State-of-the-Art Machine Learning MRI Reconstruction in 2020: Results of the Second fastMRI Challenge

A Preprint

Matthew J. Muckley[*,1], Bruno Riemenschneider[*,2], Alireza Radmanesh[2], Sunwoo Kim[3], Geunu Jeong[3], Jingyu Ko[3], Yohan Jun[4], Hyungseob Shin[4], Dosik Hwang[4], Mahmoud Mostapha[5], Simon Arberet[5], Dominik Nickel[6], Zaccharie Ramzi[7,8], Philippe Ciuciu[7], Jean-Luc Starck[8], Jonas Teuwen[9], Dimitrios Karkalousos[10], Chaoping Zhang[10], Anuroop Sriram[11], Zhengnan Huang[2], Nafissa Yakubova[1], Yvonne W. Lui[2], Florian Knoll[2]

[1]Facebook AI Research, New York, NY, USA
[2]NYU School of Medicine, New York, NY, USA
[3]AIRS Medical, Seoul, South Korea
[4]Yonsei University, Seoul, Korea
[5]Siemens Healthineers, Princeton, NJ, USA
[6]Siemens Healthcare GmbH, Erlangen, Germany
[7]CEA (NeuroSpin) & Inria Saclay (Parietal), Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
[8]Département d'Astrophysique, CEA-Saclay, 91191 Gif-sur-Yvette, France
[9]Radboud University Medical Center, Nijmegen, Netherlands
[10]Amsterdam UMC, Amsterdam, Netherlands
[11]Facebook AI Research, Menlo Park, CA, USA
*Equal contribution.

December 29, 2020

Abstract

Accelerating MRI scans is one of the principal outstanding problems in the MRI research community. Towards this goal, we hosted the second fastMRI competition targeted towards reconstructing MR images with subsampled k-space data. We provided participants with data from 7,299 clinical brain scans (de-identified via a HIPAA-compliant procedure by NYU Langone Health), holding back the fully-sampled data from 894 of these scans for challenge evaluation purposes. In contrast to the 2019 challenge, we focused our radiologist evaluations on pathological assessment in brain images. We also debuted a new Transfer track that required participants to submit models evaluated on MRI scanners from outside the training set. We received 19 submissions from eight different groups. Results showed one team scoring best in both SSIM scores and qualitative radiologist evaluations. We also performed analysis on alternative metrics to mitigate the effects of background noise and collected feedback from the participants to inform future challenges. Lastly, we identify common failure modes across the submissions, highlighting areas of need for future research in the MRI reconstruction community.

Keywords Challenge · Public Dataset · MR Image Reconstruction · Machine Learning · Parallel Imaging · Compressed Sensing · Fast Imaging · Optimization

1 Introduction

Due to advances in algorithms, software platforms [1, 2, 3] and compute hardware, over the last five years there has been a surge of research of MR image reconstruction methods based on machine learning [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Traditionally, research in MR image reconstruction methods has been conducted on small data sets collected by individual research groups with direct access to MR scanner hardware and research agreements with the scanner vendors. Data set collection is difficult and expensive, with many research groups lacking the organizational infrastructure to collect data at the scale necessary for machine learning research. Furthermore, data sets collected by individual groups are often not shared publicly for a variety of reasons. As a result, research groups lacking large-scale data collection infrastructure face substantial barriers to reproducing results and making comparisons to existing methods in the literature.

Such challenges have been seen before. In the field of computer vision, the basic principles of convolutional neural networks (CNNs) were proposed as early as 1980 [15] and became well-established for character recognition by 1998 [16]. Following Nvidia's release of CUDA in 2007, independent research groups began to use GPUs to train larger and deeper networks [17, 18]. Nonetheless, universal acceptance of the utility of CNNs did not occur until the debut of the large-scale ImageNet data set and competition [19]. The introduction of ImageNet allowed direct cross-group comparison using this well-recognized data set of a size beyond what most groups could attain individually. In 2012 a CNN-based model [20] out-performed all non-CNN models, spurring a flurry of state-of-the-art results for image recognition [21, 22, 23, 24].

Since 2018, the fastMRI project has attempted to advance community-based scientific synergy in MRI by building on two pillars. The first consists of the release of a large data set of raw k-space and DICOM images [25, 26]. This data set is available to almost any researcher, allowing them to download it, replicate results, and make comparisons. The second pillar consists of hosting public leaderboards [25] and open competitions, such as the 2019 fastMRI Reconstruction Challenge on knee data [27]. The dimension of public competitions is not new to the MR community. Other groups have facilitated challenges around RF pulse design [28], diffusion tractography reconstruction [29], and ISMRM initiatives for reconstruction competitions and reproducibility [30].

The 2020 fastMRI Challenge continues this tradition of open competitions and follows the 2019 challenge with a few key differences. First, our target anatomy has been changed to focus on images of the brain rather than knee. Second, for 2020 we updated the radiologist evaluation process, asking radiologists to rate images based on *depiction of pathology* rather than *overall image quality*, emphasizing clinical relevance in competition results. Lastly, we address a core traditional problem in MR imaging: the capacity of models to generalize across sites and vendors. We introduce a new competition track: a "Transfer" track, where participants were asked to run their models on data from vendors not included in training. This contrasts with the 2019 challenge, which only included data from a single vendor for both training and evaluation.

2 Methods

This challenge focuses on MRI scan acceleration, a topic of interest to the MR imaging community for decades. MRI scanners acquire collections of Fourier frequency "lines", commonly referred to as k-space data. Due to hardware constraints on how magnetic fields can be manipulated, the rate at which these lines are acquired is fixed, which results in relatively long scan times and has negative implications with regard to image quality, patient discomfort, and accessibility. The major way to decrease scan acquisition time is to decrease the amount of data acquired. Sampling theory [31, 32, 33, 34] states that a minimum number of lines are required for image reconstruction. This minimum requirement can be circumvented by incorporating other techniques such as parallel imaging [35, 36, 37] and compressed sensing [38]. More recently, machine learning methods have demonstrated further accelerations over parallel imaging and compressed sensing methods.

The structure of the challenge is straightforward. We used fully-sampled k-space data from brain MRIs [25]. From this fully-sampled data, we applied retrospective downsampling and provided the downsampled data to challenge participants. Challenge participants ran their models on the downsampled data and submitted it to the competition website at `https://fastmri.org`, where we evaluated it using the fully-sampled data as gold standards.

At a high level we describe the principles of our 2020 challenge as follows. Using knowledge we gained through the 2019 challenge, we identified a few key alterations for 2020. These include:

- A new imaging anatomy, the brain, the most commonly-imaged organ using MRI.
- A focus on an evaluation of pathology depiction rather than overall image quality impressions to strengthen the connection between the challenge evaluation and clinical practice.

- An emphasis on generalization with the introduction of a new "Transfer" track where participants were asked to run their models on multi-vendor data.
- We removed the single-coil track and moved to a pure multi-coil challenge to increase the clinical relevance of the submitted models.
- Due to easier practical implementation and removal of the single-coil track, we used pseudo-equispaced subsampling masks (i.e., equispaced masks with a modification for achieving exact 4X/8X sampling rates) rather than random. We maintained the fully-sampled center due to its utility for autocalibrating parallel imaging methods [35, 37, 39] and compressed sensing [38].
- In the 2019 challenge our baseline model was a U-Net [40]; however, winning models [41, 42, 43] of the 2019 challenge were variational network/cascading models [27]. For the 2020 challenge, we provided a much stronger baseline model based on an End-to-End Variational Network [14].

