Current Issues in Deep Learning for Undersampled Image Reconstruction

C.R. Noordman*

*Image Sciences Institute, UMC Utrecht, Utrecht, Netherlands

Layman's Summary

Innovations in artificial intelligence (AI) have provided new potential in the image reconstruction of undersampled k-space acquisitions in magnetic resonance imaging (MRI). Undersampled k-space acquisitions are scans where k-space, the spatial frequency of an object, is sampled below the minimum Nyquist sampling rate. By employing machine learning, which allows for learning feature extraction from data such that is it able to make decisions for new data, it is possible to reconstruct an image given an acquisition with very few data samples.

A common problem for machine learning is the inadequate amount of data the machines need to learn from. While that problem is no exception here, this paper also explores other, more MR reconstruction-specific, problems, such as the introduction of novel artifacts by the AI or the difficulties of integrating such technology into the clinic. This review provides a number of alternative solutions to the issues presented, taken from relevant literature.

Artificial intelligence has driven a new path of innovation in image reconstruction of undersampled k-space acquisitions in MRI. This review provides readers with an analysis of the current limitations and bottlenecks in deep learning-based MR image reconstruction. Literature reviews on deep learning in medical image reconstruction commonly focus on model architectures, providing the reader with an overview of the most recent technical advances in this subject. We instead take a more critical approach, focusing on recent publications that identify and analyze current bottlenecks or limitations in deep learning-based MRI image reconstruction in the literature. Research explicitly highlighting or improving our understanding of these issues is discussed in detail. The problems discussed could be an excellent starting point for any researcher or clinician looking to further their knowledge of the predicaments found in undersampled MR reconstruction.

1 Introduction

Magnetic resonance imaging (MRI) is a popular modality in medical imaging for its flexibility, safety and good soft tissue contrast. However, its relatively long acquisition time has led to a flurry of research in improving imaging speed in both the hardware and software without many compromises in image quality. The Nyquist criterion establishes a minimum sampling density required to capture all information. Undersampling gives rise to aliasing artifacts in the reconstructed image. A remarkable improvement is compressed sensing (CS), which undersamples an acquisition well below the Nyquist criterion, then reconstructs the image using a priori information [1]. The resulting reconstruction instead commonly suffers from blurring and ringing artifacts, depending on the degree of undersampling, which is considered not as detrimental to diagnostic quality [2].

Advances in machine learning techniques and development in computational infrastructures have led to deep learning (DL) techniques becoming viable candidates to aid in medical image reconstruction. At the 2016 IEEE 13th International Symposium on Biomedical Imaging, one of the first works in this field was published [3]. This work has applied deep learning, a subclass of machine learning, which is particularly useful in feature extraction and learning from a priori information. Since then, MR image acquisition time may be expected to decrease substantially using this new technology over the coming decade. Supporting the effort are the fastMRI and the Calgary-Campinas datasets, which hope to challenge and provide a consistent benchmark for new machine learning approaches [4, 5].

These advances in deep learning-based image reconstructions are commonly met with various criticisms and concerns. Most computer vision-related tasks that

use deep learning focus on improving general image quality. MRI image reconstruction is more complex, as reconstructions should also be robust in maintaining any clinical pathologies. Results from a 2020 fastMRI challenge state that the top 3 frameworks, according to qualitative radiologist evaluation, create hallucinatory features [6]. Statistical evaluation, however, using the structural similarity index measure (SSIM) as an example, shows excellent results, with up to 95% similarities to the ground truth. It is not implausible that relevant pathologies are found in the final dissimilar 5%. It appears there is discordance between the statistical methods for evaluating reconstructed images and the radiologist-defined diagnostic quality of an image.

It is challenging to find statistical evaluation methods which directly speak to the diagnostic quality of an image, but numerous other factors may also play a part. The dataset selected among research papers is hardly given motivation, and most datasets commonly originate from a single vendor. Indeed, the fastMRI dataset is sourced from Siemens, and applying the models to similar scans but using GE or Phillips scanners greatly reduced performance [6]. This concern is highlighted by a study that applies stability tests on reconstruction models to evaluate their performance after perturbations of the data [7]. They show, among other things, that a change of vendor should be considered as a perturbation the learning algorithm is not prepared for.

Other image reconstruction algorithms position themselves as models for artifact removal, implying that the subsampled acquisitions are images beset with artifacts, rather than solving an inverse reconstruction problem [8]. Using such a definition, statistical evaluations methods cannot realistically quantify the quality of artifact removal. Blind testing performed by experienced radiologists is instead employed. Equalling or exceeding the false-positive rate of radiologists while successfully accelerating the acquisition time is the natural end goal.

This does not align well with dataset projects such as the fastMRI project, which promotes that models apply a single objective and quantitative benchmark for easier comparisons. Solutions which better quantify diagnostic image quality are critical to improving the feasibility of clinical integration of these new models. Various editorials have created guidelines in an effort to improve the reporting of studies that apply artificial intelligence to radiology [9, 10, 11].

The article is organized as follows. Section 2 describes the literature selection process. Section 3 provides a synopsis of common deep-learning models, providing particular attention to the primary bottlenecks models face. Section 4 provides an overview of proposed solutions to selected issues. Section 5 provides a final discussion and conclusion.

