UncertaINR: Uncertainty Quantification of End-to-End Implicit Neural Representations for Computed Tomography

Francisca Vasconcelos [*1] **Bobby He** [*1] **Nalini Singh** [2] **Yee Whye Teh** [1]

Abstract

Implicit neural representations (INRs) have achieved impressive results for scene reconstruction and computer graphics, where their performance has primarily been assessed on reconstruction accuracy. However, in medical imaging, where the reconstruction problem is underdetermined and model predictions inform high-stakes diagnoses, uncertainty quantification of INR inference is critical. To that end, we study UncertaINR: a Bayesian reformulation of INR-based image reconstruction, for computed tomography (CT). We test several Bayesian deep learning implementations of UncertaINR and find that they achieve well-calibrated uncertainty, while retaining accuracy competitive with other classical, INR-based, and CNN-based reconstruction techniques. In contrast to the best-performing prior approaches, UncertaINR does not require a large training dataset, but only a handful of validation images.

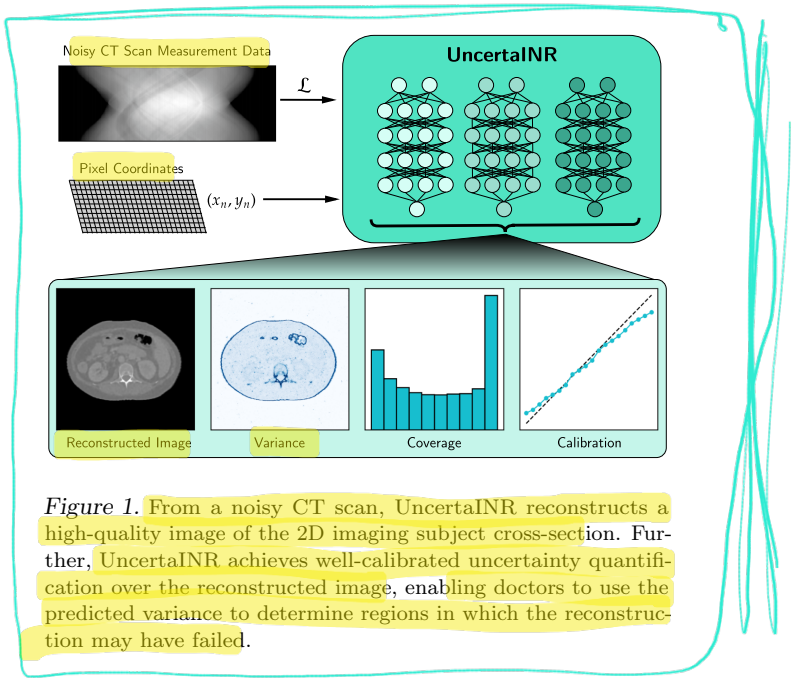

Figure 1. From a noisy CT scan, UncertaINR reconstructs a high-quality image of the 2D imaging subject cross-section. Further, UncertaINR achieves well-calibrated uncertainty quantification over the reconstructed image, enabling doctors to use the predicted variance to determine regions in which the reconstruction may have failed.

1. Introduction

In 2010, 5 billion medical imaging studies were performed worldwide, two-thirds of which employed ionizing radiation (Roobottom et al., 2010). Since then, the use of radiology has only grown, making diagnostic X-rays the largest man-made source of radiation exposure to the general population (de Gonzalez & Darby, 2004; Picano, 2004; Lin, 2010; Picano, 2004). Computed tomography (CT) comprises the majority of this exposure (Brenner & Hall, 2007), with an estimated 29,000 current or future cancer cases linked to CT scans performed in the United States of America in 2007 alone (De González et al., 2009).

Due to the aforementioned radiation risk, there is interest in CT reconstruction techniques which achieve high-quality images from few measurements. However, in the case of limited and noisy data, the CT image reconstruction problem is underdetermined. Furthermore, medical datasets are often small and apparatus-specific, posing a significant challenge for data-driven and standard deep-learning approaches. Finally, medical imaging is one of the highest-stakes image reconstruction domains, with reconstructed images informing doctor decisions. Even a small image artifact could result in misdiagnosis and life-threatening consequences. In this work, we propose UncertaINR – a potential solution to all these challenges for CT image reconstruction.

The first key component of UncertaINR is an implicit neural representation (INR). INRs represent complex coordinate-based signals as functions encoded by small neural networks (NNs). Specifically, UncertaINR lever-

[*]Equal contribution [1]Department of Statistics, University of Oxford, Oxford, UK [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: Francisca Vasconcelos <francisca.vasconcelos@keble.ox.ac.uk>, Bobby He <bobby.he@stats.ox.ac.uk>.

Preprint

ages an end-to-end INR, which represents the reconstructed image as a function mapping image coordinates to pixel values, $f : (x, y) \to [0, 1]$. In the context of CT image reconstruction, INRs present significant benefits over classical or deep-learning based approaches. First, INRs learn a functional form of the image, requiring far fewer parameters than deep-learning or classical grid/voxel approaches. Second, INRs do not require large training datasets, but only a small number of validation images for network hyperparameter tuning. Thus, the INR model is ideal for few-shot learning and easily adaptable to data coming from different imaging devices.

The second key component of UncertaINR is uncertainty quantification, achieved in this work via a Bayesian reformulation of the problem setting. Given the high-stakes nature of medical image reconstruction, we argue that, in addition to achieving decent reconstruction accuracy, models should also be well calibrated. Uncertainty estimates can further serve as important information sources for doctors. For example, if model variance is large in critical image regions, e.g. the location of a potential tumor, a doctor could order additional measurements to ensure proper diagnosis. Uncertainty quantification can also be used to reduce healthcare costs via automated triage, e.g. by assigning images with varying degrees of uncertainty to healthcare providers of relevant expertise. Finally, understanding of model uncertainty could enable techniques, such as active learning (Cohn et al., 1996), to inform more efficient measurement collection and thereby reduce patient radiation exposure.

In this work, we use UncertaINR to test the efficacy of several Bayesian deep-learning approaches – Bayes-by-backprop, Monte Carlo dropout, and deep ensembles – for uncertainty quantification of INRs. We find that UncertaINR attains well-calibrated uncertainty estimates without sacrificing reconstruction quality relative to other classical, INR-based, and CNN-based reconstruction techniques on realistic, noisy data.

2. Problem Setting and Related Works

We will now define the CT image reconstruction problem, informing a literature review of existing reconstruction approaches, including recent work on INRs.

2.1. CT Image Reconstruction

In this work, we focus on reconstruction of 2D images. Specifically, we represent a 2D image as a function $f(x, y)$ of tissue attenuation coefficients at locations (x, y) within an imaging subject cross-section. As visualized in Figure 2, CT scanners operate by rotating a set of X-ray emitters and detectors around the subject. The collected measurements do not constitute pixel values of the desired image f^* but instead its Radon transform,

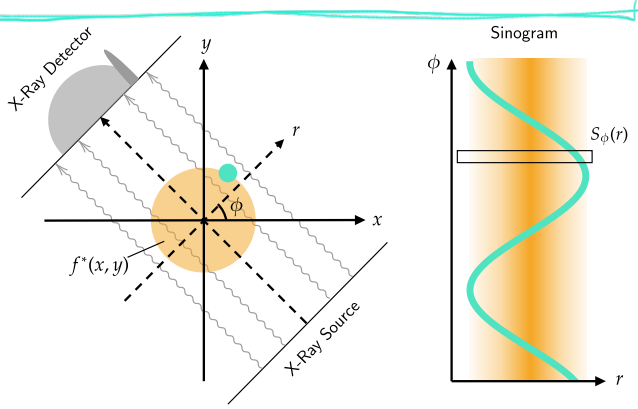

Figure 2. A CT scan of attenuation-coefficient image $f^*(x, y)$ results in sinogram $S_\phi(r)$.

$$S_\phi(r) = \int\int f^*(x, y)\, \delta(x \cos\phi + y \sin\phi - r)\, dxdy, \tag{1}$$

for view-angles ϕ and X-ray detector radii r. $S_\phi(r)$ is known as a *sinogram* and is not directly human-interpretable. The reconstruction of $f(x, y)$ from $S_\phi(r)$ is governed by the inverse Radon transform. Appendix A gives a detailed description of CT measurement physics and the image reconstruction problem. While the Projection Slice Theorem (Bracewell, 1956) ensures that f^* can be fully reconstructed from the complete sinogram, practical measurement data is finite and noisy, making the image reconstruction problem underdetermined and quantification of the reconstruction uncertainty desireable.

In the following, we assume that the image $f(x, y)$ will be reconstructed on a finite grid of pixels $\mathcal{X} \times \mathcal{Y}$, while sinogram measurements are obtained on a finite set of view-angles, Φ, and X-ray detector radii, $\mathcal{R}$. For notational convenience, we denote the resulting ith sinogram measurement by S_i, for $i \in \{1, \ldots, |\Phi \times \mathcal{R}|\}$ and jth pixel value f_j, for $j \in \{1, \ldots, |\mathcal{X} \times \mathcal{Y}|\}$. In a slight abuse of notation, we also denote the resulting vectors of pixel values and sinogram measurements as f and S respectively. The resulting discretization of Equation 1 becomes

$$\sum_{j=1}^{|\mathcal{X}\times\mathcal{Y}|} A_{ij} f_j = S_i, \forall i \; ; \quad \text{equivalently,} \quad \mathbf{A}f = S, \tag{2}$$

where the ijth entry of the discretized Radon transform matrix, $\mathbf{A}$, represents the extent to which pixel

j contributes to the ith prediction measurement, e.g. A_{ij} is zero when pixel j is not along the ray measured by S_i.

2.2. Classical CT Reconstruction Techniques

In this section we describe the most prominent classical CT reconstruction techniques (Kak & Slaney, 2001). Appendix B provides a more detailed discussion.

The simplest analytical technique for reconstructing an image f from measurements S is filtered backprojection (FBP). FBP applies a ramp filter to the 1D Fourier transforms of the projection data to approximate the surrounding, missing sectors of Fourier space, and then reconstructs the resulting image via an inverse Fourier transform. FBP provides a simple closed-form estimate of the image and, by upweighting high frequencies, enables reconstruction of detailed image structures. However this upweighting can also emphasize high-frequency noise, resulting in poor image quality.