We kept the following principles from the 2019 challenge:

- We again used a two-stage evaluation, where a quantitative metric was used to select the top 3 submissions. These finalists were then sent to radiologists to determine the winners. We used the structural similarity index (SSIM) [44] as our quantitative image quality index for ranking submissions prior to submission to clinical radiologists [27].
- We wanted to maintain realism for a straightforward, 2D imaging setting, and so all of the competition data was once again based on fully-sampled 2D raw k-space data.
- For the ground truth reference, we had discussions on alternatives to the root sum-of-squares (RSS) method used for quantitative evaluation in 2019. Although there was some consensus on the drawbacks of RSS [45, 46], there was no consensus on a single best alternative. In the following sections we discuss the impact of this choice further.

2.1 Challenge Tracks

In the 2019 challenge we included three submission tracks: multicoil with four times acceleration (Multi-Coil 4X), multicoil with eight times acceleration (Multi-Coil 8X), and single-coil with 4X acceleration (Single-Coil 4X). Among these tracks, the single-coil track garnered the most engagement, but due to its distance from clinical practice we decided to remove it from the 2020 challenge, replacing it with the Transfer track. For the standard multicoil tracks in the 2019 challenge, we observed that although there were many high-quality submissions at 4X, all of the submissions began missing pathology at 8X acceleration [27]. Since this time, 4X machine learning methods have been validated for clinical interchangeability [47]. This suggests that the current upper limit of 2D machine learning image reconstruction performance remains between 4-fold and 8-fold acceleration rates. In order to provide participants with both an obtainable target and a "reach" goal, we kept the 4-fold and 8-fold tracks for the 2020 challenge.

One frequent feedback on the 2019 challenge was on generalizability: despite the size of the data set, all of the data and results were from studies performed on MRI scanners from a single vendor at a single institution. To address this, we created the new Transfer track at 4-fold acceleration (Transfer 4X). For the Transfer track, participants were asked to run their models on data from vendors outside the main fastMRI data set. There was a caveat: we also restricted participants in the Transfer track to train their models only using available fastMRI data to ensure evaluation of transfer capability. At the time of the 2020 challenge announcement, we stated that these data would come "from another vendor" but did not specify further. At the challenge launch time, we revealed that the challenge data for this track was a mix of data from GE and Philips, providing additional difficulty for participants. As a result, submissions in the Transfer track exhibited wide deviations in performance depending on vendor.

2.2 Data Set

For the 2020 challenge we used brain MRI data. The neuroimaging subset of the fastMRI data has been described in an updated version of the arXiv paper [25]. It includes 6,970 scans (3,001 at 1.5 T, 3,969 at 3 T) collected at NYU Langone Health on Siemens scanners using T1, T1 post-contrast, T2, and FLAIR acquisitions. Unlike the knee challenge, this data set exhibits a wide variety of reconstruction matrix sizes. A summary of the data for the two main track splits is shown in Table 1. Of these 6,970 scans, 565 were withheld for evaluation in the challenge. In addition to standard HIPAA-compliant anonymization practices, all scans were cropped at the level of the orbital rim, preserving only the top part of the head.

For the challenge, the 565 scans were augmented further by 329 non-Siemens scans for the Transfer track. GE data were collected at NYU Langone Health and Philips data were collected on volunteers by clinical partner sites of Philips

Table 1: Summary of the data for the 2020 fastMRI challenge.

Split	T1	T1POST	T2	FLAIR	Total
Siemens/Main Tracks					
train	498	949	2,678	344	4,469
val	169	287	815	107	1,378
test (4X)	33	54	170	24	281
test (8X)	32	68	152	25	277
challenge (4X)	26	67	192	18	303
challenge (8X)	24	65	159	14	262
Transfer Track (4X, all challenge)					
GE	22	29	83	77	211
Philips	18	0	50	50	118

Healthcare of North America. Since the Philips data was collected on volunteers, this subsplit had no post-contrast imaging. One difficulty of the Transfer track was the fact that the GE data did not contain frequency oversampling. The lack of frequency oversampling was due to automatic removal during the analog-to-digital conversion process on the GE scanner.

In total the 2020 challenge had 6,405 scans available for training and validation (train, val, test) and there were 894 total scans for evaluation in the final challenge phase. This marked a substantial increase in scale from the 2019 challenge. For reference, the multicoil data from the 2019 challenge on knee data had 1,290 scans for training and validation (train, val, test) and 104 scans for the challenge, so the data for training increased by roughly 5-fold and the data for challenge evaluation increased by roughly 8-fold.

2.3 Evaluation Process

After submission, evaluation followed a two-stage process of comparisons to the fully-sampled "ground truth" images. For the ground truth images, we followed the previous convention [27] to use root sum-of-squares images. The advantage of this approach is that it does not bias to any one method for coil sensitivity estimation. A drawback is that RSS images can have substantial noise in the background. This noise is treated as ground truth in our quantitative evaluation, and any deviations from it influence our ranking. In planning for the challenge, we were unable to build consensus on an alternative ground truth calculation technique, but this topic could be re-examined in future challenges. For the quantitative evaluation metric, we chose to use SSIM [44].

For the qualitative assessment phase, a board-certified neuroradiologist selected six (two T1 post-contrast, two T2, and two FLAIR) cases from the challenge data set in each of the three tracks. Cases were specifically selected to represent a broad range of neuroimaging pathologies from intracranial tumors and strokes to normal and age-related changes. The selection process favored cases with more subtle pathologies for the 4X track and more obvious pathologies for the 8X track with the objective that this might yield better granularity for separating methods in the 4X track. Selected cases included both intraaxial and extraaxial tumors, strokes, microvascular ischemia, white matter lesions, edema, surgical cavities, as well as postsurgical changes and hardware including craniotomies and ventricular shunts. The Philips data set was constructed from images of volunteers, so this data set did not contain any clinically significant pathologies.

Radiologists were asked to base their overall ranking on the quality of the depiction of the pathology. For the volunteer cases in the Transfer track, small age-related imaging changes were used for ranking in place of pathology. In addition, we also asked radiologists to score each case in terms of artifacts, sharpness and contrast-to-noise ratio (CNR) using a Likert-type scale. On the Likert scale, 1 was the best (e.g., no artifacts) and 4 was the worst (e.g., unacceptable artifacts). A Likert score of 3 would affect diagnostic image quality.

2.4 Timeline

The 2020 challenge had the following timeline:

- December 19, 2019 - Release of the brain data set and update to the arXiv reference [25].
- July 9, 2020 - Announcement of the 2020 challenge.
- October 1-15, 2020 - Release of the challenge data set and submission window.

- October 16-19, 2020 - Calculation of SSIM scores. We selected the top 3 submissions for each track and forwarded them to a panel of radiologists for qualitative evaluation.
- October 19-November 1, 2020 - Radiologists evaluated submissions. They were asked to complete a score sheet for each of the 3 tracks which included ranking the submissions for each individual case in terms of overall quality of depiction of pathology.
- December 5, 2020 - Publication of the challenge leaderboard with results.
- December 12, 2020 - Official announcement of the winners of the three tracks with presentations at the Medical Imaging Meets NeurIPS Workshop.

2.5 Overview of Submission Methodologies

Here we share a brief description of the methodologies behind each of the submissions that made it to the finalist round for radiologist evaluation. The developers of these submissions are included as co-authors on this paper.

