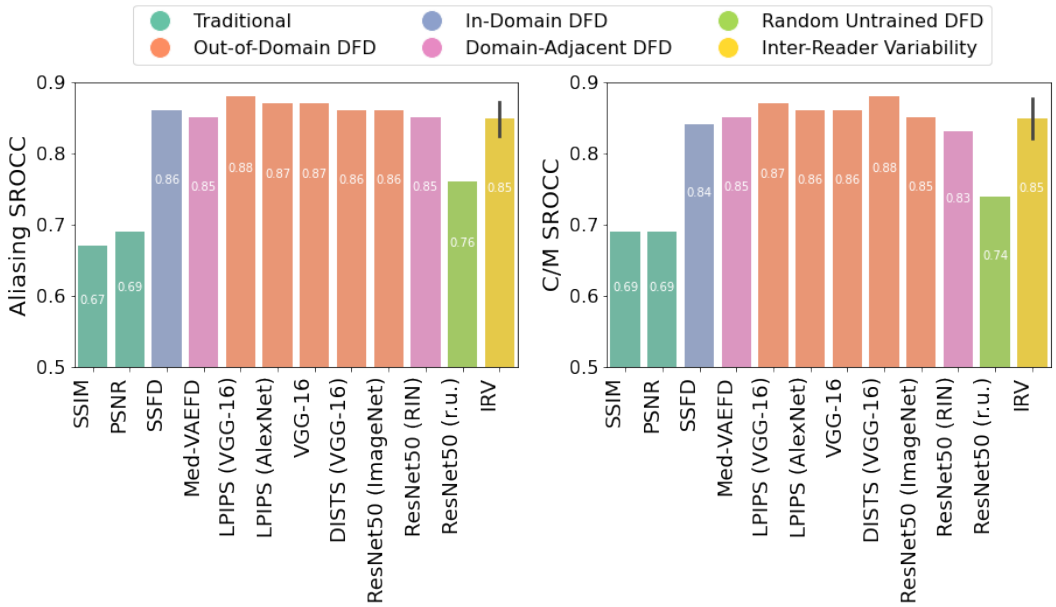

Figure 1: Mean reader score correlations to traditional and DFD IQ metrics based on encoder training data domain D. Aliasing reader score SROCC values are shown on left with cartilage and meniscus reader score SROCCs on the right. DFDs out-perform tradition IQ metrics and are comparable to inter-reader variability (IRV), but out-of-domain DFDs perform comparably as an MR reconstruction IQ metric to in-domain and domain-adjacent DFDs.

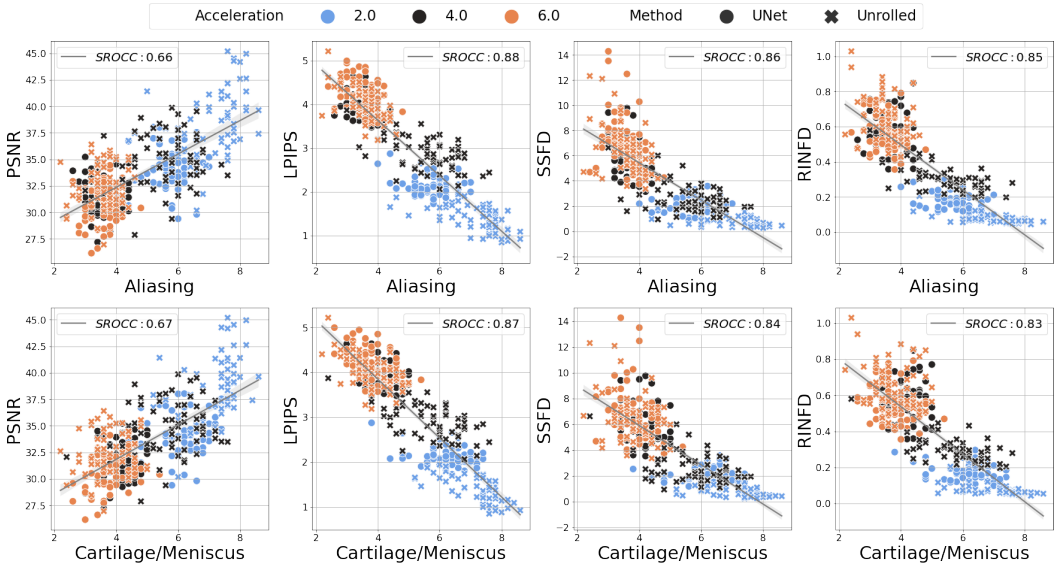

Figure 2: Example traditional (PSNR), out-of-domain (LPIPS), in-domain (SSFD) and domain-adjacent (RINFD) IQ metric values versus mean reader scores for aliasing (top) and cartilage/mensiscus assessment (bottom). Each point corresponds to a single image taken from the center slice from 61 MR reconstruction images, each with 2x (blue), 4x (black) and 6x (orange) accelerations with a UNet (circle) and unrolled (X's) networks. Higher reader score values correspond to better radiologist perceived IQ.