2 Methodology

The process used in this review uses the PRISMA protocol for identifying eligible articles for analysis (Fig. 1). Given the research question posed, the literature collection process was performed iteratively. The first iteration was collecting literature on MR image reconstruction using deep learning, with a focus on literature that introduced novel methods and strongly influenced later developments. Particular attention was devoted to the discussion sections, to extract any issues the models of the authors have encountered. Keywords used in search engines, primarily Google Scholar and Microsoft Academic, were:

- mr image reconstruction
- deep learning
- machine learning
- neural network
- undersampled k-space
- subsampled k-space

where keywords are used simultaneously with the first, "mr image reconstruction", for a more direct search. The earliest publication year was limited to 2016. Articles are excluded which did not fit the criteria.

After a thorough analysis of included literature, a second iteration is aimed at collecting literature that adequately addresses concerns voiced in the discussions of literature selected from the first iteration. This collection is used to conduct a thorough review of recent literature on the topic of MR image reconstruction using deep learning and discussing its common bottlenecks and limitations.

3 Models and their limitations

The concept of machine learning using artificial neurons was first published in 1967 [14]. Computer architectures were not as advanced, and the modern techniques applied today have only been feasible after discovering that graphics processing architectures are much more suited to machine learning [15]. Deep learning, a subset of machine learning models where multiple "deep" layers of neurons are used in a network, ideal for signal processing, has first been used to

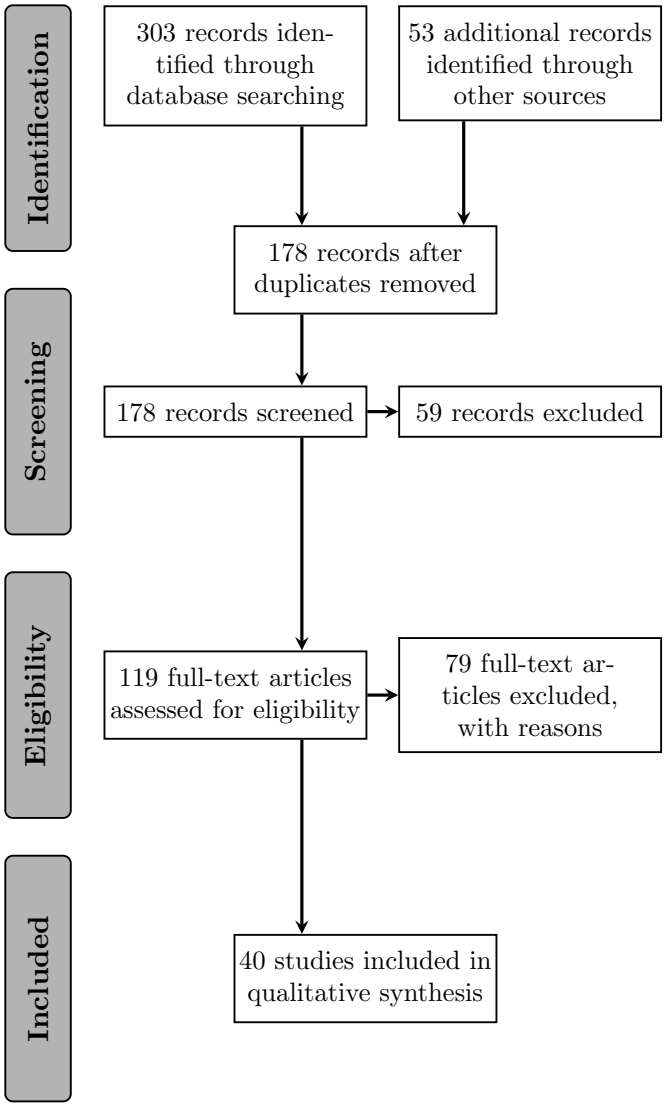

Figure 1: PRISMA flow diagram of the review process.

reconstruct undersampled acquisitions of MR data in 2016 [3].

Neural networks are highly flexible decision-making systems. Each neuron in these networks takes a vector of inputs x and a vector of weight parameters w to make a decision using some activation function $h(x)$

$$f(x) = h(w^T x + w_0) \tag{1}$$

where w_0 is a bias initialized to 1. Neurons are arranged in a layer each connected in a vector of combination parameters v. Thus, the trainable parameters of a neural network are all weight and combination parameters, traditionally represented as θ. Neurons in one layer activate the next deeper layer of neurons, creating a versatile network of decision-making. While a single layer can adequately make decisions and recognize features, using multiple layers, creating a deep structure, the network's ability to represent features increases dramatically [16].

Weight parameters are learned iteratively by using a loss function, which measures the quality of the learned parameters θ. Learning is performed by extracting features from a set of training data, then updating parameters based on the loss function.

There are many design choices to be made when developing a deep learning model: the form of the loss function, the number of layers and the number of neurons in each layer, which activation function $h(x)$ to use, and many other parameters not discussed in this brief introduction.

In MR reconstruction for undersampled MR data, deep learning entails finding weight parameters such that it can reconstruct the image using only the limited information available. The primary bottlenecks, discussed next, are the scarcity of acquisition data for training networks, the destructive role of noise found in input data, (artificially introduced) artifacts, and the difficulty of integrating this technology in existing clinical workflows.

3.1 Convolutional neural networks and data scarcity

MR data is acquired in the frequency domain, called k-space. Applying an inverse Fourier transform on this data gives a reconstructed spatial image. Low frequencies in k-space are found in the center and primarily contain contrast information of the image, while high frequencies in k-space are found away from the center and mainly contain spatial information. Undersampled k-space (that is, a sampling less than the Nyquist limit) improves acquisition time but introduces aliasing artifacts and diminishes contrast.