AIRS Medical The AIRS Medical model centered on two key factors: data standardization and multi-domain data processing. The data standardization was performed in the multi-channel coil dimension in order to accommodate diverse coil geometry and sensitivity information. The multi-channel input data were first preprocessed using GRAPPA [37] after discarding the oversampled area in image domain. The preprocessed data were reorganized into channel-combined images and residual images via ESPIRiT sensitivity maps [39], creating a consistent data format for different coil types and sensitivity information. The training input data were augmented using a different combination of sampling masks and were normalized by standard deviation of each imaging slice. For the multi-domain data processing, both image and k-space data were processed at the same convolutional layers of the baseline U-Net [40], sharing features between the two domains. The modified U-Net was cascaded for 4 stages with a data consistency path to build the entire model. The model training was optimized for SSIM loss with the Adam optimizer, taking approximately one week using 4 Nvidia V100 GPUs.

Neurospin The XPDNet [48] is a modular neural network unrolling the Chambolle-Pock algorithm [49] for 25 iterations. It is inspired by the primal only version of the Primal-Dual net [50]. In particular, a buffer of 5 previous unrolled iterates is used to make sure a complex nonlinear acceleration scheme is learned. The vanilla CNN is replaced with a more complex Multi-level Wavelet CNN [51], which has been benchmarked against other denoising sub-networks in the easier single-coil reconstruction task. Finally, the problem of estimating the sensitivity maps is handled with the solution suggested in [14], where a coarse estimate of the sensitivity maps is refined by a U-Net [40] with shared weights across all sensitivity maps. The whole architecture is trained end-to-end using a compound L1-Multi-scale SSIM loss as advised by [41]. The training lasted for one week for each acceleration factor using a single Nvidia V100 GPU. The code for training XPDNet is available online at `https://github.com/zaccharieramzi/fastmri-reproducible-benchmark`.

ATB ATB proposed Joint-ICNet, a Joint Deep Model-based MR Image and Coil Sensitivity Reconstruction Network. Joint-ICNet is based on a gradient descent algorithm based unrolled architecture [5, 14] and incorporated three different CNN-based regularizations in the model. Inspired by the approach in [8], optimization was performed using dual-domain CNN-based regularizations. One domain is a de-aliasing model in the image domain and the other is a k-space interpolation model in the k-space domain. The initial coil sensitivity maps were estimated with CNNs [14], and they were updated with a data consistency scheme combined with an update of the MR image. Joint-ICNet used the U-Net [40] architecture for CNN-based regularizations. The model was unrolled for a total of 10 iterations and the total number of trainable parameters was about 21 million. Training was performed using an Adam optimizer with β_1 = 0.9 and β_2 = 0.999 for 50 epochs with a learning rate of 0.0005 and a SSIM loss loss function. ATB separately trained the model with a different reduction factor (4X and 8X) and the training took approximately 10 days using 8 Nvidia TITAN RTX GPUs for each reduction factor.

MRRecon The proposed Momentum_DIHN k-space to image reconstruction network unrolls a proximal gradient algorithm with an extrapolation mechanism based on Nesterov momentum [52] to improve data-consistency convergence. Momentum_DIHN specializes in some of the cascades to lessen the first unrolled cascades' computational and memory requirements and provide a hard data-consistency mechanism at the network end to minimize any risk of hallucination. The deep learning regularization network is based on a novel hierarchical design of an iterative network that repeatedly decreases and increases the feature maps' resolution, allowing for a more memory-efficient model than conventional U-Nets [40]. Finally, ensemble modeling is applied to improve the overall performance of the proposed reconstruction network. The predictions of multiple diverse models trained using different architecture hyperparameters and various combinations of L1 and multi-scale SSIM losses are aggregated to produce the final prediction for the unseen challenge data. Input k-space was normalized by a factor equal to the 98th percentile of the zero-filled reconstruction. For the Transfer track, additional models were trained with random augmentation to the estimated normalization factor to improve overall system robustness.

ResoNNance The Recurrent Inference Machine (RIM) [53, 54, 43, 55] learns how to solve the inverse problem of accelerated-MRI reconstruction through the Maximum a Posteriori (MAP) estimation, using a forward model as has been described in [54]. Inputs to the model were initialized with the parallel-imaging compressed sensing (PICS) reconstruction [56] on the subsampled k-space, where the aliasing artifacts with equidistant spaced sampling patterns were reduced compared to a zero-filled reconstruction. Using a depth of 2, 128 hidden channels, and 8 time steps, ResoNNance created a RIM of 360,190 parameters. The U-Net model used for post-processing the sensitivity maps consisted of 480,900 parameters. Therefore, in total the RIM model had 841,090 parameters. ResoNNance trained separate models with regard to the acceleration factor (4x, 8x), the field strength (1.5T, 3T), and also the contrast (FLAIR, T1/T1PRE/T1POST, and T2) of the acquisition, where no cropping or padding was applied. Furthermore, ResoNNance used the Adam optimizer with initial learning rate 10^{-3}, and the SSIM as the loss function. Code for the RIM, data loaders, and documentation can be found through the DIRECT repository at `https://github.com/directgroup/direct`.

3 Results

3.1 Submission Overview

For the 2020 challenge we received a total of 19 submissions from eight different groups. Seven groups submitted to the Multi-Coil 4X and Multi-Coil 8X tracks. One of these groups chose not to submit to the Transfer track, while an eighth group submitted only to the Transfer track. As previously, we encourage all submitting groups to publish papers and code used to generate their results.

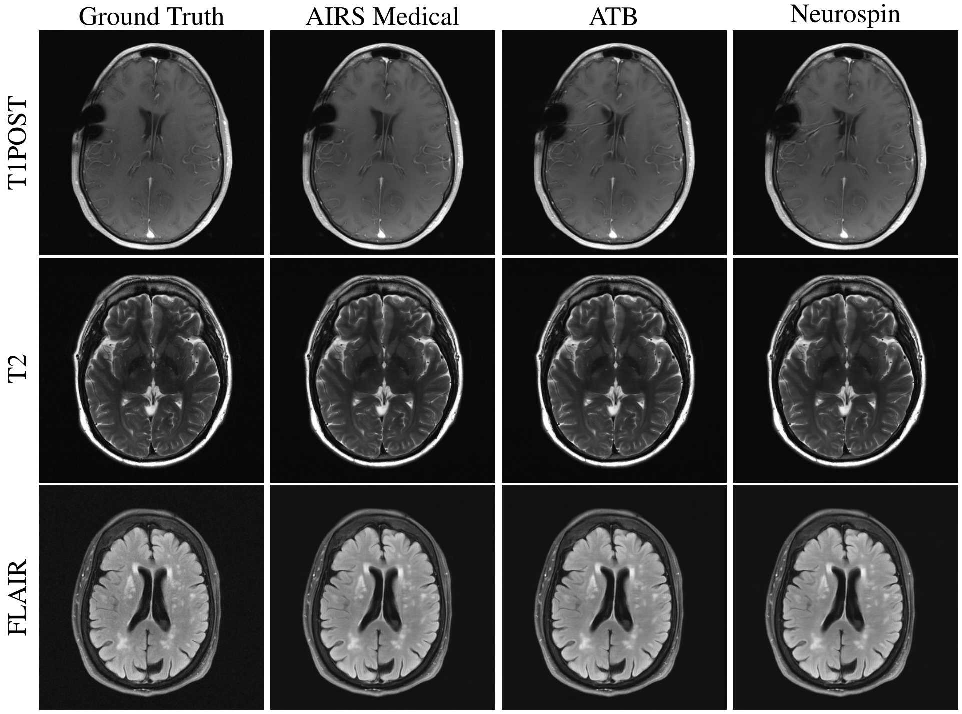

Figure 1: Examples of 4X submissions evaluated by radiologists. All methods reasonably reconstructed T2 and FLAIR images. The ATB and Neurospin methods struggled with a susceptibility region, exaggerating the focus of susceptibility and introducing a few false vessels between the susceptibility and the lateral ventricular wall. In other cases, radiologists observed mild smoothing of white matter regions on T1POST images.