Iterative reconstruction techniques can mitigate the effects of noise. These strategies solve the linear system of equations posed by the discretized Radon transform (Equation 2) via the Kaczmarz method (Kaczmarz, 1937), which initializes and then iteratively projects a solution onto the hyperplanes defined by each equation in Equation 2. The algebraic reconstruction technique (ART) (Gordon et al., 1970), simultaneous iterative reconstruction technique (SIRT) (Gilbert, 1972), simultaneous algebraic reconstruction technique (SART) (Andersen & Kak, 1984), and conjugate gradient for least squares (CGLS) (Yuan & Iusem, 1996) are variations on the relative timings, weightings, and parameterizations of these updates.

Another class of algorithms frames the reconstruction problem as minimizing the regularized objective,

$$\min_f \|\mathbf{A}f - S\|^2 + \lambda T(f), \tag{3}$$

where $T(f)$ is a regularizer encoding a prior on f, with regularization strength λ . In medical imaging, a typical regularizer is the total-variation (TV),

$$T_{\text{TV}}(f) = \sum_{x,y \in \mathcal{X}\times\mathcal{Y}} \Big(f(x+1,y) - f(x,y))^2 + (f(x,y+1) - f(x,y))^2 \Big)^{\frac{1}{2}}, \tag{4}$$

which removes unwanted image noise and artifacts, while preserving important details such as edges (Rudin et al., 1992). The minimization problem in Equation 3, with TV regularization, can be solved using proximal gradient-based techniques such as the fast iterative shrinkage-thresholding algorithm (FISTA-TV) (Beck & Teboulle, 2009). Alternatively, the expectation-maximization (EM) algorithm can be used to iteratively maximize the log likelihood of the projections given the estimated image (Dong, 2007).

2.3. Deep Learning CT Reconstruction Techniques

There has been significant recent interest in training NNs with large-scale datasets to reconstruct high quality CT images from low-dose acquisitions. These methods offer faster reconstruction times than purely iterative methods because reconstruction requires only a single forward pass through a NN, instead of a large number of optimization updates.

Some deep learning approaches use an analytical reconstruction of a low-dose CT image as input to a network trained to directly produce an artifact-free reconstruction from a higher dose acquisition (Chen et al., 2017a;b; Liu & Zhang, 2018; Yang et al., 2018). Alternatively, "unrolled" network architectures (Adler & Öktem, 2018; Jin et al., 2017; Wu et al., 2019) solve Equation 3 by chaining together network layers such that each layer of the network computes one optimization update. In a common version of this formulation, the layer first applies an analytical update based on the data-consistency term in Equation 3, and then a network is trained to approximate the update based on the regularizer. Thus, the network is used to shape a data-driven prior on the reconstructed images. Computing one inference pass through the network quickly simulates performing several sequential optimization steps. Finally, in terms of uncertainty quantification, recent work by Barbano et al. (2021) demonstrates model calibration of a CNN Unet, with a deep-image prior, for small amounts of MNIST data.

In this work, we compare our methods to the FBP-Unet (Jin et al., 2017). We also compare to GM-RED (Sun et al., 2021), which uses a deep denoiser trained on the acquired dataset to define a prior on the reconstructed images. Intuitively, this prior encourages solutions which yield themselves when denoised and thus lie on a manifold of natural images. We note that both FBP-Unet and GM-RED require a large training dataset while our methods require only a few images for validating hyperparameters.

2.4. Implicit Neural Representations

INRs have taken the field of computer graphics by storm, achieving impressive results in novel view synthesis (Mildenhall et al., 2020; Niemeyer et al., 2020; Saito et al., 2019; Sitzmann et al., 2019), shape representation (Chen & Zhang, 2019; Deng et al., 2019; Genova

et al., 2019; 2020; Jiang et al., 2020; Park et al., 2019), and texture synthesis (Henzler et al., 2020; Oechsle et al., 2019). Among the most impactful of these works are neural radiance fields (NeRF) (Mildenhall et al., 2020), which achieve state-of-the-art results in novel view synthesis by using random Fourier features (RFFs) as positional encodings, facilitating the representation of high frequency functions by the NN (Tancik et al., 2020). Since the recent publication of the original NeRF paper, there has been an explosion of literature applying and improving the technique (Dellaert & Yen-Chen, 2020). Particularly impressive among NeRF extensions is NeRF in the Wild (NeRF-W), which renders high-resolution 3D scenes from unstructured collections of 2D images 'in the wild' and encodes transient scene features as tuneable latents (Martin-Brualla et al., 2021). Among attempts to improve NeRF performance, sinusoidal representation networks (SIRENs), which use sinusoidal activation functions, were argued to outperform ReLU-based INRs (Sitzmann et al., 2020).

More recent work has also demonstrated the applicability of INRs to CT image reconstruction. Reed et al. (2021) utilizes parametric motion field warped INRs to perform limited view 4D-CT reconstruction of rapidly deforming scenes. Sun et al. (2021) proposes Coordinate-based Internal Learning (CoIL), which uses INRs to boost the performance of existing reconstruction algorithms, such as those discussed in Sections 2.2 and 2.3. Specifically, an INR is used to learn a functional form of the sinogram, receiving sinogram location (ϕ, r) as input and outputting projection measurement $S_\phi(r)$. The functional sinogram is used to generate artificial measurements from view angles not included in the original measurement sinogram. This artificially INR-enlarged measurement set is then fed to a reconstruction algorithm, which reconstructs the image f, achieving improved reconstruction performance over the same algorithm trained on the original, smaller measurement dataset.

While this work has shown the potential of INRs for medical image reconstruction, to our knowledge there have been no attempts to design INRs with uncertainty calibration. That is the main contribution of the proposed UncertaINR framework.

3. Uncertainty Quantification of INRs for CT

In order to quantify the uncertainty in reconstructing an image f given sinogram measurements S, we reformulate the problem as one of Bayesian inference. We assume a Gaussian measurement model,

$$S_i \mid f \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{A}_i f, \sigma^2) \quad \forall i \in \{1, \dots, |\Phi \times \mathcal{R}|\}, \tag{5}$$

where σ^2 is an assumed known observation noise, and $\mathbf{A}_i$ is the ith row of the discretized Radon transform $\mathbf{A}$. The end-to-end INR approach described in Tancik et al. (2020) is equivalent to maximising the likelihood resulting from the measurement model Equation 5.

A Bayesian nonparametric approach would place a prior directly on the function f, e.g. using a Gaussian process (GP). Alternatively, following the end-to-end INR approach (Tancik et al., 2020), UncertaINR parameterizes the function via a small neural network f_θ with parameters θ, and aims to approximate the Bayesian NN posterior (Hinton & Van Camp, 1993; MacKay, 1992; Graves, 2011) over θ given sinogram measurements. The NN parameterizes a mapping from pixel coordinates (x, y) to pixel values $f_\theta(x, y)$. As noted by Tancik et al. (2020), standard multilayer perceptrons (MLPs) are not able to learn the intricate details of images successfully. Instead we let $f_\theta(x, y) = h_\theta(z_\omega(x, y))$ be a composition of a fixed Random Fourier Feature (RFF) encoding z_ω (Rahimi & Recht, 2007) with frequency ω, followed by a standard MLP h_θ. We find in the ablation study detailed in Appendix E.5.1, that frequency ω has a large effect on the uncertainty calibration in INRs. We detail our architecture choices for the MLP h_θ in Appendices E.3 and E.4. Given that the INR encodes a functional representation of the learned image f_θ, a new network is trained for each set of measurement data, resulting in a new image. Thus, unlike most NN-based approaches, INRs do not require large training datasets, but only small validation sets for hyperparameter tuning.

A common prior for θ for Bayesian NNs is simply an isotropic Gaussian, which is convenient but uninformative. Inspired by the regularization-based methods in Section 2.2, we instead use a composite prior $p(\theta) \propto p_\mathcal{N}(\theta) p_I(\theta)$, where $p_\mathcal{N}(\theta)$ is a typical Gaussian prior, to constrain the NN parameter values, and $p_I(\theta) = \exp(-T(f_\theta))$ encodes a smoothing regularization on reconstructed images. We set $p_I(\theta)$ to be the TV regularizer, as in Equation 4, which is commonly used in the medical imaging domain. The resulting posterior predictive distribution $p(f_\theta|S)$ allows us both to reconstruct the image and to quantify our uncertainty over the reconstruction. Due to the high-dimensionality of θ and the nonlinearity of the neural network, the true posterior is intractable. In this work, we investigate the performance of a number of popular approximate inference techniques from the Bayesian deep learning (BDL) literature.

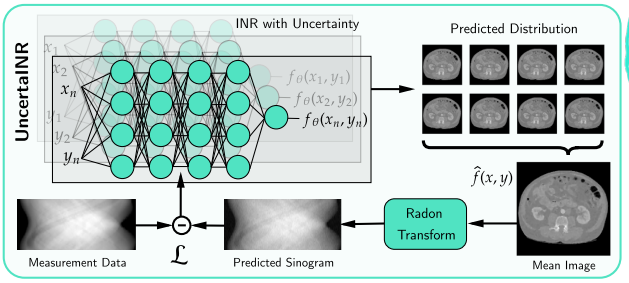

Figure 3. The UncertaINR architecture. An end-to-end INR with uncertainty quantification (BBB, MCD, and/or DEs) is sampled across all image pixels, generating a distribution of predicted images, $\{f_{\theta_n}\}_{n=1}^N$. The predicted sinogram, generated by the Radon transform of predicted mean image $\hat{f}$, is compared to the true measurement data in the INR loss, $\mathcal{L}$. Once training has completed, $\hat{f}$ is reported as the reconstructed image and the predictive distribution is used to quantify uncertainty in the model output.

The overall framework for UncertaINR is illustrated in Figure 3. At its core, the INR maps pixel coordinates (x, y) to pixel values $f_\theta(x, y)$. BDL algorithms aim to produce samples, θ, from the (approximate) posterior. This can be thought of as balancing prior $p(\theta)$ with the likelihood derived from the measurement model of Equation 5,

$$\log p(S|f_\theta) = \frac{1}{2\sigma^2} \sum_{i=1}^{|\Phi\times\mathcal{R}|} \left(S_i - \mathbf{A}_i f_\theta\right)^2 + \text{const.} \quad (6)$$

The likelihood can thus be interpreted as a comparison between the Radon transform of f_θ and the observed measurements, S.

