# Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction

Matthew J. Muckley, *Member, IEEE*, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, *Member, IEEE*, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, *Student Member, IEEE*, Philippe Ciuciu, *Senior Member, IEEE*, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll, *Member, IEEE*

***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.**

***Index Terms*— Challenge, public data set, MR image reconstruction, machine learning, parallel imaging, compressed sensing, fast imaging, optimization.**

Manuscript received March 5, 2021; revised April 21, 2021; accepted April 22, 2021. Date of publication April 30, 2021; date of current version August 31, 2021. This work was supported by the NIH under Award R01EB024532, Award P41EB017183, and Award R21EB027241. The NIH funding was awarded to NYU. *(Matthew J. Muckley and Bruno Riemenschneider contributed equally to this work.) (Corresponding author: Matthew J. Muckley.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the NYU Grossman School of Medicine Internal Review Board.

Matthew J. Muckley and Nafissa Yakubova are with Facebook AI Research, New York, NY 10003 USA (e-mail: mmuckley@fb.com; nafissay@fb.com).

Bruno Riemenschneider, Alireza Radmanesh, Zhengnan Huang, Yvonne W. Lui, and Florian Knoll are with the School of Medicine, New York University (NYU), New York, NY 10016 USA (e-mail: bruno.riemenschneider@nyulangone.org; alireza.radmanesh@nyulangone.org; zhengnan.huang@nyulangone.org; yvonne.lui@nyulangone.org; florian.knoll@nyumc.org).

Sunwoo Kim, Geunu Jeong, and Jingyu Ko are with AIRS Medical, Seoul 132-798, South Korea (e-mail: kim.sunwoo@airsmed.com; jeong.geunu@airsmed.com; ko.jingyu@airsmed.com).

Yohan Jun, Hyungseob Shin, and Dosik Hwang are with School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: yohan.jun16@gmail.com; whatzupsup@yonsei.ac.kr; dosik.hwang@yonsei.ac.kr).

Mahmoud Mostapha and Simon Arberet are with Siemens Healthineers, Princeton, NJ 08540 USA (e-mail: mahmoud.mostapha@siemens-healthineers.com; simon.arberet@siemens-healthineers.com).

Dominik Nickel is with Siemens Healthcare GmbH, 91052 Erlangen, Germany (e-mail: marcel.nickel@siemens-healthineers.com).

Zaccharie Ramzi is with CEA (NeuroSpin) and Inria Saclay (Parietal), Université Paris-Saclay, 91191 Gif-sur-Yvette, France, and also with the Département d'Astrophysique, CEA-Saclay, 91191 Gif-sur-Yvette, France (e-mail: zaccharie.ramzi@inria.fr).

Philippe Ciuciu is with CEA (NeuroSpin) and Inria Saclay (Parietal), Université Paris-Saclay, 91191 Gif-sur-Yvette, France (e-mail: philippe.ciuciu@gmail.com).

Jean-Luc Starck is with the Département d'Astrophysique, CEA-Saclay, 91191 Gif-sur-Yvette, France (e-mail: jlstarck@gmail.com).

Jonas Teuwen is with the Medical Center, Radboud University, 6525 Nijmegen, The Netherlands (e-mail: j.teuwen@nki.nl).

Dimitrios Karkalousos and Chaoping Zhang are with the Department of Biomedical Engineering and Physics, Amsterdam UMC, 1105 Amsterdam, The Netherlands (e-mail: d.karkalousos@amsterdamumc.nl; c.zhang1@amsterdamumc.nl).

Anuroop Sriram is with Facebook AI Research, Menlo Park, CA 94025 USA (e-mail: anuroops@fb.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3075856, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3075856

## I. Introduction

DUE to advances in algorithms, software platforms [1]–[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]–[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]–[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 [30], [31].

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.

## II. 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 [32]–[35] 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 [36]–[38] and compressed sensing [39]. More recently, machine learning methods have demonstrated further accelerations over parallel imaging and compressed sensing methods.

To promote the advancement of methods for accelerated MRI, we organized a public challenge. We applied retrospective downsampling to fully-sampled MRIs 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 quantitatively evaluated it using the fully-sampled data as gold standards. We selected six cases for each of the top three teams in each track of the challenge (three tracks total) and presented the cases to a group of six radiologists for qualitative evaluation. The challenge winner was selected based on the best depiction of pathology compared to the ground truth as judged by radiologists.

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. This follows more closely sampling patterns (and relaxation effects) that are used for parallel imaging in vendor sequences, facilitating easier clinical deployment. We maintained the fully-sampled center due to its utility for autocalibrating parallel imaging methods [36], [38], [40] and compressed sensing [39].
- In the 2019 challenge our baseline model was a U-Net [41]; however, winning models [42]–[44] 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) [45] 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 [46], [47], there was no consensus on a single best alternative. In the following sections we discuss the impact of this choice further.

### A. 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 [48]. 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.

### B. 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], with further information included in the supplemental material for this paper. 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 I. 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.

TABLE I
SUMMARY OF THE CHALLENGE DATA

| 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 |

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 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.

Participants were restricted to only use this data set for training the weights of their models. We did permit participants to use their own data as validation data, but not for backpropagating gradients.

