Diverse, Multi-Faceted Uncertainty Quantification in Accelerated Magnetic Resonance Imaging

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Jeffrey Wen, B.S., M.S.

Graduate Program in Department of Electrical and Computer Engineering

The Ohio State University

2025

Dissertation Committee:

Dr. Philip Schniter, Advisor

Dr. Rizwan Ahmad, Advisor

Dr. Kiryung Lee

Abstract

Magnetic resonance imaging (MRI) is an essential medical diagnostic tool that produces high-quality soft-tissue images without harmful ionizing radiation. MRI acquisition, however, is an inherently slow process, which limits patient throughput and comfort. Accelerated MRI seeks to reduce the acquisition time by collecting fewer measurements. This results in undersampled data that requires non-trivial reconstruction techniques to recover diagnostic-quality images. Yet, the reconstruction problem is fundamentally ill-posed, meaning that many plausible images can correspond to the same set of measurements and a given prior. This distribution of plausible reconstructions is known as the posterior. Despite the ill-pose nature, most existing MRI reconstruction methods are designed to provide only a single point estimate, ignoring the intrinsic uncertainty of the problem This dissertation addresses the critical need for uncertainty quantification (UQ) in accelerated MRI by proposing three methods, each quantifying the uncertainty from a different perspective.

First, we design a novel conditional normalizing flow (CNF) to approximate the posterior distribution. Then, by generating multiple reconstructions for a given measurement, we create a pixel-wise uncertainty map that highlights areas of the image with more variability. This informs practitioners about the trustworthiness of accelerated reconstructions on a pixel level.

In our second approach, we propose to quantify the uncertainty introduced when using accelerated reconstructions instead of the true image for a downstream task like pathology classification. Utilizing conformal prediction, this approach constructs a prediction interval in the task-output space that is statistically guaranteed to contain the task output given to the true image with a user-specified probability. The width of these intervals serves as a measure of the uncertainty contributed by the measurement-and-reconstruction process and offers a more direct understanding of the reliability of the reconstructed images for the downstream task.

Our final method addresses the uncertainty in evaluating the quality of reconstructed images in the absence of a ground truth. We propose to use conformal inference to construct bounds on full-reference image quality metrics such as Peak Signal-to-Noise Ratio (PSNR). These bounds come with probabilistic guarantees, thus allowing one to assess the image quality of accelerated reconstructions without relying on direct comparisons to true images.

In all, this research addresses an essential gap in the field of accelerated MRI by quantifying the uncertainty in the acceleration process through several different lenses. By providing diverse uncertainty estimates, our proposed methods aim to improve the reliability and clinical utility of accelerated MRI.

Acknowledgments

This work was supported in part by the National Institutes of Health under Grant R01-EB029957.

First, I would like to express my deep appreciation for my advisors, Dr. Rizwan Ahmad and Dr. Philip Schniter, for giving me the opportunity to continue pursuing a PhD when my academic future became unclear. Their thoughtful guidance and mentorship allowed me to develop as a researcher, writer, presenter, and learner, which I will be forever grateful for.

I would also like to acknowledge my labmate, Saurav Shastri, for his technical input and friendship. Having a friend in lab made all the difference after the isolation of the pandemic.

Of course, I could not have reached this point if not for the support of my friends and family. My parents and sister have always inspired me to pursue hard goals while providing me with the encouragement and care to succeed. I am also thankful to all the friends I've met at OSU and those from back home for making these last few years as enjoyable as possible.

Lastly, I have to thank my partner Dr. Patricia Loughney for all the meals cooked while I scrambled to meet deadlines, the open ears while I lamented about unfair reviewers, and her unwavering love during moments of self-doubt. My PhD experience would not have been the same without her.

Vita

2015-2019	B.S. Electrical Engineering, University of Wyoming
2019-2022	M.S. Electrical Engineering, The Ohio State University
2019-present	Graduate Research Associate, The Ohio State University.

Publications

Research Publications

J. Wen, R. Ahmad, P. Schniter "Task-Driven Uncertainty Quantification in Inverse Problems via Conformal Prediction". Proc. Europ. Conf. Comp. Vision, pp. 182–199, Nov. 2024.

J. Wen, R. Ahmad, P. Schniter "A Conditional Normalizing Flow for Accelerated Multi-Coil MR Imaging". Proc. Intl. Conf. Mach. Learn., pp. 36926–36939, June. 2023.

J. Wen, R. Ahmad, P. Schniter "Posterior Sampling for Accelerated Multicoil MRI Reconstruction using a Conditional Normalizing Flow". Proc. Annu. Meeting ISMRM, June. 2023.

Fields of Study

Major Field: Department of Electrical and Computer Engineering

Table of Contents

Page

List of Tables

List of Figures

Chapter 1: Introduction

Magnetic resonance imaging (MRI) is a standard diagnostic imaging tool that has become widely used in the medical field. As it provides high-quality images of soft tissue without exposing patients to harmful ionizing radiation, it is often the preferred choice over other imaging tools such as computational tomography (CT). Despite its advantages, MRI exams can be a very slow process. For example, a standard MRI scan can take anywhere from 30 minutes to over an hour depending on the resolution, number of contrasts, and different views required. This prolonged acquisition time increases the chance of motion artifacts from patient movement, reduces patient comfort, and decreases patient throughput.

Scan times can be reduced by simply collecting fewer measurements while the patient is in the scanner, often below the Nyquist rate. However, this acceleration introduces severe aliasing and can result in an image with little to no diagnostic value. A well-designed estimation method is thus required in order to recover diagnostic-quality images. The design of these estimation methods has become an active area of research [75, 55]. Figure 1.1 illustrates the pipeline of the conventional, fully-sampled MRI and the accelerated version.

Early approaches took advantage of new hardware improvements, which incorporated multiple receiver coils into the scanners. After estimating coil-sensitivity maps

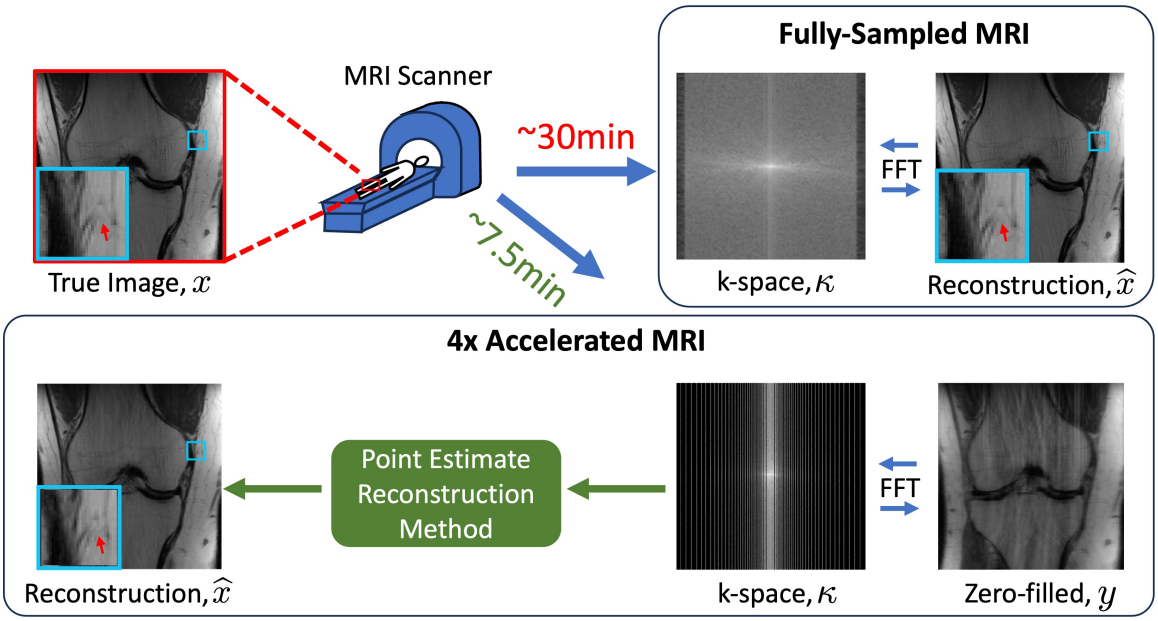


Figure 1.1: A high-level visualization of the MRI recovery problem. A fully-sampled scan recovers a high-quality, reliable image but is time consuming. Accelerated MRI reduces the scan time by only collecting a fraction of the necessary data but requires an estimation method to recover a quality image. These accelerated reconstructions can differ from the true image in subtle but meaningful ways as shown by the zoomed-in regions in blue.

or interpolation kernels, methods like SENSE [95] and GRAPPA [52] use subsampled data from multiple coils to remove aliasing artifacts in the final reconstruction. Known as parallel imaging, these reconstruction techniques are available on most modern commercial scanners but typically only enable a two- to three- fold acceleration of the acquistion process. For higher acceleration, methods based on compressed-sensing (CS) have been proposed [86]. The CS methods are framed to iteratively minimize the sum of a data-fidelity and regularization term, where the regularization term incorporates prior knowledge about the images. Prior knowledge may dictate sparsity of true images in some transform domain, as in traditional CS, or that the true images are preserved by some denoising function, as in "plug-and-play" recovery [3]. Deep neural networks have also been proposed for MRI recovery, based on end-to-end approaches [125, 47, 107] or algorithmic unrolling [54]. Yet another approach, known as compressed sensing with a generative model (CSGM) [24], trains a deep image generator and then optimizes its input to give the image that, after application of the forward model, best matches the measurements.

Although they achieve high reconstruction quality, the aforementioned methods provide only a single point estimate. Yet, accelerated MRI is an ill-posed inverse problem, where many possible reconstructions exist that are consistent with a given prior and set of subsampled measurements. This distribution of feasible reconstructions is known as the posterior. Despite this, point estimate methods do not provide any information regarding the variation in reconstructions from the posterior distribution. Since different recovery methods may be biased towards different plausible image hypotheses, this can lead to important differences in reconstruction quality. For example, modern deep-network approaches can sometimes hallucinate [35, 18, 59, 89,

21, 51, 113], i.e., generate visually pleasing recoveries that differ in important ways from the true image. An example hallucination is shown in Fig. 1.1. Since small variations in image content can impact the final diagnosis, it is crucial for radiologists to know whether a visual structure is truly reflective of the patient anatomy or merely an imaging artifact. Problems of this form fall into the realm of uncertainty quantification (UQ) [1], and their solutions are increasingly essential for the adoption of machine learning approaches in safety-critical medicine [32, 13].

The absence of a comprehensive UQ framework remains a major barrier for the integration of highly accelerated MRI. This dissertation aims to provide such a framework by proposing three diverse approaches, each quantifying the uncertainty of the measurement-and-reconstruction (acceleration) process from a different perspective. A description of each perspective and our proposed solutions is as follows:

1. Pixel-based (Ch. 3): We propose to construct a map indicating the amount of uncertainty in the prediction of each pixel. Using a novel conditional normalizing flow (CNF), we draw many reconstructions from the estimated posterior distribution and use the pixel-wise standard deviation map of these reconstructions to quantify the uncertainty on an individual pixel level.

2. Task-based (Ch. 4): We aim to evaluate how a downstream task (e.g. pathology classification) behaves differently when supplied with a reconstructed image versus the true image. Using conformal prediction, we construct an interval in the task-output space that is guaranteed to contain the task-output of the true image up to a user-specified probability. The width of the prediction interval provides a natural way to quantify the uncertainty contribution of the acceleration process on the downstream task output.

3. Image-Quality-Based (Ch. 5): Given a reconstruction, we look to quantify the uncertainty on a full-reference image-quality metric like the Peak-Signal-to-Noise Ratio (PSNR) by constructing bounds on the metric when the true image is unknown. By utilizing conformal prediction, these bounds are valid with high probability.

In the proceedings chapters, we describe our methodologies in detail and discuss how each perspective provides a different insight into the uncertainty present in accelerated MRI. By providing a more complete understanding of the associated uncertainty, our approach promises to allow practitioners to better assess the trustworthiness of accelerated images, which we hope supports fewer misdiagnoses and a wider adoption of accelerated MRI.

Chapter 2: Background

2.1 Magnetic Resonance Imaging

We provide a brief background on MRI in order to clarify certain domain-specific terms and concepts. MRI is a widely-used medical imaging technique that is often preferred due to its reputation in safety. Rather than relying on ionizing radiation, MRI exploits the magnetic properties of hydrogen protons within the body. These protons are aligned by a strong magnetic field within the scanner. Then, protons in the area of interest are temporarily excited into a high-energy state by a radio frequency pulse, and as they relax to equilibrium, they induce an electromagnetic signal in a receiver coil. These signals are later processed to generate an image. The use of multiple receiver coils is described as "multi-coil" or "parallel" MRI.

The measured signals correspond to spatial frequency information within a multidimensional spatial Fourier domain known as the "k-space". To resolve spatial localization, the k-space is filled iteratively with each new excitation-and-relaxation cycle (pulse sequence) acquiring data for a particular area of k-space. For example, Cartesian sampling, the most commonly used sampling strategy, collects an entire line of k-space at each iteration. For a given spatial resolution, a certain number of k-space lines must be collected to satisfy the Nyquist theorem. When all of these

k-space measurements are collected, the acquisition is said to be "fully-sampled", and the image can be recovered by simply performing a multidimensional inverse Fourier transform.

However, MRI acquisitions can be time-consuming due to the sequential filling of the k-space. Accelerated MRI aims to the reduce scan time by acquiring only a fraction of the fully-sampled k-space, but the application of a simple inverse Fourier transform then results in aliasing artifacts. Thus, the central challenge of accelerated MRI is to recover diagnostic-quality reconstructions from highly subsampled measurements.

2.2 Inverse Imaging Problems

Mathematically, accelerated MRI falls under the broader class of inverse imaging problems where the goal is to recover a true image x from noisy, distorted, and/or incomplete measurements $y = \mathcal{A}(x)$ [12]. These problems can be non-linear as with phase-retrieval, de-quantization, low-light imaging, and image-to-image translation or linear as in the case of accelerated MRI, limited-angle computed tomography, denoising, deblurring, inpainting, and super-resolution. Due to a partial loss of the original signal, these problems are generally ill-posed, in that it is impossible to perfectly infer x from y. For this dissertation, we focus on the linear case with an emphasis on the accelerated MRI problem but note that the methods described in Ch. 4 and 5 can also be applied more generally to non-linear problems as well.

Linear inverse problems are commonly expressed as

$$y = Ax + \epsilon \tag{2.1}$$

where A is a known forward operator and ϵ is measurement noise. For multi-coil accelerated MRI in particular, we consider the case when x is a true multi-coil image

and observed measurement y is an aliased multi-coil estimate corrupted by only collecting a $1/R$ fraction of the scan data required by the Nyquist sampling theorem. The integer R is known as the "acceleration rate", and when $R > 1$, the inverse problem is ill-posed. This formulation is made more explicit in the following section.

2.3 The Accelerated MRI Inverse Problem

As mentioned, MRI measurements of the D-pixel true image $\iota_{\mathsf{true}} \in \mathbb{C}^D$ are collected in the spatial Fourier domain known as the k-space. In a multi-coil system with B coils, measurements from the bth coil can be written as

$$\kappa_{(b)} = PFS_{(b)}\iota_{\mathsf{true}} + \varepsilon_{(b)} \in \mathbb{C}^M, \tag{2.2}$$

where $P \in \mathbb{R}^{M\times D}$ is a sampling matrix containing M rows of the $D \times D$ identity matrix I, F is the $D \times D$ 2D unitary discrete Fourier transform (DFT) matrix, $S_{(b)} \in \mathbb{C}^{D\times D}$ is the coil-sensitivity map of the bth coil, and $\varepsilon_{(b)} \in \mathbb{C}^M$ is measurement noise. We will assume that $\{S_{(b)}\}_{b=1}^B$ have been obtained from ESPIRiT [115], in which case $\sum_{b=1}^B S_{(b)}^{\mathsf{H}} S_{(b)} = I$. In the case of single-coil MRI, $B = 1$ and $S_1 = I$.

To recover the form of (2.1), we can rewrite the model in terms of the "coil images" $x_{(b)} \triangleq S_{(b)}\iota_{\mathsf{true}}$ and their corresponding "zero-filled" estimates $y_{(b)} \triangleq F^{\mathsf{H}}P^\top\kappa_{(b)}$, and stack all the coils together via $x \triangleq [x_{(1)}^\top, \ldots, x_{(B)}^\top]^\top$ and $y \triangleq [y_{(1)}^\top, \ldots, y_{(B)}^\top]^\top$. Expressing the forward operator as

$$A = \text{blkdiag}\left\{F^{\mathsf{H}}P^\top PF, \ldots, F^{\mathsf{H}}P^\top PF\right\} \tag{2.3}$$

with measurement noise $\epsilon = [(F^{\mathsf{H}}P^\top\varepsilon_{(1)})^\top, \ldots, (F^{\mathsf{H}}P^\top\varepsilon_{(B)})^\top]^\top$, we arrive at the previous formulation of a linear inverse problem (2.1).

To perform image recovery, one can first compute y from the k-space measurements, then estimate $\widehat{x} = [\widehat{x}_{(1)}^\top, \ldots, \widehat{x}_{(B)}^\top]^\top$ from y. This multi-coil estimate can be either "coil-combined" to yield a complex-valued image estimate

$$\widehat{\imath} = [S_1^{\mathsf{H}}, \ldots, S_B^{\mathsf{H}}]\widehat{x}, \tag{2.4}$$

or one can compute the root-sum-of-squares (RSS) reconstruction [96] to obtain a magnitude-only image estimate

$$|\widehat{\imath}| = \sqrt{\textstyle\sum_{b=1}^{B} |\widehat{x}_{(b)}|^2}. \tag{2.5}$$

In the fully-sampled case, $M = D$ and so $y = x + \epsilon$. As previously mentioned, fully sampled acquisition is very slow, so we are interested in accelerating the scan by collecting $M < D$ measurements per coil. This gives an "acceleration rate" of $R \triangleq D/M$, but it makes A rank deficient. In this latter case, accurate recovery of x requires the use of prior information about x, such as the knowledge that x is a vector of MRI coil images.

2.4 Data

To facilitate the study of accelerated MRI estimation, the fastMRI [125] dataset was collected and made publicly available by the NYU fastMRI initiative. Since its release, the dataset has become a primary standard for training, evaluating, and comparing recovery methods; thus, we utilize the dataset extensively for the majority of our experiments. The dataset contains scans of two different anatomies: the knee and the brain. The knee set contains 1594 scans of fully-sampled multi-coil knee MRIs, nearly half of which have fat-suppression. The brain set contains 6970 fully-sampled

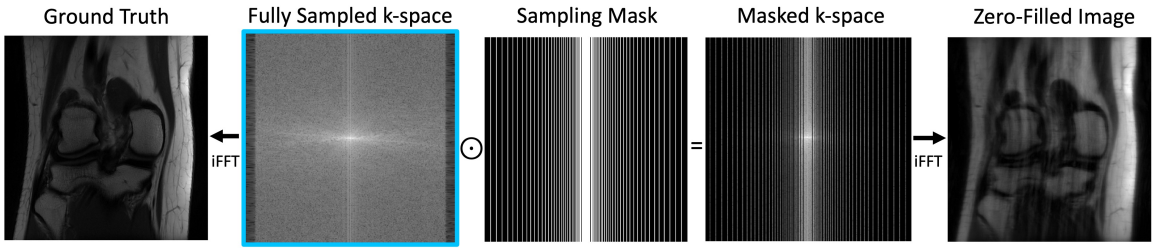


Figure 2.1: A visual illustration of simulating accelerated MRI. Given the fully sampled k-space $\kappa_{(b)}$ highlighted in blue, we obtain the ground truth $x_{(b)}$ by applying the inverse Fourier transform F^{H}. The zero-filled image $y_{(b)}$ is acquired by applying the sampling mask $P^\top P$ to fully sampled $\kappa_{(b)}$ and then taking the inverse Fourier Transform F^{H}.

scans of multi-coil brain MRIs and can be broken down further as axial T1 weighted, T2 weighted, and FLAIR images.

Since the fastMRI datasets only contain the raw fully-sampled k-space data, i.e. $\{\kappa_{(b)}\}_{b=1}^B$ with $M = D$, we simulate the acceleration process by retrospectively subsampling the k-space measurements. More explicitly, the zero-filled images $\{y_{(b)}\}_{b=1}^B$ are obtained by masking the fully-sampled k-space measurement $\kappa_{(b)}$ and taking the inverse Fourier transform, i.e., $y_{(b)} = F^{\mathsf{H}} P^\top P \kappa_{(b)}$, wherein we assume that the noise $\epsilon_{(b)}$ in (2.2) is negligible. Here, $P^\top P$ is known as the sampling mask. This mask indicates which points in the k-space are sampled. In a similar fashion, the ground truth coil-images $\{x_{(b)}\}_{b=1}^B$ are computed by taking the inverse Fourier transform of the fully sampled k-space measurement, i.e., $x_{(b)} = F^{\mathsf{H}} \kappa_{(b)}$ This procedure is illustrated in Fig. 2.1. In real-world accelerated MRI, the data acquisition process would collect masked k-space $P\kappa_{(b)}$ directly. We specify the details of the sampling mask utilized for each of our methods explicitly in the respective chapters.

Chapter 3: Conditional Normalizing Flows for Accelerated Multi-coil MRI

As previously mentioned, existing point estimate methods for accelerated MRI [95, 52, 86, 125, 107] only provide a single estimate $\widehat{x}$ without any consideration for the uncertainty inherent in the ill-posed recovery problem. One approach that facilitates UQ is Bayesian imaging, where the goal is not to compute a single "good" image estimate but rather to sample from the posterior distribution. The availability of a large batch of posterior samples enables many forms of UQ. For example, we demonstrate the generation of a pixel-wise standard-deviation map, which quantifies which pixels are more trustworthy. This gives a visual representation of the uncertainty at the individual pixel level. The generation of posterior samples also forms the basis for other methods of uncertainty quantification (Ch. 4 and 5) and facilitates future work that may use those samples for applications like adaptive sampling [100] or counterfactual diagnosis [28]. Thus, we focus on the task of sampling from the posterior in this chapter as a foundational base.

There exist several deep-learning based approaches to sample from the posterior, including those based on conditional generative adversarial networks (CGANs) [61, 2], conditional variational autoencoders (CVAEs) [43, 114], conditional normalizing flows (CNFs) [11, 121], and score/Langevin/diffusion-based approaches [64, 78, 58].

Here, we focus on the CNF approach. Compared to the other methods, CNFs yield rapid inference and require only simple, likelihood-based training. In a recent super-resolution (SR) contest [84], a CNF (by Song et al. [106]) won, beating all CGAN, CVAE, and diffusion-based competitors.

Inspired by the success of CNFs in SR, we design the first CNF for accelerated multi-coil MRI. Previous applications of CNFs to MRI [37] showed competitive results but were restricted to single-coil recovery of magnitude images. As the vast majority of modern MRI scanners capture multi-coil data, the extension to multi-coil, complex-valued data is crucial for real-world adoption. However, the order-of-magnitude increase in dimensionality makes this transition non-trivial. For this purpose, we propose a novel CNF that infers only the signal component in the nullspace of the measurement operator and combines its output with the measured data to generate complete images. Using fastMRI brain and knee data, we demonstrate that our approach outperforms existing posterior samplers based on CGANs [2] and MRI-specific score/Langevin-based approaches [62, 34] in almost all accuracy metrics, while retaining fast inference and requiring minimal hyperparameter tuning.

3.1 Background

In the case of MRI, the posterior distribution that we would ultimately like to sample from is $p_{\iota_{\mathsf{true}}|\kappa}(\cdot|\kappa)$, where $\kappa \triangleq [\kappa_{(1)}^\top, \ldots, \kappa_{(B)}^\top]^\top$. Equivalently, we could consider $p_{\iota_{\mathsf{true}}|y}(\cdot|y)$ since y and κ contain the same information. Another option is to sample from $p_{x|y}(\cdot|y)$ and then use (2.4) or (2.5) to combine coil images into a single image. We take the latter approach.

For CNFs and CGANs, posterior sampling is accomplished by designing a neural network that maps samples from an easy-to-generate latent distribution (e.g., white Gaussian) to the target distribution (i.e., the distribution of x given y, with density $p_{x|y}$). Once that network is trained, sample generation is extremely fast. For Langevin dynamic and score-based methods, an algorithm is run for hundreds or thousands of iterations to generate each sample, and each iteration involves calling a neural network. Consequently, the inference time is much longer than that of CNFs and CGANs.

Normalizing flows (NF) [40, 41, 73, 92] have emerged as powerful generative models capable of modeling complex data distributions. Normalizing flows learn an invertible mapping between a target data distribution and a simple latent distribution, generally a Gaussian. More concretely, for a latent sample v drawn from the latent distribution p_v, the normalizing flow defines an invertible transformation $h_\theta(\cdot) : \mathbb{R}^Q \to \mathbb{R}^Q$. This transformation is parameterized by θ, and $x = h_\theta(v)$ defines a sample in the target data domain. This mapping of the latent distribution induces a distribution in the target data domain with a probability density derived from the change-of-variable formula

$$\widehat{p}_x(x;\theta) = p_v(h_\theta^{-1}(x))\left|\det\left(\frac{\partial h_\theta^{-1}(x)}{\partial x}\right)\right|, \tag{3.1}$$

where $\det(\cdot)$ denotes the determinant. The goal of the normalizing flow is to approximate the underlying data distribution p_x with $\widehat{p}_x(\cdot;\theta)$. Given a set of data samples $\{x_i\}_{i=1}^{n_{\mathsf{train}}}$, the parameters θ can be fit using a maximum likelihood loss

$$L(\theta) = \sum_{i=1}^{n_{\mathsf{train}}} \ln \widehat{p}_x(x_i;\theta) \tag{3.2}$$

$$= \sum_{i=1}^{n_{\mathsf{train}}} \ln p_v(h_\theta^{-1}(x_i)) + \ln\left|\det\left(\frac{\partial h_\theta^{-1}(x_i)}{\partial x^{(i)}}\right)\right| \tag{3.3}$$

Once the training is complete, samples from the target distribution can be rapidly generated by drawing samples from the latent distribution and passing them through the normalizing flow h_θ.

It is worth noting that maximizing $L(\theta)$ is equivalent to minimizing the Kullback-Leibler (KL) divergence between $\widehat{p}_x(\cdot;\theta)$ and p_x [92], which aligns with the goal of approximating p_x with $\widehat{p}_x(\cdot;\theta)$. The maximum-likelihood loss provides stable training with minimal hyperparameter tuning and has been shown to be robust to mode collapse.

Conditional normalizing flows (CNFs) [9] generalize normalizing flows by adding a conditioning signal y. With the CNF denoted as $h_\theta(\cdot,\cdot) : \mathbb{R}^Q \times \mathbb{R}^Q \to \mathbb{R}^Q$, the forward process from the latent domain to the data domain is given by $x = h_\theta(v, y)$. For complex-valued, multi-coil MRI, we have $Q = 2BD$. The inclusion of y alters the objective of the CNF to approximating the unknown posterior distribution $p_{x|y}(\cdot|y)$ with $\widehat{p}_{x|y}(\cdot|y;\theta)$. As before, the change-of-variable formula implies the induced distribution

$$\widehat{p}_{x|y}(x|y;\theta) = p_v(h_\theta^{-1}(x,y))\left|\det\left(\frac{\partial h_\theta^{-1}(x,y)}{\partial x}\right)\right|, \tag{3.4}$$

where h_θ^{-1} refers to the inverse mapping of h_θ with respect to its first argument.

Given a dataset $\{(x_i, y_i)\}_{i=1}^{n_{\text{train}}}$, the maximum likelihood loss can be utilized to optimize the parameters θ

$$L(\theta) = \sum_{i=1}^{n_{\text{train}}} \ln \widehat{p}_{x|y}(x_i|y_i;\theta) \tag{3.5}$$

$$= \sum_{i=1}^{n_{\text{train}}} \ln p_v(h_\theta^{-1}(x_i,y_i)) + \ln\left|\det\left(\frac{\partial h_\theta^{-1}(x_i,y_i)}{\partial x_i}\right)\right|. \tag{3.6}$$

CNFs have shown promising performance in solving inverse problems, such as super-resolution [85, 72, 106], making it an exciting avenue of exploration for accelerated

MRI. Denker et al. [37] developed a CNF for single-coil, magnitude-only knee images. This study showed promising initial results, but the limited scope did not demonstrate performance in the more realistic multi-coil, complex-valued domain. As this transition increases the dimensionality by an order of magnitude, non-trivial architectural changes are required. We build on the latest advances in CNFs to create a method that is capable of generating high-quality posterior samples of multi-coil, complex-valued MRI images.

3.2 Proposed Method

Our CNF consists of two networks, a conditioning network $h_{\theta_1}^{(\text{cond})}$ and a conditional flow model $h_{\theta_2}^{(\text{flow})}$. The conditioning network takes the vector of zero-filled (ZF) coil-images y as input and produces features that are used as conditioning information by the flow model $h_{\theta_2}^{(\text{flow})}$. Aided by the conditioning information, $h_{\theta_2}^{(\text{flow})}$ learns an invertible mapping between samples in the latent space and those in the image space. Using the notation of Sec. 3.1, our overall CNF takes the form

$$h_\theta(v, y) \triangleq h_{\theta_2}^{(\text{flow})}(v, h_{\theta_1}^{(\text{cond})}(y)) \tag{3.7}$$

where $\theta = [\theta_1, \theta_2]$.