4 Conclusion

In this study, we explore the utility of DFDs as an improved IQ metric over pixel-vased metrics for MR image reconstruction. We find that all DFDs out-perform traditional IQ metrics for MR reconstruction in terms of correlation with the gold-standard radiologist perceived diagnostic IQ. Remarkably, we find that DFDs perform as well or better than radiologist IRV. We also make the surprising finding that DFDs perform comparably as an MR reconstruction IQ metric across all encoder training dataset domains, indicating that general vision encoders are sufficient for visual IQ assessment across image domains.

References

[1] Michael Lustig, David Donoho, and John M Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.

[2] Christopher M Sandino et al. "Compressed sensing: From research to clinical practice with deep neural networks: Shortening scan times for magnetic resonance imaging". In: *IEEE signal processing magazine* 37.1 (2020), pp. 117–127.

[3] Jian Sun, Huibin Li, Zongben Xu, et al. "Deep ADMM-Net for compressive sensing MRI". In: *Advances in neural information processing systems* 29 (2016).

[4] Vahid Ghodrati et al. "MR image reconstruction using deep learning: evaluation of network structure and loss functions". In: *Quantitative imaging in medicine and surgery* 9.9 (2019), p. 1516.

[5] Kerstin Hammernik et al. "Systematic evaluation of iterative deep neural networks for fast parallel MRI reconstruction with sensitivity-weighted coil combination". In: *Magnetic Resonance in Medicine* 86.4 (2021), pp. 1859–1872.

[6] Kerstin Hammernik et al. "Learning a variational network for reconstruction of accelerated MRI data". In: *Magnetic resonance in medicine* 79.6 (2018), pp. 3055–3071.

[7] Yutong Chen et al. "AI-based reconstruction for fast MRI—a systematic review and meta-analysis". In: *Proceedings of the IEEE* 110.2 (2022), pp. 224–245.

[8] Allister Mason et al. "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images". In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1064–1072.

[9] Akshay S Chaudhari et al. "Prospective deployment of deep learning in MRI: A framework for important considerations, challenges, and recommendations for best practices". In: *Journal of Magnetic Resonance Imaging* (2020).

[10] Florian Knoll et al. "Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge". In: *Magnetic resonance in medicine* 84.6 (2020), pp. 3054–3070.

[11] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[12] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[14] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[15] Jinjin Gu et al. "NTIRE 2022 challenge on perceptual image quality assessment". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 951–967.

[16] Sergey Kastryulin et al. "Image quality assessment for magnetic resonance imaging". In: *IEEE Access* 11 (2023), pp. 14154–14168.

[17] Philip M Adamson et al. "SSFD: Self-supervised feature distance as an MR image reconstruction quality metric". In: *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*. 2021.

[18] "fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning." In: *Radiol Artif Intell* 2.1 (Jan. 2020), e190007. ISSN: 2638-6100 (Electronic); 2638-6100 (Linking). DOI: 10.1148/ryai.2020190007.

[19] Liqi Xin, Dingwen Wang, and Wenxuan Shi. "FISTA-CSNet: a deep compressed sensing network by unrolling iterative optimization algorithm". In: *The Visual Computer* 39.9 (2023), pp. 4177–4193.

[20] Leslie Ying and Jinhua Sheng. "Joint image reconstruction and sensitivity estimation in SENSE (JSENSE)". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 57.6 (2007), pp. 1196–1202.

[21] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[22] Arjun D Desai et al. "Noise2Recon: Enabling SNR-robust MRI reconstruction with semi-supervised and self-supervised learning". In: *Magnetic Resonance in Medicine* 90.5 (2023), pp. 2052–2070.

[23] Keyan Ding et al. "Image quality assessment: Unifying structure and texture similarity". In: *IEEE transactions on pattern analysis and machine intelligence* 44.5 (2020), pp. 2567–2581.

[24] Deepak Pathak et al. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.

[25] Jeffrey Dominic et al. "Improving Data-Efficiency and Robustness of Medical Imaging Segmentation Using Inpainting-Based Self-Supervised Learning". In: *Bioengineering* 10.2 (2023), p. 207.

[26] Xueyan Mei et al. "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning". In: *Radiology: Artificial Intelligence* 4.5 (2022), e210315.

[27] Rogier Van der Sluijs et al. "Diagnostically Lossless Compression of Medical Images". In: *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*. 2023. URL: https://openreview.net/forum?id=ZiNFhNFxMf.