### C. Evaluation Process

Submissions were processed via https://fastmri.org. This site maintains a submission system for both the challenge and the public test leaderboard. Currently, there are no plans to make the challenge split of the data available in order to maintain the integrity of the results, but research groups may submit to the public leaderboard using the test set via the website at any time.

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. There are some drawbacks to RSS, including 1) discarding of the phase information and 2) RSS images can have substantial noise in the background.

The phase is not typically used for anatomical evaluation. The issue with noise is more fundamental as it is treated as ground truth in our quantitative evaluation, and any deviations from it influence our ranking. This is counterbalanced by using radiologist evaluation for declaring the challenge winner. 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 [45], with a script showing the script used for evaluation in the fastMRI repository at https://github.com/facebookresearch/fastMRI. SSIM has several parameters. We investigated adjusting these parameters prior to challenge launch, but found that they generally did not alter the ranking of methods evaluated in our quality control phase, and so as a result we used the default parameters in scikit-image [49].

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. Therefore small age-related imaging changes were used for ranking in place of pathology.

Six radiologists with 9-16 years of experience (two of whom are radiology division chiefs) were asked to evaluate the 18 selected image volumes for each team, basing their overall ranking on the quality of the depiction of the pathology using the ground truth as a reference. Radiologists came from a wide set of institutions, including the Mayo Clinic, Baylor College of Medicine, NYU Langone Health, the University of Pittsburgh Medical Center, Stanford University, and the University of California, Los Angeles. None of these institutions had finalist submissions. All radiologists looked at all images in the selected cases during the qualitative evaluation phase, and results were averaged. Radiologists were aware of the overarching goals of the challenge but were blinded as to which teams submitted the images. 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.

### D. 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.

### E. 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.

A summary of finalist model properties is shown in Table II. (Team names: “AIRS” is AIRS Medical, “ATB” is ATB, “MRR” is MRRecon, “Nspin” is Neurospin, “Res” is ResoNNance.) The number of model parameters ranged from 841,000 in the case of ResoNNance to 200 million in the case of AIRS. Teams applied GRAPPA [38], ESPIRiT [40], or simple zero-filled initializations. For coil estimation, teams used either ESPIRiT [40] or a simple center-based estimation with U-Net refinement similar to that in the End-to-End Variational Network [14]. Teams used 1-8 GPUs for training, and training time was between 7 and 21 days.

*AIRS Medical:* The AIRS Medical model used a combination of image- and k-space-domain processing in a fashion analogous (but distinct) from that used in KIKI-Net [8]. The model included a data consistency cascade with 4 U-Net stages. At each convolutional layer of the U-Net [41], the multi-domain processing split the channels into one group that operated in image space and one group that operated in k-space [8]. Data consistency was enforced at each layer. The network was initialized with a GRAPPA estimate [28] (pre-processed into reconstruction + residual), and coil sensitivities were estimated using ESPIRiT [40]. Since the sampling pattern was pseudo-equispaced, multiple GRAPPA kernels were used to calculate the GRAPPA images. AIRS optimized their model using Adam [50] over 20 epochs using a batch size of 4 at a learning rate of $10^{-3}$ (decayed to $10^{-4}$ after 15 epochs) with SSIM as the loss function. Optimization took approximately 7 days on four NVIDIA V100 GPUs. Code is not publicly available.

*ATB:* The ATB model, called “Joint-ICNet” [51], was a 10-iteration unrolled algorithm with CNNs replacing regularization terms in a fashion similar to other recent methods [5], [8], [14]. Joint-ICNet used the U-Net [41] at each convolutional layer with the dual-domain processing previously introduced in KIKI-net [8]. Joint-ICNet used a zero-filled reconstruction as the initial estimate and and coil sensitivities

TABLE II
FINALIST MODEL PROPERTIES

| Team | # Params | Init. | Coil Meth. | GPUs | Tr. Time |
|---|---|---|---|---|---|
| AIRS | 200 M | GRAPPA | ESPIRiT | 4 (V100) | 7 days |
| ATB | 21 M | Zero-Filled | U-Net | 8 (TITAN) | 10 days |
| MRR | 16 M | Zero-Filled | ESPIRiT | 1 (V100) | 14 days |
| Nspin | 155 M | Zero-Filled | U-Net | 1 (V100) | 7 days |
| Res | 841 k | ESPIRiT | U-Net | 4 (RTX 8000) | 21 days |

were calculated by refining a rough central k-space estimate with a U-Net [14], [41]. ATB optimized Joint-ICNet using Adam [50] over 50 epochs using a batch size of one at a learning rate of $10^{-4}$ with SSIM as the loss function. Training took approximately 10 days using 8 NVIDIA TITAN GPUs. Code is not publicly available.

