

Robustness and Inherent Ambiguity in Deep Learning-Based Accelerated MRI Reconstruction

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M.Sc. Jan Nikolas Morshuis
aus Minden

Tübingen
2026

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:

24.06.2026

Dekan:

Prof. Dr. Thilo Stehle

1. Berichterstatter:

Ass. Prof. Dr. Christian Baumgartner

2. Berichterstatter:

Prof. Dr. Matthias Hein

3. Berichterstatterin:

Prof. Dr. Kerstin Ritter

Abstract

EN: Accelerated MRI acquisition can improve the cost and efficiency of clinical imaging, thereby making high-quality treatment accessible to a larger portion of the population. However, accelerated MRI acquisitions commonly involve a trade-off: the quality and interpretability of the acquired images diminish, making subsequent manual clinical diagnosis unreliable. Modern Machine Learning (ML) methods can reconstruct accelerated MRI images and generate reconstructions with high average similarity scores to the fully sampled image. However, common similarity metrics are only a poor proxy for the clinical utility of an MRI image: Even though a reconstruction appears similar to the fully sampled image, its diagnostic content can be vastly different if, for example, a small pathology is not reconstructed correctly.

This insight leads to crucial changes in perspective regarding the robustness and reliability of commonly utilized ML-based reconstruction methods. As small changes in the reconstructed image can lead to diagnostic errors, we need to challenge common assumptions within the MRI reconstruction literature. First, we show that worst-case measurement noise should be evaluated in terms of its effect on clinical diagnosis rather than on similarity scores. We show that previous papers therefore overestimate the robustness of ML-based reconstruction methods. Second, we show that the clinical utility of a reconstruction is not accurately reflected in terms of its similarity scores, as high-quality reconstructions are not necessarily required to predict high-quality segmentations from accelerated MRI images. Surprisingly, we observe that classical reconstruction methods with low similarity scores can lead to the highest segmentation scores.

Additionally, we show that evaluating reconstruction methods in terms of their clinical utility, such as segmentation and pathology detection, necessitates consideration of the inherent reconstruction ambiguity. MRI reconstruction is an ill-posed problem, for which an infinite set of possible and plausible solutions exists. ML-based MRI reconstruction methods tend to reconstruct the most probable images or those that lead to the Minimum Mean Square Error (MMSE). In both cases, rare and small pathologies or image abnormalities, such as small cartilage volumes, are unlikely to be reconstructed correctly at high acceleration factors. We introduce methods to mitigate this problem by generating semantically diverse reconstructions that are both data-consistent and plausible, yet less probable. These methods not only provide interpretable uncertainty estimates but also detect more pathologies.

DE: Beschleunigte MRT-Messungen haben das Potenzial, die Kosten und Effizienz klinischer Abläufe zu verbessern und somit eine hochwertige diagnostische Bildgebung einem größeren Teil der Bevölkerung zugänglich zu machen. Jedoch haben beschleunigte MRT-Verfahren auch Nachteile: Die Qualität und Interpretierbarkeit der akquirierten Bilder nehmen ab, was eine nachfolgende klinische Diagnose sehr erschweren kann. Moderne Machine-Learning (ML)-Methoden können diese beschleunigten MRT-Bilder rekonstruieren und Ergebnisse generieren, die hohe mittlere Ähnlichkeitswerte gegenüber den vollständig abgetasteten (fully sampled) Bildern aufweisen. Allerdings sind gängige Ähnlichkeitsmetriken nur ein schlechter Indikator für den klinischen Nutzen eines MRT-Bildes: Selbst wenn eine Rekonstruktion eine hohe Ähnlichkeit zum originalen Bild aufweist, kann sich die semantische Bedeutung des Bildes stark unterscheiden, falls z.B. eine kleine Pathologie nicht korrekt rekonstruiert wird.

Diese Erkenntnis führt zu einem entscheidenden Perspektivwechsel hinsichtlich der Robustheit und Verlässlichkeit bekannter Rekonstruktionsmethoden. Da kleine Veränderungen im rekonstruierten Bild zu diagnostischen Fehlern führen können, ist es notwendig, übliche Annahmen der MRT-Rekonstruktionsliteratur zu hinterfragen. Erstens zeigen wir, dass Worst-Case-Rauschen in Bezug auf seine Auswirkungen auf die klinische Diagnose evaluiert werden sollte und nicht anhand von Ähnlichkeitsmetriken. Wir zeigen, dass frühere Arbeiten die Robustheit von ML-basierten Methoden daher überschätzen. Zweitens zeigen wir, dass der klinische Nutzen einer Rekonstruktion nicht akkurat durch Ähnlichkeitswerte widerspiegelt wird, da hochqualitative Rekonstruktionen nicht zwingend erforderlich sind, um präzise Segmentierungen aus beschleunigten MRT-Daten vorherzusagen. Überraschenderweise beobachten wir, dass klassische Rekonstruktionsmethoden mit geringen Ähnlichkeitswerten zu den besten Segmentierungsergebnissen führen können.

Außerdem zeigen wir, dass die Evaluation der Rekonstruktionsmethoden in Bezug auf ihren klinischen Nutzen, wie z.B. Segmentierung und Pathologiedetektion, es notwendig macht, die inhärente Ambiguität der Rekonstruktion zu berücksichtigen. MRT-Rekonstruktion ist ein inverses Problem, für das es unendlich viele mögliche und plausible Lösungen gibt. ML-basierte Methoden neigen dazu, die wahrscheinlichsten Bilder oder solche, die zum minimalen mittleren quadratischen Fehler (MMSE) führen, zu rekonstruieren. In beiden Fällen kann es vorkommen, dass bei hohen MRT-Beschleunigungen seltene und kleine Pathologien oder Abnormalitäten wie z.B. ein kleines Knorpelvolumen nicht korrekt rekonstruiert werden. Wir führen neue Methoden ein, um dieses Problem zu mildern, indem wir semantisch diverse Rekonstruktionen generieren, die sowohl datenkonsistent als auch plausibel sind, aber weniger wahrscheinlich sind. Diese Methoden bieten nicht nur interpretierbare Unsicherheitsabschätzungen, sondern detektieren auch mehr Pathologien.

Acknowledgments

First and foremost, I would like to thank my supervisors, Christian Baumgartner and Matthias Hein, for their guidance throughout my PhD journey. Our discussions and diverse points of view were instrumental in shaping papers that also offer new perspectives and novel ideas.

I am also grateful to Thomas Küstner, who, as an expert in deep learning-based MRI reconstruction, offered valuable feedback on our work. Through many discussions in Tübingen and at MICCAI conferences, he provided valuable insights that significantly improved this research. My thanks also go to Sergios Gatidis, whose radiological expertise made essential contributions regarding the clinical relevance of changes in reconstructed MRI images. I would also like to thank Wieland Brendel for the discussions we had during our three TAC meetings, which often clarified the research direction.

Furthermore, I want to thank my colleagues from both the Baumgartner and the Hein groups. The exchange of ideas with group members led to crucial insights that were often pivotal in shaping the contributions of this work. In particular, special thanks go to Christian Schlarmann and Paul Fischer for proofreading this thesis and for our successful collaboration on various papers and a MICCAI challenge, as well as to Stefano Woerner for the regular discussions throughout our PhDs.

I gratefully acknowledge the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding this project under Germany's Excellence Strategy - EXC number 2064/1 - Project number 390727645. I also thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting the research and for organizing many relevant workshops and retreats, which facilitated valuable exchange with colleagues from Tübingen and Stuttgart.

Finally, and most importantly, I want to thank my friends and my family, especially my parents and grandfather. They took a keen interest not only in my research but also in my well-being. Without their unwavering emotional support, I would not have finished this thesis.

Jan Nikolas Morshuis
Tübingen, December 2025

Abbreviations

CNN	Convolutional Neural Network
CS	Compressed Sensing
CT	Computed Tomography
DL	Deep Learning
DM	Diffusion Model
DNN	Deep Neural Network
FPN	Feature Pyramid Network
FT	Fourier Transform
IID	Independent and Identically Distributed
IoU	Intersection over Union
mAP	Mean Average Precision
ML	Machine Learning
MMSE	Minimum Mean Square Error
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NMS	Non-Maximum Suppression
NN	Neural Network
PSNR	Peak Signal-to-Noise Ratio
RoI	Region of Interest
RPN	Region Proposal Network
RR	Repeated Reconstruction
SDR	Semantically Diverse Reconstruction
SGR	Segmentation Guided Reconstruction
SOTA	State-of-the-Art
SSIM	Structural Similarity
TV	Total Variation
VarNet; VN	Variational Network
ViT	Vision Transformer

List of Publications

Peer Reviewed Publications

1. Jan Nikolas Morshuis, Sergios Gatidis, Matthias Hein, and Christian F Baumgartner. “Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations”. In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2022, pp. 24–33. DOI: https://doi.org/10.1007/978-3-031-17247-2_3

Own contribution: I developed the idea together with my co-authors, trained the neural networks, analyzed the data, created the figures and served as the primary author of the manuscript.

2. Jan Nikolas Morshuis, Matthias Hein, and Christian F Baumgartner. “Segmentation-guided MRI Reconstruction for Meaningfully Diverse Reconstructions”. In: *MICCAI Workshop on Deep Generative Models*. Springer. 2024, pp. 180–190. DOI: https://doi.org/10.1007/978-3-031-72744-3_18

Own contribution: I was responsible for training the diffusion model and the segmentation model, co-developed the method together with my supervisors, analyzed the data, and created the figures. I was the primary author of the manuscript, that was written together with my supervisors.

3. Jan Nikolas Morshuis, Christian Schlarmann, Thomas Küstner, Christian F. Baumgartner[†], and Matthias Hein[†]. “Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions”. In: *MICCAI*. Cham: Springer, 2025, pp. 356–366. DOI: https://doi.org/10.1007/978-3-032-04937-7_34. [†]Shared last authorship.

Own contribution: The idea was developed together with my co-authors, I trained the detection methods (ViTDet and Faster-RCNN) and the E2E-VarNet ensemble, collected experimental data, and created the figures. Fine-tuning of the ViTDet to make the network more robust against adversarial noise was done by Christian Schlarmann similarly to his previous publication [4]. As the primary author, I wrote the manuscript, which was revised and finalized in collaboration with all co-authors.

4. Jan Nikolas Morshuis, Matthias Hein, and Christian F. Baumgartner. “Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data”. In: *IEEE Transactions on Biomedical Engineering* (2026). DOI: 10.1109/TBME.2026.3679957

Own contribution: The idea for the paper was developed after analyzing the results of a challenge [6] that our team won (1st/12 teams). I initiated participation in the challenge and developed a solution with my supervisors and my colleague Paul Fischer. I implemented the solution, preprocessed the challenge data, analyzed validation scores, and submitted the solution. For the submitted paper, I implemented baseline methods, co-designed the experiments together with my supervisors, created the figures, and wrote the paper as first author in collaboration with my supervisors.

Contents

1	Introduction	1
1.1	MRI, Acceleration and Inverse Problems	1
1.1.1	Challenges of Long Acquisition Times	2
1.1.2	MRI Acceleration and the Inverse Problem of MRI Recon- struction	2
1.1.3	DL-Based MRI Reconstruction and SOTA Results	3
1.2	Key Challenges of DL-Based Reconstruction	4
1.2.1	Similarity Scores are Poor Proxies for Diagnostic Utility . .	4
1.2.2	Challenge of Model Robustness and Stability	5
1.2.3	Open Questions for Segmentation of Accelerated MRI	6
1.2.4	Inherent Reconstruction Ambiguity: From Point Estimates to Solution Space	7
1.3	Research Objectives and Contributions	8
1.4	Thesis Outline	10
2	Background	11
2.1	MRI Reconstruction Foundations	11
2.2	Classical Reconstruction Methods	12
2.3	Deep Learning-Based Reconstruction Methods	13
2.3.1	U-Net	13
2.3.2	Variational Network	14
2.3.3	Generative Models	15
2.4	Semantic Segmentation	17
2.5	Object Detection Methods	18
2.5.1	DL-Based Detection	18
2.5.2	Evaluation Metrics	20
2.5.3	Pathology Detection in Accelerated MRI	21
3	Adversarial Robustness of MR Image Reconstruction	23
3.1	Introduction	23
3.2	Related Work	24
3.3	Method Summary	24
3.4	Key Results	26
3.5	Conclusion	26
3.6	Recent Advances and Impact	28

4	Segmentation of Undersampled MRI data	29
4.1	Introduction	29
4.2	Related Work	30
4.3	Method Summary	31
4.4	Key Results	33
4.5	Conclusion	34
4.6	Recent Advances and Impact	35
5	Segmentation Guided MRI Reconstruction	37
5.1	Introduction	37
5.2	Related Work	39
5.3	Method Summary	39
5.4	Key Results	43
5.5	Conclusion	44
5.6	Recent Advances and Impact	44
6	Semantically Diverse Reconstructions	45
6.1	Introduction	45
6.2	Related Work	46
6.3	Method Summary	47
6.4	Key Results	51
6.5	Conclusion	52
6.6	Recent Advances and Impact	52
7	Discussion	53
7.1	Summary of Key Thesis Findings	53
7.2	Connecting the Dots	54
7.3	Recent Shifts in MRI Reconstruction Literature	55
7.4	Limitations	57
8	Conclusion and Future Work	59
8.1	Future Work	60
	Bibliography	61
A	Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations	71
B	Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data	83
C	Segmentation-guided MRI Reconstruction for Meaningfully Diverse Reconstructions	97

D Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions	109
---	------------

Chapter 1

Introduction

Magnetic Resonance Imaging (MRI) is a valuable diagnostic tool that has revolutionized medical imaging by providing non-invasive, high-contrast visualization of soft tissues. However, the physical limitations of MRI acquisition constrain the scanning speed, making MRI acceleration an important research objective. This chapter provides the motivation for this thesis: we briefly explain the MRI process and a common method to accelerate acquisition. We then formalize MRI reconstruction as an inverse problem and discuss the challenges posed by its inherent reconstruction ambiguity. We briefly explain current challenges in Deep Learning (DL)-based reconstruction methods and identify research gaps in the current literature. At the end of the chapter, we list our research objectives and contributions and provide a short outline of the remaining thesis.

1.1 MRI, Acceleration and Inverse Problems

MRI is a widely used tool in clinical practice because it enables visualization of internal anatomical structures. In contrast to other modalities such as X-rays or Computed Tomography (CT) scans, which are often used for bone visualizations, MRI offers unparalleled soft-tissue contrast and does not require ionizing radiation. MRI scans frequently form the basis for diagnosing diseases, detecting pathologies, segmenting tissue, and supporting further clinical decision-making. Given the clinical significance of the MRI image, it is crucial that the image is of high quality and accurately depicts the underlying anatomy.

MRI scanners use strong magnetic fields and radio frequency waves to interact with hydrogen atoms in the patient's body. Multiple receiver coils distributed around the patient subsequently measure these signals. By applying varying magnetic fields and radio frequencies, the scanner spatially encodes the signal. Receiver coils store raw measurement data in the frequency domain, known as **k-space**. The corresponding coil images can be calculated from this k-space using the inverse Fourier transform (FT). These coil images can be combined using coil sensitivity maps S to obtain the final MRI image. Figure 1.1 provides an example of k-space data and corresponding images.

1.1.1 Challenges of Long Acquisition Times

One of the main disadvantages of MRI is the relatively long acquisition time, which often exceeds 30 minutes per session [7]. These long scan times can lead to several negative consequences: the patient throughput of a single MRI scanner is low, resulting in longer waiting times, less efficient use of the scanner, and, consequently, higher healthcare costs. Additionally, the patient may feel uncomfortable due to the narrow bore of the scanner, the machine's loud noise, and the need to remain still throughout the scan.

Long scan times can also increase the likelihood of motion artifacts, as many patients struggle to remain still throughout the entire process, especially children and patients with certain diseases. These motion artifacts can render some MRI images non-diagnostic, potentially leading to a costly re-scan. Acceleration of the MRI scans is therefore desirable, and it is currently a major research objective to find methods that allow an acceleration of the measurement process without losing significant quality or diagnostic feasibility of the resulting MRI image [8].

1.1.2 MRI Acceleration and the Inverse Problem of MRI Reconstruction

A commonly used method to accelerate MRI acquisition is to acquire fewer measurements in k-space, in particular, fewer than required by the Nyquist sampling theorem [9]. This process leads to the so-called undersampled k-space. While fewer measurements can reduce scan time, the information collected in the undersampled k-space is insufficient to reconstruct a high-quality MRI image directly using Fourier Transforms, as this would lead to artifacts that significantly degrade the image's diagnostic value. An example of the effects of an undersampled k-space is shown in Figure 1.1.

MRI reconstruction aims to predict the missing k-space data that has not been acquired, yielding a reconstructed MR image that is plausible and also consistent with the measured k-space data, a property known as **data consistency**. We can reformulate this reconstruction problem as an ill-posed inverse problem, where an infinite set of solutions yields data-consistent, plausible reconstructions. Mathematically, we can formulate the problem as having complex-valued multi-coil undersampled k-space data $\mathbf{y} \in \mathbb{C}^m$ as well as a high-quality MRI reconstruction $\mathbf{x} \in \mathbb{C}^n$ with $m \ll n$ as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{A} \in \mathbb{C}^{m \times n}$ is the forward operator and can be described by $\mathbf{A} := \mathbf{MFS}$. $\mathbf{S}$ are the coil sensitivity maps that map the image to the individual coil-images, $\mathcal{F}$ is the Fourier transform mapping the individual coil-images to the k-space, and $\mathbf{M}$

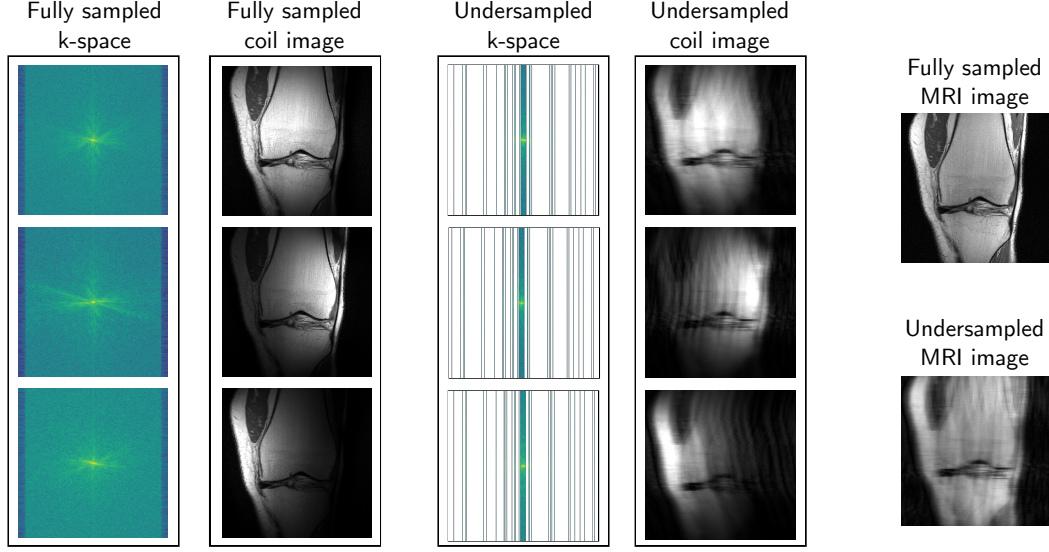


Figure 1.1: **Explanation of fully sampled and undersampled MRI data.** Multiple receiver coils (typically $\approx \mathcal{O}(10)$) acquire the magnetic resonance signal shown in the k-space data. The corresponding coil image can be calculated with an inverse Fourier Transform. The coil images can be combined using coil sensitivity masks S to obtain the final MRI images. In **accelerated MRI**, less k-space data is acquired (undersampled) and the corresponding images show artifacts and are not diagnostically useful. An acceleration of $8\times$ has been applied in this example.

is the mask-operator, masking the unmeasured parts of the k-space. The added ϵ accounts for measurement noise. The coil sensitivity maps are either known or can be estimated from the undersampled k-space [10].

The lack of a unique solution in Equation (1.1) leads to the inherent reconstruction ambiguity, which is an irreducible uncertainty. Any single reconstruction would therefore only provide an incomplete picture of the space of all possible solutions, as it does not reveal how other feasible solutions might differ semantically from the initial one. In a safety-critical field such as medical imaging, a lack of awareness of the inherent ambiguity can lead to overconfidence in the provided solution, potentially resulting in a false diagnosis with severe consequences for the patient, who might receive no treatment or wrong treatment for their symptoms. Popular and state-of-the-art (SOTA) DL methods for MRI reconstruction can potentially even worsen the problem, as we will describe in Section 1.2.

1.1.3 DL-Based MRI Reconstruction and SOTA Results

In recent years, DL-based reconstruction methods have achieved SOTA results for MRI reconstruction. They can find plausible and high-quality MRI images that are

candidate solutions to Equation (1.1) given undersampled k-space data [11, 12, 13, 14]. A key factor in the high performance of DL-based reconstruction methods is their ability to leverage large training sets to learn complex priors. During inference, this learned prior knowledge allows the model to predict a solution that is both plausible and data-consistent. These methods often achieve high performance on standard similarity metrics that quantify the similarity between the reconstructed MR image and the fully sampled MR image, such as the Structural Similarity (SSIM) [15] and the Peak Signal-to-Noise Ratio (PSNR) [16].

1.2 Key Challenges of DL-Based Reconstruction

Despite the high similarity scores achieved by current DL-based methods, they usually tend to ignore the inverse nature of the reconstruction problem, as explained in Section 1.1.2. The problem of MRI reconstruction is typically ill-posed, and there are infinitely many solutions that are both data-consistent and plausible.

Current research on DL-based reconstruction often neglects this ambiguity and typically only analyses a single-point estimate for each inverse reconstruction problem. Regression-based reconstruction models such as U-Nets (see section 2.3.1) or Variational Networks (see section 2.3.2) that are trained with paired undersampled and fully sampled data tend to predict *averaged* solutions that can yield blurry reconstructions [17, 18]. Diffusion models (see section 2.3.3) can, in principle, sample multiple different reconstructions according to the posterior distribution, yet most of these solutions to the inverse problem will be one of the most *likely* solutions. If only a limited set of reconstructions is sampled, unlikely cases can be missed, and the samples might be repetitive and close to the most likely solutions [19].

While these properties of DL-based reconstruction methods lead to high average similarity scores, both methods also pose a serious risk of omitting rare and small pathologies, which are, however, often essential for the utility of the reconstructed MRI image.

1.2.1 Similarity Scores are Poor Proxies for Diagnostic Utility

Commonly used similarity metrics, such as the SSIM and PSNR, measure the similarity of an MRI reconstruction to the fully sampled ground truth image, but can be poor proxies for diagnostic utility. Crucially, a reconstruction can achieve a high average similarity score on a test set while simultaneously distorting or even omitting subtle but clinically relevant information, such as rare pathologies or abnormal anatomical structures. Figure 1.2 illustrates a case where a pathology is omitted in the reconstruction.

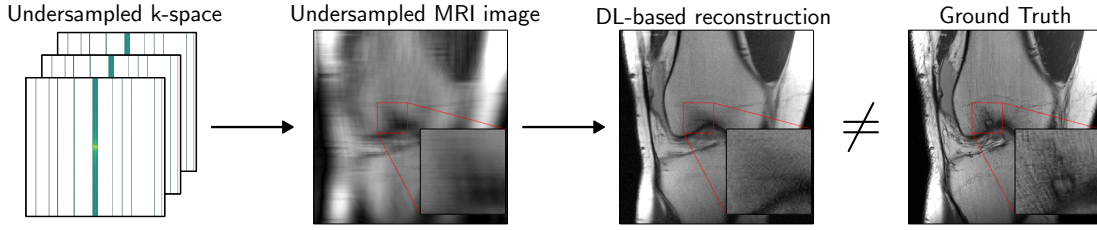


Figure 1.2: **Failure case of a DL-based reconstruction.** While the reconstruction appears visually plausible, the critical pathology has not been reconstructed correctly (see magnified view). This demonstrates the risk on relying on a single reconstruction, which can miss the pathology and can therefore lead to misdiagnosis with severe consequences for the patient.

To evaluate a reconstruction algorithm, it is therefore necessary to use metrics better aligned with the clinical objectives of the MRI scan. This shifts the evaluation focus from an average-similarity measurement to a task-based evaluation, in which the quality of the reconstruction is assessed by how well information about the underlying anatomy relevant to a downstream diagnostic task is preserved. This could mean, for example, analyzing volumetric agreement between cartilage segmentations to determine whether a patient suffers from arthrosis. It could also mean achieving diagnostic agreement between the reconstructed and fully sampled MRI images, to ensure that no pathologies are omitted or hallucinated during the reconstruction process and that they are reconstructed faithfully.

Shifting the evaluation focus from image-wide similarity metrics to a task-based evaluation framework yields a different perspective. It illuminates new challenges in MRI reconstruction: For example, a change of only a few pixels in the reconstructed image can lead to significant semantic changes if it results in the hallucination or omission of a pathology. The threshold of what would be a substantial error in a reconstructed MRI image is therefore drastically lowered. Existing problems that seem minor with respect to the similarity metrics, such as instabilities of the model to worst-case measurement noise or the inherent ambiguity of the reconstruction, can therefore have serious diagnostic consequences, as will be shown in this thesis.

1.2.2 Challenge of Model Robustness and Stability

A reconstructed MR image must be a faithful representation of the underlying patient anatomy to be clinically useful and for the reconstruction method to be trustworthy. The reconstruction method should therefore accurately reconstruct the underlying image without hallucinating or omitting any pathologies. However, it is well known that DL methods are susceptible to adversarial noise [20, 21], so that small, imperceptible perturbations to the input can lead to significant changes in the output. In the context of MR image reconstruction, this means that small amounts

of measurement noise in k-space can lead to changes in the reconstructed image that, in the worst case, can even result in hallucinations or differing pathologies. A diagnosis based on such an erroneous reconstruction will also likely be incorrect, potentially with severe consequences for the patient.

The danger of this worst-case noise is generally underestimated in the research literature, as most analyses focus on the decline in similarity metrics, such as PSNR, between a clean and an attacked reconstruction [22, 23, 24]. These approaches, therefore, do not evaluate the most critical failure cases, which are not necessarily the ones with the largest average pixel intensity deviations, but instead those that result in the largest diagnostic error. It is therefore necessary to analyze the consequences of worst-case noise on the diagnostic utility of reconstructions to better understand the clinical risk posed by instabilities in DL-based reconstruction methods.

1.2.3 Open Questions for Segmentation of Accelerated MRI

The susceptibility of DL-based reconstruction methods to worst-case noise and the potential for subsequent diagnostic errors might indicate that only small MRI acceleration factors are achievable, as the methods are otherwise too susceptible to noise. However, semantic segmentation of anatomical structures is more robust to minor changes in the image, as larger structures would need to be changed. Can higher acceleration factors be safely used if the acquisition’s clinical goal is semantic segmentation of a specific tissue? Answering this question requires a clear understanding of the capabilities of SOTA methods for segmenting accelerated MRI data, which is surprisingly difficult to obtain from the previous literature.

Previous work on semantic segmentation of accelerated MRI has focused mainly on introducing novel frameworks that combine reconstruction and segmentation in a single forward pass [25, 26]. However, these methods are highly hyperparameter-dependent, as they require finding suitable combinations of segmentation and reconstruction losses. New datasets require further grid search to find the right hyperparameters, as image resolution, contrast, or the segmented anatomies may differ. Additionally, previous work often focused on single datasets or single acceleration factors, or lacked comparisons with other SOTA methods, making it difficult to understand the field’s development and preventing a useful overview of which method choices are most beneficial for the segmentation task.

Even large-scale efforts designed to provide clarity on the current state of the field and the capabilities of segmentation methods for accelerated MRI have left many open questions. The 2022 MICCAI challenge K2S [6], for example, aimed to create an unbiased benchmark. However, the submissions varied in computational budget, model, ensemble size, and training strategies. Attributing the performance to any single methodological choice is therefore nearly impossible [27, 28]. Even though our team won the challenge, this lack of clear takeaways about why certain methods succeeded invites a more systematic investigation that compares the SOTA

methods on a more equal footing and potentially generates additional insights.

1.2.4 Inherent Reconstruction Ambiguity: From Point Estimates to Solution Space

Neglecting rare pathologies can be a problem even for ideal reconstruction models that can sample from the posterior distribution perfectly. If the pathology is sufficiently rare, it might require dozens of MRI reconstructions to obtain a single sample that displays it. This is not only time-consuming but also results in many of these reconstructed images being redundant. It would also be unfeasible for a radiologist to analyze the reconstructed images to estimate the posterior distribution. Additionally, it would lead to overconfidence in the most likely solutions, further discriminating against the rare pathologies that need to be detected.

To detect these rare pathologies, the tail cases of the posterior distribution need to be reconstructed. Tail cases are less probable reconstructions that can, however, significantly influence the diagnostic outcome. Providing not only the most probable reconstruction but also additional unlikely tail cases can lead to a better understanding of the size of the possible solution space and the inherent reconstruction ambiguity with respect to the given diagnostic task. Semantic segmentation and pathology detection require different kinds of changes in the reconstructed images (low-frequency vs. high-frequency information), which is why we consider both cases separately.

Reconstruction Ambiguity I: Semantic Segmentation

For the downstream task of semantic segmentation, information about the minimum and maximum achievable segmentation volumes, given the reconstruction ambiguity, might be most relevant for determining whether a volume is abnormal and, for example, whether arthrosis is present. To obtain reconstructions that display differences in, e.g., cartilage volume, the low-frequency elements of the images that define the larger anatomical structures have to be changed, such that the volume of the corresponding segmentation changes as well. At the same time, the image needs to remain consistent with the undersampled k-space data. This problem is addressed in Chapter 5.

Reconstruction Ambiguity II: Pathology Detection

In pathology detection, uncertainty due to reconstruction ambiguity primarily concerns the presence, absence, and type of pathology. In the case of high MRI acceleration, multiple semantically diverse reconstructions might be possible, of which some might show a pathology and some might not. If this is the case, at least both types of reconstructions need to be provided to prevent false overconfidence in any

one solution and to avoid missing existing pathologies, which can often occur when reconstructing the most likely solutions. Generating reconstructions that display changes in pathological features requires altering the high-frequency components of the image, which define the fine-grained textures essential for identifying pathologies, while remaining data-consistent. Chapter 6 focuses on this problem.

1.3 Research Objectives and Contributions

The main goal of this thesis is to evaluate and improve the clinical feasibility of DL-based MRI reconstruction methods. By shifting our perspective from similarity scores to clinical use cases, such as segmentation and pathology detection, we identified new challenges: model robustness, segmentation of highly accelerated data, and mitigation of the inherent reconstruction ambiguity. By combining insights from our research, we find concepts that led to improved segmentation quality and pathology detection reliability in accelerated MRI data compared to previous methods.

1. **Evaluating Adversarial Robustness:** The first goal of this thesis is to demonstrate the non-robust behavior of popular MRI reconstruction methods if the k-space measurement data is degraded with worst-case noise. Unlike previous work, we do not focus solely on average similarity scores; instead, we show that small amounts of worst-case noise can lead to clinically relevant semantic changes, resulting in the hallucination of non-existent pathologies. We also show that small rotations of the underlying anatomy can have similar effects, potentially resulting in different diagnostic outcomes. Apart from showing the brittleness of popular reconstruction algorithms, this work also sets the stage for one of the central topics of this thesis: MRI reconstructions should be evaluated with clinical utility as the success metric rather than relying solely on similarity scores.
2. **Benchmarking Segmentation on Accelerated MRI:** The second goal is to increase understanding of the clinical utility of MRI reconstructions. For this purpose, we systematically evaluate a variety of methods for segmenting accelerated MRI data. We compare several SOTA methods across two datasets at a wide range of acceleration factors and find that many methods do not achieve statistically significant improvements in segmentation accuracy relative to a simple baseline. Surprisingly, we find that the best segmentation performance is achieved not with SOTA methods that combine reconstruction and segmentation in a single forward pass, but rather with methods that leverage existing reconstruction techniques and train a segmentation network on these reconstructions. These two-stage methods do not introduce novel

network architectures, yet consistently beat leading SOTA methods on the tested data in our experiments.

3. **Predicting Segmentation Uncertainty:** As the third goal, we introduce a novel method to enhance the interpretability of segmentation uncertainty arising from the inherent reconstruction ambiguity. While our second goal shows that high acceleration factors are feasible for achieving high average segmentation scores, it is also essential to understand the segmentation uncertainty before applying these methods in clinical practice. We introduce an intuitive concept: generating two semantically diverse reconstructions corresponding to the minimum and maximum segmentations, enabling easy understanding of the potential volumes of the underlying anatomy and the so-called uncertainty boundary. We build our method on diffusion models, which allow us to guide the reconstruction process across multiple steps. We show that our novel method yields uncertainty boundaries that are more reliable at capturing the true segmentation (i.e., at no acceleration) than those obtained by repeated reconstruction with a diffusion model. Our method, therefore, enables more reliable and interpretable uncertainty quantification.
4. **Improving Pathology Detection:** The fourth and final goal is to find methods to improve the pathology detection in accelerated MRI. As mentioned earlier, the bias of DL-based reconstruction methods to predict the most likely solutions, or average solutions, can lead to the omission of rare pathologies, resulting in false-negative diagnoses and potentially severe consequences for the patient. Similar to the semantically diverse reconstructions with respect to the segmentation volume, we developed a method that generates semantically diverse reconstructions with respect to the reconstructed pathologies. If a pathology is plausible given the data, the clinician must be alerted to it, and the potential pathology should be shown in at least one reconstruction. To achieve this, we add small amounts of noise to an initial reconstruction and optimize the added noise so that the resulting image has a large ℓ_2 distance in the embedding space of an adversarial robust detection model, while remaining consistent with the k-space measurement data. Our method not only improves the detection of pathologies but also increases common detection metrics, such as the mean Average Precision (mAP), thereby providing uncertainty estimation and increased trustworthiness while detecting more pathologies in the underlying anatomy.

All in all, the contributions in this thesis represent a significant step forward towards making DL methods for accelerated MRI reconstruction safer and more trustworthy for clinical applications. By shifting our focus from comparing similarity metrics to performance on clinically relevant downstream tasks, problems such as hallucinated or omitted pathologies and segmentation uncertainty become more

prominent. In response, we developed and validated novel frameworks that mitigate the reconstruction ambiguity, thereby improving the reliability and transparency of anatomical segmentation and pathology detection.

1.4 Thesis Outline

The remainder of the thesis focuses on a systematic development from the initial ideas towards our newly developed methods. We will motivate the reason for focusing our evaluations on clinical metrics and we will show how our developed methods for mitigating reconstruction ambiguity lead to more reliable estimates of uncertainty for clinical diagnostic tasks.

Chapter 2 provides short explanations of the methods that are currently often used in the research literature for the reconstruction of accelerated MRI and that are partly also utilized within the thesis. It also provides a brief overview of segmentation and detection methods, providing the reader with the necessary background for a better understanding of the individual research contributions.

Chapter 3 introduces our analysis of model robustness to worst-case noise and worst-case rotation, in which we show that the diagnostic feasibility of the reconstructions is affected more than previously expected. Chapter 4 is dedicated to the diagnostic utility of accelerated MRI for semantic segmentation, in which we compare several SOTA methods in an unbiased way to obtain surprising insights about the benefits of two-stage methods that treat the reconstruction and segmentation problem independently. Chapter 5 introduces our method to mitigate the effects of reconstruction ambiguity for the segmentation uncertainty. Chapter 6 introduces our method for mitigating the inherent reconstruction ambiguity in pathology detection, demonstrating that it can help detect more existing pathologies, thereby increasing the reliability of accelerated MRI.

Chapter 7 is dedicated to the Discussion section, in which we will summarize our findings, discuss their current and potential future impact, and discuss potential limitations. We conclude the thesis in Chapter 8 with a conclusion and promising directions for future work.

Chapter 2

Background

This section is intended to provide the reader with background information on different MRI reconstruction methods, some of which are being utilized throughout the thesis. It also introduces foundations and popular methods for semantic segmentation of medical images. The chapter ends with a short introduction to DL-based object detection methods that are used in this thesis and the challenges occurring for object detection on reconstructed MR images.

2.1 MRI Reconstruction Foundations

In MRI reconstruction, the goal is to find a plausible MRI image $\mathbf{x}$ that is also consistent with the measurement data $\mathbf{y}$ of the k-space. From a Bayesian perspective, we are interested in finding the probability distribution $p(\mathbf{x}|\mathbf{y})$. This probability distribution is proportional to

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x}). \quad (2.1)$$

Here, $p(\mathbf{y}|\mathbf{x})$ is the likelihood, $p(\mathbf{x})$ is the prior and $p(\mathbf{x}|\mathbf{y})$ is the posterior. Typically, we would like to either find the maximum a posteriori estimate – the most probable reconstruction, which often has high average similarity scores – or sample from the posterior distribution. To find the maximum a posteriori estimate, we can minimize the negative log posterior for improved numerical stability, such that:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} -\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) \quad (2.2)$$

Assuming IID Gaussian noise on the measurement data, we can approximate the negative log-likelihood term $-\log p(\mathbf{y}|\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$, where $\mathbf{A}$ is the forward operator mapping image space to k-space as in Eq. 1.1. The prior can be interpreted as a regularization term such that $-\log p(\mathbf{x}) = \lambda \mathcal{R}(\mathbf{x})$. The regularization term should obtain large values when the reconstructed image is an unlikely MRI image (e.g., the zero-filled reconstruction, for which all missing values are replaced by zeros) and small values when it is a plausible MRI image. One can then formulate the search for the most probable $\mathbf{x}^*$ as a regularized optimization problem.

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \mathcal{R}(\mathbf{x}) \quad (2.3)$$

The value of $\mathbf{x}$ is often approximated in an iterative procedure, in which the gradient of the posterior is utilized to update the current value for $\mathbf{x}$. Starting from a timestep $t = 0$ and optimizing until the timestep $t = T$, we can write the optimization as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{y}) = \mathbf{x}_t - \eta_t \left(\mathbf{A}^* (\mathbf{A}\mathbf{x}_t - \mathbf{y}) + \lambda \nabla_{\mathbf{x}_t} \mathcal{R}(\mathbf{x}_t) \right), \quad (2.4)$$

with $\mathbf{A}^*$ being the Hermitian adjoint of $\mathbf{A}$ and thereby mapping from the measurement space back to the image space.

Optimization procedures such as those in Equation (2.4) are common in, e.g., compressed sensing-based reconstruction methods. Similar theoretical motivations can also be found for some DL-based reconstruction algorithms [13, 29].

2.2 Classical Reconstruction Methods

Undersampling of the k-space as a strategy to accelerate MRI acquisition has been successfully applied for several decades in clinical practice, starting with parallel imaging introduced in the 1990s [30, 31], which is still commonly used in modern MRI scanners. In parallel imaging, information about the distribution of the receiver coils, along with their known spatial sensitivities, is used to induce further information in the form of spatial localization of the received signals. In clinical practice, a 3- to 4-fold acceleration can be achieved without a significant loss of information by using this technique [32].

Compressed Sensing (CS) approaches were applied to MRI acquisition in the 2000s [33, 34]. In CS, the regularization term in Equation (2.3) is typically a sparsity-enforcing constraint such as the ℓ_1 norm in the wavelet domain. Optimizing the reconstruction to be sparse in the wavelet domain yields an MR image with large smooth regions and sharp edges, which are desirable properties of a reconstructed MR image. Another typical sparsity-enforcing regularization term is minimizing the image's Total Variation (TV) [35], which penalizes differences between adjacent voxels, encouraging larger regions of similarly valued voxels and making the image piecewise-constant and smoother. Both regularization functions can be utilized as the regularization function $\mathcal{R}(\mathbf{x})$ in Equation (2.4), and an optimized reconstruction $\mathbf{x}_0$ can be found.

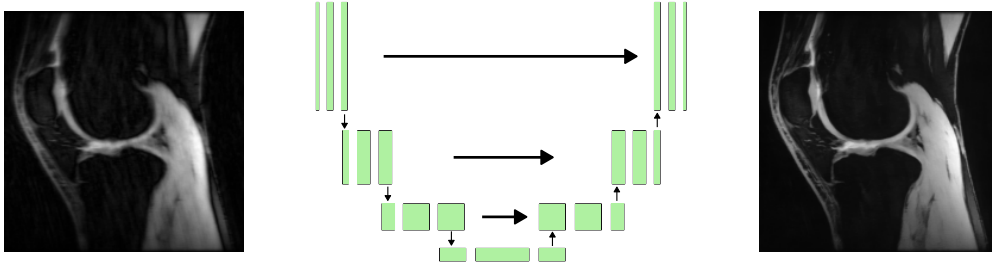


Figure 2.1: **Schematic of the U-Net architecture.** Horizontal arrows denote skip connections vertical arrows represent down- and up-sampling operations. This design enables the model to integrate fine structural details with global contextual information. The network takes an undersampled MRI image as input and reconstructs the fully sampled image.

2.3 Deep Learning-Based Reconstruction Methods

In recent years, the rise of deep learning has fundamentally advanced achievable reconstruction quality and acceleration, introducing many new state-of-the-art reconstruction methods [11, 12, 13, 14]. Unlike classical approaches, such as CS, that rely on handcrafted or at least manually selected image priors, DL methods can learn priors from large amounts of MRI data. This learning ability allows DL methods to learn more complex priors that not only optimize smooth reconstructions, but also reconstruct, e.g., specific anatomical structures accurately. This additional knowledge about the data allows, in principle, for higher reconstruction fidelity and/or higher acceleration factors than CS.