Figure 1 shows an overview of images submitted to the 4X track of the challenge with Siemens data that were forwarded to radiologists. All three top performing submissions were able to successfully reconstruct the T2 and FLAIR images with minimal artifact presentation. For some images in this track's evaluation, radiologists had difficulty perceiving substantive differences between the three top performing reconstructions in terms of their overall ability to depict the pathology. Overall, the results were better on the high signal-to-noise T2 and FLAIR contrasts compared with those on the T1POST. In the case in Figure 1, the ATB and Neurospin methods struggled with a strong susceptibility effect, introducing false vessels between the susceptibility and the lateral ventricular wall.

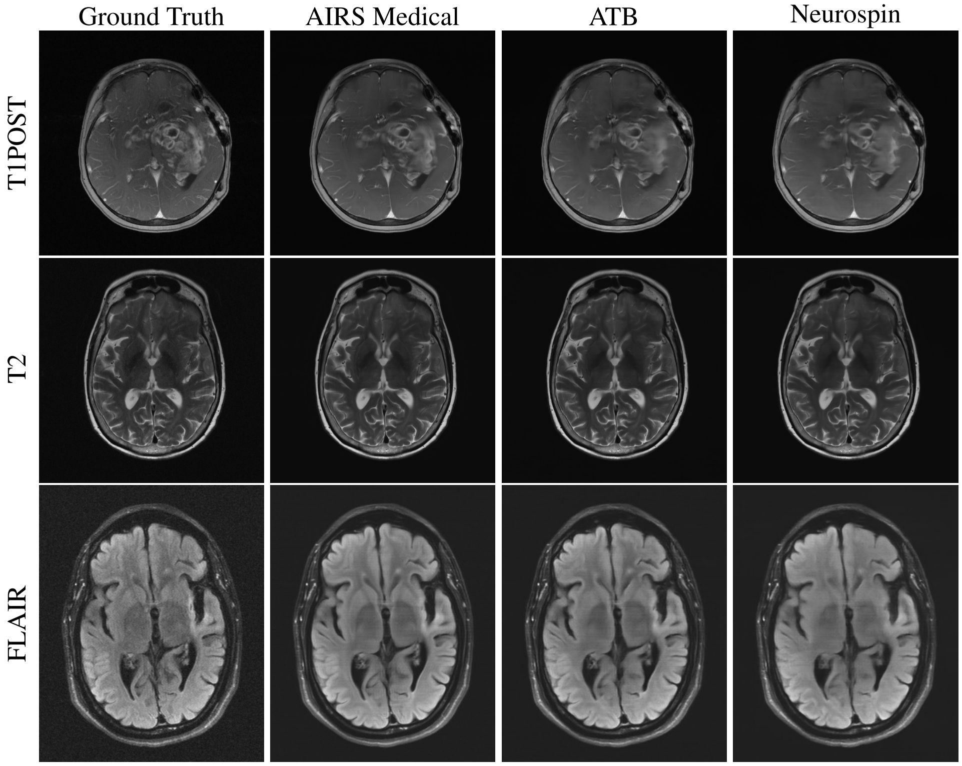

Figure 2: Examples of 8X submissions evaluated by radiologists. At this level of acceleration fine details are smoothed and obscured for all contrasts. On T1POST images, AIRS Medical was relatively more successful than ATB and Neurospin in showing fine details of the mass, particularly in its periphery. Noticeable on the FLAIR images are horizontal "banding" effects that arise from how neural networks interact with anisotropic sampling patterns.

Figure 2 shows example images for radiologist evaluation from the 8X track with Siemens data. In this track, artifacts are seen to be more severe and pronounced. For some cases radiologists stated that they were hesitant to accept any of the submissions at 8X. Over-smoothing is readily apparent in T1POST reconstructions from all three of the top performers. We noticed at this acceleration level that so-called horizontal "banding" effects [57] could be appreciated in the FLAIR images due to the extreme acceleration and the anisotropic sampling pattern.

Example images from the 4X Transfer track are shown in Figure 3. For this track, we observed the lowest SSIM values (Section 3.2). Of note, there is a divergence between performance of methods on GE versus Philips data. This can be seen the image submitted by ResoNNance in Figure 3, which introduces artifacts in its reconstructions of the GE images (T1POST and T2 in Figure 3), but less so in its Philips reconstruction (FLAIR in Figure 3). Most participant models (trained on Siemens data) were able to reconstruct Philips data with higher fidelity than GE, likely due to the fact that Philips and Siemens followed the same protocol for writing frequency-oversampled data to their raw data files. An additional factor is that GE uses a T1-based FLAIR, whereas Philips and Siemens use a T2-based FLAIR.

3.2 Quantitative Results

Figure 4 shows an overview of SSIM scores across group rankings. SSIM values were highly clustered in the 4X track, with all top 4 participants scoring between 0.955 and 0.965. We observed greater variation between submissions in the 8X track, with the top participant scoring 0.952 and the others scoring below 0.944. The greatest variation occurred in the Transfer track. Many participants struggled to adapt their models to GE data. These data did not include frequency oversampling in the raw k-space data, which we have observed can decrease SSIMs for models by as much as 0.1-0.4 if no other adjustments are made. On the other hand, the Philips data did include frequency oversampling, so adaptation here was more straightforward.

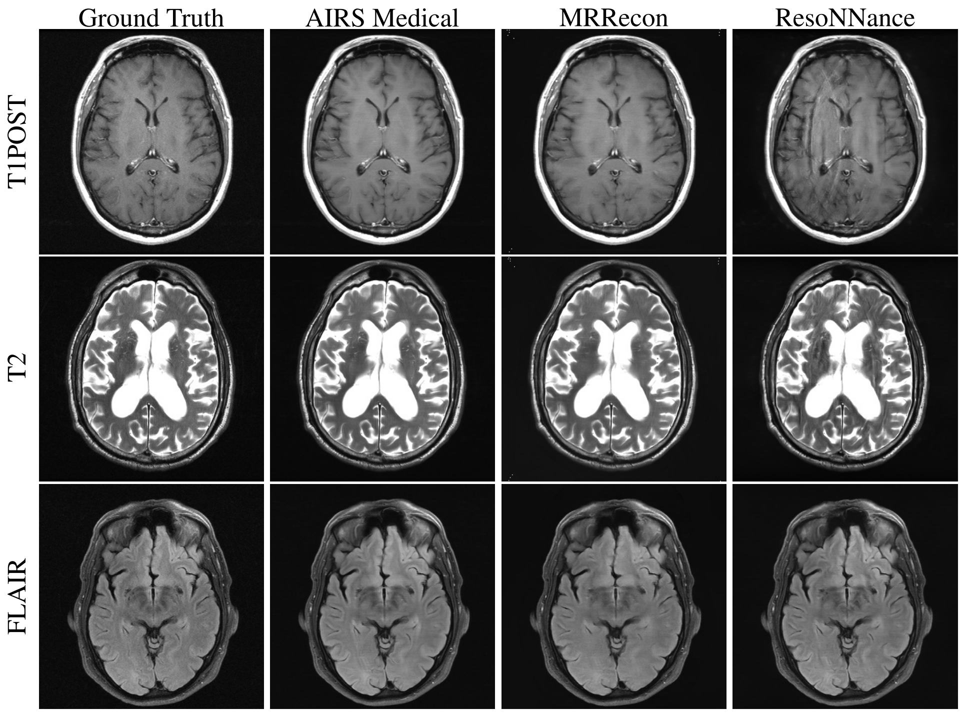

Figure 3: Examples of 4X Transfer submissions evaluated by radiologists. The T1POST and T2 examples are from GE scanners, whereas the FLAIR example is from a Philips scanner. All methods introduced blurring to the images. Several methods had trouble adapting to the GE data while performing relatively well on the Philips data, as seen in the form of aliasing artifacts in one of the T1POST images.

