SSFD: Self-Supervised Feature Distance as an MR Image Reconstruction Quality Metric

Philip M. Adamson
Stanford University
padamson@stanford.edu

Beliz Gunel
Stanford University
bgunel@stanford.edu

Jeffrey Dominic
Stanford University
jdomini@stanford.edu

Arjun Desai
Stanford University
arjundd@stanford.edu

Daniel Spielman
Stanford University
spielman@stanford.edu

Shreyas Vasanawala
Stanford University
vasanawala@stanford.edu

John M. Pauly
Stanford University
pauly@stanford.edu

Akshay S. Chaudhari
Stanford University
akshaysc@stanford.edu

Abstract

Evaluation of accelerated magnetic resonance imaging (MRI) reconstruction methods is imperfect due to the discordance between quantitative image quality metrics and radiologist-perceived image quality. Self-supervised learning (SSL) has become a popular pre-training tool due to its ability to capture generalizable and domain-specific feature representations of the underlying data for downstream tasks. In this study, we use SSL to extract image-level feature representations of MR images, and use those features to compute a self-supervised feature distance (SSFD) metric to assess MR image reconstruction quality. We demonstrate preliminary results showing the superiority of SSFD to common image quality metrics such as PSNR and SSIM, its robustness to image perturbations, and its ability to capture both pixel-level and global image quality information.

1 Introduction

Deep learning (DL) algorithms have achieved state-of-the-art performance on accelerated MR reconstruction tasks, where the goal is to reconstruct high quality images from a set of undersampled Fourier domain (k-space) measurements [1, 2, 3, 4]. While metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are routinely used to assess image quality of MR reconstructions, these do not always correlate well with clinical utility for downstream tasks and radiologist review, the gold-standard for image quality assessment [5, 6, 7]. Thus, it is imperative to develop quantitative metrics with higher concordance to clinical utility and radiologist-perceived image quality. Such a metric would allow for algorithm developers to more quickly iterate through and compare reconstruction algorithms without the need for expensive reader studies.

The VGG-16 perceptual loss (VGG-PL) is a loss function that operates in the feature (latent) space of a VGG-16 network trained on ImageNet, and has been used for optimizing perceptual image quality in inverse problems such as super-resolution [8]. VGG-PL has since been used as a distance metric for measuring image quality in natural images, outperforming traditional image quality metrics such as SSIM [9]. However, since features from MR images are fundamentally different from natural images due to their content, contrasts, and noise characteristics, the VGG-PL representations may be sub-optimal for assessing MR reconstruction quality. Furthermore, a large, labelled dataset such as

NeurIPS 2021 Workshop on Deep Learning and Inverse Problems, virtual.

ImageNet for each medical image domain is a burden to label accurately, which necessitates alternative methods for learning feature representations for medical images without requiring radiologist labels. Prior work has used an Unsupervised Feature Loss "UFLoss", whereby a patch-wise feature mapping function is learned such that each patch is maximally separated from all other patches in the feature space, but the local patch-based training cannot embed global image-level dependencies in the learned features [10]. Unsupervised image quality metrics have also been used previously, but these have been limited to assessing only blurring [11].

Self-supervised learning (SSL) uses pre-text tasks to generate labels from unlabeled data, and casts unsupervised problems into supervised ones. In this study, we propose to learn image-level feature representations of the underlying MR data via the pre-text task of context prediction, whereby a self-supervised model learns an inpainting task on the MRI domain of interest, following a pre-text task of randomly masking out small patches from the images. SSL has been shown to learn good image-level feature representations of both natural and medical images from unlabeled datasets, demonstrating that these representations may also be useful in calculating a distance metric for assessing image quality [12, 13]. We hypothesize that this distance metric will be better suited as an MR image quality metric than the pixel-wise metrics of SSIM and PSNR, or the VGG-PL based on natural image feature representations.

2 Methods

2.1 Dataset and Supervised Reconstruction Models

We used the fastMRI knee dataset with both sparse and fully acquired k-space data from a multi-coil MR scanner [14]. Reference images were computed from the fully-sampled k-space data with the JSENSE method to integrate coil sensitivities [15]. For the SSL task and DL-based undersampled MR reconstruction, 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, each of size 320x320 pixels with 0.44 mm resolution. Supervised DL-reconstruction models were trained to reconstruct 4x accelerated scans using a UNet model (with complex inputs, 2 convolutions per layer and 4 levels with 32-256 quadratically increasing filters).