Given N samples, $\{\theta_n\}_{n=1}^N$, from the approximate posterior, we obtain N reconstructed image functions, $\{f_{\theta_n}\}_{n=1}^N$. The sample average,

$$\hat{f}(x, y) = \frac{1}{N} \sum_{n=1}^{N} f_{\theta_n}(x, y) \quad (7)$$

is an estimate of the posterior mean, and can be presented to a doctor as the reconstructed image. Furthermore, the posterior variance can be estimated as

$$V(x, y) = \frac{1}{N-1} \sum_{n=1}^{N} \left(f_{\theta_n}(x, y) - \hat{f}(x, y)\right)^2. \quad (8)$$

This can be used to visualize, as in Figure 1, regions of large model uncertainty. In general, the posterior variability can be used to calculate and report various uncertainty metrics such as coverage and calibration (Section 3.2).

3.1. Bayesian Deep Learning Methods

In this section we describe the BDL methods compared in the UncertaINR framework. Implementation notes are included in Appendix D.

3.1.1. Bayes-By-Backprop

A popular method for variational approximation to exact Bayesian updates is Bayes-by-Backprop (BBB) (Blundell et al., 2015). BBB aims to find the optimal parameters ψ of an approximate distribution on the NN weights, $q(\theta|\psi)$, also known as the variational posterior. This is achieved by maximizing the variational free energy/evidence lower bound (ELBO),

$$\mathcal{L}(S, \psi) = \mathbb{E}_{q(\theta|\psi)}[\log p(S|\theta)] - \mathbb{KL}\left[q(\theta|\psi)\,||\,p(\theta)\right], \quad (9)$$

with respect to the variational parameters ψ. The ELBO can be optimized with respect to ψ using stochastic gradients estimated by Monte Carlo,

$$\nabla_\psi \mathcal{L}(S, \psi) \approx \nabla_\psi \big(\log p(S|\theta) + \log p(\theta) - \log q(\theta|\psi)\big), \quad (10)$$

where $\theta \sim q(\cdot|\psi)$ is a sample drawn from the variational posterior, and the gradient is taken through θ using the reparameterization trick (Rezende et al., 2014).

3.1.2. Monte Carlo Dropout

Another popular approach is Monte Carlo dropout (MCD) (Gal & Ghahramani, 2016). There, the authors argue that optimizing a NN regularized with dropout (Srivastava et al., 2014) applied to every layer can be interpreted as variational approximation for a deep GP. The first two moments of the corresponding variational posterior can be approximated using Monte Carlo, with N samples from NNs sampled with dropout.

3.1.3. Deep Ensembles

Alternatively, predictive uncertainty can be quantified by aggregating the outputs of several NN base learners trained for the same task from different initializations, a method known as deep ensembles (DEs) (Lakshminarayanan et al., 2017). Averaging predictions over multiple NNs consistent with the training data leads to better predictive performance, enables uncertainty quantification, makes DEs robust to model misspecification, and presents a strong baseline in out-of-distribution detection (Ovadia et al., 2019). Ensembling can also be applied on top of other methods to further improve performance. In principle, more base learners results in better performance, but, in practice, training large ensembles is computationally expensive and diminishing returns are observed after ∼10 base learners. In this work, UncertaINR leverages ensembles

of M Monte Carlo dropout base learners, such that Equations 7 and 8 are updated to,

$$\hat{f}_{\text{DE}}(x,y) = \frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N/M} f_{\theta_{m,n}}(x,y) \tag{11}$$

$$V_{\text{DE}}(x,y) = \frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=1}^{N/M}\left(f_{\theta_{m,n}}(x,y) - \hat{f}_{\text{DE}}(x,y)\right)^2.$$

The relationship between Bayesian inference and DEs is an active area of research in the BDL community. Wilson & Izmailov (2020) argue that DEs provide a more compelling approximation to the true posterior than many standard BDL approaches, whilst others have adapted DEs to provide a Bayesian interpretation (Ciosek et al., 2019; Pearce et al., 2020; D'Angelo & Fortuin, 2021). He et al. (2020) characterize how DEs relate to posterior inference in the limit of infinite NN width.

3.2. Model Performance Metrics

In this work, four metrics were selected to assess model performance: peak-signal-to-noise ratio (PSNR), signal-to-noise ratio (SNR), negative log-likelihood (NLL), and expected calibration error (ECE).

PSNR & SNR are common, similar measures of predictive accuracy. For equations, we refer the reader to Appendix C.

ECE assesses a model's ability to predict the probabilities of its outcomes, gauging reliability of the model's confidence in its predictions. Specifically, it describes the discrepancy between the *target coverage* and *achieved coverage*.

To calculate ECE for a given uncertainty quantification method, after training we sample N sets of model weights, $\{\theta_n\}_{n=1}^N$. Thus, each pixel, (x,y), has an empirical distribution of N predicted values, $\hat{F}_N(x,y) = \{f_{\theta_n}(x,y)\}_{n=1}^N$. Ideally, the pixel distribution median would be the ground truth pixel value, $f^*(x,y)$. Since this is unrealistic in practice, for given target coverage p we define,

$$C_{\hat{F}_N,f^*,p}(x,y) = \mathbb{I}\Big\{Q_{50-\frac{p}{2}}\left(\hat{F}_N(x,y)\right) \le f^*(x,y) \le Q_{50+\frac{p}{2}}\left(\hat{F}_N(x,y)\right)\Big\}, \tag{12}$$

where $Q_k(F)$ denotes the kth quantile of distriubtion F. The achieved coverage (AC) is thus defined as the percentage of pixel distributions containing $f^*(x,y)$ in that quantile,

$$\text{AC}(f^*,\hat{F}_N,p) = \frac{1}{|\mathcal{X}\times\mathcal{Y}|}\sum_{x,y} C_{\hat{F}_N,f^*,p}(x,y). \tag{13}$$

If a model is perfectly calibrated, $p\%$ of the reconstructed pixel distributions will contain $f^*(x,y)$ in their $p\%$ quantile, meaning achieved coverage equals target coverage for all quantiles. Given a finite set, $\mathcal{P}$, of percentages evenly spaced in $[0,1]$, ECE is defined as the average difference between achieved and target coverage,

$$\text{ECE}(f^*,\hat{F}_N) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}|\text{AC}(f^*,\hat{F}_N,p) - p|. \tag{14}$$

A *reliability curve*, as visualized in Figure 5, plots achieved coverage as a function of target coverage. Better model calibration results in a curve similar to the identity function, $\text{AC}(f^*,\hat{F}_N,p) = p$.

Furthermore, akin to inverse transform sampling, the marginal distribution of the inverse quantiles of calibrated pixel predictive distributions $\hat{F}_N(x,y)$, at ground truth pixel values, should be uniformly distributed in $[0,1]$. We visualize such *coverage histograms* for UncertaINR in Figure 5. For a further discussion of calibration, coverage, and implementation details, we refer the reader to Appendix C.

NLL is a common metric of probabilistic model quality, assessing both predictive accuracy and uncertainty calibration. We assume an independent Gaussian model. Each pixel, (x,y), is sampled from a Gaussian distribution with the ground truth pixel value, $f^*(x,y)$, as mean and calculated sample variance of all the network pixel responses, $\hat{\sigma}^2$, as variance. The NLL is the negative log-likelihood of all pixel values under this model, assuming independent sampling,

$$\text{NLL}(f_\theta,f^*) = \sum_{x,y}\tfrac{1}{2\hat{\sigma}^2}\left(f_\theta(x,y) - f^*(x,y)\right)^2 + \tfrac{1}{2}\log(2\pi\hat{\sigma}^2).$$

A good probabilistic model will maximize the likelihood, meaning NLL is minimized when $f_\theta(x,y) = f^*(x,y)$ for all (x,y) and variance $\hat{\sigma}^2$ is small.

4. Experimental Results

In this section we discuss two key experiments used to: (1) understand optimal UncertaINR design and (2) assess the efficacy of UncertaINR with different BDL methods relative to existing classical, INR-based, and CNN-based CT reconstruction approaches.

4.1. UncertaINR Ablation Study

We began our study of UncertaINR with a set of ablation experiments of the different UncertaINR hyperparameters, across the BDL approaches. These experiments were performed on artificial data.

Dataset. In order to train and test a large number of models, we used a dataset consisting of artificially-generated (256×256 pixel) Shepp-Logan phantom (Shepp & Logan, 1974) brain scan images. Given the simplified nature of this data, the study was performed in the extremely low measurement settings of 5- and 20-view sinograms. We tuned hyperparameters on a set of 5 validation images, and evaluated performance on 5 test images.

Ablations. We ablated the activation function (Tanh, SoftPlus, Sine, SiLU, and ReLU), depth, width, and random Fourier feature (RFF) embedding frequency. For MCD we also ablated probability of dropout, while for BBB we ablated prior standard deviation and KL factor. For the sake of brevity, we only report main findings here. A more detailed analysis is provided in Appendix E.5.

Activation function and RFF frequency were found to be the two most critical hyperparameters. Although the Sine activation function achieved the best-performing models, the resulting networks were very sensitive to hyperparameter choice. SiLU, ReLU, and Tanh networks achieved slightly lower, but more consistent reconstruction accuracies. In line with recent work demonstrating the importance of RFFs for learning high-frequency image components (Tancik et al., 2020), we found that RFF frequency significantly affected model performance. Specifically, RFF frequency must be proportional to the number of view angles – too low (high) an RFF frequency leads to blurry images (high-frequency image artifacts).

Baselines. UncertaINR was compared to several classical reconstruction baselines: FBP, CGLS, EM, SART, and SIRT. Since these approaches do not quantify uncertainty, only reconstruction accuracy is reported. Table 1 summarizes the performance of these baselines and the best-performing UncertaINR (UINR) models for each BDL approach.

Analysis. An interesting finding of this study was that MCD UINRs significantly outperform BBB UINRs. Due to their poor performance, BBB UINRs were not considered in the following experiments. With regards to reconstruction accuracy, all other UncertaINR methods significantly outperformed the classical approaches. Among the UncertaINR methods, ensembles of 5 and 10 MCD UINR base learners achieved the best performance. Similar conclusions can be drawn with respect to uncertainty quantification.

Table 1. **Ablation Study**: UncertaINR accuracy, relative to classical approaches, and calibration results on the Shepp-Logan phantom dataset. Results are averaged across 5 validation and 5 test images, with the best result for each metric (PSNR, NLL, and ECE) bolded.