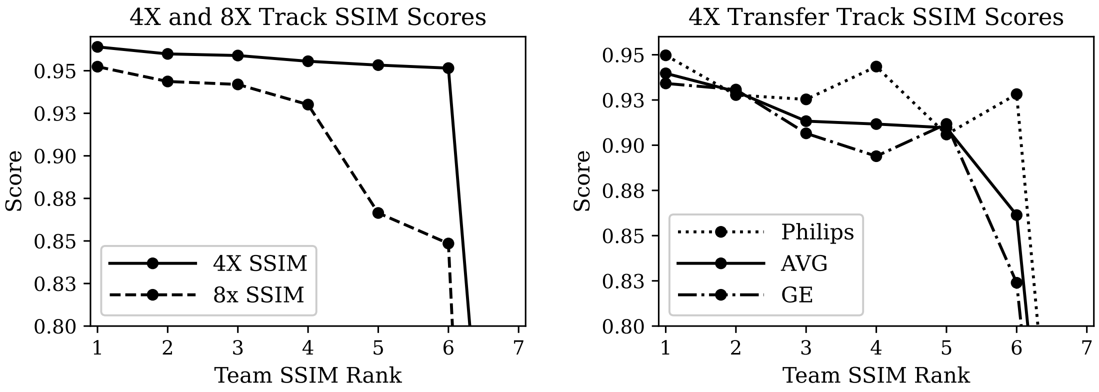

Figure 4: Summary of SSIM values across contestants. (*left*) Model perfomance for teams submitting to the main 4X and 8X Siemens competition tracks. (*right*) Model performance for teams submitting to the Transfer track (combination of GE and Philips data). The “AVG” model score for the Transfer track was a simple average across all volumes in the Transfer track.

Table 2 summarizes results by contrast for the finalists in each competition track. The strongest SSIM scores were usually recorded on T1 post-contrast images (T1POST), while the weakest scores were typically on FLAIR images. The same participant recorded the top average SSIM score for every contrast in every track except the Transfer track for T1 contrast. In this case, two other participants posted higher SSIM scores.

One team, HungryGrads, submitted to all tracks and received a very low SSIM score between 0.4 and 0.5. This team set the background air to nearly 0s, which led to a clinically irrelevant SSIM loss of approximately 0.3 for their submissions. The HungryGrads submission prompted our team to perform a post-hoc analysis where we masked both the submission and the reference RSS ground truth before calculating SSIM, with results plotted in Figure 5. Applying this mask

Table 2: Summary of SSIM scores by contrast (mean ± standard deviation).

Team	AVG	T1	T1POST	T2	FLAIR
4X Track					
AIRS Medical	**0.964 ± 0.020**	**0.967 ± 0.018**	**0.969 ± 0.013**	**0.965 ± 0.017**	**0.930 ± 0.036**
ATB	0.960 ± 0.021	0.964 ± 0.019	0.965 ± 0.015	0.961 ± 0.018	0.924 ± 0.038
Neurospin	0.959 ± 0.022	0.963 ± 0.020	0.965 ± 0.015	0.960 ± 0.018	0.920 ± 0.040
8X Track					
AIRS Medical	**0.952 ± 0.032**	**0.953 ± 0.020**	**0.969 ± 0.012**	**0.951 ± 0.036**	**0.918 ± 0.038**
ATB	0.944 ± 0.034	0.943 ± 0.022	0.954 ± 0.014	0.943 ± 0.038	0.905 ± 0.043
Neurospin	0.942 ± 0.035	0.940 ± 0.024	0.953 ± 0.015	0.942 ± 0.038	0.898 ± 0.044
4X Transfer Track					
AIRS Medical	**0.940 ± 0.053**	0.902 ± 0.071	**0.960 ± 0.010**	**0.975 ± 0.014**	**0.910 ± 0.052**
MRRecon	0.930 ± 0.048	**0.946 ± 0.020**	0.956 ± 0.010	0.950 ± 0.023	0.897 ± 0.059
ResoNNance	0.913 ± 0.045	0.936 ± 0.052	0.939 ± 0.030	0.957 ± 0.041	0.854 ± 0.071

markedly improved the SSIM scores of HungryGrads, although it would not have made this team a finalist. Applying the mask would have enabled ATB to enter the finalist round for the Transfer track, primarily due to the fact that their method introduced artifacts to GE reconstructions in the background. Our custom mask would not have changed finalist rankings otherwise.

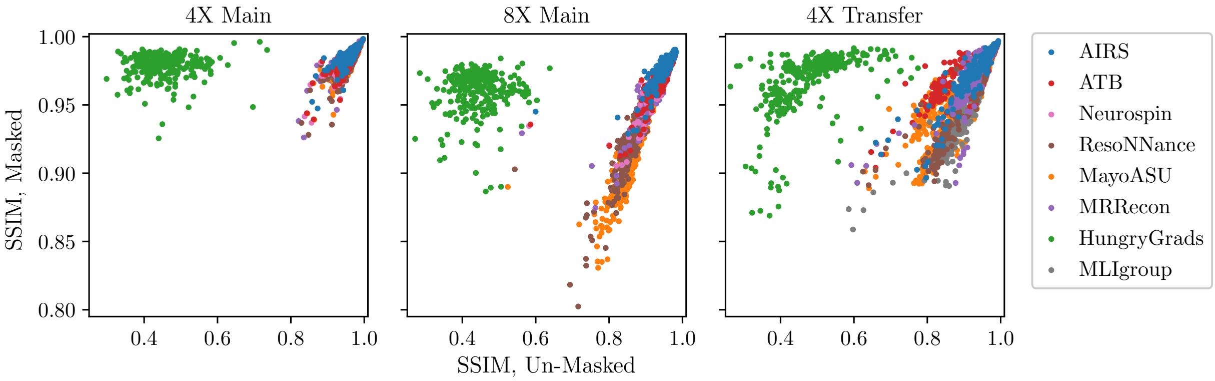

Figure 5: Overview of the impact of a masking procedure. Shown are SSIM scores incorporating masking vs. SSIM scores with no masking. Both methods used the RSS ground truth. 6 outlier points with very low SSIM on both axes were cut off for presentation in the "4X Transfer" plot.

3.3 Radiologist Evaluation Results

Radiologist rankings based on quality of pathology depiction were concordant with SSIM scores for the top submissions as shown in Figure 6. The second and third place performers for both 4X and 8X tracks were flipped between the quantitative ranking based on SSIM and the qualitative ranking based on radiologists. The SSIM difference between these two constructions methods was relatively small, out to the third decimal place. In the Transfer track, radiologist rankings matched ranking based on SSIM.

A summary of the ranks and Likert scores is shown in Table 3. Across all metrics AIRS Medical separated itself from the other submissions with the highest SSIM and best image quality. Aside from this single team, differentiation among the other teams was not strong. Of note, both the Neurospin and ATB teams had nearly identical average SSIM scores for the quantitative evaluation, with ATB presenting a slightly higher score (0.960 vs. 0.959 in 4X, 0.944 vs. 0.942 in 8X). In the radiologist evaluation phase, these ranks flipped, with Neurospin receiving slightly higher ranks (1.94 vs. 2.22 in 4X, 2.25 vs. 2.28 in 8X).