2.2 Self-supervised Learning: Context Prediction

Image corruptions for the context prediction task were generated dynamically during training by placing zero-filled image patches of size 16x16 pixels over 25% of the image area via Poisson variable density sampling (to ensure non-overlapping patches), as shown in Appendix 5.1. 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.

2.3 Self-supervised Feature Distance (SSFD)

The encoder of the pre-trained SSL UNet described in Section 2.2 was used to calculate a self-supervised feature distance (SSFD) metric. The encoder was truncated following ReLU activations of a given convolutional layer in the UNet encoder. Ground truth and reconstructed image pairs were then separately passed through this truncated model, producing two feature space outputs. The SSFD was the element-wise mean square error between the two feature representations of the image pair. Unless otherwise stated, the 9^{th} convolutional layer (network bottleneck) was used for computing SSFD on the center slice of the 3D DL-based reconstructions, from the 199 test set images for the following three studies:

SSFD under Image Perturbations: DL-reconstructions are sensitive to input perturbations [16]. To understand the robustness of metrics under known image perturbations, we explored the impact of pixel shifts, Gaussian blurring, and additive Gaussian noise on SSFD, SSIM (Fig.1) and PSNR (Appendix 5.3). Each perturbation was applied to the center slice for each of 199 scans in the test set after normalizing to zero mean, unit variance. The maximum extent of each image perturbation was chosen such that the average SSIM between the ground truth and corrupted images reached 0.3. Pixel shifts were applied by rolling pixels in the x-direction with tricubic interpolation from 0 to 0.53 mm (1.2 pixels). The standard deviation of the Gaussian kernel for blurring increased linearly between 0

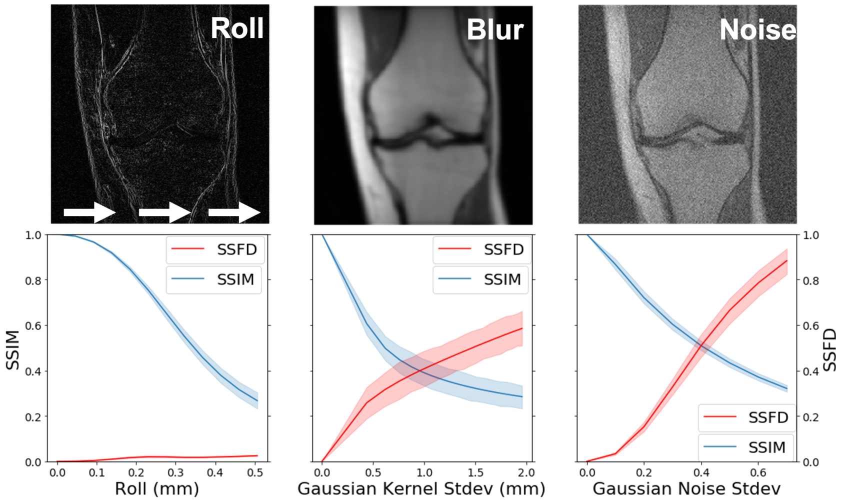

Figure 1: Example images at the 50^{th} percentile of perturbations, where the roll image is the difference between the perturbed and original scan (top). Average SSFD and SSIM with 95% confidence intervals from the 199 MR scan test set as a function of image perturbation (bottom). SSFD is less sensitive to pixel shifts compared to linearly increasing Gaussian standard deviation for blurring and noise for comparable decreases in SSIM.

and 1.97 mm (4.5 pixels). The additive Gaussian noise standard deviation increased linearly from 0 to 0.8.

SSFD versus Encoder Layer: To assess the characteristics of the features learned at different layers in the network, we computed SSFD for 4 different convolutional layers (3^{rd}, 5^{th}, 7^{th}, 9^{th}) at different resolution levels. The layer-wise SSFD was computed and compared to the overall SSIM values between the DL-based and gold-standard reconstructions (Fig. 2). We also explore the impact of fat-suppressed vs. non-fat-suppressed MR images on SSFD with a hypothesis that the feature representations of the two image types differ due to their vastly different image contrasts (as shown in Appendix 5.4).