Wang et al. [3] tries to learn an end-to-end mapping by using a convolutional neural network (CNN) and preacquired datasets to find the hidden parameters which can optimally reconstruct an MR image using undersampled k-space data. An inverse Fourier transform is then applied to the result to obtain a spatial image. A convolutional neural network is well-suited for images, as convolution layers are neurons that activate on convolved input. This allows for neurons to model local features and learn abstractions.

A major limiting factor of all learning models is the need for data. These hidden parameters are approximated more precisely and identify more features as the dataset grows. In this seminal work, 500 fully sampled MR brain images were collected. The fastMRI project holds 1594 MR knee images. Formally speak-

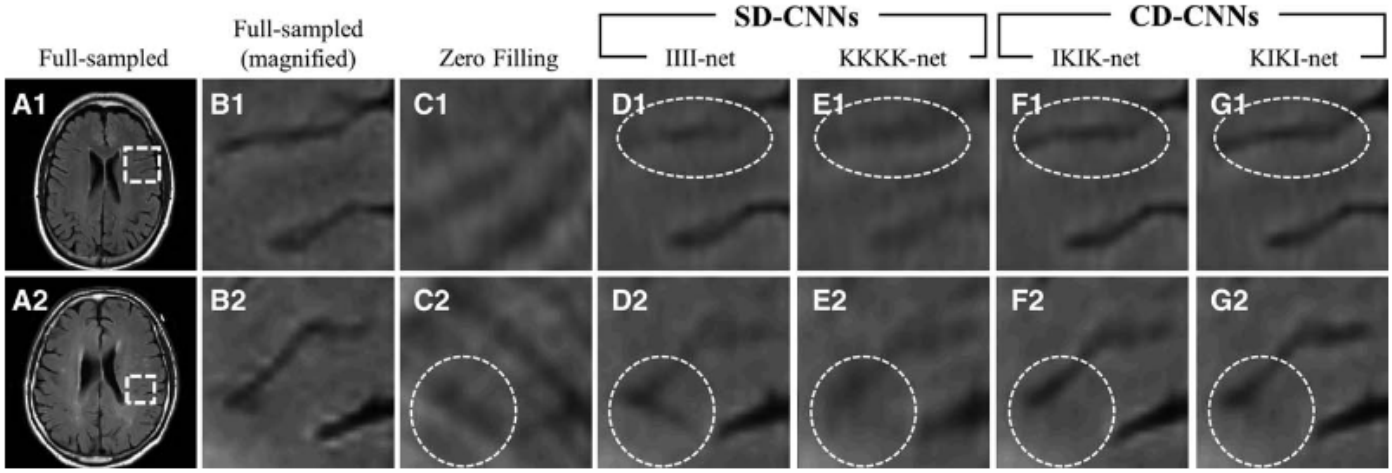

Figure 2: Reconstruction results of two undersampled acquisitions (1-2). Fully sampled image (A) and a magnified region of interest (B), zero-filled reconstruction (C) and permutations of K-nets and I-nets (D-G). The dashed circles highlight the differences of the result depending on the architecture used. Source: Eo et al. [19, Fig. 5]

ing, perfection is only obtained when there is a dataset holding all past and future acquisitions. An impossible feat, and instead, compromises are made by investigating more efficient models, and more efficient use of the datasets available.

Furthermore, MR data brings with it a few exceptional complexities. It is imaginable that using training data from healthy subjects to reconstruct images of subjects with pathology may come with an increased rate of error. Networks trained with data from one vendor may not perform as well when faced with data from another vendor, as each vendor has a different sequence in order to acquire the data. Certain attributes of data (e.g., the region of interest, the vendor that supplied it, the subject, the MR sequence used) may be more easily accessible than others, adding a concern for bias.

3.2 AUTOMAP and noisy signal

A novel framework for image reconstruction, automated transform by manifold approximation (AUTOMAP), promises to be well-protected against another common problem found in undersampled image reconstructions: the introduction of more noise [17]. Noise is enhanced in undersampled reconstruction as the training data itself is inherently noisy due to hardware limitations during acquisition. This leads to the neural network learning from noisy data, and its own imperfect learning then translates to even more noise in the result.

AUTOMAP is a more sophisticated CNN, as it performs a reconstruction without knowledge about mathematical transformations. It learns to reconstruct a spatial image directly from signal data. It does this by learning the diffeomorphism between the two manifolds which represents the signal k-space and the image space. In the context of MR image reconstruction, it learns a Fourier transform equivalent artificially.

Additionally, the framework explicitly learns a denoiser operator to alleviate any "learned" noise amplifications. Given assumed corruption in the signal domain, it attempts to find the probabilistic distribution of the input to predict it and consequently reduce any resulting noise.

A common metric in deciding on the quality of a model's output with regards to the level of noise is the peak signal-to-noise ratio (PSNR) and the root-mean-squared error (RMSE). AUTOMAP highlights the need for noise reduction in AI-generated image reconstructions. It also brings into question models which employ the standard Fourier transform reconstruction method, as a subsequent paper by the same author demonstrated further improved PSNR and RMSE of reconstructed images using this learned reconstruction model [18].

3.3 KIKI-net and sharp artifacts

KIKI-net is a 4-layer CNN that operates on k-space first (K), image space second (I), and repeat (KI).