Recently, advancements of CNFs in the super-resolution literature have revealed useful insights for more general inverse problems. First, Lugmayr et al. [85] suggested the use of a pretrained, state-of-the-art point-estimate network for the conditioning network $h_{\theta_1}^{(\text{cond})}$. This network is then trained jointly with $h_{\theta_2}^{(\text{flow})}$ using the loss in (3.6). This approach provides a functional initialization of $h_{\theta_1}^{(\text{cond})}$ and allows the conditioning network to learn to provide features that are useful for the maximum-likelihood training objective. We utilize a UNet from Zbontar et al. [125] for $h_{\theta_1}^{(\text{cond})}$ since it has

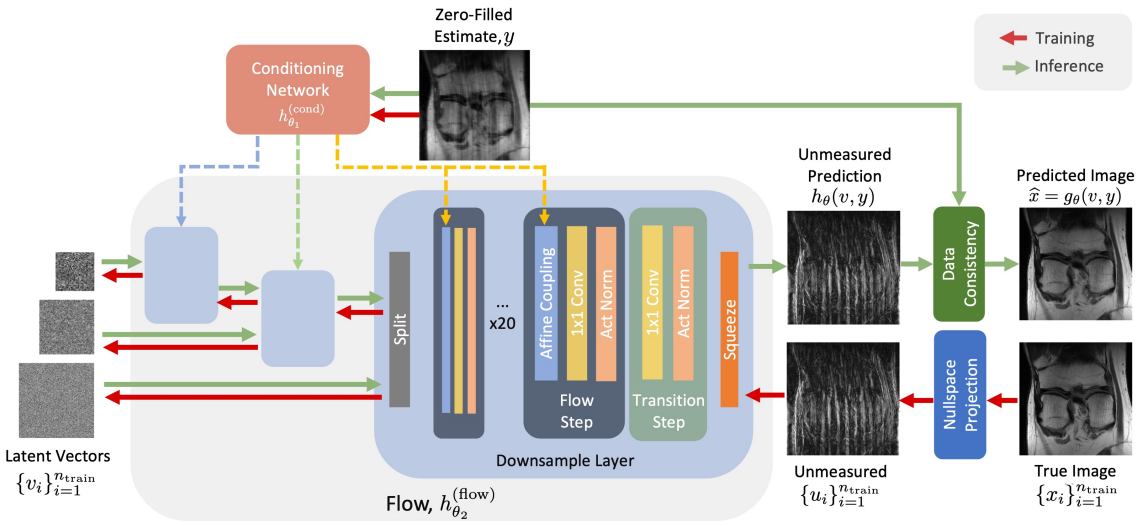


Figure 3.1: The architecture of our CNF. The conditioning network $h_{\theta_1}^{(\text{cond})}$ takes in multi-coil zero-filled image estimates y and outputs features used by the flow model $h_{\theta_2}^{(\text{flow})}$. The flow learns an invertible mapping between Gaussian random samples v_i and images u_i that are the projections of the training images x_i onto the non-measured subspace. During inference, data consistency (3.10) is applied so the final prediction $\widehat{x}$ is consistent with the observed measurements.

been shown to perform well in accelerated MRI. We first pre-train $h_{\theta_1}^{(\text{cond})}$ for MRI recovery, and later we jointly train $h_{\theta_1}^{(\text{cond})}$ and $h_{\theta_2}^{(\text{flow})}$ together.

Song et al. [106] demonstrated the benefits of using "frequency-separation" when training a CNF for super-resolution. The authors argue that the low-resolution conditional image already contains sufficient information about the low-frequency components of the image, so the CNF can focus on recovering only the high-frequency information. The CNF output is then added to an upsampled version of the conditional image to yield an estimate of the full image.

We now generalize the frequency-separation idea to arbitrary linear models of the form $y = Ax + \epsilon$ from (2.1) and apply the resulting procedure to MRI. Notice that

(2.1) implies

$$A^+y = A^+Ax + A^+\epsilon \tag{3.8}$$

where $(\cdot)^+$ denotes the pseudo-inverse. Here, A^+Ax is recognized as the projection of x onto the row-space of A, which we will refer to as the "measured space." Then

$$u \triangleq (I - A^+A)x \tag{3.9}$$

would be the projection of x onto its orthogonal complement, which we refer to as the "nullspace." Assuming that the nullspace has dimension > 0, we propose to construct an estimate $\widehat{x}$ of x with the form

$$\widehat{x} = g_\theta(v, y) = (I - A^+A)h_\theta(v, y) + A^+y, \tag{3.10}$$

where $g_\theta(v, y)$ is the complete estimation model, $h_\theta(v, y)$ is our CNF-generated estimate of u and the $(I - A^+A)$ in (3.10) strips off any part of $h_\theta(v, y)$ that has leaked into the measured space. A similar approach was used in [105] for point estimation. Given training data $\{(x_i, y_i)\}_{i=1}^{n_{\text{train}}}$, the CNF $h_\theta(\cdot, \cdot)$ is trained to map code vectors $v_i \sim p_v$ to the nullspace projections

$$u_i \triangleq (I - A^+A)x_i \tag{3.11}$$

using the measured data y_i as the conditional information. As a result of (3.10), the reconstructions $\widehat{x}$ agree with the measurements y in that $A\widehat{x} = y$. However, this also means that $\widehat{x}$ inherits the noise ϵ corrupting y, and so this data-consistency procedure is best used in the low-noise regime. In the presence of significant noise, the dual-decomposition approach [29] may be more appropriate.

In the accelerated MRI formulation (2.2)-(2.3), the matrix A is itself an orthogonal projection matrix, so that, in (3.10),

$$I - A^+A \;\; = \;\; \text{blkdiag}\big\{F^{\mathsf{H}}\widetilde{P}^\top\widetilde{P}F, \ldots, F^{\mathsf{H}}\widetilde{P}^\top\widetilde{P}F\big\}, \tag{3.12}$$

where $\widetilde{P} \in \mathbb{R}^{(D-M)\times D}$ is the sampling matrix for the non-measured k-space. Also, y is in the row-space of A, so

$$A^+y = y \tag{3.13}$$

in (3.10). Figure 3.1 illustrates the overall procedure, using "data consistency" to describe (3.10) and "nullspace projection" to describe (3.11). In Sec. 3.3.2, we quantitatively demonstrate the improvements gained from designing our CNF to estimate only the nullspace component.

3.2.1 Architecture

The backbone of $h_{\theta_1}^{(\text{cond})}$ is a UNet [98] that mimics the design in Zbontar et al. [125], with 4 pooling layers and 128 output channels in the first convolution layer. The first layer was modified to accept complex-valued coil images. The inputs have $2B$ channels, where B is the number of coils each with a real and imaginary component. The outputs of the final feature layer of the UNet are processed by a feature-extraction network with $\vartheta_{\text{layers}}$ convolution layers. Together, the feature extraction network and the UNet make up our conditioning network $h_{\theta_1}^{(\text{cond})}$. The output of each convolution layer is fed to conditional coupling blocks of the corresponding layer in $h_{\theta_2}^{(\text{flow})}$.

For the flow model $h_{\theta_2}^{(\text{flow})}$, we adopt the multi-scale RealNVP [41] architecture. This construction utilizes $\vartheta_{\text{layers}}$ layers and ϑ_{steps} flow steps in each layer. A flow step consists of an activation normalization [73], a fixed 1×1 orthogonal convolution

[11], and a conditional coupling block [9]. Each layer begins with a checkerboard downsampling (squeeze layer) [41] and a transition step made up of an activation normalization and 1×1 convolution. Layers end with a split operation that sends half of the channels directly to the output on the latent side. For all experiments, we use $\vartheta_{\text{layers}} = 3$ and $\vartheta_{\text{steps}} = 20$. The full architecture of g_θ is specified in Fig. 3.1.

Although the code that accompanies Denker et al. [37] gives a built-in mechanism to scale their flow architecture to accommodate an increased number of input and output channels, we find that this mechanism does not work well (see Sec. 3.3.2). Thus, in addition to incorporating nullspace learning, we redesign several aspects of the flow architecture and training. First, to prevent the number of flow parameters from growing unreasonably large, our flow uses fewer downsampling layers (3 vs 6) but more flow steps per downsampling layer (20 vs 5), and we utilize one-sided (instead of two-sided) affine coupling layers. Second, to connect the conditioning network to the flow, Denker et al. [37] used a separate CNN for each flow layer and adjusted its depth to match the flow-layer dimension. We use a single, larger CNN and feed its intermediate features to the flow layers with matched dimensions, further preventing an explosion in the number of parameters. Third, our conditioning network uses a large, pretrained UNet, whereas Denker et al. [37] used a smaller untrained UNet. With our modifications, we grow the conditional network more than the flow network, which allows the CNF to better handle the high dimensionality of complex-valued, multi-coil data.

3.2.2 Data

We apply our network to two datasets: the fastMRI knee and fastMRI brain datasets [125]. For the knee data, we use the non-fat-suppressed subset, giving 17286 training and 3592 validation images. We compress the measurements to $B = 8$ complex-valued virtual coils [127] and crop the images to 320×320 pixels. The sampling mask is generated using the golden ratio offset (GRO) [63] Cartesian sampling scheme with an acceleration rate $R = 4$ and autocalibration signal (ACS) region of 13 pixels.

With the brain data, we use the T2-weighted images and take the first 8 slices of all volumes with at least 8 coils. This provides 12224 training and 3352 validation images. The data is compressed to $B = 8$ virtual coils [127] and cropped to 384×384 pixels. The GRO sampling scheme is again used with an acceleration rate $R = 4$ and a 32-wide ACS region. For both methods, the coil-sensitivity maps are estimated from the ACS region using ESPIRiT [115]. All inputs to the network are normalized by the 95th percentile of the ZF magnitude images.

3.2.3 Training

For both datasets, we first train the UNet in $h_{\theta_1}^{(\text{cond})}$ with an additional 1×1 convolution layer to get the desired $2B$ channels. We train the UNet to minimize the mean-squared error (MSE) from the nullspace projected targets $\{u_i\}_{i=1}^{n_{\mathsf{train}}}$ for 50 epochs with batch size 8 and learning rate 0.003. Then, we remove the final 1×1 convolution and jointly train $h_{\theta_1}^{(\text{cond})}$ and $h_{\theta_2}^{(\text{flow})}$ for 100 epochs to minimize the negative log-likelihood (NLL) loss of the nullspace projected targets. For the brain data, we use batch size 8 and learning rate 0.0003. For the knee data, we use batch size 16 with learning rate 0.0005. All experiments use the Adam optimizer [74] with default

parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The full training takes about 4 days on 4 Nvidia V100 GPUs.

3.2.4 Comparison Methods

We compare against other methods that are capable of generating posterior samples for accelerated MRI. For the fastMRI brain data, we present results for the CGAN from Adler & Öktem [2] and the Langevin method from Jalal et al. [62]. For the fastMRI knee data, we present results for the "Score" method from Chung & Ye [34] and the "sCNF" method from Denker et al. [37].

For the CGAN, we utilize a UNet-based generator with 4 pooling layers and 128 output channels in the initial layer and a 5-layer CNN network for the discriminator. The generator takes y concatenated with a latent vector v as input. The model is trained with the default loss and hyperparameters from the authors' implementation [2] for 100 epochs with a learning rate of 0.001. For the Langevin method, we use the authors' implementation [62] but with the GRO sampling mask described in Sec. 3.2.2.

The Score method is different than the other methods in that it assumes that the k-space measurements κ are constructed from true coil images x with magnitudes affinely normalized to the interval $[0, 1]$ and phases normalized to $[0, 1]$ radians. Although this normalization cannot be enforced on prospectively undersampled MRI data, Score fails without this normalization. So, to evaluate Score, we normalize each $\kappa_{(b)}$ using knowledge of the ground-truth $x_{(b)}$, run Score, and un-normalized its output $\widehat{x}_{(b)}$ for comparison with the other methods. Since the Score paper [34] used RSS combining to compute $|\widehat{\imath}|$, we do the same. For the Score method, we use $T = 200$ iterations and not the default value of $T = 2000$. This is because, when using posterior-sample

averaging (see Sec. 3.2.5), the PSNR computed using 200 iterations is better than with 2000.

The sCNF method works only on single-coil magnitude data, and so we convert our multi-coil data to that domain in order to evaluate sCNF. To do this, we apply RSS (2.5) to ZF coil-images y and repeat the process for the true coil images x. Using those magnitude images, we train sCNF for 300 epochs with learning rate 0.0005 and batch size 32.

3.2.5 Evaluation

We report results for several different metrics, including peak-signal-to-noise ratio (PSNR), structural-similarity index (SSIM) [120], Fréchet Inception Score (FID) [57], and conditional FID (cFID) [104]. PSNR and SSIM were computed on the average of p posterior estimates $\{\widehat{\imath}^{(j)}\}_{j=1}^{p}$, i.e.,

$$\overline{\imath}_{[p]} \triangleq \frac{1}{p}\sum_{j=1}^{p}\widehat{\imath}^{(j)} \tag{3.14}$$

to approximate the posterior mean, while FID and cFID were evaluated on individual posterior samples $|\widehat{\imath}^{(j)}|$. By default, we compute all metrics using magnitude reconstructions $|\widehat{\imath}|$ rather than the complex-valued reconstructions $\widehat{\imath}$, in part because competitors like sCNF generate only magnitude reconstructions, but also because this is typical in the MRI literature (e.g., the fastMRI competition [125]). So, for example, PSNR is computed as

$$\text{PSNR} \triangleq 10\log_{10}\left(\frac{D\max_d |[\iota_{\mathsf{true}}]_d|^2}{\left\||\overline{\imath}_{[p]}| - |\iota_{\mathsf{true}}|\right\|_2^2}\right), \tag{3.15}$$

where $[\cdot]_d$ extracts the dth pixel. For FID and cFID, we use the embeddings of VGG-16 [103] as Kastryulin et al. [67] found that this helped the metrics better correlate with the rankings of radiologists.

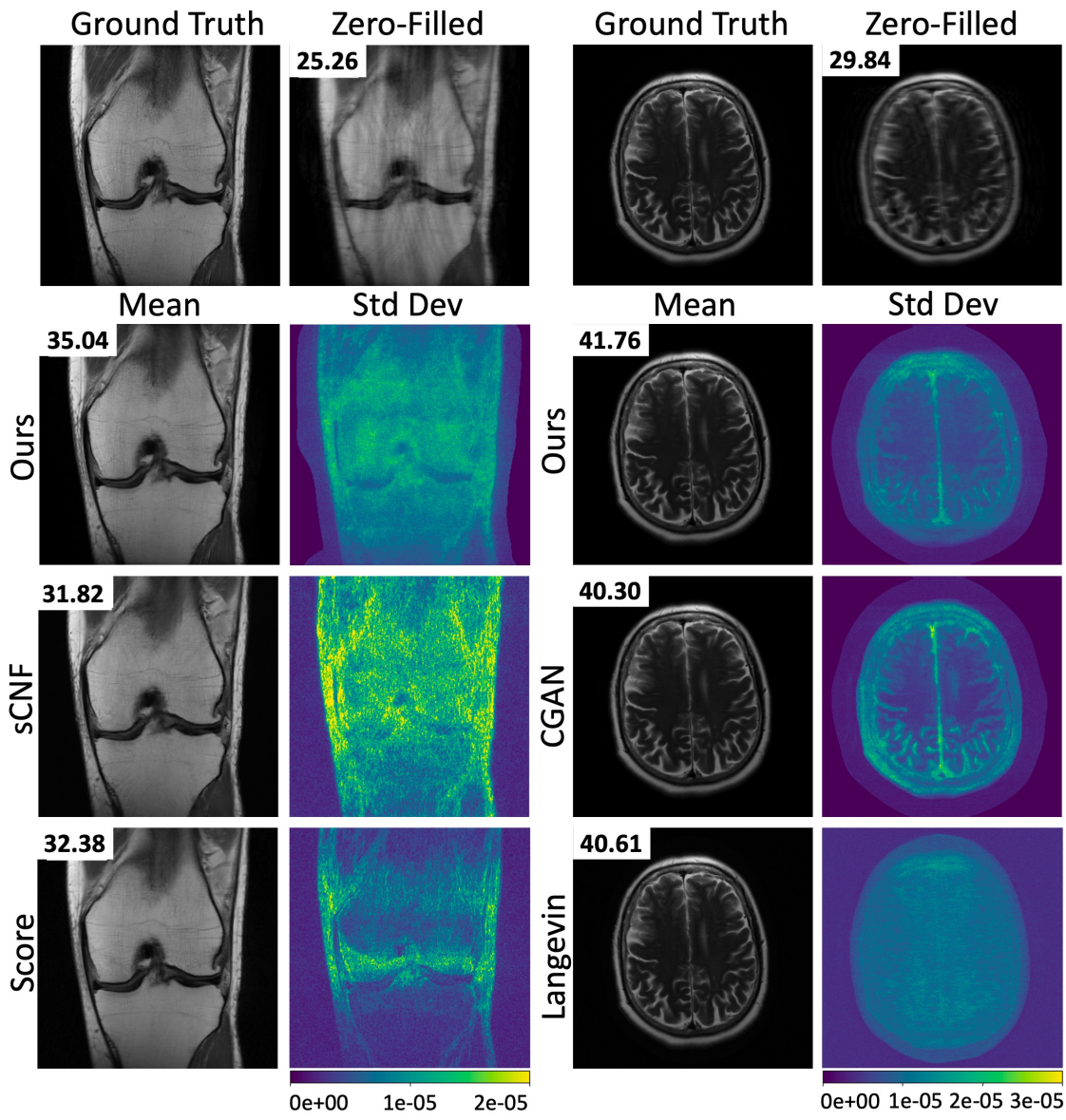


Figure 3.2: Mean images and pixel-wise standard-deviation maps computed from 8 and 32 posterior samples for the brain images and knee datasets, respectively. The standard-deviation maps show which pixels have the greatest reconstruction uncertainty. The corresponding PSNR is shown on each reconstruction.

Model	PSNR (dB)↑	SSIM↑	$\text{FID}^1\downarrow$	$\text{FID}^2\downarrow$	$\text{cFID}^1\downarrow$	$\text{cFID}^2\downarrow$	Time
Score	34.15 ± 0.19	0.8764 ± 0.0036	**4.49**	—	4.49	—	15 min
sCNF	32.93 ± 0.17	0.8494 ± 0.0047	7.32	5.78	8.49	6.51	**66 ms**
Ours	**35.23 ± 0.22**	**0.8888 ± 0.0046**	4.68	**2.55**	**3.96**	**2.44**	108 ms

Table 3.1: Average performance on non-fat-suppressed fastMRI knee data, with standard error reported after the ±. PSNR, SSIM, FID^1, and cFID^1 are computed for 72 test images and $p = 8$ posterior samples. FID^2, and cFID^2 are computed for 2188 test samples and $p = 8$ posterior samples. Time marks the generation time for one posterior sample.

For the brain data, we compute all metrics on 72 random test images in order to limit the Langevin image generation time to 4 days. We generate complex-valued images using the coil-combining method in (2.4) before computing the magnitude and use $p = 32$ posterior samples to calculate cFID^1, FID^1, PSNR, and SSIM. (For the reference statistics of FID, we use the entire training dataset.) Because FID and cFID are biased by small sample sizes, we also compute FID^2 and cFID^2 with 2484 test samples and $p = 8$ for our method and the CGAN.

With the knee data, we follow a similar evaluation procedure except that, to comply with the evaluation steps of Score, we generate magnitude-only signals using the root-sum-of-square (RSS) combining from (2.5). Also, we computed metrics on 72 randomly selected slices in order to bound the image generation time of Score to 6 days with $p = 8$. We use $p = 8$ for all metrics, but for FID^2 and cFID^2, we use 2188 test samples.

When computing inference time for all methods, we use a single Nvidia V100 with 32GB of memory and evaluate the time required to generate one posterior sample.

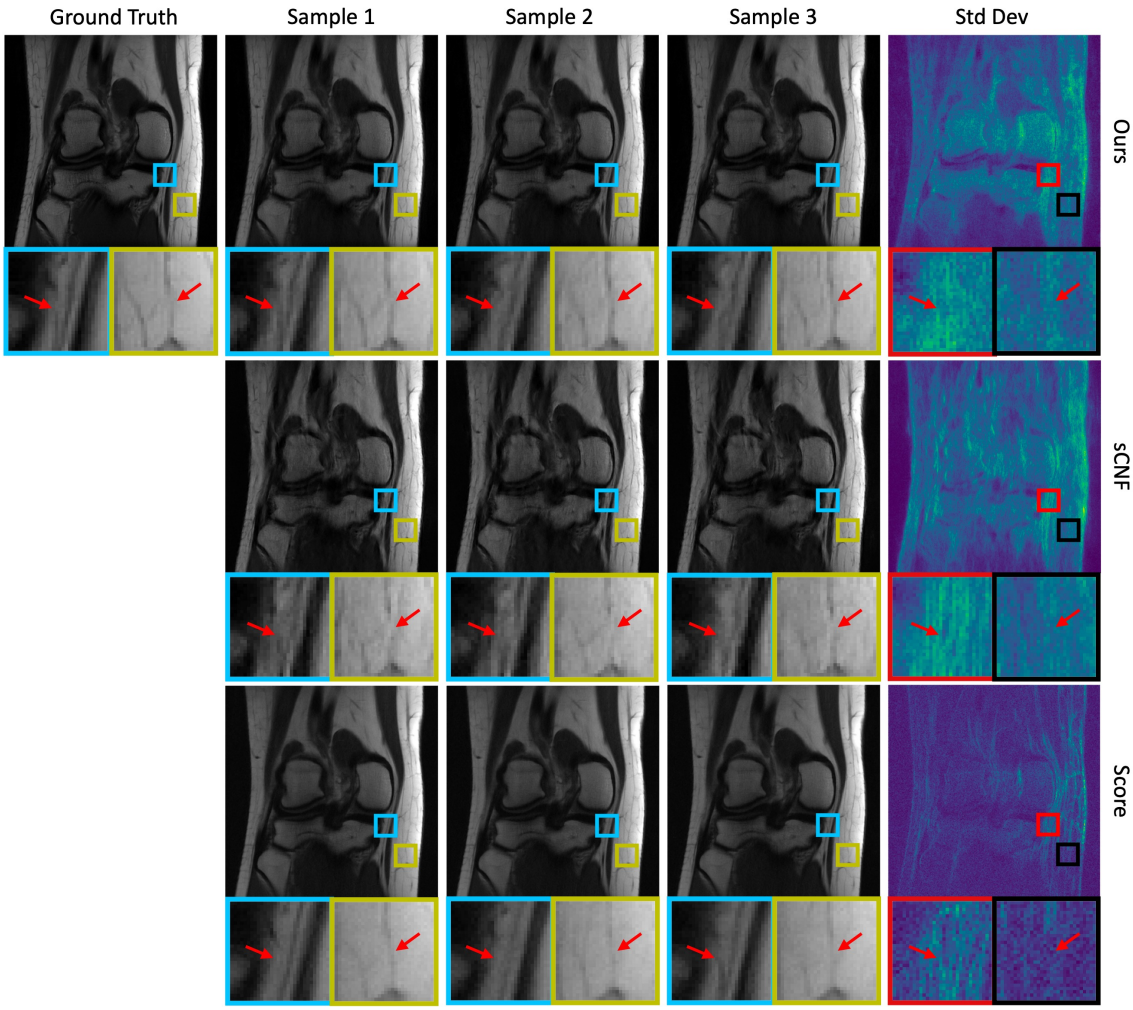


Figure 3.3: Examples of posterior samples and standard-deviation maps for the knee data. The samples show important structural variations. This demonstrates the advantages of generating multiple reconstructions and computing a pixel-wise standard-deviation map.

3.3 Results

Tab. 3.1 reports the quantitative metrics for the knee dataset. It shows that our method outperforms sCNF by a significant margin in all metrics except inference time. By using information from multiple coils and a more advanced architecture,

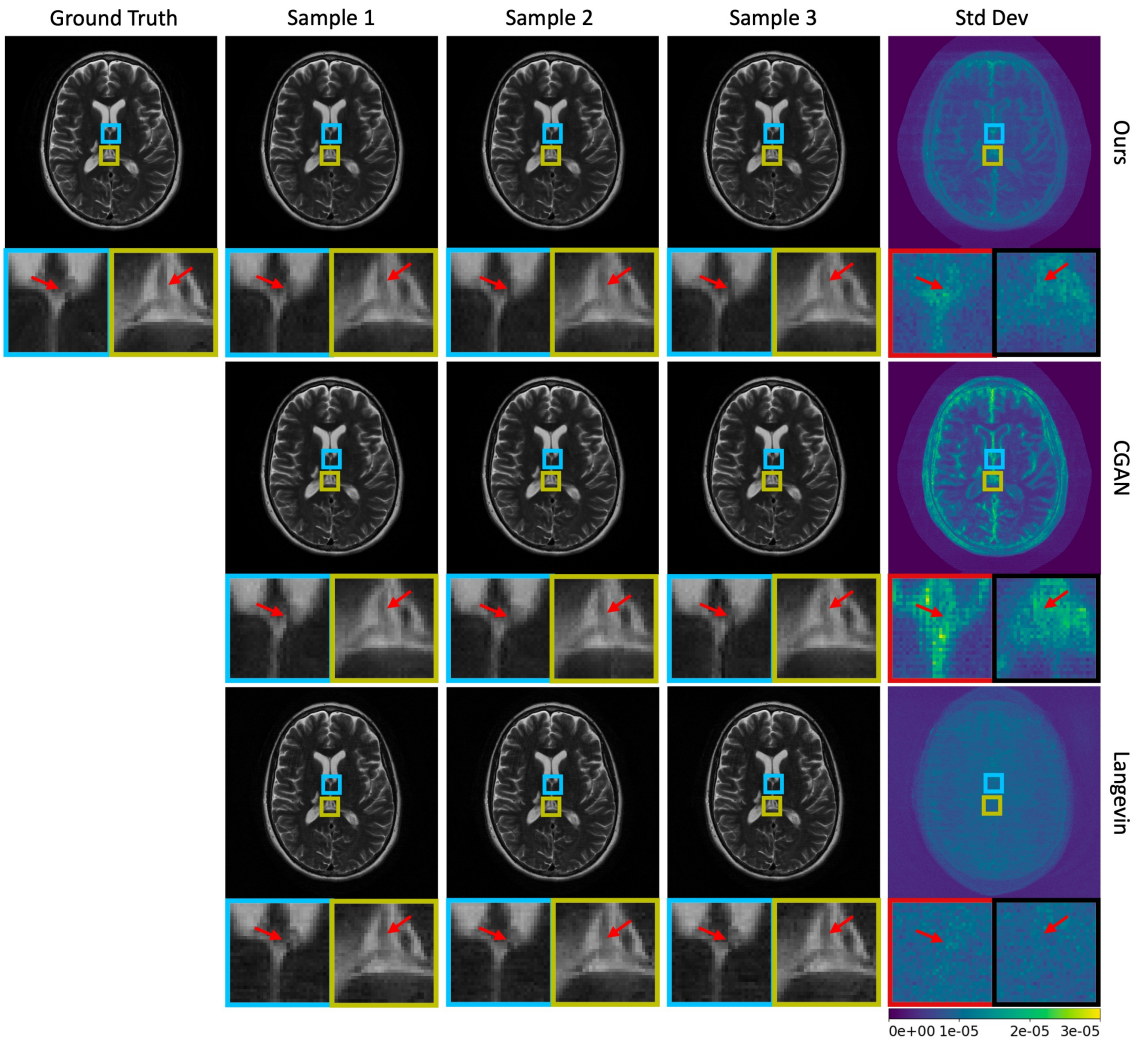


Figure 3.4: Examples of posterior samples and standard-deviation maps for the brain images, both with zoomed regions.

Model	PSNR (dB)↑	SSIM↑	FID^1↓	FID^2↓	cFID^1↓	cFID^2↓	Time
Langevin	37.88 ± 0.41	0.9042 ± 0.0062	6.12	—	5.29	—	14 min
CGAN	37.28 ± 0.19	0.9413 ± 0.0031	5.38	4.06	6.41	4.28	**112 ms**
Ours	**38.85 ± 0.23**	**0.9495 ± 0.0012**	**4.13**	**2.37**	**4.15**	**2.44**	177 ms

Table 3.2: Average performance on T2-weighted fastMRI brain data, with standard error reported after the ±. PSNR, SSIM, FID^1, and cFID^1 are computed for 72 test images and $p = 32$ posterior samples. FID^2 and cFID^2 are computed using 2484 test samples and $p = 8$. Time marks the generation time for one posterior sample.

our method shows the true competitive potential of CNFs in realistic accelerated MR imaging.

Tab. 3.1 also shows that our method surpasses Score in all metrics except FID^1, even though Score benefited from impractical ground-truth normalization. Compared to Score, our method generated posterior samples 8000× faster. Furthermore, our method (and sCNF) will see a speedup when multiple samples are generated because the conditioning network $h_{\theta_1}^{(\text{cond})}$ needs to be evaluated only once per p generated samples for a given y. For example, with the knee data, we are able to generate $p = 32$ samples in 1.41 seconds, corresponding to 44 milliseconds per sample, which is a 2.5× speedup over the value reported in Tab. 3.1.

Tab. 3.2 reports the quantitative results for the brain dataset. The table shows that we outperform the Langevin and CGAN methods in all benchmarks except inference time. While our method is a bit slower than the CGAN, it is orders of magnitude faster than the Langevin approach.

We show the mean images and standard-deviation maps for the fastMRI knee and brain experiments in Fig. 3.2. For the knee data, our method captures texture more accurately than the sCNF method and provides a sharper representation than the

Score method. All of the brain methods provide a visually accurate representation to the ground truth, but the Langevin method provides a more diffuse variation map, with energy spread throughout the image.

In Fig. 3.3 and Fig. 3.4, we plot multiple posterior samples, along with zoomed-in regions, to illustrate the changes across independently drawn samples for each method. The standard-deviation maps are generated using $p = 8$ posterior samples, three of which are shown. From the zoomed-in regions, it can be seen that several samples are consistent with the ground truth while others are not (although they may be consistent with the measured data). Regions of high posterior variation can be flagged from visual inspection of the standard-deviation map and further investigated through viewing multiple posterior samples for improved clinical diagnoses.

Our method presents observable, realistic variations of small anatomical features in the zoomed-in regions. The variations are also registered in the standard-deviation map, which illustrates the pixel-wise uncertainty. Both the posterior samples and the standard-deviation map could be used by clinicians to assess their findings. Comparatively, our method demonstrates variation that is spread across the entire knee image, while in the Score method, the variation is mostly localized to small regions. The sCNF also demonstrates variation, but it is mostly driven by residual aliasing artifacts. For the brain images, the Langevin method again gives a very diffuse standard-deviation map with no discernible features. Both our method and the CGAN highlight particular regions of high variation although the CGAN map indicates much larger standard-deviation values. Since it is difficult to say which standard-deviation map is more useful or correct, the interpretation of these maps could be an interesting direction for future work.