SSFD as a Quality Control Tool: To assess if image quality metrics may be insensitive to poor reconstructions, we qualitatively compared the MR reconstruction to quantitative metrics of SSIM, SSFD, and VGG-PL (Fig.3), as well as PSNR (Appendix 5.3). Our implementation of VGG-PL can be found in Appendix 5.2.

3 Experimental Results and Discussion

The average SSFD and SSIM metrics under image perturbations (Fig. 1) showed that SSFD increased approximately linearly under Gaussian noise and blurring perturbations, while SSIM decreased approximately as a decaying exponential. A linear relationship between image perturbation and SSFD is desirable so that SSFD penalizes both small and large perturbations proportionally, whereas a decaying loss metric saturates despite increasing perturbations. Notably, SSFD was comparatively insensitive to pixel shifts, indicating that SSFD captures more global image quality features that are less sensitive to exact pixel-level correspondence than SSIM and PSNR, and relaxing constraints that require pixel-level correspondence between scans.

SSFD behavior at different convolutional layer depths (Fig. 2) showed that earlier layers in the encoder correlate more closely with pixel-level metrics such as SSIM, indicating they capture simpler lower-level pixel information. The correlations become weaker deeper into the network where more complex representations are extracted in a way that is not possible with SSIM and PSNR. SSFD therefore provides flexibility in choosing the complexity of feature representations of interest, depending on the convolutional layer used to extract feature representations. A linear combination of SSFD values calculated from different layers then becomes a hyperparameter which could be

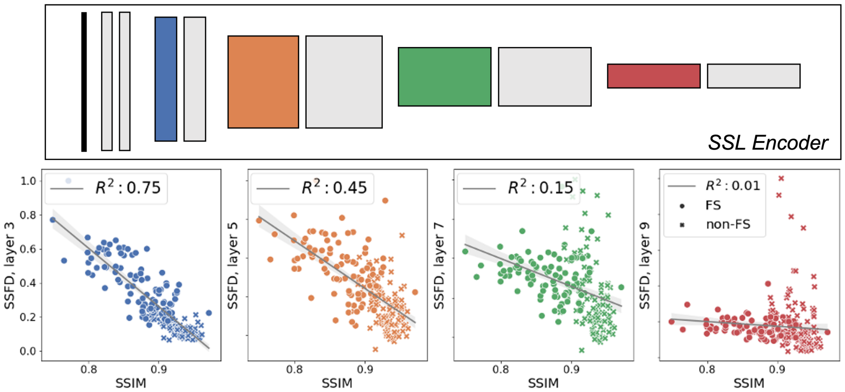

Figure 2: SSFD vs. SSIM for 4 different encoder layers. Each point represents the center slice from each of 199 MR reconstructions including both fat-suppressed (FS) and non-fat-suppressed (non-FS) acquisitions. SSFD is more highly correlated with SSIM higher up in the encoder network, indicating that layers earlier in the network learn simpler pixel-level feature representations, compared to more complex features deeper in the network. We also observe stronger clustering in terms of SSFD of FS vs. non-FS in the earlier layers of the network.