View #	Recon. Type	Validation Set PSNR	Validation Set NLL	Validation Set ECE	Test Set PSNR	Test Set NLL	Test Set ECE
5	FBP	7.68	–	–	5.15	–	–
	CGLS	16.38	–	–	14.62	–	–
	EM	21.39	–	–	19.88	–	–
	SART	21.12	–	–	19.75	–	–
	SIRT	21.12	–	–	21.12	–	–
	BBB UINR	23.26	-1.190	0.152	22.52	0.138	0.203
	MCD UINR	26.15	-1.473	0.111	24.45	-1.572	0.083
	DE-2 MCD UINR	26.31	-1.730	0.091	24.49	-1.774	0.069
	DE-5 MCD UINR	**26.44**	-1.737	0.085	**24.88**	-1.751	**0.067**
	DE-10 MCD UINR	26.36	**-2.226**	**0.075**	24.67	**-1.969**	0.068
20	FBP	17.35	–	–	15.71	–	–
	CGLS	21.85	–	–	20.82	–	–
	EM	30.22	–	–	29.11	–	–
	SIRT	31.98	–	–	30.44	–	–
	SART	31.97	–	–	30.45	–	–
	BBB UINR	28.25	1.650	**0.121**	28.16	0.562	0.119
	MCD UINR	33.74	0.701	0.135	33.08	1.093	0.113
	DE-2 MCD UINR	33.96	0.005	0.136	33.44	-0.372	0.102
	DE-5 MCD UINR	34.31	-0.364	0.134	**34.02**	-0.625	0.101
	DE-10 MCD UINR	**34.38**	**-0.529**	0.131	33.86	**-0.774**	**0.096**

Overall, these experiments show that, beyond enabling calibrated uncertainty, MCD UINRs and DEs of MCD UINR base learners outperform even the best classical reconstruction techniques in terms of reconstruction accuracy. Furthermore, despite the small size of the validation set (5 images), UncertaINR networks generalized well to the test set. For 20-views, validation and test accuracies were comparable and calibration improved for the test set. For 5-views, despite the slight degredation of PSNR and NLL, ECE improved in the test set. These results suggest that a small validation set is sufficient to tune an UncertaINR.

4.2. UncertaINR Performance Assessment

We next compared UncertaINR to state-of-the-art reconstruction approaches, on a real-world dataset.

Dataset. A validation set of 3 and test set of 8 reconstructed abdominal CT scan images (512 × 512 pixel), provided by the Mayo Clinic for the 2016 Low-dose CT AAPM Grand Challenge (McCollough et al., 2017), were used to train and assess UncertaINR. 60- and 120-view sinograms were generated from these images, with Gaussian noise added to achieve a desired 40dB SNR (relative to the original, noiseless sinogram).

Baselines. As in the ablation study, UncertaINR reconstruction was compared to our implementation of the classical FBP, CGLS, EM, SART, and SIRT methods. We also report the results presented in Sun et al. (2021) for FISTA-TV, GM-RED, FBP-UNet,

Table 2. **Performance Assessment**: Accuracy and calibration results of all reconstruction approaches are presented for the AAPM-Mayo dataset, with noise added to the sinogram to achieve a 40dB SNR. The table is divided into 5 sections: classical techniques, deep-learning methods, classical techniques with CoIL, deep-learning methods with CoIL, and INRs. (*) denotes results taken directly from the CoIL paper (Sun et al., 2021). Results are averaged over all 8 test set images and the best result for each metric (SNR, NLL, and ECE) is bolded.

Reconstruction Method	60 Views ↑ SNR	NLL	ECE	120 Views ↑ SNR	NLL	ECE
FBP	10.58	–	–	14.11	–	–
EM	14.47	–	–	15.55	–	–
CGLS	20.08	–	–	21.94	–	–
SIRT	20.89	–	–	21.36	–	–
SART	21.54	–	–	21.77	–	–
FISTA-TV*	26.08	–	–	27.59	–	–
FBP-UNet*	27.08	–	–	29.18	–	–
GM-RED*	27.12	–	–	29.30	–	–
FBP CoIL*	23.48	–	–	24.52	–	–
FISTA-TV CoIL*	26.95	–	–	28.95	–	–
FBP-UNet CoIL*	**27.93**	–	–	29.71	–	–
GM-RED CoIL*	27.42	–	–	**29.79**	–	–
INR	27.25	–	–	28.81	–	–
DE-2 UINR	27.29	24.55	0.224	28.83	18.86	0.222
DE-5 UINR	27.30	8.427	0.176	28.83	8.804	0.183
DE-10 UINR	27.28	6.882	0.144	28.82	6.346	0.162
MCD UINR	27.38	-3.447	0.078	28.65	-3.759	0.071
DE-2 MCD UINR	27.44	-3.573	0.063	28.68	-3.819	0.056
DE-5 MCD UINR	27.48	-3.660	0.051	28.70	-3.876	**0.043**
DE-10 MCD UINR	27.46	**-3.689**	**0.045**	28.74	**-4.090**	0.053

and all CoIL methods[1]. Finally, we compare to an implementation of our end-to-end INR without uncertainty quantification, comparable to the INR of Tancik et al. (2020). Table 2 presents reconstruction results of these baseline methods as well as the best-performing UncertaINRs trained with MCD, DEs of INR base learners, and DEs of MCD UINR base learners. For more information about the dataset as well as model hyperparameters, we refer the reader to Appendix F.

Analysis. The results of Table 2 show that UncertaINR achieves better reconstruction accuracy than all the classical methods and is competitive with the existing INR-based and CNN-based approaches. In fact, in the low-measurement regime (60-views), the UncertaINR of highest reconstruction accuracy (DE-5 MCD UINR) outperformed all approaches other than FBP-UNET with CoIL. Note that this was despite our focus on achieving calibrated models over optimizing reconstruction accuracy. Furthermore, unlike deep-learning methods, UncertaINR does not require a large training dataset and, unlike CoIL-based methods, it does not require pre-processing.

With regards to uncertainty calibration, the results of Table 2 are surprising. Contrary to the common deep-learning belief that DEs achieve the best model calibration (Ovadia et al., 2019), we found that MCD is more important for INR calibration. UncertaINR with MCD and with DEs of MCD base learners achieved significantly better model calibration than DEs of INRs without uncertainty. For example, the introduction of MCD in a DE-10 UINR reduced ECE from 0.144 to 0.045 (60-views) and from 0.162 to 0.053 (120-views). Although increasing ensemble sizes generally improved model calibration, the performance boost was not as significant as that of using MCD.

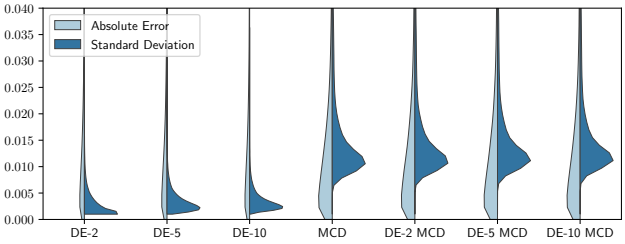

Figure 4. The distributions of absolute errors and predicted standard deviations are useful indicators of model calibration. Here, these distributions are plotted for all the UncertaINR models, reported in Table 2, for a single AAPM-Mayo test set image.

The benefits of MCD for INR calibration are further illustrated by Figure 4, which compares the distributions of pixel-wise absolute error versus predicted standard deviation across UncertaINR models. For DEs without MCD, the standard deviation distribution is skewed towards smaller values than the absolute error distribution, indicating that the model is overconfident. In contrast, models with MCD predicted larger standard deviations, resulting in a distribution more closely resembling that of the absolute error. Meanwhile, there is no significant change in the absolute error and standard deviation distribution as ensemble size increases, with and without MCD. It should also be noted that using MCD, which only requires a few extra forward passes with dropout at each training iteration, has a much smaller computational overhead than training large DEs, which require training the entire network several times. Thus, MCD proves to be a computationally-efficient, yet highly effective approach for achieving calibrated INR uncertainty quantification.

Figure 5 visualizes the model output, calibration diagnostics, and uncertainty for the best-calibrated 60-view UncertaINR model (DE-10 MCD UINR). The nearly ideal reliability curves and largely uniform coverage distributions, as well as the relatively low NLL and ECE values reported in Table 2, show that that the model is well-calibrated, though still imperfect (there are slightly higher masses near either extremes on the coverage histograms). However, these metrics as well as the reconstruction error would not be available to a

[1]Since no code is available for CoIL, we were unable to reproduce their results.

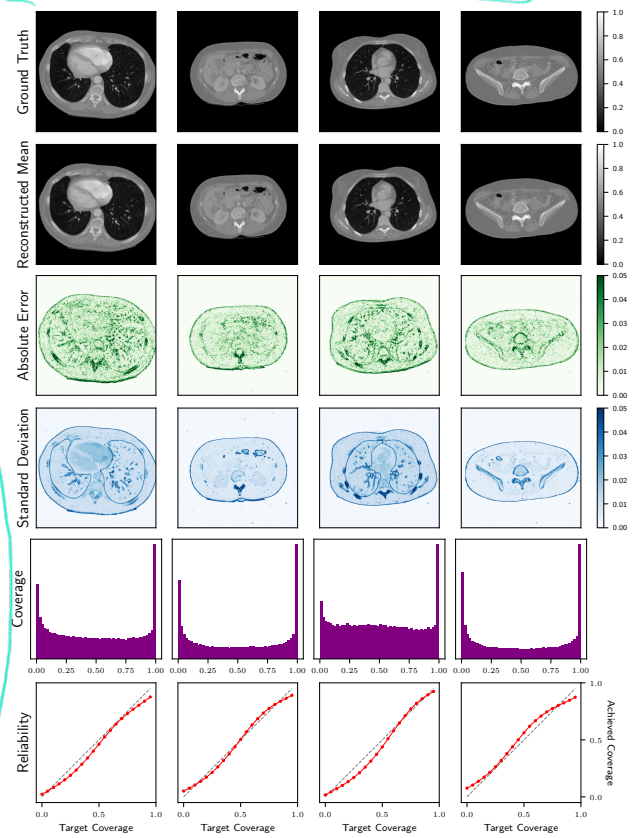

Figure 5. Visualizations of the predicted mean, absolute error, predicted standard deviation, coverage histogram, and reliability curves of the best calibrated 60-view UncertaINR (DE-10 MCD UINR) presented for 4 of the 8 AAPM-Mayo test set images.

doctor without access to the ground truth image. In this case, the only visualizations available are the reconstructed mean and standard deviation images. Given that a well-calibrated model reports larger variance in regions of larger absolute error, variance can be used as a proxy for reconstruction error. Specifically, the absolute error images of Figure 5 show that the model underperforms at predicting boundaries between different tissues. This is reflected in the corresponding higher uncertainties of the standard deviation images. When presented to a doctor, the latter would inform them to be more cautious regarding diagnoses based on, say, perceived issues in organ linings.