2.3.1 U-Net

A variety of DL-based approaches have been developed in recent years to solve the MRI reconstruction problem. A popular method that serves as the basis for many later approaches uses the U-Net architecture [36], for which a sketch is shown in Figure 2.1. While the U-Net was originally developed for segmentation tasks, it has also achieved remarkable performance for MRI reconstruction problems [37]. The U-Net learns to predict fully sampled images given the undersampled MRI images during training. The U-Net is especially well-suited for the reconstruction task due to its design, which is also beneficial for image segmentation tasks: it allows for combining both detailed image information, such as the exact boundary between two different tissues, with broader context information such as the type of tissue. The network consists of an encoder and a decoder that are connected at every stage by skip connections. The encoder allows the inclusion of information about the surrounding anatomical context for the prediction of the fully sampled image, while skip connections preserve information about details observable at high resolution.

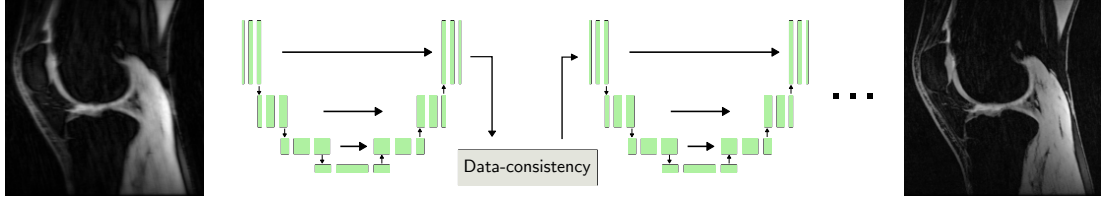


Figure 2.2: **Schematic of a Variational Network (VarNet) architecture.** A VarNet typically consists of multiple U-Net blocks in cascade. Between each U-Net block, the data-consistency of the intermediate reconstruction is increased.

2.3.2 Variational Network

Although the U-Net achieved remarkable performance in image reconstruction, it has certain limitations. The network learns to map the undersampled MRI image to the fully sampled ground-truth MRI image, yet it does not utilize the measured k-space data in its calculations. The resulting images are therefore not necessarily consistent with the measured k-space data. Consistency with the measured k-space, however, not only has the potential to increase reconstruction fidelity further but also enables the prediction to serve as a solution to the ill-posed reconstruction problem. One way to include k-space information in the prediction has been developed with unrolled networks, where variational networks (VarNets or VNs) [29] are one of the most popular applications [13, 38]. The architecture of VarNets consists of several U-Nets in cascade, as indicated in Figure 2.2. In between each of the individual U-Nets, a Fourier transform is applied to the intermediate reconstruction, and the measured part is adjusted, such that the data-consistency is increased (although it does often not need to agree perfectly at every step). An inverse Fourier Transform is applied to the 'corrected' k-space, and the image is used as input for the next U-Net in the cascade. The VarNets often achieve better reconstruction scores compared to the simple U-Net on 2-dimensional data [38, 39].

The objective of a VarNet can also be formulated in terms of the gradient update equation in Equation (2.4) as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t A^*(A\mathbf{x}_t - \mathbf{y}) + \text{CNN}(\mathbf{x}_t), \quad (2.5)$$

where the CNN (e.g., the U-Net) is learned from the training data and acts as a form of regularization, guiding the reconstruction $\mathbf{x}_t$ towards the MRI image domain [13].

Reconstruction Losses

DL-based reconstruction methods are trained using a variety of optimization losses. The most common pixel-wise loss functions are the ℓ_1 loss, utilized in [39], and the ℓ_2 loss as used in [37, 25]. Losses based on the SSIM index are also often used, with a popular example being E2E-VarNet [13]. In contrast to pixel-wise similarity

metrics such as ℓ_1 and ℓ_2 , SSIM loss encourages a network to preserve structural and contrast information. Many MRI reconstruction approaches [26, 40] linearly combine the SSIM with a ℓ_1 loss, utilizing the advantages of SSIM while preserving luminance with the ℓ_1 loss [41].

2.3.3 Generative Models

Both DL-based reconstruction methods seen previously are trained to predict a single reconstruction that ideally leads to a small average loss, e.g. the Minimum Mean Square Error (MMSE). Additionally, U-Nets and Variational Networks are trained on paired data, for which the models have learned to predict a fully sampled MR image from an undersampled one. Changes in the k-space sampling pattern or acceleration factor would result in the input data being out of distribution, leading to corresponding declines in reconstruction quality [11].

Another class of reconstruction models are the generative models. Generative models learn a prior distribution over fully sampled MRI data and are typically capable of generating realistic-looking MRI images from Gaussian noise. During the inference for an MRI reconstruction problem, an image is generated that is – ideally – consistent with the measured MRI data. In contrast to U-Nets and variational networks, generative models can typically produce multiple different reconstructions of a single undersampled MR image due to their probabilistic nature. Additionally, they are often not trained on a single undersampling pattern or acceleration factor, enabling these models to be used across a broader range of acceleration factors and undersampling patterns.

Several families of generative models exist. Variational Autoencoders (VAEs) [42, 43] have been among the earliest approaches to utilize generative models for MR image reconstruction, and variations are still successfully applied [44]. Generative Adversarial Networks (GANs) [45] have also demonstrated strong performance in MRI reconstruction [46, 47]. Current state-of-the-art results in MRI reconstruction literature are, however, obtained with diffusion models (DMs) [48, 11, 49]. As we utilize diffusion models in this thesis and adapt them to our applications (see Chapter 5), we focus our explanation of generative models on this method.

Diffusion models can learn complex data-distributions $p_{data}(\mathbf{x}_0)$, such as, e.g., knee MRI images, and, through iterative denoising, can sample from the learned prior $p_{\theta}(\mathbf{x}_0)$ during inference, generating new non-existent images. They can also be used to reconstruct MRI images by incorporating k-space information during inference, as explained below. Through the iterative denoising process, the rough image structure is reconstructed before the finer image details. An example of a diffusion model-based reconstruction process is shown in Figure 2.3.

A trained diffusion model can be used not only to generate images from the learned data distribution, but also to reconstruct MRI images from undersampled k-space data. Several methods have been developed for this purpose [11, 12, 50,

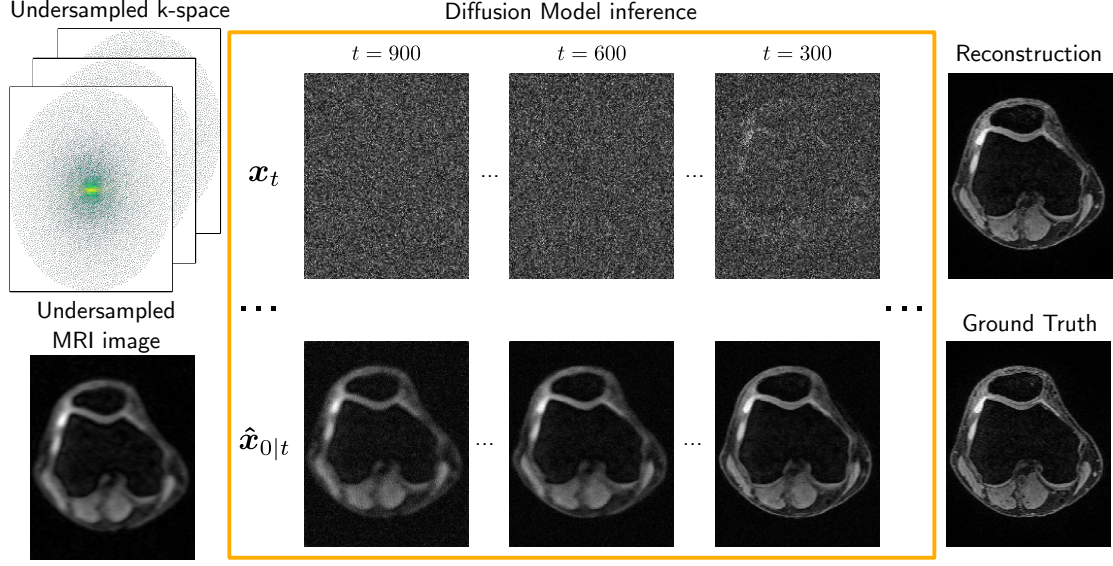


Figure 2.3: **Explanation of a diffusion model used for MRI reconstruction.** The reconstruction process consists of multiple steps t , typically starting at $t = 1000$ and ending at $t = 0$. At every step t , the diffusion model $\epsilon_\theta(\mathbf{x}_t)$ estimates the noise in the image $\mathbf{x}_t$. This prediction can be used to calculate the denoised estimate $\hat{\mathbf{x}}_{0|t}$. The reconstruction is the final prediction at $t = 0$. The example image shows a knee from the axial (top-down) direction.

51]. For our method in Chapter 5, we utilize the Decomposed Diffusion Sampler (DDS) introduced by [51], which in turn is based on Denoising Diffusion Implicit Models (DDIM) [52]. Several arguments exist that make DDS a good choice for our applications: It does not rely on many hyperparameters, it leads to SOTA results, it has a relatively short inference time (≈ 50 steps), and it provides the denoised estimate $\hat{\mathbf{x}}_{0|t}$ at every step t .

DMs generate samples from the learned data distribution by reversing a diffusion process that adds Gaussian noise to an image and reduces the signal at every timestep t . This is described by a conditional density as shown in Equation (2.6) with a pre-defined noise schedule given by a variance schedule $\beta_1, \dots, \beta_T$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) ; q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (2.6)$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

The DM is trained to predict at every timestep t the noise ϵ that has been added to the original image $\mathbf{x}_0$ according to Equation (2.6). For this purpose, a conditional U-Net ϵ_θ is utilized conditioned on the timestep t , taking as input the noisy image

$\mathbf{x}_t$. The parameters θ of the U-Net are trained via epsilon matching:

$$\min_{\theta} \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0), \mathbf{x}_0 \sim p_{data}(\mathbf{x}_0), \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})} (\|\boldsymbol{\varepsilon}_{\theta}^{(t)}(\mathbf{x}_t) - \boldsymbol{\varepsilon}\|_2^2). \quad (2.7)$$

The learned $\boldsymbol{\varepsilon}$ -matching function can be used to predict the denoised estimate $\hat{\mathbf{x}}_{0|t}$ at every step t of the diffusion process using Tweedie denoising [53]:

$$\hat{\mathbf{x}}_{0|t} = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\varepsilon}_{\theta}^{(t)}(\mathbf{x}_t)) / \sqrt{\bar{\alpha}_t}. \quad (2.8)$$

In DDS [51], conjugate gradient (CG) optimization is utilized between the timesteps t : During the CG-optimization process, starting from $\hat{\mathbf{x}}_{0|t}$ we search a $\hat{\mathbf{x}}'_{0|t}$ that minimizes the distance $\mathbf{A}^* \mathbf{A} \hat{\mathbf{x}}'_{0|t} - \mathbf{A}^* \mathbf{y}$:

$$\hat{\mathbf{x}}'_{0|t} = CG(\mathbf{A}^* \mathbf{A}, \mathbf{A}^* \mathbf{y}, \hat{\mathbf{x}}_{0|t}). \quad (2.9)$$

The update rule based on the DDIM method [52] is then given as follows:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}'_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \eta^2 \tilde{\beta}_t^2} \boldsymbol{\varepsilon}_{\theta}^{(t)}(\mathbf{x}_t) + \eta \tilde{\beta}_t \boldsymbol{\varepsilon}. \quad (2.10)$$

with $\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$.

For our work in Chapter 5, we modify the algorithm from [51] slightly to achieve full data-consistency: For the last diffusion step, we replace Eq. 2.9 and make sure that the reconstructed image $\mathbf{x}_0$ is fully data-consistent using $\mathbf{y}$ and Eq. 1.1, basically replacing the predicted k-space lines with the measured lines provided by the undersampled data. We do not encounter any artifacts in the reconstructed image when we do this, but we obtain data-consistent reconstructions. Data consistency of the predicted reconstructions is important when analyzing reconstruction ambiguity, to ensure each generated reconstruction is indeed a possible solution.

2.4 Semantic Segmentation

Semantic segmentation is a computer vision task for pixel-level classification, in which each pixel in the image is assigned a specific class (e.g., background, cartilage, bone, etc). This semantic information can be important for the clinical evaluation of the image, as it can, for example, provide information on cartilage thickness, which can be used to diagnose osteoarthritis.

For a successful semantic segmentation, it is necessary to combine the detailed pixel-level image information with the broader context information, which can be solved well by a U-Net architecture as explained in Section 2.3.1 due to the combination of an encoder, which encodes information about the anatomical surroundings, and the skip connections, that preserve high-resolution information until the final layers of the network. Even though some more recent transformer-based [54,

55] and transformer-inspired architectures [56] aim to challenge the segmentation performance of the U-Net, it was recently shown that the U-Net still achieves state-of-the-art segmentation results on many datasets [57].

Existing U-Net libraries make it particularly easy to adapt to new medical imaging datasets, automatically fitting training hyperparameters such as patch size, pre-processing parameters, dataset splits, and more. In particular, the nnU-Net library [58] is still reaching state-of-the-art performance on many datasets [57] without the need for dataset-specific manual hyperparameter adjustment. The segmentation and, to a certain extent, some reconstruction methods in Section 4 are based on the nnU-Net framework, which also enabled us to achieve top results in international MICCAI challenges (see Section 4.6). In Chapter 5, we utilized a lighter U-Net implementation from MONAI [59], as it was easier to adapt to our requirements.

Segmentation performance is most commonly measured by the Dice score [60] in medical image analysis. The formula for the Dice score is given in Equation (2.11), where TP, FP, and FN stand for True Positive, False Positive, and False Negative pixel predictions. A perfect segmentation has neither FPs nor FNs and therefore would obtain a Dice score of 1, while a segmentation without any TPs leads to a Dice score of 0.

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (2.11)$$

2.5 Object Detection Methods

In object detection, the goal is to localize and classify specific objects. In medical imaging, one might want to detect the location and class of pathologies such as meniscus tears, cartilage decline, or sprains. A DL-based detection model is trained to predict bounding boxes for every object of interest and a class label for each bounding box.

In contrast to semantic segmentation discussed in the previous section, which provides a pixel-perfect segmentation mask of an anatomy and therefore defines the exact shape and the volume of it, the detection model aims to provide an estimate of the location of the object (or the pathology) and is therefore often utilized to answer the question of whether a pathology is present or not and where it is located.

2.5.1 DL-Based Detection

In this thesis, two detection models are utilized in Chapter 6: The well-established Faster R-CNN [61] and the more recent transformer-based ViTDet [62]. Both architectures are two-stage detectors. Two-stage detectors are considered more accurate than one-stage detectors, such as YOLO [63], but typically have higher

inference time. Because the inference time of modern two-stage detectors is generally sufficiently fast for our use case (< 1 second), we focus on the two-stage detectors.

A two-stage detector typically consists of three main components: A backbone for feature extraction, a Region Proposal Network (RPN) to propose potential bounding boxes, and a Region of Interest (RoI) Head to classify the proposed bounding boxes. The individual components are explained in more detail below.

Backbone Feature Extraction

The first step is to process the input image into feature maps that encode its semantic information while simultaneously retaining the necessary spatial information. In Faster R-CNN, this is achieved using a Convolutional Neural Network (CNN) such as ResNet [64]. In CNNs, multiple layers are commonly used, for which in later layers the spatial resolution decreases while the channel dimension increases. Many modern detection networks use this behavior via a Feature Pyramid Network (FPN) [65], in which the backbone features are collected at multiple resolutions before being passed to the RPN and the RoI Head.

In contrast, ViTDet utilizes a Vision Transformer (ViT) [66] as its backbone. Unlike CNNs, a ViT typically has a single scale representation for all layers. A key contribution of ViTDet is to use only features from the ViT backbone’s last layer. This single low-dimensional representation of the image serves as a *bottleneck* in the detection pipeline, where all necessary information for further detection is encoded. In Chapter 6, we use this property to increase the robustness of the features computed by the ViT backbone to adversarial noise using a method by Schlarman et al. [4].

Region Proposal Network

The aim of the Region Proposal Network (RPN) is to act as an “attention” mechanism, proposing regions in the image that are likely to contain foreground objects rather than background. The RPN uses the backbone’s computed features as input. At each location in the feature space, it evaluates k anchor boxes that have varying sizes and aspect ratios.

For each anchor box, the RPN predicts two values: The first is an **objectness score** indicating the probability that an anchor contains a foreground object. The second value provides **box deltas** ($\Delta x, \Delta y, \Delta w, \Delta h$) to refine the position and shape of the anchor according to the underlying object.

This process generates thousands of proposal boxes. To reduce redundancy, Non-Maximum Suppression (NMS) is applied. This process reduces the number of highly overlapping boxes by keeping only the proposed boxes with the highest objectness scores if a significant overlap with other boxes exists. Finally, the top- N (typically ≈ 2000) proposal boxes are passed to the RoI Head.

RoI Head and Classification

In the second stage, features for each proposal box need to be extracted from the backbone’s feature maps for classification. The classification layers are typically fully connected and therefore require a fixed-size input, so pooling layers are needed to compute the box feature for each proposal box. These box features are then fed into the detection heads to predict the final class label (e.g., meniscus tear, cartilage decline, background, etc.) and to further refine the bounding box size and shape.

2.5.2 Evaluation Metrics

To evaluate the detection quality of DL models, two criteria must be met for successful detection: the predicted box location should be approximately correct, and the predicted class should be accurate. To measure the accuracy of the expected box location relative to the ground-truth box location, the **Intersection over Union (IoU)** is typically used. It measures the area of overlap between the boxes, divided by their combined area. Commonly, a detection is treated as a True Positive (TP) if $IoU \geq 0.5$ [61]; however, in medical image analysis, the threshold is typically reduced, as many pathologies do not show clear-cut boundaries and are often relatively small, such that slight deviations in the box-location lead to strongly reduced IoUs. Therefore, IoUs of only 0.1 are sometimes selected for pathology detection [67, 68]. In our work, we use $IoU = 0.25$ as the threshold for a positive detection.

Relevant metrics for evaluating the model’s detection quality are **precision** and **recall**. While precision ($TP/(TP + FP)$) measures the fraction of the model’s positive predictions that are indeed present in the underlying data, recall ($TP/(TP + FN)$) measures the fraction of real pathologies in the underlying anatomy that the model successfully detected. Recall is arguably the more important metric in clinical practice, as we want to avoid missing pathologies that exist in the underlying anatomy. Especially with DL models, it is critical to track this metric due to risk factors outlined in Chapter 1.

Usually, there is a trade-off between precision and recall. Increasing precision by limiting predictions to instances where the model is certain results in lower recall, as some subtle pathologies may have been missed. This trade-off is captured in the **precision-recall curve**. Using the precision-recall curve, it is possible to find a suitable model confidence threshold that achieves the desired trade-off between precision and recall. The area under the precision recall curve for a single class is the **Average Precision (AP)**. It provides a single quantity to measure overall detection model performance. The mean of the AP across all classes is called **mean Average Precision (mAP)** and is one of the most common detection metrics [61]. One can add information about the utilized IoU requirement by writing, e.g., $mAP_{0.25}$ to indicate that a box is considered successfully detected if $IoU \geq 0.25$

and the predicted class matches the ground truth.

2.5.3 Pathology Detection in Accelerated MRI

Finding subtle pathologies in medical images can be difficult, and it is well known that radiologists require a multi-year education before they are allowed to diagnose medical images. Detecting pathologies in accelerated MRI images can be particularly challenging, as some information of the image is lost due to the acceleration process, and the utilized machine learning-based reconstruction methods also have drawbacks:

1. Certain pathologies can be quite rare and therefore difficult for the ML-based reconstruction method to learn. Most slices in, e.g., the fastMRI dataset [69] show only healthy tissue. Even if a pathology exists in a given slice, the majority of the image still shows healthy tissue.
2. Reconstruction methods based on diffusion models can be biased towards the most likely solution [19]. Regression models trained on paired undersampled and fully sampled data often predict an average solution [17].
3. Pathologies in MRI images are typically small, and their shape therefore depends on the high-frequency data in the measured k-space. However, high-frequency regions in k-space are commonly measured less frequently than low-frequency regions in accelerated MRI, leading to a lack of information needed to reconstruct the MRI image correctly.

The combination of these three points makes pathology detection in accelerated MRI not only challenging, but it can also lead to dangerous outcomes for the patient. If the DL-based reconstruction methods do not reconstruct a pathology, necessary medical treatment can potentially not be started, and the patient cannot be treated, potentially leading to extended and preventable patient suffering.

Apart from the practical challenges that methods for detecting pathologies in accelerated MRI data face, research in this field also faces the problem of a very limited number of publicly available datasets suitable for this task. A dataset usable for detection on accelerated MRI needs to provide k-space data for the post-hoc undersampling simulation, as well as bounding boxes for training DL-based detection models. The only suitable publicly available datasets for this task – to the best of our knowledge – are fastMRI-plus [70] and SKM-TEA [39]. However, the SKM-TEA dataset is not applicable to our combined reconstruction + segmentation task, as the plane for detection is sagittal. In contrast, the k-space was acquired in the axial direction.

Most work on accelerated MRI reconstruction fails to explicitly assess whether pathologies are successfully preserved, despite this arguably being the most essential criterion for clinical utility. Instead, evaluations mostly rely on global similarity

metrics such as SSIM and PSNR. However, pathologies are typically rare and small, so these global metrics can be insensitive to local errors that compromise diagnostic feasibility. Consequently, downstream tasks such as object detection and semantic segmentation might be considered more relevant and clinically informed metrics [8].

Chapter 3

Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations

Source Publication

Publication:

Jan Nikolas Morshuis, Sergios Gatidis, Matthias Hein, and Christian F Baumgartner. “Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations”. In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2022, pp. 24–33. DOI: https://doi.org/10.1007/978-3-031-17247-2_3

3.1 Introduction

As established in the introduction, a critical barrier to clinical transition of DL-based reconstruction methods is their potential vulnerability to adversarial or “worst-case” noise in the measured k-space data. This chapter addresses the first research objective of systematically evaluating the brittleness of these DL-based methods. Instead of considering only global similarity metrics such as SSIM and PSNR, we focus on clinically relevant changes that can be hallucinated under worst-case noise. We found that even though image-wide similarity metrics can be deceptively high, adversarial perturbations to the measurement noise can lead to realistic-looking, yet hallucinated, pathologies that could alter the image’s diagnosis and, in turn, result in severe consequences, including false treatment decisions for the patient.

In our experiments, we analyzed the behavior of two popular DL-based reconstruction methods (U-Net and E2E-VN) under worst-case noise. We showed that for both algorithms, even small amounts of such noise can lead to clinically significant changes, as confirmed by a radiologist. These findings show that applying DL-based reconstruction methods in clinical practice poses a serious risk and underscore one

of the central themes of this thesis: the need to evaluate reconstruction based on its clinical utility rather than solely on average similarity scores.

3.2 Related Work

The effects of worst-case noise on MRI reconstruction have been investigated in several prior works. However, most prior work has focused on image-level metrics such as the MAE, rather than clinically relevant metrics, such as the preservation of pathologies. As we show in our work, the effects of attacking image-wide metrics are far milder than the effects of targeting pathology regions. A frequently cited paper by Antun et al. [22] showed that small perturbations applied to the input of a reconstruction model can indeed lead to image-wide artifacts in the reconstructions of DL models. Yet, a quantitative analysis of the effects of different perturbation sizes is missing.

Similarly, Zalbagi Darestani et al. [23] also tested the effects of worst-case noise on the reconstructed images. They have shown that the reconstruction quality of all ML-based reconstruction methods degrades with higher noise levels. However, as they measure only image-wide PSNR decline, their studied noise levels are relatively high ($\varepsilon \in [0.02, 0.08]$), in contrast to the noise levels we study ($\varepsilon \in [0.0025, 0.025]$). The qualitative examples shown in their work are exclusively from the higher end of their noise distribution ($\varepsilon = 0.08$), showing severe artifacts. In our work, we are more interested in silently failing reconstructions with very low noise levels, which can lead to false predictions because no obvious reconstruction error is visible.

Effects of smaller amounts of noise ($\varepsilon \in [0.0025, 0.025]$) have been analyzed by Genzel et al. [24]. Surprisingly, the authors conclude that standard DL-based reconstruction methods are remarkably resilient against even adversarial noise. The authors, however, only considered image-wide degradations that are indeed not seriously affected by the adversarial noise, as also confirmed by our work. However, the authors failed to see that even a small amount of noise can lead to clinically significant changes, as we have shown in our paper.

3.3 Method Summary

Dataset

Despite their relevance to evaluating the clinical utility of reconstruction methods, datasets that combine k-space data and pathology annotations are rare. The fastMRI-plus dataset [70] provides one of only two publicly available datasets that provide both k-space data and pathology annotations. The other one – SKM-TEA [39] – is not applicable for our purposes, as the annotations are provided from a different anatomical plane (sagittal) than the acquired k-space data (axial). We

therefore conduct our analysis on the fastMRI-plus data. We selected three patients from the validation set, yielding a combined total of 149 distinct annotated regions. As a single slice can contain multiple annotations, we evaluate each annotation entry individually to obtain the quantitative results.

Tested Models

We considered the effects on two popular DL-based reconstruction methods f_ϕ , the U-Net [37] as well as the E2E-VN [13] that are described in more detail in Chapter 2. A core difference between the two methods is that the U-Net is applied directly in image space, taking the undersampled image as input and predicting the reconstructed image as output. In contrast, the E2E-VN uses k-space data to improve data consistency (see Section 2.3). Even though pre-trained weights are available for both models, these models were trained on the publicly available train and validation splits of the fastMRI dataset [69]. We therefore had to retrain both models using only the training set to evaluate on the validation set. Both models require substantial computational resources, and even after reducing model sizes slightly to fit our available compute infrastructure, we still required 32 NVIDIA 2080 TI GPUs in parallel to train the E2E-VN.

Optimization Objectives for Adversarial Attacks

In the paper corresponding to this chapter, we explored two types of adversarial attacks. First, we analyzed the problem of adversarial noise added to the k -space measurements. Second, we evaluated adversarial rotations by simulating a slightly rotated position of the patient’s knee during acquisition, which can be interpreted as a more likely clinical scenario.

For the first adversarial attack method, we aimed to find a bounded adversarial noise vector $\mathbf{z}_i$ for every k -space measurement $\mathbf{k}_i$ of every coil i , such that the resulting model output $f_\phi(\cdot)$ changes as much as possible in a user-defined target region S . More specifically, we have the following optimization problem shown in Equation (3.1):

$$\max_{\mathbf{z}} \|S \odot [f_\phi(\mathcal{M}(\mathbf{k} + \mathbf{z})) - \mathbf{X}]\|_2 \quad : \quad \|\mathbf{z}_i\|_2 \leq \eta \|\mathbf{k}_i\|_2. \quad (3.1)$$

Here, $\mathbf{k}$ is the vector of all k -space data $\mathbf{k}_i$ from all coils i , η defines the ℓ_2 -adversarial noise level relative to every coil measurement $\mathbf{k}_i$, and S indicates the target region. Specifically, we used the bounding boxes indicating pathologies from the fastMRI+ dataset [70] as the selected target region for S . In contrast to [24, 23], we restricted the noise vector $\mathbf{z}_i$ for every coil individually instead of the complete measurement, thereby allowing for differences in the coil sensitivities.

For the second adversarial attack method, we can formulate a similar optimization objective as shown in Equation (3.2).

$$\max_{\theta} \|S \odot [(\mathcal{R}_{\theta}^{-1} \circ f_{\phi} \circ \mathcal{M} \circ \mathcal{R}_{\theta})(\mathbf{k}) - \mathbf{X}]\|_2 \quad \text{s.t.} \quad \theta \in [-\theta_{max}, \theta_{max}]. \quad (3.2)$$

Here, we introduced an operator $\mathcal{R}_{\theta}$ which denotes a rotation of the MRI data by θ . In practice, we rotated the data in the spatial domain and then applied a Fourier transform to obtain $\mathbf{k}$, as in [71]. After the neural network produces a prediction, we apply the inverse rotation $\mathcal{R}_{\theta}^{-1}$ to transform the image back to its original orientation for comparison with the original fully sampled image.

3.4 Key Results

We showed that optimizing for worst-case noise indeed leads to significant changes in the resulting reconstructions for both U-Net and E2E-VN. An example is shown in Figure 3.1. The fully sampled image shows a low-grade sprain of the anterior cruciate ligament (ACL), which is also visible in the reconstructed images (top row). However, adding a relative noise level of $\eta = 2\%$ can lead to significant changes in the targeted region (bottom row), making the original diagnosis infeasible based on the reconstructions, as confirmed by a board-certified radiologist. If the same amount of noise is applied to the fully sampled image, nearly no changes are detectable (bottom left).

Statistical results across the validation images are shown in Figure 3.2. As trained pathology detection models were not available at the time of the study (having been developed later for our subsequent work [3]), we evaluated the statistics using the SSIM similarity score for the given annotated pathology region. We can see that for both reconstruction methods and both acceleration factors, the decline in similarity is comparable (left). We can also see that attacking only the target regions leads to more substantial similarity declines in the target regions compared to attacking the entire image (right).

These results are also illustrated in Figure 3.3, in which the effects of focusing on a target region vs. optimizing the adversarial attack for the entire image are shown. While clear changes are visible in the targeted attack (third column), nearly no changes to the original image are detectable when targeting the whole image (fourth column).

We have obtained similar, albeit slightly weaker results for worst-case rotation and refer to our paper in Appendix A for further details.

3.5 Conclusion

We have shown that modern and popular DL methods are susceptible to "worst-case noise". Only small amounts of noise in the k-space measurement data are sufficient

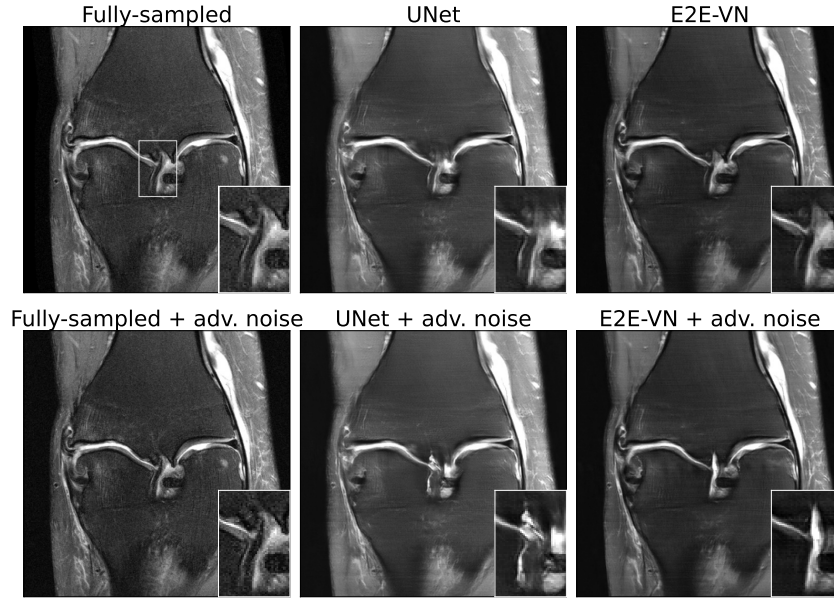


Figure 3.1: **Comparison of reconstruction robustness.** Results at $8\times$ acceleration with a relative noise level of 2%. Top row: Reconstructions with Gaussian noise. Bottom row: Reconstructions with adversarial noise

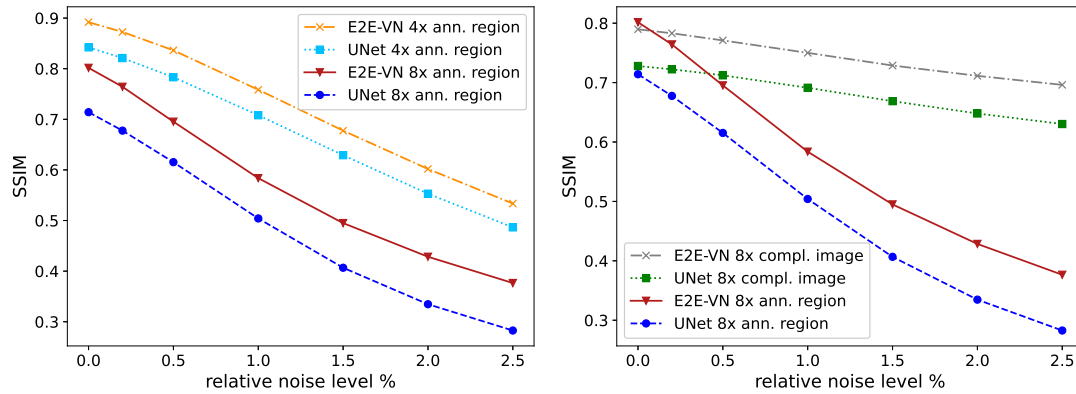


Figure 3.2: **Quantitative analysis of the adversarial attack.** Left: SSIM decline vs. noise level for both methods and acceleration factors. Right: Comparison of SSIM decline when attacking only the annotated region versus the complete image.

to alter the diagnostic interpretation of the reconstructed image significantly. As a consequence, the susceptibility of DL-based reconstruction methods to this worst-case noise needs to be taken into account and, if possible, mitigated before applying these methods in clinical practice. Furthermore, this work establishes the need to consider the clinical consequences of reconstruction errors in experimental evaluations, rather

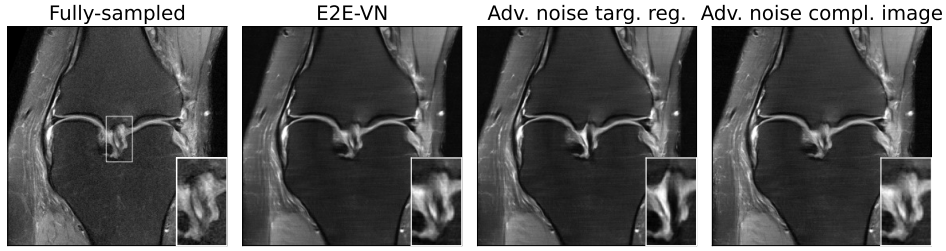


Figure 3.3: **Impact of targeted versus whole-image adversarial attack.** Reconstructions using the E2E-VN at $8\times$ acceleration. Third column shows the effect of restricting the attack to the targeted region, while the fourth column shows the result when attacking the complete image. Notably, attacking only the annotated region changes the ligament and it appears more homogeneous. This can lead to a misdiagnosis of the underlying ACL rupture.

than solely considering the average global similarity metrics.

3.6 Recent Advances and Impact

Recent work on the adversarial robustness of MRI reconstruction methods indicates that this remains an open problem [8]. Experiments by Krainovic et al. [72] indicate that jittering – the process of adding isotropic Gaussian noise to the data – improves robustness of reconstruction methods. However, only U-Net-based reconstruction (see Section 2.3.1) has been analyzed. This confirms insights by Genzel et al. [24], who recommended the same approach earlier. Even though both works [72, 24] argue that training on real data containing real measurement noise should improve robustness, our work shows that, even when trained on real-world data, reconstruction failures can occur with only small amounts of noise.

Most of the work on adversarial attacks in medical imaging still focuses on classification and segmentation tasks, with only a small minority analyzing the MRI reconstruction problem [73, 74] (our approach is the only one listed for MRI reconstruction in both papers). In parallel with our paper publication, Alaifari et al. [75] also found that localized reconstruction artifacts due to adversarial noise are a more relevant failure mode than an image-wide decline in similarity metrics, confirming our findings. Recently, Gottschling et al. [76] provided theoretical foundations for the empirical evidence that DL-based reconstruction methods are generally unstable against adversarial noise and can hallucinate non-existent structures. The authors note that methods claiming reconstruction robustness are in stark contrast to research such as ours, which demonstrates that it is easy to fool modern reconstruction methods, and conclude that robustness in image reconstruction and avoidance of hallucinations remains an open problem.

Chapter 4

Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data

Source Publication

Preprint:

Jan Nikolas Morshuis, Matthias Hein, and Christian F. Baumgartner. “Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data”. In: *IEEE Transactions on Biomedical Engineering* (2026). DOI: 10.1109/TBME.2026.3679957

4.1 Introduction

In the previous chapter, we saw a shortcoming of DL-based reconstruction methods for the downstream task of pathology detection: Only small perturbations to the measurement data can cause the models to hallucinate non-existent pathologies, as these often require only minor changes to the reconstructed MRI image. This behavior also limits the achievable MRI acceleration, which still allows robust reconstructions suitable for pathology detection. However, these restrictions on the acceleration can potentially be relaxed for the downstream task of segmenting larger anatomical structures, as these are easier to reconstruct and would require larger changes in the reconstructed image to be perturbed.

This chapter, therefore, shifts the focus towards the downstream task of semantic segmentation and the achievable acceleration limits while maintaining high-quality segmentations. To address this, the work presented in this chapter provides a systematic investigation into several key, and to our knowledge, previously unanswered questions: Are high-quality reconstructions necessary for feasible segmentations? Is

enforcing data consistency for the reconstruction beneficial? And should we combine reconstruction and segmentation into a one-stage end-to-end task, or use a two-stage approach and treat both tasks independently?

The motivation for this study stems from the K2S-challenge [6], for which we have developed a method specifically designed to segment undersampled MRI data. Even though our method won the challenge (1st of 12 teams), the generated insights are limited, as is often the case for biomedical image analysis challenges [77, 78]. The reason is that the individual solutions have too many factors of variation to be comparable, including architecture, preprocessing, sampling strategy, and computational resources. The questions from the previous paragraph, therefore, remained unanswered by the challenge.

In the paper corresponding to this chapter, we systematically analyzed current SOTA methods for the segmentation of undersampled MRI data to identify which concepts actually improve segmentation scores. We include the three top-performing methods of the K2S-challenge [79, 80, 81] and also include two SOTA end-to-end methods from prior work [25, 26]. Additionally, we analyze additional two-stage methods that treat the reconstruction and segmentation tasks independently, yet lead to surprisingly good reconstruction and segmentation scores. To provide optimal comparability of the methods, we fix the preprocessing, sampling strategy, training time, and model size of the segmentation models as much as possible. We analyze behavior across different acceleration factors and also include, to the best of our knowledge, the only publicly available dataset SKM-TEA [39] that combines MRI k-space data and segmentations in the analysis.

Our findings provide insights that surprisingly lead to different recommendations for future work than those of the K2S-challenge. While the results and discussion of the challenge’s summary paper suggest looking further into end-to-end approaches that combine reconstruction and segmentation tasks [6], we find that treating the reconstruction and segmentation tasks independently can lead to better reconstructions and higher segmentation scores while also being less hyperparameter-sensitive. We also found that reconstructions that enforced data consistency with the available k-space measurements generally lead to the best Dice scores.

4.2 Related Work

Before the K2S-challenge [6] mentioned above, other end-to-end frameworks for the segmentation of accelerated MRI data were developed. Sui et al. [25] combined a reconstruction and a segmentation U-Net in a cascade, combining the loss of reconstruction and segmentation as training loss. The authors observed that combining the reconstruction and segmentation tasks increased both reconstruction and some segmentation scores. However, no increase in Dice score has been observed in some datasets compared to a single segmentation U-Net, leaving the results inconclusive.

Calivá et al. [26] trained a shared encoder alongside two decoders, one for the reconstruction task and one for the segmentation task. The authors also report higher segmentation scores for their method compared with implemented baselines, yet the reconstruction scores reportedly did not improve compared to reconstructions from a specialized network. The authors did not evaluate the segmentation network trained on these improved reconstructions, leaving open the question of whether a simple two-stage method (first reconstruction, then segmentation) might yield better results than the specialized method. This is one of the questions that we aim to answer in this chapter.

4.3 Method Summary

Datasets

We conduct the experiments in this section on two datasets that include both multicoil k-space measurements and ground-truth segmentations. The first one is the K2S-data [6], which was released during the K2S-challenge in 2022, but is currently unavailable online. It consists of 300 3D-isotropic MRI scans of the knee. We split the data into 240, 30, and 29 scans for the train-, val-, and test-splits, respectively. One scan was not spatially aligned with its segmentation mask, so we excluded the scan from evaluation. The dataset contains segmentation annotations for six classes: the three bones of the knee (Patella, Femur, and Tibia) and their corresponding three cartilages. Note that this dataset requires 3-dimensional Fourier Transforms, which are computationally expensive, making it infeasible to compute them during a neural network’s forward pass, a requirement for some unrolled architectures (see Section 2.3.2).

The second dataset, SKM-TEA [39], is – to the best of our knowledge – the only publicly available dataset that includes k-space measurement data and annotations for the segmentation classes, namely patella-, femoral-, tibial cartilage, and meniscus of the knee. This dataset consists of 155 high-resolution MRI scans, with a train-, val-, and test-split of 86, 33, and 36 scans, respectively.

Tested Methods

We analyzed several state-of-the-art methods explicitly developed for accelerated MRI segmentation and also included new combinations of independent reconstruction and segmentation methods in our evaluation. We distinguished between **one-stage** approaches, which combine the reconstruction and segmentation task in a single forward pass, and **two-stage** approaches, which treat the reconstruction and segmentation task independently and consecutively.