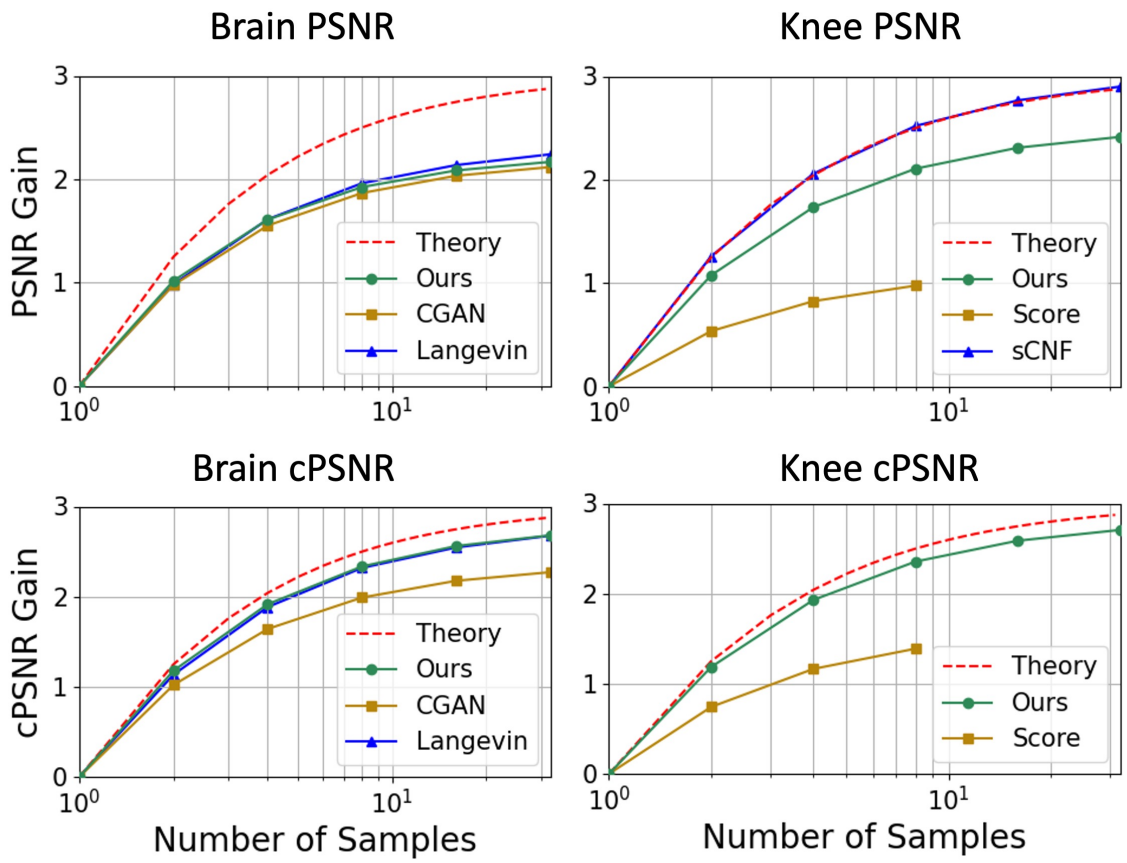


Figure 3.5: The gain in (magnitude) PSNR and complex PSNR of the p-sample mean estimate $\bar{\iota}_{[p]}$ versus p, for both brain and knee data. Note the ≈ 3 dB increase as p grows from 1 to infinity.

3.3.1 PSNR Gain versus Number of Posterior Samples

It is well known that the minimum mean-squared error (MMSE) estimate of ι from y equals the conditional mean $\mathrm{E}\{\iota|y\}$, i.e., the mean of the posterior distribution $p_{\iota|y}(\cdot|y)$. Thus, one way to approximate the MMSE estimate is to generate many samples from the posterior distribution and average them, as in (3.14). Bendel et al.

[19] showed that the MSE

$$\mathcal{E}_p \triangleq \mathrm{E}\left[\|\bar{\iota}_{[p]} - \iota_{\mathsf{true}}\|_2^2 \big| y\right] \tag{3.16}$$

of the p-posterior-sample average $\bar{\iota}_{[p]}$ obeys $\mathcal{E}_1/\mathcal{E}_p = 2p/(p+1)$. So, for example, the SNR increases by a factor of two as p grows from 1 to ∞. The same thing should happen for PSNR, as long as the PSNR definition is consistent with (3.16). For positive signals (i.e., magnitude images) the PSNR definition from (3.15) is consistent with (3.16), but for complex signals we must use "complex PSNR"

$$\text{cPSNR} \triangleq 10\log_{10}\left(\frac{D\max_d |[\iota_{\mathsf{true}}]_d|^2}{\|\bar{\iota}_{[p]} - \iota_{\mathsf{true}}\|_2^2}\right). \tag{3.17}$$

As RSS combining provides only a magnitude estimate, we compute the coil-combined estimate for our method and Score to evaluate cPSNR behavior for the knee dataset.

One may then wonder whether a given approximate posterior sampler has a PSNR gain versus p that matches the theory. In Fig. 3.5, we answer this question by plotting the PSNR gain and the cPSNR gain versus $p \in \{1, 2, 4, 8, 16, 32\}$ for the various methods under test (averaged over all 72 test samples). There we see that our method's cPSNR curve matches the theoretical curve well for both brain and knee data. As expected, our (magnitude) PSNR curve does not match the theoretical curve. The cPSNR curves of the Score and CGAN methods fall short of the theoretical curve by a large margin, but interestingly, the Langevin method's cPSNR curve matches ours almost perfectly. sCNF's PSNR gain curve matches the theoretical one almost perfectly, which provides further empirical evidence that CNF methods accurately sample from the posterior distribution.

Model	PSNR (dB) ↑	SSIM ↑	FID^2 ↓	cFID^2 ↓
Denker et al. [37]	17.61 ± 0.20	0.6665 ± 0.0072	16.02	16.68
+ Data Consistency	27.27 ± 0.21	0.7447 ± 0.0061	16.92	18.56
+ Architectural Changes	33.87 ± 0.23	0.8715 ± 0.0049	4.48	4.50
+ Nullspace Learning	**35.23 ± 0.22**	**0.8888 ± 0.0046**	**2.55**	**2.44**

Table 3.3: Ablation Study: Performance on non-fat-suppressed fastMRI knee data, with standard error reported after the ±. Each line adds a new contribution to the model of the previous line. Metrics are computed as described in Sec. 3.2.5

3.3.2 Ablation Study

To evaluate the impact of our contributions to CNF architecture and training design, we perform an ablation study using the fastMRI knee dataset. We start with the baseline model in Denker et al. [37], modified to take in 16 channels instead of 1, and scale it up using the built-in mechanism in the author's code. We train this model for 300 epochs with batch size 32 and learning rate 0.0001 to minimize the NLL of the multi-coil targets $\{x_i\}_{i=1}^{n_{\text{train}}}$, since higher learning rates were numerically unstable. Table 3.3 shows what happens when we add each of our contributions. First, we add data consistency (3.10) to the evaluation of the baseline. We then add the architectural changes described in Sec. 3.2.1, and finally we add nullspace learning to arrive at our proposed method. From Tab. 3.3, it can be seen that each of our design contributions yielded a significant boost in performance, and that nullspace learning was a critical ingredient in our outperforming the Score method in Tab. 3.1. For this ablation study, all models were trained following the procedure outlined in Sec. 3.2.3 (except for the learning rate of the baseline).

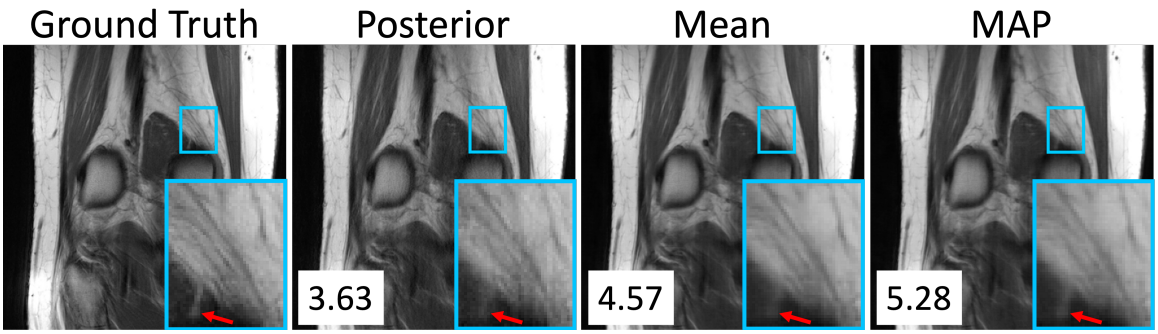


Figure 3.6: Examples of a ground-truth image, one posterior sample, an average of $p = 8$ posterior samples, and a MAP estimate. The log posterior density in units of bits-per-dimension is shown in the bottom right corner of each image.

3.3.3 Maximum a Posteriori (MAP) Estimation

Because CNFs can evaluate the posterior density of a signal hypothesis (recall (3.4)), they can be used for posteriori (MAP) estimation, unlike CGANs.

Due to our data-consistency step (3.10), we find the MAP estimate of x using

$$\widehat{x}_{\mathsf{MAP}} = \widehat{u}_{\mathsf{MAP}} + A^{+}y \tag{3.18}$$

$$\widehat{u}_{\mathsf{MAP}} = \arg \max_{u \in \mathrm{null}(A)} \ln \widehat{p}_{u|y}(u|y). \tag{3.19}$$

Note the CNF output u is constrained to the nullspace of A. From (3.12), this nullspace is spanned by the columns of

$$W \triangleq \mathrm{blkdiag}\left\{F^{\mathsf{H}}\widetilde{P}^{\top}, \ldots, F^{\mathsf{H}}\widetilde{P}^{\top}\right\}, \tag{3.20}$$

which are orthonormal, and so $\widehat{u}_{\mathsf{MAP}} = W\widetilde{\kappa}_{\mathsf{MAP}}$ with

$$\widetilde{\kappa}_{\mathsf{MAP}} = \arg\max_{\widetilde{\kappa}} \ln \widehat{p}_{u|y}(W\widetilde{\kappa}|y;\theta) \tag{3.21}$$

$$= \arg\max_{\widetilde{\kappa}} \left[\ln p_v(h_\theta^{-1}(W\widetilde{\kappa}, y)) + \ln\left|\det\left(\frac{\partial h_\theta^{-1}(\widetilde{u}, y)}{\partial \widetilde{u}}\bigg|_{\widetilde{u}=W\widetilde{\kappa}}\right)\right|\right]. \tag{3.22}$$

For this maximization, we use the Adam optimizer with 5000 iterations and a learning rate of 1×10^{-8}. Above, $\widetilde{\kappa}$ can be recognized as the unmeasured k-space samples.

In Figure 3.6, we show an example of a MAP estimate along with the ground truth image, one sample from the posterior, a $p = 8$ posterior-sample average, and their corresponding log-posterior-density values. As expected, the MAP estimate has a higher log-posterior-density that the other estimates. Visually, the MAP estimate is slightly sharper than the sample average but contains less texture details than the single posterior sample.

3.4 Conclusion

In this work, we present the first conditional normalizing flow for posterior sample generation in multi-coil accelerated MRI. To do this, we designed a novel conditional normalizing flow (CNF) that infers the signal component in the measurement operator's nullspace, whose outputs are later combined with information from the measured space. In experiments with fastMRI brain and knee data, we demonstrate improvements over existing posterior samplers for MRI. Compared to score/Langevin-based approaches, our inference time is four orders-of-magnitude faster. We also illustrate how the posterior samples can be used to quantify uncertainty in MR imaging on an individual pixel level. This provides radiologists with additional tools to enhance the robustness

of clinical diagnoses and serves as a foundational component for the methods presented in the subsequent chapters.

Chapter 4: Task-based UQ via Conformal Prediction

Pixel-wise uncertainty quantification has been a very common presentation of uncertainty in inverse imaging problems. As seen in Ch. 3, posterior sampling methods [42, 2, 10, 43, 62, 33] allow one to easily draw many samples from the distribution $p(x|y)$ of plausible x given y and construct a pixel-wise uncertainty map. This quantifies the uncertainty that the measurement process imposes on x, known as aleatoric uncertainty. Another approach is to utilize Bayesian Neural Networks (BNNs), which treat the reconstruction network parameters as random variables [71, 122, 14, 44, 90]. This allows one to quantify epistemic (i.e., model) uncertainty by measuring the variation over reconstructions generated by different draws from the parameter distribution. It's possible to combine BNNs with posterior sampling as well, as in Ekmekci & Cetin [45]. Recently, conformal prediction [117, 5] has also been shown to be a promising avenue as it computes pixel-wise uncertainty intervals with particular statistical guarantees [7, 60, 111, 77].

However, the value of these pixel-wise uncertainty maps is not clear. For example, when recovering images, we are usually concerned about many-pixel visual structures (e.g., lesions in MRI, hallucinations) that single-pixel statistics say little about. Secondly, uncertainty maps are not easy to interpret. They often convey little beyond the notion that there is less pixel-wise uncertainty in smooth regions as compared to near

edges (see, e.g., Figure 4.8). Lastly, it's not clear how pixel-wise uncertainty relates to the overall imaging goal, which is often task-oriented, such as detecting whether a tumor is present or not.

One approach to assess multi-pixel uncertainty is Bayesian Uncertainty Quantification by Optimization (BUQO) [110], which aims to test whether a particular "structure of interest" in the maximum a-posteriori (MAP) reconstruction is truly present. However, inpainting is used to hypothesize what the image would look like without the structure, the correctness of which is difficult to guarantee. Alternatively, Belhasin et al. [17] compute conformal prediction intervals on the principal components of the posterior covariance matrix. This allows for the visualization of the uncertainty on multi-pixel structures, but it is challenging to know if such structures are relevant to an imaging task like pathology detection.

In this work, we propose a novel UQ framework for imaging inverse problems that aims to provide a more impactful measure of uncertainty. In particular, we aim to quantify to what extent a downstream task behaves differently when supplied with the reconstructed image versus the true image. Our framework supports any measurement-and-reconstruction procedure and any downstream task that outputs a real-valued scalar. Our contributions are as follows.

1. We propose to construct, using conformal prediction, an interval in the task-output space that is guaranteed to contain the true task output up to a user-specified probability. The prediction interval width provides a natural way to quantify the uncertainty that measurement-and-reconstruction contributes to the downstream task output. (See Figure 4.1.)

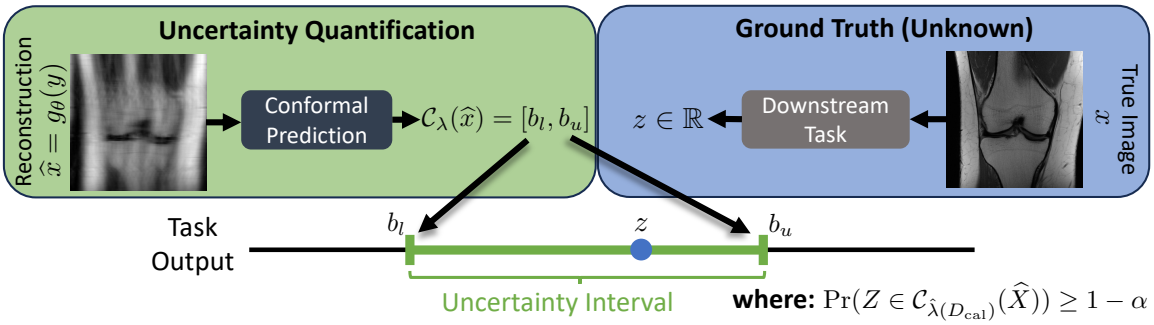


Figure 4.1: High-level overview of our approach: For true image x, measurement $y = \mathcal{A}(x)$, recovery $\widehat{x} = g_\theta(y)$, and task output $\widehat{z} = f(\widehat{x})$, we use conformal prediction to construct an interval $\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\widehat{x}) \subset \mathbb{R}$ that is guaranteed to contain the true task output $z = f(x)$ in the sense that $\Pr(Z \in \mathcal{C}_{\widehat{\lambda}(D_{\text{cal}})}(\widehat{X})) \geq 1 - \alpha$ for some chosen error-rate α.

2. For posterior-sampling-based image reconstruction, we propose to construct adaptive uncertainty intervals that shrink when the measurements offer more certainty about the true output of the downstream task. (See Figure 4.2.)

3. We propose a multi-round acquisition protocol whereby measurements are accumulated until the task uncertainty is acceptably low.

4. We demonstrate our approach on accelerated MRI with the task of soft-output-classifying a meniscus tear. Several conformal predictors are evaluated and compared.

4.1 Background

Conformal prediction [117, 5] is a framework for generating uncertainty sets with prescribed statistical guarantees. Notably, it can be applied to any black-box predictor without making any distributional assumptions about the data.

We now explain the basics of conformal prediction or, more precisely, the common variant known as split conformal prediction [91, 79]. Say that we have a black-box model $f : \mathcal{X} \to \mathcal{Z}$ that predicts a target $z \in \mathcal{Z}$ from features $x \in \mathcal{X}$. Say that we also have a calibration dataset $d_{\mathsf{cal}} \triangleq \{(x_i, z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ that was unseen when training f, as well as a test feature x_0 and an unknown test target z_0. In split conformal prediction, we define a prediction set $\mathcal{C}_\lambda(x_0) \subset 2^{\mathcal{Z}}$ that grows in size as the parameter λ increases and then use the calibration data to select a value of λ that provides the "marginal coverage" [80] guarantee

$$\Pr\left(Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(X_0)\right) \geq 1 - \alpha, \tag{4.1}$$

where $\alpha \in [0, 1]$ is a user-specified error rate. In (4.1) and the remainder of the chapter, we use capital letters to denote random variables and lower-case to denote their realizations. Thus (4.1) can be interpreted as follows: When averaged over random calibration data D_{cal} and test data (X_0, Z_0), the set $\mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(X_0)$ is guaranteed to contain the correct target Z_0 with probability no less than $1 - \alpha$. Although we would prefer a "conditional coverage" guarantee of the form $\Pr\left(Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(X_0) \middle| X_0 = x_0\right) \geq 1 - \alpha$, this is generally impossible to achieve [116, 80].

We now describe the standard recipe for constructing a prediction set $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(x_0)$ and selecting $\widehat{\lambda}(d_{\mathsf{cal}})$, a process known as calibration. First one chooses a nonconformity score $s(x, z; f) \in \mathbb{R}$ that assigns higher values to worse predictions. Then one computes the empirical quantile

$$\widehat{\lambda}(d_{\mathsf{cal}}) \triangleq \text{EmpQuant}\left(\tfrac{\lceil (1-\alpha)(n+1) \rceil}{n}; s_1, \ldots, s_{n_{\mathsf{cal}}}\right) \tag{4.2}$$

from the calibration scores $s_i = s(x_i, z_i; f)$. Finally one constructs

$$\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(x_0) = \{z : s(x_0, z; f) \leq \widehat{\lambda}(d_{\mathsf{cal}})\}. \tag{4.3}$$

Under these choices, it can be proven [118, 79] that the marginal coverage guarantee (4.1) holds when $(X_1, Z_1), \ldots, (X_n, Z_n), (X_0, Z_0)$ are i.i.d., and even under the weaker condition that they are exchangeable [117].

There are many ways to construct the nonconformity score $s(x, z; f)$. For real-valued targets z, the simplest choice would be the absolute residual

$$s(x, z; f) = |z - f(x)| \quad \Rightarrow \quad \mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(x) = \big[f(x) - \widehat{\lambda}(d_{\mathsf{cal}}), f(x) + \widehat{\lambda}(d_{\mathsf{cal}})\big], \tag{4.4}$$

which gives an x-invariant interval length of $|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(x)| = 2\widehat{\lambda}(d_{\mathsf{cal}})$. We will discuss a few other choices in the sequel. For more on conformal prediction, we suggest the excellent overviews [117, 5].

4.2 Proposed Method

Suppose that we collect measurements $y = \mathcal{A}(x)$ of a true image x as in a standard inverse imaging problem. Using an arbitrary recover method g_θ, we can compute an image recovery $\widehat{x} = g_\theta(y)$. Ideally, we would like that $\widehat{x} = x$, but this is impossible to guarantee with an ill-posed inverse problem. Although there are many ways to quantify the difference between $\widehat{x}$ and x (e.g., PSNR, SSIM [120], LPIPS [126], DISTS [39]), we will instead assume that we are primarily interested in using $\widehat{x}$ for some downstream task $f(\widehat{x}) \in \mathbb{R}$. As a running example, we consider x to be a medical image, y to be accelerated MRI measurements, and $f(\cdot) \in [0, 1]$ to be the soft output of a classifier that aims to detect the presence or absence of a pathology. For example, when $f(\widehat{x}) = 0.7$, the classifier believes that there is a 70% chance that the pathology exists.

When image recovery is imperfect (i.e., $\widehat{x} \neq x$), we expect the task output to also be imperfect, in the sense that $\widehat{z} = f(\widehat{x}) \neq f(x) = z$. We are thus strongly motivated

to understand how close $\widehat{z}$ is to the true z or, even better, to construct a prediction interval $\mathcal{C}_\lambda(\widehat{x}) \subset \mathbb{R}$ that contains the true z with some guarantee. The interval width $|\mathcal{C}_\lambda(\widehat{x})|$ would then quantify the uncertainty that the measurement-and-reconstruction process contributes to predicting the true task output z.

We emphasize that our approach makes *no assumptions about the task* $f(\cdot)$ beyond it producing a real number. For example, if $f(\cdot)$ is a soft-output classifier, we do not assume that it is accurate or even calibrated [53]. Likewise, our approach does not aim to assess the uncertainty implicit in the task, but rather the *additional uncertainty that measurement-and-reconstruction contributes to the task*. For a soft-output classifier, a (true) output of $z = f(x) = 0.7$ would express considerable uncertainty about the presence of a pathology in x. But if the true z could be perfectly predicted from $\widehat{x}$, then the measurement-and-reconstruction process would bring no *additional* uncertainty.

To construct the interval $\mathcal{C}_\lambda(\widehat{x})$, we use conformal prediction. Adapting the methodology from Section 4.1 to the current setting, we use a calibration set $d_{\mathsf{cal}} = \{(\widehat{x}_i, z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ of (recovered-image, true-task-output) pairs, and we expect to satisfy the marginal coverage guarantee

$$\Pr\left(Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{X}_0)\right) \geq 1 - \alpha \tag{4.5}$$

when $(\widehat{X}_1, Z_1), \ldots, (\widehat{X}_n, Z_n), (\widehat{X}_0, Z_0)$ are exchangeable. In (4.5) and in the sequel, we explicitly denote the dependence of $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\widehat{x}_0)$ on the calibration data. To construct the calibration set d_{cal}, we assume access to ground-truth examples $\{x_i\}_{i=1}^{n_{\mathsf{cal}}}$, a measurement model $\mathcal{A}(\cdot)$, a reconstruction model $g_\theta(\cdot)$, and a task function $f(\cdot)$. From these, we can construct $y_i = \mathcal{A}(x_i)$, $\widehat{x}_i = g_\theta(y_i)$, and $z_i = f(x_i)$ for $i = 1, \ldots, n_{\mathsf{cal}}$.

In some cases we may instead have access to a posterior-sampling-based image reconstruction model (as in Ch. 3) that generates c recoveries $\{\widehat{x}_i^{(j)}\}_{j=1}^{c}$ from every

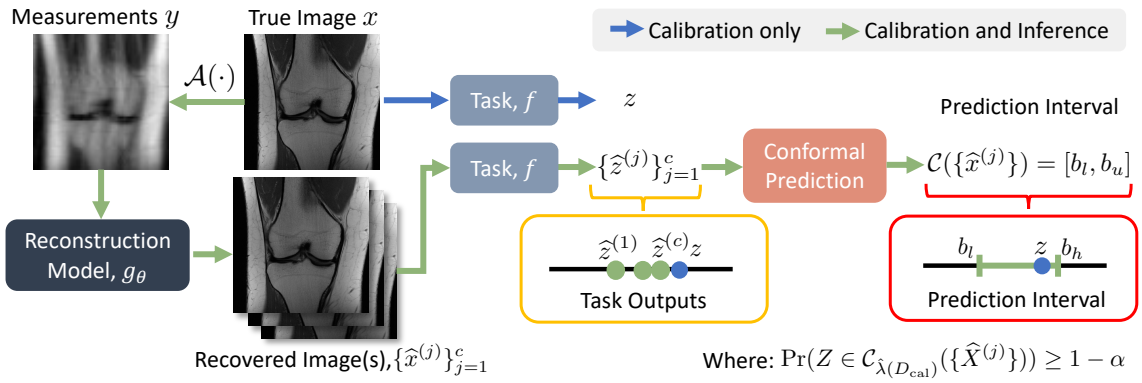


Figure 4.2: Detailed overview of our approach: For true image x, measurement $y = \mathcal{A}(x)$, reconstructions $\{\widehat{x}^{(j)}\}_{j=1}^{c}$, and task outputs $\widehat{z}^{(j)} = f(\widehat{x}^{(j)})$, we use conformal prediction with a calibration set $d_{\text{cal}} = \{(\{\widehat{x}_i^{(j)}\}_{j=1}^{c}, z_i)\}_{i=1}^{n_{\text{cal}}}$ to construct an interval $\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\{\widehat{x}^{(j)}\}) = [b_l, b_h]$ that is guaranteed to contain the true task output $z = f(x)$ in the sense that $\Pr(Z \in \mathcal{C}_{\widehat{\lambda}(D_{\text{cal}})}(\{\widehat{X}^{(j)}\})) \geq 1-\alpha$ for some chosen error-rate α.

measurement y_i via $\widehat{x}_i^{(j)} = g_\theta(v_i^{(j)}, y_i)$, where $\{v_i^{(j)}\}_{j=1}^{c}$ are i.i.d. code vectors and, typically, $v_i^{(j)} \sim \mathcal{N}(0, I)$. In this case, the prediction interval becomes $\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\{\widehat{x}_i^{(j)}\}_{j=1}^{c})$. As we will see, posterior sampling facilitates locally adaptive prediction sets.

Next we describe different ways to construct the prediction intervals $\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\widehat{x})$ and $\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\{\widehat{x}_i^{(j)}\}_{j=1}^{c})$, and later we describe a multi-round measurement protocol that exploits locally adaptive prediction intervals. See Figure 4.2 for a detailed overview of our approach.

4.2.1 Method 1: Absolute Residuals (AR)

We first consider the case where image recovery yields a point-estimate $\widehat{x} = g_\theta(y)$ of the true x. As described in Section 4.1, a simple way to construct a nonconformity score is through the absolute residual (recall (4.4))

$$s(\widehat{x}, z; f) = |z - f(\widehat{x})|. \tag{4.6}$$

Evaluating this score on the calibration set d_{cal} gives $\{s_i\}_{i=1}^{n_{\mathsf{cal}}}$, whose empirical quantile $\widehat{\lambda}(d_{\mathsf{cal}})$ can be computed as in (4.2) and used to construct the prediction interval

$$\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\widehat{x}) = \big[f(\widehat{x}) - \widehat{\lambda}(d_{\mathsf{cal}}), f(\widehat{x}) + \widehat{\lambda}(d_{\mathsf{cal}})\big], \tag{4.7}$$

which then provides the marginal coverage property (4.5) [79].

Note that, with this choice of score, the interval width $|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\widehat{x})| = 2\widehat{\lambda}(d_{\mathsf{cal}})$ varies with the calibration set d_{cal} but not with $\widehat{x}$. Thus, for a fixed d_{cal}, the score (4.6) provides no way to tell whether one $\widehat{x}$ will yield more task-output uncertainty than a different $\widehat{x}$.

4.2.2 Method 2: Locally-Weighted Residuals (LWR)

We now consider the case where we have a posterior-sampling-based recovery method that yields c recoveries $\{\widehat{x}^{(j)}\}_{j=1}^{c}$ per measurement y. We make no assumption on how accurate or diverse these c samples are, other than assuming that the corresponding task outputs $\widehat{z}^{(j)} = f(\widehat{x}^{(j)})$ are not all identical.

Suppose that we choose the nonconformity score

$$s(\{\widehat{x}^{(j)}\}, z; f) = \frac{|z - \overline{z}|}{\sigma_z} \text{ with } \begin{cases} \overline{z} \triangleq \frac{1}{c}\sum_{j=1}^{c} f(\widehat{x}^{(j)}) \\ \sigma_z \triangleq \sqrt{\frac{1}{c}\sum_{j=1}^{c}\big(f(\widehat{x}^{(j)}) - \overline{z}\big)^2} \end{cases}, \tag{4.8}$$

evaluate it on the calibration set d_{cal} to get scores $\{s_i\}_{i=1}^{n_{\mathsf{cal}}}$, and compute their empirical quantile $\widehat{\lambda}(d_{\mathsf{cal}})$ as in (4.2). Then the prediction interval

$$\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\{\widehat{x}^{(j)}\}) = \big[\overline{z} - \sigma_z\widehat{\lambda}(d_{\mathsf{cal}}), \overline{z} + \sigma_z\widehat{\lambda}(d_{\mathsf{cal}})\big] \tag{4.9}$$

of this "locally weighted residual" (LWR) method provides the marginal coverage property in (4.5) [79].

In words, this method first computes (approximate) posterior samples $\widehat{z}^{(j)} \sim Z|Y = y$, which are then averaged to approximate the conditional mean $\widehat{z}_{\mathsf{mmse}} \triangleq \mathrm{E}(Z|Y =$

$y) \approx \overline{z}$ and square-root conditional covariance $\sqrt{\mathrm{cov}(Z|Y=y)} \approx \sigma_z$. When exactly computed, the conditional covariance gives a meaningful uncertainty metric on how well the true Z can be estimated from measurements y, because $\mathrm{cov}(Z|Y=y) = \mathrm{E}((Z-\widehat{z}_{\mathsf{mmse}})^2|Y=y)$. However, the σ_z that we compute is merely an approximation. So, with the aid of the calibration set, σ_z is adjusted by the scaling $\widehat{\lambda}(d_{\mathsf{cal}})$ to yield a prediction interval $\left[\overline{z} - \sigma_z\widehat{\lambda}(d_{\mathsf{cal}}), \overline{z} + \sigma_z\widehat{\lambda}(d_{\mathsf{cal}})\right]$ that satisfies the marginal coverage criterion (4.5).

Importantly, the interval width $|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\{\widehat{x}^{(j)}\})|$ now varies with $\{\widehat{x}^{(j)}\}_{j=1}^{c}$ through σ_z. This latter property is known as "local adaptivity" [79].