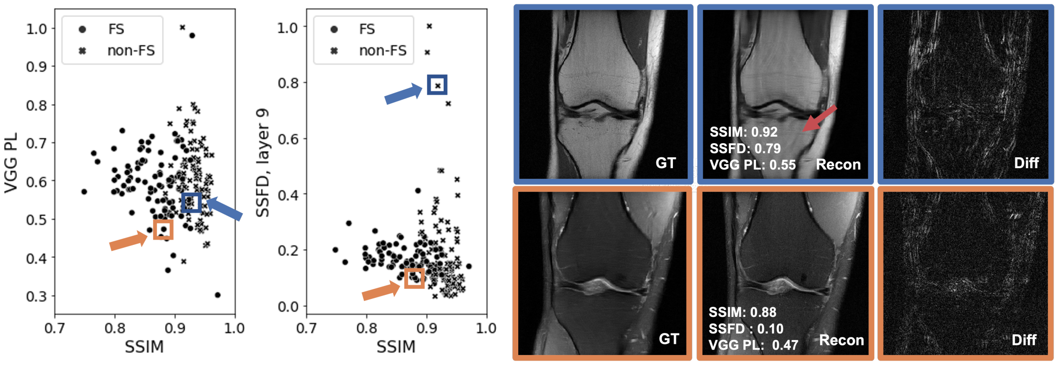

Figure 3: SSIM versus VGG-PL and SSFD plots for the test set (left), with two representative examples of MR reconstructions (right) showing the ground truth fully-sampled reconstruction (left), DL-based reconstruction with 4x acceleration (middle), and difference images (right). The top image (blue) has a qualitatively poor reconstruction that is also captured by SSFD, but SSIM and VGG-PL values that indicate a good reconstruction. Note the aliasing artifact (red arrow), as well as the large error on the top left of the difference image. The bottom image (orange) has a comparatively good qualitative reconstruction quality, captured by both SSFD and traditional metrics.

fine-tuned to correlate with radiologist-perceived image quality for various clinical tasks.

Two representative example MR reconstructions and their quantitative metrics are shown in Fig. 3. We find that SSFD can highlight examples of poor MR reconstruction quality, despite high SSIM, PSNR, and VGG-PL values. This is promising for using SSFD as an improved MR reconstruction quality control tool than traditional image quality metrics, and illustrates the potential benefit of image-domain-specific training.

4 Conclusion

This work introduces the SSFD image quality metric based on MR domain-specific feature representations learned from a self-supervised learning task. We show that SSFD provides flexibility to capture different complexities of feature representations, and highlights examples of poor image quality when traditional image quality metrics fail to do so.

References

[1] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.

[2] Dongwook Lee, Jaejun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing mri. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 15–18. IEEE, 2017.

[3] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[4] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.

[5] Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images. *IEEE transactions on medical imaging*, 39(4):1064–1072, 2019.

[6] 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*, 2020.

[7] 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.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[10] Ke Wang, Jonathan I Tamir, X Yu Stella, and Michael Lustig. High-fidelity reconstruction with instance-wise discriminative feature matching loss. In *Proc. Intl. Soc. Mag. Reson. Med*, 2020.

[11] Akshay S Chaudhari, Kathryn J Stevens, Jeff P Wood, Amit K Chakraborty, Eric K Gibbons, Zhongnan Fang, Arjun D Desai, Jin Hyung Lee, Garry E Gold, and Brian A Hargreaves. Utility of deep learning super-resolution in the context of osteoarthritis mri biomarkers. *Journal of Magnetic Resonance Imaging*, 51(3):768–779, 2020.

[12] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[13] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*, 2020.

[14] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzalv, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C Lawrence Zitnick, Michael P Recht, Daniel K Sodickson, and Yvonne W Lui. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiol Artif Intell*, 2(1):e190007, Jan 2020.

[15] 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.

[16] Mohammad Zalbagi Darestani, Akshay S Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2433–2444. PMLR, 18–24 Jul 2021.

5 Appendix

5.1 Self-Supervised Learning Framework

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 using a dropout rate of 0.1 and Group Normalization layers, and an Adam Optimizer with $|l_2|$ loss.