Traditional CNNs are trained end-to-end. In KIKI-

net, the individual CNNs are trained in an incremental form. The first K-net was trained, then the next I-net was trained using the inverse Fourier transform output of the K-net. The output is the corresponding fully sampled reconstructed k-space or image, respectively. This process continued until the entire KIKI-net was fully trained.

Evaluation of K-net and I-net separately highlights a bottleneck in deep learning image reconstruction somewhat exclusive to the application of AI in image reconstruction: artifacts. According to the authors, the I-nets were especially strong in restoring detailed structures, but failed to remove artifacts. Instead, they enhanced the artifacts. The K-nets removed artifacts successfully, but have much weaker restoring capabilities. Fig. 2 gives a visual example demonstrating the complexity of the problem. The columns depict the fully sampled image (A-B), a zero-filled reconstruction (C), along with a number of different permutations of K-nets and I-nets (D-G). Compared with Fig. 2B2, Fig. 2D2 shows an enhancement of structure that does not exist. Fig. 2E2 instead is much too blurry to be useful for diagnosis. Fig. 2F2-G2 show improvements, but there remains low detail and contrast at the interface of the structures.

Unlike noise, artifacts are difficult to quantify. Common statistical metrics used for noise, such as PSNR and RMSE, are not useful here. SSIM is a more appropriate metric, as it would likely deteriorate when evaluating Fig. 2D2. However, artifacts are very localized problems, and the regions of interest for diagnosis are usually a small fraction of the entire acquisition. An SSIM of up to 95%, which are relatively common results, could still be entirely unsatisfactory if the dissimilar 5% manages to obfuscate the disease.

3.4 Clinical integration

While a large set of models have been proposed, true integration of deep learning algorithms used in real-world applications of image reconstruction has barely begun [20]. This slow crawl of integration comes with good reason. Disparities found in gender classification AIs of various vendors, where darker skin complexions and/or gender resulted in significantly increased error rates, raise questions on its reliability and concurrently give a false sense of progress [21]. Such concerns are all the more important in a clinical setting, where error rates have very strict limitations. Obfuscating pathology due to an unreliable image reconstruction technique would be a critical failure.

In an effort to bring evaluation of reconstructions closer to the intended end goal of clinical integration, subjective opinions of radiologists being included in evaluation are becoming increasingly common [22, 23, 24, 25]. Radiologists are asked to blindly rank ground truth from reconstructions. There exists no standardized evaluation of subsampled MR reconstruction, which complicates the comparison of models. However, even when including expert blinded reviewing, the acceptance rate of proposed AI-based algorithms implementing remains low [26].

The primary bottlenecks for clinical integration are the guarantee of reliability and safety of the method, easy integration into current workflows, and the building of trust [27, 28]. A difficult hurdle that is especially important in healthcare, is making any employed AI explainable [29]. When an AI is explainable, end users are much quicker to trust algorithms as they have a method to further their sense of understanding when observing end results.

4 Investigations and proposed solutions

The above bottlenecks are not at all unknown to authors working on developing MR reconstruction models, and many of the more recent publications describe their method of tackling one or more of the issues described. The articles discussed are summarized in Table 3.

Ref.	Remarks
[3]	One of the earliest publications to employ deep learning
[5]	W-net, can be summarized as a optimized "IKIKII" model.
[8]	Demonstrates domain transferring by using synthetic radial k-space input.
[17]	AUTOMAP, see 3.2.
[18]	Improved the performance of [17].
[19]	KIKI, see 3.3.
[30]	DeepCascade, heavily influenced by CS and applies interleaved data consistency stages.
[31]	Improved the computing performance of [30].
[32]	Demonstrates that focusing on subsampling the center k-lines improves performance.
[33]	Demonstrates applying small offsets to subsampling strategies improves performance.

(Continued)

Ref.	Remarks
[34]	Experiments with sampling k-lines on the fly during acquisition.
[35]	Employs reinforcement learning for on-the-fly k-line sampling.
[36]	Uses a recurrent neural network to sweep k-space in all cardinal directions.
[37]	Deep residual learning on both Cartesian and radial k-space data.
[38]	Recurrent inference machine trained with radial k-space data.
[39]	A generalist network is finetuned for undersampled MR reconstruction.
[40]	Improves [30] by reutilizing intermediary data consistency layers in the final layer.
[41]	Implements the GRAPPA technique in their deep learning algorithm.
[42]	Uses sensitivity maps as additional input in their model.
[43]	Attempts to predict sensitivity maps based on reconstruction estimations.
[44]	Implicitly learns sensitivity map estimation.
[45]	Demonstrates the value of learning the complex representations of multicoil images.
[46]	Applies interpolated CS to multislice acquisition data.
[47]	Multislice undersampled image reconstruction by exploiting correlations among and within slices.
[48]	Implements a GAN for undersampled image reconstruction.
[49]	Improves the performance of [48] by implementing a data consistency constraint.
[50]	Combines the GAN architecture with KIKI to create KIGAN.

Table 3: Results of deep learning-based image reconstruction articles discussed, in order of appearance.

4.1 Model sophistication

Since the early deep learning-based architecture, such as those proposed by Wang et al. [3], new models have been proposed and show more promising results every year. AUTOMAP and KIKI-net are both models which have strongly influenced the field of MR reconstruction using deep learning. The fastMRI project has given an excellent boost to research on reducing noise and improving performance both statistically and computationally. Models become more sophisticated as research in the field progresses, which are expected to reduce any error caused by data scarcity and improvements to reconstructions for both noise reduction and artifact removal. A number of approaches to the problem from the last few years are described and discussed.