4.2.3 Method 3: Conformalized Quantile Regression (CQR)

Another popular locally adaptive method is known as conformalized quantile regression (CQR) [97]. The idea is to construct the nonconformity score using two quantile regressors [76], one which estimates the $\frac{\alpha}{2}$th quantile of $Z|Y=y$ and the other which estimates the $(1-\frac{\alpha}{2})$th quantile.

To compute these quantile estimates, we will once again assume access to a posterior-sampling-based recovery method that yields c recoveries $\{\widehat{x}^{(j)}\}_{j=1}^{c}$ per measurement y. From the corresponding task-outputs $\widehat{z}^{(j)} = f(\widehat{x}^{(j)})$, we compute the empirical quantiles $\widehat{q}(\frac{\alpha}{2})$ and $\widehat{q}(1-\frac{\alpha}{2})$ using

$$\widehat{q}(\omega) \triangleq \mathrm{EmpQuant}\left(\omega; \widehat{z}^{(1)}, \ldots, \widehat{z}^{(c)}\right). \tag{4.10}$$

From these quantile estimates, we construct the nonconformity score

$$s(\{\widehat{x}^{(j)}\}, z; f) = \max\left\{\widehat{q}(\tfrac{\alpha}{2}) - z, z - \widehat{q}(1-\tfrac{\alpha}{2})\right\}, \tag{4.11}$$

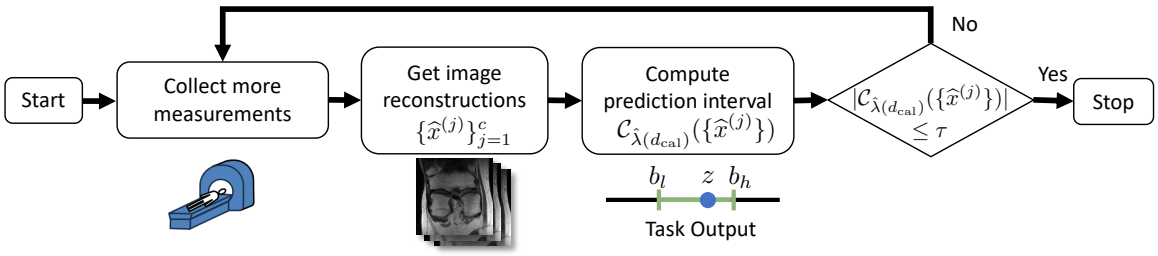


Figure 4.3: Proposed multi-round measurement protocol. In each round, measurements are collected and reconstructions and conformal intervals are computed. If the length of the interval falls below a user-set threshold τ, the procedure stops. Otherwise, more measurements are collected, and the process repeats until the threshold has been met.

evaluate it on the calibration set d_{cal} to obtain $\{s_i\}_{i=1}^{n_{\text{cal}}}$, and compute their $\lceil (1-\alpha)(n+1) \rceil / n$-empirical quantile $\widehat{\lambda}(d_{\text{cal}})$ as in (4.2). Then the prediction interval

$$\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\{\widehat{x}^{(j)}\}) = \left[\widehat{q}(\tfrac{\alpha}{2}) - \widehat{\lambda}(d_{\text{cal}}), \widehat{q}(1-\tfrac{\alpha}{2}) + \widehat{\lambda}(d_{\text{cal}})\right] \tag{4.12}$$

provides the marginal coverage in (4.5) [97]. Like (4.9), this interval is locally adaptive. We will compare these three conformal prediction methods in Section 4.3.

4.2.4 Multi-Round Measurement Protocol

In many applications, there is a significant cost to collecting a large number of measurements (i.e., acquiring a high-dimensional y). One example is MRI as discussed in Ch. 1. For these applications, we propose to collect measurements over multiple rounds, stopping as soon as the task uncertainty falls below a prescribed level τ. The goal is to collect the minimal number of measurements that accomplishes the task with probability of at least $1-\alpha$.

Our approach is to use the prediction interval width $|\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\{\widehat{x}^{(j)}\})|$ as the metric for task uncertainty. This requires the interval to be locally adaptive, as with LWR

and CQR above. The details are as follows. First, a sequence of $K > 1$ nested measurement configurations is chosen, so that the resulting measurement sets obey $\mathcal{Y}^{[1]} \subset \mathcal{Y}^{[2]} \subset \cdots \subset \mathcal{Y}^{[K]}$. Then, for each configuration $k = 1, \ldots, K$, a calibration set $d_{\mathsf{cal}}^{[k]}$ is collected, from which the set-valued function $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}^{[k]})}(\cdot)$ is constructed. At test time, we begin by collecting measurements $y \in \mathcal{Y}^{[1]}$ according to the first (i.e., minimal) configuration. From y we compute the reconstructions $\{\widehat{x}^{(j)}\}_{j=1}^{c}$ and, from them, the task uncertainty $|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}^{[1]})}(\{\widehat{x}^{(j)}\})|$. If this uncertainty falls below the desired τ, we stop collecting measurements. If not, we would collect the additional measurements in $\mathcal{Y}^{[2]} \setminus \mathcal{Y}^{[1]}$, and repeat the procedure. Figure 4.3 summarizes the proposed multi-round protocol.

4.3 Numerical Experiments

We now demonstrate our task-based uncertainty quantification framework on MRI [75]. As before, we are interested in accelerated MRI, which speeds up the acquisition process by collecting a fraction $1/R$ of the measurements specified by the Nyquist sampling theorem. This comes at the expense of making the inverse problem ill-posed when $R > 1$.

In MRI, a typical task is to diagnose the presence or absence of a pathology. Although this task is typically performed by a radiologist, neural-network-based classification is expected to play a significant role in aiding radiologists [23]. Thus, in our experiments, we implement the task $f(\cdot)$ using a neural network. Details are given below.

Data: We use the multi-coil fastMRI knee dataset [124] and in particular the non-fat-suppressed subset, which includes 484 training volumes (17286 training slices,

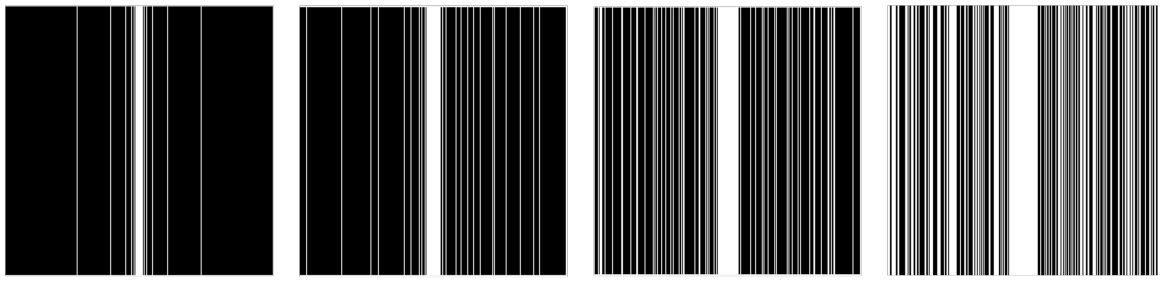


Figure 4.4: Sampling masks in the 2D spatial Fourier domain for each acceleration rate R. Each white line indicates the subset of collected samples. The masks are nested, in that the mask at a given R contains all samples in all masks at higher R.

or images) and 100 validation volumes. For the validation set, we use the central 80% of slices in a volume, which provides 2188 validation images. We use pathology labels from fastMRI+ [128]. For knee-MRI, meniscus tears yield the largest fastMRI+ label set, and so we choose meniscus-tear-detection as our task. In total, there are 1921 training images and 324 validation images that contain a meniscus tear. To collect measurements, we retrospectively subsample the fastMRI data in the k-space using a set of random nested masks (shown in Figure 4.4) that yield acceleration rates $R \in \{16, 8, 4, 2\}$. The details for the generation of these masks are described in App. A.1.

Image Recovery: We consider two recovery networks $g_\theta(\cdot)$. As a point estimator, we use the state-of-the-art E2E-VarNet from Sriram et al. [107] and, as a posterior sampler, we use the conditional normalizing flow (CNF) from Ch. 3. Both were specifically designed around the fastMRI dataset. Another option would be the MRI diffusion sampler [34], but its performance is a bit worse than the CNF and its sampling

Table 4.1: Number of images in each data fold.

Training	Validation		
	Calibration	Testing	Total
17286	1531	656	2188

speed is 8000× slower. The E2E-VarNet and CNF were each trained to handle all four acceleration rates with a single model.

Task Network: We used a ResNet50 [56] for the task network $f(\cdot)$. Starting from an ImageNet-based initialization, we pretrained the weights to minimize the unsupervised SimCLR loss [31] and later minimized binary-cross-entropy loss using the fastMRI+ labels. See App. A.2 for details on all of the networks.

Empirical Validation: Recall that the marginal coverage guarantee (4.5) holds on average over random test samples $(\widehat{X}_0, Z_0)$ and random calibration data $D_{\mathsf{cal}} = \{(\widehat{X}_1, Z_1), \ldots, (\widehat{X}_n, Z_n)\}$. To empirically validate marginal coverage and evaluate other average-performance metrics, we perform Monte-Carlo averaging over $T = 10000$ trials as follows. In each trial t, we randomly partition the 2188-sample validation dataset into a 70% calibration fold with indices $i \in \mathcal{I}_{\mathsf{cal}}[t]$ and a 30% test fold with indices $i \in \mathcal{I}_{\mathsf{test}}[t]$, construct conformal predictors using the calibration data $d_{\mathsf{cal}}[t] = \{(\{\widehat{x}_i^{(j)}\}_{j=1}^{c}, z_i)\}_{i \in \mathcal{I}_{\mathsf{cal}}[t]}$, and evaluate performance on the test fold of trial t. Finally, we average performance over the T trials. The dataset splits are summarized in Table 4.1. Further details are given below.

4.3.1 Effect of Acceleration Rate and Conformal Prediction Scheme

We have seen that the interval length $|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\{\widehat{x}^{(j)}\})|$ provides a way to quantify the uncertainty that the measurement-and-reconstruction scheme contributes to the meniscus-classification task. So a natural question is: How is the interval length affected by the MRI acceleration R? We study this question below.

For a fixed acceleration R, the interval length is also affected by the choice of conformal predictor. All else being equal, better conformal predictors yield smaller uncertainty sets [5]. So another question is: How is the interval length affected by selecting among the AR, LWR, or CQR conformal methods?

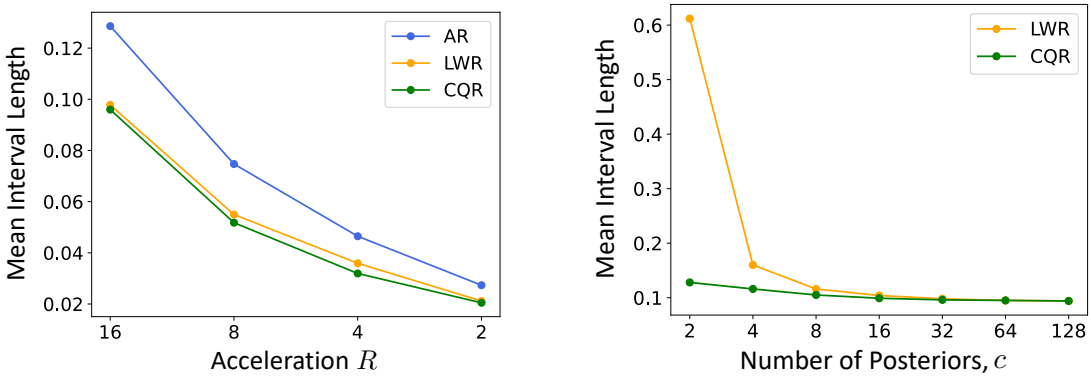
To answer these questions, we compute the "average mean interval length" $\overline{\mathrm{MIL}} \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathrm{MIL}[t]$ using the trial-t mean interval length

$$\mathrm{MIL}[t] \triangleq \frac{1}{|\mathcal{I}_{\mathsf{test}}[t]|}\sum_{i\in\mathcal{I}_{\mathsf{test}}[t]}\left|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}[t])}\left(\{\widehat{x}_i^{(j)}\}_{j=1}^{c}\right)\right|. \tag{4.13}$$

Figure 4.5a plots the average mean interval length versus R for the AR, LWR, and CQR conformal predictors using $T = 10000$ trials, $c = 32$ posterior samples, and error-rate $\alpha = 0.05$. The figure shows that, as expected, the average mean interval length decreases as more measurements are collected (i.e., as R decreases). The figure also shows that, as expected, the locally adaptive LWR and CQR methods give consistently smaller average mean interval lengths than the non-adaptive AR method. In this sense, posterior sampling is advantageous over point sampling.

4.3.2 Effect of Number of Posterior Samples

Above, we saw that the measurement process and conformal method both affect the prediction-interval length. We conjecture that the image reconstruction process



(a) Mean Interval Length vs. R

(b) Mean Interval Length vs c

Figure 4.5: a) Average mean interval length versus acceleration R with $c = 32$ samples. b) Mean interval length versus c with acceleration $R = 16$. All results use error-rate $\alpha = 0.05$ and $T = 10000$ trials.

will also affect the prediction-interval length. To investigate this, we vary the number of samples c produced by the posterior-sampling scheme, reasoning that smaller values of c correspond to less accurate recoveries (e.g., a less accurate posterior mean approximation).

Figure 4.5b plots the average mean interval length versus c for the LWR and CQR conformal predictors using $T = 10000$ trials, acceleration $R = 16$, and error-rate $\alpha = 0.05$. As expected, the interval length decreases as the posterior sample size c grows. But interestingly, LWR is much more sensitive to small values of c than CQR. One implication is that small values of c may suffice when used with an appropriate conformal prediction method.

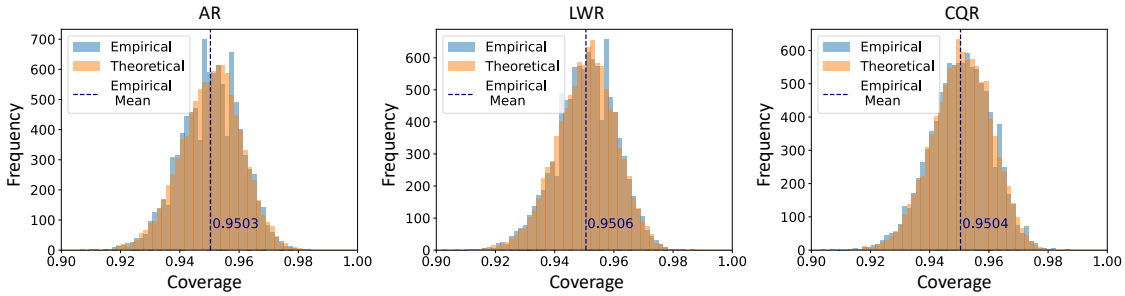


Figure 4.6: For the AR, LWR, and CQR conformal methods, each subplot shows the histograms of the empirical and theoretical empirical-coverage samples $\{\mathrm{EC}[t]\}_{t=1}^{T}$ across $T = 10000$ Monte-Carlo trials using $\alpha = 0.05$, $R = 8$, and $c = 32$. The subplots are also labelled with the empirical mean of $\{\mathrm{EC}[t]\}_{t=1}^{T}$, which is very close to the target value of $1 - \alpha = 0.95$.

4.3.3 Empirical Validation of Coverage

To verify that the marginal coverage guarantee (4.5) holds, we compute the empirical coverage of Monte-Carlo trial t as

$$\mathrm{EC}[t] \triangleq \frac{1}{|\mathcal{I}_{\mathsf{test}}[t]|} \sum_{i \in \mathcal{I}_{\mathsf{test}}[t]} \mathbb{1}\{z_i \in \mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}[t])}(\{\widehat{x}_i^{(j)}\})\}, \tag{4.14}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Existing theory (see, e.g., [5]) says that when $(\{\widehat{X}_i^{(j)}\}, Z_i)$ and $D_{\mathsf{cal}}[t]$ in (4.14) are exchangeable pairs of random variables, $\mathrm{EC}[t]$ is random and distributed as

$$\mathrm{EC}[t] \sim \mathrm{BetaBin}(n_{\mathsf{test}}, n_{\mathsf{cal}} + 1 - l_{\mathsf{cal}}, l_{\mathsf{cal}}) \quad \text{for} \quad l_{\mathsf{cal}} \triangleq \lceil (n_{\mathsf{cal}} + 1)\alpha \rceil, \tag{4.15}$$

where $n_{\mathsf{test}} \triangleq |\mathcal{I}_{\mathsf{test}}[t]|$ and $n_{\mathsf{cal}} \triangleq |\mathcal{I}_{\mathsf{cal}}[t]|$.

For each of the three conformal methods, Figure 4.6 shows the histogram of $\{\mathrm{EC}[t]\}_{t=1}^{T}$ from (4.14) for $T = 10000$, error-rate $\alpha = 0.05$, acceleration $R = 8$, and $c = 32$ posterior samples. The figure shows that this histogram is close to the histogram created from T samples of the theoretical distribution in (4.15). Figure 4.6 also prints

the average empirical coverage $\frac{1}{T}\sum_{t=1}^{T}\mathrm{EC}[t]$ for each method, which is very close to the target value of $1-\alpha=0.95$. Thus, we see that, in practice, conformal prediction behaves close to the theory.

4.3.4 Multi-Round Measurements

We now investigate the application of the multi-round measurement protocol from Section 4.2.4 to accelerated MRI. For this, we simulated the collection of MRI slices over rounds $k=1,\ldots,5$, stopping as soon as the $\alpha=0.01$ interval width $|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}^{[k]})}(\{\widehat{x}^{(j)}\})|$ falls below the threshold of $\tau=0.1$. The first round collects k-space measurements at acceleration rate $R^{[1]}=16$, and the remaining rounds each collect additional k-space measurements to yield $R^{[2]}=8$, $R^{[3]}=4$, $R^{[4]}=2$, and $R^{[5]}=1$ respectively. For quantitative evaluation, we randomly selected 8 multi-slice volumes from the 100-volume fastMRI validation set to act as test volumes (half of which were labeled as meniscus tears and half of which were not), and we used the remaining 92 volumes for calibration. We will refer to the corresponding index sets as $\mathcal{I}_{\mathsf{test}}$ and $\mathcal{I}_{\mathsf{cal}}$.

We begin by discussing the AR conformal prediction method, which uses the point-sampling E2E-VarNet [107] for image recovery. The AR method produces prediction intervals $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}^{[k]})}(\widehat{x})$ that are $\widehat{x}$-invariant (i.e., not locally adaptive). Thus, immediately after calibration, it is known that $k=4$ measurement rounds (i.e., $R=2$) are necessary and sufficient to achieve the $\tau=0.1$ threshold at error-rate $\alpha=0.01$.

The LWR and CQR conformal prediction methods both use the CNF from Ch. 3 with $c=32$ posterior samples and yield locally adaptive prediction intervals $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}^{[k]})}(\{\widehat{x}^{(j)}\})$. This allows them to evaluate the interval length for each $\{\widehat{x}^{(j)}\}$ and stop the measurement process as soon as that length falls below the threshold τ. For

test image $i \in \mathcal{I}_{\mathsf{test}}$, we denote the final measurement round as

$$k_i \triangleq \min\left\{k : |\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}^{[k]})}(\{\widehat{x}_i^{(j)}\})| < \tau\right\}. \tag{4.16}$$

(Note that $\{\widehat{x}_i^{(j)}\}$ also changes with the measurement round k, although the notation does not explicitly show this.) The average acceleration is then

$$\overline{R} = \left(\frac{1}{|\mathcal{I}_{\mathsf{test}}|}\sum_{i\in\mathcal{I}_{\mathsf{test}}}\frac{1}{R^{[k_i]}}\right)^{-1}. \tag{4.17}$$

Table 4.2 shows the average acceleration $\overline{R}$ for the AR, LWR, and CQR conformal methods. We see that $\overline{R} = 2$ for the AR method because it always uses four measurement rounds. The LWR and CQR methods achieve higher average accelerations $\overline{R}$ because fewer measurement rounds suffice in a large fraction of cases. Table 4.2 also shows that the empirical coverage is close to what we would expect given this relatively small test set.

Figure 4.7 plots the distribution of final-round $\{k_i\}_{i\in\mathcal{I}_{\mathsf{test}}}$ for the AR, LWR, and CQR conformal methods. It too shows that the AR method always uses four rounds (i.e., $R = 2$), while the LWR and CQR methods typically use fewer rounds. However, this plot also shows that the LWR method sometimes uses five measurement rounds. This may seem counter-intuitive but can be explained as follows. At $k = 4$, the AR method is calibrated so that the true score z lands in the prediction interval in all but $\alpha = 1\%$ of the cases, where the length of that interval is small enough to meet the $\tau = 0.1$ threshold. Meanwhile, the LWR (and CQR) methods adapt the prediction interval based on the difficulty of $\{\widehat{x}^{(j)}\}$. In most cases, the LWR prediction interval is smaller than the AR interval, but for a few "difficult" cases the LWR prediction interval is wider, and in fact too wide to meet the $\tau = 0.1$ threshold. For these difficult cases, the LWR method moves on to the fifth measurement round.

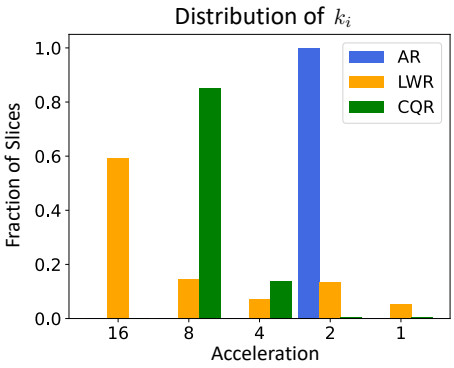


Figure 4.7: Fraction of slices accepted after a given acceleration rate.

Table 4.2: Average metrics for the multi-round MRI simulation (± standard error).

Method	Average Acceleration	Empirical Coverage	Average Max Center Error
AR	2.000	0.991 ± 0.008	0.032 ± 0.017
LWR	5.157	0.992 ± 0.005	0.020 ± 0.002
CQR	6.762	0.987 ± 0.008	0.044 ± 0.009

Based on the previous discussion, one might conjecture that the prediction intervals accepted by the AR method at round $k = 4$ will be somehow worse than those accepted by LWR at $k = 4$, even though their lengths all meet the threshold. We can confirm this by interpreting the midpoint of the prediction interval as an estimate of z and evaluating the absolute error on that estimate, which we call the "center error" (CE):

$$\mathrm{CE}(\{\widehat{x}^{(j)}\}, z) \triangleq \left| z - \frac{b_l + b_u}{2} \right| \quad \text{where} \quad [b_l, b_u] = \mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\{\widehat{x}^{(j)}\}). \tag{4.18}$$

When evaluating the center error, we take the maximum over the slices in each volume. Table 4.2 lists the average maximum center error and confirms that it is smaller for LWR than for AR.

Figure 4.8 shows examples of image reconstructions, pixel-wise standard deviation maps, and CQR prediction intervals for a test image labeled with a meniscus tear. At higher accelerations like $R = 16$, relatively large variations across posterior samples $\{\widehat{x}^{(j)}\}$ result in relatively large variations across classifier outputs $\{\widehat{z}^{(j)}\}$, which result in a large prediction interval, i.e., high uncertainty about the ground-truth classifier

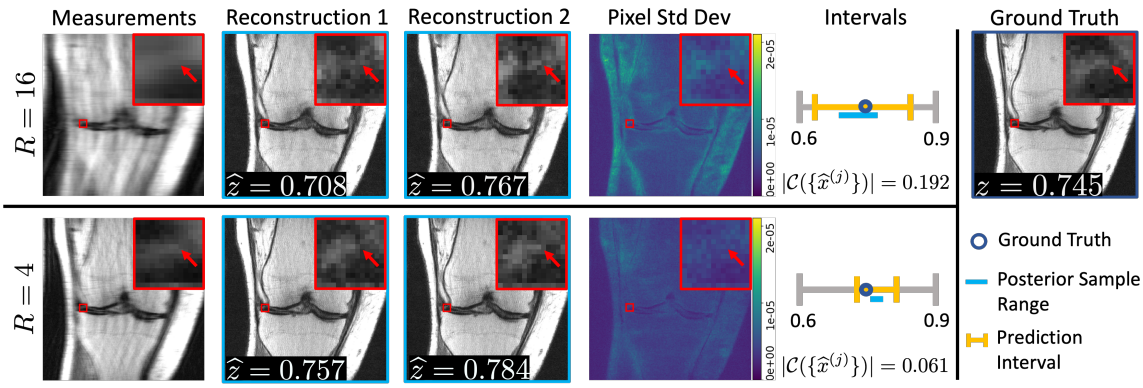


Figure 4.8: MR image reconstructions and CQR prediction intervals at accelerations $R = 16$ and $R = 4$ with error-rate $\alpha = 0.01$ and a total of $c = 32$ posterior samples. The fastMRI+ bounding box around the meniscus tear is magnified in red. The prediction intervals shrink as the posterior samples become more consistent in the meniscus region. The standard-deviation maps show areas of high pixel-wise uncertainty but are difficult to connect to the downstream task. Note, image brightness was increased to better highlight the tear. Best viewed when zoomed.

output z. At lower accelerations like $R = 4$, relatively small variations across posterior samples yield smaller prediction intervals, i.e., less uncertainty about z. While the pixel-wise standard-deviation maps also show reduced variation across posterior samples, it's difficult to draw conclusions about uncertainty in the downstream task from them. For example, the same pixel-wise variations could result from a set of reconstructions that show clear evidence for a tear in some cases and clear evidence to the contrary in others, or from a set of reconstructions that show clear evidence for a tear in all cases but are corrupted by different noise realizations. Our uncertainty quantification methodology circumvents these issues by focusing on the task itself. Furthermore, by leveraging the framework of conformal prediction, it ensures that the uncertainty estimates are statistically meaningful.

As far as practical implementation is concerned, for each slice in a volume, the CNF reconstructions, ResNet-50 classifier outputs, and conformal prediction intervals can be computed in 414 milliseconds for $c = 32$ samples, or 7.4 milliseconds for $c = 2$ samples, on a single NVIDIA A100 GPU.

4.4 Discussion

A number of works on uncertainty quantification for MRI have been proposed based on Bayesian neural networks and posterior sampling, e.g., [102, 43, 37, 62, 44, 90, 34, 20]. They produce a set of possible reconstructions $\{\widehat{x}^{(j)}\}$, from which a pixel-wise uncertainty map is typically computed. Conformal prediction methods have also been proposed to generate pixel-wise uncertainty maps for MRI and other imaging inverse problems [7, 77, 111, 60], but with statistical guarantees. However, when imaging is performed with the eventual goal of performing a downstream task, pixel-wise uncertainty maps are of questionable value. In this work, we construct a conformal prediction interval that is statistically guaranteed to contain the task output from the true image. We focus on tasks that output a real-valued scalar, such as soft-output binary classification.

Other works have applied conformal prediction to MRI tasks. Lu et al. [83] consider a dataset $\{(x_i, z_i)\}$ with MRI images x_i and discrete ordinal labels $z_i \in \{1, \dots, K\}$ that rate the severity of a pathology. They design a predictor that, given test x, outputs a set $\mathcal{Z}(x) \in 2^K$ that is guaranteed to contain the true label z with probability $1 - \alpha$. Different from our work, Lu et al. [83] involves no inverse problem and aims to quantify the uncertainty in a discrete z. Sankaranarayanan et al. [101] compute uncertainty intervals on the presence/absence of semantic attributes in images, and

mention that one application could be pathology detection in MRI (although they do not pursue it). Although their high-level goal is similar to ours, their solution requires a trained "disentangled" generative network that, in the case of MRI, would generate MRI images from pathology probabilities. To our knowledge, no such networks exist for MRI. In contrast, our method requires only a trained pathology classifier $f(\cdot)$, which should be readily available.

Limitations: First, our method requires a downstream task, which is not always available. Second, we demonstrated our method on only a single inverse problem and task; validation on other applications is needed. Third, our MRI application ideas are preliminary and not ready for clinical use. Since we use the conformal prediction interval width as a proxy for the diagnostic value of the reconstructed image(s), several aspects of our design (e.g., the choice of classifier $f(\cdot)$, recovery algorithm $g_\theta(\cdot)$, conformal prediction method, threshold τ, and error-rate α) would need to be tuned and validated through rigorous clinical studies. Fourth, for ease of exposition, the conformal methods that we use (AR, LWR, CQR) are somewhat simple. More advanced methods, like risk-controlling prediction sets (RCPS) [16], may perform better. Fifth, we considered only tasks that output a single real-valued scalar, such as soft-output binary classification. Extensions to more general tasks would be useful. Lastly, our posterior sampler only considers aleatoric uncertainty. In principle, epistemic uncertainty could be included by sampling the generator's weights from a distribution, as in Ekmekci & Cetin [45], but more work is needed in this direction.

4.5 Conclusion

For imaging inverse problems, we proposed a method to quantify how much uncertainty the measurement-and-reconstruction process contributes to a downstream task, such as soft-output classification. In particular, we use conformal prediction to construct an interval that is guaranteed to contain the task-output from the true image with high probability. We showed that, with posterior-sampling-based image recovery methods, the prediction intervals can be made adaptive, and we proposed a multi-round measurement protocol that stops acquiring new data when the task uncertainty is sufficiently small. We applied our method to meniscus-tear detection in accelerated knee MRI and demonstrated significant gains in acceleration rate.

Chapter 5: Conformal Bounds on Full-Reference Image Quality for Inverse Problems

While the methodology in Ch. 4 presents one option to provide UQ beyond the pixel level, it requires a downstream task of interest, which outputs a real-valued scalar. In some contexts, a suitable downstream task may be difficult to identify. For example, a denoising algorithm on a consumer camera may just need to provide the best "quality" image since the resulting output is not used to perform a particular task. This notion of quantifying "quality", or better yet "accuracy" in the setting of medical image recovery, is a much more general objective that is relevant in almost all inverse problems. Thus, we now consider how one may quantify the uncertainty on the "accuracy" of a reconstruction.