5. Conclusion

We proposed to address challenges in CT image reconstruction through INRs with well-calibrated uncertainty. Through the UncertaINR framework, a Bayesian formulation of the CT image reconstruction problem, we assessed the relative performance of several BDL methods – BBB, MCD, and DEs – for uncertainty quantification of INRs in the context of CT image reconstruction. In terms of model calibration, we found that MCD significantly outperformed BBB and DEs of INR base learners without uncertainty quantification. DEs of MCD base learners achieved the best overall performance, enabling calibrated uncertainty quantification, while retaining reconstruction accuracy competitive with state-of-the-art approaches.

References

Adler, J. and Öktem, O. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37 (6):1322–1332, 2018.

Andersen, A. H. and Kak, A. C. Simultaneous Algebraic Reconstruction Technique (SART): A Superior Implementation of the ART Algorithm. *Ultrasonic Imaging*, 6(1):81–94, 1984.

Barbano, R., Antoran, J., Hernández-Lobato, J. M., and Jin, B. A probabilistic deep image prior over image space. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Berger, M., Coursey, J., Zucker, M., and Chang, J. *ESTAR, PSTAR, and ASTAR: Computer Programs for Calculating Stopping-Power and Range Tables for Electrons, Protons, and Helium Ions*. National Institute of Standards and Technology, Gaithersburg, MD, 2004. (version 1.2.3) Availabile: http://physics.nist.gov/Star [2021, 08, 22]. Originally published as: Berger, M.J., NISTIR 4999, National Institute of Standards and Technology, Gaithersburg, MD (1993).

Biewald, L. Experiment Tracking with Weights and Biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight Uncertainty in Neural Network. In *International Conference on Machine Learning*, pp. 1613–1622. PMLR, 2015.

Bracewell, R. N. Strip Integration in Radio Astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956.

Brenner, D. J. and Hall, E. J. Computed tomography — an increasing source of radiation exposure. *New England Journal of Medicine*, 357(22):2277–2284, 2007.

Bull, D. R. Chapter 4 - Digital Picture Formats and Representations. In Bull, D. R. (ed.), *Communicating Pictures*, pp. 99–132. Academic Press, Oxford, 2014. ISBN 978-0-12-405906-1.

Bust, G. S. and Mitchell, C. N. History, Current State, and Future Directions of Ionospheric Imaging. *Reviews of Geophysics*, 46(1), 2008.

Chen, H., Zhang, Y., Kalra, M. K., Lin, F., Chen, Y., Liao, P., Zhou, J., and Wang, G. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36 (12):2524–2535, 2017a.

Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., and Wang, G. Low-dose ct via convolutional neural network. *Biomedical optics express*, 8(2):679–694, 2017b.

Chen, Z. and Zhang, H. Learning Implicit Fields for Generative Shape Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.

Ciosek, K., Fortuin, V., Tomioka, R., Hofmann, K., and Turner, R. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*, 2019.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active Learning with Statistical Models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

Cybenko, G. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

D'Angelo, F. and Fortuin, V. Repulsive deep ensembles are bayesian. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

de Gonzalez, A. B. and Darby, S. Risk of Cancer from Diagnostic X-Rays: Estimates for the UK and 14 Other Countries. *The Lancet*, 363(9406):345–351, 2004.

De González, A. B., Mahesh, M., Kim, K.-P., Bhargavan, M., Lewis, R., Mettler, F., and Land, C. Projected Cancer Risks from Computed Tomographic Scans Performed in the United States in 2007. *Archives of Internal Medicine*, 169(22):2071–2077, 2009.

Deans, S. R. *The Radon Transform and Some of Its Applications*. Courier Corporation, 2007.

Dellaert, F. and Yen-Chen, L. Neural Volume Rendering: NeRF And Beyond. *arXiv preprint arXiv:2101.05204*, 2020.

Deng, B., Lewis, J. P., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., and Tagliasacchi, A. NASA: Neural Articulated Shape Approximation. *arXiv preprint arXiv:1912.03207*, 2019.

Der Kiureghian, A. and Ditlevsen, O. Aleatory or Epistemic? Does It Matter? *Structural Safety*, 31 (2):105–112, 2009.

Dong, B.-Y. Image Reconstruction using EM Method in X-Ray CT. In *2007 International Conference on Wavelet Analysis and Pattern Recognition*, volume 1, pp. 130–134. IEEE, 2007.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. Incorporating Second-Order Functional Knowledge for Better Option Pricing. *Advances in Neural Information Processing Systems*, pp. 472–478, 2001.

Dupont, E., Golinski, A., Alizadeh, M., Teh, Y. W., and Doucet, A. COIN: COmpression with implicit neural representations. In *Neural Compression: From Information Theory to Applications – Workshop @ ICLR 2021*, 2021.

Esposito, P. BLiTZ - Bayesian Layers in Torch Zoo (a Bayesian Deep Learing library for Torch). https://github.com/piEsposito/blitz-bayesian-deep-learning/, 2020.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W. T., and Funkhouser, T. Learning Shape Templates with Structured Implicit Functions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7154–7164, 2019.

Genova, K., Cole, F., Sud, A., Sarna, A., and Funkhouser, T. Local Deep Implicit Functions for 3D Shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4857–4866, 2020.

Gilbert, P. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of theoretical biology*, 36(1):105–117, 1972.

Glorot, X., Bordes, A., and Bengio, Y. Deep Sparse Rectifier Neural Networks. In *Proceedings of the*

Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Gordon, R., Bender, R., and Herman, G. T. Algebraic Reconstruction Techniques (ART) for Three-Dimensional Electron Microscopy and X-Ray Photography. *Journal of Theoretical Biology*, 29(3):471–481, 1970.

Graves, A. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017a.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On Calibration of Modern Neural Networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017b.

Hanin, B. Universal Function Approximation by Deep Neural Nets with Bounded Width and ReLU Activations. *Mathematics*, 7(10), 2019. ISSN 2227-7390.

Hara, K., Saito, D., and Shouno, H. Analysis of Function of Rectified Linear Unit used in Deep Learning. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2015. doi: 10.1109/IJCNN.2015.7280578.

He, B., Lakshminarayanan, B., and Teh, Y. W. Bayesian deep ensembles via the neural tangent kernel. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1010–1022. Curran Associates, Inc., 2020.

Helgason, S. *Groups & Geometric Analysis: Radon Transforms, Invariant Differential Operators and Spherical Functions: Volume 1*. Academic Press, 1984.

Henzler, P., Mitra, N. J., and Ritschel, T. Learning a Neural 3d Texture Space from 2D Exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8356–8364, 2020.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Hornik, K. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2):251–257, 1991.

Hornik, K., Stinchcombe, M., and White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989.

Hsieh, J. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*, volume 114. SPIE Press, 2003.

Hubbell, J. and Seltzer, S. *Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients*. National Institute of Standards and Technology, Gaithersburg, MD, 2004. (version 1.4) Availabile: http://physics.nist.gov/xaamdi [2021, 08, 22]. Originally published as NISTIR 5632, National Institute of Standards and Technology, Gaithersburg, MD (1995).

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 6–6, 2021.

Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., Funkhouser, T., et al. Local Implicit Grid Representations for 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6001–6010, 2020.

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

Kaczmarz, S. Angenaherte auflosung von systemen linearer gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, 1937.

Kak, A. C. and Slaney, M. *Principles of computerized tomographic imaging*. SIAM, 2001.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30:5574–5584, 2017.

Kidger, P. and Lyons, T. Universal Approximation with Deep Narrow Networks. In *Conference on Learning Theory*, pp. 2306–2327. PMLR, 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *stat*, 1050:1, 2014.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate Uncertainties for Deep Learning using Calibrated Regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Lin, E. C. Radiation risk from medical imaging. *Mayo Clinic proceedings*, 85 12:1142–6; quiz 1146, 2010.

Lindell, D. B., Martel, J. N., and Wetzstein, G. AutoInt: Automatic Integration for Fast Neural Volume Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14556–14565, 2021.

Liu, Y. and Zhang, Y. Low-dose ct restoration via stacked sparse denoising autoencoders. *Neurocomputing*, 284:80–89, 2018.

Lu, L., Shin, Y., Su, Y., and Karniadakis, G. E. Dying ReLU and Initialization: Theory and Numerical Examples. *arXiv preprint arXiv:1903.06733*, 2019.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The Expressive Power of Neural Networks: A View from the Width. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The Expressive Power of Neural Networks: A View from the Width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6232–6240, 2017b.

MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural computation*, 4 (3):448–472, 1992.

Martin-Brualla, R., Radwan, N., Sajjadi, M. S., Barron, J. T., Dosovitskiy, A., and Duckworth, D. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7210–7219, 2021.

McCollough, C. H., Bartley, A. C., Carter, R. E., Chen, B., Drees, T. A., Edwards, P., Holmes III, D. R., Huang, A. E., Khan, F., Leng, S., et al. Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. *Medical physics*, 44(10):e339–e352, 2017.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, pp. 405–421. Springer, 2020.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Niemeyer, M., Mescheder, L., Oechsle, M., and Geiger, A. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3504–3515, 2020.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring Calibration in Deep Learning. In *CVPR Workshops*, volume 2(7), 2019.

Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., and Geiger, A. Texture Fields: Learning Texture Representations in Function Space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4531–4540, 2019.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32:13991–14002, 2019.

Pan, X., Sidky, E. Y., and Vannier, M. Why Do Commercial CT Scanners Still Employ Traditional, Filtered Back-Projection for Image Reconstruction? *Inverse Problems*, 25(12):123009, 2009.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. DeepSDF: Learning Continuous

Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pearce, T., Leibfried, F., and Brintrup, A. Uncertainty in neural networks: Approximately bayesian ensembling. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 234–244, 26–28 Aug 2020.