One-stage methods are methods that have been developed explicitly for the task of segmenting undersampled MRI and can predict a segmentation mask (and

optionally a reconstruction) in a single forward pass. The primary motivation for many one-stage methods is that both reconstruction and segmentation tasks can benefit from training signals obtained by the other one [25, 26, 6]. We include the following one-stage methods in this research:

- **RecSeg** [25]: This method utilizes two U-Nets in cascade. The first U-Net specializes in reconstruction, while the second U-Net focuses on segmentation.
- **RegSeg** [6, 79]: This method is our submission to the K2S-challenge that has won the challenge and has set the new state-of-the-art for accelerated MRI segmentation on the K2S dataset. It also consists of two U-Nets in cascade, but the training scheme differs from RecSeg: both U-Nets are initially trained independently on either reconstruction or segmentation. After this pretraining, the U-Nets are put into a cascade and finetuned exclusively on the segmentation task, which has been the success metric in the challenge.
- **TwoDec** [26]: The TwoDec consists of a single encoder and two separate decoders, where one decoder is responsible for the reconstruction and one decoder for the segmentation of the accelerated MRI data.

Two-stage methods treat the reconstruction and segmentation tasks independently, first generating a reconstruction using existing reconstruction methods before predicting the segmentation based on the reconstruction with a standard U-Net. We include the following two-stage methods in our research:

- **NaïveSeg** [81]: This method trains a 3D segmentation U-Net on top of zero-filled MRI reconstruction. It is the most straightforward method, and we utilize it as a baseline. We also use the segmentation network’s architecture for all segmentations in the two-stage methods, training it directly on the reconstructions.
- **UNet-Rec** [37]: For this method, we utilize a 3D U-Net trained for MRI reconstruction (see Section 2.3.1).
- **CS-Rec** [34, 33, 80]: In this approach, first a Compressed Sensing reconstruction (CS-Rec) is applied to the undersampled image¹.
- **Unr-Rec** [82]: We trained the unrolled network architecture provided by the SKM-TEA repository [39] for the reconstruction task. The implementation is similar to other VarNets (see Section 2.3.2), as it uses multiple U-Nets in cascade, with data-consistency enhancing layers between them.

¹The code for the CS-Rec method was kindly provided by the team of Razumov and Dylov [80] that ranked second in the K2S-challenge.

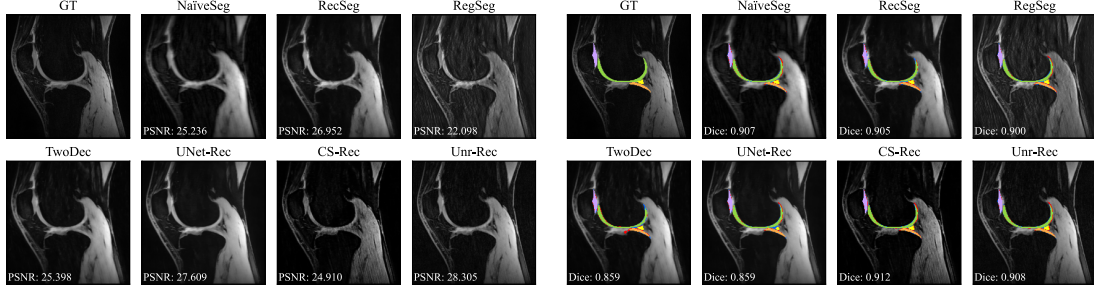


Figure 4.1: **Qualitative example from the SKM-TEA dataset.** The figure shows reconstructions (left) and segmentations (right) of the evaluated methods at $16\times$ acceleration. Red regions correspond to false positive predictions, blue regions to false negative predictions.

We re-implemented the methods to maximize the comparability of all U-Net-based approaches. To this end, we utilized the same nnU-Net-based encoder and decoder architectures, fixed the number of iterations, data augmentation, loss functions, and data preprocessing. As many methods use an ℓ_2 loss [37, 25, 79], we also used the ℓ_2 loss in our experiments.

Statistical Evaluation

To see which methods provide a statistically significant improvement in Dice score compared to the simple baseline NaïveSeg, we conducted a Wilcoxon-Signed-Rank test [83]. We marked the statistically significant improved Dice scores with an asterisk (*) in Table 4.1 for the K2S-data as well as Table 4.2 for SKM-TEA.

4.4 Key Results

We observed that most methods achieve Dice scores that are highly comparable across a wide range of accelerations. This is evident in the left plot of Figure 4.2 for the K2S dataset and in Table 4.2 for the SKM-TEA dataset. Surprisingly, only a few methods achieve statistically significant improvement compared to the baseline of NaïveSeg. For the SKM-TEA example in Table 4.2, we see that only Unr-Rec consistently provides a statistically significant improvement over NaïveSeg, even though the Dice score improvements are still minor (≤ 0.01).

In general, data-consistent two-stage methods yield the best Dice score. As shown in Figure 4.2 for the K2S data, CS-Rec achieves the highest Dice scores across most acceleration factors, while also having the lowest SSIM reconstruction scores (right plot). CS-Rec also obtains the second-highest scores in SKM-TEA, only surpassed

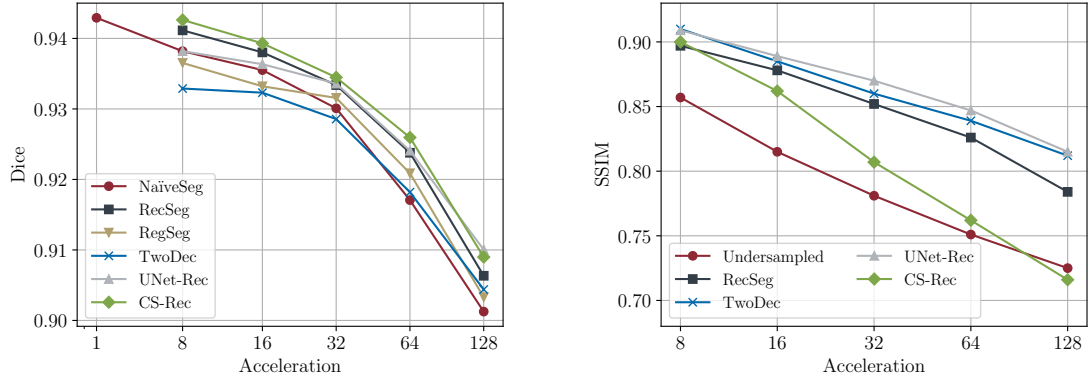


Figure 4.2: **Quantitative comparison on the K2S dataset.** Left: Dice score versus acceleration factor. Right: SSIM score versus acceleration factor. Note that CS-Rec achieves the highest Dice scores despite obtaining the lowest SSIM scores among the evaluated methods.

Table 4.1: **Dice scores (mean $\pm$ std.) for the K2S dataset.** The Dice score of NaïveSeg trained on fully sampled data is 0.943 ± 0.013 . Models with a statistically significant ($p < 0.01$) better Dice score compared to the NaïveSeg model at the corresponding acceleration factors are marked by an asterisk *.

	Dice					
	1 $\times$	8 $\times$	16 $\times$	32 $\times$	64 $\times$	128 $\times$
NaïveSeg	0.943 ± 0.013	0.938 ± 0.018	0.936 ± 0.015	0.930 ± 0.015	0.917 ± 0.020	0.901 ± 0.022
RecSeg	-	0.941 ± 0.012	0.938 ± 0.013	$0.933 \pm 0.013^*$	$0.924 \pm 0.018^*$	$0.906 \pm 0.024^*$
RegSeg	-	0.937 ± 0.016	0.933 ± 0.020	0.932 ± 0.016	$0.921 \pm 0.020^*$	0.903 ± 0.022
TwoDec	-	0.933 ± 0.021	0.932 ± 0.019	0.929 ± 0.018	0.918 ± 0.019	$0.904 \pm 0.024^*$
UNet-Rec	-	0.938 ± 0.014	0.936 ± 0.020	$0.934 \pm 0.015^*$	$0.924 \pm 0.016^*$	$0.910 \pm 0.019^*$
CS-Rec	-	$0.943 \pm 0.012^*$	$0.939 \pm 0.013^*$	$0.934 \pm 0.016^*$	$0.926 \pm 0.017^*$	$0.909 \pm 0.025^*$

by Unr-Rec, which also enforces data consistency implicitly.

Example results are shown in Figure 4.1. All methods provide reasonable segmentation predictions, only differing in minor details marked in red (false positives) or blue (false negatives). The original undersampled image is shown under NaïveSeg. Note that RegSeg is not optimized to generate high-quality reconstructions; instead, it is optimized solely for segmentation performance. The relevant cartilage regions, however, are still relatively sharp. The two-stage UNet-Rec method achieves the best reconstruction performance on both datasets.

4.5 Conclusion

We have shown that, contrary to previous findings, the best methods for segmenting accelerated MRI images are not specialized one-stage methods, but rather data-

Table 4.2: **Scores for the SKM-TEA dataset.** While Unr-Rec achieves the highest Dice scores in our experiments, UNet-Rec achieves the highest SSIM and PSNR scores.

	Dice			SSIM		PSNR	
	1×	8×	16×	8×	16×	8×	16×
NaïveSeg	0.854 ± 0.053	0.843 ± 0.044	0.837 ± 0.045	0.736 ± 0.024	0.692 ± 0.026	28.6 ± 1.1	27.6 ± 1.1
RecSeg	-	$0.845 \pm 0.047^*$	0.816 ± 0.049	0.799 ± 0.024	0.739 ± 0.029	30.9 ± 1.0	29.4 ± 1.0
RegSeg	-	$0.847 \pm 0.042^*$	0.838 ± 0.044	0.582 ± 0.036	0.509 ± 0.040	25.5 ± 1.4	24.8 ± 1.3
TwoDec	-	0.842 ± 0.043	0.811 ± 0.060	0.817 ± 0.021	0.737 ± 0.026	30.8 ± 1.1	28.1 ± 1.1
UNet-Rec	-	$0.845 \pm 0.044^*$	0.839 ± 0.045	0.826 ± 0.024	0.750 ± 0.042	31.6 ± 1.0	29.7 ± 1.0
CS-Rec	-	$0.847 \pm 0.044^*$	0.839 ± 0.046	0.680 ± 0.028	0.632 ± 0.030	29.0 ± 1.2	27.9 ± 1.2
Unr-Rec	-	$0.848 \pm 0.045^*$	$0.843 \pm 0.041^*$	0.704 ± 0.035	0.645 ± 0.036	31.2 ± 1.1	29.5 ± 1.1

consistent two-stage methods such as CS-Rec or Unr.-Rec. Not only are these models the only ones achieving consistent statistically significant improvements compared to NaïveSeg, but these methods are also easier to apply to new datasets: These two-stage methods require considerably less dataset-specific hyperparameter tuning during training, as the models do not need to find the right trade-off between reconstruction and segmentation loss. We have also shown that these simpler and data-consistent methods outperform our previously thought-to-be SOTA results in the K2S-challenge. More details can be found in the corresponding paper in Appendix B.

4.6 Recent Advances and Impact

The need to analyze reconstruction quality with respect to downstream tasks, such as semantic segmentation, has been noted in several review papers [8, 84]. Heckel et al. [8] pointed out that the best performing method of the K2S-challenge [6] generates reconstructions with the lowest similarity scores while still obtaining the highest segmentation scores.

We utilized insights from this research to develop high-scoring challenge contributions for two recent international MICCAI 2025 challenges: The Ultra-Low-Field MRI Image Enhancement Challenge (ULF-EnC) ² and the Low-Field Pediatric brain magnetic resonance Image Segmentation and quality Assurance Challenge 2025 (LISA 2025) ³. As observed in our work, UNet-Rec achieved the highest reconstruction similarity scores. We therefore applied it to the ultra-low-field MRI images of the ULF-EnC challenge, in which the aim was to predict the standard high-field MRI images. In this challenge, our method also led to high-quality predictions, placing 2nd among the 18 participants who submitted Docker containers at the final stage. For the LISA 2025 challenge, we utilized our observation that NaïveSeg

²<https://www.synapse.org/Synapse:syn65485242/wiki/631224>

³<https://www.synapse.org/Synapse:syn65670170/wiki/631438>

obtains segmentation scores comparable to the more complicated methods and submitted a conceptually similar method. Our submission ranked 1st out of 14 teams in task 2b of Basal Ganglia segmentation and 3rd out of 15 teams in task 2a of Hippocampus segmentation. These strong results provide external validation that the insights from this chapter’s research generalize to related tasks in low-field MRI, namely image enhancement and segmentation.

The high accelerations achievable while still obtaining segmentations with high average Dice scores also led to research analyzing the segmentation uncertainty due to the inherent reconstruction ambiguity [44, 2]. Fischer et al. [44] developed a method (PHiRec) that showed how uncertainties due to the MR reconstruction propagate to the downstream segmentation task. We [2] introduced a method to guide the reconstruction process of a diffusion model towards data-consistent reconstructions with larger or smaller segmentation volumes, yielding upper and lower segmentation boundaries that also provide intuitive segmentation uncertainties. We explain the method in more detail in the following chapter (Chapter 5).

Chapter 5

Segmentation Guided MRI Reconstruction for Meaningfully Diverse Reconstructions

Source Publication

Publication:

Jan Nikolas Morshuis, Matthias Hein, and Christian F Baumgartner. “Segmentation-guided MRI Reconstruction for Meaningfully Diverse Reconstructions”. In: *MICCAI Workshop on Deep Generative Models*. Springer. 2024, pp. 180–190. DOI: https://doi.org/10.1007/978-3-031-72744-3_18

5.1 Introduction

As seen in the previous chapter, high-quality segmentations are achievable even for high MRI acceleration factors, making the semantic segmentation downstream task an interesting application for accelerated MRI. However, analyzing only average segmentation scores across subjects does not do justice to the quality requirements of safety-critical applications such as medical imaging. Instead of just optimizing average scores, in practice, the method must work well for each patient individually, especially if the patient has outlier characteristics, such as smaller cartilage than the general population. As some outliers are typically rare, they are generally not well represented in the reported average scores, which is why analyzing only average scores is often insufficient to determine whether an algorithm is reliable enough for clinical practice.

To make matters worse, MRI reconstruction is an inverse problem as explained in Section 1.1.2, such that infinite possible and plausible solutions exist. It can, therefore, be insufficient for clinical use cases to provide only a single solution. Having a single solution does not allow the user to infer any information about the inherent reconstruction uncertainty. Since this uncertainty is intrinsic to inverse

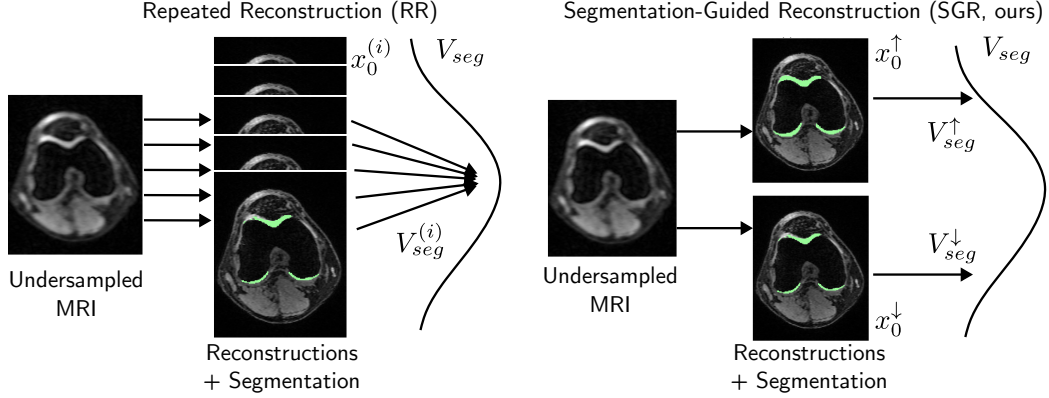


Figure 5.1: **Problem overview.** Repeatedly generating MRI reconstructions with a diffusion model tends to result in conceptually similar and not meaningfully diverse reconstructions. Similar results are also observed in downstream tasks such as segmentation in our experiments. Our method in contrast only reconstructs two images, an upper bound ($x_0^\uparrow$) and a lower bound ($x_0^\downarrow$) reconstruction, corresponding to an upper ($V_{seg}^\uparrow$) and lower bound ($V_{seg}^\downarrow$) on the segmentation volume, giving an intuitive understanding of the segmentation uncertainty.

problems such as MRI reconstruction, the user should be aware of it and its consequences for the intended clinical task, especially in safety-critical applications such as medical imaging.

From this perspective, providing only a single reconstruction and/or segmentation prediction, as done by the methods shown in Chapter 4, ignores the other feasible solutions. However, simply generating multiple reconstructions might also not be sufficient either; Many reconstructions will be close to the most-likely reconstruction [19] or to an averaged solution [17]. However, in medical imaging, we are especially interested in the tail cases, which are typically rare but still possible solutions to the reconstruction problem. In the context of segmentation, this could mean that we want to know whether the patient has reduced knee cartilage volume and therefore potentially suffers from osteoarthritis. Finding these rare abnormalities by repeatedly sampling from a posterior distribution could mean that an infeasible amount of samples is required in order to find, e.g., one of the smallest and one of the largest segmentations possible.

This paper, therefore, introduces a novel method that we term ”**Segmentation Guided MRI Reconstruction (SGR)**” to provide information about the tail-cases with regard to the segmentation volume as shown in Figure 5.1. SGR generates two data-consistent reconstructions per segmentation class, corresponding to the minimum and maximum feasible segmentations. The true segmentation will then most likely be between these extreme predictions. The predicted min- and max-segmentations, therefore, act as a boundary for the segmentation prediction.

Higher accelerations lead to larger boundaries, while smaller accelerations reduce the region in between the boundaries, which we call the "uncertainty region".

We showed that the predicted minimum and maximum segmentations yield better precision and recall than repeated sampling from the diffusion model, suggesting that the true segmentation lies within the predicted uncertainty region, indicating faithful uncertainty boundaries. Generating the two tail cases of reconstructions that lead to minimum and maximum segmentations also introduces the concept of meaningfully diverse reconstructions, which we also utilize in a later publication, summarized in Chapter 6, for pathology detection in accelerated MRI.

5.2 Related Work

The primary focus of this paper is the propagation of uncertainty to downstream tasks, such as segmentation, for which only a few specialized prior works exist. Feiner et al. [85] propagated the reconstruction uncertainty from medical image reconstruction to downstream regression and classification tasks by generating multiple reconstructions and calculating the pixel-wise variance [86], which is then utilized by the downstream model. Cohen et al. [19] showed that repeated sampling from the posterior of diffusion models for image reconstruction does generally not provide meaningful diversity in the reconstructions. Cheung et al. [87] have used repeated sampling of CT reconstructions to obtain distributions for several downstream metrics, which they use to implement conformal prediction methods. Fischer et al. [44] introduced a probabilistic reconstruction model (PHiRec), which they used to sample multiple reconstructions and, subsequently, multiple segmentations, allowing an estimate of segmentation uncertainty. However, no data consistency has been enforced in their method, and no upper or lower bound segmentation has been estimated.

5.3 Method Summary

Dataset

We evaluated our method on the SKM-TEA dataset [39], which, as mentioned earlier in Section 3.3, is to the best of our knowledge the only publicly available dataset containing both raw MRI k-space data and segmentations. We performed our evaluation on the official test split. To manage the computational costs, we select only slices from the relevant anatomical region (defined as slices containing at least one annotated pixel). To further reduce redundancy and inference time, we subsampled the volumes by choosing every 8th slice, resulting in an effective slice spacing of 2.5mm and a total of 869 test slices.

Method Description

The key idea of our method is to guide the reconstruction process of a diffusion model towards either larger ($S^\uparrow$) or smaller ($S^\downarrow$) segmentations. Diffusion models are particularly suitable for this task, as they use multiple steps: earlier steps reconstruct mainly the low-frequency components of the image, such as the general anatomical structure, while later steps focus on the details of the underlying MRI image. An example of this behavior is shown in Figure 2.3 in Chapter 2. To find $S^\uparrow$ and $S^\downarrow$, guidance of the reconstruction process is needed already in early steps, such that the reconstruction of anatomical information can be guided towards larger or smaller anatomical structures.

We start the summary of our method by defining the necessary concepts. We want to generate meaningfully diverse reconstructions $\mathbf{x}_0^{(i)}$ that differ in their segmentation volume. To achieve this, we reconstruct two images per segmentation class ($\mathbf{x}_0^\uparrow$ and $\mathbf{x}_0^\downarrow$) with corresponding upper and lower bound segmentations ($S^\uparrow$ and $S^\downarrow$) as well as segmentation volumes ($V_{seg}^\uparrow$ and $V_{seg}^\downarrow$). The volume between $S^\uparrow$ and $S^\downarrow$ is the uncertainty region S_{unc} . The true segmentation, which we define as the segmentation obtained on the fully sampled MRI image, will most likely be somewhere between $S^\uparrow$ and $S^\downarrow$.

Let $p_\phi^c(\mathbf{x}_{0,jk})$ be the predicted probability of pixel (j, k) in the image $\mathbf{x}_0$ belonging to the segmentation class $c \in \mathcal{C}$, where p_ϕ is the segmentation network. We abbreviate the term as $\mathbf{p}_{jk}^{(c)}$. The volume of the segmentation mask of class c can be calculated by summing the number of pixels that belong to c multiplied by voxel volume V_{voxel} :

$$V_c(\mathbf{x}_0) = \sum_{j,k} (\mathbb{I}_c(\mathbf{p}_{jk}^{(c)})) \cdot V_{voxel}, \quad (5.1)$$

where $\mathbb{I}_c$ is an indicator function

$$\mathbb{I}_c = \begin{cases} 1 & \text{if } \mathbf{p}_{jk}^{(c)} > \mathbf{p}_{jk}^{(c')} \ \forall c' \in \mathcal{C}, c' \neq c \\ 0 & \text{otherwise} \end{cases}. \quad (5.2)$$

During optimization, we consider only one class at a time, thereby simplifying the objective to a binary segmentation task. We build our guidance loss based on the widely used Binary-Cross-Entropy loss:

$$\mathcal{L}_{BCE}(\mathbf{y}, \mathbf{p}, c) = - \sum_{j,k} \left(\mathbf{y}_{jk} \log(\mathbf{p}_{jk}^{(c)}) + (1 - \mathbf{y}_{jk}) \log(1 - \mathbf{p}_{jk}^{(c)}) \right). \quad (5.3)$$

We can maximize the optimization objective $\mathcal{L}^\uparrow$ to find the upper bound segmentation, as well as $\mathcal{L}^\downarrow$ to find the lower bound segmentation:

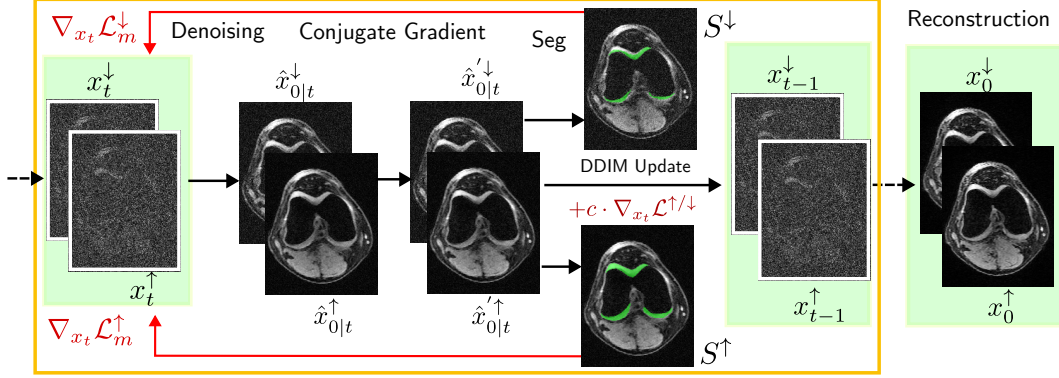


Figure 5.2: **Method explanation.** For each inverse diffusion step t , we calculate a loss $\mathcal{L}^{\uparrow/\downarrow}$ and include its gradient ($\nabla_{\mathbf{x}_t} \mathcal{L}^{\uparrow/\downarrow}$) in the calculation of $\mathbf{x}_{t-1}$. This guides the generation trajectory towards slightly increased or decreased segmentation volumes in subsequent steps. After the diffusion process, we obtain the upper- and lower-bound segmentation $S^\uparrow$ and $S^\downarrow$.

$$\mathcal{L}^\uparrow = \mathcal{L}_{BCE}(\mathbf{0}, \mathbf{p}, c) = - \sum_{j,k} \log(1 - \mathbf{p}_{jk}^{(c)}), \quad (5.4)$$

$$\mathcal{L}^\downarrow = \mathcal{L}_{BCE}(\mathbf{1}, \mathbf{p}, c) = - \sum_{j,k} \log(\mathbf{p}_{jk}^{(c)}). \quad (5.5)$$

The binary cross-entropy loss is unbounded, meaning that pixel probabilities are further optimized, even if the target class has already been reached. Following Croce et al. [88], we can avoid this behavior by using a masked cross-entropy loss, which excludes pixels already segmented as the target class from the optimization process. The optimization objectives can therefore be written as follows:

$$\mathcal{L}_m^\downarrow = \mathbb{I}_c \mathcal{L}^\downarrow; \quad \mathcal{L}_m^\uparrow = (1 - \mathbb{I}_c) \mathcal{L}^\uparrow. \quad (5.6)$$

The derived optimization objective can be used during the reverse diffusion process [51]. We therefore calculate the gradient of $\mathcal{L}_m^\uparrow$ (or $\mathcal{L}_m^\downarrow$, respectively) with respect to $\mathbf{x}_t$ at every step t , to find the perturbation necessary to increase the optimization loss. We add the gradient $\nabla_{\mathbf{x}_t} \mathcal{L}_m^\uparrow$ (or $\nabla_{\mathbf{x}_t} \mathcal{L}_m^\downarrow$, respectively) during the DDIM update that we have seen in Eq. 2.10 to arrive at Eq. 5.7:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}'_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \eta^2 \tilde{\beta}_t^2} (\boldsymbol{\varepsilon}_\theta^{(t)}(\mathbf{x}_t) + \gamma \nabla_{\mathbf{x}_t} \mathcal{L}_m^{\uparrow/\downarrow}) + \eta \tilde{\beta}_t \boldsymbol{\varepsilon}, \quad (5.7)$$

where the parameter γ ensures that the ℓ_2 norm of $\nabla_{\mathbf{x}_t} \mathcal{L}_m^{\uparrow/\downarrow}$ stays small with respect to $\boldsymbol{\varepsilon}_\theta$. We set γ to the following value:

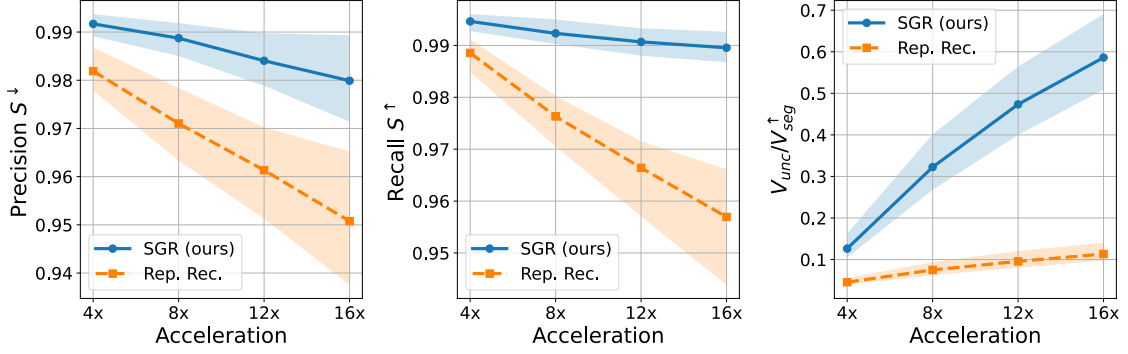


Figure 5.3: **Results.** Median precision and recall scores and percentile boundaries (25%, 75%) for different acceleration factors. The fraction $V_{unc}/V_{seg}^\uparrow$ in the right image indicates the ratio of the segmented volume for which uncertainty exists.

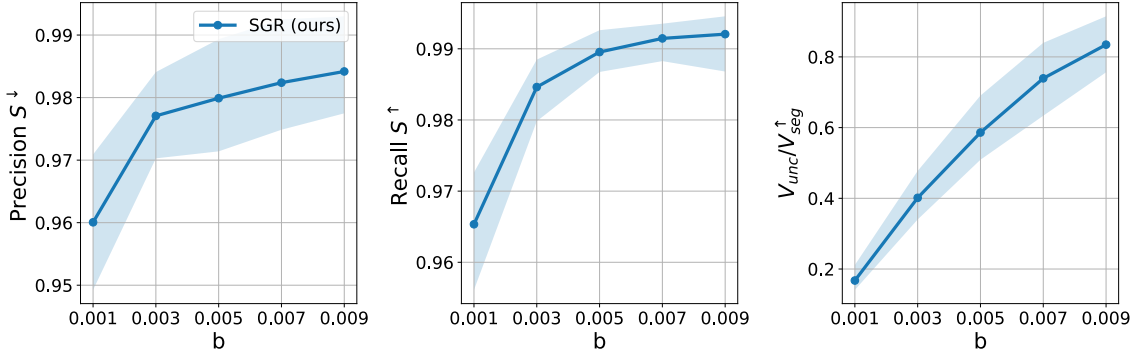


Figure 5.4: **Ablation.** Effects of the hyperparameter b in Equation 5.8. Note that the value of b largely influences the uncertainty volume V_{unc} , as well as the precision of $S^\downarrow$ and the recall of $S^\uparrow$. We set the acceleration to $16\times$ for the ablation experiments.

$$\gamma = \begin{cases} b \cdot \|\epsilon_\theta^{(t)}\|_2 / \|\nabla_{x_t} \mathcal{L}_m^{\uparrow/\downarrow}\|_2 & \text{if } \|\nabla_{x_t} \mathcal{L}_m^{\uparrow/\downarrow}\|_2 > b \|\epsilon_\theta^{(t)}\|_2 \\ 1 & \text{otherwise,} \end{cases} \quad (5.8)$$

where b is a small constant that we set to $b = 0.005$ in our experiments. An ablation study on the effects of the parameter value can be found in Figure 5.4. A graphical summary of our method is shown in Fig. 5.2.

For the baseline of Repeated Reconstruction (RR) with the diffusion model, we sample 16 reconstructions per slice using the same diffusion model [51] and select $\mathbf{x}_0^\uparrow$ and $\mathbf{x}_0^\downarrow$ for every segmentation class as the reconstruction leading to the highest segmentation volume ($V_{seg}^\uparrow$) and lowest segmentation volume ($V_{seg}^\downarrow$).

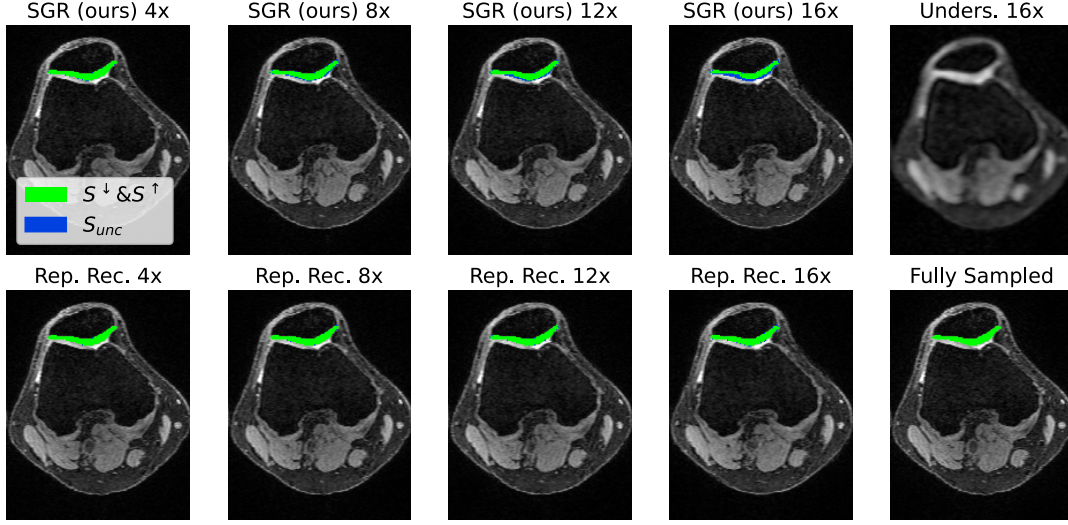


Figure 5.5: **Analysis of different acceleration factors.** For our baseline of Repeated Reconstructions, we observe that the uncertainty does not increase noticeably, even though high accelerations up to $16\times$ are tested. Our method, in contrast, generates more reliable uncertainty boundaries.

5.4 Key Results

We found that the min- and max-segmentation provided by our method better captures the true segmentation uncertainty due to reconstruction ambiguity than simple repeated reconstruction with the diffusion model [51]. This is shown in Figure 5.3, which shows the results for all tested acceleration factors ($4\times$, $8\times$, $12\times$, and $16\times$). Note that we compare 16 samples generated by the diffusion model with only two samples generated by our method. We obtain high precision scores for $S^\downarrow$, meaning that only a few false positive voxels are segmented, as well as the high recall scores for $S^\uparrow$, indicating that we do not miss many segmented voxels. The uncertainty volume V_{unc} compared to $V_{seg}^\uparrow$ is also growing for higher acceleration factors, as expected. The comparably high ratio likely results from small cartilage segmentation volumes, where a small absolute uncertainty represents a large relative fraction. Note that in all experiments, we compare against the predicted segmentation of the fully sampled image to disentangle the uncertainty arising from the reconstruction from other potential sources of uncertainty.

An example showing the union of $S^\downarrow$ and $S^\uparrow$ as well as S_{unc} can be found in Figure 5.5. We can see that the uncertainty region barely increases for the Repeated Reconstruction method at higher accelerations, thereby underestimating the true uncertainty as shown in Figure 5.3. Our method (SGR), in contrast, correctly displays larger uncertainty regions for higher accelerations.

5.5 Conclusion

We have introduced the novel method "Segmentation Guided MRI Reconstruction" for generating meaningfully diverse reconstructions in the context of the segmentation task. We have shown that two images reconstructed with our method better capture uncertainty than 16 images sampled from the posterior distribution of the diffusion model via repeated reconstruction.

Our method is one of the first to quantify the segmentation uncertainty due to the ill-posed nature of the MRI reconstruction problem. The original paper is provided in Appendix C. In Chapter 6, we will extend the concept of generating multiple semantically diverse reconstructions to the task of pathology detection in MRI.

5.6 Recent Advances and Impact

Analyzing the uncertainty for downstream tasks arising from reconstruction ambiguity has recently attracted growing interest in the research community [89, 90, 3]. Trautmann et al. [89] also found that measuring voxel-level uncertainty does not provide insights into morphological uncertainty. Instead, the authors propose to use segmentation as a downstream task and to evaluate an ensemble of reconstructions with respect to the segmentation volumes of relevant anatomical structures. Fischer et al. [90] also analyzed the reconstruction uncertainty with respect to a clinically relevant downstream task (patellar cartilage volume and cardiac ejection fraction), using their previously introduced PHiRec model [44] to obtain several reconstructions. The authors also use their task-based uncertainty estimate as a stopping criterion, allowing the MRI acquisition to stop when the uncertainty is sufficiently low. We [3] extended the idea of generating semantically diverse reconstructions, focusing on the semantic diversity with respect to pathology detection instead of segmentation. More details on our Semantically Diverse Reconstruction method for the pathology detection task can be found in the following chapter (Chapter 6).

Chapter 6

Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions

Source Publication

Publication:

Jan Nikolas Morshuis, Christian Schlarmann, Thomas Küstner, Christian F. Baumgartner[†], and Matthias Hein[†]. “Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions”. In: *MICCAI*. Cham: Springer, 2025, pp. 356–366. DOI: https://doi.org/10.1007/978-3-032-04937-7_34. [†]Shared last authorship.

6.1 Introduction

As seen in the previous chapter, it is important to predict the segmentation uncertainty due to the inherent reconstruction ambiguity in order to obtain trustworthy results. It is arguably even more important to account for reconstruction ambiguity for the task of pathology detection for several reasons:

- Some pathologies can be rare in the training set. Consequently, reconstruction models that sample from the posterior, such as diffusion models, will sample these pathologies less frequently if other solutions (e.g., healthy tissue) are also feasible given the data.

- Many pathologies are small and depend on the high-frequency information that is often measured less precisely compared to the low-frequency information in accelerated MRI. Slight variations in the predictions can lead to large changes in the reconstructed pathology.
- Fully-supervised regression-based reconstruction methods trained on paired undersampled and fully sampled data can lead to regression to the mean or MMSE reconstructions [17]. Diffusion models can be biased to the most likely solution [19]. Both methods can lead to incorrectly reconstructed pathologies.

For these reasons, a single prediction can be of limited use if significant inherent ambiguity exists in the reconstruction process. Even if the prediction is consistent with the measured k-space data, a better understanding of the solution space is essential to judge whether other data-consistent solutions are feasible, which might lead to a different diagnosis. Neglecting this additional information can lead to a false-negative diagnosis with potentially severe consequences for the patient.

To prevent omitting pathologies in the MRI reconstruction, the corresponding paper to this chapter introduced a novel method to mitigate these potential risks, which we term **”Semantically Diverse Reconstruction” (SDR)**: Instead of only generating a single data-consistent MRI reconstruction, our method additionally generates two semantically diverse reconstructions, that are also data-consistent and that therefore form possible solutions to the inverse reconstruction problem. These semantically diverse reconstructions should exhibit different pathologies compared to the first reconstruction, if feasible, given the measured data. To this end, we utilize a ViT-based detection network (ViTDet [62]). We alter the reconstruction by adding a small amount of noise. We then optimize the noise so that the distance between the ViTDet embeddings of the altered reconstructions increases and, consequently, the ViTDet detections are likely to change. Our method not only provides interpretable uncertainty estimates but also increases the number of pathologies detected by a different detection model used for validation only (Faster R-CNN [61]).

6.2 Related Work

Automatic pathology detection in MRI is a relatively underexplored research topic. Harman et al. [91] analyzed the detection of meniscus tears in MRI images in accelerated MRI, utilizing existing reconstruction methods such as U-Net (Section 2.3.1) and E2E-VarNet (Section 2.3.2) as well as a detection method nnDetection [67]. Chung and Ye [92] evaluated their proposed diffusion model-based MRI reconstruction for pathology detection, showing that their method found more pathologies than their baselines. Yen et al. [93] proposed a classifier that classifies selected underlying pathologies directly from the k-space data. However, they focus only on single-coil MRI data, which is no longer common in clinical practice.

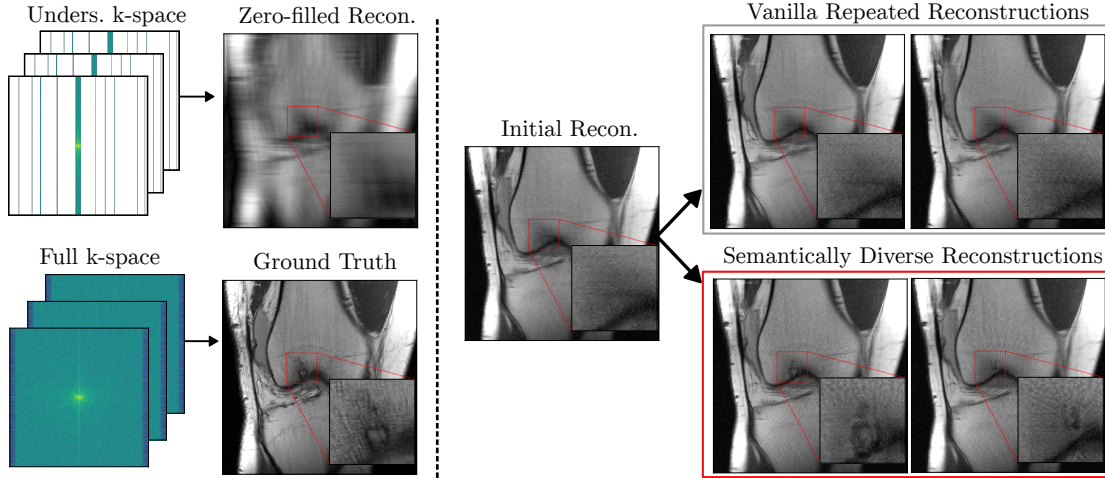


Figure 6.1: **Method overview.** Recent DL-based reconstruction methods, such as DDIP [14], can generate realistic-looking MRI images from undersampled MRI data. However, these images can lack crucial details, and pathologies might be omitted in the reconstruction. This problem can persist even when multiple solutions are sampled (top right). Our proposed SDR approach produces semantically diverse reconstructions, helping to detect possible pathologies that might otherwise be missed (bottom right). The example shows a subchondral edema pathology.

Küstner et al. [94] analyzed MRI reconstruction uncertainty using ensembles of trained reconstruction models and observed – for reconstructions that are only trained on healthy patients – that uncertainty is higher in regions containing pathologies compared to the healthy areas. It is therefore promising to analyze the effects of the reconstruction uncertainty on the clinical downstream task of pathology detection.