*MRRecon:* The MRRecon model, called "Momentum_DIHN," unrolled the Nesterov momentum algorithm with CNN-based regularization for 12 cascades, 6 pre-cascades, and 0 or 1 post-cascade. The CNN module is a "Deep, Iterative, Hierarchical Network" (DIHN) that extends the Down-Up network [52] with a hierarchical block design, facilitating memory efficiency over a standard U-Net [41]. Momentum_DIHN used a zero-filled image as the initial estimate and ESPIRiT for calculating coil sensitivities. To improve transfer track performance, models with several hyperparameters were ensembled to generate the final images. MRRecon optimized Momentum_DIHN using Adam [50] for less than 5 epochs using a batch size of one at a learning rate of $10^{-4}$ with a compound L1/MS-SSIM loss function [42]. Training took approximately 14 days on an NVIDIA V100 GPU. Code is not publicly available.

*Neurospin:* The Neurospin model, called XPDNet [53], [54], is a modular neural network unrolling the Chambolle-Pock algorithm [55] for 25 iterations. The model was inspired by the primal-only version of the Primal-Dual net [56], replacing the vanilla CNN with a multi-level wavelet CNN [57]. XPDNet used a zero-filled image as the initial estimate and calculated coil sensitivities using a rough central k-space estimate refined by a U-Net [14], [41]. Neurospin optimized XPDNet using Rectified Adam [58] over 100 epochs using a batch size of one at a learning rate of $10^{-4}$ with a compound L1/MS-SSIM loss function [42] (98% MS-SSIM weight). Training took approximately 7 days on an NVIDIA V100 GPU. Code is available at https://github.com/zaccharieramzi/fastmri-reproducible-benchmark.

*ResoNNance:* ResoNNance used a Recurrent Inference Machine (RIM) that has been previously described [31], [44], [59], [60]. Coil sensitivities were calculating using the center of k-spaced followed by U-Net refinement [41]. RIM used ESPIRiT as the model input calculated from the BART toolbox [61]. ResoNNance optimized RIM using Adam [50] over 90 epochs using a batch size of one at a learning rate of $10^{-3}$ with an SSIM loss function. Separate models were trained for every field strength (1.5 T, 3 T) and contrast (FLAIR, T1/T1PRE/T1POST, and T2). Code for the RIM, data loaders, and documentation can be found through the DIRECT repository at https://github.com/directgroup/direct.

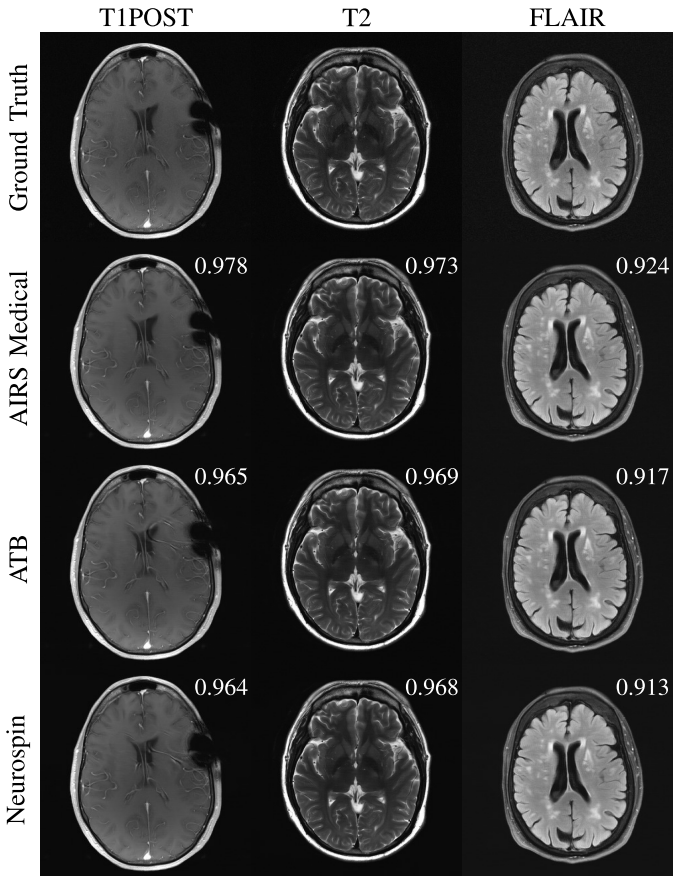


Fig. 1. Examples of 4X submissions evaluated by radiologists with slice-level SSIM scores. All methods reasonably reconstructed T2 and FLAIR images. The ATB and Neurospin methods struggled with a susceptibility region, exaggerating the focus of susceptibility and introducing a few false vessels between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