4.1.1 Compressed Sensing

An early model for a CS-based reconstruction algorithm was proposed by Schlemper et al. [30], DeepCascade. In CS, incoherent subsampling provides a k-space $y \in \mathbb{C}^M$, which, when reconstructed, results in an image with noise superimposed over an image. The image reconstruction problem is described by

$$y = F_u x \tag{2}$$

where $x \in \mathbb{C}^N$ is the fully sampled k-space (with $N \ll M$). This inverse problem is solved by minimizing x for $F_u x - y$ using regularization. DeepCascade closely follows the original CS methods on iterative reconstruction, but employs a CNN which allows for end-to-end optimization of the algorithm.

Various improvements to CS-based reconstructions have been proposed. KIKI-net has already been discussed earlier, but is an improvement to DeepCascade, but with a prohibitive number of parameters. Sun et al. uses a recursive dilated network to drastically reduce the number of parameters [31]. The W-net model proposed by Souza et al., also employs dual-domain learning (image- and k-space) and is in essence an IKIKII model, but with significantly fewer parameters [5].

4.1.2 K-space sampling

CS with random sampling has some limitations in preserving fine details and introduces noise-like textures [32]. Instead, performing a uniform subsampling strategy with factor 2 (that is, sample every other k-line) is known not to work, as it produces an unsolvable inverse problem. Several alternatives have been proposed.

Hyun et al. uses a uniform subsampling factor 4 and includes the center k-lines, which are low-frequency lines [32]. Taking into account the conjugate symmetry property of Fourier space, Defazio notes that when using uniform subsampling, including an offset reduces the amount of symmetric redundant information sampled [33]. Zhang et al. instead argues k-line sampling should be adapted on the fly [34]. It achieves this by employing an evaluator network, which can distinguish whether a k-line is a true measurement, or from a re-

construction. The next k-line to be acquired is therefore the k-line which the evaluator is most certain to be from a reconstruction. Pineda et al. improves on this work, by instead casting the problem as a partially observable Markov decision process, and proposes to use reinforcement learning to solve it [35].

New methods are also introduced that experiment with alternative k-space trajectories. Instead of a Cartesian trajectory, radial and spiral trajectories are also possible. Methods that only use a synthesized radial profile have already shown promising improvements compared to a normal Cartesian approach [36, 37, 38].

4.1.3 Domain transferring

While radial profiles may positively augment reconstructions, most existing clinical data are acquired in a Cartesian trajectory. Given the data scarcity problem, this might result in a net loss in performance when learning a network today. However, there is evidence of successful models which transfer data from one domain to another. Transfer learning is where a network is trained on a domain with large datasets available, then transferred to a domain with scarce datasets through the use of finetuning the network. Han et al. demonstrates this potential, by using synthetic radial MR data from MR image data, or using pre-trained networks, which has learned radial computed tomography data [8]. These pre-trained networks are then fine-tuned using real radial k-space data. Transfer learning has also shown promise when transferring from large collections of natural images and public brain MR images [39]. Such methods provide a solution to the data scarcity problem.

4.1.4 Data consistency

The data consistency (DC) layer in CNNs enforces that reconstruction does not deviate from acquired k-space data. DeepCascade introduced the idea by intermittently inserting DC layers between each block [30]. Data consistency is further developed by Kocanaogullari and Eksioglu, where they included a second output to the DC layer, which holds residual image projecting the innovations made at that point [40]. The reconstruction is then based on both the intermediate rectified images and the residual images.

4.1.5 Parallel imaging

Parallel imaging exploits the relative intensities of each coil to predict the spatial location of the signal. These sensitivity maps are traditionally exploited to more accurately interpolate the missing k-space data, by using, for example, the generalized autocalibrating partial parallel acquisition (GRAPPA) technique [51]. Sriram et al. proposes GrappaNet, which are two U-nets with the GRAPPA operator in between, which leverages the additional information found in sensitivity maps [41]. Hammernik et al. also uses sensitivity maps for their VarNet as input [42]. In this model, a loss function is employed which optimizes the parameters by comparing the reconstruction with an artifact-free reference image. The sensitivity map input is estimated from fully sampled k-space, but the performance degrades significantly at higher acceleration rates. Sriram et al. proposes a solution to this problem by including a U-net, which more accurately predicts the sensitivity map after estimation [43]. Wang et al. evades the sensitivity map estimation problem with their proposed DeepcomplexMRI model, allowing for parallel imaging by implicitly learning the sensitivity maps [44]. Feng et al. similarly proposed DONet, which is capable of learning complex representations of multicoil MR images, demonstrating the value of combining frequency domains [45].

4.1.6 Multislice image reconstruction

Consecutive slices in a multislice MR acquisition have strong interslice correlation, and this correlation is exploited by using interpolated CS [46]. Xiao et al. proposes SR-net, which is uniquely capable of performing multislice undersampled image reconstruction by using deformable convolutions, which jointly exploits correlations among and within slices [47]. The results show an improvement over earlier single slice reconstruction models, which suggests that neighboring slices help with image reconstruction.