Table 3: Summary of quality ranks and Likert scores (mean ± standard deviation, lower is better).

Team	Rank	Artifacts	Sharpness	CNR
4X Track				
AIRS Medical	**1.36 ± 0.64**	**1.53 ± 0.70**	**1.53 ± 0.51**	**1.53 ± 0.51**
Neurospin	1.94 ± 0.86	1.81 ± 1.01	1.72 ± 0.66	1.75 ± 0.84
ATB	2.22 ± 0.87	1.75 ± 0.97	1.97 ± 0.65	1.86 ± 0.80
8X Track				
AIRS Medical	**1.28 ± 0.64**	**1.67 ± 0.68**	**1.89 ± 0.75**	**1.94 ± 0.75**
Neurospin	2.25 ± 0.77	1.86 ± 0.83	2.72 ± 0.81	2.28 ± 0.81
ATB	2.28 ± 0.70	1.92 ± 0.94	2.56 ± 0.77	2.42 ± 0.84
4X Transfer Track				
AIRS Medical	**1.11 ± 0.32**	**1.42 ± 0.50**	**1.83 ± 0.65**	**1.81 ± 0.62**
MRRecon	1.97 ± 0.56	1.61 ± 0.55	2.41 ± 0.69	2.22 ± 0.64
ResoNNance	2.78 ± 0.54	3.08 ± 0.84	2.86 ± 0.76	3.06 ± 0.86

A case-wise breakdown of the ranks for all 3 finalists and all rated cases is shown in Figure 6. For second and third-place metrics as rated by SSIM, radiologist assessment was discordant between the two methods. However, in 16 out of 18 cases the highest SSIM score within the finalists' batches also received the highest radiologists' rating. A similar relation - not shown here - was found for the other used metrics such as normalized mean-squared error (NMSE) and peak signal-to-noise ratio (PSNR).

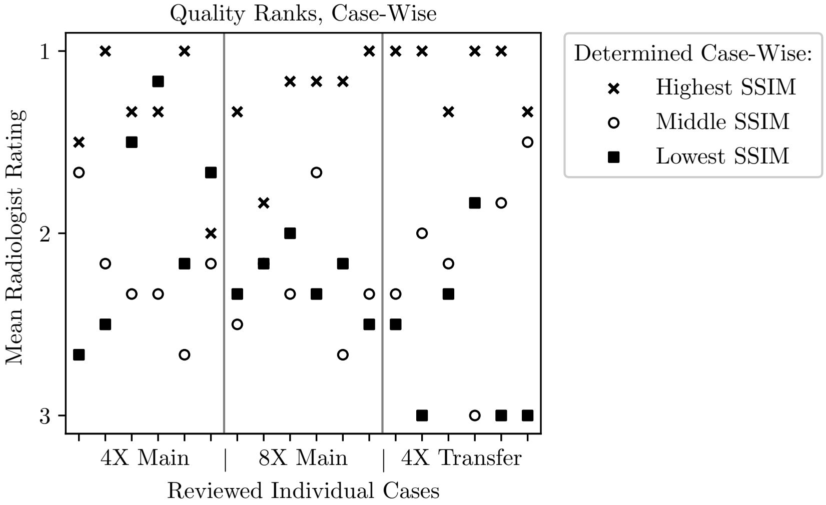

Figure 6: Scatter plot of mean radiologist rank across cases. The horizontal axis has a separate tick for each case evaluated by the radiologist cohort. The scatter plot markers indicate whether that method was from the team with the highest, middle, or lowest SSIM scores. We generally observed radiologists awarding the best ranks to models with the best SSIM score.

Radiologists did take note of hallucinatory effects introduced by the submission models. Figure 7 shows hallucination examples from all three tracks. In some cases methods created artifact-mimics. In other examples, models morphed an abnormality into a more normal brain structure, such as a sulcus or vessel. Finally, we observed at least one example combining these two where an artifact was created at some intermediate layer of a model and then processed by the remaining portions of the network into a normal structure mimic.

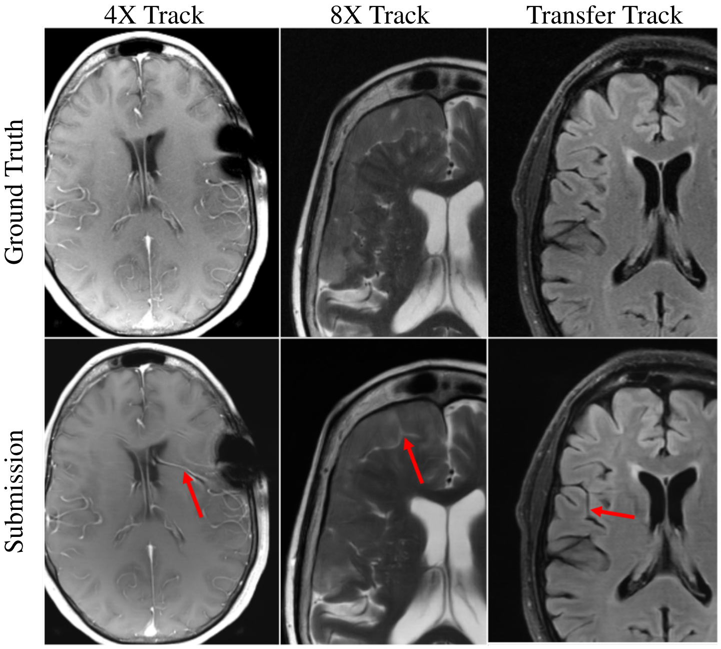

Figure 7: Examples of reconstruction hallucinations among challenge submissions. (*left*) A 4X submission from Neurospin generated a false vessel, possibly related to susceptibilities introduced by surgical staples. (*center*) An 8X submission from ATB introduced a linear bright signal mimicking a cleft of cerebrospinal fluid, as well as blurring of the boundaries of the extra-axial mass. (*right*) A submission from ResoNNance introduced a false sulcus or prominent vessel.

4 Discussion

4.1 Submission Overview

In the 2019 challenge all three tracks were very closely contested, with little separation between teams either in the quantitative or the radiologist evaluation phases. We observed this pattern to be reversed in the 2020 challenge, with one team assertively scoring the best in all evaluation phases. For some images in the 4X track, multiple radiologists said that they did not observe major differentiating aspects affecting the depiction of pathology in the submissions. However, when averaging the radiologists' rankings, radiologists preferred the method that had the highest-scoring on SSIM from AIRS Medical. We further observed that the AIRS model scored highest on Likert-type ratings of artifacts, sharpness, and CNR. This model also provided improvement over the baseline [14], which had previously been demonstrated for clinical interchangeability at 4X for knee imaging [47]. Outside of the AIRS model, in the 4X and 8X tracks the second and third-place models scored very close together in both the quantitative and the qualitative evaluation phase. In some cases the SSIM scores for these two models were identical out to three decimal places.

We observed decreases in performance in the Transfer track. Many participants struggled to adapt their models to the GE data with its lack of disk-written frequency oversampling. Although technically the GE scanner did not operate in any majorly different way than Philips and Siemens scanners (all use frequency oversampling), this simple aspect rendered many models useless in this track without modification. Another factor was a divergence in FLAIR methodology: our Philips and Siemens data used T2 FLAIR images, whereas the GE data had T1 FLAIR images. Modifications for correcting these effects seem not to be straightforward.