[28] Maya Varma et al. "CompRx: A Benchmark for Diagnostically Lossless Compression of Medical Images". In: 2023.

5 Acknowledgments and Disclosure of Funding

This work was supported by NIH R01 AR077604, R01EB002524, R01 AR079431. It was also supported by the Radiological Sciences Laboratory Seed Grant from Stanford University, and the NSF Graduate Research Fellowship under Grant No. DGE-2146755.

6 Appendix

6.1 DL Reconstruction Models

All DL MR reconstruction models were trained in meddlr, a PyTorch based ML framework for medical image reconstruction problems [22]. The DL reconstruction models for the reader study were trained with a 54 image subset of the fastMRI training dataset described in Section 6.2.2. This was to reduce computation costs of training reconstruction models as the aim of the reader study was to generate a range of realistic image qualities, rather than achieve the best possible reconstructions in each setting.

6.1.1 UNet

The UNet models were trained with complex inputs and outputs, 2 convolutions per layer and 4 levels with 32-256 quadratically increasing filters as per the fastMRI paper implementation [18]. Each model was trained for 129,000 iterations using the Adam optimizer with a learning rate of 1e-4, weight decay of 1e-4, and a complex L1 image loss function.

6.1.2 Unrolled

The unrolled models follow the fast iterative shrinkage-thresholding algorithm (FISTA) unrolled architecture [19] implemented in [2]. The network was unrolled for 8 steps, and a ResNet with two residual blocks (2 convolutional layers and 128 filters each) was used to model the proximal update operation at each step. Each model was trained for 129,000 iterations using the Adam optimizer with a learning rate of 1e-4, weight decay of 1e-4, and a complex L1 k-space loss function.

6.2 DFD Implementation Details

All DFDs were implemented in meddlr [22].

6.2.1 Out-of-Domain DFDs

LPIPS and DISTS were both implemented in meddlr using their respective Python packages. For both LPIPS and DISTS pre-processing, the magnitude of the complex MR reconstruction was taken, followed by replicating the image 3x channel wise to create a pseudo-rgb image. The images were then preprocessed to values between -1 and 1 for LPIPS and between 0 and 1 for DISTS, per their respective Github documentations [11, 23]. Three configurations of LPIPS were tested - one with a VGG-16 backbone, one with the AlexNet backbone, and third with the VGG-16 backbone but without the linear layer fine-tuned on reader scores from the LPIPS study.

6.2.2 In-Domain DFD

The self-supervised model was trained using the full training dataset described in Section with a masked inpainting pre-text task. Image corruptions for the context prediction task were generated dynamically during training by placing zero-filled image patches of size 16x16 pixels over 50% of the image area via Poisson variable density sampling (to ensure non-overlapping patches). A self-supervised UNet model (with 2 convolutions per level and 5 levels with 20-320 quadratically increasing filters) was trained to in-paint the zero-filled patches and restore the original image subject to an L2 loss. Parameters were initialized based on a prior transfer learning study for knee cartilage segmentation [25] and modified empirically to maximize SROCC. We use SSFD with the 7th convolutional layer for l and the MSE distance function for G, determined empirically.

6.2.3 Domain-Adjacent DFDs

RINFD uses a ResNet50 pretrained on the RIN dataset $\phi_D^{(l)}$, convolutional layer $l = 23$ (chosen empirically), and the MSE distance function for G. The DFDs using ResNet50 models trained on ImageNet and with randomly initialized, untrained weights use the same l and G as RINFD. Note that these parameter choices maximized SROCC for all 3 choices of D. Images were preprocessed by first taking the magnitude of the complex MR reconstruction, followed by replicating the image 3x channel wise to create a pseudo-rgb image, followed by normalization between 0 and 1 per the RadImageNet Github documentation [26].

Med-VAEFD specifically uses the neural compressor with an in-plane compression factor of 8, and 4 latent channels from [28]. Images were pre-processed by first taking the magnitude of the complex MR reconstruction, followed by replicating the image 3x channel wise to create a pseudo-rgb image. Images were then normalized to 0.5 mean, 0.5 variance.

6.3 Reader Study MR Reconstruction Examples

Example images of each reconstruction type used in the reader study for a single MR slice along with several IQ metrics and mean reader scores are shown in Figure 3.