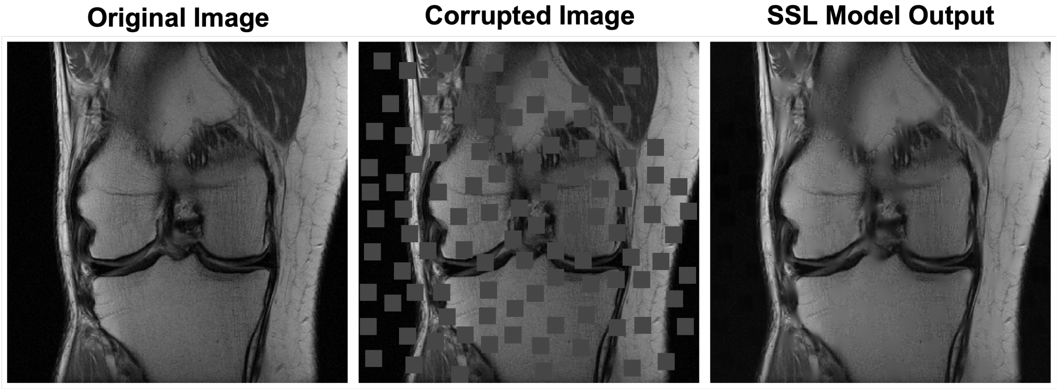

Figure 4: Self-supervised learning context prediction pre-task example framework. The original image belonging to the validation set (left), corrupted image with zero-filled image patches of size 16x16 pixels covering 25% of the image area via Poisson random density sampling (middle), and output of the trained SSL UNet model that corrected the corrupted image by in-painting (right).

5.2 VGG-16 Perceptual Loss

The VGG-16 Perceptual Loss was implemented using the Keras VGG-16 model pre-trained on ImageNet. The 3 fully-connected layers at the top of the network were excluded in order to preserve the 320x320 input resolution. Input MR images were converted to magnitude images, replicated along the 3rd axis to create pseudo-RGB inputs, scaled between 0 and 255, and then zero-centered with respect to the ImageNet dataset. The model was truncated after the ReLu in the 3rd convolution in the 3rd downsampling layer, as was done for the Content Perceptual Loss in the original implementation [8].

5.3 SSFD vs. PSNR

This section demonstrates the benefit of SSFD compared to PSNR, an additional traditional image quality metric. Fig. 5 and Fig. 6 are identical to Fig. 1 (SSFD under Image Perturbations) and Fig. 3 (SSFD as a Quality Control Tool), respectively, but compare SSFD to PSNR rather than SSIM.

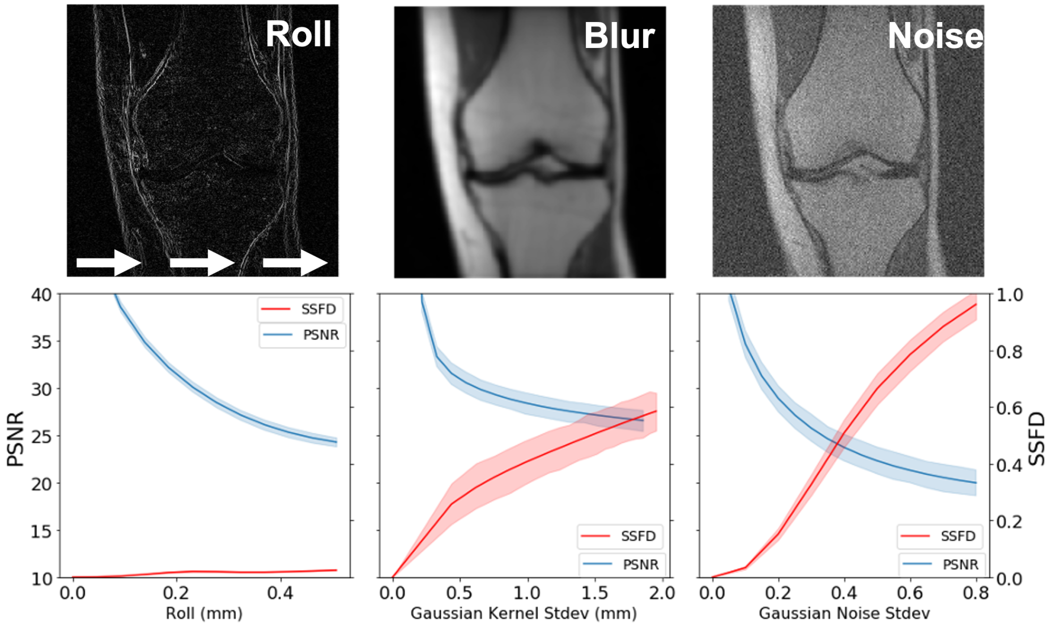

Figure 5: Example images at the 50^{th} percentile of perturbations, where the roll image is the difference between the perturbed and original scan (top). Average SSFD and PSNR with 95% confidence intervals from the 199 MR scan test set as a function of image perturbation (bottom). SSFD is less sensitive to pixel shifts compared to linearly increasing Gaussian standard deviation for blurring and noise for comparable decreases in PSNR.