## III. RESULTS

### A. 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 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.

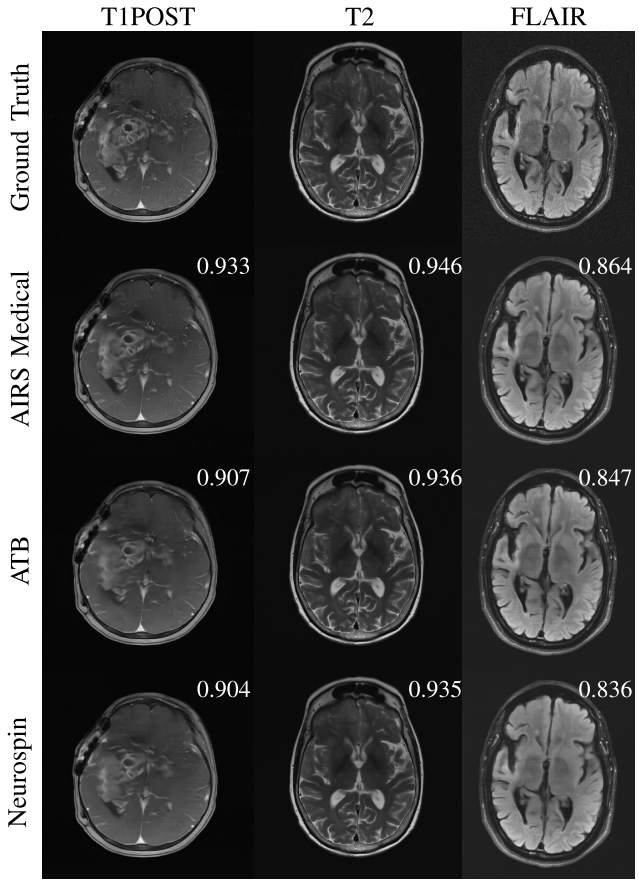


Fig. 2. Examples of 8X submissions evaluated by radiologists with slice-level SSIM scores. At this level of acceleration fine details are smoothed and obscured for all contrasts. On T1POST images, AIRS Medical was relatively more successful than ATB and Neurospin in showing fine details of the mass,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

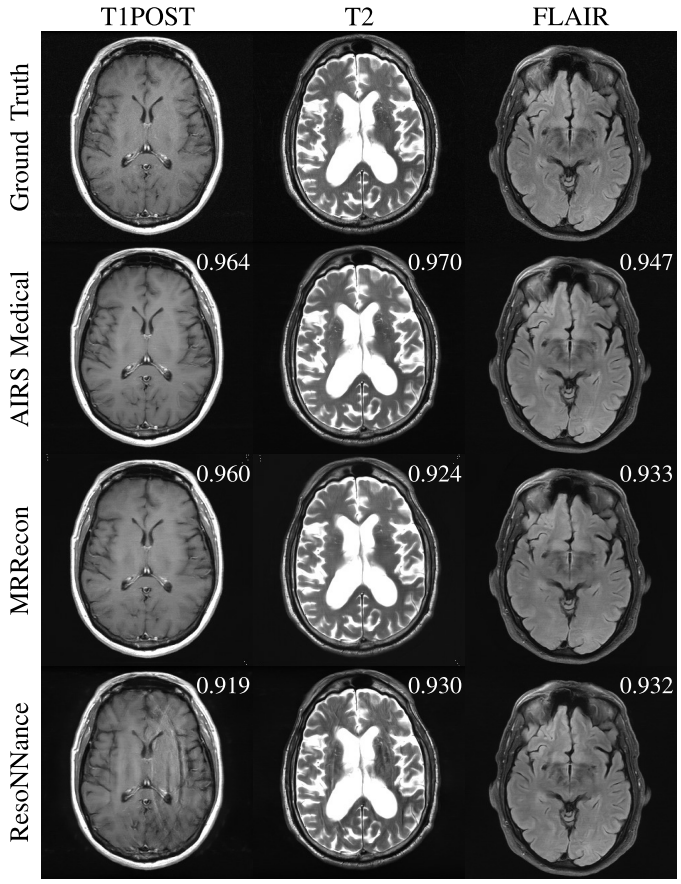


Fig. 3. Examples of 4X Transfer submissions evaluated by radiologists with slice-level SSIM scores. The T1POST and T2 examples are from GE scanners, whereas the FLAIR example is from a Philips scanner. All methods introduced blurring to the images. Several methods had trouble adapting to the GE data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

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 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 [62] 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 III-B). 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.

### *B. 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.

TABLE III
SUMMARY OF AVERAGE SSIM SCORES WITH 95% CONFIDENCE INTERVALS

| Team | AVG | T1 | T1POST | T2 | FLAIR |
|---|---|---|---|---|---|
| **4X Track** | | | | | |
| AIRS Medical | **0.964 (0.914, 0.987)** | **0.967 (0.926, 0.985)** | **0.969 (0.935, 0.986)** | **0.965 (0.933, 0.988)** | **0.930 (0.862, 0.976)** |
| ATB | 0.960 (0.908, 0.985) | 0.964 (0.922, 0.983) | 0.965 (0.928, 0.984) | 0.961 (0.926, 0.986) | 0.924 (0.850, 0.971) |
| Neurospin | 0.959 (0.904, 0.984) | 0.963 (0.920, 0.982) | 0.965 (0.926, 0.984) | 0.960 (0.924, 0.986) | 0.920 (0.844, 0.970) |
| **8X Track** | | | | | |
| AIRS Medical | **0.952 (0.894, 0.979)** | **0.953 (0.906, 0.979)** | **0.963 (0.939, 0.981)** | **0.951 (0.899, 0.978)** | **0.918 (0.857, 0.973)** |
| ATB | 0.944 (0.881, 0.973) | 0.943 (0.892, 0.973) | 0.954 (0.926, 0.977) | 0.943 (0.889, 0.973) | 0.905 (0.835, 0.965) |
| Neurospin | 0.942 (0.874, 0.973) | 0.940 (0.888, 0.972) | 0.953 (0.924, 0.975) | 0.942 (0.887, 0.972) | 0.898 (0.824, 0.957) |
| **4X Transfer Track** | | | | | |
| AIRS Medical | **0.940 (0.802, 0.989)** | 0.902 (0.795, 0.983) | **0.960 (0.942, 0.969)** | **0.975 (0.934, 0.992)** | **0.910 (0.728, 0.979)** |
| MRRecon | 0.930 (0.846, 0.981) | **0.946 (0.910, 0.975)** | 0.956 (0.935, 0.969) | 0.950 (0.889, 0.984) | 0.897 (0.671, 0.956) |
| ResoNNance | 0.913 (0.791, 0.977) | 0.936 (0.892, 0.967) | 0.939 (0.920, 0.956) | 0.957 (0.911, 0.980) | 0.855 (0.695, 0.951) |