Pelt, D. M., Gürsoy, D., Palenstijn, W. J., Sijbers, J., De Carlo, F., and Batenburg, K. J. Integration of TomoPy and the ASTRA Toolbox for Advanced Processing and Reconstruction of Tomographic Synchrotron Data. *Journal of Synchrotron Radiation*, 23(3):842–849, 2016.

Picano, E. Sustainability of Medical Imaging. *Bmj*, 328(7439):578–580, 2004.

Pryse, S., Kersley, L., Rice, D., Russell, C., and Walker, I. Tomographic Imaging of the ionospheric Mid-Latitude Trough. *Annales Geophysicae*, 11(2-3):144–149, 1993.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the Expressive Power of Deep Neural Networks. In *International Conference on Machine Learning*, pp. 2847–2854. PMLR, 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1177–1184, 2007.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for Activation Functions. *arXiv preprint arXiv:1710.05941*, 2017.

Reed, A. W., Kim, H., Anirudh, R., Mohan, K. A., Champley, K., Kang, J., and Jayasuriya, S. Dynamic CT Reconstruction from Limited Views with Implicit Neural Representations and Parametric Motion Fields. *arXiv preprint arXiv:2104.11745*, 2021.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 1278–1286, 2014.

Roobottom, C., Mitchell, G., and Morgan-Hughes, G. Radiation-reduction strategies in cardiac computed tomographic angiography. *Clinical Radiology*, 65(11): 859–867, 2010. ISSN 0009-9260.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2304–2314, 2019.

Shepp, L. A. and Logan, B. F. The Fourier Reconstruction of a Head Section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

Sitzmann, V., Zollhoefer, M., and Wetzstein, G. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. *Advances in Neural Information Processing Systems*, 32:1121–1132, 2019.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. Implicit Neural Representations with Periodic Activation Functions. *Advances in Neural Information Processing Systems*, 33, 2020.

Smith-Bindman, R., Lipson, J., Marcus, R., Kim, K.-P., Mahesh, M., Gould, R., De González, A. B., and Miglioretti, D. L. Radiation Dose Associated with Common Computed Tomography Examinations and the Associated Lifetime Attributable Risk of Cancer. *Archives of Internal Medicine*, 169(22):2078–2086, 2009.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Sun, Y., Liu, J., Xie, M., Wohlberg, B., and Kamilov, U. S. Coil: Coordinate-based internal learning for tomographic imaging. *IEEE Transactions on Computational Imaging*, 7:1400–1412, 2021.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Wilson, A. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 2020-December, 2020. ISSN 1049-5258.

Wu, D., Kim, K., and Li, Q. Computationally efficient deep neural network for computed tomography image reconstruction. *Medical physics*, 46(11):4763–4776, 2019.

Yadan, O. Hydra - A Framework for Elegantly Configuring Complex Applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., and Wang, G. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6): 1348–1357, 2018.

Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., and Kanazawa, A. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

Yuan, J.-Y. and Iusem, A. N. Preconditioned Conjugate Gradient Method for Generalized Least Squares Problems. *Journal of Computational and Applied Mathematics*, 71(2):287–297, 1996.

Zaidi, S., Zela, A., Elsken, T., Holmes, C., Hutter, F., and Teh, Y. W. Neural Ensemble Search for Uncertainty Estimation and Dataset Shift. *arXiv preprint arXiv:2006.08573*, 2020.

A. Computed Tomography

Computed tomography (CT), also known as computed axial/assisted tomography (CAT), is a noninvasive medical imaging technique frequently used in radiology to generate detailed images of the body. Since its original development in the 1970s, CT has become widespread in medical imaging – with over 70 million CT scans taken and reported annually in the United States, since 2007 (Smith-Bindman et al., 2009). There are multiple types of CT scanners, such as spiral CT, electron beam CT, and CT perfusion imaging. In this work, we focus on spiral, also know as spinning tube or helical, CT.

In spiral CT, illustrated in Fig. 6, the patient lies along the central axis of a cylindrical measurement tube. As the scan is performed, an X-ray generator rotates around the patient while moving along the axis of measurement. X-rays are emitted, which pass through the patient and are attenuated at various rates by the different types of tissues in the body, as described in Sec. A.1. After exiting the body, the attenuated X-rays are measured by X-ray detectors positioned and moving opposite to the X-ray source. These measurements are used to create a sinogram, as described in Section A.2, which is not understandable by doctors. This sinogram is then input to a reconstruction algorithm, which solves an under-determined inverse problem, described in Section A.3, to generate a human-understandable 2D or 3D image of the organ of interest. This image can then be used by the doctor for medical diagnosis.

A.1. X-Ray Attenuation

X-rays produced by CT scanners can interact with matter via the photoelectric effect, the Compton effect, and coherent scattering. Through these interactions, some of the emitted X-ray photons are absorbed or scattered when passing through different tissues in the body. The attenuation is described by the Beer-Lambert Law,

$$J = J_0 e^{-fL}, \tag{15}$$

where J and J_0 are the incident and transmitted X-ray intensities; L is the material thickness; and f is the linear attenuation coefficient of the material,

$$f = \tau_1 + \tau_2 + \tau_3, \tag{16}$$

with photoelectric (τ_1), Compton (τ_2), and coherent scattering (τ_3) attenuation coefficients. Attenuation coefficients for common materials in the body – iodine, bone, water, and soft-tissue – are plotted, in Fig. 7, over the range of incident X-ray energies used in CT imaging. It is clear that the attenuation of X-ray photons can be used to distinguish and, thus, image various tissues in the body (Hsieh, 2003).

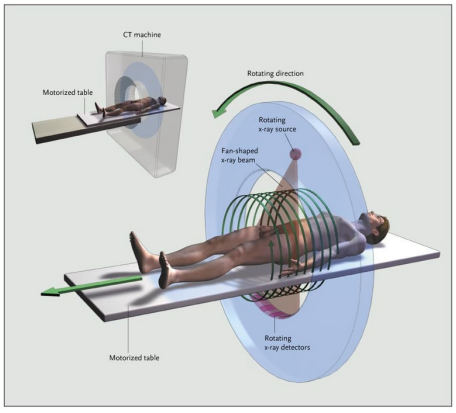

Figure 6. Illustration of a CT scanning device. Reproduced with permission from (Brenner & Hall, 2007), Copyright Massachusetts Medical Society.

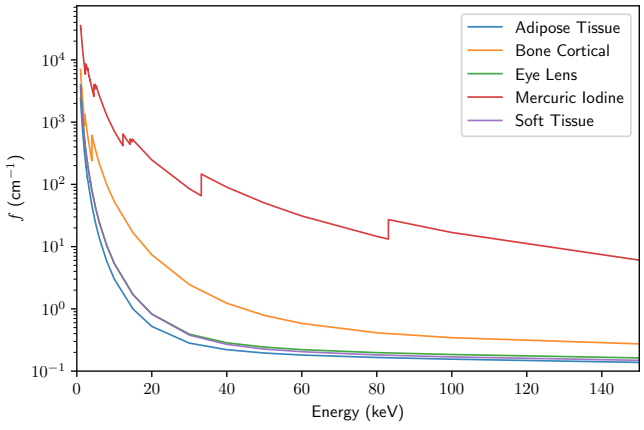

Figure 7. The attenuation coefficient (f) of different tissues found in the body, as a function of energy (keV). The plot was generated using X-ray mass attenuation coefficients (f/ρ) and densities (ρ) from the NIST Standard Reference Database (Hubbell & Seltzer, 2004; Berger et al., 2004).

A.2. CT Projection Measurements

As previously described, the CT scan relies on an X-ray generator which rotates around the patient, emitting X-ray photons. In this work, we only consider a restricted case of the spiral CT setup, in which there is no motion along the patient axis. Instead, we focus on reconstructing singular 2D image cross-sections and assume a parallel-beam geometry, in which photons are emitted and detected with the linear geometry of Figure 2.

We begin by considering the measurements of a single detector, measuring at angle ϕ. Assume that the X-ray generator outputs monoenergetic X-rays of intensity J_0. If the patient were simply a homogeneous block of tissue, with length $\Delta\ell$ and attenuation coefficient f^*, we could directly apply the Beer-Lambert law (Equation 15),

$$J = J_0 e^{-f^* \Delta\ell}, \tag{17}$$

to solve for the output attenuated X-ray intensity J. In reality, several blocks of tissue will be present in the patient, each with its own attenuation coefficient. However, since the exit X-ray flux from one block of tissue is the entrance X-ray flux to its neighboring block, we can simply apply the Beer-Lambert law in a cascading fashion over intervals of length $\Delta\ell$ and attenuation coefficients $(f_1^*, f_2^*, ..., f_n^*)$,

$$J = J_0 e^{-f_1^* \Delta\ell} e^{-f_2^* \Delta\ell} ... e^{-f_n^* \Delta\ell} = J_0 e^{-\sum_{i=1}^{n} f_i^* \Delta\ell}. \tag{18}$$

As $\Delta\ell \to 0$, the summation term becomes an integration over the length, L, of the patient,

$$J = J_0 e^{-\int_L f^*(\ell) d\ell}. \tag{19}$$

Finally, dividing both sides of the expression by J_0 and taking a negative logarithm, we define the **projection measurement** term,

$$S_\phi = -\ln\left(\frac{J}{J_0}\right) = \int_L f^*(\ell) d\ell. \tag{20}$$

As illustrated in Figure 2, a CT scanner with a parallel beam geometry contains several detectors side-by-side, collectively measuring a plane of attenuated X-ray photons. In this 2D imaging space, the projection measurement becomes a function of detector position, r. Thus, Equation 20 is re-expressed as the line integral,

$$S_\phi(r) = -\ln\left(\frac{J}{J_0}\right) = \int f^*(\phi, r) ds, \tag{21}$$

known as a **forward-projection (FP)**. It follows from the coordinate system of Figure 2 that, for measurements at angle ϕ, point (x, y) within the patient cross-section is projected onto detector position

$$r' = x\cos\phi + y\sin\phi. \tag{22}$$

Combining this with Equation 21, we derive the **Radon transform** (Deans, 2007) of the patient cross-section,

$$\begin{aligned} S_\phi(r) &= -\ln\left(\frac{J}{J_0}\right) \\ &= \int_{\mathcal{Y}} \int_{\mathcal{X}} f(x, y)\ \delta(x\cos\phi + y\sin\phi - r)\ dxdy, \end{aligned} \tag{23}$$

where δ is the Dirac delta function and $\mathcal{X} \times \mathcal{Y}$ is the set of image pixels (x, y). Sweeping over all the angles, these projective measurements are stacked to form a Radon transform image. As depicted in Figure 2, the projection measurements of the blue point across several angles produces a sinusoidal curve. The representation of all the CT scan measurements is thus known as a **sinogram**.