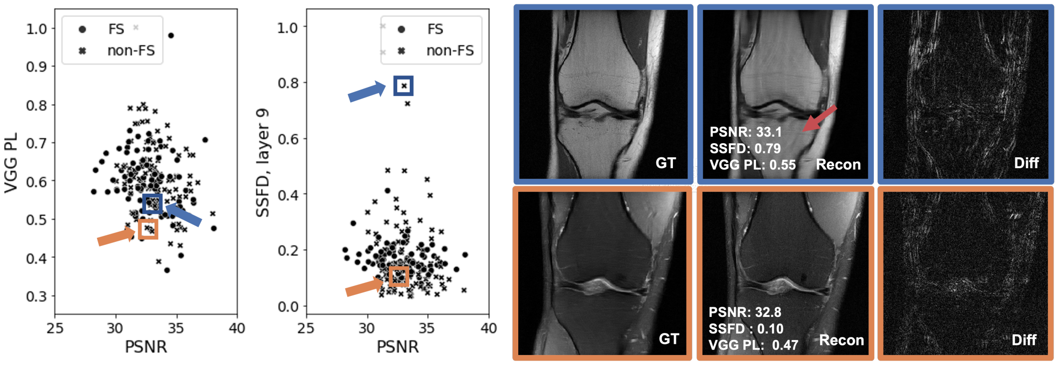

Figure 6: PSNR versus VGG-PL and SSFD plots for the test set (left), with two representative examples of MR reconstructions (right). The examples include the ground truth fully-sampled reconstruction (left), DL-based reconstruction with 4x acceleration (middle), and the resultant difference images (right). The top image (blue) has a qualitatively poor reconstruction that is also captured by SSFD, but PSNR and VGG-PL values that indicate a good reconstruction. Note the aliasing artifact (red arrow), as well as the large error in the top left of the difference image. The bottom image (orange) has a comparatively good qualitative reconstruction quality, captured by both SSFD and traditional metrics.

5.4 Fat-suppressed vs. non-fat-suppressed scans

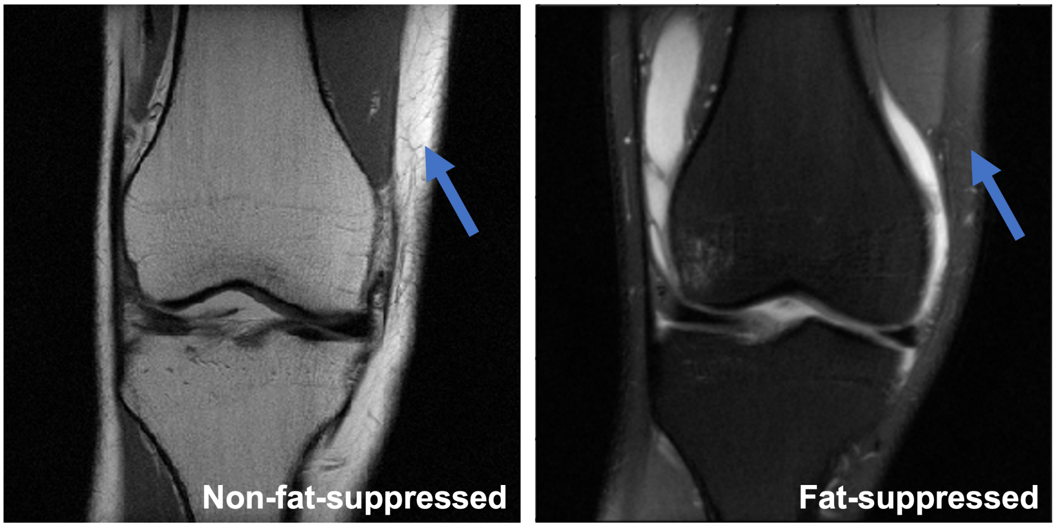

Figure 7: An example of a non-fat-suppressed (left) and a fat-suppressed (right) MR scan, each from a different patient in the fastMRI knee dataset. Note that the fatty regions, including bone and subcutaneous fat (blue arrow), appear bright on the non-fat-suppressed scan compared to the fat-suppressed scan, where fluid appears bright.