4.1.7 Alternative deep learning frameworks

Deep de-aliasing generative adversarial networks (DAGAN) is a deep learning architecture also based on CS, but instead using a conditional GAN to obtain a de-aliased reconstruction [48]. GANs consist of a generator network and a discriminator network. The discriminator is employed to distinguish between true data and data provided by the generator. A GAN is considered solved when the discriminator can no longer (or barely) discriminate between real and synthesized data. A GAN is conditional when a priori information is included. In DAGAN, only the generator receives the undersampled image as additional input. The results suggest an improvement in quality while perform-

ing reconstructions more quickly than conventional CS methods.

Since DAGAN, other methods have continued to use GAN for DL reconstruction. ReconGAN/RefineGAN is a method that adds a constraint that is intended to minimize any data consistency loss that occurred during a reconstruction [49]. KIGAN is reminiscent of KIKI-net, where two cascading networks, one for reconstruction (K) and one for image restoration (I), are used [50]. According to Lv et al., RefineGAN outperforms other methods in all metrics but one, the Visual Information Fidelity (VIF) [52]. KIGAN outperforms using the VIF metric. Paradoxically, by inspection, the authors find obvious residual artifacts in KIGAN reconstructions, while VIF is known to be highly correlated with radiologist assessment for image quality [53].

4.2 Robustness and testing

As touched on before, global quantitative methods such as SSIM, NMSE, or PSNR fail to properly reflect deficiencies in reconstructing fine details [54]. Zhao et al. shows no significant differences in SSIM in various slices with various ratios of true positive and false negatives. Their framework provides testing and comparison of algorithms based on their ability to preserve clinically important details, in addition to the common performance metrics. Instead, the much less commonly used metric VIF has shown to be highly correlated with radiological assessment of image quality [53].

Improvements specifically targetting robustness are also proposed. Defazio et al.[23] introduces the concept of adversarial loss to combat streaking artifacts, which are of particular strength in low SNR regions. The goal of the adversarial model is to predict whether a given image contains streaking artifacts. In adversarial training, a predictor loss function is focused on fooling the adversary model.

Reliability tests have been proposed by Antun et al. by using their set of instability tests [7]. Their test has three parts. Part one introduces tiny perturbations to the image, which during reconstruction is shown to lead to severe error. Part two adds significant structural change, such as a small heart shape, to the image, to see if the reconstruction algorithm can accurately recover this shape. The degree of failure in this part is varied among the algorithms tested, which included AUTOMAP. The third and final step tests the result when changing the rate of undersampling. Here, it was concluded that networks likely need to be retrained for different combinations of acquisition size, undersampling ratio, and other such parameters.

4.3 Safety and trust

Introducing novel algorithms to the clinic should follow a continual trust-building approach. McIntosh et al. promotes a framework where physicians evaluate performance both quantitatively and qualitatively, creating a feedback loop between end-user and researcher. In this framework, the goal is a prospective clinical deployment using output from AI only, emerging from a retrospective clinical simulation with expert review [26].

Topol is in disagreement, and believes AI in healthcare will not surpass beyond conditional automation — relying on a human as backup. The next step, one of high automation, where humans are only relied upon in limited circumstances, is unlikely as "human health is too precious" [20].

On the subject of trust lies an important unanswered question: when do we trust the systems we build? At what level of explainability, reliability, and accuracy are we satisfied? These are best answered in real-world settings, and Wiens et al. suggests testing systems in silence: results from MR reconstructions might be exposed to physicians and prospectively validated, then discarded [27].

5 Discussion and conclusion

Deep learning in MRI has reinvigorated the potential of undersampled MR reconstructions and has shown improvements over existing model-based solutions such as CS. Innovation is going at a rapid pace, and the ease of implementing the often open-source codebase as well as the learned parameters has allowed further development to be based only on preprints rather than peer-reviewed publications. A number of deep learning models do not provide the essential reasoning for their architectural decisions, but instead, simply provide their results with an accompaniment of various quantitative statistics. This critique can be alleviated if papers could provide a better discussion as to why certain architectures work better than others.

A clear shift can be appreciated, however, in the literature on undersampled MR reconstruction. Earlier works have primarily focused on taking an undersampled Cartesian k-space as input, and providing a synthetic k-space as output. Models such as VarNet and SR-net demonstrate that deep learning is able to extract information from multiple sources to provide a better result [47, 42]. Multidomain inputs do come

with much greater computing requirements, and there is reason to consider whether the benefits outweigh the costs. Nevertheless, more recent works are investigating how supplementary data, such as sensitivity maps or coil configurations, may result in more radiologically sound images.

An image is only useful for diagnosis if it is able to extract anomalies. Radiology is about classifying whether an image contains pathology, and the goal is therefore not to paint a pretty picture, but to ease classification. From a sample of 26 deep learning image reconstruction papers which attempted to outperform others, 22 used SSIM, 18 used PSNR and 17 used MSE in their results to conclusively demonstrate the performance of their methods. However, each of these metrics is all strongly correlated with the level of noise, and do not properly reflect radiologically sound image quality [55].

The method by which we measure the quality of our reconstructions may indeed be flawed. The goal of a reconstruction is to find pathology using only an undersampled k-space. Instead, we are deceiving ourselves by understanding the problem as one of minimizing noise across the entire image. If there were two reconstructions, one with a 99% SSIM but the remaining percent holds pathological information, and another with a significantly lower SSIM but does accurately provide pathology, the preference would be the latter. Perhaps by aiming for high SSIM, important details are obscured in the name of noise reduction. A suggestion would be to design models which explicitly reconstruct for specific pathology (e.g. for the benefit of detecting prostate cancer).