6.3 Method Summary

Dataset

We tested our method on the fastMRI [69] knee dataset. The corresponding pathological annotations were taken from the fastMRI-plus [70] extension, which provides labels for 21 different classes. From the 34,742 multi-coil slices in the fastMRI training dataset, only 8057 slices ($\approx 23\%$) contain one or more labeled pathologies. The majority of the data, therefore, shows only healthy tissue. For evaluation, we focused on the 10 most frequent pathology classes, as the remaining classes provided insufficient samples in the test set for a statistically meaningful evaluation. Furthermore, to evaluate our method’s ability to recover missing pathologies (avoiding false negatives and increasing recall), our test set consists of

slices containing at least one annotated pathology.

Because the fastMRI-plus dataset provides annotations only for the training and validation split of the fastMRI data, the official fastMRI test set could not be used for evaluation. To address this, we re-purposed the validation set. We split the validation set patient-wise, using 33% of the data for internal validation and 67% for the hold-out test set. We used our internal validation set to validate the detection models and the E2E-VarNet during training and to determine early-stopping criteria. We used our hold-out test set to obtain the qualitative and quantitative results shown in this chapter. From the 1005 slices with pathologies in our test set, we selected every second slice to reduce the computational cost of generating multiple reconstructions using diffusion models, resulting in 503 slices being evaluated in our experiments.

Method Description

Instead of generating many redundant semantically similar reconstructions, we wanted to generate a few semantically diverse reconstructions $\hat{\mathbf{x}}^{(i)}$, $i = 1, 2, \dots, N_{rec}$ that correspond to the more unlikely tail-cases of the solution space $\mathcal{X}$. This not only reduces the computational cost but also allows for a visual inspection by an expert to understand the inherent reconstruction ambiguity better. We have also seen in Chapter 5 that a few reconstructions from the tail of the posterior distribution can provide better uncertainty estimates than many reconstructions sampled IID from the posterior distribution. Furthermore, it is known that diffusion models have a bias towards the more likely solutions when using IID sampling [19].

Optimizing image generation to obtain a reconstruction from the tail of the distribution is more complex than in semantic segmentation, where we could optimize for easily measurable anatomical volume. As we have seen in Chapter 3, even small changes to the image can affect the clinical diagnosis, thereby leading to a large semantic distance. To better measure (and later increase) the semantic distance between two images, we proposed utilizing the embedding space of a ViT-based detection model (ViTDet) [62] that we trained on the pathology detection task.

To obtain semantically diverse reconstructions, we first obtain an initial reconstruction $\hat{\mathbf{x}}^{(1)}$ using existing methods, either based on diffusion models [14, 11] or the E2E-VarNet [13]. Note that even though the diffusion model by Chung and Ye [14] can generalize out of distribution, we use a model trained on in-distribution knee MRI data for better results. The goal is to find a new $\hat{\mathbf{x}}^{(i \neq 1)}$, such that $\|\hat{\mathbf{x}}^{(i)} - \hat{\mathbf{x}}^{(1)}\|_2 \leq r$ for a small r (we select $r = 3$) while at the same time the semantic distance $d(\hat{\mathbf{x}}^{(i)}, \hat{\mathbf{x}}^{(j)})$ is large, where we define the semantic distance as:

$$d(\hat{\mathbf{x}}^{(i)}, \hat{\mathbf{x}}^{(j)}) = \sum_{b \in \mathcal{B}} \|f_\phi(\hat{\mathbf{x}}^{(i)}, b) - f_\phi(\hat{\mathbf{x}}^{(j)}, b)\|_2. \quad (6.1)$$

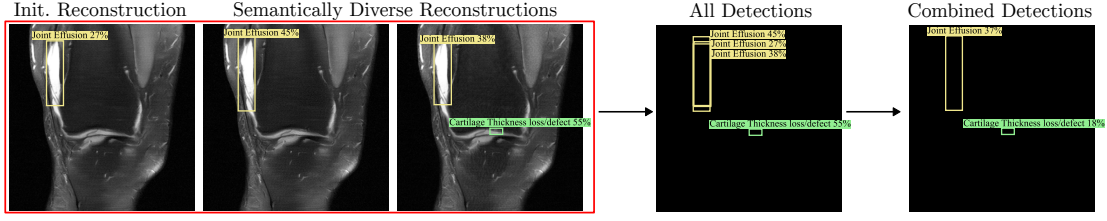


Figure 6.2: **Combining predictions of multiple reconstructions.** We combine overlapping detection boxes and average their probability score across reconstructed images to compute the mAP of our detections.

Here f_ϕ is the detection network and $\mathcal{B} = \{b_1, b_2, \dots\}$ are the bounding boxes that can either be defined manually (SDR-M) or automatically (SDR-A), as will be explained later in this section.

We can now maximize the semantic distance d of $\hat{\mathbf{x}}^{(i)}$ to all previous reconstructions $\hat{\mathbf{x}}^{(j)}$ in an iterative procedure, where $P_{B(\hat{\mathbf{x}}^{(1)}, r)}$ projects $\hat{\mathbf{x}}^{(i)}$ to an ℓ_2 ball of radius r around $\hat{\mathbf{x}}^{(1)}$, $\mathcal{DC}$ maps $\hat{\mathbf{x}}^{(i)}$ to the space of data-consistent solutions, and η is the step size:

$$\hat{\mathbf{x}}^{(i)} \leftarrow \mathcal{DC} \left(P_{B(\hat{\mathbf{x}}^{(1)}, r)} \left(\hat{\mathbf{x}}^{(i)} + \eta \nabla_{\hat{\mathbf{x}}^{(i)}} \sum_{j \neq i} d(\hat{\mathbf{x}}^{(i)}, \hat{\mathbf{x}}^{(j)}) \right) \right), \quad (6.2)$$

After a one-step optimization, we optimize the next reconstruction for one step, and so on, until 10 iterations are completed for each reconstruction.

Adversarial Robustness

We need to ensure that the ViTDet network f_ϕ , used to compute the semantic distance $d(\cdot, \cdot)$ in Equation (6.2), is robust to adversarial perturbations. This prevents the problem of small changes to the reconstruction, which are imperceptible to the human eye, leading to large distances in the embedding space. Furthermore, adversarial robust models possess semantically more meaningful gradients [95, 96, 97, 98, 99]. However, as many pathologies are rather small, training the model to be robust to these small perturbations risks unlearning the prediction of small pathologies. We solved this issue by utilizing a recent fine-tuning method by Schlarmann et al. [4] (who also implemented and ran the fine-tuning process) to make the ViT-backbone g_ϕ robust against adversarial perturbations on ImageNet data [100] using Equation (6.3). This way, the ViT-backbone learns to be robust against adversarial noise, without unlearning the detection of small pathologies.

$$\min_{\phi} \sum_i \max_{\|\delta\|_2 \leq r} \|g_\phi(\mathbf{x}_i + \delta) - g_{\phi_{\text{org}}}(\mathbf{x}_i)\|_2^2 \quad (6.3)$$

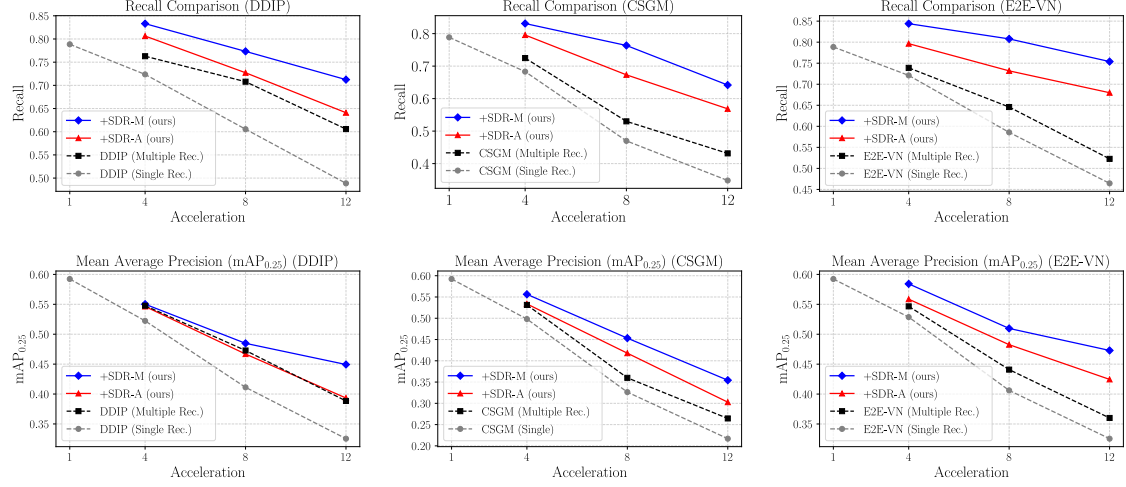


Figure 6.3: **Evaluation of reconstruction methods.** Both SDR-A and SDR-M outperform single and multiple reconstructions using the original reconstruction techniques (DDIP, CSGM, E2E-VN) in terms of recall (better detection of pathologies) while achieving similar or better mAP_{0.25}.

Box-Proposal Selection

We propose two methods for box-proposal selection: an automatic method, **SDR-A**, and a manual method, **SDR-M**. These methods cover the two use cases: either some information about the potential location of the pathology is known and bounding boxes can be set manually (SDR-M), or no additional information is assumed and the bounding boxes need to be found automatically (SDR-A). For automatic bounding box selection, we use the Region Proposal Network (RPN) implemented in the ViTDet detection model and described in more detail in Section 2.5.1. To reduce the number of proposal boxes, we apply stricter filtering compared to the standard RPN. We remove 25% of boxes with the lowest objectness score and additionally remove boxes with considerable overlap to other boxes ($IoU \geq 0.05$) that have a higher objectness score. We evaluate SDR-M by perturbing the bounding boxes slightly by changing their size and shifting their position, to better simulate manual annotations with only approximate information of potential pathology location.

For evaluation, we combined the predicted detections of the three generated images by averaging the probability scores of the bounding boxes throughout all three images, as shown in Figure 6.2. The probability scores of overlapping boxes are averaged if they belong to the same class. If a box does not occur in a reconstruction, we assign the probability score of 0. The evaluation model was a Faster R-CNN [61]. Using a different detection network than the ViTDet we used for optimization is necessary to ensure that changes in the image are not hallucinated by the ViTDet but are also observable by a different detection network.

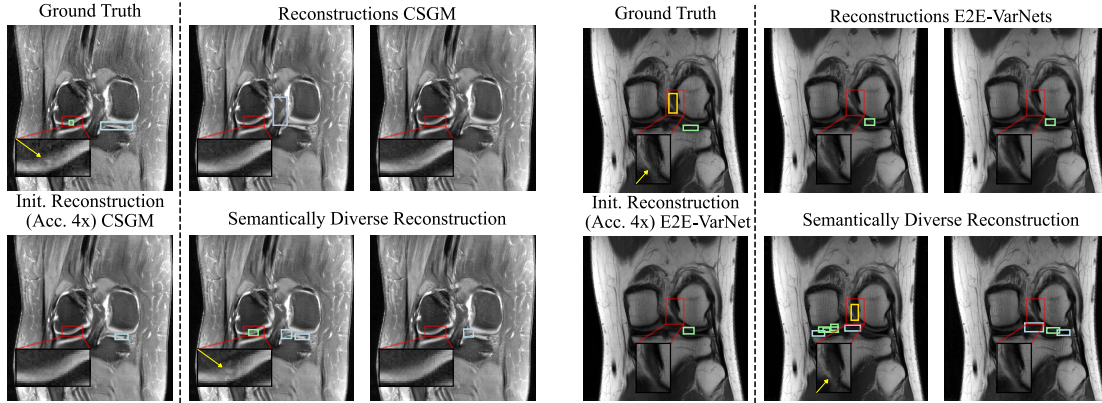


Figure 6.4: **Qualitative Results** The highlighted box in the ground truth shows a cartilage thickness loss (left) and a high grade ligament sprain (right). The semantic changes of SDR-M (left) and SDR-A (right) reveal the pathological changes. Yellow arrows indicate pathologies. Note that the pathologies can sometimes be hard to recognize for non-experts.

6.4 Key Results

Interestingly, our method not only provides interpretable uncertainty estimates for pathology detection but also increases detection scores in accelerated MRI on our test data. Figure 6.3 displays the quantitative results measured by the recall and the mean Average Precision (mAP) detection score. As mentioned in Section 2.5, a $\text{mAP}_{0.25}$ utilizes an IoU threshold of 0.25 to define positive detection. A higher recall indicates that the model finds more of the existing pathologies. A higher mAP score indicates better overall detection quality. Both our SDR-A and SDR-M models improve performance compared to generating multiple reconstructions for all tested reconstruction methods. We can see that simply generating multiple reconstructions and averaging the probability scores already yields better recall and mAP scores than evaluating just a single reconstruction for all tested reconstruction methods and acceleration factors. SDR-M obtains higher scores than SDR-A, as additional information about the approximate location of the pathology can be used. More qualitative examples are shown in Figure 6.4.

The importance of the robustness of the ViT-backbone of the ViTDet according to Eq. 6.3 is indicated by an ablation study. The results shown in Figure 6.5 indicate the higher recall obtained by our method, which has been fine-tuned on ImageNet data for increased robustness, compared to methods with no additional fine-tuning (left plot) or with fine-tuning on the fastMRI data (right plot).

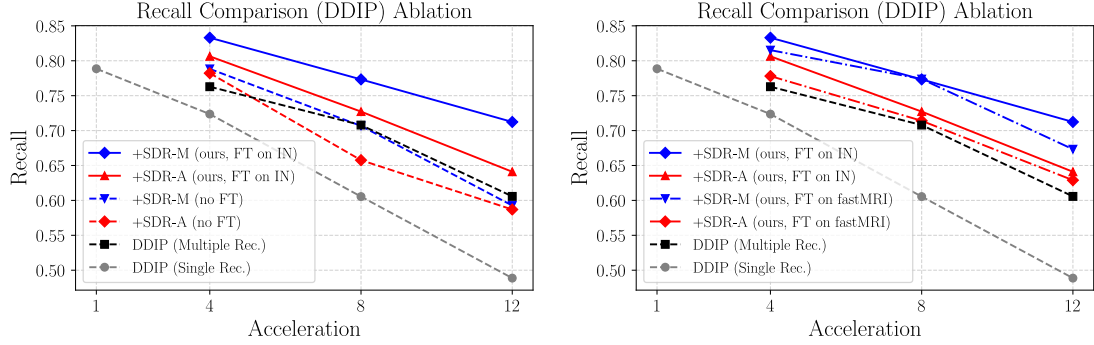


Figure 6.5: **Ablation study** for the finetuning of the ViT-backbone for increased adversarial robustness and increased recall. The left plot compares our method, which has been fine-tuned on ImageNet (IN) to no adversarial fine-tuning. The right plot compares our method to the same method but fine-tuned on fastMRI data. We observe higher recall scores for our method than for the tested alternatives across most acceleration factors.

6.5 Conclusion

We have shown that our method, SDR, can generate semantically diverse reconstructions, which lead to more pathologies being detected and higher detection scores on our test data. As the reconstruction problem is ill-posed and infinite solutions exist, we believe that a reasonable estimate of the uncertainty with respect to the target task will be necessary for many safety-critical future applications of highly accelerated MRI. Our method, to the best of our knowledge, is the first to provide an algorithm to mitigate reconstruction ambiguity specifically for the task of pathology detection and shows that this mitigation leads to improved detection results in our experiments as measured by the $\text{mAP}_{0.25}$ score.

6.6 Recent Advances and Impact

As this work was published only a few months before the writing of the thesis, not much closely related work has been published in the field since then. Notably, Fischer et al. [90] utilized segmentation uncertainty due to reconstruction ambiguity as a stopping criterion, such that the MRI acquisition can be stopped if sufficiently low uncertainty in a clinically relevant task is achieved. Similarly, future work can build upon our SDR method to find the precise acceleration, for which the MRI acquisition can be stopped as no further ambiguity remains in the reconstructed pathologies and a confident diagnosis can be reached.

Chapter 7

Discussion

7.1 Summary of Key Thesis Findings

In this thesis, we have shown that MRI reconstruction should be evaluated with respect to the targeted downstream tasks rather than solely using the currently standard similarity metrics, such as SSIM and PSNR. When evaluating MRI reconstruction methods for their utility for downstream tasks, it becomes apparent that some assumptions need to be shifted and that neglected factors, such as the inherent ambiguity of inverse MRI reconstruction problems, need careful consideration and mitigation if the algorithms are applied in clinical practice.

We have seen in Chapter 3 that only small perturbations in the MRI measurement data with "worst-case noise" can lead to clinically significant changes in the resulting reconstruction. In contrast, when targeting the similarity metrics for the entire image with the same noise budget, the resulting changes are challenging to detect. This observation indicates that the susceptibility of reconstruction models to worst-case noise, as well as their adversarial robustness, might need more consideration than previously thought.

Another interesting finding that can challenge common assumptions in MRI reconstruction research emerged from the analysis of methods for segmenting under-sampled MRI in Chapter 4. Despite the development of several specialized methods for segmenting accelerated MRI images, the best methods appear to be those that treat reconstruction and segmentation as independent problems. Surprisingly, combining classical CS reconstruction with a subsequent U-Net segmentation yields better results than specialized methods that combine reconstruction and segmentation in a single forward pass. This is the case even though the CS reconstruction achieves the lowest similarity scores among the tested methods, therefore calling into question whether similarity scores are a meaningful proxy for clinical utility. It also shows that the most effective methods do not need to be the most complex ones – for the segmentation of undersampled MRI images, the opposite appears to be true.

During our research on segmentation for accelerated MRI images, we have shown that high MRI accelerations are feasible while still obtaining high average Dice scores.

However, high accelerations lead to greater inherent reconstruction ambiguities due to the ill-posed nature of the problem. Even though average segmentation scores might be high, it is also essential to understand the reconstruction ambiguity and how it propagates to the segmentation task. Especially high accelerations require an estimate of the segmentation uncertainty to prevent overconfidence and to prevent missing the more unlikely cases, such as arthrosis, which typically corresponds to low cartilage volumes.

We have developed a novel method, 'Segmentation Guided Reconstruction' (Chapter 5), to generate two semantically diverse reconstructions that correspond to large or small segmentation volumes. Visualizing segmentation uncertainty through two semantically diverse reconstructions might be an intuitive way for a user to better understand the range of possible segmentations, which will need to be confirmed in future user studies. It also ensures that both reconstructions corresponding to large and small segmentation are indeed feasible, as they are consistent with the measured k-space data. Our method allows the clinician not only to see the most likely reconstruction together with the corresponding semantic segmentation, but instead provides an uncertainty boundary that includes the possibility of smaller or larger segmentation volumes. Being able to consider potential tail cases can be beneficial for any machine learning application in medical image analysis, as the detection of potentially unlikely abnormalities is often the primary reason for conducting MR image acquisition and subsequent analysis.

We have also applied the concept of generating multiple semantically diverse reconstructions to the task of pathology detection in Chapter 6. In contrast to the task of segmentation, for which anatomical structures are already visible early on in the reconstruction process of a diffusion model, many pathologies consist of fine details and high-frequency information that are reconstructed at the late stage of the diffusion process. We have therefore developed a novel method that leverages the embedding space of a detection model that we trained on the pathology detection task. Finding semantically diverse reconstructions then amounts to finding slight perturbations of an initial reconstruction that lead to a large ℓ_2 distance in the embedding space. We have shown that by generating only three semantically diverse reconstructions, we can increase the number of pathologies detected and also increase the general detection score ("mean Average Precision") in our experiments. This showcases the capability of our method not only to provide interpretable uncertainty estimates but also to detect additional pathologies that would otherwise be missed, and improve detection scores on the test data.

7.2 Connecting the Dots

One of the main findings of this thesis is that clinical utility is not accurately captured by commonly used metrics such as SSIM and PSNR. For example, our

study on the effects of worst-case noise (Chapter 3) has shown that a reduction in similarity score is not a good metric for evaluating the impact of worst-case noise. Instead, analyzing the clinical implications of the resulting reconstruction errors better captures the clinical risk posed by worst-case noise. Similarly, DL-based reconstructions achieving high SSIM and PSNR scores do not necessarily preserve clinically relevant information better than classical compressed sensing (Chapter 4): We have shown that segmentation networks trained on CS-reconstructions can obtain the highest Dice scores despite the reconstructions obtaining the worst similarity scores of the tested methods. The good segmentation results obtained with reconstructions from unrolled network architectures indicate that data consistency might be a more important property of the reconstruction than obtaining high similarity scores.

Focusing the evaluation not on average similarity scores but instead on clinical utility emphasizes the importance of considering the inherent ambiguity in MRI reconstruction when applying DL-based reconstruction methods. As discussed in Section 1.2.4, DL-based MRI reconstruction methods may tend to omit rare and small pathologies, especially when data-consistent solutions that display healthy tissues are feasible given the measurement data. While this problem cannot entirely be solved for high accelerations due to the inherent ambiguity of the ill-posed reconstruction problem, we have shown that it is possible to mitigate it by generating reconstructions corresponding to the tail cases of the posterior distribution, for both segmentation (Chapter 5), in which the predicted low-frequency image information needs to be varied, and pathology detection (Chapter 6), in which the predicted high-frequency image information needs to be varied.

7.3 Recent Shifts in MRI Reconstruction Literature

One of the most common methods to quantify uncertainty for the ill-posed problem of MRI reconstruction is the sampling of multiple reconstructions from a posterior distribution and calculating the corresponding variance, a process used to calculate the epistemic uncertainty [86]. There is no shortage of methods that provide this uncertainty quantification: Either through repeated sampling of diffusion models [11, 12], using MC-dropout [101], by utilizing intrinsic probabilistic models [44], or by an ensemble of multiple trained models [94]. Despite significant research effort in this area, it remains unclear whether pixel-wise variance maps provide meaningful insights for clinicians who need to base clinical decisions on the quality and uncertainty estimation of the reconstructions.

To visualize one example in which the pixel-wise variance of multiple reconstructions generated by a diffusion model does not provide the necessary information

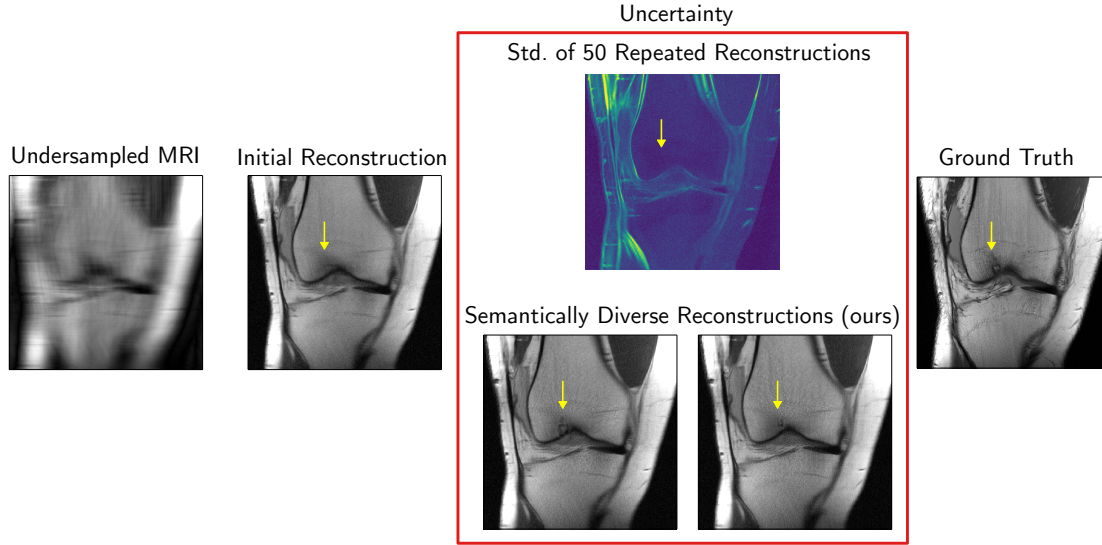


Figure 7.1: **Comparison of pixel-wise uncertainty versus semantic diversity.** The pixel-wise standard deviation map (top row), derived from 50 DDIP reconstructions, fails to indicate the location of the missed pathology. High variance is observed mainly at tissue boundaries rather than the location of the pathology. In contrast, the proposed SDR method (Chapter 6) generates semantically diverse samples (bottom row) that visualize different pathological structures.

about the reconstruction ambiguity, we generated such a standard deviation map from the example in Figure 6.1 of Chapter 6. We generated 50 repeated reconstructions of the DDIP diffusion model [14], and we calculated the standard deviation, following the methodology of [92, 11]. Surprisingly, as shown in Figure 7.1, no increased standard deviation is observed at the location where the pathology was not reconstructed correctly. Such uncertainty maps might therefore be of limited use if the reconstruction method cannot correctly reconstruct these rare pathologies. In contrast, our method “Semantically Diverse Reconstructions” (SDR from Chapter 6) shows semantically meaningful variations with just two additional reconstructions.

Recently, the first signs of a shift in the research landscape away from pixel-wise intensity variance as a reconstruction uncertainty estimation towards more task-based uncertainty are emerging. Fischer et al. [90] have analyzed the reconstruction uncertainty obtained by multiple reconstructions generated by the probabilistic model PHiRec [44] with respect to clinically relevant metrics such as patellar cartilage volume or cardiac ejection fraction, two tasks based on the predicted segmentation volumes.

In a recent study by Trautmann et al. [89], the authors utilized ensembles of common reconstruction models to analyze downstream segmentation-volume variation of various brain anatomies. The authors observed that this uncertainty estimation can

fail to reveal critical errors and biases that occur due to the reconstruction process. They also observed that neither voxel-level uncertainty metrics nor measures of variation in segmentation volume capture these effects accurately. The authors conclude that these uncertainty measures might be fundamentally inadequate when segmentations are the targets, confirming our observation from Chapter 5. Based on their observations, the authors argue that uncertainty in MRI reconstruction should be evaluated in the context of clinical tasks and not only in pixel space, also confirming our observations and conclusions throughout the thesis.

Insights from this thesis can help guide the MRI reconstruction research field toward methods that increase clinical utility and ensure safe application. Crucially, the inherent reconstruction ambiguity for the clinical task at hand must be correctly estimated and communicated to the user. Our contribution of providing semantically diverse reconstructions could serve as a suitable framework for estimating and interpreting reconstruction ambiguities and might inspire future work toward safer DL-based MRI reconstruction.

7.4 Limitations

This thesis has several limitations that need to be discussed. First, all experiments on accelerated MRI have been conducted with retrospective k-space masking. While this is standard practice in machine learning research for accelerated MRI, such ‘simulated’ data potentially differs from prospective undersampled measurements due to magnetization dynamics and underlying physics. Consequently, a method trained on retrospectively masked data might fail to generalize to real-world measured accelerated MRI data without sacrificing prediction quality.

Furthermore, all our methods have not been tested in clinical practice, and no clinical guideline exists on how to utilize the extra information from semantically diverse reconstructions. If a repeated scan is frequently necessary, it might reduce the benefits provided by the MRI acceleration process. However, knowing that a repeated scan might be necessary due to the inherent reconstruction ambiguity is already a valuable insight; Our method can help prevent false overconfidence in the most likely solutions, potentially making the application of DL-based MRI reconstruction methods safer and more trustworthy.

A further limitation of this work is due to the small number of publicly available datasets suitable for research on MR image reconstruction. Throughout the thesis, we have utilized many of the datasets that include both k-space data as well as some downstream diagnostic analysis, such as segmentation or pathology detection [8] (fastMRI [69, 70], SKM-TEA [39], K2S [6]). These datasets primarily focus on knee anatomy. Many abnormalities in these datasets, such as small cartilage volumes or multiple pathology classes, are challenging for non-experts to detect, making it hard to judge by visual inspection whether a reconstruction has been successful.

Another limitation concerns the experimental design in Chapter 6. Our main objective was to mitigate the risk of missing pathologies in accelerated MRI, thereby reducing false negatives and increasing recall. Consequently, our test data consisted only of slices containing at least one pathology. While our method demonstrates improved mAP and recall on this pathological dataset, we did not quantify the false positive rate on healthy samples. The method might therefore trade off higher sensitivity for a slight reduction in specificity, potentially leading to more false-positive detections.

Chapter 8

Conclusion and Future Work

This thesis challenges the prevailing paradigm in MRI reconstruction research, arguing that relying solely on average similarity metrics, such as SSIM and PSNR, is insufficient to evaluate the clinical utility of reconstruction methods. We have demonstrated that methods achieving high similarity scores can still omit subtle or rare pathologies and may yield inferior segmentation performance compared to classical reconstruction methods. This observation highlights the critical discrepancy between common model evaluation metrics and actual clinical needs. Consequently, this work argues that reconstruction algorithms should be evaluated and optimized based on their clinical utility rather than solely on their average similarity to the fully sampled reference images.

Through the lens of clinical utility, we show that a better understanding of the inherent reconstruction uncertainty is necessary to obtain more accurate segmentation uncertainties and more reliable pathology detections. Current MRI reconstruction research often neglects this ambiguity; algorithms typically generate only a single ‘most likely’ solution or an averaged solution, leaving the user unaware of alternative diagnostic possibilities and potentially leading to missed abnormalities. In contrast, our novel methods generate semantically diverse reconstructions that mitigate this risk. By generating plausible yet statistically less probable reconstructions (tail cases), we provide more reliable estimates of potential minimum and maximum segmentation volumes and recover pathologies that standard models might suppress.

These findings also suggest a necessary shift in evaluation priorities: instead of focusing on marginally increasing population-wide average scores, model development should concentrate on reliability for difficult and rare patient anatomies. While DL-based reconstruction methods offer unprecedented capabilities, their tendency to predict ‘most likely’ or averaged solutions poses a safety risk for these critical abnormal patient anatomies. Therefore, future MRI reconstruction development should focus not on ever higher acceleration factors or slightly increased similarity scores, but instead on the reliability of algorithms for the rarest and most complex patient anatomies and on providing an accurate and interpretable quantification of the inherent reconstruction uncertainty.

8.1 Future Work

Our work offers a new perspective on accelerated MRI reconstruction by prioritizing the utility of these reconstructions for downstream clinical tasks. To translate the concepts developed in this thesis into clinical practice, several critical steps remain.

First, it is necessary to validate our proposed methods for generating semantically diverse reconstructions on prospectively undersampled MRI data. This would allow for an analysis of whether the methods are robust to changes in noise levels and scanner physics that are not fully captured in the retrospective simulations we use. Additionally, acquiring more datasets that indicate rare anatomical abnormalities and pathologies would allow for a more rigorous analysis of our method’s ability to capture these critical tail cases.

Second, future work should address integrating the introduced methods into clinical practice. While generating semantically diverse reconstructions helps to prevent overconfidence, it does not always allow for a confident diagnosis if the inherent ambiguity of the reconstruction process is too high. Future clinical studies are required to investigate how clinicians utilize this additional information and whether presenting semantically diverse reconstructions can improve confidence and trust in the reconstruction algorithms. Crucially, these studies also need to determine whether occasional false positive pathology reconstructions are acceptable in this context, provided that these potential pathologies are feasible given the measured data.

Finally, future work could apply our methods in an active acquisition setting. By estimating task-specific uncertainty during MRI acquisition, it is possible to continue acquiring k-space data until the diagnostic ambiguity drops below a sufficiently small threshold, so that no image with an alternative clinical interpretation can be reconstructed and the semantic content of the image becomes unambiguous. Furthermore, by analyzing how these semantically diverse reconstructions differ, it may be possible to acquire the k-space regions that yield the most information gain for resolving the existing ambiguities. Ultimately, gaining a precise understanding of the inherent ambiguity of the reconstruction problem, along with the capabilities and limitations of DL-based reconstruction methods, is paramount for the robust and trustworthy application of these techniques in clinical practice.

Bibliography

- [1] Jan Nikolas Morshuis, Sergios Gatidis, Matthias Hein, and Christian F Baumgartner. “Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations”. In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2022, pp. 24–33. DOI: https://doi.org/10.1007/978-3-031-17247-2_3.
- [2] Jan Nikolas Morshuis, Matthias Hein, and Christian F Baumgartner. “Segmentation-guided MRI Reconstruction for Meaningfully Diverse Reconstructions”. In: *MICCAI Workshop on Deep Generative Models*. Springer. 2024, pp. 180–190. DOI: https://doi.org/10.1007/978-3-031-72744-3_18.
- [3] Jan Nikolas Morshuis, Christian Schlarmann, Thomas Küstner, Christian F. Baumgartner[†], and Matthias Hein[†]. “Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions”. In: *MICCAI*. Cham: Springer, 2025, pp. 356–366. DOI: https://doi.org/10.1007/978-3-032-04937-7_34. [†]Shared last authorship.
- [4] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. “Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models”. In: *ICML* (2024).
- [5] Jan Nikolas Morshuis, Matthias Hein, and Christian F. Baumgartner. “Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data”. In: *IEEE Transactions on Biomedical Engineering* (2026). DOI: 10.1109/TBME.2026.3679957.
- [6] Aniket A. Tolpadi et al. “K2S Challenge: From Undersampled K-Space to Automatic Segmentation”. In: *Bioengineering* 10.2 (2023).
- [7] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. “fastMRI: An open dataset and benchmarks for accelerated MRI”. In: *arXiv preprint arXiv:1811.08839* (2018).
- [8] Reinhard Heckel, Mathews Jacob, Akshay Chaudhari, Or Perlman, and Efrat Shimron. “Deep learning for accelerated and robust MRI reconstruction”. In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 37.3 (July 2024), pp. 335–368.

- [9] H. Nyquist. “Certain Topics in Telegraph Transmission Theory”. In: *Transactions of the American Institute of Electrical Engineers* 47.2 (1928), pp. 617–644.
- [10] Martin Uecker, Frank Ong, Jonathan I Tamir, Dara Bahri, Patrick Virtue, Joseph Y Cheng, Tao Zhang, and Michael Lustig. “Berkeley advanced reconstruction toolbox”. In: *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 23. 2486. 2015, p. 9.
- [11] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jonathan I Tamir. “Robust Compressed Sensing MRI with Deep Generative Priors”. In: *NeurIPS* 34 (2021), pp. 14938–14954.
- [12] Hyungjin Chung and Jong Chul Ye. “Score-based diffusion models for accelerated MRI”. In: *Medical Image Analysis* (2022), p. 102479.
- [13] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. “End-to-End Variational Networks for Accelerated MRI Reconstruction”. In: *MICCAI*. Ed. by Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz. Cham: Springer International Publishing, 2020, pp. 64–73.
- [14] Hyungjin Chung and Jong Chul Ye. “Deep Diffusion Image Prior for Efficient OOD Adaptation in 3D Inverse Problems”. In: *Computer Vision – ECCV 2024*. Ed. by Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol. Cham: Springer Nature Switzerland, 2025, pp. 432–455.
- [15] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.
- [16] Rafael Gonzalez and Richard Woods. *Digital Image Processing Global Edition*. Pearson Deutschland, 2017, p. 1024.
- [17] Hyungjin Chung and Jong Chul Ye. “Diffusion Models for Inverse Problems in Medical Imaging”. In: *Generative Machine Learning Models in Medical Image Computing*. Cham: Springer Nature Switzerland, 2025, pp. 129–148.
- [18] Yochai Blau and Tomer Michaeli. “The perception-distortion tradeoff”. In: *CVPR*. 2018, pp. 6228–6237.
- [19] Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. “From Posterior Sampling to Meaningful Diversity in Image Restoration”. In: *ICLR*. 2024.
- [20] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing properties of neural networks”. In: *ICLR* (2014).

-
- [21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *ICLR* (2015).
 - [22] 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”. In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30088–30095.
 - [23] Mohammad Zalbagi Darestani, Akshay S. Chaudhari, and Reinhard Heckel. “Measuring Robustness in Deep Learning Based Compressive Sensing”. In: *ICML*. 2021.
 - [24] Martin Genzel, Jan Macdonald, and Maximilian März. “Solving inverse problems with deep neural networks—robustness included?”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2022), pp. 1119–1134.
 - [25] Bin Sui, Jun Lv, Xiangrong Tong, Yan Li, and Chengyan Wang. “Simultaneous image reconstruction and lesion segmentation in accelerated MRI using multitasking learning”. In: *Medical Physics* 48.11 (2021), pp. 7189–7198.
 - [26] Francesco Calivá, Andrew P. Leynes, Rutwik Shah, Upasana Upadhyay Bharadwaj, Sharmila Majumdar, Peder E. Z. Larson, and Valentina Pedoia. “Breaking Speed Limits with Simultaneous Ultra-Fast MRI Reconstruction and Tissue Segmentation”. In: *Medical Imaging with Deep Learning*. Proceedings of Machine Learning Research. PMLR, July 2020.
 - [27] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. “Why rankings of biomedical image analysis competitions should be interpreted with care”. In: *Nature communications* 9.1 (2018), p. 5217.
 - [28] Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Matthias Eisenmann, Laura Aguilera Saiz, M Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. “Methods and open-source toolkit for analyzing and visualizing challenge results”. In: *Scientific reports* 11.1 (2021), p. 2369.
 - [29] 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”. In: *Magnetic resonance in medicine* 79.6 (2018), pp. 3055–3071.
 - [30] Daniel K Sodickson and Warren J Manning. “Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays”. In: *Magnetic resonance in medicine* 38.4 (1997), pp. 591–603.

- [31] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. “SENSE: sensitivity encoding for fast MRI”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pp. 952–962.
- [32] Martin Blaimer, Felix Breuer, Matthias Mueller, Robin M Heidemann, Mark A Griswold, and Peter M Jakob. “SMASH, SENSE, PILS, GRAPPA: how to choose the optimal method”. In: *Topics in Magnetic Resonance Imaging* 15.4 (2004), pp. 223–236.
- [33] Michael Lustig, David Donoho, and John M Pauly. “Sparse MRI: The application of compressed sensing for rapid MR imaging”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.
- [34] David L Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.
- [35] Leonid I Rudin, Stanley Osher, and Emad Fatemi. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *MICCAI*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241.
- [37] Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. “Deep learning for undersampled MRI reconstruction”. In: *Physics in Medicine & Biology* 63.13 (2018), p. 135007.
- [38] Matthew J. Muckley et al. “Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction”. In: *IEEE Transactions on Medical Imaging* 40.9 (2021), pp. 2306–2317.
- [39] Arjun D Desai, Andrew M Schmidt, Elka B Rubin, Christopher Michael Sandino, Marianne Susan Black, Valentina Mazzoli, Kathryn J Stevens, Robert Boutin, Christopher Re, Garry E Gold, et al. “SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation”. In: *NeurIPS Datasets and Benchmarks Track (Round 2)*. 2021.
- [40] Nicola Pezzotti, Sahar Yousefi, Mohamed S. Elmahdy, Jeroen Hendrikus Fransiscus Van Gemert, Christophe Schuelke, Mariya Doneva, Tim Nielsen, Sergey Kastruyulin, Boudewijn P. F. Lelieveldt, Matthias J. P. Van Osch, Elwin De Weerd, and Marius Staring. “An Adaptive Intelligence Algorithm for Undersampled Knee MRI Reconstruction”. In: *IEEE Access* 8 (2020), pp. 204825–204838.

-
- [41] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. “Loss Functions for Image Restoration With Neural Networks”. In: *IEEE Transactions on Computational Imaging* 3.1 (2017), pp. 47–57.
 - [42] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *ICLR*. 2014.
 - [43] Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu. “MR image reconstruction using deep density priors”. In: *IEEE Transactions on Medical Imaging* 38.7 (2018), pp. 1633–1642.
 - [44] Paul Fischer, K. Thomas, and Christian F. Baumgartner. “Uncertainty Estimation and Propagation in Accelerated MRI Reconstruction”. In: *MICCAI Workshop: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Cham: Springer Nature Switzerland, 2023, pp. 84–94.
 - [45] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets”. In: *NeurIPS* 27 (2014).
 - [46] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin. “DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction”. In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pp. 1310–1321.
 - [47] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. “Deep generative adversarial neural networks for compressive sensing MRI”. In: *IEEE Transactions on Medical Imaging* 38.1 (2018), pp. 167–179.
 - [48] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *NeurIPS* 33 (2020), pp. 6840–6851.
 - [49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *ICLR*. 2021.
 - [50] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. “Solving Inverse Problems in Medical Imaging with Score-Based Generative Models”. In: *ICLR*. 2022.
 - [51] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. “Decomposed Diffusion Sampler for Accelerating Large-Scale Inverse Problems”. In: *ICLR*. 2024.
 - [52] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising Diffusion Implicit Models”. In: *ICLR*. 2021.

- [53] Bradley Efron. “Tweedie’s formula and selection bias”. In: *Journal of the American Statistical Association* 106.496 (2011), pp. 1602–1614.
- [54] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. “Self-supervised pre-training of swin transformers for 3d medical image analysis”. In: *CVPR*. 2022, pp. 20730–20740.
- [55] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. “nnformer: Volumetric medical image segmentation via a 3d transformer”. In: *IEEE transactions on image processing* 32 (2023), pp. 4036–4045.
- [56] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. “MedNeXt: transformer-driven scaling of convnets for medical image segmentation”. In: *MICCAI*. Springer. 2023, pp. 405–415.
- [57] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. “nnU-Net revisited: A call for rigorous validation in 3d medical image segmentation”. In: *MICCAI*. Springer. 2024, pp. 488–498.
- [58] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pp. 203–211.
- [59] M. Jorge Cardoso et al. “MONAI: An open-source framework for deep learning in healthcare”. In: *arXiv preprint arXiv:2211.02701* (2022).
- [60] Lee R Dice. “Measures of the amount of ecologic association between species”. In: *Ecology* 26.3 (1945), pp. 297–302.
- [61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2016), pp. 1137–1149.
- [62] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. “Exploring plain vision transformer backbones for object detection”. In: *ECCV*. Springer. 2022, pp. 280–296.
- [63] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection”. In: *CVPR*. 2016, pp. 779–788.
- [64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *CVPR*. 2016, pp. 770–778.