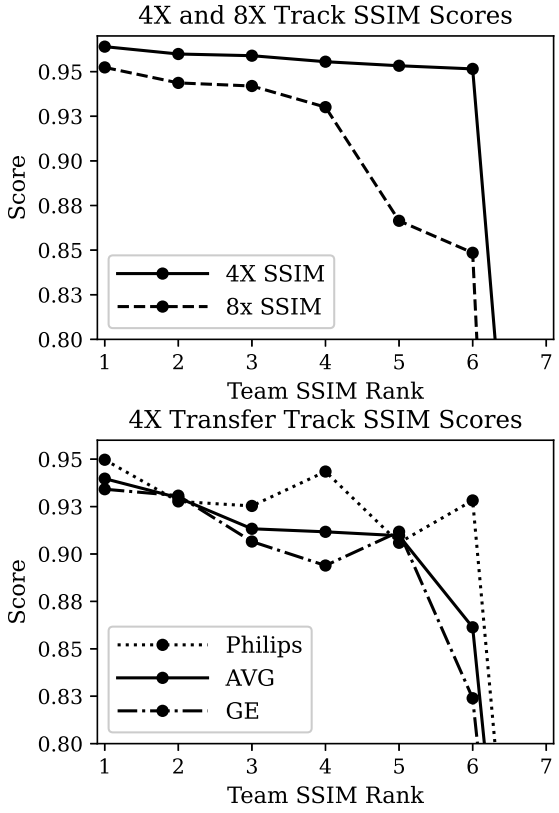


Fig. 4. Summary of SSIM values across contestants. (*top*) Model perfomance for teams submitting to the main 4X and 8X Siemens competition tracks. (*bottom*) Model performance for teams submitting to the Transfer track (combination of GE and Philips data). The "AVG" model score for the Transfer track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

Table III summarizes results by contrast for the finalists in each competition track with means and 95% confidence intervals based on 2.5% and 97.5% quantiles. 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 S1 in the supplementary material. Applying this mask 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. Our custom mask would not have changed finalist rankings otherwise.

### C. Radiologist Evaluation Results

Radiologist rankings based on quality of pathology depiction were concordant with SSIM scores for the top submissions as shown in Figure 5. 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 reconstruction methods was relatively small, out to the third decimal place. In the Transfer track, radiologist rankings matched rankings based on SSIM.

A summary of the ranks and Likert scores with means and standard deviations is shown in Table IV (AIRS = AIRS Medical, Nspin = Neurospin, MRR = MRRecon, and Res. = ResoNNance). We applied standard deviations for Table IV (instead of quantiles as used in Table III) to show more information on the variability. 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 IV
SUMMARY OF QUALITY RANKS AND LIKERT SCORES

| Team | Rank | Artifacts | Sharpness | CNR |
|---|---|---|---|---|
| **4X Track** | | | | |
| AIRS | **1.36 ± 0.64** | **1.53 ± 0.70** | **1.53 ± 0.51** | **1.53 ± 0.51** |
| Nspin | 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 | **1.28 ± 0.64** | **1.67 ± 0.68** | **1.89 ± 0.75** | **1.94 ± 0.75** |
| Nspin | 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 | **1.11 ± 0.32** | **1.42 ± 0.50** | **1.83 ± 0.65** | **1.81 ± 0.62** |
| MRR | 1.97 ± 0.56 | 1.61 ± 0.55 | 2.41 ± 0.69 | 2.22 ± 0.64 |
| Res. | 2.78 ± 0.54 | 3.08 ± 0.84 | 2.86 ± 0.76 | 3.06 ± 0.86 |

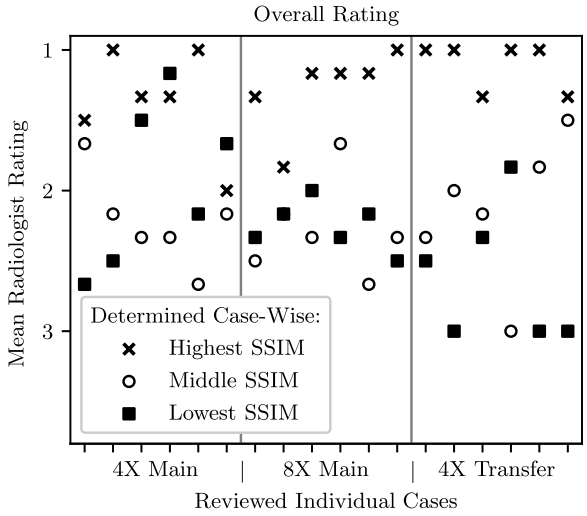


Fig. 5. Scatter plot of mean radiologist rank across cases. The horizontal axis has a separate tick for each case evaluated by the radiologist cohort. The scatter plot markers indicate whether that method was from the team with the highest, middle, or lowest SSIM scores. We generally observed radiologists awarding the best ranks to models with the best SSIM score.