The true metric towards measuring the performance of a reconstruction model is its ability to provide a radiologist with the information they need to perform diagnosis. While initial learning can presumably be done using global metrics, fine-tuning should ideally be an iterative process between radiologist interpretations and the AI engineer, slowly and carefully improving upon the model. Such a methodology neatly falls in line with the discussion on trust, as expert radiologists can actively engage and observe the improvements, which further develops trust.

MR reconstruction from undersampled acquisitions has been given a restart due to new innovations in AI. We have addressed a number of prevailing bottlenecks which impede progression: data scarcity, the complicated nature of noise in MR, (artificially introduced) artifacts, and deploying the technology in the clinical workflow. However, very recent publications show promise that the solutions presented in this paper are contemplated upon. The issues and solutions presented in this paper are not to be considered exhaustive, but are recommended to be acknowledged for the benefit of any future development. With this publication, we hope to have made a successful emphasis on the concerns summarized, in an effort to further improve deep learning models for undersampled k-space reconstruction.

References

[1] E J Candes, J Romberg, and T Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, February 2006.

[2] Oren N Jaspan, Roman Fleysher, and Michael L Lipton. Compressed sensing MRI: a review of the clinical literature. *Br. J. Radiol.*, 88(1056): 20150487, September 2015.

[3] Shanshan Wang, Zhenghang Su, Leslie Ying, Xi Peng, Shun Zhu, Feng Liang, Dagan Feng, and Dong Liang. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 514–517, April 2016.

[4] 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. November 2018.

[5] Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *Neuroimage*, 170:482–494, April 2018.

[6] Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik

Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W Lui, and Florian Knoll. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. *IEEE Trans. Med. Imaging*, 40(9):2306–2317, September 2021.

[7] 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. *Proc. Natl. Acad. Sci. U. S. A.*, 117(48):30088–30095, December 2020.

[8] Yoseob Han, Jaejun Yoo, Hak Hee Kim, Hee Jung Shin, Kyunghyun Sung, and Jong Chul Ye. Deep learning with domain adaptation for accelerated projection-reconstruction MR. *Magn. Reson. Med.*, 80(3):1189–1205, September 2018.

[9] Issam El Naqa, John M Boone, Stanley H Benedict, Mitchell M Goodsitt, Heang-Ping Chan, Karen Drukker, Lubomir Hadjiiski, Dan Ruan, and Berkman Sahiner. AI in medical physics: guidelines for publication. *Med. Phys.*, 48(9): 4711–4714, September 2021.

[10] John Mongan, Linda Moy, and Charles E Kahn, Jr. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol Artif Intell*, 2(2):e200029, March 2020.

[11] David A Bluemke, Linda Moy, Miriam A Bredella, Birgit B Ertl-Wagner, Kathryn J Fowler, Vicky J Goh, Elkan F Halpern, Christopher P Hess, Mark L Schiebler, and Clifford R Weiss. Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and Readers-From the radiology editorial board. *Radiology*, 294 (3):487–489, March 2020.

[12] Shanshan Wang, Taohui Xiao, Qiegen Liu, and Hairong Zheng. Deep learning for fast MR imaging: A review for learning reconstruction from incomplete k-space data. *Biomed. Signal Process. Control*, 68:102579, July 2021.

[13] Emmanuel Ahishakiye, Martin Bastiaan Van Gijzen, Julius Tumwiine, Ruth Wario, and Johnes Obungoloch. A survey on deep learning in medical image reconstruction. *Intelligent Medicine*, May 2021.

[14] Aleksei Grigorevich Ivakhnenko, A G Ivakhnenko, Valentin Grigorevich Lapa, and Valentin Grigorevich Lapa. *Cybernetics and forecasting techniques*, volume 8. American Elsevier Publishing Company, 1967.

[15] Rajat Raina, Anand Madhavan, and Andrew Y Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880, 2009.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[17] 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, March 2018.

[18] N Koonjoo, B Zhu, G Cody Bagnall, D Bhutto, and M S Rosen. Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction. *Sci. Rep.*, 11(1):8248, April 2021.

[19] 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. *Magn. Reson. Med.*, 80(5):2188–2201, November 2018.

[20] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.*, 25(1):44–56, January 2019.

[21] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[22] Seok Hahn, Jisook Yi, Ho-Joon Lee, Yedaun Lee, Yun-Jung Lim, Jin-Young Bang, Hyunwoong Kim, and Joonsung Lee. Image quality and diagnostic performance of accelerated shoulder MRI with deep Learning-Based reconstruction. *AJR Am. J. Roentgenol.*, September 2021.

[23] Aaron Defazio, Tullie Murrell, and Michael P Recht. MRI banding removal via adversarial training. January 2020.

[24] 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, Mohammad 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 3 t: Results of an interchangeability study. *AJR Am. J. Roentgenol.*, 215(6):1421–1429, December 2020.

[25] Sebastian Gassenmaier, Saif Afat, Dominik Nickel, Mahmoud Mostapha, Judith Herrmann, and Ahmed E Othman. Deep learning-accelerated t2-weighted imaging of the prostate: Reduction of acquisition time and improvement of image quality. *Eur. J. Radiol.*, 137:109600, April 2021.