-
- [65] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. “Feature pyramid networks for object detection”. In: *CVPR*. 2017, pp. 2117–2125.
- [66] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *ICLR*. 2021.
- [67] Michael Baumgartner, Paul F. Jäger, Fabian Isensee, and Klaus H. Maier-Hein. “nnDetection: A Self-configuring Method for Medical Object Detection”. In: *MICCAI*. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 530–539.
- [68] Paul F Jaeger, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. “Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection”. In: *Machine learning for health workshop*. PMLR. 2020, pp. 171–183.
- [69] Florian Knoll et al. “fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning”. In: *Radiology: Artificial Intelligence* 2.1 (2020). PMID: 32076662, e190007.
- [70] Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W. Lui, Michael S. Hansen, and Matthew P. Lungren. “fastMRI+, Clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data”. In: *Scientific Data* 9.1 (Apr. 2022), p. 152.
- [71] Zalan Fabian, Reinhard Heckel, and Mahdi Soltanolkotabi. “Data augmentation for deep learning based accelerated MRI reconstruction with limited data”. In: *ICML*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 3057–3067.
- [72] Anselm Krainovic, Mahdi Soltanolkotabi, and Reinhard Heckel. “Learning provably robust estimators for inverse problems via jittering”. In: *NeurIPS* 36 (2023), pp. 57276–57305.
- [73] Gladys W Muoka, Ding Yi, Chiagoziem C Ukwuoma, Albert Mutale, Chukwuebuka J Ejiyi, Asha Khamis Mzee, Emmanuel SA Gyarteng, Ali Alqahtani, and Mugahed A Al-antari. “A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense”. In: *Mathematics* 11.20 (2023), p. 4272.

- [74] Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. “Survey on adversarial attack and defense for medical image analysis: Methods and challenges”. In: *ACM Computing Surveys* 57.3 (2024), pp. 1–38.
- [75] Rima Alaifari, Giovanni S. Alberti, and Tandri Gauksson. “Short Communication: Localized Adversarial Artifacts for Compressed Sensing MRI”. In: *SIAM Journal on Imaging Sciences* 16.4 (2023), SC14–SC26.
- [76] Nina M. Gottschling, Vegard Antun, Anders C. Hansen, and Ben Adcock. “The Troublesome Kernel: On Hallucinations, No Free Lunches, and the Accuracy-Stability Tradeoff in Inverse Problems”. In: *SIAM Review* 67.1 (2025), pp. 73–104.
- [77] Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L. Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, Julio Saez-Rodriguez, Bram van Ginneken, Annette Kopp-Schneider, and Bennett A. Landman. “BIAS: Transparent reporting of biomedical image analysis challenges”. In: *Medical Image Analysis* 66 (2020), p. 101796.
- [78] Lena Maier-Hein et al. “Why rankings of biomedical image analysis competitions should be interpreted with care”. In: *Nature Communications* 9.1 (Dec. 2018), p. 5217.
- [79] Jan Nikolas Morshuis, Matthias Hein, and Christian F. Baumgartner. *RegSeg Method contribution to the K2S Challenge*. Described in the K2S Challenge overview paper [6]. 2023.
- [80] Artem Razumov and Dmitry Dylov. *CS-Rec Method contributed to the K2S Challenge*. Described in the K2S Challenge overview paper [6]. 2023.
- [81] Quintin van Lohuizen and Stefan J. Fransen. *NaïveSeg Contribution to the K2S Challenge*. Described in the K2S Challenge overview paper [6]. 2023.
- [82] Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. “Unrolled optimization with deep priors”. In: *arXiv preprint arXiv:1705.08041* (2017).
- [83] Frank Wilcoxon. “Individual comparisons by ranking methods”. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.
- [84] Mojtaba Safari, Zach Eidex, Chih-Wei Chang, Richard LJ Qiu, and Xiaofeng Yang. “Advancing MRI reconstruction: a systematic review of deep learning and compressed sensing integration”. In: *Biomedical Signal Processing and Control* 111 (2026), p. 108291.

-
- [85] Leonhard F. Feiner, Martin J. Menten, Kerstin Hammernik, Paul Hager, Wenqi Huang, Daniel Rueckert, Rickmer F. Braren, and Georgios Kaissis. “Propagation and Attribution of Uncertainty in Medical Imaging Pipelines”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Cham: Springer Nature Switzerland, 2023, pp. 1–11.
 - [86] Alex Kendall and Yarin Gal. “What uncertainties do we need in bayesian deep learning for computer vision?” In: *NeurIPS* 30 (2017).
 - [87] Matt Y Cheung, Tucker J Netherton, Laurence E Court, Ashok Veeraraghavan, and Guha Balakrishnan. “Metric-guided Image Reconstruction Bounds via Conformal Prediction”. In: *arXiv preprint arXiv:2404.15274* (2024).
 - [88] Francesco Croce, Naman D Singh, and Matthias Hein. “Towards Reliable Evaluation and Fast Training of Robust Semantic Segmentation Models”. In: *ECCV*. 2024, pp. 180–197.
 - [89] Luca LC Trautmann, Peter A Wijeratne, Itamar Ronen, and Ivor JA Simpson. “Evaluating Structural Uncertainty in Accelerated MRI: Are Voxelwise Measures Useful Surrogates?” In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2025, pp. 508–519.
 - [90] Paul Fischer, Jan Nikolas Morshuis, Thomas Küstner, and Christian Baumgartner. “CUTE-MRI: Conformalized Uncertainty-based framework for Time-adaptive MRI”. In: *arXiv preprint arXiv:2508.14952* (2025).
 - [91] Fatma Harman, Mustafa Alper Selver, Mustafa Mahmut Baris, Ali Canturk, and Ilkay Oksuz. “Deep Learning-Based Meniscus Tear Detection From Accelerated MRI”. In: *IEEE Access* 11 (2023), pp. 144349–144363.
 - [92] Hyungjin Chung and Jong Chul Ye. “Score-based diffusion models for accelerated MRI”. In: *Medical Image Analysis* (2022), p. 102479.
 - [93] Chen-Yu Yen, Raghav Singhal, Umang Sharma, Rajesh Ranganath, Sumit Chopra, and Lerrel Pinto. “Adaptive Sampling of k-Space in Magnetic Resonance for Rapid Pathology Prediction”. In: *ICML*. 2024.
 - [94] Thomas Küstner, Kerstin Hammernik, Daniel Rueckert, Tobias Hepp, and Sergios Gatidis. “Predictive uncertainty in deep learning-based MR image reconstruction using deep ensembles: Evaluation on the fastMRI data set”. In: *Magnetic Resonance in Medicine* 92.1 (2024), pp. 289–302.
 - [95] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. “Robustness May Be at Odds with Accuracy”. In: *ICLR*. 2019.
 - [96] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. “Image Synthesis with a Single (Robust) Classifier”. In: *NeurIPS*. Vol. 32. 2019.

- [97] Maximilian Augustin, Alexander Meinke, and Matthias Hein. “Adversarial Robustness on In- and Out-Distribution Improves Explainability”. In: *ECCV*. 2020.
- [98] Roy Ganz and Michael Elad. “Clipag: Towards generator-free text-to-image generation”. In: *WACV*. 2024.
- [99] Francesco Croce, Christian Schlarmann, Naman Deep Singh, and Matthias Hein. “Adversarially Robust CLIP Models Can Induce Better (Robust) Perceptual Metrics”. In: *IEEE SaTML*. 2025.
- [100] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *CVPR*. Ieee. 2009, pp. 248–255.
- [101] Jo Schlemper, Daniel C. Castro, Wenjia Bai, Chen Qin, Ozan Oktay, Jinming Duan, Anthony N. Price, Jo Hajnal, and Daniel Rueckert. “Bayesian Deep Learning for Accelerated MR Image Reconstruction”. In: *Machine Learning for Medical Image Reconstruction*. Ed. by Florian Knoll, Andreas Maier, and Daniel Rueckert. Cham: Springer International Publishing, 2018, pp. 64–71.

Appendix A

Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations

Source Publication

Publication:

Jan Nikolas Morshuis, Sergios Gatidis, Matthias Hein, and Christian F Baumgartner. “Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations”. In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2022, pp. 24–33. DOI: https://doi.org/10.1007/978-3-031-17247-2_3



Adversarial Robustness of MR Image Reconstruction Under Realistic Perturbations

Jan Nikolas Morshuis¹(✉), Sergios Gatidis², Matthias Hein¹,
and Christian F. Baumgartner¹

¹ Cluster of Excellence Machine Learning, University of Tübingen, Tübingen, Germany

nikolas.morshuis@uni-tuebingen.de

² Max-Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract. Deep Learning (DL) methods have shown promising results for solving ill-posed inverse problems such as MR image reconstruction from undersampled k -space data. However, these approaches currently have no guarantees for reconstruction quality and the reliability of such algorithms is only poorly understood. Adversarial attacks offer a valuable tool to understand possible failure modes and worst case performance of DL-based reconstruction algorithms. In this paper we describe adversarial attacks on multi-coil k -space measurements and evaluate them on the recently proposed E2E-VarNet and a simpler UNet-based model. In contrast to prior work, the attacks are targeted to specifically alter diagnostically relevant regions. Using two realistic attack models (adversarial k -space noise and adversarial rotations) we are able to show that current state-of-the-art DL-based reconstruction algorithms are indeed sensitive to such perturbations to a degree where relevant diagnostic information may be lost. Surprisingly, in our experiments the UNet and the more sophisticated E2E-VarNet were similarly sensitive to such attacks. Our findings add further to the evidence that caution must be exercised as DL-based methods move closer to clinical practice.

Keywords: MRI reconstruction · MR imaging · Adversarial attacks

1 Introduction

MR image acquisition is notoriously limited by long scan times adding to patient discomfort and healthcare costs. By acquiring only a subset of measurements in the k -space this process can be substantially accelerated. However, reconstructing images from such undersampled k -space data requires solving an ill-posed inverse problem. The problem was initially approached using techniques rooted in the theory of compressed sensing which offers provable error bounds (see e.g. [14]). In recent years, it has been shown that MRI reconstruction algorithms based on deep learning (DL) can substantially outperform classical reconstruction methods both in accuracy and speed [7, 11, 15–17, 21]. Such approaches,

however, currently lack provable robustness guarantees. Moreover, their failure cases are not sufficiently understood, putting into question the trustworthiness of such approaches in real-world scenarios.

Recently, there has been an increasing focus on better understanding the robustness properties of DL-based reconstruction algorithms. For instance, in the 2020 fastMRI challenge it was reported that top performing deep learning based methods can sometimes miss small abnormalities at $8\times$ acceleration [2, 11]. A priori it is unclear whether the information of these abnormalities is lost in the k -space subsampling step, which would make a correct reconstruction physically impossible, or whether the information is contained in the measurements but reconstruction algorithms fail to retrieve it. Cheng et al. [3] provided compelling evidence that in many cases the latter is the case by inserting adversarially generated false positives which fail to be reconstructed. This suggests that robustness properties of DL-based reconstruction algorithms deserve closer scrutiny. In another line of work, a large analysis of the 2019 fastMRI challenge submissions [8] found that the performance of DL-based reconstruction methods can suffer if a lower SNR is encountered during test-time than during training time. Darestani et al. [18] further showed a sensitivity to different acquisition techniques and different anatomies.

While it seems well documented that clinically relevant failure modes do exist, comparatively little work has been dedicated to understanding the worst case performance. Better understanding the maximum harm an algorithm can do is crucial for any clinical application and an integral part of the medical auditing process [9]. In absence of theoretical guarantees a number of works have attempted to provide empirical bounds on the worst case performance using adversarial attacks. While it is well established that DL-based classification networks are very sensitive to such attacks due to the “vulnerable” decision boundary stemming from the discrete nature of the problem, MR reconstruction algorithms do not necessarily suffer from the same problem [1, 6]. Firstly, MR reconstruction can be seen as a regression problem lacking discrete boundaries. Furthermore, many DL-based algorithms, such as the state-of-the-art E2E-VarNet (E2E-VN) [16], which we study in this paper, are inspired by classical algorithms with documented robustness properties.

Antun et al. [1] demonstrated that, indeed, small adversarial perturbations applied to the input of reconstruction algorithms may result in severe reconstruction artifacts. Similarly, Darestani et al. [18] investigated adding adversarial noise to the k -space measurements, observing that DL-based methods can be sensitive to the adversarial noise and that such attacks can also substantially degrade diagnostically relevant regions. Genzel et al. [6] also investigated adversarial attacks with bounded noise on the k -space measurements. Interestingly, the authors came to the conclusion that the investigated reconstruction networks are relatively stable to such attacks, though results varied with different undersampling patterns. In particular, the authors note that diagnostically relevant regions remain largely unaffected. We thus observe that the evidence for adversarial robustness of MR reconstruction algorithms remains inconclusive.

In this paper, we aim to shed further light on this question by analysing adversarial robustness on the commonly investigated fastMRI knee dataset [19, 20]. In contrast to prior work we show that when attacking state-of-the-art MRI reconstruction algorithms in a way targeting regions with diagnosed pathologies directly, very small levels of adversarial noise on the k -space (of comparable magnitude to the thermal noise always present in MR acquisition) can substantially degrade those regions and in some instances alter diagnostically relevant features. Going beyond the standard noise based adversarial attacks, we furthermore show that, somewhat worryingly, very small adversarial rotations can have similar effects. These findings add additional evidence that the worst case performance of DL-based MR reconstruction algorithms may currently be unacceptable for clinical use.

2 Methods

Problem Setting. In MR image acquisition typically multiple receiver coils are used to measure the data necessary to produce an image. Each coil i produces a k -space measurement $\mathbf{k}_i \in \mathbb{C}^n$. Given $\mathbf{k}_i$, a coil image $\mathbf{x}_i \in \mathbb{C}^n$ can be calculated using the inverse Fourier transform, such that

$$\mathbf{x}_i = \mathcal{F}^{-1}(\mathbf{k}_i + \mathbf{z}), \quad (1)$$

where $\mathbf{z} \in \mathcal{N}(0, \sigma \cdot \mathbf{I})$ is additive Gaussian thermal noise which is unavoidable in MR image acquisition [10]. In the fastMRI setting only a subset of the k -space data is measured. This can be simulated by applying a mask function $\mathcal{M}$ that sets the unmeasured part of the k -space data to 0, i.e. $\mathcal{M}(\mathbf{k}) = \mathbf{k} \odot M$, for a mask M and with $\odot$ the element-wise matrix multiplication. We employ the usual Cartesian undersampling pattern also used for the fastMRI challenge in all our experiments. Thus, a single undersampled coil-image can be expressed as

$$\tilde{\mathbf{x}}_i = \mathcal{F}^{-1}(\mathcal{M}(\mathbf{k}_i + \mathbf{z})). \quad (2)$$

The acquired coil-images $\mathbf{x}_i$ or $\tilde{\mathbf{x}}_i$ can be combined using the voxelwise root-sum-of-squares (RSS) method [12] to obtain the fully-sampled image $\mathbf{X}$ (or the undersampled image $\tilde{\mathbf{X}}$, respectively):

$$\mathbf{X} = \sqrt{\sum_{i=0}^N |\mathbf{x}_i|^2}, \quad (3)$$

where N indicates the number of coils.

MR reconstruction algorithms aim to reconstruct $\mathbf{X}$ given only the undersampled k -space data $\mathcal{M}(\mathbf{k}_i)$ for $i \in \{1, 2, \dots, N\}$. DL-based approaches either start by obtaining the image $\tilde{\mathbf{X}}$ created from undersampled measurements according to Eq. (3) and treat the reconstruction as a de-aliasing problem (e.g. [7, 15]), or use the k -space data as input directly (e.g. [16, 21]). Here, we define the reconstruction network f as a mapping from (partial) k -space to image space.

Investigated Reconstruction Methods. In this paper we investigated two different neural network architectures. Namely a method based on the image-to-image UNet approach proposed in [13] as well as the iterative network E2E-VN [16], which reported state-of-the-art performance on the fastMRI challenge data. Although pretrained weights for both these models exist in the fastMRI-repository [11], they were trained using both train and validation sets of the fastMRI data, and evaluated on a private challenge test set. Since the ground-truth data of the private test set is not publicly available, we conduct our analysis of the model robustness on the validation set. The models therefore had to be retrained using data from the training set only.

Both networks have been trained using a batch-size of 32, like in their original implementation [11]. To train with the original batch-size, the E2E-VN requires 32 GPUs of 32 GB memory each, which was not feasible with our local compute infrastructure. This necessitated adapting the network size such that it could fit into the 11 GB memory of our local GPUs. For the E2E-VN, we reduced the number of unrolled iterations from 12 to 8 and halved the number of channels in each UNet. The size of the UNet for estimating the sensitivity maps was also reduced from 8 top-level channels to 6. Due to the computational constraints, we also halved the number of channels in the baseline UNet approach. The final training of our E2E-VN adjustment was carried out with 32 GPUs in parallel and took around one day to train, our adjusted UNet used 8 GPUs in parallel and also took around one day to train. Other than the reduction of the number of parameters, our version of the two networks is identical to the originals used for the fastMRI challenge and we expect our insights into its robustness properties to carry over to their larger counterparts.

Adversarial Attacks on the Reconstruction Methods. When using adversarial attacks, we are interested in finding a worst-case perturbation according to some attack model and within given boundaries. In this paper we explore two types of attack: 1) Adversarial noise added to the k -space measurements, which is bounded by a maximum L_2 -norm, 2) Adversarial rotations where we simulate a slightly rotated position of the patient during acquisition. In the following both methods are described in more detail.

For the first attack model, we aim to find bounded adversarial noise-vectors $\mathbf{z}_i$ for every k -space measurement $\mathbf{k}_i$ of every coil i , such that the resulting output changes as much as possible in a user-defined target region. More specifically, we have the following optimization problem:

$$\max_{\mathbf{z}} \|S \odot [f(\mathcal{M}(\mathbf{k} + \mathbf{z})) - \mathbf{X}]\|_2 : \|\mathbf{z}_i\|_2 \leq \eta \|\mathbf{k}_i\|_2. \quad (4)$$

Here, $\mathbf{k}$ is the vector of all k -space data $\mathbf{k}_i$ from all coils i , η defines the L_2 -adversarial noise level relative to every coil measurement $\mathbf{k}_i$, and S is a binary region selection mask to limit the attack to diagnostically relevant regions. Specifically, we use the bounding boxes indicating pathologies provided by the fastMRI+ dataset [20] to define S . In contrast to [6, 18], we restrict the

noise-vector $\mathbf{z}_i$ for every coil individually instead of the complete measurement, thereby allowing for differences in the coil-sensitivities.

The proposed realistic rotation attacks can be expressed similarly as

$$\max_{\theta} \|S \odot [(\mathcal{R}_{\theta}^{-1} \circ f \circ \mathcal{M} \circ \mathcal{R}_{\theta})(\mathbf{k}) - \mathbf{X}]\|_2 : d \in [-\theta_{max}, \theta_{max}]. \quad (5)$$

Here, we introduce an operator $\mathcal{R}_{\theta}$ which denotes a rotation of the MRI data by θ . In practice, we rotate the data in the spatial domain and then Fourier-transform back to obtain $\mathbf{k}$ similar to [4]. After a prediction is created by the neural network, the inverse rotation $\mathcal{R}_{\theta}^{-1}$ is used to transform the image back. It is then compared to the original untransformed target-image.

The optimization process in Eq. (4) was implemented using 10-step projected gradient descent. In every step, the adversarial method adds a noise-vector $\mathbf{z}_i$ to the k -space $\mathbf{k}_i$ for every coil i , using this transformed k -space data as a new input for the reconstruction method f . After the calculation of the main objective in Eq. (4) and successive backpropagation, the resulting gradients are added to the noise-vector using a step-size of 0.5. The noise-vector $\mathbf{z}_i$ is then renormalized to stay within the pre-defined boundaries. The worst-case rotation angle θ in Eq. (5) was found using a grid search using an evenly spaced grid from $-\theta_{max}$ to θ_{max} with a step-size of 0.1° . Our code is publicly be available online¹.

3 Experiments and Results

We investigated adversarial noise and rotation attacks as described in Sect. 2 on the fastMRI knee dataset [19]. We explored attacking the entire center-cropped image, as well as targeted attacks on diagnostically relevant regions, which we obtained from the pathology annotations in the fastMRI+ dataset [20]. We evaluated the adversarial robustness for the E2E-VN [16] as well as the UNet [7] and for acceleration factors of $4\times$ and $8\times$.

To explore attacks of varying severity, we repeated all experiments for a range of parameters η and θ_{max} . Similar to [6] and [18], we set the upper bound of the noise vector η relative to the k -space norm $\|\mathbf{k}_i\|_2$. We explored a range of $\eta \in [0\%, 2.5\%]$. We empirically verified that the adversarial noise levels have a similar scale to thermal noise in our data, by measuring the real noise in the background regions of the input images as described in [5]. This makes it plausible that similar noise could also occur in practice. For the maximum rotation parameter θ_{max} we explored a range of $[0^\circ, 5^\circ]$.

Figure 1 shows the relative degradation of the structural similarity (SSIM) computed over the diagnostically relevant regions for the noise attack model. Overall E2E-VN yields consistently better results compared to the simpler UNet-based reconstruction method. Nevertheless, the slope of the degradation is similar for both techniques, suggesting that both methods are equally susceptible to the attack. These findings also hold for the rotation attacks shown in Fig. 4a.

¹ <https://github.com/NikolasMorshuis/AdvRec>

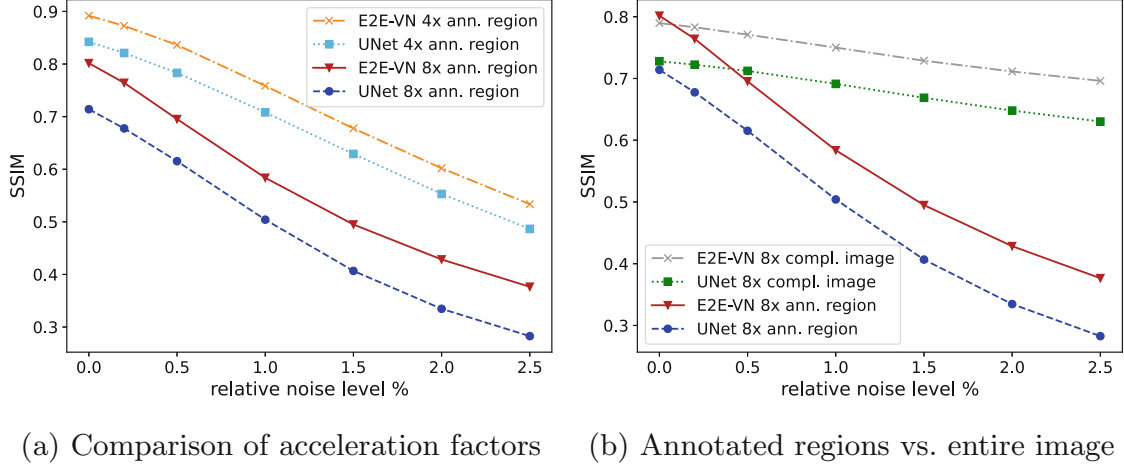


Fig. 1. SSIM between the full k -space reconstruction and DL-based reconstruction with adversarial noise. (a) shows the behaviour with respect to the acceleration factor, while (b) compares attacks on annotated, diagnostically relevant regions with attacks on the entire image.

The similar sensitivity to adversarial attacks is surprising given that the E2E-VN is implicitly an iterative approach with enforced data consistency.

We further analysed the results qualitatively in collaboration with a senior radiologist. In Fig. 2, reconstructions of the UNet, the E2E-VN as well as reconstructions of both methods with adversarial noise are compared for a sample image. While the fully sampled reconstruction shows a low grade sprain of the anterior cruciate ligament (ACL), the UNet reconstruction with adversarial noise shows a thickening and distention, which could mistakenly be interpreted as an acute injury. The E2E-VN reconstruction with adversarial noise on the other hand could lead to a false diagnosis of an old ACL tear. It can thus be concluded that even small amounts of adversarial noise can lead to a false diagnosis with potential clinical consequences.

Figure 1b compares attacks targeted on annotated regions with attacks on entire images (the latter being the standard approach e.g. [1, 6, 18]). While [6] reported that diagnostically relevant regions remain relatively unaffected by adversarial noise attacks, we are able to show that when the attack is targeting those regions directly, the effect is much more severe. A qualitative example demonstrating this effect is shown in Fig. 3. While the annotated region noticeably changed when the adversarial attack was focused on it (third column), the effect is much smaller when the attack was performed on the complete image (fourth column). Clinically interpreting the results, we note that when only the annotated region is attacked (third column), it starts to look more defined and homogeneous. This could lead to a failure to detect the ACL rupture which is present in the image.

The SSIM scores after adversarially attacking our models with rotations are shown in Fig. 4. Comparing Fig. 4a with Fig. 1a above, it can be seen that rotation attacks are less severe than attacks with adversarial noise which is not unex-

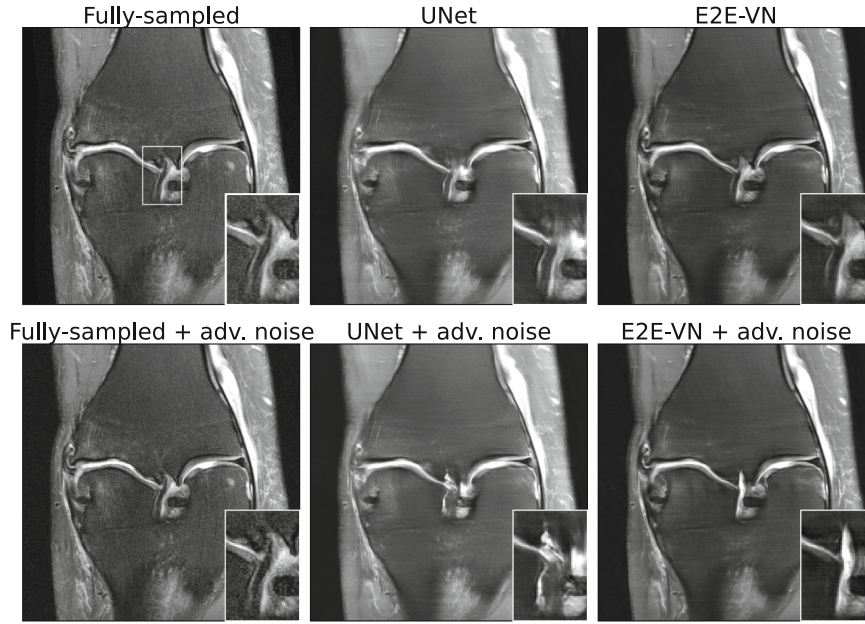


Fig. 2. Comparison of reconstruction algorithms with a relative noise-level of 2% and $8\times$ -acceleration (top row: unperturbed reconstruction, bottom row: reconstruction with adversarial noise). The fully sampled image shows a low grade sprain of the anterior cruciate ligament (ACL). Both the UNet and the E2E-VN reconstructions with adv. noise could lead to clinically relevant misdiagnoses. Note that changes are barely visible if the adversarial noise is added to the fully-sampled k -space (bottom left).

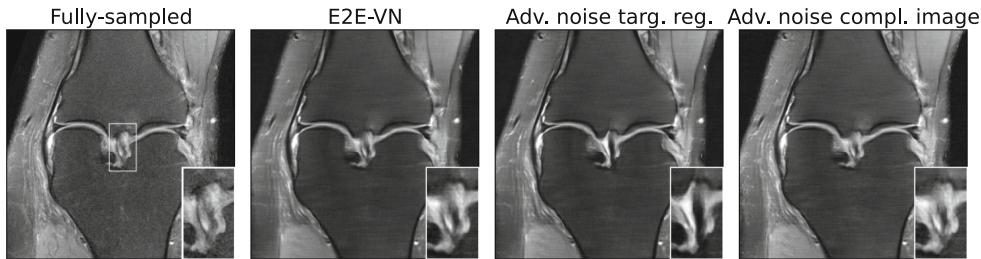


Fig. 3. Comparison of attacking the targeted region only (third column) versus attacking the complete image (fourth column) using the E2E-VN trained for $8\times$ acceleration. Attacking only the annotated region, the ligament appears more homogeneous, which could lead to a misdiagnosis of the ACL rupture.

pected given the much higher degree of freedom of the noise attacks. However, the rotation attacks, although very small and purportedly benign in nature, can still manage to substantially degrade the output of the algorithm. From Fig. 4b it can be observed that the attack is continuous in θ and a non-negligible range of rotation angles leads to substantial drops in SSIM. This indicates that such degradation could realistically occur in clinical practice. A qualitative example of a UNet reconstruction is shown in Fig. 5. While the rotation has negligible effects on the fully-sampled reconstruction (third column), it can be seen that the UNet attacked with the same rotation (fourth column) produces an image

where the ACL appears fuzzy (similar to the example in Fig. 3). As in the previous examples, this could potentially lead to a misdiagnosis.

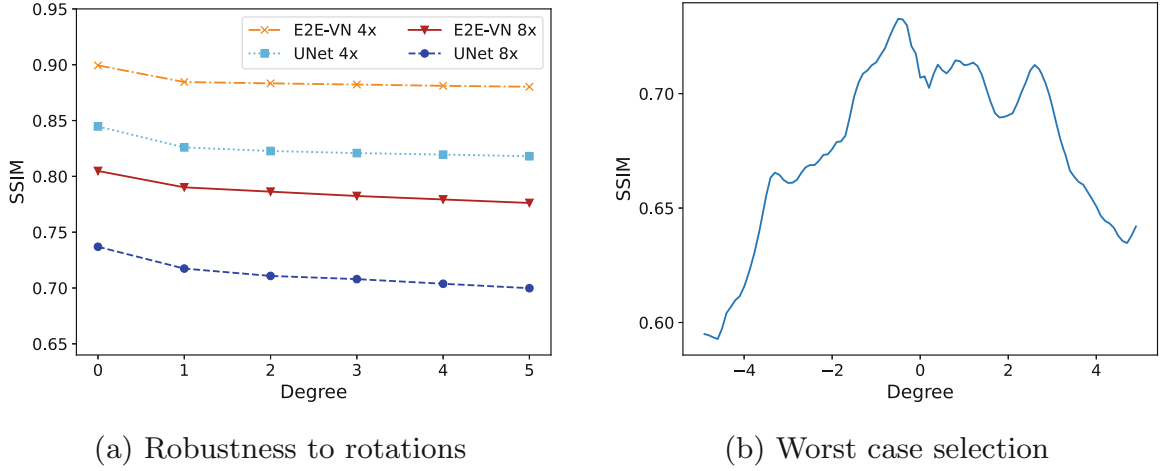


Fig. 4. Results of adversarial rotation experiments. (a) contains the methods’ SSIM with increasing severity of the attack. (b) visualises the sensitivity of the SSIM score with respect to θ for the qualitative example shown in Fig. 5.

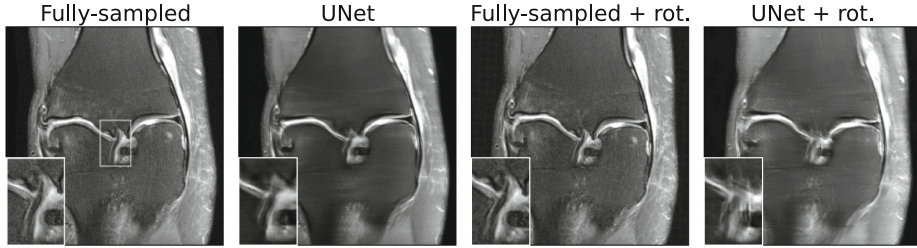


Fig. 5. Effects of small rotation on the reconstruction quality (8 $\times$ acceleration). For the worst-case angle of -4.9° , the annotated region gets more fuzzy, potentially leading to a misdiagnosis of an ACL tear instead of a low grade sprain.

4 Conclusion

In this work we have shown potential pitfalls of using DL-based MR-reconstruction methods. By creating realistic adversarial changes to the k -space data we have shown that the reconstruction quality of common reconstruction methods declines significantly for both an iterative network (i.e. the E2E-VN) and a UNet-based approach. We observed that both methods, although the architecture and reconstruction method are quite different, encounter similar levels of reconstruction quality decline when adversarial perturbations are applied. By further targeting only specific regions we have shown that diagnostically relevant information can – contrary to the findings of [6] – fall victim to adversarial attacks for the same noise levels. Lastly, we observed that both examined reconstruction algorithms are also sensitive to very small adversarial rotations in some

instances also leading to loss of diagnostic information. We believe future work should increasingly incorporate such worst case analyses and focus on the creation of models that are relatively stable under such perturbations. A limitation of our work is that we explored reduced versions of the UNet and E2E-VN reconstruction architectures due to computational constraints. In future work, we aim to verify that our insights carry-over to the full size models.

Acknowledgements. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Jan Nikolas Morshuis.

References

1. Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A.C.: On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci.* **117**(48), 30088–30095 (2020). <https://doi.org/10.1073/pnas.1907377117>
2. Calivá, F., Cheng, K., Shah, R., Pedoia, V.: Adversarial robust training of deep learning MRI reconstruction models. *arXiv preprint arXiv:2011.00070* (2020)
3. Cheng, K., Calivá, F., Shah, R., Han, M., Majumdar, S., Pedoia, V.: Addressing the false negative problem of deep learning MRI reconstruction models by adversarial attacks and robust training. In: Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C. (eds.) *Proceedings of the Third Conference on Medical Imaging with Deep Learning*. *Proceedings of Machine Learning Research*, vol. 121, pp. 121–135. PMLR, 06–08 July 2020. <https://proceedings.mlr.press/v121/cheng20a.html>
4. Fabian, Z., Heckel, R., Soltanolkotabi, M.: Data augmentation for deep learning based accelerated MRI reconstruction with limited data. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 139, pp. 3057–3067. PMLR, 18–24 July 2021. <https://proceedings.mlr.press/v139/fabian21a.html>
5. Firbank, M., Coulthard, A., Harrison, R., Williams, E.: A comparison of two methods for measuring the signal to noise ratio on MR images. *Phys. Med. Biol.* **44**(12), N261 (1999)
6. Genzel, M., Macdonald, J., März, M.: Solving inverse problems with deep neural networks - robustness included? (2020). <https://doi.org/10.48550/ARXIV.2011.04268>. <https://arxiv.org/abs/2011.04268>
7. Hyun, C.M., Kim, H.P., Lee, S.M., Lee, S., Seo, J.K.: Deep learning for undersampled MRI reconstruction. *Phys. Med. Biol.* **63**(13), 135007 (2018). <https://doi.org/10.1088/1361-6560/aac71a>
8. Johnson, P.M., et al.: Evaluation of the robustness of learned MR image reconstruction to systematic deviations between training and test data for the models from the fastMRI challenge. In: Haq, N., Johnson, P., Maier, A., Würfl, T., Yoo, J. (eds.) *MLMIR 2021*. *LNCS*, vol. 12964, pp. 25–34. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88552-6_3
9. Liu, X., Glocker, B., McCradden, M.M., Ghassemi, M., Denniston, A.K., Oakden-Rayner, L.: The medical algorithmic audit. *Lancet Digit. Health* **4**(5), e384–e397 (2022)

10. Macovski, A.: Noise in MRI. *Magn. Reson. Med.* **36**(3), 494–497 (1996)
11. Muckley, M.J., et al.: Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. *IEEE Trans. Med. Imaging* **40**(9), 2306–2317 (2021). <https://doi.org/10.1109/TMI.2021.3075856>
12. Roemer, P.B., Edelstein, W.A., Hayes, C.E., Souza, S.P., Mueller, O.M.: The NMR phased array. *Magn. Reson. Med.* **16**(2), 192–225 (1990)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.* **60**(1–4), 259–268 (1992)
15. Schlemper, J., Caballero, J., Hajnal, J.V., Price, A., Rueckert, D.: A deep cascade of convolutional neural networks for MR image reconstruction. In: Niethammer, M., et al. (eds.) *IPMI 2017*. LNCS, vol. 10265, pp. 647–658. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_51
16. Sriram, A., et al.: End-to-end variational networks for accelerated MRI reconstruction. In: Martel, A.L., et al. (eds.) *MICCAI 2020*. LNCS, vol. 12262, pp. 64–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_7
17. Tezcan, K.C., Baumgartner, C.F., Luechinger, R., Pruessmann, K.P., Konukoglu, E.: MR image reconstruction using deep density priors. *IEEE Trans. Med. Imaging* **38**(7), 1633–1642 (2018)
18. Darestani, M.Z., Chaudhari, A.S., Heckel, R.: Measuring robustness in deep learning based compressive sensing. In: *International Conference on Machine Learning (ICML)* (2021)
19. Zbontar, J., et al.: fastMRI: an open dataset and benchmarks for accelerated MRI (2018)
20. Zhao, R., et al.: fastMRI+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Sci. Data* **9**(1), 152 (2022). <https://doi.org/10.1038/s41597-022-01255-z>
21. Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S.: Image reconstruction by domain-transform manifold learning. *Nature* **555**(7697), 487–492 (2018)

Appendix B

Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data

Source Publication

Jan Nikolas Morshuis, Matthias Hein, and Christian F. Baumgartner. “Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data”. In: *IEEE Transactions on Biomedical Engineering* (2026). DOI: 10.1109/TBME.2026.3679957

Understanding Benefits and Pitfalls of Current Methods for the Segmentation of Undersampled MRI Data

Jan Nikolas Morshuis, Matthias Hein, and Christian F. Baumgartner

Abstract—MR imaging is a valuable diagnostic tool allowing to non-invasively visualize patient anatomy and pathology with high soft-tissue contrast. However, MRI acquisition is typically time-consuming, leading to patient discomfort and increased costs to the healthcare system. Recent years have seen substantial research effort into the development of methods that allow for accelerated MRI acquisition while still obtaining a reconstruction that appears similar to the fully-sampled MR image. However, for many applications a perfectly reconstructed MR image may not be necessary, particularly, when the primary goal is a downstream task such as segmentation. This has led to growing interest in methods that aim to perform segmentation directly on undersampled MRI data. Despite recent advances, existing methods have largely been developed in isolation, without direct comparison to one another, often using separate or private datasets, and lacking unified evaluation standards. To date, no high-quality, comprehensive comparison of these methods exists, and the optimal strategy for segmenting accelerated MR data remains unknown. This paper provides the first unified benchmark for the segmentation of undersampled MRI data comparing 7 approaches. A particular focus is placed on comparing *one-stage approaches*, that combine reconstruction and segmentation into a unified model, with *two-stage approaches*, that utilize established MRI reconstruction methods followed by a segmentation network. We test these methods on two MRI datasets that include multi-coil k-space data as well as a human-annotated segmentation ground-truth. We find that simple two-stage methods that consider data-consistency lead to the best segmentation scores, surpassing complex specialized methods that are developed specifically for this task. Our code is available at <https://github.com/NikolasMorshuis/UndersampledMRISeg>

Index Terms—Deep learning, Image reconstruction, Magnetic resonance imaging, Semantic segmentation

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) has proven to be a powerful modality to segment anatomical structures and pathologies due to its high soft-tissue contrast. However, a

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2064/1, Project number 390727645. We also utilized the Tübingen Machine Learning Cloud, supported by DFG FKZ INST 37/1057-1 FUGG. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for their support of JNM, MH, and CFB.

JNM, MH, and CFB are with the University of Tübingen. CFB additionally is with the Faculty of Health Sciences and Medicine, University of Lucerne.

major disadvantage of MRI acquisition are the long scan times required. This does not only lead to high healthcare costs and longer waiting times but also to increased patient discomfort [28] and potential artifacts in the MRI image due to patient movement [42]. Faster MRI acquisition is therefore desirable and methods such as parallel imaging [13], [29] and compressed sensing [21] have achieved significant accelerations in the last two decades. Recent deep learning based reconstruction methods can lead to a further reduction in scan times and are therefore currently a major research topic [14].

A widely used strategy to accelerate MRI acquisition is to reduce the amount of data collected, resulting in an under-sampled k-space. As the undersampling factor increases, the reconstruction problem becomes increasingly ill-conditioned, which can lead to noise amplification and potential artifacts in the reconstruction. To solve the ill-posed reconstruction problem some prior information is necessary. This prior can be based on known properties of MR images like sparsity and smoothness constraints as used in compressed sensing methods [11], [21], or the prior can also be learned as is the case for most machine learning-based reconstruction approaches [14].

A tradeoff exists between the quality of the reconstruction and the undersampling factor – higher accelerations allow shorter measurement times but typically also lead to worse reconstructions. What undersampling factor to choose largely depends on the downstream task: For instance, [30] have shown that even though an accurate reconstruction might not be possible for high accelerations, the reconstructed images can still be used for initial screening by radiologists, who can decide which patients need further examination. Therefore, the acceleration factor of the MRI can be increased significantly if the purpose of the acquisition is not the creation of accurate reconstructions but rather the downstream task of initial screening. Similarly, the ultimate purpose of an MRI acquisition might be to find segmentations of anatomical or pathological structures to track disease progressions [41]. For many segmentation tasks, not all details displayed in a fully-sampled MRI image are required, and larger structures are still recognizable even for high acceleration. The task of segmenting undersampled MRI images is therefore of high interest due to its applicability and its potential to save costly measurement time.