A case-wise breakdown of the ranks for all 3 finalists and all rated cases is shown in Figure 5. 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).

Radiologist agreement according to Kendall's coefficient of concordance generally improved as SSIM scores diverged. We calculated the concordance using radiologist rankings of teams for quality of depiction of pathology vs. the ground truth. For each case in the radiologist evaluation phase, we evaluated Kendall's coefficient of concordance with tie correction, and then aggregated over all cases by averaging. This resulted in values of 0.457 for the 4X track, 0.386 for the 8X track, and 0.781 for the 4X Transfer track (where 0 indicates complete disagreement and 1 indicates complete agreement).

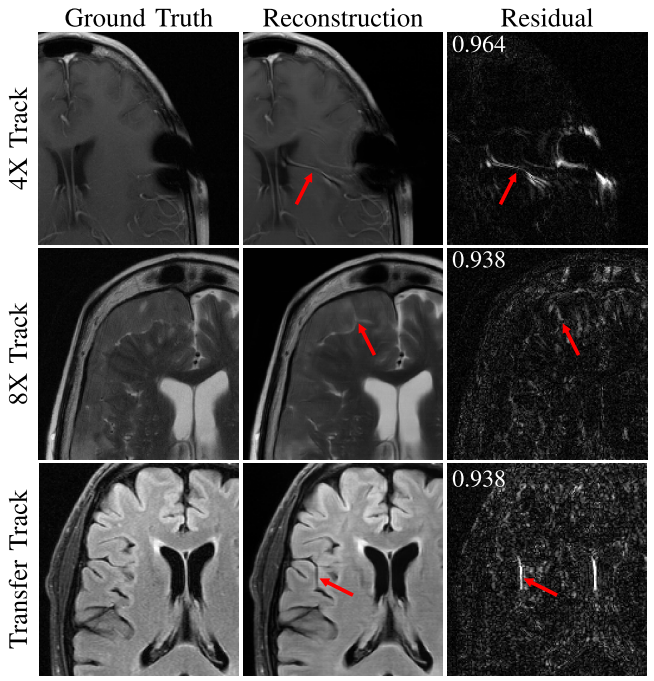


Fig. 6. Examples of reconstruction hallucinations among challenge submissions with SSIM scores over residual plots (residuals magnified by 5). (*top*) A 4X submission from Neurospin generated a false vessel, possibly related to susceptibilities introduced by surgical staples. (*middle*) An 8X submission 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. (*bottom*) A submission from ResoNNance introduced a false sulcus or prominent vessel.

In the 4X and 8X tracks, discordance was primarily driven by two submissions (Neurospin and ATB) that were very close in SSIM score. For the Transfer track, separation among the teams was more clear, and we observed corresponding increases in concordance.

Radiologists did take note of hallucinatory effects introduced by the submission models. Figure 6 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.

## IV. DISCUSSION

### A. 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 [48]. 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. We note that as designed the Transfer track primarily evaluated one type of transfer: generalization across vendors. This was the most commonly-cited type of transfer in feedback from the 2019 challenge, but future challenges may investigate other types of transfer.

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 acceptable, indicating that these two tasks may remain open problems going into 2021.

We note that the results of this paper were observed within the regime of retrospective undersampling. Retrospective undersampling does not consider potential differences in signal relaxation along the echo trains. Even though echo trains can be designed for most sequences in such a manner that undersampling does not lead to a change in overall relaxation weighting along the phase dimension, we have not shown equivalency in our work. We would recommend that researchers confirm the results of the methods in this paper with prospective sampling prior to clinical use.

The results of the challenge suggests a few conclusions on approaches. The first: cascaded models with a data fidelity term and CNN regularization continue to dominate the submission field, as occurred in the previous challenge [11]. Second: AIRS, the team that won all three tracks, had the largest model. However, large models were not always better, with ATB having a model with similar performance to Neurospin despite having 87% less parameters. Lastly, we note the AIRS model used a normalization routine to get the data into a consistent format for all coil configurations, as well as being the only team to use a GRAPPA [38] reconstruction as the initialization.

### B. 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 [63]. 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 details in Supplementary Material. 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 alternative reconstruction techniques such as adaptive combine that implicitly suppress background noise [47]. 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. Another alternative to adaptive combine would be to use other metrics that are more aware of the noise properties, such as Stein's Unbiased Risk Estimate (SURE).