[26] Chris McIntosh, Leigh Conroy, Michael C Tjong, Tim Craig, Andrew Bayley, Charles Catton, Mary Gospodarowicz, Joelle Helou, Naghmeh Isfahanian, Vickie Kong, Tony Lam, Srinivas Raman, Padraig Warde, Peter Chung, Alejandro Berlin, and Thomas G Purdie. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat. Med.*, 27(6):999–1005, June 2021.

[27] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.*, 25(9):1337–1340, September 2019.

[28] Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36, 2019.

[29] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[30] 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 Trans. Med. Imaging*, 37(2):491–503, February 2018.

[31] Liyan Sun, Zhiwen Fan, Yue Huang, Xinghao Ding, and John Paisley. Compressed sensing MRI using a recursive dilated network. In *Thirty-Second AAAI Conference on Artificial Intelligence*. aaai.org, April 2018.

[32] Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. Deep learning for undersampled MRI reconstruction. *Phys. Med. Biol.*, 63(13):135007, June 2018.

[33] Aaron Defazio. Offset sampling improves deep learning based accelerated MRI reconstructions by exploiting symmetry. December 2019.

[34] Zizhao Zhang, Adriana Romero, Matthew J Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled MRI reconstruction with active acquisition. February 2019.

[35] Luis Pineda, Sumana Basu, Adriana Romero, Roberto Calandra, and Michal Drozdzal. Active MR k-space sampling with reinforcement learning. July 2020.

[36] Changheun Oh, Dongchan Kim, Jun-Young Chung, Yeji Han, and Hyunwook Park. ETER-net: End to end MR image reconstruction using recurrent neural network. In *Machine Learning for Medical Image Reconstruction*, pages 12–20. Springer International Publishing, 2018.

[37] Soumick Chatterjee, Mario Breitkopf, Chompunuch Sarasaen, Hadya Yassin, Georg Rose, Andreas Nürnberger, and Oliver Speck. ReconResNet: Regularised residual learning for MR image reconstruction of undersampled cartesian and radial data. March 2021.

[38] George Yiasemis, Chaoping Zhang, Clara I Sánchez, Jan-Jakob Sonke, and Jonas Teuwen. Deep MRI reconstruction with radial subsampling. August 2021.

[39] Salman Ul Hassan Dar, Muzaffer Özbey, Ahmet Burak Çatlı, and Tolga Çukur. A transfer-learning approach for accelerated MRI using deep neural networks. *Magn. Reson. Med.*, 84(2):663–685, August 2020.

[40] Deniz Kocanaogullari and Ender M Eksioglu. Deep learning for mri reconstruction using a novel projection based cascaded network. In *2019 IEEE*

29th International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6, October 2019.

[41] Anuroop Sriram, Jure Zbontar, Tullie Murrell, C Lawrence Zitnick, Aaron Defazio, and Daniel K Sodickson. GrappaNet: Combining parallel imaging with deep learning for Multi-Coil MRI reconstruction. October 2019.

[42] 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. *Magn. Reson. Med.*, 79(6):3055–3071, June 2018.

[43] 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. April 2020.

[44] Shanshan Wang, Huitao Cheng, Leslie Ying, Taohui Xiao, Ziwen Ke, Hairong Zheng, and Dong Liang. DeepcomplexMRI: Exploiting deep residual network for fast parallel MR imaging with complex convolution. *Magn. Reson. Imaging*, 68: 136–147, May 2020.

[45] Chun-Mei Feng, Zhanyuan Yang, Huazhu Fu, Yong Xu, Jian Yang, and Ling Shao. DONet: Dual-Octave network for fast MR image reconstruction. *IEEE Trans Neural Netw Learn Syst*, PP, July 2021.

[46] Y Pang and X Zhang. Interpolated compressed sensing mr image reconstruction using neighboring slice k-space data. In *Proceedings of the 20th Annual Meeting of ISMRM*, page 2275, 2012.

[47] Zhiyong Xiao, Nianmao Du, Jianjun Liu, and Weidong Zhang. SR-Net: A sequence offset fusion net and refine net for undersampled multislice MR image reconstruction. *Comput. Methods Programs Biomed.*, 202:105997, April 2021.

[48] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, David Firmin, Jennifer Keegan, Greg Slabaugh, Simon Arridge, Xujiong Ye, Yike Guo, Simiao Yu, Fangde Liu, David Firmin, Pier Luigi Dragotti, Guang Yang, and Hao Dong. DAGAN: Deep De-Aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imaging*, 37(6):1310–1321, June 2018.

[49] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss. *IEEE Trans. Med. Imaging*, 37(6): 1488–1497, June 2018.

[50] Roy Shaul, Itamar David, Ohad Shitrit, and Tammy Riklin Raviv. Subsampled brain MRI reconstruction by generative adversarial neural networks. *Med. Image Anal.*, 65:101747, October 2020.

[51] 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: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.

[52] Jun Lv, Jin Zhu, and Guang Yang. Which GAN? a comparative study of generative adversarial network-based fast MRI reconstruction. *Philos. Trans. A Math. Phys. Eng. Sci.*, 379(2200): 20200203, June 2021.

[53] Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. *IEEE Trans. Med. Imaging*, 39(4):1064–1072, April 2020.

[54] Ruiyang Zhao, Yuxin Zhang, Burhaneddin Yaman, Matthew P Lungren, and Michael S Hansen. End-to-end AI-based MRI reconstruction and lesion detection pipeline for evaluation of deep learning image reconstruction. (arXiv:2109.11524), September 2021.

[55] Richard Dosselmann and Xue Dong Yang. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5(1): 81–91, 2011.