A number of works have explored methods to jointly optimize MRI reconstruction for optimal segmentation and

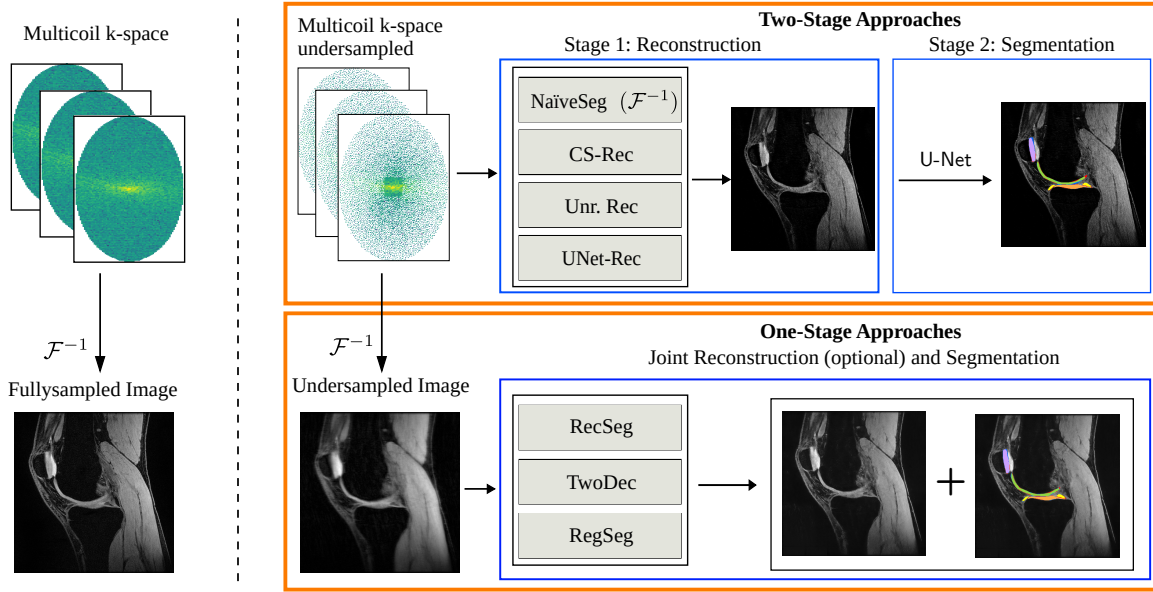


Fig. 1: Overview of tested reconstruction and segmentation methods. We separate modern methods that are optimized to segment undersampled MRI data into two categories: Two-Stage Approaches and One-Stage Approaches. Two-stage approaches first reconstruct the undersampled MR image and a simple UNet is then used to segment the reconstructed image. One-stage approaches try to combine both tasks and it is often claimed that this combination of tasks is helpful to achieve better reconstructions and segmentations. Our results suggest the contrary finding.

reconstructions rather than for reconstruction fidelity alone [4], [35], [37]. Unfortunately, the majority of these methods suffer from common validation pitfalls [18] and do not compare to each other directly and instead only compare to simple baseline methods [4], [35]. Moreover, most methods only consider single and private datasets [4], [35], [37] and sometimes only single acceleration factors [37], such that the behavior of the algorithms across datasets and across acceleration factors remains unknown.

To improve comparability between methods, a MICCAI challenge—the K2S Challenge [37]—invited participants to perform segmentation given only undersampled k-space data and evaluated submissions based on the quality of the resulting segmentations. Interestingly, the winning method that predicted the best segmentations also obtained the lowest reconstruction fidelity of all submissions and did not explicitly optimize for high reconstruction scores. Moreover, the third-placed method skipped the reconstruction altogether and trained a segmentation network on the undersampled data directly.

However, results from challenges such as the K2S Challenge should be interpreted with caution [23], [39], as it is often unclear which specific factors within a method's implementation are responsible for its performance [12]. Apart from the network architecture, implementation decisions like data augmentation, number of iterations, and ensembling also play a crucial role and might be even more important than the underlying architecture [17]. Moreover, the K2S challenge was performed on a single and now private dataset. Therefore neither prior work nor the results of the K2S challenge can be used to answer the question of which method for segmenting

undersampled MRI data performs the best and if high-quality reconstructions are indeed needed for better segmentation performance.

In order to address this research gap and generate additional insight about segmenting undersampled MRI data, this paper provides the first fair comparison and a unified evaluation framework for many state-of-the-art methods in the current literature. We note that this paper does not aim to introduce a novel method for the segmentation of accelerated MRI images. Instead, in this paper, we create a rigorous benchmark which is currently lacking from the literature. Specifically, we perform an extensive analysis of several methods on two knee MRI datasets and for several acceleration factors. We show that there is indeed no correlation between reconstruction scores in terms of PSNR and segmentation scores as measured by the Dice score [10]. Our results also show that most methods do not achieve statistically significant improvements compared to the simple baseline of segmenting the image calculated from undersampled k-space data directly. However, two-stage methods where the reconstruction methods enforce k-space consistency yield small yet still statistically significant better segmentation performance. This suggests that incorporating k-space consistency is a critical component for effective segmentation from undersampled MRI data.

II. RELATED WORK

Only a small number of papers have focused on the task of obtaining segmentations directly from undersampled MRI data. In [34], the authors trained a cardiac-segmentation network on undersampled MRI-data directly, encouraging the

creation of a consistent segmentation over different acceleration factors. In [15] the authors used a combination of iterative reconstruction, an attention mechanism, a recurrent module, as well as a segmentation module to create adequate segmentations. They restricted their analysis, however, to one private single-coil dataset and did not analyze the effect of different acceleration factors on the segmentation quality. The authors of [36] combined the reconstruction with the segmentation task by using U-Net-like networks, in which the representation created by a shared encoder is used as input for the reconstruction and segmentation decoder. In a similar manner, [4] have combined the MRI reconstruction and segmentation into one V-Net-like architecture that consists of a shared encoder and utilizes two decoders, one decoder for the reconstruction task and one for the segmentation task. In this method, the skip connection for the segmentation method is obtained from the reconstruction decoder instead of the encoder. Another line of work [1], [35] indicates that good reconstruction and segmentation results can be achieved by making use of a reconstruction network in cascade with a segmentation network. The analyses, however, focus on the reconstruction scores and no in-depth analysis of effects on the Dice score is provided.

In a recent work by [25] the authors considered the uncertainty that results from the segmentation of undersampled MRI data. The authors analyzed the ill-posed nature of the MRI reconstruction problems: As there are infinite possible reconstructions that are data-consistent with the measured k-space, the authors generated several reconstructions corresponding to minimum and maximum segmentations. The true segmentation lies in-between these extreme cases. However, the authors did not try to maximize segmentation performance and instead focused on providing reliable uncertainty estimates.

Predicting high-quality segmentations has been the objective in the K2S challenge [37], where several methods for segmenting undersampled MRI data have been submitted. While reconstruction fidelity was also analyzed for the methods that did provide reconstructions, it was not used as an evaluation criterion. A drawback of the challenge is that only a single acceleration factor (8x) was considered, leaving the behavior of the submitted algorithms on higher acceleration factors unknown. Moreover, submitted methods differed not only in architecture, but also in other crucial aspects of the training pipeline such as augmentation and optimization strategies, or ensembling of multiple networks. It is well-known that these factors can preclude any meaningful conclusions from challenge results [23], [39] and the success of a method can potentially be attributed to the wrong component of a system [12].

III. METHODS

A. Data

In order to evaluate methods that create segmentations for undersampled MRI-data, datasets are required that provide both the raw k-space measurements and ground-truth segmentations. It is then possible to simulate undersampled MRI

data by retrospectively masking the determined fraction of k-space points according to some undersampling pattern such as Poisson undersampling [3].

The K2S dataset [37] consists of multi-coil k-space data of 300 3D isotropic MRI scans of the knee and segmentations of the bones and cartilage. One scan of the dataset was excluded in the following experiments, because the MRI-image provided has not been correctly aligned with the corresponding segmentation. From the remaining 299 images, a train-, val-, test-split of 240, 30, 29 has been created. An Poisson-disc undersampling mask with a factor 8 is provided with the challenge data. Masks of higher acceleration factors have been created using the *sigpy* library [27], with default parameters and a calibration region (fully-sampled center region) according to the original mask provided by the challenge. The fully-sampled center region was reduced for higher acceleration factors proportionally. The segmentation regions contained a total of 6 classes: The 3 bone-regions of the knee Patellar, Femoral and Tibial as well as the corresponding 3 cartilages.

Furthermore, we conducted experiments on the publicly available SKM-TEA dataset [8]. Similar to the K2S dataset mentioned above, this dataset consists of 155 3-dimensional multi-coil k-space data samples as well as their corresponding segmentations. Using the k-space data and the provided Poisson-disc undersampled masks, we selected acceleration factors of $8\times$ and $16\times$, where $16\times$ corresponds to the maximum acceleration factor for which an undersampling mask is provided by the dataset. The segmentation classes contain the patellar, femoral, and tibial cartilage as well as the meniscus. We use the official data split described in [8], which results in a training, test, and validation set of 86, 33, and 36 images, respectively.

The SKM-TEA MR images can be calculated from the k-space data by making use of 2D inverse Fast Fourier Transforms (iFFT), while the MRI images of the K2S data can only be calculated using a 3D iFFT. Image-voxels of the K2S images therefore depend on all measured k-space points of the MRI volume, while voxels in the SKM-TEA images only depend on the k-space points of the respective 2D MRI slice.

B. Reconstruction and Segmentation Methods

We analyze several state-of-the-art methods that have been developed for the segmentation of undersampled MRI data (see summary in Figure 1). We separate the methods into two-stage and one-stage methods: Two-stage methods generate the reconstruction in a first step before predicting the segmentation in a second step. For the segmentation prediction, a standard U-Net as provided by nnU-Net [17] is utilized. The one-stage methods are specialized methods for segmenting undersampled MRI data and combine both tasks in a single forward pass. Details on our tested methods are provided below.

All segmentation methods are trained using the same loss-function, which is a combination of Dice and Cross-Entropy and is also utilized in the nnU-Net framework [17]. If a reconstruction loss is required for the method, we apply the Mean-Squared-Error (MSE). We unify these losses for better comparability of the methods.

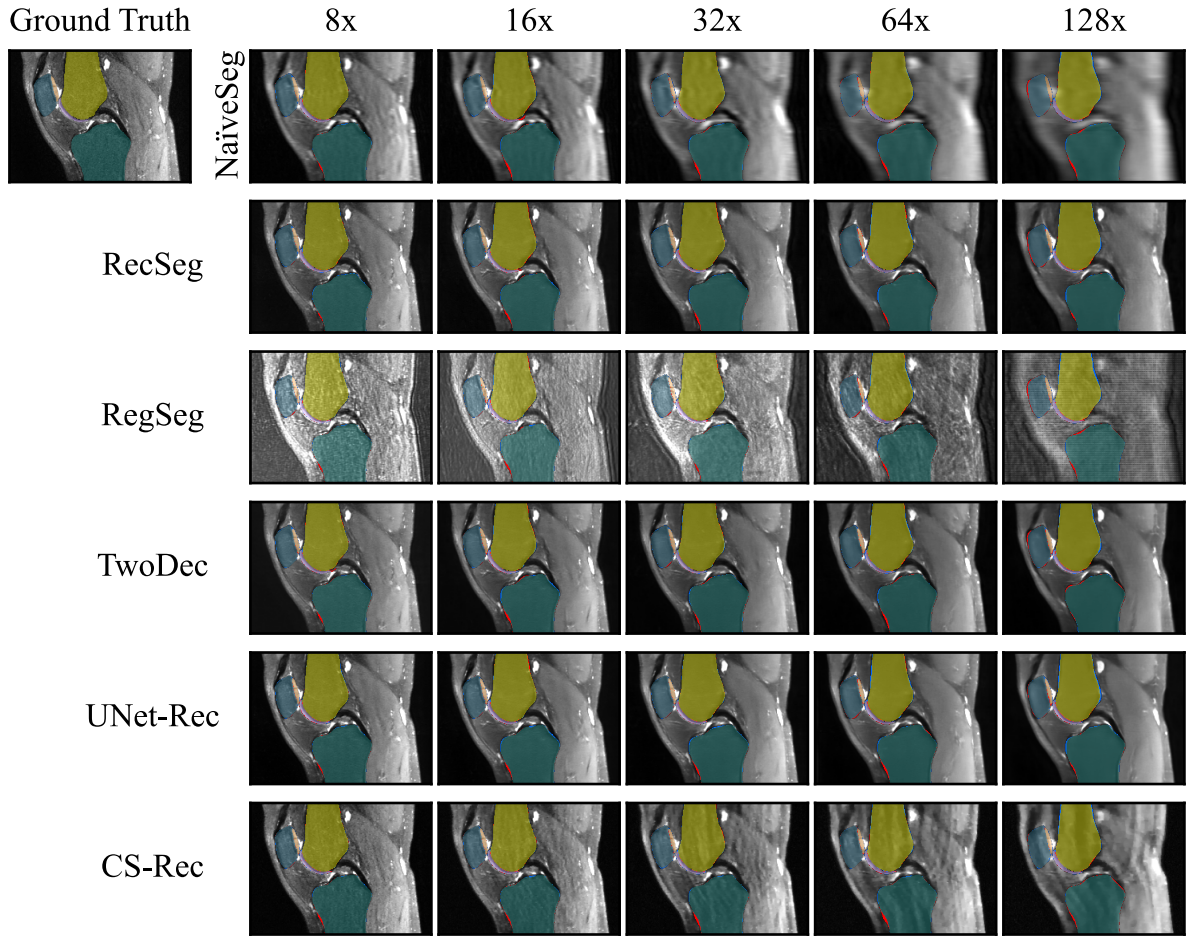


Fig. 2: Example reconstructions and segmentations for the tested methods on the K2S dataset. False-positive segmentations are marked in red, and false negative segmentations are marked in blue. The RegSeg method does not necessarily generate reconstructions similar to the ground-truth, as this is no objective during training and instead the method focuses solely on the generation of good segmentations.

1) Two-stage methods: **NaïveSeg:** This method relies on the creation of a zero-filled reconstruction by replacing the missing k-space measurements with zeros and applying an inverse Fourier transform. In a following step, we train a U-Net [32] based segmentation network directly on the zero-filled reconstructions. Because we utilize the method as a simple baseline method, we refer to this method as NaïveSeg. This approach was also submitted to the K2S challenge [20] where it reached third place.

UNet-Rec [16]: This two-stage approach utilizes an U-Net based reconstruction, originally proposed similarly by [16], yet we work with 3-dimensional (3D) reconstruction U-Nets with the same architectures as the ones used in nnU-Net [17]. First the images are reconstructed with a 3D U-Net. As reconstruction loss we use the Mean-Squared-Error. After the reconstruction, we train the U-Net based segmentation network in the same manner as NaïveSeg on the reconstructed images.

CS-Rec. [11], [21]: In this approach, first a Compressed Sensing reconstruction (CS-Rec.) is applied to the image reconstructed from undersampled data. CS-based reconstruction is a well-established method that is also applied in

clinical practice [5], [11], [22]. We use a combination of Total Variation [2] and L1-Wavelet [6] as the optimization objective. In a second step, a segmentation network is trained to segment the anatomical structures from the reconstructed image. We utilize the original code used by [31] the second-place submission in the K2S challenge [37] for the image reconstruction. However, in contrast to the original method, we replace the V-Net segmentation architecture with the same U-Net architecture as NaïveSeg for better comparability to the other methods in our benchmark.

Unr-Rec [9]: Similarly to CS-Rec here we first generate reconstructions and then segment the image in a two-stage process. The reconstructions are obtained using the unrolled reconstruction method (Unr-Rec) proposed in [9], [33] and also implemented as a baseline in the SKM-TEA dataset [8] that we utilize in our implementation. Unrolled reconstruction networks consist of multiple U-Nets in cascade with data-consistency elements in between. Due to computational constraints, unrolled reconstruction is currently limited to 2-dimensional imaging, as backpropagating through multiple 3D Fourier Transforms is computationally infeasible. The k-

space structure of the SKM-TEA data allows for 2-dimensional Fourier transforms, but for the K2S data the full 3-dimensional k-space is required in order to calculate the inverse Fourier transform as mentioned in Section III-A. We therefore were only able to evaluate this method on the SKM-TEA data.

2) One-stage methods: **RecSeg** [35]: This method combines both the reconstruction and segmentation in a single forward pass. Specifically, the RecSeg model [35] uses two U-Nets in a cascade, in which the first U-Net learns to reconstruct the MR image, while the second U-Net utilizes this reconstruction to estimate the segmentation. The difference to the UNet-Rec approach above is that here the two networks are trained simultaneously in an end-to-end fashion. The reconstruction network therefore also receives training gradients from the segmentation loss, potentially providing reconstructions that lead to better segmentations. Following the original method we use an MSE-loss for training the reconstruction network, but use a Dice plus Cross-Entropy loss for the segmentation network as we also do for all other approaches.

RegSeg [24]: This was the winning method of the K2S challenge [37]. Similar to RecSeg, the RegSeg approach places a reconstruction and a segmentation network in a cascade. There are however two key differences to RecSeg, which is the optimization objective and the training method: RegSeg pre-trains both U-Nets independently for half the training epochs on the reconstruction and segmentation tasks respectively. For the pre-training of the segmentation network, fully-sampled MR images were utilized in combination with the images reconstructed from undersampled data. After pre-training, the two U-Nets are placed in a cascade, where the segmentation network takes as input the reconstructed image as well as the zero-filled reconstruction of the original undersampled data. The two U-Nets are then trained for the second half of the epochs in an end-to-end manner on the segmentation task only. This allows the intermediate reconstructions to deviate from the fully-sampled image, thereby creating images that might be better suited for segmentation. These generated images will, however, potentially deviate strongly from the original reconstruction, as the generation of high-quality reconstructions is not the goal of RegSeg and instead the method focuses only on the segmentation quality.

TwoDec [4]: This method is an adaptation of the TB-Recon method originally proposed by [4]. It consists of a shared encoder and two decoders that solve the reconstruction and the segmentation, respectively. While the original method employs a V-Net architecture, we opted for a U-Net-like base architecture. We also changed the loss function to include the Dice + Cross-Entropy loss as in all other segmentation models and adjusted the reconstruction function to MSE for better comparability to the other methods. Moreover, we observed that a V-Net sometimes led to unstable training on the K2S dataset. As in the original publication we set the weight for the reconstruction loss and the segmentation loss to be equal for the K2S dataset and the SKM-TEA dataset at 8x acceleration, but do an hyperparameter search for the SKM-TEA dataset at 16x acceleration in Section IV-A.

C. Experimental Setup

In MRI reconstruction, the goal is to find a reconstruction $\hat{x}$ that approximates the ground-truth image x while remaining consistent with the measured k-space data y :

$$y = A\hat{x} \quad \text{with} \quad y \in \mathbb{C}^m, \quad A \in \mathbb{C}^{m \times n}, \quad \hat{x} \in \mathbb{C}^n, \quad (1)$$

where $A = MFS$ with the coil sensitivity maps S , the Fourier transform $\mathcal{F}$ and the undersampling mask M . In this work, our focus is not on predicting the most accurate reconstructions $\hat{x}$ but on predicting high-quality segmentations $\hat{s}$ that closely match the ground truth segmentations s . In the **one-stage** approaches, the methods are applied directly on the MRI image that is derived from the zero-filled k-space data y , such that the segmentation $\hat{s}$ is predicted as $\hat{s} = f_\theta(y)$. In the **two-stage** approaches the image is first reconstructed using existing reconstruction methods $\hat{x} = g_\psi(y)$ and then segmented using a standard 3D U-Net framework (f_ϕ) to estimate the segmentation $\hat{s} = f_\phi(\hat{x})$. We investigate which approach leads to the most accurate segmentations $\hat{s}$ under potentially highly accelerated MRI acquisitions (see Fig. 1 for a concept overview). Note that not all reconstruction methods require strict data-consistency as expressed in Equation (1).

The primary goal of this paper is to provide a rigorous benchmark of a representative set of recent methods that can segment undersampled MRI images. To enable meaningful comparisons, we standardize the evaluation setup to ensure maximal comparability across approaches. This, in turn, allows us to draw clear conclusions from our experiments and highlight promising directions for future research in this important yet underexplored field. In order to make the tested approaches maximally comparable, we modify the networks' architectures such that the employed components are the same and the differences in the score originate from changes in the tested method instead of the design of the network.

To achieve comparability in the other aspects of the training setup like data-augmentation, training loss-function, data-preprocessing etc., we base all re-implemented methods on the well-established nnU-Net framework [17], that was also utilized by the winning method of the K2S challenge [37]. For all segmentation methods we therefore utilize a training loss consisting of a combination of the Dice loss and the Cross-Entropy loss. An high-level overview of similarities and differences is provided in Figure 1.

D. Metrics

We apply the Dice score to evaluate the predicted segmentations and utilize the SSIM [38] and PSNR scores in order to evaluate the quality of the predicted reconstructions. The metrics have been calculated using implementations provided by the MONAI framework [7].

E. Implementation Details

We train and test all methods described in Section III-B from scratch on both the K2S and SKM-TEA datasets for all tested acceleration factors using a comparable training pipeline. All segmentation models are trained for 1000 epochs

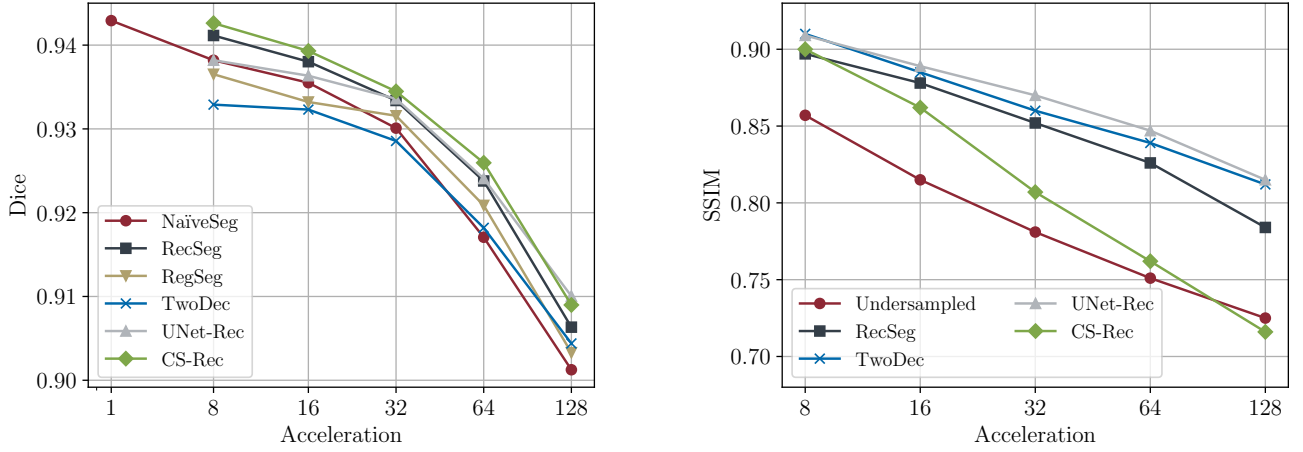


Fig. 3: Dice (left) and SSIM (right) scores for different accelerations on the K2S dataset. Note that for better readability we did not include RegSeg scores in the SSIM plot on the right, as the method has solely been developed to improve segmentation scores and does not aim at creating a good reconstruction. We provide SSIM scores for RegSeg in Table II. Interestingly, CS-Rec achieves the highest Dice scores while only achieving relatively low SSIM scores. The PSNR score behaves similarly as the SSIM score across accelerations.

of 250 iterations on an Nvidia Geforce RTX 2080Ti GPU using a batch-size of 2 and the Adam optimizer [19]. We use a learning rate of $1e-2$ for most methods and $1e-3$ for RecSeg and TwoDec for training stability. The data-augmentation strategy is identical for all methods and consists of mirroring in the sagittal direction.

IV. EXPERIMENTS AND RESULTS

A. Segmentation Quality

The segmentation scores for all methods and all acceleration factors for the K2S dataset are shown in Figure 3 (left). The quantitative values for the K2S, and the SKM-TEA datasets are further reported in Tables I and IV, respectively. We observed that, overall, the tested methods predicted segmentations of comparable quality and reached similar Dice scores. For both datasets, the difference in Dice score is within one standard deviation between the best and the worst performing method for all tested acceleration factors. The segmentations shown in Figure 2 for the K2S dataset and in Figure 6b for the SKM-TEA dataset show examples of the similar segmentation quality between the different methods. It can be seen that the differences in the segmentations are rather small within the same acceleration factor for all tested methods. We further note that the segmentations for all examined methods are of high quality and the errors are often of a similar nature for all methods, like the too large segmentation of the tibia bone as shown in Figure 2.

In order to better understand the differences between the examined methods we analyzed them using a statistical significance test. Specifically, we compared the Dice scores of the tested models to the ones obtained by the simplest baseline method NaiveSeg using a Wilcoxon-Signed-Rank test [40]. Statistically significant improvements in Dice score ($p < 0.01$) are indicated by a star ‘*’ in Tables I and IV. Even though the models are conceptually different, the Dice scores are rather

similar despite varying SSIM scores of the reconstruction. Surprisingly, the CS-Rec method leads to the segmentation scores that are among the best for both datasets over nearly all acceleration factors. Furthermore, the results for CS-Rec are nearly always significantly better compared to NaiveSeg. Another surprising result is the good segmentation quality of NaiveSeg compared to the other more specialized methods.

The method achieving the best Dice scores for the SKM-TEA dataset is Unr-Rec, which achieves the best scores for both tested acceleration factors and is the only method achieving statistically significantly better segmentation results compared to NaiveSeg at $16\times$ acceleration. Also, most other methods except for CS-Rec are statistically significant worse compared to Unr-Rec at $16\times$ acceleration for the SKM-TEA dataset (see Table VI). Note that Unr-Rec cannot be tested on the K2S data, because of the 3-dimensional k-space of the MRI images, as explained in Section III-A. A class-wise comparison of the Dice scores at $16\times$ acceleration for the SKM-TEA data is provided in Table V.

We found that the performance of methods that combine the reconstruction and segmentation tasks such as RecSeg and TwoDec can be dependent on the dataset and the acceleration factor. We observed that both methods are comparable to other methods on the K2S dataset, yet perform worse than other methods on the SKM-TEA dataset for $16\times$ -acceleration (see Table IV). We have analyzed the behavior of the TwoDec method on SKM-TEA with $16\times$ acceleration when using higher weight (higher λ_{seg}) for the segmentation loss on the validation set

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{seg}\mathcal{L}_{seg} \quad (2)$$

The results are shown in Figure 5. Note that higher segmentation weight leads to higher Dice scores, but the reconstruction scores degrade. For the results in Table IV we chose $\lambda_{seg} = 10$, as it provides a good tradeoff between segmentation and

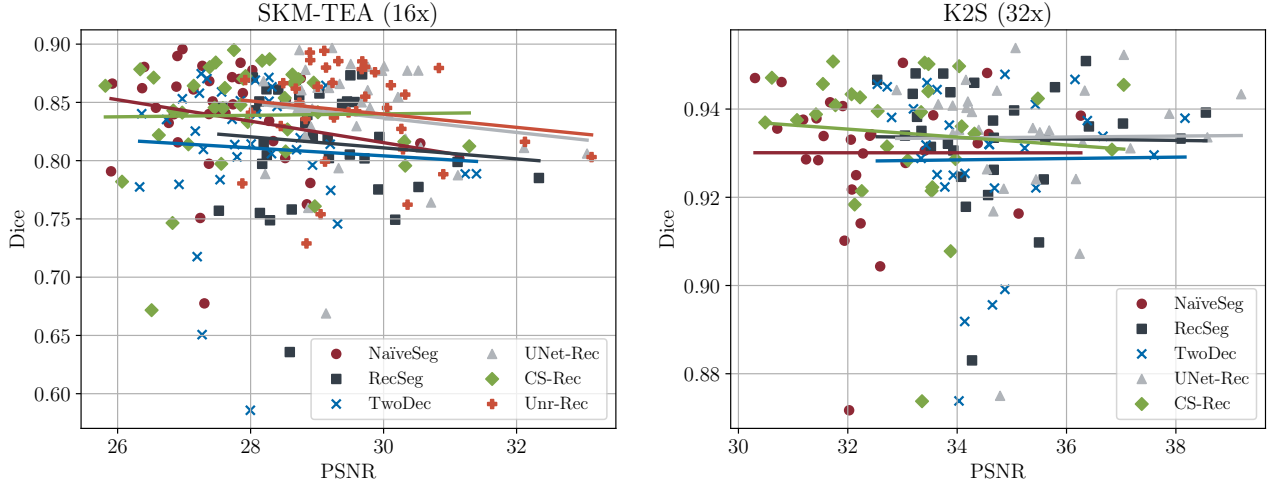


Fig. 4: Relation between PSNR values and the Dice scores. The left plot shows results for the SKM-TEA dataset with $16\times$ acceleration, the right plot shows the K2S dataset with $32\times$ acceleration. Every reconstructed image on the test set is a single data point. It can be seen by the regression fit that no positive correlation exists between the PSNR and Dice scores within the methods. This observation is confirmed quantitatively with Spearman's ρ ranging from 0 to -0.20 (all p -values > 0.2 , indicating no statistically significant correlation).

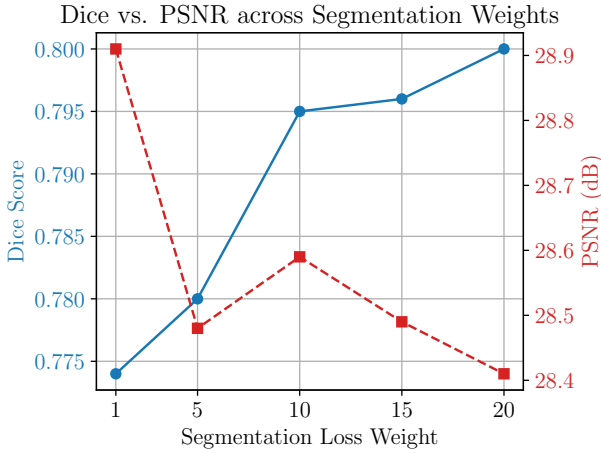


Fig. 5: Hyperparameter analysis. We analyzed segmentation and reconstruction results for different segmentation weights in the TwoDec method on the validation set of SKM-TEA $16\times$ data. Higher weights for the weight of the segmentation loss lead to higher segmentation scores, but the reconstruction quality declines.

reconstruction performance for the SKM-TEA dataset at $16\times$ acceleration. Note that even though we have optimized the segmentation weight hyperparameter for the TwoDec method on the $16\times$ acceleration experiments on SKM-TEA and thereby providing the TwoDec method with an advantage, the results are still worse compared to the other methods.

B. Reconstruction Quality

In contrast to the Dice scores, the reconstruction scores vary substantially between the methods. Interestingly, even

though the segmentation scores are good for CS-based reconstructions, CS-Rec does not achieve very high SSIM and PSNR scores in both datasets. In Figure 3 the SSIM score is shown for several acceleration factors. It can be seen that all methods that aim to generate a reconstruction of the MR image improve SSIM for nearly all acceleration factors compared to the zero-filled reconstruction used in the method NaiveSeg. The reconstruction scores of CS-Rec decline more for higher accelerations compared to the other methods, even leading to lower reconstruction scores than the original zero-filled reconstruction.

As can be seen in Figure 4, no correlation exists between the reconstruction score in terms of PSNR and the segmentation score when considering all predicted images and segmentations from the different methods individually. This indicates that the segmentation does not benefit from reconstructions that achieve higher PSNR values.

Tables II and IV provide PSNR and SSIM scores for the reconstruction of different methods. The best reconstruction scores are achieved by the UNet-Rec method for the majority of cases. Qualitative examples are shown in Figures 2 and 6. It can be seen that the UNet-Rec method also generates smooth reconstructions, which is known to be beneficial for the SSIM and PSNR metrics [26]. Lower PSNR scores are reached with the SKM-TEA data in Table IV compared to the K2S data shown in Table III.

V. DISCUSSION

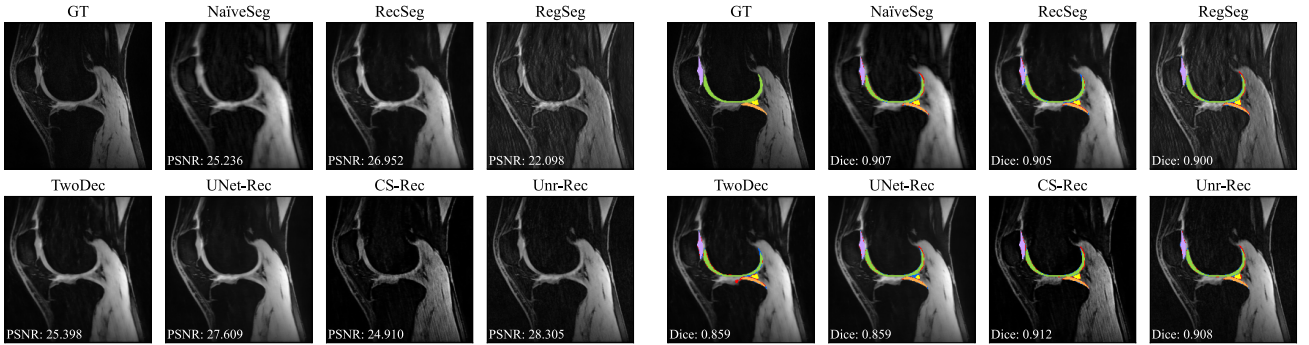
In our work, we have shown that, in contrast to previous findings [37], most methods for segmenting undersampled MRI data perform similarly for the same acceleration factors if the methods are trained in a comparable manner. Crucially, more complex approaches do not necessarily improve the segmentation score in a statistically significant way.

TABLE I: Dice scores for the K2S dataset. Values represent the Mean $\pm$ Standard Deviation. The fully-sampled NaïveSeg baseline ($1\times$ acceleration) achieved 0.943 ± 0.013 . Models with a statistically significantly ($p < 0.01$) higher Dice score compared to the NaïveSeg model at the respective acceleration factor are marked with an asterisk (*). The two-stage approach CS-Rec generally achieves the highest Dice scores. The scores remain stable up to $32\times$ acceleration.

		Dice					
Type		$1\times$	$8\times$	$16\times$	$32\times$	$64\times$	$128\times$
NaïveSeg	2-stage	0.943 ± 0.013	0.938 ± 0.018	0.936 ± 0.015	0.930 ± 0.015	0.917 ± 0.020	0.901 ± 0.022
RecSeg	1-stage	-	0.941 ± 0.012	0.938 ± 0.013	$0.933 \pm 0.013^*$	$0.924 \pm 0.018^*$	$0.906 \pm 0.024^*$
RegSeg	1-stage	-	0.937 ± 0.016	0.933 ± 0.020	0.932 ± 0.016	$0.921 \pm 0.020^*$	0.903 ± 0.022
TwoDec	1-stage	-	0.933 ± 0.021	0.932 ± 0.019	0.929 ± 0.018	0.918 ± 0.019	$0.904 \pm 0.024^*$
UNet-Rec	2-stage	-	0.938 ± 0.014	0.936 ± 0.020	$0.934 \pm 0.015^*$	$0.924 \pm 0.016^*$	$0.910 \pm 0.019^*$
CS-Rec	2-stage	-	$0.943 \pm 0.012^*$	$0.939 \pm 0.013^*$	$0.934 \pm 0.016^*$	$0.926 \pm 0.017^*$	$0.909 \pm 0.025^*$

TABLE II: SSIM scores for the K2S dataset. The reconstructions generated by the UNet-Rec methods usually achieve the highest SSIM scores. Note that the low scores of RegSeg can be explained by the fact that RegSeg does not aim to create a quality reconstruction and instead focuses on the segmentation.

		SSIM				
Type		$8\times$	$16\times$	$32\times$	$64\times$	$128\times$
NaïveSeg	2-stage	0.857 ± 0.024	0.815 ± 0.029	0.781 ± 0.034	0.751 ± 0.037	0.725 ± 0.039
RecSeg	1-stage	0.897 ± 0.020	0.878 ± 0.023	0.852 ± 0.027	0.826 ± 0.032	0.784 ± 0.037
RegSeg	1-stage	0.477 ± 0.074	0.603 ± 0.063	0.511 ± 0.076	0.610 ± 0.059	0.405 ± 0.071
TwoDec	1-stage	0.910 ± 0.018	0.885 ± 0.021	0.860 ± 0.027	0.839 ± 0.031	0.812 ± 0.036
UNet-Rec	2-stage	0.909 ± 0.018	0.889 ± 0.022	0.870 ± 0.025	0.847 ± 0.031	0.815 ± 0.033
CS-Rec	2-stage	0.900 ± 0.019	0.862 ± 0.027	0.807 ± 0.038	0.762 ± 0.044	0.716 ± 0.048



(a) Example of Reconstructions on the SKM-TEA dataset. (b) Example of Reconstructions and Segmentations on the SKM-TEA dataset. Segmentations: Patellar Cartilage (purple), Femoral Cartilage (green), Tibial Cartilage (orange), Meniscus (yellow), False Positives (red), False Negatives (blue). NaïveSeg shows the image reconstructed from undersampled data. RegSeg does not aim to generate perfect reconstructions but rather generates images helpful for the downstream segmentation task.

Fig. 6: Qualitative results from the SKM-TEA dataset showing (a) reconstructions and (b) segmentations at $16\times$ acceleration.

TABLE III: PSNR-scores for the K2S dataset. The Reconstruction provided by the UNet-Rec achieve the highest PSNR scores.

		PSNR				
Type		$8\times$	$16\times$	$32\times$	$64\times$	$128\times$
NaïveSeg	2-stage	35.0 ± 1.4	33.7 ± 1.4	32.5 ± 1.4	31.2 ± 1.4	30.2 ± 1.4
RecSeg	1-stage	37.1 ± 1.5	36.0 ± 1.5	34.7 ± 1.5	33.5 ± 1.5	32.0 ± 1.4
RegSeg	1-stage	26.0 ± 1.7	28.4 ± 1.6	27.4 ± 1.7	27.1 ± 1.7	25.3 ± 1.6
TwoDec	1-stage	36.7 ± 1.4	35.7 ± 1.4	34.6 ± 1.4	33.2 ± 1.4	32.3 ± 1.4
UNet-Rec	2-stage	37.6 ± 1.6	36.5 ± 1.5	35.4 ± 1.5	34.2 ± 1.5	32.8 ± 1.4
CS-Rec	2-stage	37.5 ± 1.6	35.5 ± 1.5	33.1 ± 1.6	31.3 ± 1.6	29.5 ± 1.5

We observed that two-stage methods which first generate a reconstruction before predicting the segmentations generally perform better, both creating reconstructions leading to higher similarity scores as measured in PSNR and SSIM and also pre-

dicting more accurate segmentations. The best segmentation scores were obtained with the CS-Rec and Unr-Rec method that both (explicitly or implicitly) enforce data-consistency with the k-space data when predicting the reconstructed image.

TABLE IV: Scores for the SKM-TEA data. While Unr-Rec achieves the highest Dice scores in our experiments, UNet-Rec achieves the highest SSIM and PSNR scores. Dice scores that are statistically significant better ($p < 0.01$) compared to NaïveSeg are marked with an asterisk (*).

	Type	Dice			SSIM		PSNR	
		1×	8×	16×	8×	16×	8×	16×
NaïveSeg	2-stage	0.854 ± 0.053	0.843 ± 0.044	0.837 ± 0.045	0.736 ± 0.024	0.692 ± 0.026	28.6 ± 1.1	27.6 ± 1.1
RecSeg	1-stage	-	0.845 ± 0.047*	0.816 ± 0.049	0.799 ± 0.024	0.739 ± 0.029	30.9 ± 1.0	29.4 ± 1.0
RegSeg	1-stage	-	0.847 ± 0.042*	0.838 ± 0.044	0.582 ± 0.036	0.509 ± 0.040	25.5 ± 1.4	24.8 ± 1.3
TwoDec	1-stage	-	0.842 ± 0.043	0.811 ± 0.060	0.817 ± 0.021	0.737 ± 0.026	30.8 ± 1.1	28.1 ± 1.1
UNet-Rec	2-stage	-	0.845 ± 0.044*	0.839 ± 0.045	0.826 ± 0.024	0.750 ± 0.042	31.6 ± 1.0	29.7 ± 1.0
CS-Rec	2-stage	-	0.847 ± 0.044*	0.839 ± 0.046	0.680 ± 0.028	0.632 ± 0.030	29.0 ± 1.2	27.9 ± 1.2
Unr-Rec	2-stage	-	0.848 ± 0.045*	0.843 ± 0.041*	0.704 ± 0.035	0.645 ± 0.036	31.2 ± 1.1	29.5 ± 1.1

TABLE V: Class-wise Dice scores for the SKM-TEA dataset at 16× acceleration. Dice scores that are statistically significantly better ($p < 0.01$) compared to NaïveSeg are marked with an asterisk (*). Unr-Rec achieves the highest dice-scores in all segmentation classes.