In terms of radiologist evaluations, despite the drawbacks of SSIM and RSS ground truths, we observed a correlation between radiologist scores and SSIM scores for large SSIM separations. Multiple radiologists found images at 4X to be similar in terms of depiction of the pathology, although artifacts tended to be more problematic in T1POST images. When it came to the 8X and Transfer tracks, radiologist sentiment became more negative. Multiple radiologists in both of these tracks offered feedback that none of the submitted images would be clinically acceptable, indicating that these two tasks may remain open problems going into 2021.

4.2 Quantitative Evaluation Process

Discussions around the quantitative evaluation process primarily concerned the presence of background noise during both the planning and execution stage of the challenge. The influence of background noise on SSIM scores is substantial. One participant in the 2019 challenge had a dedicated style transfer model in order to add this noise back into the reconstructed images [58]. Despite the drawbacks to SSIM, we were unable to agree on an alternative for the 2020 challenge.

In the 2020 challenge, the HungryGrads team submitted images with backgrounds of nearly zeroes, which penalized their scores. Prompted by this submission we investigated the effect a masked metric might have had on their scores in Section 3.2. We opted for a masking algorithm that removes most background pixels and altered the algorithm parameters for low-SNR edge cases where it did not perform well. Due to the relatively small size of the challenge data set visual inspection of the validity of the masks was feasible. The ranking of the challenge did not change dramatically due to masking, but masking made metrics less prone to a specific reconstruction method's impact on the background.

Another intriguing alternative would be to use adaptive combine reconstructions as a ground truth that implicitly suppress background noise [46]. We considered using adaptive combine reconstructions in our evaluation, but concluded that the results were not particularly meaningful as most models had been trained with RSS backgrounds. Future use of adaptive combine for both training and validation remains feasible.

4.3 Qualitative Radiologist Evaluation

For the 2020 challenge, we altered the radiologist questionnaire to focus their ranking on the depiction of pathologies rather than general image quality. Some radiologists found the focus helpful, commenting specifically that the images at 8X and in the 4X Transfer track might not be acceptable for clinical use. As this task aligns more closely with the normal clinical workflow, we would encourage future competition organizers to use this approach for their radiologist evaluation procedures.

In the 4X track, there were specific cases where the radiologist rankings were concordant and others where the rankings were discordant. Discordant cases tended, upon review, to show that the main abnormalities were similarly well depicted across the top 3 reconstructions, though there were oftentimes concordant estimations of differences between reconstructions in terms of artifacts, sharpness and CNR.

Radiologist sentiment was affected by hallucinations such as those in Figure 7. Such hallucinatory features are not acceptable and especially problematic if they mimic normal structures that are either not present or actually abnormal. Neural network models can be unstable as demonstrated via adversarial perturbation studies [59]. Despite the lack of realism in some of these perturbations, our results indicate that hallucination and artifacts remain a real concern, particularly at higher accelerations. Initial work based on bootstrap or jackknife methods has been done to predict the possible magnitude of neural network errors [60], but this topic remains in major need for further development.

4.4 Feedback from Participants

Participants were generally enthusiastic about being able to participate in the challenge. We received positive feedback on our communication via the fastMRI GitHub repository at `https://github.com/facebookresearch/fastMRI` and the forum associated with the web site at `https://fastmri.org`. We also received positive feedback around the challenge's realism in focusing on multi-coil data, as well as the challenge's generalizability initiative in focusing on the Transfer track.

Still, participants felt the realism could be improved in other areas. In particular, the sampling mask used for the challenge used pseudo-regular sampling in order to achieve exact 4X and 8X sampling rates. This sampling pattern is not equivalent to the perfectly equidistant sampling pattern used on MRI systems, which give acceleration rates slightly less than the target rate due to the densely-sampled center. As a result, challenge models are likely to require further fine-tuning training before application to clinical data.

Another point of feedback centered on the storage and compute resources necessary to participate in the challenge. In the 2019 challenge, the storage aspect was mitigated by the inclusion of the single-coil track (which had a smaller download size). The single-coil track attracted a lot of engagement, with 25 out of 33 groups submitting to it [27]. From the compute angle, the trend towards larger models requires costly hardware. Training the baseline End-to-End Variational Network [14] requires 32 GPUs, each with 32 GB of memory, for about 3.5 days. This level of compute power is not available at many academic centers. By comparison, one participant submitted a model trained on only a single GPU. For the future, researchers felt it would be helpful for the barriers to entry were lower, particularly for academic groups that might have innovative methods but less compute or storage.

As always, the selection of best quantitative evaluation metrics to use is extremely difficult and there are potential drawbacks to many or all. Participants did provide feedback concerning the use of SSIM and the use of RSS for ground truth images. Although groups acknowledged efforts to seek superior metrics, they felt that settling for this particular metric was disappointing. Some participants felt there was a tradeoff between optimizing for SSIM (which promotes smoothing) vs. radiologist interpretation. We did not allow secondary submissions from participants that might enhance the images for human perception, such as those based on noise dithering or inspired by stochastic resonance [61, 62, 63, 47]. Allowing secondary submissions for radiologist interpretation may be beneficial for future challenges, provided ground truth images are also included to allow radiologists to watch for hallucination. In compiling the results for this challenge we have attempted to investigate some other options that would at least mitigate the effects of background noise and feel that this is an important topic for further investigation.

5 Conclusion

The 2020 fastMRI reconstruction challenge featured two core modifications from its 2019 predecessor: 1) a new competition Transfer track to evaluate model generalization and 2) adjusting the radiologist evaluation to focus on pathology depiction. In addition to these, we extended our competition to a new anatomy with much larger data sets for both training and competition evaluation. The competition resulted in a new state-of-the-art model. Our challenge confirmed areas in need of research, particularly those along the lines of evaluation metrics, error characterization, and AI-generated hallucinations. Radiologist sentiment was mixed for images submitted to the 8X and the Transfer tracks; these may remain open research frontiers going into 2021. We hope that researchers and future challenge organizers find the results of the 2020 fastMRI challenge helpful in their future endeavors.

6 Acknowledgments

We would like to thank all research groups that participated in the challenge. We thank the scientific advisors that gave feedback during challenge organization. We thank the judges for the contribution of their clinical expertise and time. We would also like to thank Tullie Murrell, Mark Tygert, and C. Lawrence Zitnick for many valuable contributions to the launch and execution of this challenge. We give special thanks to Philips North America and their clinical partner sites for the data they contributed for use as part of the Transfer track. We acknowledge grant support from the NIH under awards R01EB024532, P41EB017183, and R21EB027241.

References

[1] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cuDNN: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, pages 265–283, 2016.

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[4] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep ADMM-Net for compressive sensing MRI. In *Advances in Neural Information Processing Systems*, pages 10–18, 2016.

[5] 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.

[6] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2017.

[7] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1310–1321, 2017.

[8] Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. KIKI-net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic Resonance in Medicine*, 80(5):2188–2201, 2018.

[9] 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, 2018.

[10] 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.

[11] Florian Knoll, Kerstin Hammernik, Erich Kobler, Thomas Pock, Michael P Recht, and Daniel K Sodickson. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic Resonance in Medicine*, 81(1):116–128, 2019.

[12] Florian Knoll, Kerstin Hammernik, Chi Zhang, Steen Moeller, Thomas Pock, Daniel K. Sodickson, and Mehmet Akçakaya. Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues. *IEEE Signal Processing Magazine*, 37(1):128–140, 2020.

[13] Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020.

[14] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–73, 2020.

[15] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*, pages 267–285, 1982.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Rajat Raina, Anand Madhavan, and Andrew Y Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the International Conference on Machine Learning*, pages 873–880, 2009.