One area lacking in our quantitative analysis was hallucination detection. This is an area of great interest to the community, but as of the end of our challenge we were unaware of automated, quantitative methods for detecting lesions or characterizing stability (although some methods can demonstrate instability qualitatively [64]). Detection of automated stability/hallucination analysis remains a topic of great interest for future challenges.

### C. 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 6. Such hallucinatory features are not acceptable and especially problematic if they mimic normal structures that are either not present or actually abnormal. These images had high SSIM scores, indicating that even though these images are considered well-optimized according to this metric, they are not optimized regarding hallucination features. Neural network models can be unstable as demonstrated via adversarial perturbation studies [64]. 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. This topic is in major need for further development.

We note that we did not perform intensity correction for either the participant submissions or the ground truth reconstructions. None of the radiologists commented on intensity inhomogeneity in either the initial testing phase or the final evaluation phase. However, intensity correction is routinely done by vendor scanners and could have affected the evaluation, as some pathologies manifest via varying tissue intensities.

### D. Feedback From Participants

We asked participants for feedback regarding challenge organization. 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 gives 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, multiple participants submitted models trained on only a single GPU. This was also a topic of feedback from non-participants, with some telling us informally that they did not participate due to compute or storage requirements. 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. Most vendors have variations in their post-processing pipelines for precisely this reason. Some vendor post-processing methods even allow for radiologists to adjust the strength of the regularization. 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 [48], [65]. 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. Consensus around evaluations–for ground truth calculations, metrics, and radiologist presentation–would substantially aid the organization of future challenges.

## V. 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.

## Acknowledgment

The authors would like to thank all research groups that participated in the challenge. They thank the scientific advisors that gave feedback during challenge organization. They thank Girish M. Fatterpekar, Nael Kambiz, Amy L. Kotsenas, Christie Mary Lincoln, Joseph Mettenburg, and Max Wintermark for judging the submissions and the contribution of their clinical expertise and time. They 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. They 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. The ResoNNance team's submission was conducted under the STAIRS Project under the TKI-PPP Program. The ResoNNance collaboration project is co-funded by the PPP Allowance made available by Health Holland, Top Sector Life Sciences and Health, to stimulate public-private partnerships.

## REFERENCES

[1] S. Chetlur *et al.*, "cuDNN: Efficient primitives for deep learning," 2014, *arXiv:1410.0759*. [Online]. Available: http://arxiv.org/abs/1410.0759

[2] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX Conf. Oper. Syst. Design Implement.*, 2016, pp. 265–283.

[3] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[4] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-Net for compressive sensing MRI," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 10–18.

[5] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, Jun. 2018.

[6] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 491–503, Feb. 2018.

[7] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[8] T. Eo, Y. Jun, T. Kim, J. Jang, H. Lee, and D. Hwang, "KIKI-Net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2188–2201, Nov. 2018.

[9] H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 394–405, Feb. 2019.

[10] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, Mar. 2018.

[11] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson, "Assessment of the generalization of learned image reconstruction and the potential for transfer learning," *Magn. Reson. Med.*, vol. 81, no. 1, pp. 116–128, Jan. 2019.

[12] F. Knoll *et al.*, "Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues," *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 128–140, Jan. 2020.

[13] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magn. Reson. Med.*, vol. 84, no. 6, pp. 3172–3191, Dec. 2020.

[14] A. Sriram *et al.*, "End-to-end variational networks for accelerated MRI reconstruction," in *Proc. Int. Conf. Med. Imag. Comput. Comput.-Assist. Intervent.*, 2020, pp. 64–73.

[15] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, pp. 193–202, 1980.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[17] R. Raina, A. Madhavan, and A. Y. Ng, "Large-scale deep unsupervised learning using graphics processors," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 873–880.

[18] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural Comput.*, vol. 22, no. 12, pp. 3207–3220, Dec. 2010.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[25] J. Zbontar *et al.*, "fastMRI: An open dataset and benchmarks for accelerated MRI," 2018, *arXiv:1811.08839*. [Online]. Available: http://arxiv.org/abs/1811.08839

[26] F. Knoll *et al.*, "fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190007.

[27] F. Knoll *et al.*, "Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge," *Magn. Reson. Med.*, vol. 84, no. 6, pp. 3054–3070, Dec. 2020.

[28] W. A. Grissom, K. Setsompop, S. A. Hurley, J. Tsao, J. V. Velikina, and A. A. Samsonov, "Advancing RF pulse design using an open-competition format: Report from the 2015 ISMRM challenge," *Magn. Reson. Med.*, vol. 78, no. 4, pp. 1352–1361, Oct. 2017.

[29] K. G. Schilling *et al.*, "Challenges in diffusion MRI tractography—Lessons learned from international benchmark competitions," *Magn. Reson. Imag.*, vol. 57, pp. 194–209, Apr. 2019.

[30] O. Maier *et al.*, "CG-SENSE revisited: Results from the first ISMRM reproducibility challenge," *Magn. Reson. Med.*, vol. 85, no. 4, pp. 1821–1839, Apr. 2021.