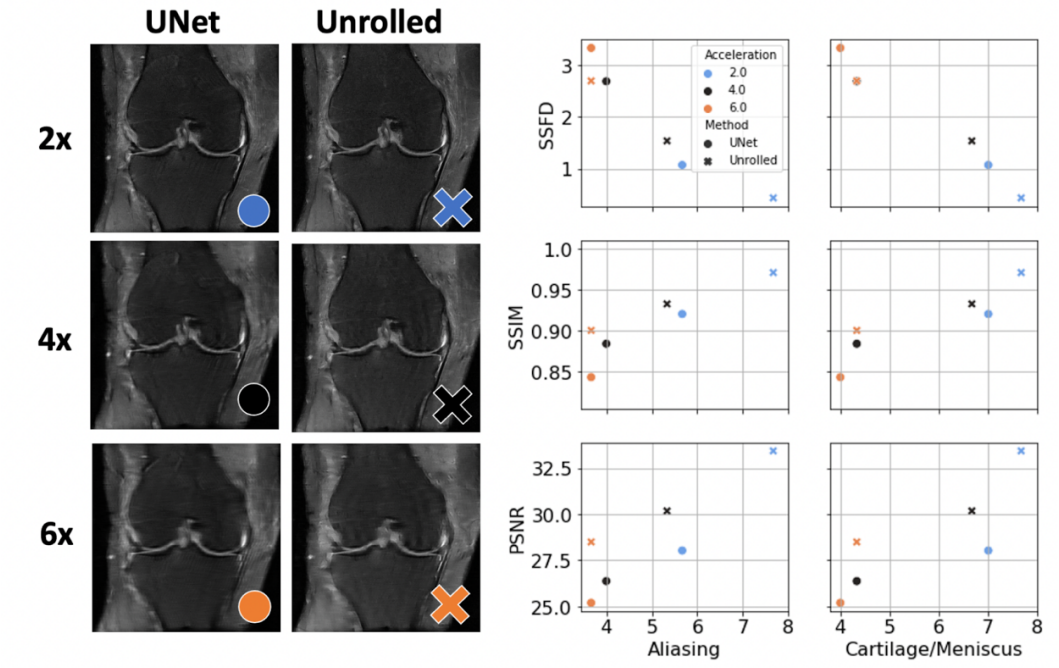

Figure 3: Example of the six reconstruction techniques (left), and plots of their IQ metrics versus mean radiologist IQ score (right). Reader scores decrease for higher accelerations (blue to orange) and are higher for the unrolled (X) versus UNet (circle) generated images.

6.4 Inter-reader Variability

Per-reader inter-reader correlations are shown in Figure 4. The reader score of each reader is plotted against the mean reader score of the other four readers for each image in the reader study. The mean correlation across all readers is the inter-reader variability (IRV), an SROCC of 0.85.

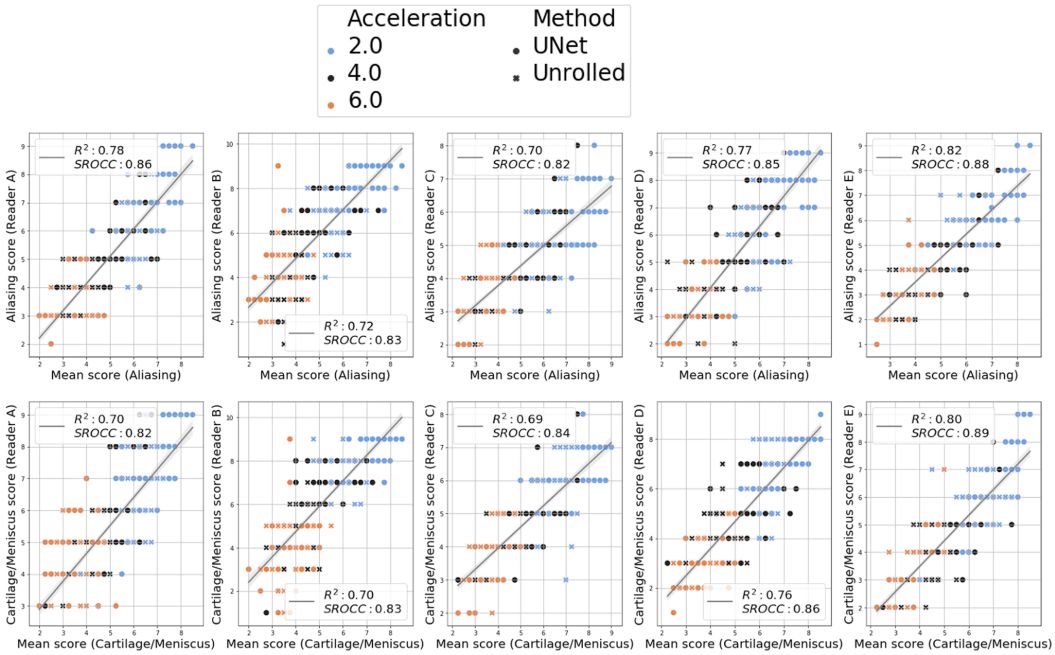

Figure 4: Per-reader correlations with the mean of the withheld readers. The reader score of each reader is plotted against the mean reader score of the other four readers for each image in the reader study.