	Type	Dice			
		Patellar Cartilage	Femoral Cartilage	Tibial Cartilage	Meniscus
NaïveSeg	2-stage	0.858 ± 0.077	0.846 ± 0.043	0.830 ± 0.045	0.815 ± 0.054
RecSeg	1-stage	0.810 ± 0.097	0.826 ± 0.048	0.825 ± 0.042	0.802 ± 0.058
RegSeg	1-stage	0.860 ± 0.072	0.846 ± 0.045	0.832 ± 0.040	0.813 ± 0.057
TwoDec	1-stage	0.807 ± 0.138	0.824 ± 0.043	0.823 ± 0.041	0.788 ± 0.079
UNet-Rec	2-stage	0.858 ± 0.076	0.848 ± 0.042	0.833 ± 0.043	0.816 ± 0.056
CS-Rec	2-stage	0.859 ± 0.081	0.848 ± 0.044*	0.832 ± 0.042	0.816 ± 0.059
Unr-Rec	2-stage	0.864 ± 0.061	0.852 ± 0.040*	0.835 ± 0.042*	0.819 ± 0.058

TABLE VI: Statistical comparison to best performing 2-stage model for the SKM-TEA dataset. Dice scores that are statistically significantly worse ($p < 0.01$) compared to Unr-Rec are marked with a (†).

Model	Type	8×		16×	
		Dice	p-value	Dice	p-value
NaïveSeg	2-stage	0.846†	< 0.001	0.837†	< 0.001
RecSeg	1-stage	0.845	0.014	0.816†	< 0.001
RegSeg	1-stage	0.847	0.169	0.838†	< 0.001
TwoDec	1-stage	0.842†	< 0.001	0.811†	< 0.001
UNet-Rec	2-stage	0.845†	0.003	0.839†	0.005
CS-Rec	2-stage	0.847	0.157	0.839	0.011
Unr-Rec	2-stage	0.848	<i>Ref.</i>	0.843	<i>Ref.</i>

The high segmentation quality obtained with CS-Rec is particularly interesting as the predicted reconstructions did not achieve high reconstruction scores compared to other approaches. The remaining methods that did not enforce k-space consistency often led to worse segmentation results despite having better reconstructions qualitatively and quantitatively in terms of SSIM (see for example the TwoDec approach). This suggests that reconstructions that are faithful to the measured k-space data may be more important for achieving good segmentations than reconstructions that obtain high visual fidelity. Thus, enforcing k-space consistency is likely to be an important feature for future work in this domain.

Another research question we investigated was whether high-quality reconstructions provide an advantage for accurate segmentation prediction. The missing correlation between reconstruction and segmentation scores shown in Figure 4 suggests that a high-quality reconstruction is, indeed, not required to predict high-quality segmentations. Beyond CS-Rec, discussed above, two illustrative examples are RegSeg and

NaïveSeg. Although neither method aims to predict a reconstruction, both achieve segmentation performance comparable to methods that explicitly optimize for image reconstruction, such as TwoDec or UNet-Rec. This suggests that future solutions aiming exclusively at obtaining high segmentation performance from highly accelerated k-space data do not necessarily need to additionally predict the reconstruction. However, we note that depending on the application high-quality reconstructions may be required for other reasons such as human interpretability.

A disadvantage of the methods that combine the reconstruction with the segmentation task like RecSeg and TwoDec is the dependence on the dataset and acceleration factor. While the methods perform relatively well on the K2S Dataset, their performance degrades for the SKM-TEA dataset at accelerations of 16x as seen in Table IV. The TwoDec method is highly dependent on the selection of the loss weights. Careful tuning of those weights improved the segmentation score, yet the score was still worse compared to the other methods.

In our experiments we have observed differences in the similarity scores of the predicted reconstructions on the two examined datasets. PSNR and SSIM values for the reconstructions on the SKM-TEA dataset were much lower compared to the K2S Dataset. This may indicate that the images of the SKM-TEA dataset were more difficult to reconstruct compared to the images of the K2S dataset. A possible reason for this could be that the SKM-TEA data is of higher resolution compared to the K2S data (512x512x160 vs 256x256x196) and can therefore display additional anatomical details that can make the reconstruction process difficult.

A. Limitations

The evaluation in this paper focused on knee MRI data due to the limited availability of public datasets that contain both k-space data as well as ground-truth segmentations. As segmentation results depend on multiple factors such as the size of the structure and whether or not pathologies are present, our results might vary for other datasets that aim to reconstruct and segment i.e. brain MRI data.

The tested segmentation method in this paper is limited to the nnU-Net framework [17], as it still performs very well on 3D segmentation tasks, even compared to transformer-based segmentation methods [18]. We did not utilize transformer-based or diffusion model-based reconstruction methods as those methods are too computationally expensive to be applied to 3D reconstruction tasks at scale.

We also only utilized the Poisson-disc sampling pattern provided by the datasets [8], [37] as the effects of different undersampling patterns are not the focus of this work. Future work can test reconstruction and segmentation methods with different sampling patterns such as acquiring the image at a lower resolution, spiral- or uniform undersampling. Uniform undersampling masks might cause some of the models to fail for high accelerations, as the image becomes distorted too heavily for patch-based segmentation approaches.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown that good segmentation quality of undersampled MRI images is feasible even for high acceleration of up to 32x. Contrary to the findings in the K2S challenge [37], we have observed that reconstruction methods that enforce data-consistency - either strictly, as in CS-Rec, or implicitly, as in Unr-Rec - lead to the highest segmentation scores. However, we have also shown that the difference between the tested methods can be small, potentially making it difficult to see real-world benefits of the more complex and specialized methods. Experiments have also shown that the performance of some one-stage methods like RecSeg and TwoDec can depend on the dataset and acceleration factor.

In the context of future studies on the segmentation of undersampled MR images, we recommend focusing on two primary directions: The task of reconstruction and segmentation can be treated independently, where first the reconstruction is generated and afterwards utilized as input for the 'naïve' UNet-based segmentation network as done in the

tested two-stage methods. This leads to fewer hyperparameters that require tuning, to more stable results and to the best reconstruction (UNet-Rec) and segmentation scores (CS-Rec and Unr-Rec). Furthermore, we have seen that it is beneficial for the segmentation quality if the reconstruction method utilizes a data-consistency component, as is the case for CS-Rec and Unr-Rec.

Our research indicates that not all of the k-space information obtained in the fully-sampled image is required for an accurate segmentation prediction and high acceleration factors can be utilized. In the future, we believe it will be necessary to find a suitable trade-off between acceleration and segmentation quality that fits clinical standards while concurrently utilizes the MRI acquisition time more efficiently, allowing more patients to be treated and healthcare costs to be reduced.

ACKNOWLEDGEMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Jan Nikolas Morshuis.

REFERENCES

- [1] Acar, M., Çukur, T., Öksüz, İ.: Segmentation-aware MRI reconstruction. In: Haq, N., Johnson, P., Maier, A., Qin, C., Würfl, T., Yoo, J. (eds.) *Machine Learning for Medical Image Reconstruction*. pp. 53–61. Springer International Publishing, Cham (2022)
- [2] Block, K.T., Uecker, M., Frahm, J.: Undersampled radial MRI with multiple coils. iterative image reconstruction using a total variation constraint. *Magnetic Resonance in Medicine* **57**(6), 1086–1098 (2007)
- [3] Bridson, R.: Fast poisson disk sampling in arbitrary dimensions. *SIGGRAPH sketches* **10**(1), 1 (2007)
- [4] Calivá, F., Leynes, A.P., Shah, R., Upadhyay Bharadwaj, U., Majumdar, S., Larson, P.E.Z., Pedoia, V.: Breaking speed limits with simultaneous ultra-fast MRI reconstruction and tissue segmentation. In: Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C. (eds.) *Proceedings of the Third Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research*, vol. 121, pp. 94–110. PMLR (06–08 Jul 2020), <https://proceedings.mlr.press/v121/caliva20a.html>
- [5] Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory* **52**(2), 489–509 (2006)
- [6] Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications* **14**, 877–905 (2008)
- [7] Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: Monai: An open-source framework for deep learning in healthcare (2022). <https://doi.org/10.48550/ARXIV.2211.02701>, <https://arxiv.org/abs/2211.02701>
- [8] Desai, A.D., Schmidt, A.M., Rubin, E.B., Sandino, C.M., Black, M.S., Mazzoli, V., Stevens, K.J., Boutin, R., Re, C., Gold, G.E., et al.: SKM-TEA: A dataset for accelerated MRI reconstruction with dense image labels for quantitative clinical evaluation. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021)

- [9] Diamond, S., Sitzmann, V., Heide, F., Wetzstein, G.: Unrolled optimization with deep priors. arXiv preprint arXiv:1705.08041 (2017)
- [10] Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302 (1945)
- [11] Donoho, D.L.: Compressed sensing. *IEEE Transactions on information theory* **52**(4), 1289–1306 (2006)
- [12] Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Ali, S., Andrearczyk, V., Aubreville, M., Baid, U., Bakas, S., Balu, N., Bano, S., Bernal, J., Bodenstedt, S., Casella, A., Cheplygina, V., Daum, M., de Bruijne, M., Depeursinge, A., Dorent, R., Egger, J., Ellis, D.G., Engelhardt, S., Ganz, M., Ghatwary, N., Girard, G., Godau, P., Gupta, A., Hansen, L., Harada, K., Heinrich, M.P., Heller, N., Hering, A., Huaiumé, A., Jannin, P., Kavur, A.E., Kodym, O., Kozubek, M., Li, J., Li, H., Ma, J., Martín-Isa, C., Menze, B., Noble, A., Oreiller, V., Padoy, N., Pati, S., Payette, K., Radsch, T., Rafael-Patiño, J., Bawa, V.S., Speidel, S., Sudre, C.H., van Wijnen, K., Wagner, M., Wei, D., Yamlahi, A., Yap, M.H., Yuan, C., Zenk, M., Zia, A., Zimmerer, D., Aydogan, D.B., Bhattarai, B., Bloch, L., Brüngel, R., Cho, J., Choi, C., Dou, Q., Ezhov, I., Friedrich, C.M., Fuller, C.D., Gaire, R.R., Galdran, A., Faura, A.G., Grammatikopoulou, M., Hong, S., Jahanifar, M., Jang, I., Kadkhodamohammadi, A., Kang, I., Kofler, F., Kondo, S., Kuijff, H., Li, M., Luu, M., Martinčić, T., Morais, P., Naser, M.A., Oliveira, B., Owen, D., Pang, S., Park, J., Park, S.H., Plotka, S., Puybureau, E., Rajpoot, N., Ryu, K., Saeed, N., Shephard, A., Shi, P., Štepec, D., Subedi, R., Tochon, G., Torres, H.R., Urien, H., Vilaça, J.A.L., Wahid, K.A., Wang, H., Wang, J., Wang, L., Wang, X., Wiestler, B., Wodzinski, M., Xia, F., Xie, J., Xiong, Z., Yang, S., Yang, Y., Zhao, Z., Maier-Hein, K., Jäger, P.F., Kopp-Schneider, A., Maier-Hein, L.: Why is the winner the best? In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 19955–19966 (June 2023)
- [13] Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., Haase, A.: 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)
- [14] Heckel, R., Jacob, M., Chaudhari, A., Perlman, O., Shimron, E.: Deep learning for accelerated and robust MRI reconstruction. *Magnetic Resonance Materials in Physics, Biology and Medicine* **37**(3), 335–368 (2024)
- [15] Huang, Q., Chen, X., Metaxas, D., Nadar, M.S.: Brain segmentation from k-space with end-to-end recurrent attention network. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. pp. 275–283. Springer International Publishing, Cham (2019)
- [16] Hyun, C.M., Kim, H.P., Lee, S.M., Lee, S., Seo, J.K.: Deep learning for undersampled MRI reconstruction. *Physics in Medicine & Biology* **63**(13), 135007 (2018)
- [17] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: A self-configuring method for deep learning-based biomedical image segmentation (Dec 2020), <https://www.nature.com/articles/s41592-020-01008-z>
- [18] Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 488–498. Springer (2024)
- [19] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)* (2015)
- [20] van Lohuizen, Q., Fransen, S.J.: Naïveseg contribution to the K2S challenge. Described in the K2S Challenge overview paper (2023), see [37]
- [21] Lustig, M., Donoho, D., Pauly, J.M.: Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine* **58**(6), 1182–1195 (2007)
- [22] Lustig, M., Donoho, D.L., Santos, J.M., Pauly, J.M.: Compressed sensing MRI. *IEEE signal processing magazine* **25**(2), 72–82 (2008)
- [23] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications* **9**(1), 5217 (2018)
- [24] Morshuis, J.N., Hein, M., Baumgartner, C.F.: Regseg method contribution to the K2S challenge. Described in the K2S Challenge overview paper (2023), see [37]
- [25] Morshuis, J.N., Hein, M., Baumgartner, C.F.: Segmentation-guided MRI reconstruction for meaningfully diverse reconstructions. In: *MICCAI Workshop on Deep Generative Models*. pp. 180–190. Springer (2024)
- [26] Muckley, M.J., Riemenschneider, B., Radmanesh, A., Kim, S., Jeong, G., Ko, J., Jun, Y., Shin, H., Hwang, D., Mostapha, M., Arberet, S., Nickel, D., Ramzi, Z., Ciuciu, P., Starck, J.L., Teuwen, J., Karkalousos, D., Zhang, C., Sriram, A., Huang, Z., Yakubova, N., Lui, Y.W., Knoll, F.: Results of the 2020 fastMRI challenge for machine learning mr image reconstruction. *IEEE Transactions on Medical Imaging* **40**(9), 2306–2317 (2021). <https://doi.org/10.1109/TMI.2021.3075856>
- [27] Ong, F., Lustig, M.: Sippy: a python package for high performance iterative reconstruction. In: *Proceedings of the ISMRM 27th Annual Meeting*, Montreal, Quebec, Canada. vol. 4819 (2019)
- [28] Oztek, M.A., Brunnquell, C.L., Hoff, M.N., Boulter, D.J., Mossa-Basha, M., Beauchamp, L.H., Haynor, D.L., Nguyen, X.V.: Practical considerations for radiologists in implementing a patient-friendly mri experience. *Topics in Magnetic Resonance Imaging* **29**(4), 181–186 (2020)
- [29] Pruessmann, K.P., Weiger, M., Scheidegger, M.B., Boesiger, P.: Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **42**(5), 952–962 (1999)
- [30] Radmanesh, A., Muckley, M.J., Murrell, T., Lindsey, E., Sriram, A., Knoll, F., Sodickson, D.K., Lui, Y.W.: Exploring the acceleration limits of deep learning variational network-based two-dimensional brain MRI. *Radiology: Artificial Intelligence* **4**(6) (Nov 2022). <https://doi.org/https://doi.org/10.1148/ryai.210313>
- [31] Razumov, A., Dylov, D.: Cs-rec method contributed to the K2S challenge. Described in the K2S Challenge overview paper (2023), see [37]
- [32] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
- [33] Sandino, C.M., Cheng, J.Y., Chen, F., Mardani, M., Pauly, J.M., Vasanawala, S.S.: Compressed sensing: From research to clinical practice with deep neural networks: Shortening scan times for magnetic resonance imaging. *IEEE signal processing magazine* **37**(1), 117–127 (2020)
- [34] Schlemper, J., Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V., Rueckert, D.: Cardiac mr segmentation from undersampled k-space using deep latent representation learning. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. pp. 259–267. Springer International Publishing, Cham (2018)
- [35] Sui, B., Lv, J., Tong, X., Li, Y., Wang, C.: Simultaneous image reconstruction and lesion segmentation in accelerated MRI using multitasking learning. *Medical Physics* **48**(11), 7189–7198 (Nov 2021). <https://doi.org/https://doi.org/10.1002/mp.15213>, <https://pubmed.ncbi.nlm.nih.gov/34542180/>
- [36] Sun, L., Fan, Z., Ding, X., Huang, Y., Paisley, J.: Joint CS-MRI reconstruction and segmentation with a unified deep network. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) *Information Processing in Medical Imaging*. pp. 492–504. Springer International Publishing, Cham (2019)
- [37] Tolpadi, A.A., Bharadwaj, U., Gao, K.T., Bhattacharjee, R., Gassert, F.G., Luitjens, J., Giesler, P., Morshuis, J.N., Fischer, P., Hein, M., Baumgartner, C.F., Razumov, A., Dylov, D., Lohuizen, Q.V., Fransen, S.J., Zhang, X., Tibrewala, R., de Moura, H.L., Liu, K., Zibetti, M.V.W., Regatte, R., Majumdar, S., Pedoia, V.: K2s challenge: From undersampled k-space to automatic segmentation. *Bioengineering* **10**(2) (2023). <https://doi.org/10.3390/bioengineering10020267>, <https://www.mdpi.com/2306-5354/10/2/267>
- [38] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004). <https://doi.org/10.1109/TIP.2003.819861>
- [39] Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A.: Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific reports* **11**(1), 2369 (2021)
- [40] Wilcoxon, F.: Individual comparisons by ranking methods. Springer (1992)
- [41] Wirth, W., Eckstein, F., Kemnitz, J., Baumgartner, C.F., Konukoglu, E., Fuerst, D., Chaudhari, A.S.: Accuracy and longitudinal reproducibility of quantitative femorotibial cartilage measures derived from automated u-net-based segmentation of two different mri contrasts: Data from the osteoarthritis initiative healthy reference cohort - magnetic resonance materials in physics, biology and medicine (Oct 2020), <https://link.springer.com/article/10.1007/s10334-020-00889-7>

- [42] Zaitsev, M., Maclaren, J., Herbst, M.: Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging* **42**(4), 887–901 (2015)

Appendix C

Segmentation-guided MRI Reconstruction for Meaningfully Diverse Reconstructions

Source Publication

Publication:

Jan Nikolas Morshuis, Matthias Hein, and Christian F Baumgartner. “Segmentation-guided MRI Reconstruction for Meaningfully Diverse Reconstructions”. In: *MICCAI Workshop on Deep Generative Models*. Springer. 2024, pp. 180–190. DOI: https://doi.org/10.1007/978-3-031-72744-3_18



Segmentation-Guided MRI Reconstruction for Meaningfully Diverse Reconstructions

Jan Nikolas Morshuis¹(✉), Matthias Hein¹, and Christian F. Baumgartner^{1,2}

¹ University of Tübingen, Tübingen, Germany
`nikolas.morshuis@uni-tuebingen.de`

² University of Lucerne, Lucerne, Switzerland

Abstract. Inverse problems, such as accelerated MRI reconstruction, are ill-posed and an infinite amount of possible and plausible solutions exist. This may not only lead to uncertainty in the reconstructed image but also in downstream tasks such as semantic segmentation. This uncertainty, however, is mostly not analyzed in the literature, even though probabilistic reconstruction models are commonly used. These models can be prone to ignore plausible but unlikely solutions like rare pathologies. Building on MRI reconstruction approaches based on diffusion models, we add guidance to the diffusion process during inference, generating two meaningfully diverse reconstructions corresponding to an upper and lower bound segmentation. The reconstruction uncertainty can then be quantified by the difference between these bounds, which we coin the ‘uncertainty boundary’. We analyzed the behavior of the upper and lower bound segmentations for a wide range of acceleration factors and found the uncertainty boundary to be both more reliable and more accurate compared to repeated sampling. Code is available at <https://github.com/NikolasMorshuis/SGR>.

1 Introduction

Accelerated MRI reconstruction has substantially improved in recent years through the use of machine learning methods. Currently, diffusion models (DMs) [13, 19] achieve state-of-the-art MRI reconstruction scores [4, 5, 14]. An often neglected area in the literature is the inherent uncertainty of the reconstruction process. Accelerated MRI reconstruction is an ill-posed problem and infinite plausible reconstructions exist for an undersampled MRI image. However, in the majority of recent work only a single point estimate (a single reconstruction) of the solution space is provided. This common practice is problematic in safety-critical applications like medical imaging, as it does not accurately reflect the inherent uncertainty and no estimation of the confidence of the solution can be provided.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72744-3_18.

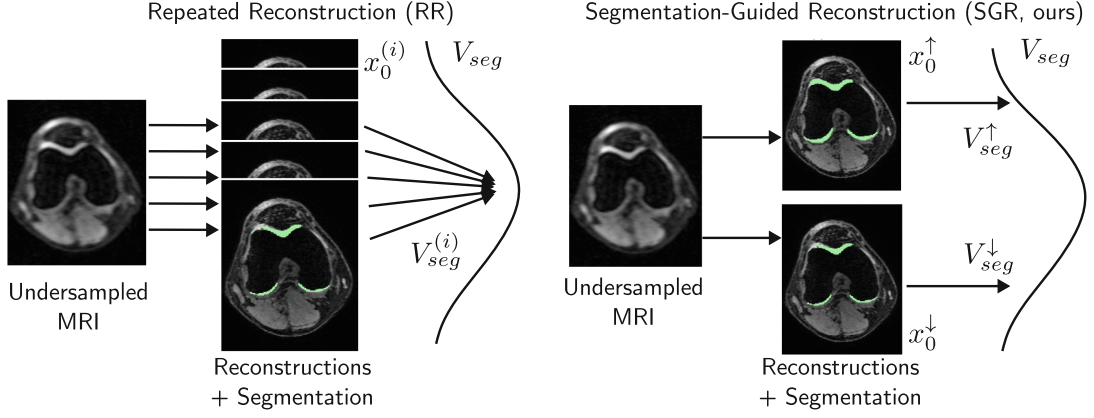


Fig. 1. Motivation. Repeated reconstruction tends to create MRI reconstructions that are conceptually similar and not meaningfully diverse. Similar results are also observed in downstream tasks like segmentation in our experiments. Our method in contrast only reconstructs two images, an upper bound ($x_0^\uparrow$) and a lower bound ($x_0^\downarrow$) reconstruction, corresponding to an upper ($V_{seg}^\uparrow$) and lower bound ($V_{seg}^\downarrow$) on the segmentation volume, giving an intuitive understanding of the segmentation uncertainty.

A common approach to quantify the reconstruction uncertainty is to reconstruct a large number of samples and analyze their pixel-wise variance. This kind of uncertainty prediction has been investigated with probabilistic networks [3, 12], Monte Carlo dropout [16] ensembling [15] or DMs [5, 14, 24].

Recently however, [6] has shown that Repeated Reconstruction (RR) in natural image reconstruction with DMs leads to similar images which display a lack of meaningful diversity. Even though the diffusion model is trained to model the posterior distribution accurately, the heavy tail of the true posterior distribution is often largely ignored in practice and plausible yet unlikely solutions with different semantic meaning are not found. This is a problem for medical imaging, where certain pathologies might be rare and unlikely. However, it is important to recover these cases precisely and include them as possible solutions when reconstructing multiple samples for uncertainty prediction.

The definition of meaningful diversity can depend on the downstream task and might be different for e.g. pathology detection (a pathology exists or not) or semantic segmentation (the segmented tissue might be smaller or larger). In this work we focus on diversity of semantic segmentation, as it is an important task in clinical practice and the results are more easily interpretable for non-experts.

In this paper, instead of random sampling, we propose to reconstruct two MRI-images per segmentation class corresponding to the upper and lower bound segmentation volume. To this end, we introduce a novel method called Segmentation Guided Reconstruction (SGR). SGR builds on DM-based MRI reconstruction. However, rather than standard sampling, we guide the reconstruction process with the gradient of segmentation losses that either penalize small segmentation volumes or large segmentation volumes. The resulting lower and upper bound reconstructions and segmentations are an intuitive way of understanding the segmentation uncertainty which arises due to the ill-posed nature of the

reconstruction problem. See Fig. 1 for a visual explanation of the motivation. In summary, our contributions are:

- We introduce a novel method Segmentation-Guided Reconstruction (SGR) to guide the reconstruction process of undersampled MRI data, leading to meaningfully diverse reconstructions.
- We show that the diverse reconstructions lead to diverse segmentations that correspond to an upper and lower bound on the segmentation. We introduce the concept of ‘uncertainty boundary’ for the volume between both bounds.
- We show that our method captures the inherent uncertainty more faithfully than the commonly used Repeated Reconstruction (RR), potentially leading to safer MRI reconstruction methods

2 Related Work

Diffusion based inverse problem solvers have seen large success in MRI reconstruction [5, 14, 18, 24], achieving state-of-the-art reconstruction scores [4, 5, 14]. In contrast to supervised learning-based methods [20, 25], trained DMs can be used for several acceleration factors. Recent reconstruction methods [4] allow for more than $80\times$ faster inference time, reducing the computational burden for RR. With RR it is possible to reconstruct several distinct images, allowing to estimate the reconstruction uncertainty by analyzing the variance of the voxel values. However, the distinct images do not necessarily display meaningful diversity [6].

While the reconstruction and segmentation problem are mostly tackled separately, some work exists aiming to improve segmentation results by combining reconstruction and segmentation in an end-to-end fashion [1, 2, 21–23]. Wu et al. [23] show that fine-tuning a pre-trained reconstruction network with a task-specific network head leads to improved segmentation results. Similarly, the winning method of the K2S challenge [22] also first trains a reconstruction network before fine-tuning the network for segmentation. These approaches, however, lead to degraded reconstructions. Moreover, all of these methods lack an analysis of the effects of the inherent reconstruction uncertainty on segmentation.

The literature on reconstruction uncertainty propagation to downstream tasks is relatively sparse and most methods utilize repeated sampling of probabilistic networks that might not offer meaningful diversity [6]. Feiner et al. [11] propose to utilize Monte Carlo sampling to estimate the reconstruction uncertainty that can then be propagated to classification tasks. Cheung et al. [3] use repeated sampling of CT reconstructions to obtain a distribution of several downstream metrics, which they use to calculate statistically valid prediction sets using conformal prediction. Fischer et al. [12] are the first to consider the propagation of the aleatoric uncertainty due to the reconstruction process in the context of segmentation. By repeated sampling of a probabilistic reconstruction network and subsequent segmentation, they were able to quantify the variance of the segmentation at pixel-level. However, data-consistency was not enforced and no upper or lower bound of the segmentation has been estimated.

3 Method

3.1 Background Diffusion Models and MRI-Reconstruction

In the inverse problem of MRI reconstruction, we aim to reconstruct a plausible image $\mathbf{x}_0$ that is coherent with the measurement data $\mathbf{y}$ when passing through a forward operator (i.e. Fourier Transform + Masking) $\mathbf{A}$, such that the following equation holds:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0^{(i)} . \quad (1)$$

The problem is typically ill-posed and infinite $\mathbf{x}_0^{(i)}$ exist that are possible solutions for Eq. 1. In recent years DMs have shown to perform exceptionally well on the MRI-reconstruction tasks, being able to sample a set of distinct solutions $\mathbf{x}_0^{(i)}$ from the posterior.

DMs generate samples from the data distribution $p_{data}(\mathbf{x}_0)$ by reversing a diffusion process from timestep $t = 0$ to $t = T$ described by a conditional density

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (2)$$

with a pre-defined defined $\bar{\alpha}_t$ schedule.

A DM can then be trained by predicting the noise through epsilon matching:

$$\min_{\theta} \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0), \mathbf{x}_0 \sim p_{data}(\mathbf{x}_0), \epsilon \sim \mathcal{N}(0, \mathbf{I})} (\|\epsilon_{\theta}^{(t)}(\mathbf{x}_t) - \epsilon\|_2^2) . \quad (3)$$

The learned ϵ -matching function can then be used to predict the clean image $\hat{\mathbf{x}}_{0|t}$ at every step t of the diffusion process using Tweedie denoising [10]:

$$\hat{\mathbf{x}}_{0|t} = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_{\theta}^{(t)}(\mathbf{x}_t))/\sqrt{\bar{\alpha}_t} . \quad (4)$$

Recently, Chung et al. [4] introduced a DM based reconstruction method that utilizes conjugate gradient (CG) optimization: During the CG-optimization process, starting from $\hat{\mathbf{x}}_{0|t}$ we search a $\hat{\mathbf{x}}'_{0|t}$ that minimizes the distance $\mathbf{A}^*\mathbf{A}\hat{\mathbf{x}}'_{0|t} - \mathbf{A}^*\mathbf{y}$ with $\mathbf{A}^*$ being the backward operator mapping from the measurement space back to the image space:

$$\hat{\mathbf{x}}'_{0|t} = CG(\mathbf{A}^*\mathbf{A}, \mathbf{A}^*\mathbf{y}, \hat{\mathbf{x}}_{0|t}) . \quad (5)$$

The update rule using the DDIM method [17] is then given as follows:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}'_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \eta^2\tilde{\beta}_t^2}\epsilon_{\theta}^{(t)}(\mathbf{x}_t) + \eta\tilde{\beta}_t\epsilon . \quad (6)$$

We modify the algorithm from [4] slightly to achieve full data-consistency. For the last diffusion step $t = 1 \rightarrow t = 0$, we replace Eq. 5 and make sure that the reconstructed image $\mathbf{x}_0$ is fully data-consistent using $\mathbf{y}$ and Eq. 1.

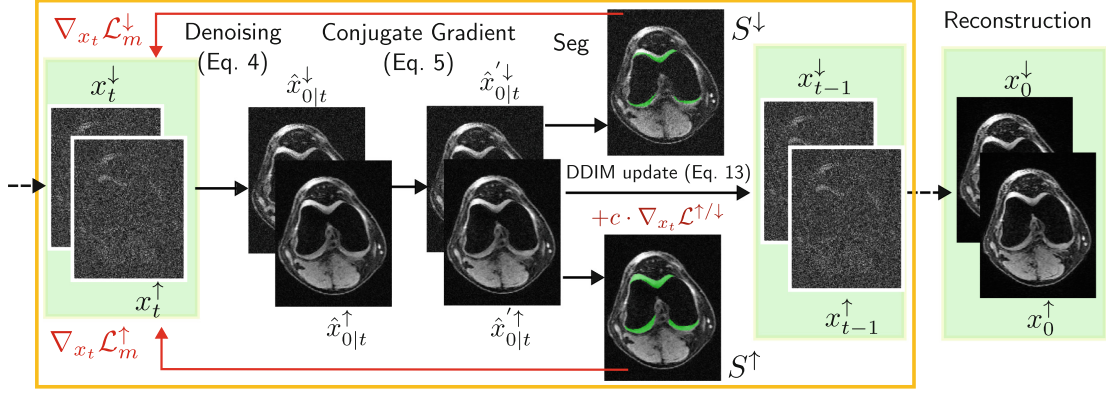


Fig. 2. Method explanation. For each inverse diffusion step t , we calculate $\nabla_{x_t} \mathcal{L}^{\uparrow/\downarrow}$ and include the gradient in the calculation of x_{t-1} , in order to increase (decrease) the segmentation volume and to get an upper- and lower-bound segmentation ($S^\uparrow$ and $S^\downarrow$).

3.2 Segmentation Guidance for Diverse Sampling

In order to generate a set of meaningfully diverse solutions $\mathbf{x}_0^{(i)}$, we reconstruct two images per segmentation class ($\mathbf{x}_0^\uparrow$ and $\mathbf{x}_0^\downarrow$) with corresponding upper and a lower bound segmentations ($S^\uparrow$ and $S^\downarrow$) as well as segmentation volumes ($V_{seg}^\uparrow$ and $V_{seg}^\downarrow$). The volume between $S^\downarrow$ and $S^\uparrow$ is the uncertainty boundary S_{unc} . We guide the reconstruction process of the diffusion model using segmentation guidance. To achieve this, we propose two loss-functions $\mathcal{L}^\uparrow$ and $\mathcal{L}^\downarrow$ that are suitable for segmentation-based model guidance.

Let $p_\phi^c(\mathbf{x}_{0,jk})$ be the predicted probability of pixel (j,k) in the image $\mathbf{x}_0$ belonging to the segmentation class $c \in \mathcal{C}$, and let p_ϕ be a segmentation network. We abbreviate the term as $\mathbf{p}_{jk}^{(c)}$. The volume of c can be calculated by summing the number of pixels that belong to c multiplied by voxel-volume V_{voxel} :

$$V_c(\mathbf{x}_0) = \sum_{j,k} (\mathbb{I}_c(\mathbf{p}_{jk}^{(c)})) \cdot V_{voxel}, \quad (7)$$

where $\mathbb{I}_c$ is an indicator function

$$\mathbb{I}_c = \begin{cases} 1 & \text{if } \mathbf{p}_{jk}^{(c)} > \mathbf{p}_{jk}^{(c')} \forall c' \in \mathcal{C}, c' \neq c \\ 0 & \text{otherwise} \end{cases}. \quad (8)$$

During optimization, we consider only one class at a time, such that we can simplify the objective to a binary segmentation class. We build our guidance-loss based on the widely-used Binary-Cross-Entropy loss:

$$\mathcal{L}_{BCE}(\mathbf{y}, \mathbf{p}, c) = - \sum_{j,k} \left(\mathbf{y}_{jk} \log(\mathbf{p}_{jk}^{(c)}) + (1 - \mathbf{y}_{jk}) \log(1 - \mathbf{p}_{jk}^{(c)}) \right). \quad (9)$$

We can maximize the optimization objective $\mathcal{L}^\uparrow$ to find the upper bound segmentation, as well as $\mathcal{L}^\downarrow$ to find the lower bound segmentation:

$$\mathcal{L}^\uparrow = \mathcal{L}_{BCE}(\mathbf{0}, \mathbf{p}, c) = - \sum_{j,k} \log(1 - \mathbf{p}_{jk}^{(c)}), \quad (10)$$

$$\mathcal{L}^\downarrow = \mathcal{L}_{BCE}(\mathbf{1}, \mathbf{p}, c) = - \sum_{j,k} \log(\mathbf{p}_{jk}^{(c)}). \quad (11)$$

The binary cross-entropy loss is unbounded, meaning that pixel probabilities are further optimized, even if the target class has already been reached. Following Croce et al. [7] we can avoid this behavior by using a masked cross-entropy loss, which excludes pixels already segmented as the target class from the optimization process. The optimization objectives can therefore be written as follows:

$$\mathcal{L}_m^\downarrow = \mathbb{I}_c \mathcal{L}^\downarrow; \quad \mathcal{L}_m^\uparrow = (1 - \mathbb{I}_c) \mathcal{L}^\uparrow. \quad (12)$$

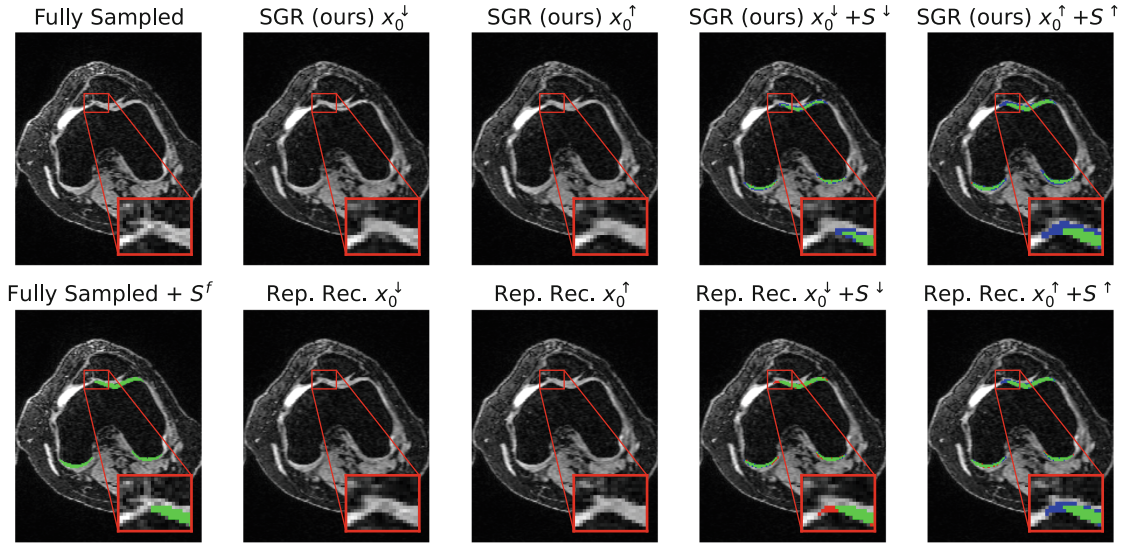


Fig. 3. Sampling vs. adversarial guidance. Examples of lower- ($x_0^\downarrow$) and upper-bound ($x_0^\uparrow$) reconstructions (16x acc.) and segmentations ($S^\downarrow$, $S^\uparrow$) using our SGR method and the standard RR method. (S^f is the segmentation of the fully-sampled image. Green: $S^f \& S^\uparrow$ or $S^f \& S^\downarrow$, Blue: $S^f > S^\downarrow$ or $S^f < S^\uparrow$, Red: $S^f < S^\downarrow$ or $S^f > S^\uparrow$)

The derived optimization objective can be used during the reverse diffusion process. We therefore calculate the gradient of $\mathcal{L}_m^\uparrow$ (or $\mathcal{L}_m^\downarrow$, respectively) with respect to x_t at every step t , to find the perturbation necessary to increase the optimization loss. We add the gradient $\nabla_{x_t} \mathcal{L}_m^\uparrow$ (or $\nabla_{x_t} \mathcal{L}_m^\downarrow$, respectively) during the DDIM update in Eq. 6:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}'_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \eta^2 \tilde{\beta}_t^{(t)}(\boldsymbol{\epsilon}_\theta^{(t)}(x_t) + \gamma \nabla_{x_t} \mathcal{L}_m^{\uparrow/\downarrow}) + \eta \tilde{\beta}_t \boldsymbol{\epsilon}}, \quad (13)$$

where the parameter γ makes sure that the l_2 norm of $\nabla_{x_t} \mathcal{L}_m^{\uparrow/\downarrow}$ stays small with respect to ϵ_θ . We set γ to the following value:

$$\gamma = \begin{cases} b \cdot \|\epsilon_\theta^{(t)}\|_2 / \|\nabla_{x_t} \mathcal{L}_m^{\uparrow/\downarrow}\|_2 & \text{if } \|\nabla_{x_t} \mathcal{L}_m^{\uparrow/\downarrow}\|_2 > b \|\epsilon_\theta^{(t)}\|_2 \\ 1 & \text{otherwise,} \end{cases} \quad (14)$$

where b is a small constant that we set to $b = 0.005$ in our experiments. An ablation study on the effects of the parameter value can be found in the Appendix. A summary of our method is shown in Fig. 2.

For the baseline RR, we sample 16 images per slice and select $x_0^\uparrow$ and $x_0^\downarrow$ for every segmentation class as the reconstruction leading to the highest segmentation volume ($V_{seg}^\uparrow$) and lowest segmentation volume ($V_{seg}^\downarrow$), respectively.

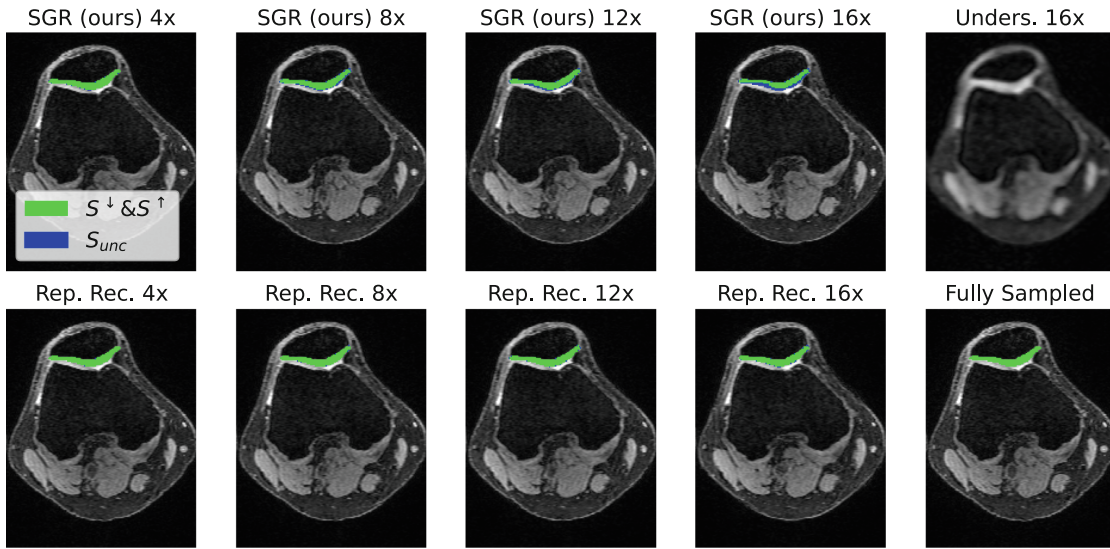


Fig. 4. Analysis of different acceleration factors. For RR, we see that the uncertainty does not increase, even though very high accelerations are tested. Our method, in contrast, generates more reliable uncertainty boundaries.

3.3 Data and Models

We train and test our method on the publicly available SKM-TEA [8] dataset and focus only on the $E1$ -echoes. As the spacing of the SKM-TEA data is very fine-grained (0.3125 mm in the axial (=sampling) direction), neighboring images tend to be similar. We therefore save computational resources during testing by only considering every 8th slice in the relevant regions, corresponding to a spacing of 2.5 mm between the slices. In total we analyze 869 slices for each acceleration factor and each method.