[31] Y. Beauferris *et al.*, "Multi-channel MR reconstruction (MC-MRRec) challenge—Comparing accelerated MR reconstruction models and assessing their genereralizability to datasets collected with different coils," 2020, *arXiv:2011.07952*. [Online]. Available: http://arxiv.org/abs/2011.07952

[32] E. T. Whittaker, "XVIII.—On the functions which are represented by the expansions of the interpolation-theory," *Proc. Roy. Soc. Edinburgh*, vol. 35, pp. 181–194, Sep. 1915.

[33] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. Amer. Inst. Electr. Eng.*, vol. 47, no. 2, pp. 617–644, Apr. 1928.

[34] V. A. Kotelnikov, "On the carrying capacity of the 'ether' and wire in telecommunications," in *Proc. 1st All-Union Conf. Questions Commun.*, 1933.

[35] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[36] D. K. Sodickson and W. J. Manning, "Simultaneous acquisition of spatial harmonics (SMASH): Fast imaging with radiofrequency coil arrays," *Magn. Reson. Med.*, vol. 38, no. 4, pp. 591–603, Oct. 1997.

[37] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, no. 5, pp. 952–962, Nov. 1999.

[38] M. A. Griswold *et al.*, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Magn. Reson. Med.*, vol. 47, no. 6, pp. 1202–1210, Jun. 2002.

[39] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

[40] M. Uecker *et al.*, "ESPIRiT—An eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA," *Magn. Reson. Med.*, vol. 71, no. 3, pp. 990–1001, Mar. 2014.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Imag. Comput. Comput.-Assisted Intervent.*, 2015, pp. 234–241.

[42] N. Pezzotti *et al.*, "An adaptive intelligence algorithm for undersampled knee MRI reconstruction," *IEEE Access*, vol. 8, pp. 204825–204838, 2020.

[43] P. Wang, E. Z. Chen, T. Chen, V. M. Patel, and S. Sun, "Pyramid convolutional RNN for MRI reconstruction," 2019, *arXiv:1912.00543*. [Online]. Available: http://arxiv.org/abs/1912.00543

[44] P. Putzky *et al.*, "i-RIM applied to the fastMRI challenge," 2019, *arXiv:1910.08952*. [Online]. Available: http://arxiv.org/abs/1910.08952

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[46] P. B. Roemer, W. A. Edelstein, C. E. Hayes, S. P. Souza, and O. M. Mueller, "The NMR phased array," *Magn. Reson. Med.*, vol. 16, no. 2, pp. 192–225, 1990.

[47] D. O. Walsh, A. F. Gmitro, and M. W. Marcellin, "Adaptive reconstruction of phased array MR imagery," *Magn. Reson. Med.*, vol. 43, no. 5, pp. 682–690, May 2000.

[48] M. P. Recht *et al.*, "Using deep learning to accelerate knee MRI at 3T: Results of an interchangeability study," *Amer. J. Roentgenol.*, vol. 215, no. 6, pp. 1421–1429, Dec. 2020.

[49] S. van der Walt *et al.*, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[51] Y. Jun, H. Shin, and D. Hwang, "Joint deep model-based MR image and coil sensitivity reconstruction network (joint-ICNet) for fast MRI," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021.

[52] S. Yu, B. Park, and J. Jeong, "Deep iterative down-up CNN for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[53] Z. Ramzi, P. Ciuciu, and J.-L. Starck, "Benchmarking MRI reconstruction neural networks on large public datasets," *Appl. Sci.*, vol. 10, no. 5, p. 1816, Mar. 2020.

[54] Z. Ramzi, J.-L. Starck, and P. Ciuciu, "XPDNet for MRI reconstruction: An application to the 2020 fastMRI challenge," in *Proc. Int. Soc. Magn. Res. Med.*, 2021.

[55] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, May 2011.

[56] J. Adler and O. Oktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, Jun. 2018.

[57] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782.

[58] L. Liu *et al.*, "On the variance of the adaptive learning rate and beyond," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14.

[59] P. Putzky and M. Welling, "Recurrent inference machines for solving inverse problems," 2017, *arXiv:1706.04008*. [Online]. Available: http://arxiv.org/abs/1706.04008

[60] K. Lønning, P. Putzky, J.-J. Sonke, L. Reneman, M. W. A. Caan, and M. Welling, "Recurrent inference machines for reconstructing heterogeneous MRI data," *Med. Image Anal.*, vol. 53, pp. 64–78, Apr. 2019.

[61] M. Uecker *et al.*, "Berkeley advanced reconstruction toolbox," in *Proc. Int. Soc. Magn. Res. Med.*, 2015, p. 2486.

[62] A. Defazio, T. Murrell, and M. P. Recht, "MRI banding removal via adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7660–7670.

[63] K. Hammernik, J. Schlemper, C. Qin, J. Duan, R. M. Summers, and D. Rueckert, "Σ-Net: Systematic evaluation of iterative deep neural networks for fast parallel MR image reconstruction," 2019, *arXiv:1912.09278*. [Online]. Available: https://arxiv.org/abs/1912.09278

[64] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of AI," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30088–30095, Dec. 2020.

[65] F. Moss, "Stochastic resonance and sensory information processing: A tutorial and review of application," *Clin. Neurophysiol.*, vol. 115, no. 2, pp. 267–281, Feb. 2004.