In image recovery, "accuracy" can be defined in different ways. Classical metrics like mean-squared error (MSE), or its scaled counterpart peak signal-to-noise ratio (PSNR), are convenient for theoretical analysis but do not always correlate well with human perceptions of image quality. This fact inspired the field of full-reference image-quality (FRIQ) assessment [81, 119], which led to the well-known Structural Similarity Index Measure (SSIM) [120] that is still popular today. However, progress continues to be made. Most recent methods leverage the internal features of deep neural networks, which are said to mimic the processing architecture of the human

visual cortex [123, 82]. A popular example of the latter is Learned Perceptual Image Patch Similarity (LPIPS) [126]. In the end though, the best choice of metric may depend on the application. For example, in magnetic resonance imaging (MRI), the goal is to provide the radiologist with an image recovery that leads to an accurate diagnosis. A recent clinical MRI study [68] found that, among 35 tested metrics, Deep Image Structure and Texture Similarity (DISTS) [39] correlated best with radiologists' perceptions.

In this chapter, our goal is to provide rigorous bounds on the FRIQ $m(\widehat{x}_0, x_0)$ of a recovery $\widehat{x}_0 = g_\theta(y_0)$ relative to the true image x_0. Here, $g_\theta(\cdot)$ is an arbitrary image-recovery scheme and $m(\cdot, \cdot)$ is an arbitrary FRIQ metric. The key challenge is that x_0 is unknown. To our knowledge, there exists no prior work on providing FRIQ guarantees in image recovery. Our contributions are as follows.

1. We propose a framework to bound the FRIQ $m(\widehat{x}_0, x_0)$ of a recovered image $\widehat{x}_0$ without access to the true image x_0. Our framework uses conformal prediction [117, 5] to construct bounds that hold with probability at least $1 - \alpha$ under certain exchangeability assumptions and where $\alpha \in (0, 1)$ is chosen by the user.

2. We show how posterior-sampling-based image recovery can be used to construct conformal bounds that adapt to the measurements y_0 and reconstruction $\widehat{x}_0$.

3. We demonstrate our approach on two linear inverse problems: denoising of FFHQ faces [66] faces and recovery of fastMRI knee images [125] from accelerated multi-coil measurements.

5.1 Background

We now describe split CP [91, 79] from a slightly different but equivalent perspective. This new description aligns more clearly with the formulation of our proceeding methodology. Given features $w_0 \in \mathcal{W}$, the goal of CP is to construct a set $\mathcal{C}_\lambda(\widehat{z}_0)$ that contains an unknown target $z_0 \in \mathcal{Z}$ with high probability. Here, $\mathcal{C}_\lambda(\cdot)$ is constructed so that $|\mathcal{C}_\lambda(\widehat{z}_0)|$ is monotonically non-decreasing in $\lambda \in \mathbb{R}$ for any fixed $\widehat{z}_0$, and $\widehat{z}_0 = f(w_0)$ is some prediction from a black-box model $f(\cdot)$. Split CP accomplishes this goal by calibrating λ using a dataset of feature and target pairs $\{(w_i, z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ that has not been used to train $f(\cdot)$. In particular, it first constructs the set $d_{\mathsf{cal}} \triangleq \{(\widehat{z}_i, z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ using $\widehat{z}_i = f(w_i)$ and then finds a $\widehat{\lambda}(d_{\mathsf{cal}})$ to provide the marginal coverage guarantee [80]

$$\Pr\left\{Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{Z}_0)\right\} \geq 1 - \alpha, \tag{5.1}$$

where α is a user-chosen error rate. Here and in the sequel, we use capital letters to denote random variables and lower-case letters to denote their realizations. In words, (5.1) guarantees that the unknown target Z_0 falls within the interval $\mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{Z}_0)$ with probability at least $1 - \alpha$ when averaged over the randomness in the test data $(Z_0, \widehat{Z}_0)$ and calibration data D_{cal}.

While there are a number of ways to describe CP calibration of λ [117, 5], we utilize the method from Angelopoulos et al. [6], which applies to general risk control. It starts by defining the risk as the empirical miscoverage

$$\widehat{r}_n(\lambda; d_{\mathsf{cal}}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{z_i \notin \mathcal{C}_\lambda(\widehat{z}_i)\}, \tag{5.2}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. The empirical miscoverage measures the proportion of targets z_i that land outside of $\mathcal{C}_\lambda(\widehat{z}_i)$ in the calibration set d_{cal}. Note the

dependence on λ, which controls the size of the prediction interval. The calibration procedure then sets λ at

$$\widehat{\lambda}(d_{\mathsf{cal}}) = \inf\left\{\lambda : \widehat{r}_n(\lambda; d_{\mathsf{cal}}) \leq \alpha - \tfrac{1-\alpha}{n}\right\}, \tag{5.3}$$

which can be found using a simple binary search. Intuitively, the λ chosen in (5.3) yields an empirical miscoverage that is slightly more conservative than the desired α in order to handle the finite size of the calibration set. When $\{(Z_0, \widehat{Z}_0), (Z_1, \widehat{Z}_1), \ldots, (Z_n, \widehat{Z}_n)\}$ are exchangeable (a weaker condition than i.i.d.), (5.3) ensures that (5.1) holds [6].

5.2 Proposed Method

Consider an imaging inverse problem, where we observe incomplete and/or noisy measurements $y_0 = \mathcal{A}(x_0)$ of a true image x_0. Suppose that $\widehat{x}_0 = g_\theta(y_0)$ is a reconstruction of x_0 provided by some image recovery method $g_\theta(\cdot)$ and that $z_0 = m(\widehat{x}_0, x_0) \in \mathbb{R}$ is some FRIQ metric on $\widehat{x}_0$ with respect to the true x_0. We would like to know z_0, especially in safety critical applications. For example, if z_0 was unacceptable, then perhaps we could use a different recovery method $g_\theta(\cdot)$ or collect more measurements y_0. But z_0 cannot be directly computed because x_0 is unknown.

Our key insight is that it's possible to construct a set $\mathcal{C}_\lambda(\widehat{z}_0)$ that is guaranteed to contain the unknown FRIQ z_0 with high probability. This can be done using CP, at least when one has access to calibration data $\{(x_i, y_i)\}_{i=1}^{n_{\mathsf{cal}}}$ of true image and measurement pairs that agrees with the test (x_0, y_0) in the sense that the resulting FRIQ pairs $\{(\widehat{z}_i, z_i)\}_{i=0}^{n_{\mathsf{cal}}}$ are statistically exchangeable.

Our general approach is as follows. Using $\{(x_i, y_i)\}_{i=1}^{n_{\mathsf{cal}}}$, we compute the image recovery $\widehat{x}_i = g_\theta(y_i)$ and the corresponding true FRIQ $z_i = m(\widehat{x}_i, x_i)$ for each $i = 1, \ldots, n$. Then we construct an estimator $f(\cdot)$ that produces an FRIQ estimate

$\widehat{z}_i = f(w_i)$ for some choice of w_i. Several choices of $f(\cdot)$ and w_i will be described in the sequel. We then collect the results into the set $d_{\mathsf{cal}} = \{(\widehat{z}_i, z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ and calibrate the λ parameter of the FRIQ prediction interval $\mathcal{C}_\lambda(\widehat{z}_i)$ using CP.

We now describe our choice of prediction interval $\mathcal{C}_\lambda(\cdot)$. In the sequel, we will refer to those metrics $m(\cdot,\cdot)$ for which a higher value indicates better image quality (e.g., PSNR, SSIM) as Higher-Preferred (HP) metrics, and those for which a lower value indicates better image quality (e.g., LPIPS, DISTS) as Lower-Preferred (LP) metrics. We choose to construct the prediction set for the i-th sample as

$$\mathcal{C}_\lambda(\widehat{z}_i) = [\beta(\widehat{z}_i, \lambda), \infty) \;\text{ for HP metrics and }\; \mathcal{C}_\lambda(\widehat{z}_i) = (-\infty, \beta(\widehat{z}_i, \lambda)] \;\text{ for LP metrics,} \tag{5.4}$$

where we choose the lower/upper bound $\beta(\cdot,\cdot)$ as

$$\beta(\widehat{z}_i, \lambda) = \widehat{z}_i - \lambda \;\text{ for HP metrics and }\; \beta(\widehat{z}_i, \lambda) = \widehat{z}_i + \lambda \;\text{ for LP metrics.} \tag{5.5}$$

By calibrating the bound parameter λ as $\widehat{\lambda}(d_{\mathsf{cal}})$ using (5.3), we obtain the following marginal coverage guarantee for the test sample $(\widehat{Z}_0, Z_0)$:

$$\Pr\left\{Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{Z}_0)\right\} \geq 1 - \alpha, \tag{5.6}$$

which holds when $\{(Z_0, \widehat{Z}_0), (Z_1, \widehat{Z}_1), \ldots, (Z_n, \widehat{Z}_n)\}$ are exchangeable [6]. In particular, $\beta(\widehat{Z}_0, \widehat{\lambda}(D_{\mathsf{cal}}))$ lower-bounds the unknown true HP metric value Z_0, or upper-bounds the unknown true LP metric value Z_0, with probability at least $1 - \alpha$, where α is selected by the user. A smaller error-rate α will tend to yield a looser bound, but—importantly—the coverage guarantee (5.6) will hold for any chosen $\alpha \in (0, 1)$. In the sequel, we will refer to $\beta(\widehat{z}_0, \widehat{\lambda}(d_{\mathsf{cal}}))$ as the "conformal bound" on z_0. Note that the conformal bound can "adapt" to the test measurements y_0 and reconstruction $\widehat{x}_0$ through $\widehat{z}_0 = f(w_0)$ for appropriate choices of $f(\cdot)$ and w_0.

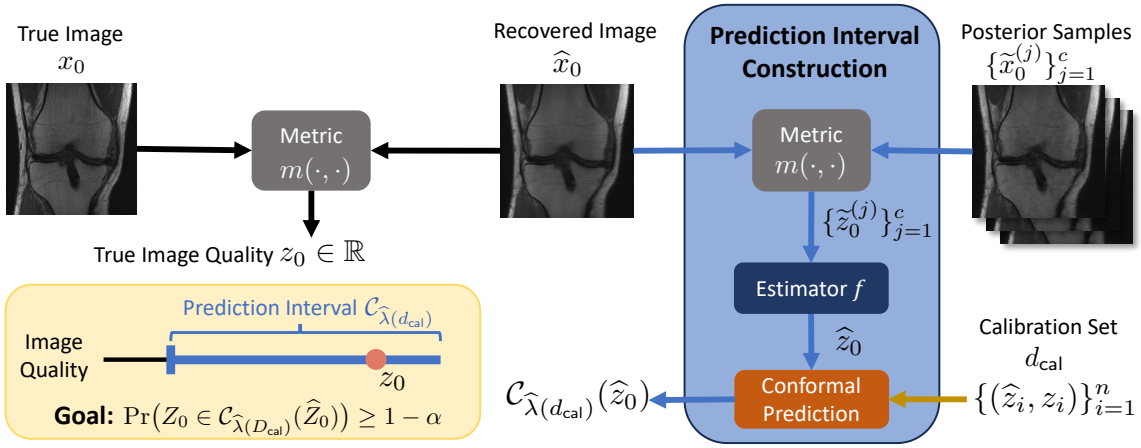


Figure 5.1: Overview of method: Given a recovery $\widehat{x}_0$ of true image x_0, approximate posterior samples $\{\widetilde{x}_0^{(j)}\}_{j=1}^c$, and a calibration set d_{cal}, we construct a prediction interval $\mathcal{C}_{\widehat{\lambda}(d_{\text{cal}})}(\widehat{z}_0)$ that is guaranteed to contain the unknown true FRIQ $z_0 = m(\widehat{x}_0, x_0)$ with probability at least $1-\alpha$.

Below we describe different ways to construct $f(\cdot)$ and w_0, which in turn yield conformal bounds with different properties. Appendix D.4 investigates violations of the exchangeability assumption.

5.2.1 A Non-adaptive Bound on Recovered-image FRIQ

As a simple baseline, we start with the choice $f(\cdot) = 0$. In this case, w_0 is inconsequential and $\widehat{z}_0 = 0$, and so the conformal bound $\beta(\widehat{z}_0, \widehat{\lambda}(d_{\text{cal}}))$ will depend on the calibration set d_{cal} but not the test measurements y_0 or reconstruction $\widehat{x}_0$. We refer to such bounds as "non-adaptive." As we demonstrate in Sec. 5.3, non-adaptivity leads to conservative bounds. Still, this non-adaptive bound is valid in the sense of guaranteed marginal coverage (5.6) under the exchangeability assumption.

5.2.2 Intuitions on Constructing Adaptive FRIQ Bounds

Our approach to constructing adaptive FRIQ bounds is based on the following probabilistic viewpoint. Conditioned on the observed measurements y_0, we can model the unknown FRIQ as $Z_0 = m(\widehat{x}_0, X_0)$ for $\widehat{x}_0 = g_\theta(y_0)$ and $X_0 \sim p_{X_0|Y_0}(\cdot|y_0)$. The distribution $p_{X_0|Y_0}(\cdot|y_0)$ is often referred to as the posterior distribution on X_0 given the measurements $Y_0 = y_0$.

Let us first consider the ideal and unrealistic case that the y_0-conditional FRIQ distribution $p_{Z_0|Y_0}(\cdot|y_0)$ is known. And let's consider the case of HP metrics, noting that similar arguments can be made for LP metrics. If $p_{Z_0|Y_0}(\cdot|y_0)$ was known, then constructing a lower-bound β on Z_0 that holds with probability $1-\alpha$ could be directly accomplished by finding the $\beta \in \mathbb{R}$ that satisfies $\Pr\{Z_0 \geq \beta | Y_0 = y_0\} = 1 - \alpha$, which is known as the αth quantile of $Z_0|Y_0 = y_0$.

Now suppose that the distribution of $Z_0|Y_0 = y_0$ was unknown, but instead one had access to an infinite number of perfect posterior image samples $\{\widetilde{x}_0^{(j)}\}_{j=1}^{\infty}$. By "perfect" we mean that $\widetilde{x}_0^{(j)}$ are independent realizations of $X_0|Y_0 = y_0$. From them, one could construct posterior FRIQs $\{\widetilde{z}_0^{(j)}\}_{j=1}^{c}$ using $\widetilde{z}_0^{(j)} \triangleq m(\widehat{x}_0, \widetilde{x}_0^{(j)})$. Importantly, $\{z_0, \widetilde{z}_0^{(1)}, \widetilde{z}_0^{(2)}, \widetilde{z}_0^{(3)}, \dots\}$ are i.i.d. realizations of $Z_0|Y_0 = y_0$. Thus, to construct a lower bound β on $Z_0|Y_0 = y_0$ that holds with probability $1 - \alpha$, one could use the empirical quantile of $\{\widetilde{z}_0^{(j)}\}$, i.e.,

$$\beta = \lim_{c \to \infty} \text{EmpQuant}\left(\alpha, \{\widetilde{z}_0^{(j)}\}_{j=1}^{c}\right), \tag{5.7}$$

which converges to the αth quantile of $Z_0|Y_0 = y_0$ [50].

In practice, one will not have access to an infinite number of perfect posterior image samples. However, it is not difficult to obtain a finite number of approximate posterior

samples $\{\widetilde{x}_0^{(j)}\}_{j=1}^c$. From them, one could estimate the αth quantile of $Z_0|Y_0=y_0$ and subsequently calibrate that (imperfect) estimate using conformal prediction. Two such strategies are described below.

5.2.3 An Adaptive Bound on Recovered-image FRIQ

Suppose that, for each $i \in \{0, 1, \ldots, n\}$, we have access to $c \geq 1$ approximate posterior image samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^c$ produced by a black-box posterior image sampler such as those listed in Ch. 3. Guided by the intuitions from Sec. 5.2.2, we propose the following for HP metrics. For each i, we first compute the corresponding approximate posterior FRIQs $\{\widetilde{z}_i^{(j)}\}_{j=1}^c$ using $\widetilde{z}_i^{(j)} = m(\widehat{x}_i, \widetilde{x}_i^{(j)})$ and then set $\widehat{z}_i$ at their empirical quantile

$$\widehat{z}_i = \text{EmpQuant}\left(\alpha, \{\widetilde{z}_i^{(j)}\}_{j=1}^c\right) = f(w_i) \quad \text{for} \quad \begin{cases} f(\cdot) = \text{EmpQuant}(\alpha, \cdot) \\ w_i = [\widetilde{z}_i^{(1)}, \ldots, \widetilde{z}_i^{(c)}]^\top \in \mathbb{R}^c. \end{cases} \tag{5.8}$$

We then use $d_{\mathsf{cal}} = \{(\widehat{z}_i, z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ to calibrate the bound parameter λ using (5.3), yielding $\widehat{\lambda}(d_{\mathsf{cal}})$. Finally, we plug this λ and $\widehat{z}_0$ into (5.5) to get $\beta(\widehat{z}_0, \widehat{\lambda}(d_{\mathsf{cal}}))$, which is our conformal bound on the true FRIQ z_0. From Sec. 5.1, we know that this conformal bound satisfies the coverage guarantee (5.6) under the exchangeability assumption. Furthermore, it adapts to the measurements y_0 and reconstruction $\widehat{x}_0$ through their effect on $\widehat{z}_0$ and $\{\widetilde{z}_0^{(j)}\}_{j=1}^c$, unlike the non-adaptive bound from Sec. 5.2.1. We refer to these conformal bounds as the "quantile" bounds.

Recalling Sec. 5.2.2, one could interpret $\widehat{z}_0$ as a rough estimate of the αth quantile of $Z_0|Y_0 = y_0$ and $\widehat{\lambda}(d_{\mathsf{cal}})$ as an additive correction that accounts for the finite and approximate nature of the posterior image samples $\{\widetilde{x}_0^{(j)}\}_{j=1}^c$ used to construct $\widehat{z}_0$. For LP metrics, we would instead compute the $(1-\alpha)$-empirical quantile in (5.8). Figure 5.1 illustrates the overall methodology.

5.2.4 A Learned Adaptive Bound on Recovered-image FRIQ

In Sec. 5.2.2, we reasoned that the αth quantile of $Z_0|Y_0=y_0$ yields a valid HP FRIQ bound, but we noted that this quantile is not directly observable. Thus, in Sec. 5.2.3, we used the αth empirical quantile of $\{\widetilde{z}_i^{(j)}\}_{j=1}^{c}$ as a rough estimate "$\widehat{z}_i$" of the desired quantile, after which we used CP to correct this estimate and obtain a valid HP FRIQ conformal bound. However, it is well known from the CP literature that inaccurate base estimators cause loose conformal bounds [5]. Thus, in this section, we aim to improve our estimate of the αth quantile of $Z_0|Y_0=y_0$.

Inspired by conformalized quantile regression [97], we propose to estimate the αth quantile of $Z_0|Y_0=y_0$ using

$$\widehat{z}_i = f(w_i;\xi) \;\; \text{with} \;\; w_i = [\widetilde{z}_i^{(1)},\ldots,\widetilde{z}_i^{(c)}]^\top \in \mathbb{R}^c, \tag{5.9}$$

where ξ are predictor parameters trained using quantile regression (QR) [76]. An example $f(\cdot;\xi)$ is given in App. C.1. In the case of an HP metric, this manifests as

$$\arg\min_{\xi} \sum_{i=n_{\mathsf{cal}}+1}^{n_{\mathsf{cal}}+n_{\mathsf{train}}} \big(\alpha \max(0, z_i - \widehat{z}_i(\xi)) + (1-\alpha)\max(0, \widehat{z}_i(\xi) - z_i)\big) + \gamma\rho(\xi), \tag{5.10}$$

using a training set $d_{\mathsf{train}} = \{(w_i,z_i)\}_{i=n_{\mathsf{cal}}+1}^{n_{\mathsf{cal}}+n_{\mathsf{train}}}$ that is independent of the calibration samples $\{(w_i,z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ and test sample (w_0,z_0).

The first term in (5.10) is the pinball loss [76], which encourages an α-fraction of training samples to violate the HP bound $\widehat{z}_i \le z_i$. The $\rho(\cdot)$ term in (5.10) is regularization that avoids overfitting ξ to the training set. The regularization weight γ can be tuned using k-fold cross-validation. The ξ-dependence of $\widehat{z}_i$ is made explicit in (5.10).

Once the predictor $f(\cdot;\xi)$ is trained, it can be used to obtain the quantile estimates $\{\widehat{z}_i\}_{i=0}^{n_{\mathsf{cal}}}$. Then $d_{\mathsf{cal}} \triangleq \{(\widehat{z}_i,z_i)\}_{i=1}^{n_{\mathsf{cal}}}$ can be used to calibrate the bound parameter λ

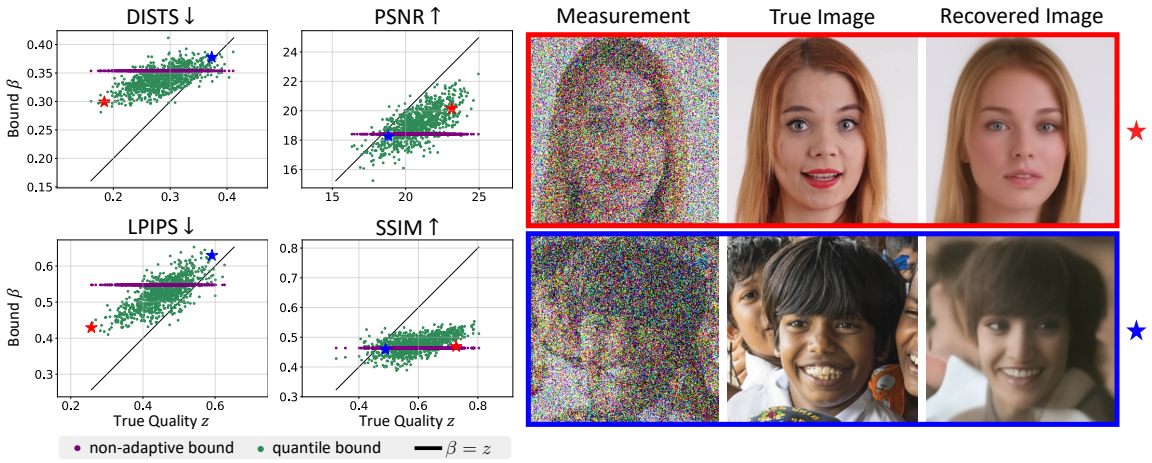


Figure 5.2: Scatter plots show the non-adaptive (purple) and quantile (green) bounds $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$ versus the true FRIQ z_k over FFHQ test indices $k \in \mathcal{I}_{\mathsf{test}}[t]$. The black line shows where $\beta = z$, and a fraction $\alpha = 0.05$ of samples are on the side of the line that violates the bound. The quantile bound tracks the true z_k much better than the non-adaptive bound. The red and blue stars correspond to the images in the red and blue boxes: the red recovery represents better FRIQs and blue represents worse.

using (5.3). As before, the resulting conformal bound $\beta(\widehat{z}_0, \widehat{\lambda}(d_{\mathsf{cal}}))$ will enjoy the coverage guarantee (5.6) under the exchangeability assumption. To handle LP metrics, we would swap α with $1 - \alpha$ in (5.10). Note that any estimation function $f(\cdot; \xi)$ can be used in (5.9), and the best choice will vary with the application. Through the remainder of the paper, we describe these bounds as the "regression" bounds.

5.3 Numerical Experiments

We now consider two imaging inverse problems: image denoising and accelerated MRI. For each, we evaluate the proposed bounds using the PSNR, SSIM [120], LPIPS [126], and DISTS [39] metrics.

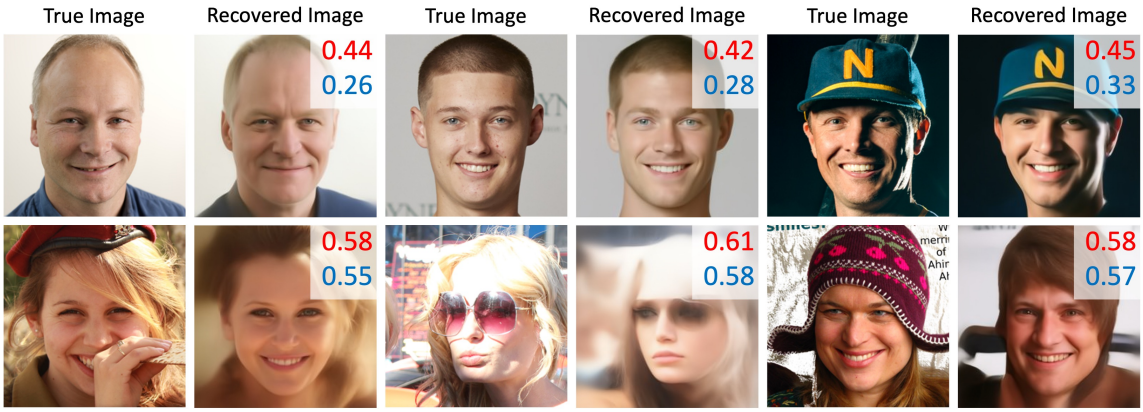


Figure 5.3: Examples from the FFHQ denoising experiment. Top row: true image and low-LPIPS recovery. Bottom row: true image and high-LPIPS recovery. True LPIPS reported in blue and quantile upper-bound in red. (Recall that LPIPS assigns lower values to better recoveries.)

5.3.1 Denoising

Data: For true images, we use a random subset of 4000 images from the Flickr Faces HQ (FFHQ) [66] validation dataset, to which we added white Gaussian noise of standard deviation $\sigma = 0.75$ to create the measurements y_0. The first 1000 images were used to train the predictor $f(\cdot;\xi)$ in (5.9) and the remaining 3000 were used for calibration and testing.

Recovery: To recover $\widehat{x}_0$ from y_0, a denoising task, we use the Denoising Diffusion Restoration Model (DDRM) [69]. Following [69], we run DDRM with a Denoising Diffusion Probabilistic Model (DDPM) [58] pretrained on the CelebA-HQ dataset [65]. To increase sampling diversity, we used $\eta = 1$ and $\eta_b = 0.5$ but set all other hyperparameters at their default values. For each measurement y_i, we use one DDRM sample for the image estimate $\widehat{x}_i$ and c independent samples for $\{\widetilde{x}_i^{(j)}\}_{j=1}^{c}$.

Conformal bounds: We evaluate the proposed bounding methods from Secs. 5.2.1, 5.2.3, and 5.2.4, which we refer to as the **non-adaptive**, **quantile**, and **regression** bounds, respectively. For the regression bound, we use a quantile predictor $f(\cdot, \xi)$ that takes the form of a linear spline with two knots (see Appendix C.1 for more details).

Validation procedure: Because the coverage guarantee (5.6) involves random calibration data and test data, we evaluate our methods using T Monte-Carlo trials. For each trial $t \in \{1, \ldots, T\}$, we randomly select 70% of the 3000 non-training samples to create the calibration set $d_{\mathsf{cal}}[t]$ with indices $i \in \mathcal{I}_{\mathsf{cal}}[t]$, and we use the remaining 30% of the non-training samples for a test fold with indices $k \in \mathcal{I}_{\mathsf{test}}[t]$. In particular, we compute $\widehat{\lambda}$ using $d_{\mathsf{cal}}[t]$ and then, for each sample index $k \in \mathcal{I}_{\mathsf{test}}[t]$, we compute the bound $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$. Finally, we compute performance for each test fold t and average the results across all T trials. Unless specified otherwise, we used error rate $\alpha = 0.05$, $T = 10\,000$, and $c = 32$ samples for the adaptive bounds.

Bound versus true metric: To be useful for individual sample assessment, the bounds should ideally track the true FRIQ such that the bounds are small when the true FRIQ is small and large when the true FRIQ is large. Figure 5.2 shows scatter plots of the non-adaptive and quantile bounds $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$ versus the true FRIQ z_k for the test indices $k \in \mathcal{I}_{\mathsf{test}}[t]$ of a single Monte Carlo trial, along with the true image x_k and recovery $\widehat{x}_k$ for two test samples. The sample highlighted in red has better subjective visual quality compared to the one in blue, and this is reflected in both the true FRIQ metrics z_k and the corresponding quantile bounds, but not the non-adaptive bound. In Fig. 5.3, we show six additional samples from the FFHQ denoising experiment, three with low (true) LPIPS and three with high (true) LPIPS, along with the respective true images. The quantile upper-bound on

Table 5.1: Mean empirical coverage for all bounds with $\alpha = 0.05$ and $T = 10\,000$ on the FFHQ denoising task ($\pm$ standard error). Quantile and regression bounds are computed with $c = 32$.

Bound	DISTS	LPIPS	PSNR	SSIM
Non-adaptive	0.95000 ± 0.00009	0.95016 ± 0.00009	0.95004 ± 0.00009	0.95010 ± 0.00009
Quantile	0.95002 ± 0.00009	0.95013 ± 0.00009	0.95003 ± 0.00009	0.95006 ± 0.00009
Regression	0.95013 ± 0.00009	0.95026 ± 0.00009	0.95001 ± 0.00009	0.95006 ± 0.00009

LPIPS is superimposed on each recovery. We see that the bounds are valid in the sense that they did not under-predict the true LPIPS, and adaptive in the sense that the bounding value is lower when the true LPIPS is lower.

Empirical Coverage: To verify the coverage guarantee in (5.6) is satisfied, we compute the empirical coverage as

$$\mathrm{EC}[t] \triangleq \frac{1}{|\mathcal{I}_{\mathsf{test}}[t]|} \sum_{k \in \mathcal{I}_{\mathsf{test}}[t]} \mathbb{1}\{z_k \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{z}_k)\}, \tag{5.11}$$

for each Monte Carlo trial t. In Table 5.1, we report the average empirical coverage and standard error across $T = 10\,000$ trials for all three methods on the FFHQ data using $\alpha = 0.05$. For all methods, the average empirical coverage is very close to the theoretical coverage $1 - \alpha = 0.95$ regardless of the metric, demonstrating close adherence to the theory. In Appendix D.1, we further demonstrate that this adherence holds independent of the choice of c.