A.3. CT Image Reconstruction Problem

Sinograms are not human-interpretable. They depict the integrated attenuation coefficients, or projection measurements $(S_\phi(r))$, of the patient cross-section from several angles (ϕ) over all detector positions (r). Instead,

the desired outcome of a CT scan is a reconstructed image of the cross-section itself. This corresponds to the attenuation coefficient function, $f(x, y)$, which is the inverse of the Radon transform of Equation 23, or

$$f(x, y) = \frac{1}{2\pi} \int_0^{\pi} u_\phi(x \cos\phi + y \sin\phi)\, d\phi, \tag{24}$$

where u_ϕ is the derivative of the Hilbert transform of $S_\phi(r)$ (Helgason, 1984). The **Projection-Slice theorem** (Bracewell, 1956) ensures that f^* can be fully reconstructed with infinite measurement angles, ϕ. In practice, however, it is not possible to acquire infinite measurements. Typically, reconstruction quality improves with number of measurements, but this increases radiation exposure. In practice, hundreds of measurements are performed in a CT scan, but there is interest in reducing this number. In this work, we study algorithm performance in the very low measurement data regime, where uncertainty quantification over the value of $f(x, y)$ becomes especially important. The combination of limited and noisy real-world data renders the reconstruction of the desired image much more complex than simply evaluating the integral of Equation 24. Leveraging assumptions or data-driven insights about the measurements and physics at play, several statistical models have been developed and used to derive various image reconstruction algorithms, as discussed in the next section.

B. Classical Reconstruction Algorithms

In this appendix, we provide brief descriptions of the classical approaches implemented in this work via the TomoPy Astra (Pelt et al., 2016) software package. These algorithms are clinically approved and widely used in medical imaging, serving as a basis of comparison for the methods developed in this work. Note that although our approach used NNs, it was not a data-driven approach, but rather learned image functions from small amounts of data. Thus, we do not discuss data-driven techniques.

B.1. Filtered Back-Projection (FBP)

Filtered back-projection (FBP) (Pan et al., 2009) is an analytic algorithm which calculates a stable, discretized version of the inverse Radon transform. As the name implies, there are two key steps: filtering and back-projection.

The forward-projection of Equation 21 describes how X-rays passing through the object domain create a measurement. In back-projection (BP), this measurement is integrated back along the X-ray path across the object domain. This is done over all projection angles ϕ, using

$$f_{\mathrm{BP}}(x, y) = \int S_\phi(x \cos\phi + y \sin\phi)\, d\phi \tag{25}$$

to reconstruct the object attenuation coefficient image. As the number of projection angles increases, the image reconstruction improves. However, as shown in Figure 8, this back-projection is insufficient to guarantee a clear image. While information about the low frequencies of the object are captured in measurements at several view angles, that of high frequencies may only be captured in a few view-angles. Thus, the low frequencies are sampled far more densely than the higher frequencies, resulting in a blurry image. This can be corrected by suppressing the lower frequencies with filtering, by applying to each projective measurement, $S_\phi(r)$, the sequence of a Fourier transform (FFT), high-pass filter, and an inverse Fourier transform (iFFT). While several high-pass filters can be used, a popular choice is the Ram-Lak filter, which generates the filtered projective measurement

$$\tilde{S}_\phi(r) = \int \mathcal{F}[S_\phi](\omega)|\omega| e^{i2\pi\omega r} d\omega, \tag{26}$$

where $\mathcal{F}[S_\phi](\omega)$ is the Fourier transform of $S_\phi(r)$ and $|\omega|$ the frequency response of the filter. Performing back-projections of all the filtered projective measurements,

$$f_{\mathrm{FBP}}(x, y) = \int \tilde{S}_\phi(x \cos\phi + y \sin\phi)\, d\phi, \tag{27}$$

results in a sharper object attenuation coefficient image. Figure 8 visualizes the difference in reconstruction performance of BP and FBP for increasing measurement view angles, ϕ. Although the analytic FBP algorithm is fast and numerically stable, it suffers from poor resolution-noise trade-off.

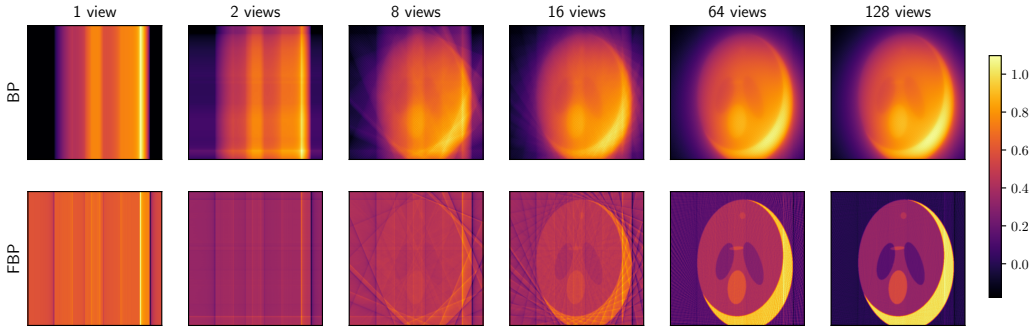

Figure 8. Comparison of the reconstruction quality, as a function of the number of views, of the BP (top) and FBP (bottom) algorithms.

B.2. Algebraic and Iterative Reconstruction

The reconstruction problem can be formulated as a system of linear equations

$$\mathbf{W}\vec{f} = \vec{S}, \tag{28}$$

where $\vec{S}$ is an $m \times 1$ vector of the m projective measurement values in the sinogram; $\vec{f}$ is an $n \times 1$ vector of the n attenuation coefficient pixel values in the reconstruction image; and $\mathbf{W}$ is a, typically sparse, $m \times n$ weight matrix representing the contribution of each of the m sinogram values to each of the n image pixel values. Given the vector $\vec{S}$, the goal is to solve for $\vec{f}$. If $\mathbf{W}$ were invertible, $\vec{f}$ would simply be $\mathbf{W}^{-1}\vec{S}$. However, because n is usually much larger than m, the system of equations of (28) is underconstrained. In algebraic reconstruction, this problem is addressed by using iterative algorithms that pose the reconstruction of $\vec{f}$ as the solution of a constrained optimization problem,

$$\vec{f}^{\,*} = \arg\min_{\vec{f}} ||\, \vec{S} - \mathbf{W}\vec{f}\,||, \text{ subject to } f_i \geq 0 \; \forall i. \tag{29}$$

Several families of iterative solvers can be used to solve this optimization, such as Landweber, Krylov subspaces, and expectation maximization (EM). The key benefit of iterative methods is that prior system knowledge can be integrated, via the cost function and initialization of $\mathbf{W}$. Their down-side is that they are not necessarily stable, may not converge, and are much slower than analytic techniques, such as FBP.

B.3. Simultaneous Iterative Reconstruction Technique (SIRT)

The simultaneous iterative reconstruction technique (SIRT) (Pryse et al., 1993; Bust & Mitchell, 2008) is a Landweber iterative method that updates the image reconstruction using all available sinogram projection data, $\vec{S}$, simultaneously. The optimization update at step k is defined as

$$\vec{f}^{\,(k+1)} = \vec{f}^{\,(k)} + \mathbf{B}\mathbf{W}^T\mathbf{D}\big(\,\vec{S} - \mathbf{W}\vec{f}^{\,(k)}\,\big), \tag{30}$$

where $\mathbf{D} \in \mathbb{D}^{m\times m}$ is a diagonal matrix containing the inverse row sums, $d_{ii} = (\sum_{j=0}^{n-1} w_{ij})^{-1}$, and $\mathbf{B} \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the inverse column sums, $b_{ii} = (\sum_{i=0}^{m-1} w_{ij})^{-1}$. The weighted projection difference, $\mathbf{D}\big(\,\vec{S} - \mathbf{W}\vec{f}^{\,(k)}\,\big)$, corresponds to the inverse of the length each X-rays passes through the volume. Shorter rays have a higher contribution, with the weighting required to guarantee convergence. This difference is forward-passed back to the image domain, using the weighted back-projection term, $\mathbf{B}\mathbf{W}^T$, where it can be used to update the reconstruction. These updates iteratively solve the problem

$$\vec{f}^{\,*} = \arg\min_{\vec{f}} ||\, \vec{S} - \mathbf{W}\vec{f}\,||_{\mathbf{D}} = \arg\min_{\vec{f}} \big(\,\vec{S} - \mathbf{W}\vec{f}\big)^T\mathbf{D}\,\big(\,\vec{S} - \mathbf{W}\vec{f}\big), \tag{31}$$

converging to a weighted least-squares solution, with weights given by the inverse row sums of $\mathbf{W}$.

B.4. Simultaneous Algebraic Reconstruction Technique (SART)

The algebraic reconstruction technique (ART) (Gordon et al., 1970) was one of the first proposed algebraic iterative algorithms for CT image reconstruction. It is a Landweber technique almost identical to the SIRT algorithm. However, a single projective measurement is used to update the reconstruction image per update step. Generally, the ART algorithm reaches a solution much faster than SIRT, but does not have stable convergence if the system of equations is inconsistent, for example due to measurement noise.

The simultaneous algebraic reconstruction technique (SART) (Andersen & Kak, 1984) was proposed in 1984, as an improvement to ART, and is also a Landweber algebraic iterative algorithm. It combines the reduced runtimes of ART with the improved convergence of SIRT, by using all the projective measurements from a single view angle in each optimization iteration. The update of image vector index i at step n is defined as

$$f_i^{(n+1)} = f_i^{(n)} + \frac{\lambda_n}{\sum_{j=0}^{n-1} w_{ij}} \sum_{j=\phi_n L+1}^{\phi_n L+L} w_{ij} \frac{S_j - \hat{S}_j}{\sum_{g=0}^{m-1} w_{gj}}, \tag{32}$$

where $\lambda_n << 1$ is a, potentially dynamic, relaxation parameter; ϕ_n is the $(n \mod N)^{\text{th}}$ measurement angle of the sinogram, assuming N total measurement angles; and L is the number of projective measurements taken at each angle. SART typically converges to a good reconstruction within a few iterations.