All experiments are based on the same diffusion model. We train the diffusion model on the fully-sampled images of the SKM-TEA data. We normalize the data by dividing by the 0.99 quantile of every slice. The diffusion model is based

on the one introduced in [9]. We train the model for 1 million steps using a batch-size of 2 on a 3080TI GPU. For segmentation, we train a U-Net on the fully-sampled MRI-data in axial direction. Due to the strong class-imbalance on the axial imaging plane, where only a minority of slices show elements of tibial cartilage or meniscus as these parts tend to be relatively flat, we perform a foreground oversampling of 40%. To reduce overfitting, we apply mirroring and Gaussian Noise augmentations during training.

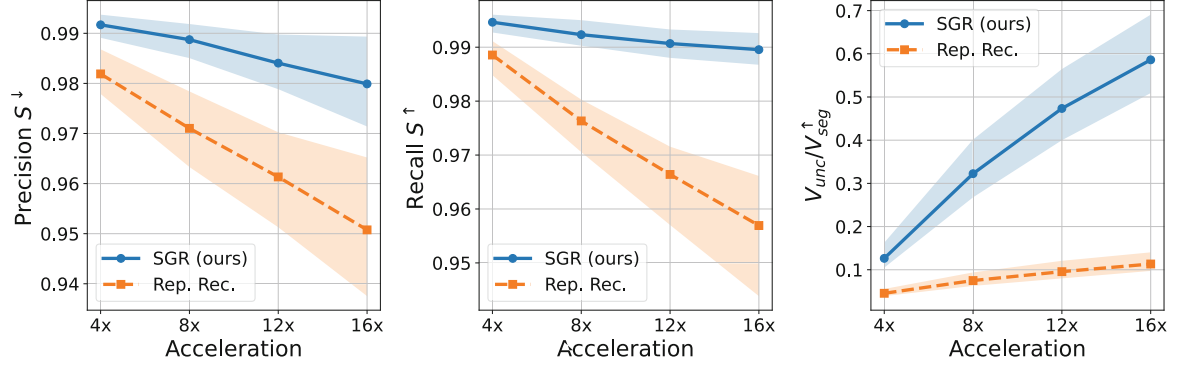


Fig. 5. Results. Median score and percentile boundaries (25%, 75%) for different acceleration factors. Note that the upper and lower bound segmentation of our method are mostly correctly capturing the inherent uncertainty even for high acceleration factors: we have high precision (small amount of false-positives in $S^\downarrow$) and recall (small amount of false-negatives in $S^\uparrow$). The ratio of uncertain volume $V_{unc} = V_{seg}^\uparrow - V_{seg}^\downarrow$ over $V_{seg}^\uparrow$ also increases, better reflecting the true underlying uncertainty.

4 Experiments and Results

We perform experiments on 4 acceleration factors: $4\times$, $8\times$, $12\times$, and $16\times$, analyzing the behavior of the uncertainty boundaries of SGR and RR. We assess the quality of the generated segmentations $S^\downarrow$ and $S^\uparrow$ by calculating the precision for $S^\downarrow$, recall for $S^\uparrow$ and the uncertainty volume V_{unc} . We focus on precision on $S^\downarrow$, as we want to avoid overestimating the segmentation and creating false-positives, and measure recall for $S^\uparrow$, because we want to avoid segmentations containing false-negatives. An example image where RR overestimates $S^\downarrow$ while our method predicts accurate boundaries is shown in Fig. 3. A similar observation can be found in Fig. 5, where our method sustains high precision and recall values, while these scores degrade with higher acceleration for the baseline method RR. Reconstruction quality as measured by SSIM and PSNR remains very high for SGR compared to RR, as shown in Table 1.

In Fig. 4, we see that the uncertainty boundary increases with higher acceleration for our method, while it stays approximately constant for the baseline method RR. This result is also shown in Fig. 5, where the ratio $V_{unc}/V_{seg}^\uparrow$ increases faster for SGR compared to RR. Given the relatively low precision

and recall scores for RR and the higher scores for our method, one can conclude that RR underestimates the uncertainty boundaries, while this problem is largely reduced for the uncertainty boundaries estimated by SGR.

5 Conclusion and Discussion

We have introduced a novel method for segmentation-guided MRI reconstruction and have shown that it can generate meaningfully diverse reconstructions that lead to upper and lower bound segmentations. These segmentation can offer an intuitive way of understanding the segmentation uncertainty that is caused by the ill-posed nature of the reconstruction problem. As our method also offers more reliable uncertainty boundaries compared to the standard method of repeated reconstruction, we believe that it can be helpful in clinical practice, whenever a trustworthy uncertainty estimation is required.

Table 1. Quantitative results for experiments for different accelerations. All reported segmentation metrics are calculated on the 3D-volumes of all 4 segmentation classes. Median values are reported. SSIM and PSNR values are calculated as the average over all reconstructed volumes. Note that it is not clear if larger or smaller ratios of $V_{unc}/V^\uparrow$ are preferable, due to the inherent uncertainty of the reconstruction problem.

Acc.	Method	Segmentation			Reconstruction	
		Prec. $S^\downarrow$	Recall $S^\uparrow$	$V_{unc}/V^\uparrow$	SSIM	PSNR
4x	SGR (ours)	0.992	0.995	0.126	0.914	34.6
	RR	0.982	0.989	0.045	0.915	34.6
8x	SGR (ours)	0.989	0.992	0.322	0.857	30.8
	RR	0.971	0.976	0.075	0.857	30.9
12x	SGR (ours)	0.984	0.991	0.473	0.818	29.1
	RR	0.961	0.966	0.096	0.820	29.1
16x	SGR (ours)	0.980	0.990	0.586	0.788	27.9
	RR	0.951	0.957	0.113	0.789	28.0

In future work, we want to analyze if robust segmentation networks introduced in [7] can help to reduce V_{unc} while maintaining high precision and recall scores. In this work, however, we have focused on more commonly used neural networks, to show the easy applicability of our method for existing pipelines.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Jan Nikolas Morshuis.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Acar, M., Çukur, T., Öksüz, İ.: Segmentation-aware mri reconstruction. In: *Machine Learning for Medical Image Reconstruction*. pp. 53–61. Springer International Publishing, Cham (2022)
2. Calivá, F., Leynes, A.P., Shah, R., Upadhyay Bharadwaj, U., Majumdar, S., Larson, P.E.Z., Padoia, V.: Breaking speed limits with simultaneous ultra-fast mri reconstruction and tissue segmentation. In: *MIDL. Proceedings of Machine Learning Research*, PMLR (06–08 Jul 2020)
3. Cheung, M.Y., Netherton, T.J., Court, L.E., Veeraraghavan, A., Balakrishnan, G.: Metric-guided image reconstruction bounds via conformal prediction. *arXiv preprint [arXiv:2404.15274](https://arxiv.org/abs/2404.15274)* (2024)
4. Chung, H., Lee, S., Ye, J.C.: Decomposed diffusion sampler for accelerating large-scale inverse problems. In: *ICLR* (2024)
5. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated mri. *Medical Image Analysis* p. 102479 (2022)
6. Cohen, N., Manor, H., Bahat, Y., Michaeli, T.: From posterior sampling to meaningful diversity in image restoration. In: *ICLR* (2024)
7. Croce, F., Singh, N.D., Hein, M.: Towards reliable evaluation and fast training of robust semantic segmentation models. In: *ECCV* (2024)
8. Desai, A.D., Schmidt, A.M., Rubin, E.B., Sandino, C.M., Black, M.S., Mazzoli, V., Stevens, K.J., Boutin, R., Re, C., Gold, G.E., et al.: Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. In: *NeurIPS Datasets and Benchmarks Track (Round 2)* (2021)
9. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: *NeurIPS* (2021)
10. Efron, B.: Tweedie’s formula and selection bias. *Journal of the American Statistical Association* **106**(496), 1602–1614 (2011)
11. Feiner, L.F., Menten, M.J., Hammernik, K., Hager, P., Huang, W., Rueckert, D., Braren, R.F., Kaissis, G.: Propagation and Attribution of Uncertainty in Medical Imaging Pipelines, pp. 1–11. Springer Nature Switzerland, Cham (2023)
12. Fischer, P., Thomas, K., Baumgartner, C.F.: Uncertainty estimation and propagation in accelerated mri reconstruction. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. pp. 84–94. Springer Nature Switzerland, Cham (2023)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *NeurIPS* (2020)
14. Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A., Tamir, J.: Robust compressed sensing MRI with deep generative priors. In: *NeurIPS* (2021)
15. Küstner, T., Hammernik, K., Rueckert, D., Hepp, T., Gatidis, S.: Predictive uncertainty in deep learning-based mr image reconstruction using deep ensembles: Evaluation on the fastmri data set. *Magnetic Resonance in Medicine* (2024)
16. Schlemper, J., Castro, D.C., Bai, W., Qin, C., Oktay, O., Duan, J., Price, A.N., Hajnal, J., Rueckert, D.: Bayesian deep learning for accelerated mr image reconstruction. In: *Machine Learning for Medical Image Reconstruction*. pp. 64–71. Springer International Publishing, Cham (2018)
17. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: *ICLR* (2021)
18. Song, Y., Shen, L., Xing, L., Ermon, S.: Solving inverse problems in medical imaging with score-based generative models. In: *ICLR* (2022)

19. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
20. Sriram, A., Zbontar, J., Murrell, T., Defazio, A., Zitnick, C.L., Yakubova, N., Knoll, F., Johnson, P.: End-to-end variational networks for accelerated mri reconstruction. In: MICCAI. Springer (2020)
21. Sui, B., Lv, J., Tong, X., Li, Y., Wang, C.: Simultaneous image reconstruction and lesion segmentation in accelerated mri using multitasking learning. *Medical Physics* **48**(11), 7189–7198 (2021)
22. Tolpadi, A.A., Bharadwaj, U., Gao, K.T., Bhattacharjee, R., Gassert, F.G., Luitjens, J., Giesler, P., Morshuis, J.N., Fischer, P., Hein, M., et al.: K2s challenge: From undersampled k-space to automatic segmentation. *Bioengineering* **10**(2), 267 (2023)
23. Wu, Z., Yin, T., Sun, Y., Frost, R., van der Kouwe, A., Dalca, A.V., Bouman, K.L.: Learning task-specific strategies for accelerated mri. *IEEE Transactions on Computational Imaging* (2024)
24. Xie, Y., Li, Q.: Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In: MICCAI. Springer Nature Switzerland, Cham (2022)
25. Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al.: fastmri: An open dataset and benchmarks for accelerated mri. arXiv preprint [arXiv:1811.08839](https://arxiv.org/abs/1811.08839) (2018)

Appendix D

Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions

Source Publication

Publication:

Jan Nikolas Morshuis, Christian Schlarmann, Thomas Küstner, Christian F. Baumgartner[†], and Matthias Hein[†]. “Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions”. In: *MICCAI*. Cham: Springer, 2025, pp. 356–366. DOI: https://doi.org/10.1007/978-3-032-04937-7_34. [†]Shared last authorship.



Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions

Jan Nikolas Morshuis^{1(✉)}, Christian Schlarmann¹, Thomas Küstner^{1,2},
Christian F. Baumgartner^{1,3}, and Matthias Hein¹

¹ University of Tübingen, Tübingen, Germany

nikolas.morshuis@uni-tuebingen.de

² University Hospital of Tübingen, Tübingen, Germany

³ University of Lucerne, Lucerne, Switzerland

Abstract. In recent years, accelerated MRI reconstruction based on deep learning has led to significant improvements in image quality with impressive results for high acceleration factors. However, from a clinical perspective image quality is only secondary; much more important is that all clinically relevant information is preserved in the reconstruction from heavily undersampled data. In this paper, we show that existing techniques, even when considering resampling for diffusion-based reconstruction, can fail to reconstruct small and rare pathologies, thus leading to potentially wrong diagnosis decisions (false negatives). To uncover the potentially missing clinical information we propose “Semantically Diverse Reconstructions” (SDR), a method which, given an original reconstruction, generates novel reconstructions with enhanced semantic variability while all of them are fully consistent with the measured data. To evaluate SDR automatically we train an object detector on the fastMRI+ dataset. We show that SDR significantly reduces the chance of false-negative diagnoses (higher recall) and improves mean average precision compared to the original reconstructions. The code is available on <https://github.com/NikolasMorshuis/SDR>

Keywords: MRI Reconstruction · Detection · Uncertainty

1 Introduction

Accelerated MRI reconstruction from undersampled k-space data is an inverse problem with infinitely many solutions. However, it is common to reconstruct only a single solution, thereby ignoring potentially different yet still plausible reconstructions. Deep-learning-based approaches, including novel techniques based on diffusion models, have led to impressive reconstruction results in terms

C. F. Baumgartner and M. Hein—Shared last authorship.

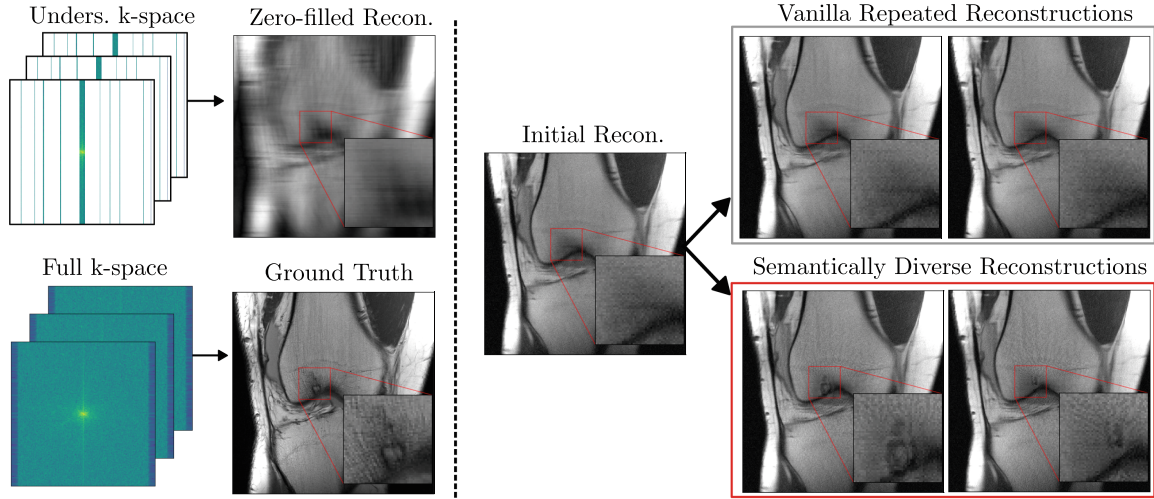


Fig. 1. Method overview. At high acceleration factors (12x shown here), naive MRI reconstruction suffers from artifacts (left). Recent reconstruction methods like DDIP [4] can recover image details and generate multiple reconstructions, but these often lack diversity, potentially leading to missed pathologies (top right). Our proposed SDR approach produces semantically diverse reconstructions, helping detect pathologies that might otherwise be missed (bottom right).

of image quality even for high acceleration factors. On the other hand, in clinical practice, some pathological tissue changes can be small and may occur rarely, making them potentially underrepresented in the training data. As a result, overconfident reconstruction models may favor more likely solutions of healthy tissue if consistent with the measurement data, potentially leading to missed pathologies. Missing clinically relevant pathologies in the reconstruction can be a major obstacle to the application of accelerated MRI in clinical practice.

While some recent reconstruction methods such as diffusion-models [3, 4, 11] are capable of generating multiple different solutions by repeated sampling, it has been shown that these models tend to sample the most likely solutions frequently but only rarely sample less likely solutions that might still be diagnostically relevant [5] (see Fig. 1 top right). Moreover, sampling multiple reconstructions can be time-consuming as it requires repeating the denoising process.

To address this problem, we propose a technique to generate semantically diverse reconstructions (SDR) that are fully consistent with the measured k-space data. Our SDR approach can be applied post-hoc to initial reconstructions obtained using any reconstruction method, producing a set of alternative solutions that preserve data consistency while potentially revealing pathologies not present in the original reconstruction. This offers the unique possibility to study a diverse set of data-consistent solutions and their inherent uncertainty. Understanding the inherent uncertainty can improve the safety of the reconstruction process because the risk of a missed pathology and therefore a false negative diagnosis is reduced. If high uncertainty prevents a confident diagnosis from being made, additional measurements can be acquired. Our method generates these alternative solutions in ≈ 3 seconds per image, which is faster

than the sampling time of diffusion models that can require several minutes per image [11]. This fast inference time could potentially enable an application of our method during the scanning process in the future.

Our technique relies on identifying candidate regions in reconstructed images and modifying them to alter their semantic content while ensuring consistency with the measured k-space data. To generate semantically diverse reconstructions, we develop an adversarially robust version of the detection algorithm ViTDet [14], leveraging techniques from [20] for the ViT-Backbone. We then use the features of candidate regions to maximize the semantic distance from the existing reconstructions (see Fig. 1 bottom right). This is motivated by the observation that the Euclidean distance in the feature space of adversarially robust vision encoders is a good semantic metric and that maximizing distance in this feature space leads to semantically meaningful changes [6]. We explore two versions of our method: i) an automatic selection of regions to be optimized using the proposal network of the object detector, ii) manual annotation of a candidate image region by the human user, e.g. a doctor.

In experiments on the fastMRI+ dataset [27], we demonstrate that our method is able to produce a semantically diverse set of possible MRI reconstructions that capture more possible pathologies compared to alternative sampling techniques based on diffusion models [4] and the E2E-Varnet [22]. We quantitatively verify this by showing that pathology recall – obtained with a separate Faster-RCNN detector [18] – outperforms the baseline techniques, while mean average precision is not negatively affected.

2 Related Work

The generation of multiple solutions in deep-learning-based MRI reconstructions is a relatively recent area of research. Morshuis et al. [17] proposed a method to produce diverse MRI-reconstructions corresponding to minimum and maximum segmentations in a downstream task. This approach estimates the segmentation uncertainty due to the ill-posed nature of the reconstruction problem. Küstner et al. [13] trained models with different initial seeds to generate multiple reconstructions, which were then used to estimate reconstruction uncertainty. They observed that uncertainty was highest in image regions containing pathologies. Recent studies have demonstrated the strong performance of diffusion models in MR reconstruction [3, 4, 11, 25]. Apart from producing high-quality reconstructions, these techniques can also be used to generate multiple solutions by repeated sampling from the posterior distribution of images that are consistent with the measurement data.

Only few works have focused on pathology detection in MRI [10, 26]. Other work has used the pathology annotations provided by some datasets [7, 27] to compare reconstruction metrics in these relevant areas [7, 13] or to show that measurement noise can lead to changes in reconstructions in these areas [16]. To our knowledge, ours is the first work to consider pathology detection in accelerated MRI reconstruction utilizing multiple reconstructions.

3 Method

3.1 Background and Notation

The goal of MRI reconstruction is to find a reconstructed image $\hat{\mathbf{x}}$ that is consistent with the measured k-space data $\mathbf{y}$ subject to the following equation:

$$\mathbf{y} = \mathbf{A}\hat{\mathbf{x}} \quad \text{with} \quad \mathbf{y} \in \mathbb{C}^m, \quad \mathbf{A} \in \mathbb{C}^{m \times n}, \quad \hat{\mathbf{x}} \in \mathbb{C}^n, \quad (1)$$

where $\mathbf{A} = MFS$ with the coil sensitivity maps S , the Fourier transform $\mathcal{F}$ and the undersampling mask M . MRI acquisitions are often accelerated by only measuring a fraction of the k-space data $\mathbf{y}$.

The fully sampled image, which would have been obtained by measuring the entire k-space, contains a (possibly empty) set of pathologies denoted as $\mathcal{S} = \{s_1, s_2, \dots\}$. However, due to image degradation caused by acceleration, a different set of pathologies, $\mathcal{S}' = \{s'_1, s'_2, \dots\}$, is visible in the reconstructed image $\hat{\mathbf{x}}$. Ideally, $\mathcal{S}$ and $\mathcal{S}'$ should be identical to ensure accurate diagnosis. However, there are infinitely many reconstructions that satisfy Eq. 1 with no guarantee that they correctly reconstruct pathologies. As a result, $\mathcal{S}'$ may fail to include all pathologies present in the patient, particularly for high acceleration factors.

In this work, we aim to generate a set of reconstructions, $\mathcal{X} = \{\hat{\mathbf{x}}^{(1)}, \dots, \hat{\mathbf{x}}^{(N_{rec})}\}$, such that when considered jointly, the reconstructed images capture a set of pathologies $\mathcal{S}'$ that includes as many of the ground truth pathologies $\mathcal{S}$ as possible (few false negatives) while not hallucinating pathologies (few false positives).

3.2 Generating Diverse Solutions Maximizing Semantic Distance

Our key contribution is a method which, given an initial reconstruction $\hat{\mathbf{x}}^{(1)}$, can efficiently generate a set of reconstructions $\mathcal{X}$ that are semantically diverse in the regions where pathologies are likely to occur. To achieve this, we iteratively generate additional reconstructions $\hat{\mathbf{x}}^{(i)}$, $i = 2, \dots, N_{rec}$, by maximizing semantic distance d to all previous reconstructions $\hat{\mathbf{x}}^{(j)}$ in an iterative procedure

$$\hat{\mathbf{x}}^{(i)} \leftarrow \mathcal{DC} \left(P_{B(\hat{\mathbf{x}}^{(1)}, r)} \left(\hat{\mathbf{x}}^{(i)} + \eta \nabla_{\hat{\mathbf{x}}^{(i)}} \sum_{j \neq i} d(\hat{\mathbf{x}}^{(i)}, \hat{\mathbf{x}}^{(j)}) \right) \right), \quad (2)$$

where η is the step size, $\nabla_{\hat{\mathbf{x}}^{(i)}}$ is the gradient operator, $P_{B(\hat{\mathbf{x}}^{(1)}, r)}$ is the projection onto the ℓ_2 -ball of radius r around $\hat{\mathbf{x}}^{(1)}$, and $\mathcal{DC}$ is the projection operator mapping the reconstruction $\hat{\mathbf{x}}^{(i)}$ on the data-consistent space. We ensure data-consistency by replacing the k-space data with the original measured data at the locations where a measurement was acquired. To prevent potential artifacts from occurring, we run M-Step Conjugate Gradient before, minimizing the distance of the predicted k-space to the measurement. Furthermore, the projection $P_{B(\hat{\mathbf{x}}^{(1)}, r)}$ ensures that $\|\hat{\mathbf{x}}^{(i)} - \hat{\mathbf{x}}^{(1)}\|_2 \leq r$, where r is the allowed perturbation radius set to $r = 3$ in our experiments. The value has been chosen, as it allows for meaningful perturbations while keeping the number of new false positives low.

Input: $y, \hat{x}^{(1)}, f_\phi, RPN, N_{opt}, N_{rec}, r$
Output: $\mathcal{X} = \{\hat{x}^{(1)}, \dots, \hat{x}^{(N_{rec})}\}$
 $\mathcal{B} = \{b_1, b_2, \dots\} \leftarrow RPN(\hat{x}^{(1)})$ or manually ; // Generate Proposal Boxes
for $i \leftarrow 1$ **to** N_{rec} **do**
 $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ with $\epsilon^{(1)} = 0$; // Keep init. recon.
 $\hat{x}^{(i)} \leftarrow P_{B(\hat{x}^{(1)}, r)}(\hat{x}^{(1)} + \epsilon^{(i)})$, where $B(\hat{x}^{(1)}, r) = \{x \mid \|x - \hat{x}^{(1)}\|_2 \leq r\}$,
 $\hat{x}^{(i)} \leftarrow \mathcal{DC}(\hat{x}^{(i)})$; // Data consistency
 $\mathbf{a}^{(i)} \leftarrow f_\phi(\hat{x}^{(i)}, \mathcal{B})$; // Get box-features
end
for $k \leftarrow 0$ **to** N_{opt} **do**
 for $i \leftarrow 2$ **to** N_{rec} **do**
 $d^{(i)} = \sum_{j \neq i} \|\mathbf{a}^{(i)} - \mathbf{a}^{(j)}\|_2$; // Calculate feature distances
 $\hat{x}^{(i)} \leftarrow P_{B(\hat{x}^{(1)}, r)}(\hat{x}^{(i)} + \eta \nabla_{\hat{x}^{(i)}} d^{(i)})$
 $\hat{x}^{(i)} \leftarrow \mathcal{DC}(\hat{x}^{(i)})$; // Data consistency
 $\mathbf{a}^{(i)} \leftarrow f_\phi(\hat{x}^{(i)}, \mathcal{B})$; // Get box-features
 end
end

Algorithm 1: Generating semantically diverse reconstructions by optimizing the distance between box features $\mathbf{a}^{(i)}$ obtained by the robust ViTDet.

To measure semantic distance between images, we utilize the features of a ViTDet detection network f_ϕ [14], which we have finetuned on the fastMRI+ data. Relevant proposal bounding boxes $\mathcal{B} = \{b_1, b_2, \dots\}$ can be selected using methods described in Sect. 3.4. Given the box-features $\mathbf{a}^{(i)} = f_\phi(\hat{x}^{(i)}, b)$ we define

$$d(\hat{x}^{(i)}, \hat{x}^{(j)}) = \sum_{b \in \mathcal{B}} \|f_\phi(\hat{x}^{(i)}, b) - f_\phi(\hat{x}^{(j)}, b)\|_2. \quad (3)$$

The proposed algorithm is summarized in Alg. 1.

3.3 Adversarially Robust Backbone

Our application relies on gaining semantically meaningful gradients from the box-feature encoder f_ϕ in Eq. (3). While gradients of deep learning models are generally not interpretable [15], it is well-known that adversarially robust models possess semantically more meaningful gradients [1, 6, 9, 19, 24]. To obtain an adversarially robust detection model, we leverage the recently proposed adversarial fine-tuning method of [20]. We apply this method to the ViT-backbone g_ϕ of the ViTDet. The ViT-backbone also serves as the backbone of the box-feature encoder $f_\phi = h_\phi \circ g_\phi$. The method ensures that the backbone features, and thus also the box features, remain stable under small adversarial perturbations. The training objective is

$$\min_{\phi} \sum_i \max_{\|\delta\|_2 \leq r} \|g_\phi(\mathbf{x}_i + \delta) - g_{\phi_{\text{org}}}(\mathbf{x}_i)\|_2^2, \quad (4)$$

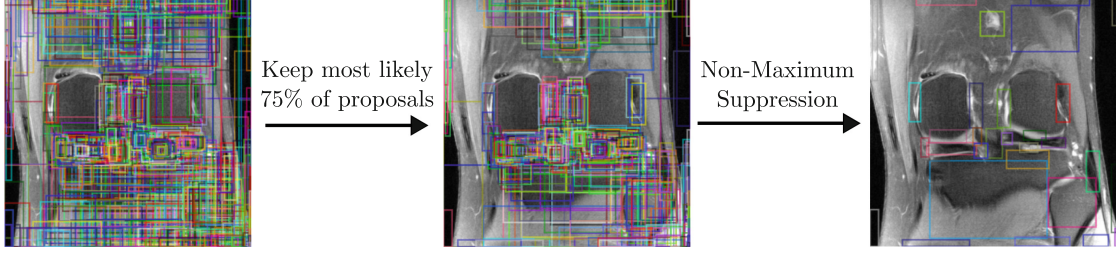


Fig. 2. Automatic Proposal Generation. Given an initial reconstruction, we filter the proposal boxes given by the Region Proposal Network (RPN) of the ViTDet to only keep the 75% of boxes with the highest score. We then apply Non-Maximum Suppression (NMS) on the remaining proposals. The remaining proposal boxes are then used for optimization with SDR-A.

where $g_{\phi_{\text{org}}}$ is the frozen original backbone, r is the adversarial perturbation budget, and $\mathbf{x}_i$ are unlabeled training images. In contrast to [20], we use the ℓ_2 -norm instead of the ℓ_∞ -norm to bound the perturbation budget, as this generally yields better generative capabilities [1]. We set r to 10 and train for 20,000 steps on ImageNet data at resolution 640 using a batch size of 16. While it would be natural to train on fastMRI images, we have observed a stronger performance when training on a much larger set of diverse natural images.

3.4 Box-Proposal Selection

We introduce two approaches for identifying the box regions to optimize, which are motivated by two different use-cases.

With **Manual (SDR-M) Proposal Box Generation** we address the scenario where a user suspects a pathology in a specific area and manually draws a bounding box. We simulate user annotations by perturbing the ground-truth boxes in size (75%-125%) and position (up to 25% of width or height).

With **Automatic (SDR-A) Proposal Box Generation** we address the case when no patient-specific information on the location of pathologies is available. In this scenario, we use the Region Proposal Network (RPN) of the ViTDet. We filter out the 25% of boxes that have the lowest probability, as these are often in the background of the MRI image. We then utilize Non-Maximum-Suppression (NMS) to remove boxes with significant overlap ($\text{IoU} > 0.05$), see Fig. 2.

After applying either method, we get a set of proposal boxes $\mathcal{B} = \{b_1, b_2, \dots\}$ in which we optimize the reconstructions to be semantically different.

4 Data, Evaluation and Results

Assessing clinical relevance of reconstructions would ideally require input from clinicians. As it is difficult to do such an evaluation at scale, instead, we evaluate by training an object detector and use it as a proxy for human annotators.

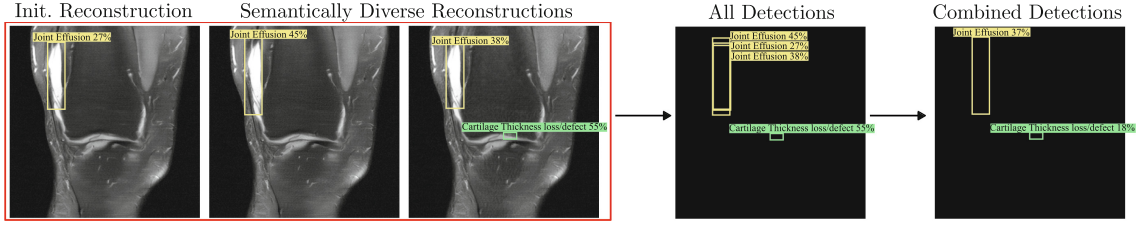


Fig. 3. Combining predictions of multiple reconstructions. We combine overlapping detection boxes and average their probability score across reconstructed images to compute the mAP of our detections.

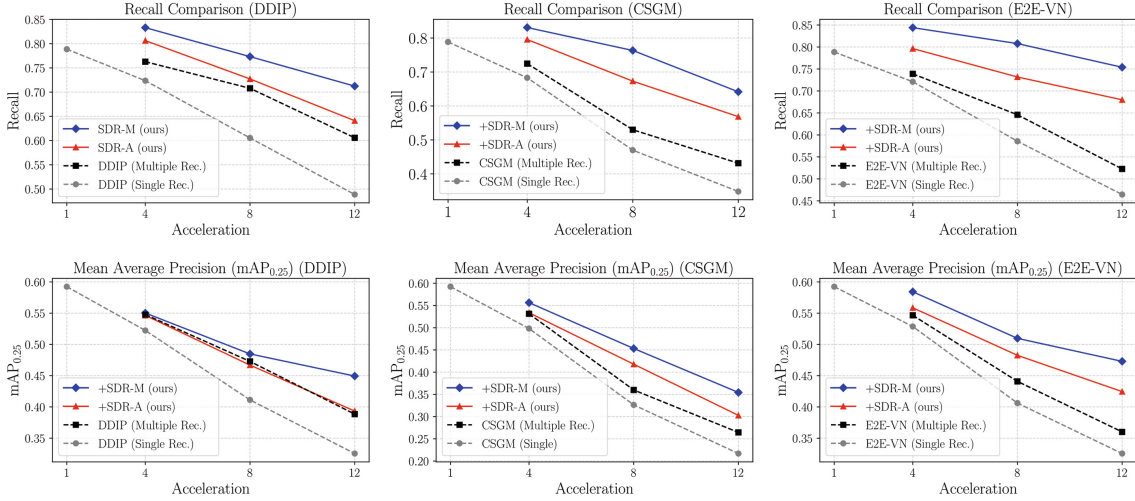


Fig. 4. Evaluation of reconstruction methods. Both SDR-A and SDR-M outperform single and multiple reconstructions using the original reconstruction techniques (DDIP, CSGM, E2E-VN) in terms of recall (better detection of pathologies) while achieving similar or better $mAP_{0.25}$. Thus SDR generates clinically relevant reconstructions outperforming resampling.

Data and Training: Open datasets of MRI reconstruction data are rare [7, 12, 21, 23] and only a few provide annotated pathologies [7, 27]. For our experiments we use the fastMRI+ [27] knee data, as it provides enough annotated slices to train large detection networks. We split the fastMRI+ validation set subject-wise into test and validation set. For evaluation, we considered the 10 most frequent pathology classes in fastMRI+, as the remaining classes have not enough test samples.

Object Detector for Evaluation: To ensure an objective evaluation independent of the ViTDet used for generating the reconstructions of SDR, we trained a separate Faster-RCNN detection network [18] which we use for evaluation.

Combining Detections of Multiple Images: Given a set of potential pathology detections $\mathcal{S}' = \{s'_1, s'_2, \dots\}$ that was extracted from a set of reconstructions $\mathcal{X}$, we need to remove near duplicates and estimate the probability of each

bounding box for the calculation of mean average precision (mAP). To remove near duplicates, we used NMS to eliminate overlapping bounding boxes (Intersection-over-Union (IoU) ≥ 0.25) of the same class. To estimate the probability scores of the remaining boxes, we averaged the probability for a given bounding box s'_i across all reconstructed images and set the probability score to 0 for reconstructions that did not contain the bounding box. An example is shown in Fig. 3.

Reconstruction Techniques: We analyzed the benefits of SDR with three base reconstruction methods: the diffusion-based DDIP [4], and CSGM [11] approaches, as well as the E2E-VarNet [22]. We produce 3 total reconstructions for each image using the baselines as well as SDR, i.e. SDR ($N_{rec} = 3$). While the diffusion models can generate multiple reconstructions by design, we trained three separate E2E-VarNets with three different random seeds for each acceleration to reconstruct three different images during inference.

Performance Measures: We measured the methods' ability to produce sets of reconstructions that contain the ground truth pathologies using the recall, where a recall of 1 indicates no missed pathologies. As more reconstructions might lead to more false positives, we also measured mean Average Precision (mAP_{0.25}) [8]. For both metrics, we consider an IoU of 0.25 as correct detection. Small IoU thresholds are commonly used in medical detection tasks [2] to account for uncertain border definitions and the typically small size of pathologies.

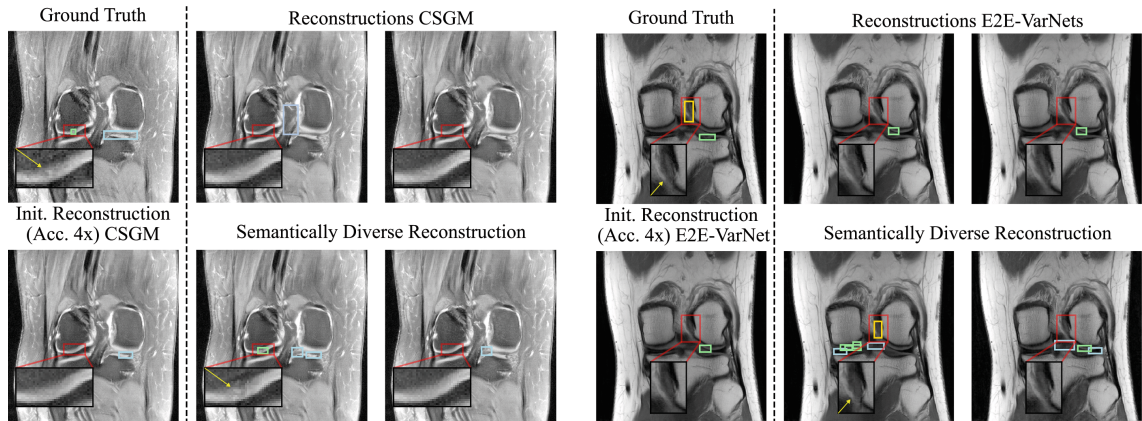


Fig. 5. The highlighted box in the ground truth shows a cartilage thickness loss (left) and a high grade ligament sprain (right). The initial reconstructions are often too smooth, preventing the detection of the pathologies. The semantic changes of SDR-M (left) and SDR-A (right) reveal the pathological changes. Yellow arrows indicate pathologies. (Color figure online)

5 Results

The quantitative results in Fig. 4 show that SDR-A and SDR-M have higher recall than the original single and multiple reconstructions. Thus SDR allows to uncover pathologies which can be missed with the original reconstruction method. A qualitative example of this is shown in Fig. 5. This is true for the diffusion-based reconstruction methods CSGM and DDIP as well as the often applied technique E2E-VarNet. At the same time SDR even improves $\text{mAP}_{0.25}$ and thus SDR is not simply achieving higher recall by introducing more false positive detections. We observed that it is necessary to use a ViTDet with robust backbone for the generation in SDR, as the original ViTDet did not improve in recall or mAP compared to repeated sampling.

6 Discussion and Conclusion

We have shown that existing accelerated MRI reconstruction methods can miss pathologies and have suggested SDR to explore semantically different but fully data consistent reconstructions. The evaluation of SDR reconstructions with an object detector shows fewer missed pathologies (higher recall) while improving mAP. We believe that SDR can lead to a safer and more reliable development of image reconstructions for accelerated MRI.

The generated semantically diverse reconstructions can also be utilized by a clinician to better understand the reconstruction uncertainty, as the diverse reconstructions visualize disparities between the solutions. This would allow the clinician to acquire further measurements if necessary.

A key next step is to integrate our approach into clinical workflows and to observe how radiologists interact with semantically diverse reconstructions in practice. Furthermore, deploying our method during the scanning process could allow the system to automatically determine whether additional data acquisition is needed to further reduce semantic uncertainty.

Acknowledgment. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting JNM and CS.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Augustin, M., Meinke, A., Hein, M.: Adversarial robustness on in- and out-distribution improves explainability. In: ECCV (2020)
2. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nnDetection: a self-configuring method for medical object detection. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 530–539. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_51

3. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated mri. *Med. Image Anal.* 102479 (2022)
4. Chung, H., Ye, J.C.: Deep diffusion image prior for efficient ood adaptation in 3d inverse problems. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *ECCV 2024*, pp. 432–455. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-73226-3_25
5. Cohen, N., Manor, H., Bahat, Y., Michaeli, T.: From posterior sampling to meaningful diversity in image restoration. In: *ICLR* (2024)
6. Croce, F., Schlarmann, C., Singh, N.D., Hein, M.: Adversarially robust clip models can induce better (robust) perceptual metrics. In: *IEEE SaTML* (2025)
7. Desai, A.D., et al.: Skm-tea: a dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. In: *NeurIPS Datasets and Benchmarks Track* (2021)
8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *IJCV* **88**, 303–338 (2010)
9. Ganz, R., Elad, M.: Clipag: towards generator-free text-to-image generation. In: *WACV* (2024)
10. Harman, F., Selver, M.A., Baris, M.M., Canturk, A., Oksuz, I.: Deep learning-based meniscus tear detection from accelerated mri. *IEEE Access* **11**, 144349–144363 (2023)
11. Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A.G., Tamir, J.I.: Robust compressed sensing mri with deep generative priors. In: *NeurIPS* (2021)
12. Knoll, F., 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.* **2**(1), e190007 (2020). PMID: 32076662
13. Küstner, T., Hammernik, K., Rueckert, D., Hepp, T., Gatidis, S.: Predictive uncertainty in deep learning-based mr image reconstruction using deep ensembles: evaluation on the fastmri data set. *Magn. Reson. Med.* **92**(1), 289–302 (2024)
14. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: *ECCV*, pp. 280–296. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20077-9_17
15. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: *CVPR* (2015)
16. Morshuis, J.N., Gatidis, S., Hein, M., Baumgartner, C.F.: Adversarial robustness of mr image reconstruction under realistic perturbations. In: Haq, N., Johnson, P., Maier, A., Qin, C., Würfl, T., Yoo, J. (eds.) *Machine Learning for Medical Image Reconstruction*, pp. 24–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17247-2_3
17. Morshuis, J.N., Hein, M., Baumgartner, C.F.: Segmentation-guided mri reconstruction for meaningfully diverse reconstructions. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Mehrof, D., Yuan, Y. (eds.) *Deep Generative Models*, pp. 180–190. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72744-3_18
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2016)
19. Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Image synthesis with a single (robust) classifier. In: *NeurIPS* (2019)
20. Schlarmann, C., Singh, N.D., Croce, F., Hein, M.: Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In: *ICML* (2024)

21. Solomon, E., et al.: fastmri breast: a publicly available radial k-space dataset of breast dynamic contrast-enhanced mri. *Radiol. Artif. Intell.* **7**(1), e240345 (2025)
22. Sriram, A., et al.: End-to-End variational networks for accelerated MRI reconstruction. In: Martel, A.L., et al. (eds.) *MICCAI 2020*. LNCS, vol. 12262, pp. 64–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_7
23. Tibrewala, R., et al.: Fastmri prostate: a public, biparametric mri dataset to advance machine learning for prostate cancer imaging. *Sci. Data* **11**(1), 404 (2024)
24. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: *ICLR* (2019)
25. Xie, Y., Li, Q.: Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *MICCAI 2022*, pp. 655–664. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_62
26. Yen, C.Y., Singhal, R., Sharma, U., Ranganath, R., Chopra, S., Pinto, L.: Adaptive sampling of k-space in magnetic resonance for rapid pathology prediction. In: *ICML* (2024)
27. Zhao, R., et al.: fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Sci. Data* **9**(1), 152 (2022)