MCB versus bounding method and number of posterior samples c: To assess the tightness of the conformal bounds, we average the bound $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$ over the test indices $k \in \mathcal{I}_{\mathsf{test}}[t]$ and the Monte Carlo trials t to yield the "mean conformal bound" (MCB). Figure 5.4 plots MCB versus the number of posterior samples c used for the adaptive bounds. The figure shows that the non-adaptive bound is looser (i.e.,

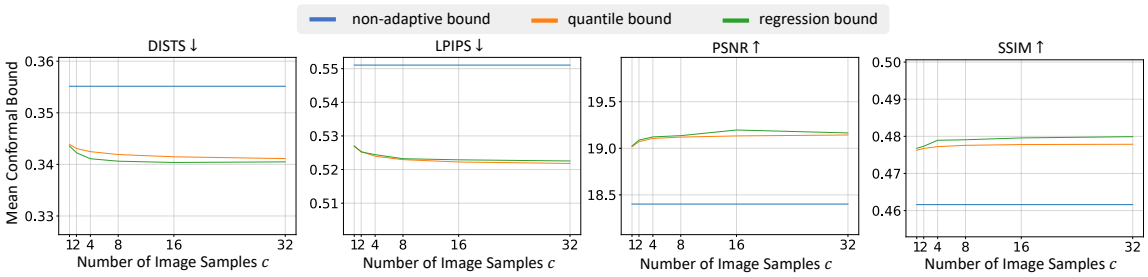


Figure 5.4: Mean conformal bound versus number of posterior samples c for FFHQ denoising.

smaller for the HP metrics PSNR and SSIM and larger for the LP metrics DISTS and LPIPS) than the two adaptive bounds. For both adaptive bounds, Fig. 5.4 shows only minor bound improvement with increasing c, suggesting that the adaptive bounds are robust to the choice of c, and that small values of c could suffice if sample-generation was computationally expensive.

Interestingly, Fig. 5.4 shows relatively little improvement when going from the quantile bound to the regression bound. This may be due to our choice of a linear spline with two knots for $f(\cdot;\xi)$, but experiments with higher spline orders and/or more knots did not yield improved results, and neither did experiments with XGBoost [30] models for $f(\cdot;\xi)$. Additional experiments that hold the number of test samples at 900 and vary n_{train} and n_{cal} such that $n_{\mathsf{train}} + n_{\mathsf{cal}} = 3100$ (see Appendix D.2) also show little change in the performance of the quantile and regression bounds. Thus, for our experimental data, the effort to train the estimation function $f(\cdot;\xi)$ from (5.9) may not be justified, given the good performance of the simple empirical-quantile estimation function $f(\cdot)$ from (5.8). But the behavior may be different with other datasets.

Computation time: Computing a single DDRM sample takes approximately 2.73 seconds. Once the calibration constant $\widehat{\lambda}(d_{\mathsf{cal}})$ is known, computing $c = 32$ FRIQ samples $\{\widetilde{z}_0^{(j)}\}_{j=1}^{c}$ and $\beta(\widehat{z}_0, \widehat{\lambda}(d_{\mathsf{cal}}))$ takes around 217ms, 320ms, 5ms, and 6ms for DISTS, LPIPS, PSNR, and SSIM, respectively. All times pertain to a single NVIDIA V100 with 32GB of memory.

5.3.2 Accelerated MRI

We now simulate our methods on accelerated multi-coil MRI [75, 55]. As before, when the acceleration rate $R > 1$, the inverse problem may become ill-posed, in which case one may be interested in bounding the FRIQ of the recovered image.

Data: We utilize the non-fat-suppressed subset of the multi-coil fastMRI knee dataset [125], yielding 17286 training images and 2188 validation images. To simulate the imaging process, we retrospectively sub-sample in the spatial Fourier domain (the "k-space") using random Cartesian masks that give acceleration rates $R \in \{16, 8, 4, 2\}$. We use the same nested sampling masks as Sec. 4.3.

Recovery: For the recovery network $g_\theta(\cdot)$ of all methods, we use the well-known E2E-VarNet [107], which is a deterministic reconstruction approach. To generate approximate posterior samples for the adaptive bounds, we utilize the conditional normalizing flow (CNF) from Ch. 3 with the modifications in Appendix A.2. Both networks are trained to work well with all four acceleration rates R. (See App. A.2 for training details.) To work with multi-coil MRI, we first compute the magnitude images using the RSS (2.5) before computing any metric. Since DISTS and LPIPS require a 3-channel image, we repeat the magnitude image for all three channels and normalize the values to be between 0 and 1 before computing either metric. Similar to

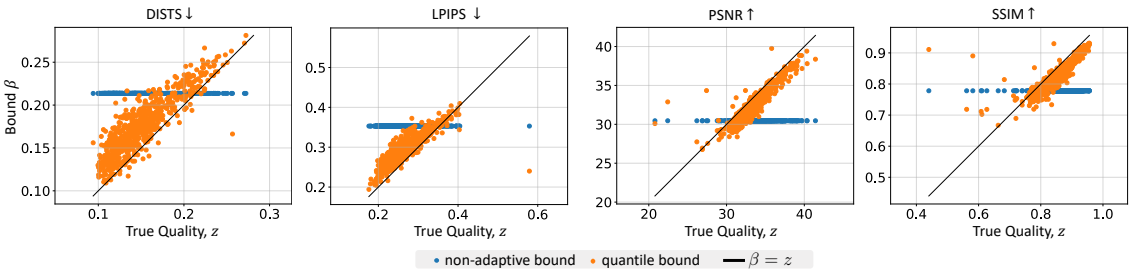


Figure 5.5: Scatter plots show the non-adaptive (blue) and quantile (orange) bounds $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$ versus the true FRIQ z_k over MRI test indices $k \in \mathcal{I}_{\mathsf{test}}[t]$ at acceleration $R = 8$. The black line shows where $\beta = z$. A fraction of $\alpha = 0.05$ samples are on the side of the line that violates the bound. Note that the quantile bound tracks the true z_k much better than the non-adaptive bound.

Sec. 5.3.1, we found that the regression bound did not provide significant gain over the quantile bound and so, to streamline our discussion, we consider only the quantile and non-adaptive bounds for MRI. As before, we evaluate performance over $T = 10\,000$ Monte Carlo trials with a random 70% calibration and 30% test split of the validation data. All experiments use an error-rate $\alpha = 0.05$. Methods are separately calibrated for each acceleration rate.

Bound versus true-metric: Figure 5.5 shows scatter plots of the true FRIQ z_k versus the non-adaptive and quantile bounds $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$ for the test indices $k \in \mathcal{I}_{\mathsf{test}}[t]$ in a single Monte-Carlo trial t. The results are shown for $R = 8$ acceleration and $c = 32$ samples in the adaptive bounds. Except for a few outliers, the quantile bound closely tracks the true FRIQ z_k, demonstrating good adaptivity, while the non-adaptive bounds remain constant with z_k.

Multi-round Measurement: To showcase the practical impact of our bounds, we adapt the multi-round measurement protocol from Ch. 4, where measurements are collected over multiple rounds until the uncertainty bound falls below a threshold. In

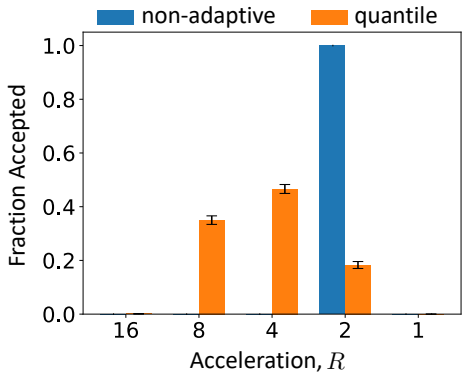


Figure 5.6: Fraction of accepted slices versus final acceleration rate for multi-round MRI using DISTS with $\tau = 0.16$. Error bars show standard deviation.

Table 5.2: Average results for a multi-round MRI simulation where measurement collection stop once bounds are below a user-set threshold τ. Results shown for $T = 10\,000$ trials using the DISTS metric with $\alpha = 0.05$, $\tau = 0.16$, and $c = 32$ ($\pm$ standard error).

Method	Average Acceleration	Acceptance Empirical Coverage
Non-adaptive	2.000 ± 0.000	0.9504 ± 0.0001
Quantile	3.973 ± 0.001	0.9323 ± 0.0001

our setting, measurements are first collected at acceleration $R = 16$, an image recovery is computed, and a conformal upper-bound on its DISTS is computed. If the bounding value is lower than a pre-determined threshold τ, signifying that the recovery is (with probability $1 - \alpha$) of sufficient diagnostic quality [68], then measurement collection stops. If not, additional measurements are collected and combined with the previous ones to yield an acceleration of $R = 8$, and the process repeats. We allow up to five measurement rounds, corresponding to final accelerations of $R \in \{16, 8, 4, 2, 1\}$.

Once again, we report average results across $T = 10\,000$ trials. We set $\tau = 0.16$, which is where we find the non-adaptive approach requires all slices to be collected at $R = 2$. Figure 5.6 plots the fraction of test image slices accepted by the multi-round protocol at each acceleration rate R with $\tau = 0.16$. With the quantile bound, the measurements stop after three of fewer rounds (i.e., $R \geq 4$) in more than 80% of the cases. With the non-adaptive bound, the measurements stop after four rounds

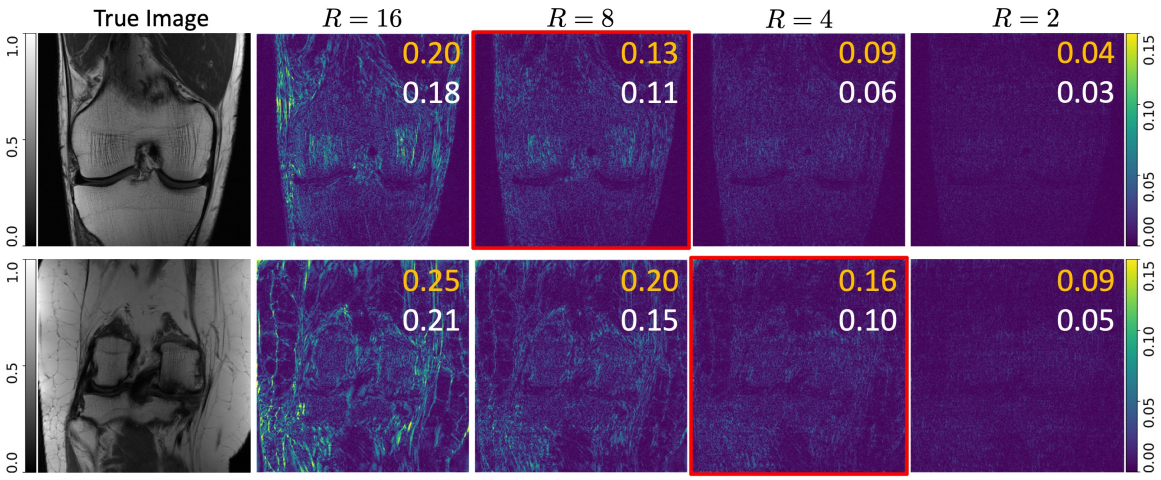


Figure 5.7: Examples of the multi-round MRI measurement procedure with DISTS at $\alpha = 0.05$, $\tau = 0.16$, and $c = 32$. Error images at each acceleration R are shown with the quantile bound (orange) and true metric (white). The red box indicates the measurement round at which the bound falls below the threshold τ and the measurement procedure concludes.

(i.e., $R = 2$) in all cases. Table 5.2 shows that, with the quantile bound, the multi-round protocol attains an average acceleration of $R = 3.973$, which far surpasses the $R = 2$ acceleration achieved with the non-adaptive bound. Table 5.2 also shows that the empirical coverage of the multi-round accepted slices is very close to $1 - \alpha$, despite having only coverage guarantees (5.6) for a single-round measurement at each acceleration rate. Figure 5.7 shows examples of the image-error, the true DISTS, and its quantile upper-bound for each measurement round. With the threshold set at $\tau = 0.16$, the example on the top would collect two rounds of measurements (i.e., $R = 8$) while the example at the bottom would collect three rounds of measurements (i.e., $R = 4$), as demarcated by the red squares. See App. D.3 for additional qualitative results.

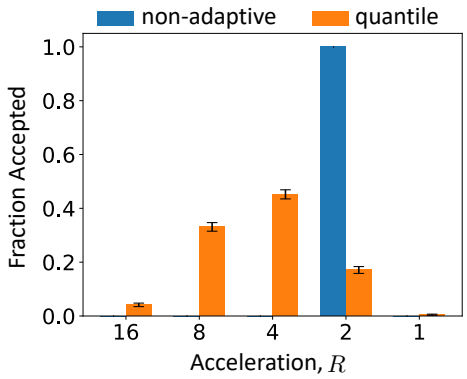


Figure 5.8: Fraction of accepted slices versus final acceleration rate for multi-round MRI using PSNR with $\tau = 33$dB. Error bars show standard deviation.

Table 5.3: Average results for a multi-round MRI simulation where measurement collection stop once bounds are above a user-set threshold τ. Results shown for $T = 10\,000$ trials using the PSNR metric with $\alpha = 0.05$, $\tau = 33$dB, and $c = 32$ ($\pm$ standard error).

Method	Average Acceleration	Acceptance Empirical Coverage
Non-adaptive	2.000 ± 0.000	0.9503 ± 0.0001
Quantile	4.048 ± 0.001	0.9514 ± 0.0001

As PSNR is often a more recognized metric, we additionally perform the multi-round experiments using PSNR as our metric of interest and instead cease measurement collection once the conformal lower-bound exceeds the threshold τ. We set $\tau = 33$dB, which we find as a threshold value that requires the non-adaptive approach to collect all slices at $R = 2$. Following a similar trend as with DISTS, we see a large proportion of the slices collected at $R \geq 4$ for the quantile bounds in Fig. 5.8. In Tab. 5.3, we see that this results in an average accepted acceleration rate of $R = 4.048$, over twice the acceleration achieved with the non-adaptive bounds.

Computation time: The E2E-VarNet takes approximately 104ms to generate a single posterior sample, while the CNF take about 1.22 seconds to generate 32 posterior samples (corresponding to $c = 32$) on a single NVIDIA V100. The computation time of the metrics and bounds is on par with the times reported for the FFHQ experiments.

Limitations: We acknowledge multiple limitations in our proposed methodology. 1) Our methods require access to calibration data $\{(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$ that is similar enough

to the test data (x_0, y_0) for the FRIQ pairs $\{(\widehat{z}_i, z_i)\}_{i=0}^{n_{\mathsf{cal}}}$ to be modeled as statistically exchangeable. More work is required to make our methods robust to distribution shift (see App. D.4), although several works [112, 15, 27] have proposed modifications to the conformal procedure that may suggest some paths forward. 2) Our methods will be most impactful when there exists evidence that the FRIQ metric is well matched to the application (e.g., DISTS for MRI [68]). For some applications, additional work is required to determine which metrics are more appropriate. 3) Our MRI application ideas are preliminary and not ready for practical use; rigorous clinical trials are needed to tune and validate the methodology on a much larger and diverse cohort of data. 4) The learned adaptive bound from Sec. 5.2.4 requires training a quantile regression model, and our FFHQ denoising experiment suggests that it may not be easy to significantly outperform the simpler adaptive bound from Sec. 5.2.3. 5) The posterior samplers that we considered in our numerical experiments target only aleatoric uncertainty, and sharper conformal bounds might be attained if epistemic uncertainty was also considered (e.g., Ekmekci et al. [45]). 6) Because our methods are based on CP (or, equivalently, conformal risk control under the indicator loss [6]), the marginal guarantee (5.6) holds with probability $1-\alpha$ over random test data (e.g., $\widehat{Z}_0, Z_0$) and calibration sets D_{cal}. A more fine-grained coverage could be achieved via the Risk-Controlling Prediction Sets (RCPS) framework from Bates et al. [16], which employs two user-selected error rates $\alpha, \delta \in (0,1)$ to yield coverage guarantees like

$$\Pr\left[\Pr\left\{Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}\big(\widehat{Z}_0\big) \big| D_{\mathsf{cal}}\right\} \geq 1-\alpha\right] \geq 1-\delta \tag{5.12}$$

in place of (5.6). In (5.12), α controls the D_{cal}-conditional error while δ controls the error over D_{cal}.

5.4 Conclusion

For imaging inverse problems, we used conformal prediction to construct bounds on the FRIQ of a recovered image relative to the unknown true image. When constructed using a calibration set that is statistically exchangeable with the test sample, our bounds are guaranteed to hold with high probability. Two of our methods leveraged approximate-posterior-sampling schemes to yield tighter conformal bounds that adapt to the measurements and reconstruction. Our approaches were demonstrated on image denoising and accelerated multi-coil MRI, illustrating the broad applicability of our work.

Chapter 6: Final Thoughts

6.1 Future Work

The methods presented in this dissertation should serve as a solid foundation for future works in UQ for accelerated MRI. For one, our task-based (Ch. 4) and image-quality-based (Ch. 5) approaches construct a single prediction set that contains a single target with high probability. A natural extension would be to construct multiple prediction sets that bound multiple targets simultaneously with statistical guarantees. This could be prediction sets for multiple soft-output pathology classifiers, multiple FRIQ metrics, or a mix of both. Several works [87, 88, 49, 93, 99] on multi-target conformal prediction may give insight into how this may be best performed. Integrating adaptive sampling [100] within our multi-round measurement protocol could be a promising direction as well. This would not only provide an uncertainty-based stopping criterion but also actively choose which measurements to collect next to best reduce the uncertainty. With proper implementation, one would expected to see even further accelerations enabled while maintaining sufficient uncertainty levels. As mentioned previously, the application of any of our proposed methods to safety-critical domains like MRI requires substantially more validation to ensure high performance across protected attributes and scenarios. In the case of medical imaging, this would

take the form of rigorous clinical trials to tune hyperparameters and assess real-world performance.

6.2 Conclusion

Advancements in accelerated magnetic resonance imaging hold promise in reducing MRI scan times by a significant margin without reducing the diagnostic utility. However, due to the consequential nature of medical imaging, the lack of understanding in the uncertainty of the recovered images remains a major barrier for the adoption of highly-accelerated MRI. To fill this gap, we proposed a multi-faceted approach to tackle uncertainty quantification in multi-coil accelerated magnetic resonance imaging.

In Ch. 3, we designed a novel conditional normalizing flow to generate many posterior sample estimates for a given measurement. By computing a pixel-wise standard-deviation map across several estimates, we could visualize which areas of the image contain more certainty and are more trustworthy for diagnostic conclusions. Our CNF demonstrated fast inference speeds while outperforming existing posterior-sampling-based methods in almost all metrics on the fastMRI brain and knee datasets.

In Ch. 4, we quantified the additional uncertainty contributed by the acceleration process to a downstream task like soft-output pathology classification. This assessed how well the downstream task could be performed given posterior estimates relative to having access to the true fully-sampled image. Using conformal prediction, we constructed a prediction interval from posterior estimates that was guaranteed to contain the task-output assigned to the true image with high probability. From this, we introduced a multi-round measurement protocol to collect data until the task uncertainty fell below a desired level. Following this protocol, we showed high

acceleration rates could often be utilized while limiting the uncertainty on meniscus-tear classification.

Lastly, Ch. 5 detailed our approach to quantify the uncertainty on the image quality of an accelerated reconstruction without access the true image. With only posterior estimates, we construct bounds on quality metrics like PSNR, which hold with conventional conformal guarantees. Applying the multi-round protocol, we demonstrated high acceleration rates could be achieved while ensuring the image quality was beyond a sufficient level with high probability.

By representing the uncertainty from three different perspectives, we provide a more comprehensive understanding of the uncertainty inherent in accelerated MRI. This provides radiologist with more information to better assess the diagnostic validity and trustworthiness of accelerated reconstructions. By establishing more confidence on the side of practitioners, our hope is that highly-accelerated MRI may be safely adopted, giving patients better access to critical imaging tests without sacrificing quality-of-care.

Appendix A: Task-based UQ Details

A.1 Mask Details

In Section 4.3.4, we simulate a multi-round measurement process whereby, in round $k = 1, \ldots, 5$, k-space samples are collected so that the accumulated samples up to round k correspond to acceleration rate $R^{[k]} \in \{16, 8, 4, 2, 1\}$. The pattern of collected 2D samples is known as the "sampling mask," and the choice of the mask can have a great impact on recovery quality.

In our work, we use Cartesian masks, where the sampling pattern consists of entire lines in 2D k-space. For round $k = 1$, we use a Golden Ratio Offset (GRO) mask [63] designed for rate $R^{[1]} = 16$ with the "α" and "s" parameters from Joshi et al. [63] chosen as $s = 15$ and $\alpha = 8$ (not to be confused with the meanings of α and s elsewhere in this paper). This provides a densely sampled rectangle in the center of the k-space, known as the autocalibration signal (ACS) region, which is 9 lines wide.

For round $k = 2$, we first collect 7 additional lines in the center of k-space, yielding an ACS region of width 16. We then sample additional k-space lines, using a sampling probability that is inversely proportional to the distance from the center, until an accumulated acceleration rate of $R^{[2]} = 8$ is achieved. In the next two rounds, this procedure is repeated to obtain ACS regions of width 24 and 32, respectively, and

accumulated accelerations of $R^{[3]} = 4$ and $R^{[4]} = 2$, respectively. The four resulting masks are shown in Figure 4.4. The last round samples everywhere in k-space, achieving acceleration $R^{[5]} = 1$.

A.2 Network and Training Details

For all models, we use an Adam optimizer [74] with the default parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train each reconstruction network with all four accelerations. More specifically, one of the four sampling masks is drawn uniformly at random for every sample in each epoch. This allows the model to see each training sample at a different acceleration during training.

For the E2E VarNet, we utilize the author's implementation at [108]. We keep the default parameters listed for the fastMRI knee leaderboard and train to minimize the structural similarity (SSIM) [120] loss for 50 epochs with a learning rate of 0.0001 and batch size of 16. This takes around 38 hours on a single NVIDIA V100 with 32GB of memory.

For the CNF, we modify the parameters and architecture from Ch. 3 slightly. To better handle multiple accelerations, we increase the size of the conditioning network to have 256 initial channels. We also add an iMAP [109] invertible attention module to the end of each flow step, and use 2 layers and 10 flow steps per layer. The CNF is trained to minimize the negative log-likelihood for 150 epochs with a learning rate of 0.0001 and batch size of 8. On a single NVIDIA V100, this takes around 335 hours.

For the soft-output binary classification network, we start with a ResNet50 [56] that is initialized with weights from a network trained on ImageNet [36]. This task network takes in 3-channel images, so we convert the multi-coil image to a magnitude

image using RSS (2.5) and feed the magnitude image into all three input channels of the classifier. We pretrain the network with self-supervision using SimCLR [31] for 500 epochs with a learning rate of 0.0002 and batch size of 128. Next, we train the network in a supervised fashion to minimize the binary cross-entropy loss for 100 epochs. During the supervised training, we use an ℓ_2-bounded projected gradient descent attack with 10 steps and a perturbation budget of 1.5 in order to make our network robust to ℓ_2-bounded adversarial attacks. The adversarial training is implemented using the robustness package [46]. The classifier is trained using a learning rate of 0.0001 and batch size of 128 with a weight decay of 0.01. To prevent overfitting to the training data, we early-stop at the epoch that maximizes the area-under-the-receiver-operating-characteristic (AUROC) on the validation data.

All networks were implemented with PyTorch [94] and PyTorch Lightning [48]. Code is available at `https://github.com/jwen307/TaskUQ`.

Appendix B: Task-based UQ Additional Results

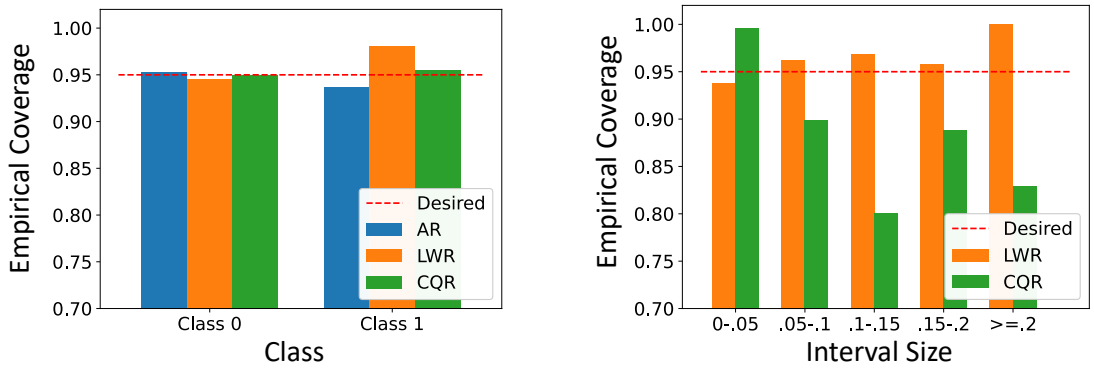
B.1 Conditional Coverage Experiments

As described in Section 4.1, conformal prediction typically provides only marginal coverage guarantees. As a result, a conformal-based approach may under or over cover test samples with certain attributes; thus, it is important to evaluate conformal methods under different conditioning scenarios. In this section, we empirically evaluate different forms of conditional coverage. To do this, we use the Monte-Carlo testing procedure described in Section 4.3 with $T = 10000$, $R = 8$, $c = 32$, and $\alpha = 0.05$.

First, we evaluate the class-conditional coverage

$$\Pr\left(Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\{\widehat{X}_0^{(j)}\}) \,\middle|\, X_0 \text{ labeled as } l\right) \tag{B.1}$$

for $l = 0$ (no pathology) and $l = 1$ (pathology). This amounts to calculating the empirical coverage for all images of class 0 (no pathology) and class 1 (pathology) separately, so we can analyze if the coverage drops severely for either of the classes. Figure B.1a shows that, although all three methods provide close to the desired coverage of 95%, the AR method shows the worst under-coverage, which is $\approx 93\%$ for class 1. Since a missed detection may carry higher consequence than a false alarm, under-coverage of class 1 should be considered carefully. Better class-conditional coverage could likely be attained using class-conditional conformal prediction [116].



(a) Class-Conditional Coverage

(b) Size-Stratified Coverage

Figure B.1: Coverage conditioned on a) class and b) interval-size for $T = 10000$, $R = 8$, $c = 32$, and $\alpha = 0.05$. Empirical coverage is close to $1 - \alpha = 0.95$ across both classes for each method. LWR maintains higher coverage across interval-sizes compared to CQR.

Next we evaluate the size-stratified coverage [8]

$$\Pr\left(Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\text{cal}})}(\{\widehat{X}_0^{(j)}\}) \,\middle|\, |\mathcal{C}_{\widehat{\lambda}(D_{\text{cal}})}(\{\widehat{X}_0^{(j)}\})| \in \mathcal{S}\right) \tag{B.2}$$

for size intervals $\mathcal{S} \in \{[0, 0.05], [0.05, 0.1], [0.1, 0.15], [0.15, 0.2], [0.2, 1]\}$. Figure B.1b shows that LWR demonstrates much more consistent size-stratified coverage than CQR. However, in cases such as multi-round sampling with $\tau \leq 0.05$, one may be concerned only with the coverage of small intervals, such as those with lengths ≤ 0.05, where Figure B.1b suggests that CQR's coverage is very good.

B.2 Image Recovery Performance

Since our method focuses on uncertainty metrics like prediction-interval length, one may wonder how well the E2E-VarNet and CNF recovery approaches work according to traditional image-recovery metrics like peak-signal-to-noise ratio (PSNR), structural

similarity index (SSIM) [120], and Fréchet Inception Score (FID) [57]. We provide those details in this section.

When computing these metrics, we used the RSS magnitude approximation from (2.5). Also, following the approach of the fastMRI paper [124], we compute PSNR and SSIM across entire volumes, rather than for each image slice separately. For the CNF method, PSNR and SSIM are computed on the posterior-mean approximation computed by averaging $c = 32$ posterior samples. When computing FID, we use the VGG-16 embedding [103] and $c = 1$ samples for the CNF, and we compute reference statistics using the entire training set.

Table B.1 shows the PSNR, SSIM, and FID performance of the E2E-VarNet and the CNF evaluated on the entire validation set at several accelerations R. There we see that the E2E-VarNet slightly outperforms the CNF in PSNR and SSIM, but not FID. Because the models were trained to handle four different accelerations, the results in Table B.1 are slightly below those reported in the original E2E-VarNet paper [107] and Ch. 3.

Figure B.2 shows example reconstructions from the E2E-VarNet and CNF approaches, as well as standard-deviation maps for the CNF. As expected, the posterior standard deviation decreases with the acceleration factor R.

B.3 Performance of Classifier

One may also wonder about the performance of our soft-output meniscus-tear classifier according to standard classification metrics. Table B.2 shows the accuracy, precision, recall, and AUROC evaluated on the validation set described in Section 5.3 with meniscus tear annotations from fastMRI+ [128]. From the table, we see that our

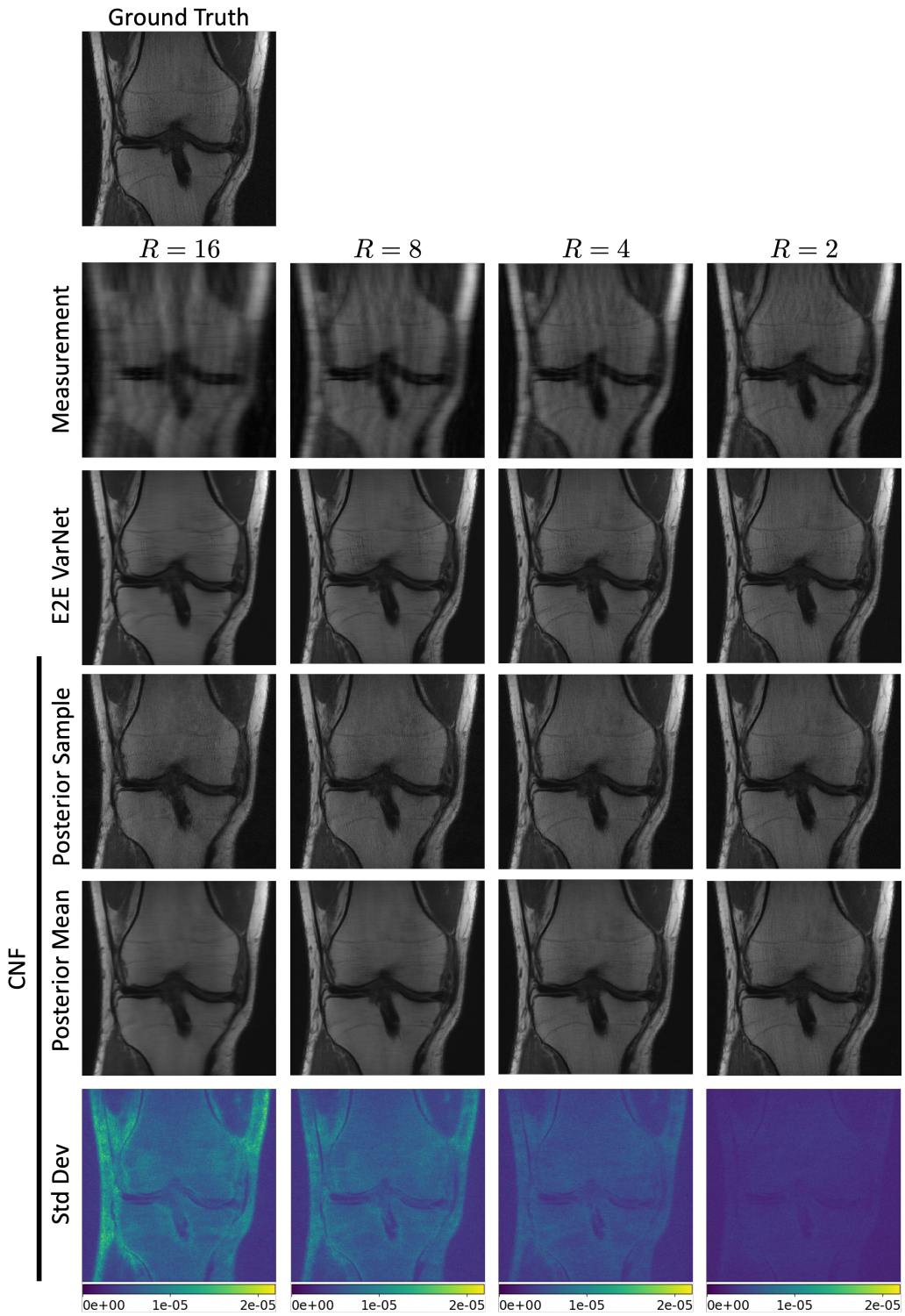


Figure B.2: Example MRI reconstructions and standard-deviation maps for several accelerations R.