[18] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, 22(12):3207–3220, 2010.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[25] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

[26] 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 Drozdzal, 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. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

[27] Florian Knoll, Tullie Murrell, Anuroop Sriram, Nafissa Yakubova, Jure Zbontar, Michael Rabbat, Aaron Defazio, Matthew J. Muckley, Daniel K. Sodickson, C. Lawrence Zitnick, and Michael P. Recht. 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.

[28] William A Grissom, Kawin Setsompop, Samuel A Hurley, Jeffrey Tsao, Julia V Velikina, and Alexey A Samsonov. Advancing RF pulse design using an open-competition format: Report from the 2015 ISMRM challenge. *Magnetic Resonance in Medicine*, 78(4):1352–1361, 2017.

[29] Kurt G Schilling, Alessandro Daducci, Klaus Maier-Hein, Cyril Poupon, Jean-Christophe Houde, Vishwesh Nath, Adam W Anderson, Bennett A Landman, and Maxime Descoteaux. Challenges in diffusion MRI tractography–lessons learned from international benchmark competitions. *Magnetic Resonance Imaging*, 57:194–209, 2019.

[30] Oliver Maier, Steven Hubert Baete, Alexander Fyrdahl, Kerstin Hammernik, Seb Harrevelt, Lars Kasper, Agah Karakuzu, Michael Loecher, Franz Patzig, Ye Tian, Ke Wang, Daniel Gallichan, Martin Uecker, and Florian Knoll. CG-SENSE revisited: Results from the first ISMRM reproducibility challenge. *Magnetic Resonance in Medicine*, 2020.

[31] Edmund T Whittaker. XVIII.—On the functions which are represented by the expansions of the interpolation-theory. *Proceedings of the Royal Society of Edinburgh*, 35:181–194, 1915.

[32] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

[33] Vladimir A Kotelnikov. On the carrying capacity of the 'ether' and wire in telecommunications. In *Material for the First All-Union Conference on Questions of Communication (Russian), Izd. Red. Upr. Svyzai RKKA, Moscow*, 1933.

[34] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[35] Daniel K Sodickson and Warren J Manning. Simultaneous acquisition of spatial harmonics (SMASH): Fast imaging with radiofrequency coil arrays. *Magnetic Resonance in Medicine*, 38(4):591–603, 1997.

[36] 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.

[37] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.

[38] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008.

[39] 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.

[40] 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*, pages 234–241, 2015.

[41] Nicola Pezzotti, Sahar Yousefi, Mohamed S. Elmahdy, Jeroen van Gemert, Christophe Schülke, Mariya Doneva, Tim Nielsen, Sergey Kastryulin, Boudewijn P. F. Lelieveldt, Matthias J. P. van Osch, Elwin de Weerdt, and Marius Staring. An adaptive intelligence algorithm for undersampled knee MRI reconstruction: Application to the 2019 fastMRI challenge. *arXiv preprint arXiv:2004.07339*, 2020.

[42] Puyang Wang, Eric Z Chen, Terrence Chen, Vishal M Patel, and Shanhui Sun. Pyramid convolutional RNN for MRI reconstruction. *arXiv preprint arXiv:1912.00543*, 2019.

[43] Patrick Putzky, Dimitrios Karkalousos, Jonas Teuwen, Nikita Miriakov, Bart Bakker, Matthan Caan, and Max Welling. i-RIM applied to the fastMRI challenge. *arXiv preprint arXiv:1910.08952*, 2019.

[44] 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.

[45] 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.

[46] David O Walsh, Arthur F Gmitro, and Michael W Marcellin. Adaptive reconstruction of phased array MR imagery. *Magnetic Resonance in Medicine*, 43(5):682–690, 2000.

[47] Michael P Recht, Jure Zbontar, Daniel K Sodickson, Florian Knoll, Nafissa Yakubova, Anuroop Sriram, Tullie Murrell, Aaron Defazio, Michael Rabbat, Leon Rybak, Mitchell Kline, Gina Ciavarra, Erin F. Alaia, Mohammed Samim, William R. Walter, Dana J. Lin, Yvonne W. Lui, Matthew Muckley, Zhengnan Huang, Patricia Johnson, Ruben Stern, and C. Lawrence Zitnick. Using deep learning to accelerate knee MRI at 3T: Results of an interchangeability study. *American Journal of Roentgenology*, 215(6):1421–1429, 2020.

[48] Zaccharie Ramzi, Philippe Ciuciu, and Jean-Luc Starck. XPDNet for MRI reconstruction: An application to the fastMRI 2020 brain challenge. *arXiv preprint arXiv:2010.07290*, 2020.

[49] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[50] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018.

[51] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level Wavelet-CNN for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 773–782, 2018.

[52] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate O(1/k^2). In *Dokl. AN USSR*, volume 269, pages 543–547, 1983.

[53] Patrick Putzky and Max Welling. Recurrent inference machines for solving inverse problems. *arXiv preprint arXiv:1706.04008*, 2017.

[54] Kai Lønning, Patrick Putzky, Jan-Jakob Sonke, Liesbeth Reneman, Matthan WA Caan, and Max Welling. Recurrent inference machines for reconstructing heterogeneous MRI data. *Medical Image Analysis*, 53:64–78, 2019.

[55] Youssef Beauferris, Jonas Teuwen, Dimitrios Karkalousos, Nikita Moriakov, Mattha Caan, Lívia Rodrigues, Alexandre Lopes, Hélio Pedrini, Letícia Rittner, Maik Dannecker, Viktor Studenyak, Fabian Gröger, Devendra Vyas, Shahrooz Faghih-Roohi, Amrit Kumar Jethi, Jaya Chandra Raju, Mohanasankar Sivaprakasam, Wallace Loos, Richard Frayne, and Roberto Souza. Multi-channel MR reconstruction (MC-MRRec) challenge – comparing accelerated MR reconstruction models and assessing their genereralizability to datasets collected with different coils. *arXiv preprint arXiv:2011.07952*, 2020.

[56] Martin Uecker, Frank Ong, Jonathan I Tamir, Dara Bahri, Patrick Virtue, Joseph Y Cheng, Tao Zhang, and Michael Lustig. Berkeley advanced reconstruction toolbox. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2015.

[57] Aaron Defazio, Tullie Murrell, and Michael P Recht. MRI banding removal via adversarial training. In *Advances in Neural Information Processing Systems*, 2020. To appear.

[58] Kerstin Hammernik, Jo Schlemper, Chen Qin, Jinming Duan, Ronald M Summers, and Daniel Rueckert. Σ-net: Systematic evaluation of iterative deep neural networks for fast parallel MR image reconstruction. *arXiv preprint arXiv:1912.09278*, 2019.

[59] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

[60] Mark Tygert, Rachel Ward, and Jure Zbontar. Compressed sensing with a jackknife and a bootstrap. *arXiv preprint arXiv:1809.06959*, 2018.

[61] Enrico Simonotto, Massimo Riani, Charles Seife, Mark Roberts, Jennifer Twitty, and Frank Moss. Visual perception of stochastic resonance. *Physical Review Letters*, 78(6):1186, 1997.

[62] Frank Moss, Lawrence M Ward, and Walter G Sannita. Stochastic resonance and sensory information processing: a tutorial and review of application. *Clinical Neurophysiology*, 115(2):267–281, 2004.

[63] Munendra Singh, Ashish Verma, and Neeraj Sharma. Optimized multistable stochastic resonance for the enhancement of pituitary microadenoma in MRI. *IEEE Journal of Biomedical and Health Informatics*, 22(3):862–873, 2017.