Table B.1: Image-recovery metrics (± standard error) versus acceleration R.

R	Network	PSNR ↑	SSIM ↑	FID ↓
2	E2E-VarNet	42.341 ± 0.273	0.966 ± 0.002	2.847
	CNF	41.366 ± 0.248	0.960 ± 0.002	2.575
4	E2E-VarNet	38.700 ± 0.238	0.937 ± 0.003	4.048
	CNF	37.974 ± 0.216	0.926 ± 0.003	3.349
8	E2E-VarNet	36.120 ± 0.212	0.906 ± 0.003	5.680
	CNF	35.593 ± 0.196	0.892 ± 0.004	4.346
16	E2E-VarNet	32.911 ± 0.194	0.859 ± 0.004	9.668
	CNF	32.684 ± 0.177	0.842 ± 0.004	6.233

Table B.2: Validation performance of the meniscus-tear classifier

Accuracy	Precision	Recall	AUROC
0.775	0.391	0.929	0.922

classifier exhibits relatively high recall but low precision, which may be preferable in the context of meniscus-tear diagnosis, where missed detections might be more costly than false alarms. Given that the dataset used to train the classifier was relatively small, we conjecture that these performance metrics could be greatly improved with more data and a better balancing across classes. That said, we believe that this classifier suffices as a task-based uncertainty evaluation tool.

Appendix C: Image-quality-based UQ Details

C.1 Training/Model details

For the regression bound in Sec. 5.2.4, where $w_i \in \mathbb{R}^c$, we use a quantile predictor of the form

$$f(w_i;\xi) = \psi(w_i)^\top \xi_1 + \xi_2 \;\; \text{with} \;\; \xi = [\xi_1, \xi_2]^\top, , \tag{C.1}$$

where $\psi(\cdot)$ is a linear spline with two knots, t_1 and t_2, implemented via the truncated power basis

$$\psi(w_i) = [w_i; (w_i - t_1\underline{1})_+; (w_i - t_2\underline{1})_+] \in \mathbb{R}^{3c}, \tag{C.2}$$

with $\underline{1}$ the c-dimensional vector of ones and $(x)_+ \triangleq \max(x, 0)$. The two knots were placed at the $\frac{1}{3}$ and $\frac{2}{3}$ empirical quantiles of the mean training feature computed as $\{\frac{1}{c}\sum_{j=1}^{c} \widetilde{z}_i^{(j)}\}_{i=n_{\mathsf{cal}}+1}^{n_{\mathsf{cal}}+n_{\mathsf{train}}}$, respectively. Essentially, for each feature in w_i, (C.1) implements a piece-wise-linear regression function with three distinct pieces. To promote consistency in $w_i = [\widetilde{z}_i^{(1)}, \widetilde{z}_i^{(2)}, \ldots, \widetilde{z}_i^{(c)}]^\top$ across different i, the spline function $\phi(\cdot)$ first sorts the values $\{\widetilde{z}_i^{(j)}\}_{j=1}^{c}$ within each w_i. For $\rho(\xi)$ in (5.10), we use ridge regularization on the weights w. The resulting (5.10) is a quadratic program, which can be optimized using any convex solver. To tune the regularization weight γ, we use K-fold cross

validation with $K = 5$ folds and select the weight that provides the lowest mean pinball loss across the 5 folds.

For DDRM, we use the author's implementation [70], which is publicly available under an MIT license.

To compute the quadratic program for Sec. 5.3.1, we use the qpsolver [26] package under a LGPL 3.0 license along with the CVXOPT [4] package under a GNU General Public License.

We use the TorchMetrics [25] package under the Apache 2.0 license to compute PSNR, SSIM, and LPIPS. We use the author's code at [38] for DISTS under a MIT license.

All models use the PyTorch [94] framework with a custom license allowing open use. The E2E-VarNet and CNF are implemented using PyTorch Lightning [48] under an Apache 2.0 license.

Appendix D: Image-quality-based UQ Additional Results

D.1 Empirical coverage

In Sec. 5.3.1, we empirically demonstrated that the coverage guarantees in (5.6) are met for the non-adaptive, quantile, and regression bounds in the FFHQ denoising experiments. Here, we further demonstrate that these guarantees hold regardless of the number of posterior samples c used to compute the adaptive bounds. Tables D.1 and D.2 show the average empirical coverage for the quantile and regression method, respectively, across $T = 10\,000$ trials for different values of c and $\alpha = 0.05$. The same number of posterior samples c is used during calibration and to compute the adaptive bounds during testing. Again, we observe that the average empirical coverage is very close to the desired $1-\alpha$ in all cases though there are very slight deviations as a result of finite trials, number of calibration samples, and number of testing samples.

In Table D.3, we report the mean empirical coverage for the quantile method in the MRI experiments with $\alpha = 0.05$, $c = 32$, and acceleration rate $R \in \{2, 4, 8, 16\}$ across $T = 10\,000$ trials. For any value of R, we see the empirical coverage is very close to the theoretical $1 - \alpha = 0.95$ coverage; thus, once again, our method shows close compliance to the theory.

Table D.1: Mean empirical coverage for the quantile method with $\alpha = 0.05$ and $T = 10\,000$ on the FFHQ denoising task ($\pm$ standard error)

c	DISTS	LPIPS	PSNR	SSIM
1	0.95002 ± 0.00009	0.94997 ± 0.00009	0.95013 ± 0.00009	0.94989 ± 0.00009
2	0.95006 ± 0.00009	0.95003 ± 0.00009	0.95001 ± 0.00009	0.95022 ± 0.00009
4	0.94997 ± 0.00009	0.95008 ± 0.00009	0.94986 ± 0.00009	0.94999 ± 0.00009
8	0.95020 ± 0.00009	0.95015 ± 0.00009	0.95019 ± 0.00009	0.94991 ± 0.00009
16	0.94998 ± 0.00009	0.94999 ± 0.00009	0.95009 ± 0.00009	0.95008 ± 0.00009
32	0.95002 ± 0.00009	0.95013 ± 0.00009	0.95003 ± 0.00009	0.95006 ± 0.00009

Table D.2: Mean empirical coverage for the regression method with $\alpha = 0.05$ and $T = 10\,000$ on the FFHQ denoising task ($\pm$ standard error)

c	DISTS	LPIPS	PSNR	SSIM
1	0.94994 ± 0.00009	0.94970 ± 0.00009	0.95009 ± 0.00009	0.95014 ± 0.00009
2	0.95011 ± 0.00009	0.94953 ± 0.00009	0.94985 ± 0.00009	0.95004 ± 0.00009
4	0.94996 ± 0.00009	0.94946 ± 0.00009	0.95003 ± 0.00009	0.94995 ± 0.00009
8	0.95004 ± 0.00009	0.94964 ± 0.00009	0.94999 ± 0.00009	0.95017 ± 0.00009
16	0.94986 ± 0.00009	0.94964 ± 0.00009	0.95007 ± 0.00009	0.94987 ± 0.00009
32	0.95013 ± 0.00009	0.95026 ± 0.00009	0.95001 ± 0.00009	0.95006 ± 0.00009

D.2 Additional FFHQ denoising experiments

Effect of training and calibration set size: For FFHQ denoising, we now investigate how the amount of training and calibration data affect the mean conformal bound. Following the same Monte Carlo procedure as Sec. 5.3.1, we fix the number of testing samples to 900 but change the proportion of n_{train} versus n_{cal} for the remaining 3100 samples. In Fig. D.1, we show the mean conformal bounds as the proportion of training samples varies, starting with 0.1 and going up to 0.95, for $T = 10\,000$,

Table D.3: Mean empirical coverage for the quantile method across accelerations with $\alpha = 0.05$, $c = 32$, and $T = 10\,000$ on the accelerated MRI task ($\pm$ standard error). All coverages are above the expected coverage of $1 - \alpha = 0.95$

R	DISTS	LPIPS	PSNR	SSIM
2	0.9503 ± 0.0001	0.9503 ± 0.0001	0.9504 ± 0.0001	0.9504 ± 0.0001
4	0.9504 ± 0.0001	0.9503 ± 0.0001	0.9505 ± 0.0001	0.9504 ± 0.0001
8	0.9503 ± 0.0001	0.9504 ± 0.0001	0.9503 ± 0.0001	0.9503 ± 0.0001
16	0.9504 ± 0.0001	0.9504 ± 0.0001	0.9505 ± 0.0001	0.9506 ± 0.0001

$c = 32$, and $\alpha = 0.05$. Both adaptive methods still provide noticeable gains over the non-adaptive bound. Even with additional training samples, however, the regression bounds show relatively little improvement over the quantile bounds. Based on (5.3), the conformal bounds should grow more conservative as the number of calibration points decreases for the non-adaptive and quantile bounds. However, this effect is not evident until very small calibration set sizes (e.g., when the fraction of calibration samples is 0.05).

Correlation between conformal bound and true FRIQ: Figure 5.2 visually demonstrates that the quantile bound tracks the true FRIQ much better than the non-adaptive bound. To quantify this tracking behavior, we compute the Pearson correlation coefficient between each conformal bound $\beta(\widehat{z}_k, \widehat{\lambda}(d_{\mathsf{cal}}[t]))$ and the true FRIQ z_k over the test indices $k \in \mathcal{I}_{\mathsf{test}}[t]$ for each Monte-Carlo trial t. In Fig. D.2, we plot the mean (across $T = 10000$ trials) Pearson correlation coefficient versus c for each bound. Since the non-adaptive bound is constant with z_k, its correlation equals 0. However, the two adaptive approaches demonstrate a correlation coefficient above 0.5, and up to 0.7, depending on the metric. These correlation coefficients quantify the adaptivity of our bounds and explain, in part, why the adaptive bounds led to

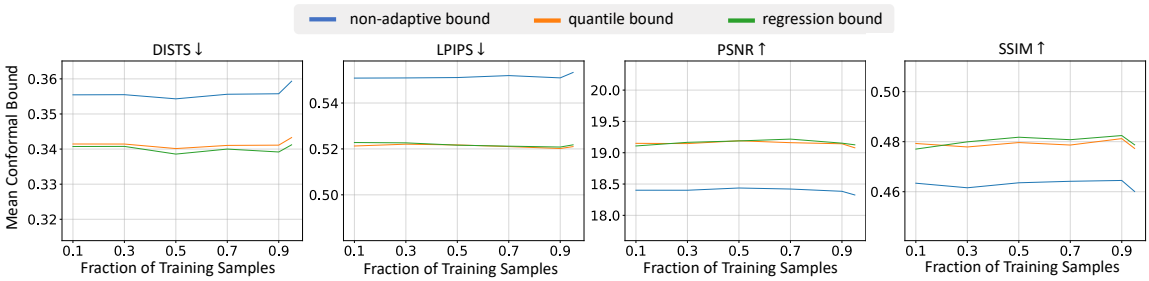


Figure D.1: Mean conformal bound versus the proportion of training samples for FFHQ denoising with $n_{\text{train}} + n_{\text{cal}} = 3100$ samples.

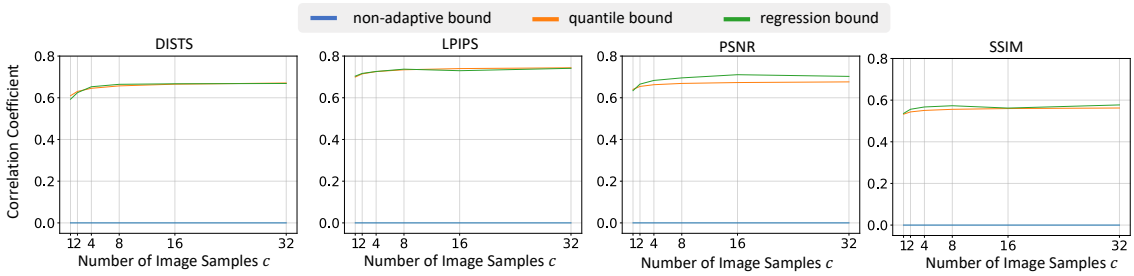


Figure D.2: Mean Pearson correlation coefficient between each conformal bound and the true FRIQ versus the number of posterior samples c for FFHQ denoising.

better average acceleration rates than the non-adaptive bound in the multi-round measurement experiment of Sec. 5.3.2.

D.3 Additional MRI experiments

Effect of number of posterior samples c in conformal bound: For the case of FFHQ denoising, Sec. 5.3.1 demonstrated the number of posterior samples c has a limited effect on the conformal bounds for the FFHQ experiments. We now investigate whether the same occurs with MRI. Figure D.3 plots the percent improvement in MCB as c increases relative to the MCB for $c = 1$. From the figure, we see less than a

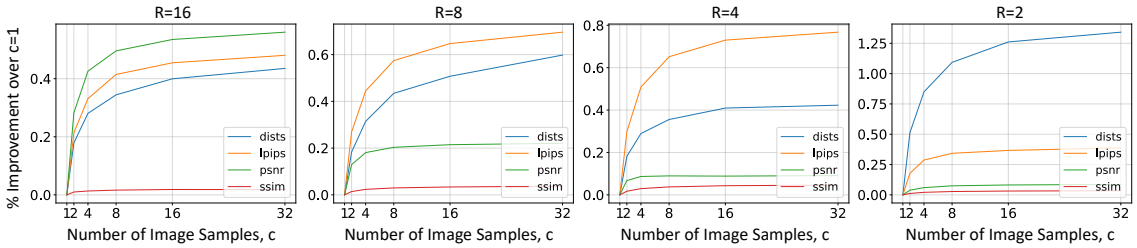


Figure D.3: Percent improvement in MCB versus number of samples c used in the quantile bound for the accelerated MRI experiments.

1.5% improvement over $c = 1$ for any metric, suggesting that the quantile method is indeed robust to the choice of c for both experiments.

Multi-round measurement samples: In Fig. D.4, we show the zero-filled measurement, recovered image, and absolute-error map at each acceleration rate. The conformal bound is imposed on the reconstructions for the case when $\alpha = 0.05$, $\tau = 0.16$, and $c = 32$. Following the multi-round measurement protocol described in Sec. 5.3.2, the reconstruction at $R = 8$ (marked in red) would be deemed sufficient ($\beta_i < \tau$), and the measurement collection would end.

D.4 Empirical investigation of distribution shift

As previously mentioned, a general limitation of CP methods like [6] is the requirement of exchangeability, which in our case applies to the pairs $\{(\widehat{Z}_i, Z_i)\}_{i=0}^n$. This requirement may be violated when there is a distributional shift between the test data (x_0, y_0) and the calibration data $\{(x_i, y_i)\}_{i=1}^n$, which can then cause a distributional shift between the corresponding FRIQ quantities $(\widehat{z}_0, z_0)$ and $\{(\widehat{z}_i, z_i)\}_{i=1}^n$.

In the case of MRI, such distributional shifts may arise for various reasons, some of which would be easy to prevent while others would be more difficult. For example,

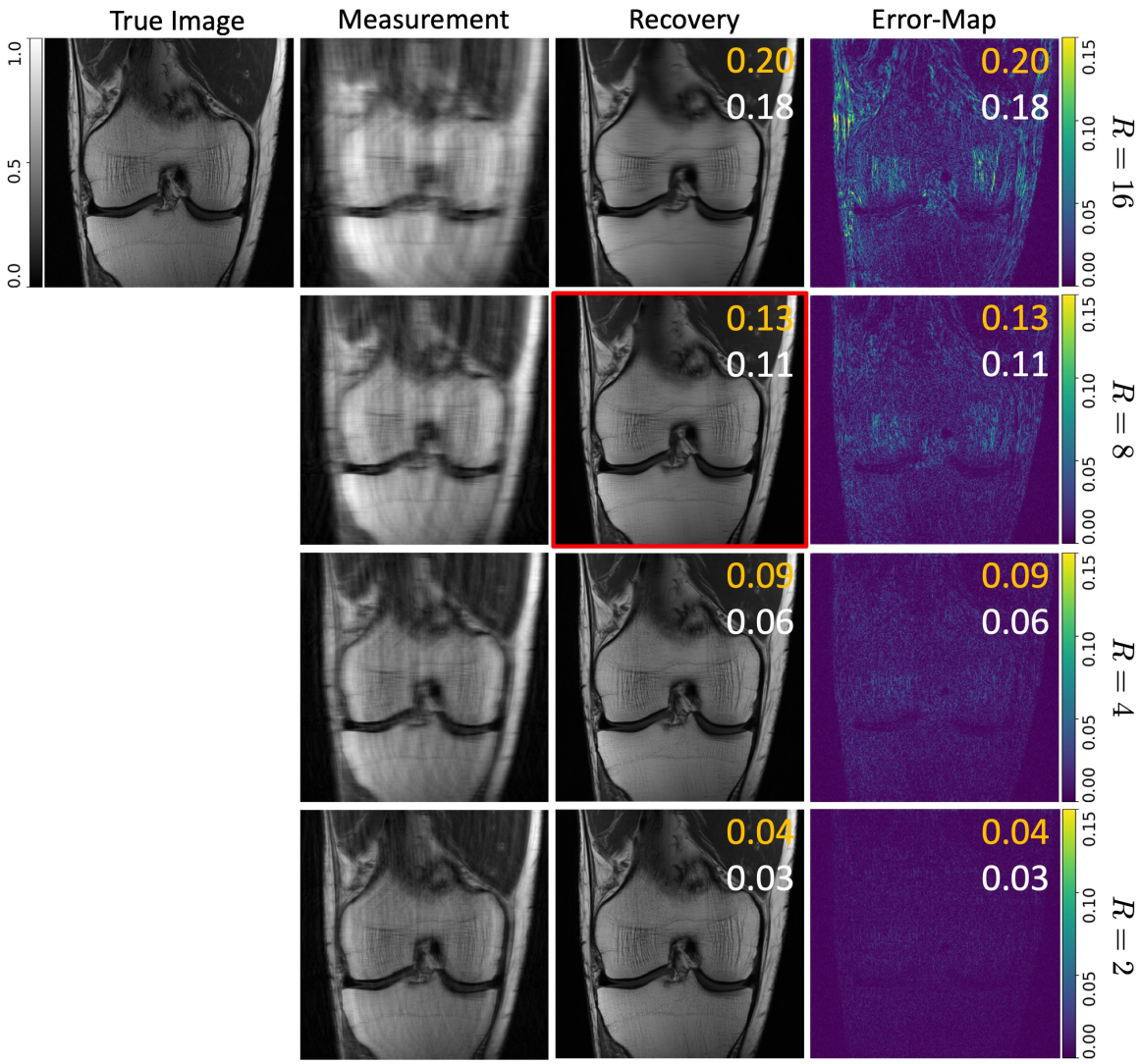


Figure D.4: Qualitative example of the multi-round MRI experiment with DISTS at $\alpha = 0.05$, $\tau = 0.16$, and $c = 32$. The measurement, recovery, and absolute error are shown for all accelerations. The quantile bound (orange) and true DISTS (white) are imposed on the reconstructions. The red box indicates the accepted reconstruction where the bound first falls below the threshold τ.

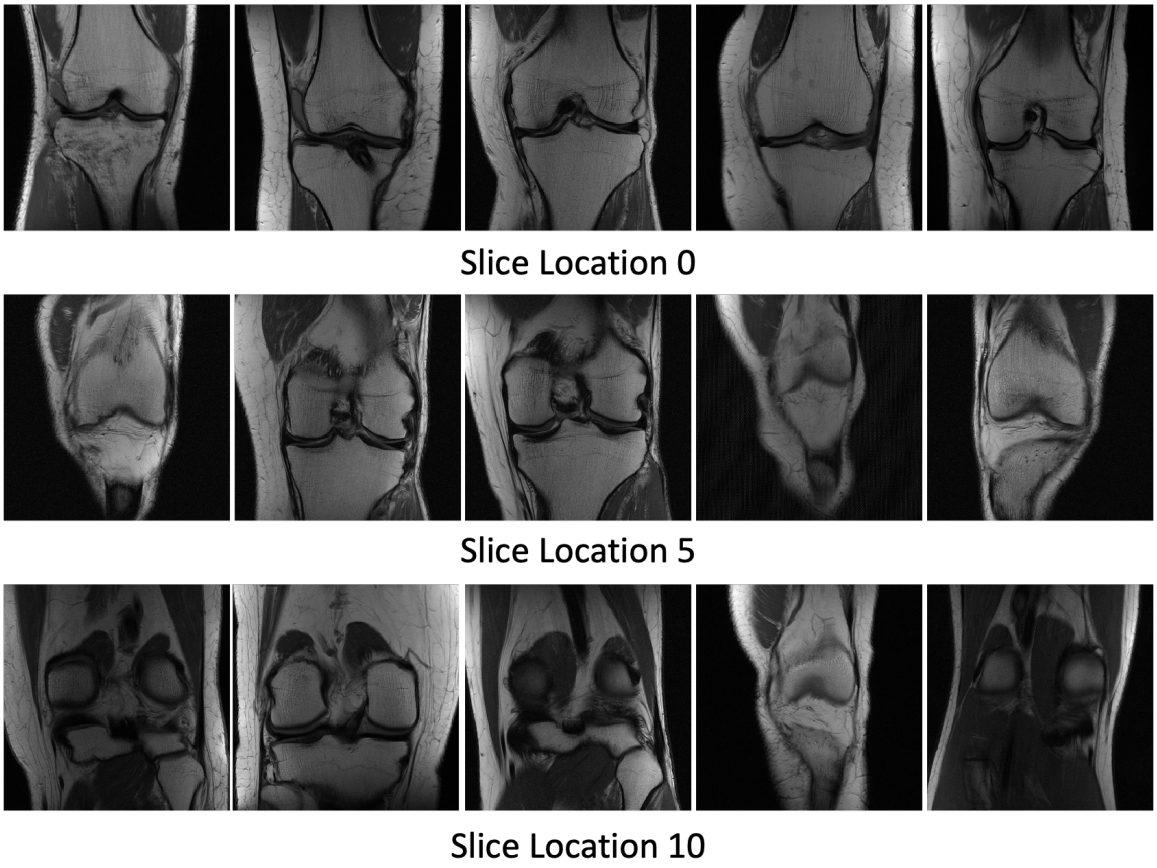


Figure D.5: Qualitative examples of images from different slice locations. Slice location 0 indicates the center slice of a volume while larger slice locations are further towards the edges of a volume.

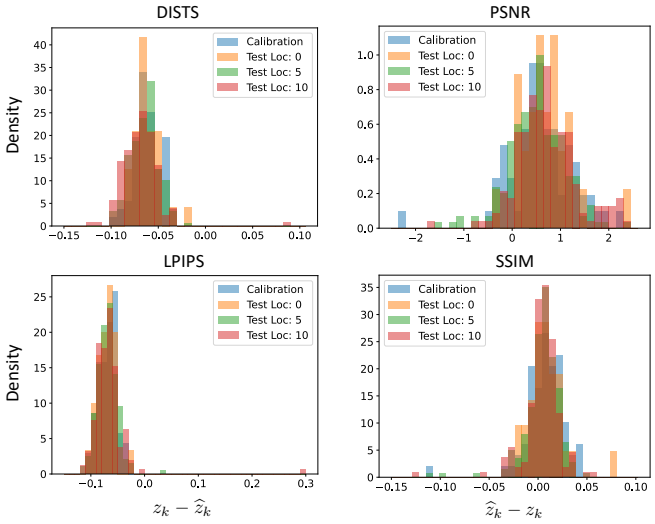


Figure D.6: Histograms of the difference between the true FRIQ z_k and the FRIQ estimate $\widehat{z}_k$ for test indices k in the test fold $\mathcal{I}_{\text{test}}[t]$ of a single trial. Histograms are shown for test slice locations $l = 0, 5, 10$. Note the increasing shift in distribution from the calibration set as l increases.

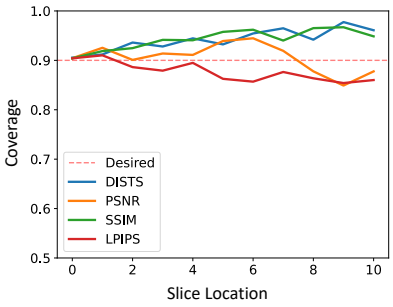


Figure D.7: The average empirical coverage across $T = 10000$ trials for test sets at different slice locations. All trials are calibrated with images from slice location 0 with $\alpha = 0.1$, $R = 8$, and $c = 32$.

if the CP method was calibrated on knee images, one would not want to immediately test on brain images, but instead recalibrate a CP method on brain images. Likewise, if the CP method was calibrated with data from one manufacturer and/or strength of scanner, then it would be best to test on data from the same manufacturer and/or strength of scanner. Still, due to limited calibration data, situations may arise where a distribution shift is inevitable. Thus, we perform a study to analyze the sensitivity of our proposed method to distribution shifts.

For this study, we use the validation fold of the non-fat-suppressed multi-coil fastMRI knee dataset [125], which contains 100 3D volumes. A volume contains all the images collected for a single patient, with each image showing a different slice of

the knee (from front to back). To induce a realistic yet controllable distribution shift, we choose calibration images from only the center slices of these volumes, and refer to the center slices as "location $l = 0$." We then create one test set with images from slice locations $l = 0$, another test set with images from slice location $l = 1$, and so on, until slice location $l = 10$ (which typically corresponds to an edge slice). Example images from various slice locations are shown in Fig. D.5.

We first evaluate the coverage of the quantile bound using $T = 10\,000$ Monte Carlo trials, error-rate $\alpha = 0.1$, acceleration $R = 8$, an E2E-VarNet [107] sample for $\widehat{x}_i$, and $c = 32$ posterior samples for w_i. For each trial $t \in \{1, \ldots, T\}$, we construct the calibration set by randomly sampling 70 of the 100 center slices. For the same t, we form the test data at location $l = 0$ using the remaining 30 slices, and we form the test data at locations $l > 0$ by randomly sampling 30 of the 200 available slices. Figure D.7 plots the mean empirical coverage over the T trials as a function of test slice location l. As expected, the desired $1 - \alpha$ coverage is met when $l = 0$. However, the behavior of the empirical coverage for $l > 0$ varies depending on the metrics. The coverage for LPIPS tends to decrease slightly as the slice location l increases, and the coverage for PSNR only falls below $1 - \alpha$ after $l = 7$. Surprisingly, for the DISTS and SSIM metrics, the coverage remains well above $1 - \alpha$ for all slice locations, suggesting the bounds remain valid, but become slightly over-conservative for $l > 0$. Overall, the results demonstrate our bounds are quite robust to small distributional shifts with only a minor loss in coverage for certain metrics.

To visualize the distribution shift versus test location l, we consider the difference between the true FRIQ z_k and the FRIQ estimate $\widehat{z}_k$ for each test index $k \in \mathcal{I}_{\mathsf{test}}[t]$ in a single trial t. This difference is $z_k - \widehat{z}_k$ for LP metrics and $\widehat{z}_k - z_k$ for HP metrics.

Figure D.6 shows the histogram of this difference for test locations $l \in \{0, 5, 10\}$. As expected, these histograms deviate more as the test location l increases, although the amount of deviation depends on the FRIQ metric. For PSNR, we see the histogram shifting slightly to the right, while for SSIM, the histogram starts to shrink in width.

Figure D.7 suggests that one could select a more conservative α to ensure sufficiently high coverage under small distributional shifts, but at the cost of more conservative bounds. In fact, this is largely the mechanism behind distributionally robust CP extensions like Cauchois et al. [27]. We leave such generalizations to future work.

D.5 Posterior averaging for image estimates

As described above in Secs. 5.2.2–5.2.4, a posterior-sampling-based image recovery method allows one to construct adaptive bounds using image samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^{c}$. But, as we now discuss, a posterior-sampling-based image recovery method also provides flexibility in how $\widehat{x}_i$ itself is constructed.

For example, when one is interested in constructing $\widehat{x}_i$ with high PSNR, or equivalently low MSE, it makes sense to set $\widehat{x}_i$ as the minimum MSE (MMSE) or conditional-mean estimate $\mathrm{E}\{X_i | Y_i = y_i\}$. This can be approximated by the empirical mean of p posterior samples, i.e.,

$$\widehat{x}_i = \frac{1}{p} \sum_{j=c+1}^{c+p} \widetilde{x}_i^{(j)}, \tag{D.1}$$

with large p. The indices on j in (D.1) are chosen to avoid the samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^{c}$ used for the adaptive bounds. However, because the MMSE estimate can look unrealistically smooth, smaller values of p are appropriate when constructing an $\widehat{x}_i$ with good SSIM, DISTS, or LPIPS performance. For example, Bendel et al. [20] found that, for multi-coil brain MRI at acceleration $R = 8$ with a particular posterior sampler, the

Table D.4: Mean empirical coverage for the quantile method with $\alpha = 0.05$, $c = 32$, and $T = 10\,000$ on the $R = 8$ accelerated MRI task ($\pm$ standard error). All coverages are above the expected coverage of $1 - \alpha = 0.95$

p	DISTS	LPIPS	PSNR	SSIM
1	0.9503 ± 0.0001	0.9503 ± 0.0001	0.9505 ± 0.0001	0.9504 ± 0.0001
2	0.9505 ± 0.0001	0.9503 ± 0.0001	0.9504 ± 0.0001	0.9505 ± 0.0001
4	0.9503 ± 0.0001	0.9503 ± 0.0001	0.9505 ± 0.0001	0.9504 ± 0.0001
8	0.9505 ± 0.0001	0.9504 ± 0.0001	0.9504 ± 0.0001	0.9505 ± 0.0001
16	0.9505 ± 0.0001	0.9502 ± 0.0001	0.9504 ± 0.0001	0.9504 ± 0.0001
32	0.9504 ± 0.0001	0.9506 ± 0.0001	0.9504 ± 0.0001	0.9505 ± 0.0001

best choice of p is 8 for SSIM and 2 for both DISTS and LPIPS. This is can explained by the perception-distortion tradeoff [22], which says that, as p increases and the MSE distortion decreases, the perceptual quality must also decrease. In the end, each FRIQ metric prefers a particular tradeoff between perceptual quality and distortion.

To assess the flexibility this affords, we now use p additional posterior samples from the CNF in Ch. 3 to construct the image estimate $\widehat{x}$ and utilize the same Monte Carlo validation as before to compute the mean empirical coverage and mean conformal bounds (MCB) over $T = 10000$ trials.

Empirical Coverage: Table D.4 reports the average empirical coverage for the quantile bounds with different values of p. As with all previous experiments, the coverage is above and very close to the desired $1 - \alpha$ value for all values of p.

Effect of acceleration rate R and choice of recovery method: Figure D.8 plots the quantile MCB with $c = 32$ versus the acceleration rate R for different image estimates $\widehat{x}_i$. The image estimate $\widehat{x}$ is computed using either the E2E-VarNet point estimate or a p-sample average from the CNF with different values of p. In all cases, MCB improves as the acceleration R decreases, as expected. However, as discussed

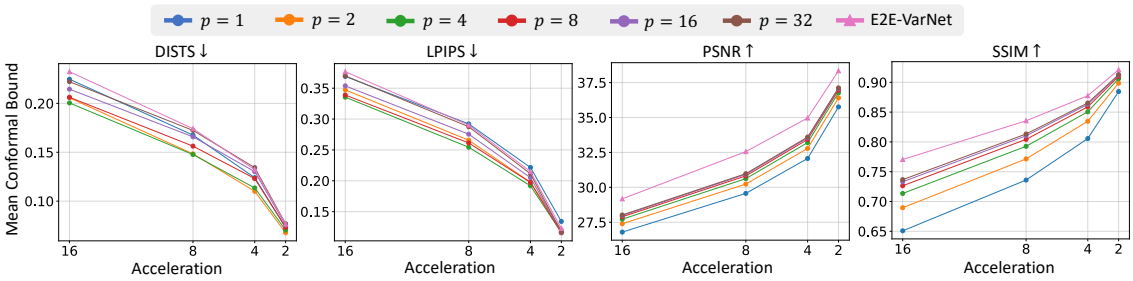


Figure D.8: Mean conformal bound versus acceleration R for accelerated MRI. Results shown for the quantile bound with $\widehat{x}$ computed from the E2E-VarNet point estimate (shown in pink) or the p-sample average from CNF posteriors. Various p shown.

in Sec. D.5, each metric benefits from a different choice of p. DISTS and LPIPS prefer $p \in \{2, 4\}$ while PSNR and SSIM prefer $p = 32$. The figure also shows that the MCB for the p-optimized CNF-based method is better than the MCB for the E2E-VarNet-based method with both DISTS and LPIPS but not with PSNR and SSIM. Thus, the recovery method that yields the tightest bounds may depend on the metric of interest.

Multi-round MRI: Since Figure D.8 reveals the tightest bounds on DISTS are obtained when the CNF posterior average with $p = 4$ is used for the image estimate $\widehat{x}$, we repeat the multi-round experiment from Sec. 5.3.2 to see how this translates practically. Again, we set $\alpha = 0.05$ and $c = 32$. The threshold is set at $\tau = 0.11$ where the non-adaptive approach requires $R = 2$ for acceptance. In Figure D.9, we plot the distribution of slices that were collected at each acceleration rate. Here, the non-adaptive approach always accepts slices at $R = 2$ while the quantile bound accepts nearly 50% of the slices at $R = 4$. Table D.5 shows that this equates to an average accepted acceleration rate of 2.596 with an empirical coverage of 0.9434 at acceptance. This demonstrates that the multi-round performance difference between

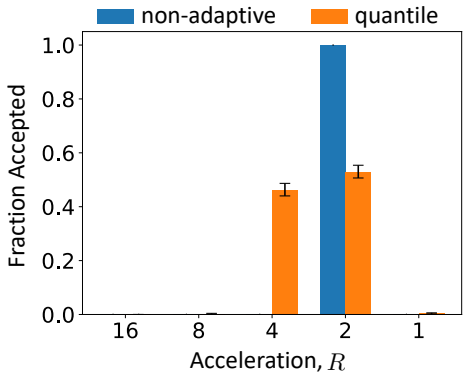


Figure D.9: Fraction of accepted slices versus final acceleration rate for multi-round MRI using DISTS. Both methods use a p-posterior average for the recovery with $p = 4$, $c = 32$, $\alpha = 0.05$, and $\tau = 0.11$. Error bars show standard deviation.

Table D.5: Average results for a multi-round MRI simulation where measurement collection stop once bounds are below a user-set threshold τ. Results shown for $T = 10\,000$ trials using the DISTS metric with $\alpha = 0.05$, $\tau = 0.11$, $p = 4$, and $c = 32$ ($\pm$ standard error).

Method	Average Acceleration	Acceptance Empirical Coverage
Non-adaptive	2.000 ± 0.000	0.9504 ± 0.0001
Quantile	2.596 ± 0.001	0.9434 ± 0.0001

bound types varies with the recovery model, and thus, the modularity of our proposed method allows improvements along multiple dimensions (i.e. the recovery model, the posterior sampler, the conformal bound) all of which can be easily swapped to suit the particular problem.

Distribution Shift: We repeat the distribution shift analysis from Appendix D.4 using a single CNF posterior sample as the image estimate $\widehat{x}$, i.e. $p = 1$, with $\alpha = 0.1$, $c = 32$, and $R = 8$. As before, we can visualize the distribution shift by looking at the histograms of the difference between the true FRIQ z_k and FRIQ estimate $\widehat{z}_k$ for each test index $k \in \mathcal{I}_{\text{test}}[t]$ as the test location l increases. Figure D.10 shows the histograms for test locations $l \in \{0, 5, 10\}$. We see more distinct distributional shifts compared to Appendix D.4 as the histogram for PSNR noticeably shifts to the right and widens while the histogram for LPIPS becomes bimodal as l increases. Not

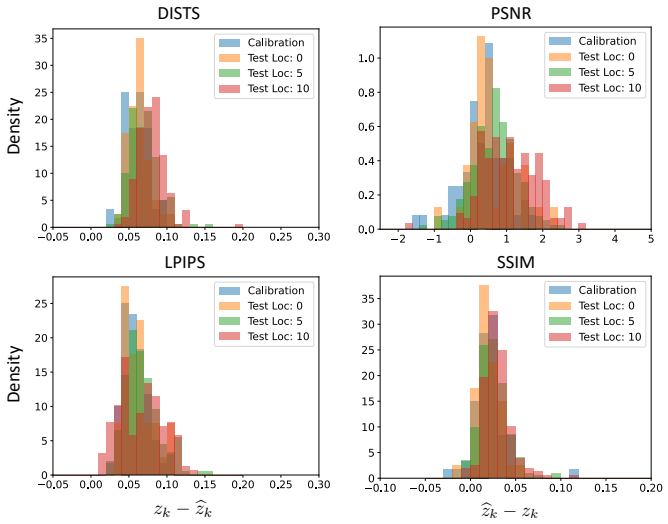


Figure D.10: Histograms of the difference between the true FRIQ z_k and the FRIQ estimate $\widehat{z}_k$ for test samples k in the test fold of a single trial. Histograms are shown for test slice locations $l = 0, 5, 10$. Note the increasing shift in distribution from the calibration set as l increases.

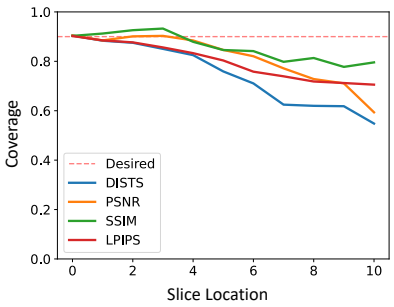


Figure D.11: The average empirical coverage across $T = 10000$ trials for test sets at different slice locations. All trials are calibrated with images from slice location 0 with $\alpha = 0.1$, $R = 8$, $p = 1$, and $c = 32$.

surprisingly, the more dramatic shifts lead to a decrease in coverage for all metrics in Fig. D.11 as l increases. We do note, however, that both SSIM and PSNR retain a coverage at or above $1 - \alpha$ until $l = 4$, demonstrating a level of robustness.

D.6 Average fastMRI reconstruction performance

To get a sense of the average reconstruction performance for the accelerated MRI task, we report the average metrics for both the E2E-VarNet and CNF on the non-fat-suppressed subset of the fastMRI knee validation set. Results for acceleration rates $R = 16$, 8, 4, and 2 are shown in Tables D.6, D.7, D.8, D.9, respectively. The E2E-VarNet outperforms the CNF in PSNR and SSIM across all accelerations. The

Table D.6: Average reconstruction performance on the fastMRI [125] knee validation set for $R = 16$ ($\pm$ standard error)

Network	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
E2E-VarNet	0.209 ± 0.001	0.354 ± 0.001	$\mathbf{30.301 \pm 0.043}$	$\mathbf{0.807 \pm 0.001}$
CNF ($p = 1$)	0.183 ± 0.001	0.312 ± 0.001	28.244 ± 0.039	0.688 ± 0.002
CNF ($p = 2$)	0.167 ± 0.001	0.292 ± 0.001	29.091 ± 0.039	0.730 ± 0.001
CNF ($p = 4$)	$\mathbf{0.165 \pm 0.001}$	$\mathbf{0.287 \pm 0.001}$	29.588 ± 0.039	0.755 ± 0.001
CNF ($p = 8$)	0.173 ± 0.001	0.296 ± 0.001	29.862 ± 0.039	0.770 ± 0.001
CNF ($p = 16$)	0.184 ± 0.001	0.314 ± 0.001	30.006 ± 0.039	0.777 ± 0.001
CNF ($p = 32$)	0.193 ± 0.001	0.333 ± 0.001	30.080 ± 0.039	0.781 ± 0.001

Table D.7: Average reconstruction performance on the fastMRI [125] knee validation set for $R = 8$ ($\pm$ standard error)

Network	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
E2E-VarNet	0.151 ± 0.001	0.262 ± 0.001	$\mathbf{33.459 \pm 0.047}$	$\mathbf{0.864 \pm 0.001}$
CNF ($p = 1$)	0.136 ± 0.000	0.248 ± 0.001	30.796 ± 0.044	0.761 ± 0.002
CNF ($p = 2$)	$\mathbf{0.118 \pm 0.000}$	0.225 ± 0.001	31.754 ± 0.044	0.799 ± 0.001
CNF ($p = 4$)	0.119 ± 0.000	$\mathbf{0.219 \pm 0.001}$	32.329 ± 0.043	0.821 ± 0.001
CNF ($p = 8$)	0.128 ± 0.001	0.228 ± 0.001	32.650 ± 0.043	0.834 ± 0.001
CNF ($p = 16$)	0.138 ± 0.001	0.243 ± 0.001	32.819 ± 0.043	0.840 ± 0.001
CNF ($p = 32$)	0.145 ± 0.001	0.255 ± 0.001	32.907 ± 0.043	0.843 ± 0.001

CNF, on the other hand, provides lower DISTS and LPIPS values in all cases other than for LPIPS at acceleration $R = 2$.

Table D.8: Average reconstruction performance on the fastMRI [125] knee validation set for $R = 4$ ($\pm$ standard error)

Network	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
E2E-VarNet	0.110 ± 0.001	0.181 ± 0.001	$\mathbf{36.030 \pm 0.053}$	$\mathbf{0.905 \pm 0.001}$
CNF ($p = 1$)	0.100 ± 0.000	0.191 ± 0.001	33.090 ± 0.048	0.826 ± 0.001
CNF ($p = 2$)	$\mathbf{0.087 \pm 0.000}$	0.170 ± 0.001	34.073 ± 0.048	0.856 ± 0.001
CNF ($p = 4$)	0.090 ± 0.000	$\mathbf{0.166 \pm 0.001}$	34.666 ± 0.048	0.873 ± 0.001
CNF ($p = 8$)	0.099 ± 0.000	0.171 ± 0.001	34.998 ± 0.047	0.882 ± 0.001
CNF ($p = 16$)	0.106 ± 0.001	0.178 ± 0.001	35.174 ± 0.047	0.887 ± 0.001
CNF ($p = 32$)	0.110 ± 0.001	0.184 ± 0.001	35.265 ± 0.047	0.889 ± 0.001

Table D.9: Average reconstruction performance on the fastMRI [125] knee validation set for $R = 2$ ($\pm$ standard error)

Network	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
E2E-VarNet	0.059 ± 0.000	$\mathbf{0.094 \pm 0.001}$	$\mathbf{39.692 \pm 0.060}$	$\mathbf{0.947 \pm 0.001}$
CNF ($p = 1$)	0.059 ± 0.000	0.118 ± 0.000	36.810 ± 0.054	0.907 ± 0.001
CNF ($p = 2$)	$\mathbf{0.054 \pm 0.000}$	0.105 ± 0.000	37.667 ± 0.054	0.923 ± 0.001
CNF ($p = 4$)	0.055 ± 0.000	0.100 ± 0.000	38.171 ± 0.054	0.931 ± 0.001
CNF ($p = 8$)	0.058 ± 0.000	0.099 ± 0.000	38.448 ± 0.054	0.935 ± 0.001
CNF ($p = 16$)	0.060 ± 0.000	0.099 ± 0.000	38.593 ± 0.054	0.937 ± 0.001
CNF ($p = 32$)	0.061 ± 0.000	0.099 ± 0.000	38.668 ± 0.054	0.939 ± 0.001

Bibliography

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Info. Fusion, 76:243–297, 2021.

[2] Jonas Adler and Ozan Öktem. Deep Bayesian inversion. arXiv:1811.05910, 2018.

[3] R. Ahmad, C. A. Bouman, G. T. Buzzard, S. Chan, S. Liu, E. T. Reehorst, and P. Schniter. Plug and play methods for magnetic resonance imaging. IEEE Signal Process. Mag., 37(1):105–116, March 2020.

[4] Martin S. Andersen, Joachim Dahl, and Lieven Vandenberghe. CVXOPT, 2023.

[5] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. Found. Trends Mach. Learn., 16(4):494–591, 2023.

[6] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. arXiv:2208.02814, 2022.

[7] Anastasios N. Angelopoulos, Amit P. Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In Proc. Intl. Conf. Mach. Learn., 2022.

[8] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In Proc. Intl. Conf. Learn. Rep., 2020.

[9] Lynton Ardizzone, Jakob Kruse, Carsten Lüth, Niels Bracher, Carsten Rother, and Ullrich Köthe. Conditional invertible neural networks for diverse image-to-image translation. arXiv:2105.02104, 2021.

[10] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. In Proc. Intl. Conf. Learn. Rep., 2019.

[11] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. arXiv:1907.02392, 2019.

[12] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. Acta Numerica, 28:1–174, June 2019.

[13] Christopher R. S. Banerji, Tapabrata Chakraborti, Chris Harbron, and Ben D. MacArthur. Clinical AI tools must convey predictive uncertainty for each individual patient. Nature Medicine, 29(12):2996–2998, 2023.

[14] Riccardo Barbano, Chen Zhang, Simon Arridge, and Bangti Jin. Quantifying model uncertainty in inverse problems via Bayesian deep gradient descent. In Proc. IEEE Intl. Conf. Pattern Recog., pages 1392–1399, 2021.

[15] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. Ann. Statist., 51(2):816–845, 2023.

[16] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. J. ACM, 68(6), 2021.

[17] Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. IEEE Trans. Pattern Anal. Mach. Intell., 46:3321–3333, 2023.

[18] Chinmay Belthangady and Loic A Royer. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. Nature Methods, 16(12):1215–1225, 2019.

[19] Matthew Bendel, Rizwan Ahmad, and Philip Schniter. A regularized conditional GAN for posterior sampling in inverse problems. arXiv:2210.13389, 2022.

[20] Matthew Bendel, Rizwan Ahmad, and Philip Schniter. A regularized conditional GAN for posterior sampling in inverse problems. In Proc. Neural Info. Process. Syst. Conf., 2023.

[21] Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. IEEE Trans. Med. Imag., 40(11):3249–3260, 2021.

[22] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 6228–6237, 2018.

[23] Tom Boeken, Jean Feydy, Augustin Lecler, Philippe Soyer, Antoine Feydy, Maxime Barat, and Loïc Duron. Artificial intelligence in diagnostic and interventional radiology: Where are we now? Diagnostic and Interventional Imaging, 104(1):1–5, 2023.

[24] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In Proc. Intl. Conf. Mach. Learn., pages 537–546, 2017.

[25] Nicki Skafte Detlefsenand Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. Torchmetrics, 2022.

[26] Stéphane Caron, Daniel Arnström, Suraj Bonagiri, Antoine Dechaume, Nikolai Flowers, Adam Heins, Takuma Ishikawa, Dustin Kenefake, Giacomo Mazzamuto, Donato Meoli, Brendan O'Donoghue, Adam A. Oppenheimer, Abhishek Pandala, Juan José Quiroz Omaña, Nikitas Rontsis, Paarth Shah, Samuel St-Jean, Nicola Vitucci, Soeren Wolfers, and Fengyu Yang. qpsolvers: Quadratic programming solvers in python, 2024.

[27] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. J. Am. Statist. Assoc., pages 1–66, 2024.

[28] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In Proc. Intl. Conf. Learn. Rep., 2019.

[29] Dongdong Chen and Mike E Davies. Deep decomposition learning for inverse imaging problems. In Proc. Europ. Conf. Comp. Vision, pages 510–526, 2020.

[30] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proc. Intl. Conf. Knowl. Disc. Data Mining, pages 785–794, 2016.

[31] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Proc. Intl. Conf. Mach. Learn., pages 1597–1607, 2020.

[32] Linda C Chu, Anima Anandkumar, Hoo Chang Shin, and Elliot K Fishman. The potential dangers of artificial intelligence for radiology and radiologists. J. Amer. Col. Radiology, 17(10):1309–1311, 2020.

[33] Hyungjin Chung, Jeongsol Kim, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In Proc. Intl. Conf. Learn. Rep., 2023.

[34] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. Med. Image Analysis, 80:102479, 2022.

[35] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In Proc. Intl. Conf. Med. Image Comput. Comput. Assist. Intervent., pages 529–536, 2018.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 248–255, 2009.

[37] Alexander Denker, Maximilian Schmidt, Johannes Leuschner, and Peter Maass. Conditional invertible neural networks for medical imaging. J. Imaging, 7(11):243, 2021.

[38] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. DISTS, 2020.

[39] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE Trans. Pattern Anal. Mach. Intell., 44(5):2567–2581, 2020.

[40] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In Proc. Intl. Conf. Learn. Rep. Workshops, 2015.

[41] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In Proc. Intl. Conf. Learn. Rep., 2017.

[42] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. SIAM J. Imag. Sci., 11(1):473–506, 2018.

[43] Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep MRI reconstruction. IEEE Trans. Med. Imag., 40(1):239–250, January 2021.

[44] Canberk Ekmekci and Mujdat Cetin. Uncertainty quantification for deep unrolling-based computational imaging. IEEE Trans. Comput. Imag., 8:1195–1209, 2022.

[45] Canberk Ekmekci and Mujdat Cetin. Quantifying generative model uncertainty in posterior sampling methods for computational imaging. In Proc. Neural Info. Process. Syst. Workshop, 2023.

[46] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

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

[48] William Falcon et al. Pytorch lightning, 2019.

[49] Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. J. Mach. Learn. Res., 24(24):1–48, 2023.

[50] Bert E Fristedt and Lawrence F Gray. A Modern Approach to Probability Theory. Springer, 2013.

[51] Nina M Gottschling, Vegard Antun, Anders C Hansen, and Ben Adcock. The troublesome kernel—On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems. arXiv:2001.01258, 2023.

[52] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn. Reson. Med., 47(6):1202–1210, 2002.

[53] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In Proc. Intl. Conf. Mach. Learn., volume 70, pages 1321–1330, 2017.

[54] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. Magn. Reson. Med., 79(6):3055–3071, 2018.

[55] Kerstin Hammernik, Thomas Küstner, Burhaneddin Yaman, Zhengnan Huang, Daniel Rueckert, Florian Knoll, and Mehmet Akçakaya. Physics-driven deep learning for computational magnetic resonance imaging: Combining physics and machine learning for improved medical imaging. IEEE Signal Process. Mag., 40(1):98–114, 2023.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 770–778, 2016.

[57] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Proc. Neural Info. Process. Syst. Conf., volume 30, 2017.

[58] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Proc. Neural Info. Process. Syst. Conf., volume 33, pages 6840–6851, 2020.

[59] David P Hoffman, Isaac Slavitt, and Casey A Fitzpatrick. The promise and peril of deep learning in microscopy. Nature Methods, 18(2):131–132, 2021.

[60] Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. arXiv.2211.09795, 2022.

[61] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 1125–1134, 2017.

[62] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alex Dimakis, and Jonathan Tamir. Robust compressed sensing MRI with deep generative priors. In Proc. Neural Info. Process. Syst. Conf., 2021.

[63] Mihir Joshi, Aaron Pruitt, Chong Chen, Yingmin Liu, and Rizwan Ahmad. Technical report (v1.0)–pseudo-random cartesian sampling for dynamic MRI. arXiv:2206.03630, 2022.

[64] Zahra Kadkhodaie and Eero P Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. arXiv:2007.13640, 2020.

[65] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In Proc. Intl. Conf. Learn. Rep., 2018.

[66] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 4396–4405, 2019.

[67] Segrey Kastryulin, Jamil Zakirov, Nicola Pezzotti, and Dmitry V. Dylov. Image quality assessment for magnetic resonance imaging. arXiv:2203.07809, 2022.

[68] Sergey Kastryulin, Jamil Zakirov, Nicola Pezzotti, and Dmitry V Dylov. Image quality assessment for magnetic resonance imaging. IEEE Access, 11:14154–14168, 2023.

[69] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In Proc. Neural Info. Process. Syst. Conf., 2022.

[70] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. Downloaded from `https://github.com/bahjat-kawar/ddrm`, May 2022.

[71] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In Proc. Neural Info. Process. Syst. Conf., 2017.

[72] Younggeun Kim and Donghee Son. Noise conditional flow model for learning the super-resolution space. arXiv:1606.02838, 2021.

[73] Diederik Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In Proc. Neural Info. Process. Syst. Conf., pages 10236–10245, 2018.

[74] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proc. Intl. Conf. Learn. Rep., 2015.

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

[76] Roger Koenker and Gilbert Bassett. Regression quantiles. Econometrica, 46(1), 1978.

[77] Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In Proc. Intl. Conf. Learn. Rep., 2023.

[78] Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play priors: When Langevin meets Tweedie. SIAM J. Imag. Sci., 15(2):701–737, 2022.

[79] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. J. Am. Statist. Assoc., 2018.

[80] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. J. Roy. Statist. Soc., 76, 2014.

[81] Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. J. Vis. Commun. Image Rep., 22(4):297–312, 2011.

[82] Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. J. Cogn. Neurosci., 33(10):2017–2031, 2021.

[83] Charles Lu, Anastasios N. Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. Proc. Intl. Conf. Med. Image Comput. Comput. Assist. Intervent., 2022.

[84] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Kang-wook Kim, Younggeun Kim, Jae-young Lee, Zechao Li, Jinshan Pan, Dongseok Shim, Ki-Ung Song, et al. NTIRE 2022 challenge on learning the super-resolution space. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 786–797, 2022.

[85] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. SRFlow: Learning the super-resolution space with normalizing flow. In Proc. Europ. Conf. Comp. Vision, 2020.

[86] Michael Lustig, David Donoho, and John M Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. Magn. Reson. Med., 58(6):1182–1195, 2007.

[87] Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. Pattern Recognition, 120:108101, 2021.

[88] Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In Proc. Conformal and Probabilistic Prediction and Applications, pages 294–306. PMLR, 2022.

[89] Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. IEEE Trans. Med. Imag., 40(9):2306–2317, 2021.

[90] Dominik Narnhofer, Alexander Effland, Erich Kobler, Kerstin Hammernik, Florian Knoll, and Thomas Pock. Bayesian uncertainty estimation of learned variational MRI reconstruction. IEEE Trans. Med. Imag., 41(2):279–291, 2022.

[91] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Proc. Europ. Conf. Mach. Learn., pages 345–356, 2002.

[92] George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. J. Mach. Learn. Res., 22(57):1–64, 2021.

[93] Ji Won Park, Robert Tibshirani, and Kyunghyun Cho. Semiparametric conformal prediction. arXiv:2411.02114, 2024.

[94] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In Proc. Neural Info. Process. Syst. Conf., pages 8024–8035, 2019.

[95] Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. Magn. Reson. Med., 42(5):952–962, 1999.

[96] Peter B Roemer, William A Edelstein, Cecil E Hayes, Steven P Souza, and Otward M Mueller. The NMR phased array. Magn. Reson. Med., 16(2):192–225, 1990.

[97] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In Proc. Neural Info. Process. Syst. Conf., pages 3543–3553, 2019.

[98] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In Proc. Intl. Conf. Med. Image Comput. Comput. Assist. Intervent., pages 234–241, 2015.

[99] Max Sampson and Kung-Sik Chan. Conformal multi-target hyperrectangles. Stat. Anal. and Data Mining, 17(5), 2024.

[100] Thomas Sanchez, Igor Krawczuk, Zhaodong Sun, and Volkan Cevher. Uncertainty-driven adaptive sampling via GANs. In Proc. Neural Info. Process. Syst. Workshop, 2020.

[101] Swami Sankaranarayanan, Anastasios N. Angelopoulos, Stephen Bates, Yaniv Romano, and Phillip Isola. Semantic uncertainty intervals for disentangled latent spaces. In Proc. Neural Info. Process. Syst. Conf., 2022.

[102] 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 Proc. Machine Learning for Medical Image Reconstruction Workshop, pages 64–71, 2018.

[103] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[104] Michael Soloveitchik, Tzvi Diskin, Efrat Morin, and Ami Wiesel. Conditional Frechet inception distance. arXiv:2103.11521, 2021.

[105] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised MAP inference for image super-resolution. In Proc. Intl. Conf. Learn. Rep., 2017.

[106] Ki-Ung Song, Dongseok Shim, Kang-wook Kim, Jae-young Lee, and Younggeun Kim. FS-NCSR: Increasing diversity of the super-resolution space via frequency separation and noise-conditioned normalizing flow. In Proc. IEEE Conf. Comp. Vision Pattern Recog. Workshop, pages 968–977, June 2022.

[107] 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 Proc. Intl. Conf. Med. Image Comput. Comput. Assist. Intervent., pages 64–73, 2020.

[108] 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. `https://github.com/facebookresearch/fastMRI`, 2020.

[109] Rhea Sanjay Sukthanker, Zhiwu Huang, Suryansh Kumar, Radu Timofte, and Luc Van Gool. Generative flows with invertible attentions. In Proc. IEEE Conf. Comp. Vision Pattern Recog., 2022.

[110] Michael Tang and Audrey Repetti. A data-driven approach for Bayesian uncertainty quantification in imaging. arXiv, 2023.

[111] Jacopo Teneggi, Matthew Tivnan, J. Webster Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control, 2023.

[112] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In Proc. Neural Info. Process. Syst. Conf., volume 32, 2019.

[113] Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination index: An image quality metric for generative reconstruction models. In Proc. Intl. Conf. Med. Image Comput. Comput. Assist. Intervent., pages 449–458, 2024.

[114] Francesco Tonolini, Jack Radford, Alex Turpin, Daniele Faccio, and Roderick Murray-Smith. Variational inference for computational imaging inverse problems. J. Mach. Learn. Res., 21(179):1–46, 2020.

[115] Martin Uecker, Peng Lai, Mark J Murphy, Patrick Virtue, Michael Elad, John M Pauly, Shreyas S Vasanawala, and Michael Lustig. ESPIRiT–an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA. Magn. Reson. Med., 71(3):990–1001, 2014.

[116] Vladimir Vovk. Conditional validity of inductive conformal predictors. In Asian Conf. Mach. Learn., pages 475–490, 2012.

[117] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic Learning in a Random World. Springer, New York, 2005.

[118] Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In Proc. Intl. Conf. Mach. Learn., pages 444–453, 1999.

[119] Zhou Wang. Applications of objective image quality assessment methods. IEEE Signal Process. Mag., 28(6):137–142, 2011.

[120] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process., 13(4):600–612, April 2004.

[121] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. arXiv:1912.00042, 2019.

[122] Yujia Xue, Shiyi Cheng, Yunzhe Li, and Lei Tian. Reliable deep-learning-based phase imaging with uncertainty quantification. Optica, 6(5), 2019.

[123] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3):356–365, 2016.

[124] Jure Zbontar et al. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv:1811.08839, 2018.

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

[126] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proc. IEEE Conf. Comp. Vision Pattern Recog., pages 586–595, 2018.

[127] Tao Zhang, John M Pauly, Shreyas S Vasanawala, and Michael Lustig. Coil compression for accelerated imaging with Cartesian sampling. Magn. Reson. Med., 69(2):571–582, 2013.

[128] 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. Scientific Data, 9(1):152, 2022.