B.5. Conjugate Gradient Least Squares (CGLS)

The conjugate gradient least squares (CGLS) (Yuan & Iusem, 1996) algorithm is a Krylov subspace iterative method. Since it requires a positive-definite system matrix, the CT image reconstruction problem is reformulated in terms of the set of normal equations

$$\mathbf{W}^T\mathbf{W}\vec{f} = \mathbf{W}^T\vec{S}. \tag{33}$$

Due to the positive-definiteness of $\mathbf{W}^T\mathbf{W}$, there exists a set of conjugate normal vectors $\mathcal{Q} = \{\vec{q}_1, ..., \vec{q}_n\}$, where $\vec{q}_i^T\mathbf{W}^T\mathbf{W}\vec{q}_j = 0$, $\forall\, i \neq j \in (1, n)$. Since $\mathcal{Q}$ forms a basis for $\mathbb{R}^n$, the image vector $\vec{f}$ can be rexpressed as a linear combination of these conjugate normal vectors,

$$\vec{f} = \sum_{i=1}^{n} \alpha_i \vec{q}_i. \tag{34}$$

Thus, solving for $\vec{f}$ becomes a problem of solving for the conjugate normal basis vector directions, $\vec{q}_i$, and their corresponding weights, α_i. This can be achieved iteratively by expressing the problem as a quadratic least-squares minimization of the function

$$L(\vec{f}) = \frac{1}{2}\vec{f}^T\mathbf{W}^T\mathbf{W}\vec{f} - \vec{f}^T\mathbf{W}^T\vec{S}, \tag{35}$$

which has gradient $\nabla L(\vec{f}) = \mathbf{W}^T\mathbf{W}\vec{f} - \mathbf{W}^T\vec{S}$ and a guaranteed unique minimizer because the Hessian $\nabla^2 L(\vec{f}) = \mathbf{W}^T\mathbf{W}$ is symmetric positive-definite. The name conjugate gradient least squares comes from the fact that, in each iteration, a conjugate basis vector and its weight are found by taking a gradient step in the direction that minimizes the least-squares function, $L(\vec{f})$, as

$$\vec{f}^{(k+1)} = \vec{f}^{(k)} + \alpha_k\vec{q}_k \tag{36}$$

$$\vec{e}_k = \mathbf{W}^T\vec{S} - \mathbf{W}^T W\vec{f}^{(k)} \tag{37}$$

$$\vec{q}_k = \vec{e}_k - \sum_{i<k} \frac{\vec{q}_i^T\mathbf{W}^T\mathbf{W}\vec{e}_k}{\vec{q}_i^T\mathbf{W}^T\mathbf{W}\vec{q}_i}\vec{q}_i \tag{38}$$

$$\alpha_k = \frac{\vec{q}_k^T\vec{e}_k}{\vec{q}_k^T\mathbf{W}^T\mathbf{W}\vec{q}_k} \tag{39}$$

where $\vec{e}_k$ is the residual at step k. Thus, the main difference between SIRT/SART and CGLS is that the search direction in SIRT/SART is determined only by the projection difference at that point, while in CGLS the search directions of all the previous iterations are also taken into account. CGLS typically converges much faster than SIRT, but has a large memory footprint.

B.6. Expectation Maximization (EM)

The final classical approach to CT image reconstruction that we consider is a statistical iterative method known as expectation maximization (EM) (Dong, 2007). This technique explicitly encodes prior knowledge about the X-ray physics at hand. Each projective measurement, S_j is modeled as a Poisson distribution,

$$S_j \sim \mathcal{S}_j = \text{Poisson}(\lambda_j) = \frac{\lambda_j^{S_j}e^{-\lambda_j}}{p_j!}, \tag{40}$$

where the distribution mean $\lambda_j = \mathbb{E}[\mathcal{S}_j]$ is the function

$$\lambda_j = \sum_i w_{ij}f_i \tag{41}$$

of the probability w_{ij} that an X-ray photon penetrating image pixel i was measured at detector location j; and the underlying attenuation coefficient function f to reconstruct.

The measurement sinogram is modeled as the likelihood

$$p(\vec{S}|\vec{f}) = \prod_j \frac{\lambda_j^{S_j}e^{-\lambda_j}}{S_j!} = \prod_j \frac{(\sum_i w_{ij}f_i)^{S_j}e^{-(\sum_i w_{ij}f_i)}}{S_j!}. \tag{42}$$

The EM algorithm computes the maximum likelihood estimate of f,

$$\hat{f}_{\text{MLE}} = \underset{f}{\text{argmax}} \left\{ \log\left(p(S|f)\right) \right\}, \tag{43}$$

by alternating between expectation and maximization steps. These can be combined into the update-step

$$\hat{f}_i^{(k+1)} = \frac{\hat{f}_i^{(k)}}{\sum_j w_{ij}} \sum_j \frac{w_{ij} S_j}{\sum_i w_{ij} \hat{f}_i^{(k)}}. \tag{44}$$

The EM algorithm is computationally intensive, but guaranteed to converge to a local optimum of the likelihood. Further, although a Poisson distribution was assumed for S_j in this discussion, further knowledge of the detector noise can be easily incorporated into the model.

B.7. Performance of Classical Reconstruction Techniques on Shepp-Logan

Finally, we evaluate the performance of these classical reconstruction algorithms on the Shepp-Logan validation set, depicted in Figure 13. Figure 9 shows average PSNR (for the 5 validation images) as a function of view angle, for each of the 5 reconstruction methods. Also shown are the values below which PSNR is usually considered unacceptable (20dB), at which reconstruction is considered lossy (30dB), and above which has high quality (40dB). FBP has the worst performance, which is particularly poor in the low-view regime (< 30 views). The iterative reconstruction algorithms perform better. CGLS has slow convergence to high quality image reconstruction. EM, SIRT, and SART all converge with far fewer views, achieving lossy compression with only $\sim$20 views. EM levels out and requires around 180 views to achieve high-quality reconstruction. SIRT and SART have near identical performance, passing the high-quality reconstruction threshold with only $\sim$75 views and SIRT performing slightly better for larger view numbers.

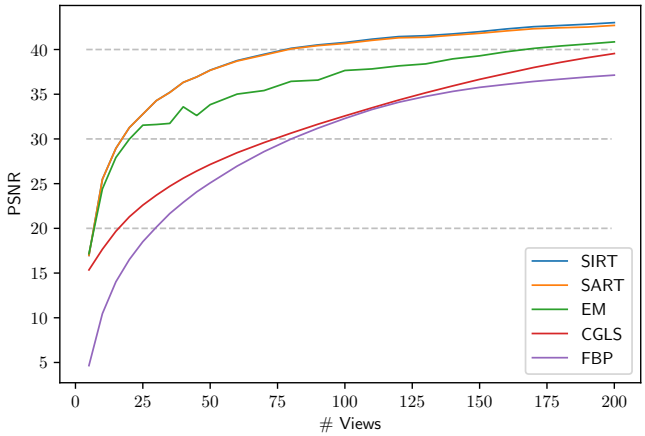

Figure 9. PSNR as a function of number of views for the classical reconstruction algorithms.

Table 3. Classical reconstruction PSNR, averaged across the five validation set images, in the 5-, 20-, and 180-view cases. The best achieved PSNR for each view-# is bolded.

# Views	FBP	CGLS	EM	SART	SIRT
5	7.68	16.38	**21.39**	21.12	21.12
20	17.35	21.85	30.22	**31.98**	31.97
180	36.74	38.6	40.46	42.51	**42.76**

Table 3 reports the PSNR values obtained for 5, 20, and 180 views. EM is the best performer for 5 views, SART for 20, and SIRT for 180. However, while the performance of the three algorithms is similar for 5, SART and SIRT perform better than EM for larger number of views. In the low-view regime, roughly an order of magnitude is required to achieve a 10dB improvement in average PSNR. Figure 10 shows a reconstructed image for each of

these algorithm-view combinations. With 5 views the reconstruction algorithms are able to capture low-frequency object structure, but the image would not be useful for medical diagnosis. With 20 views it is clear that the algorithms are already capturing high-frequency components of the object, but the images have many artifacts. By 180 views, the reconstructed images are nearly identical to the ground truth images of Figure 13, with any discrepancies in PSNR due mostly to minor image reconstruction artifacts or imprecisions.

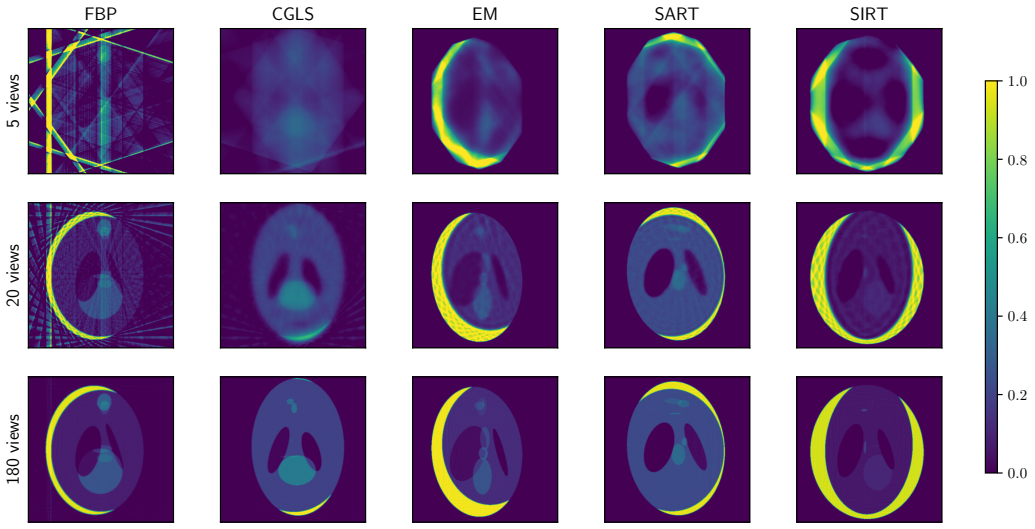

Figure 10. Reconstructions, generated with varying view angles, by the 5 classical reconstruction algorithms on the Shepp-Logan validation set.

C. Metrics and Uncertainty

C.1. PSNR & SNR

PSNR is defined as

$$\mathrm{PSNR}\big(f^*, f\big) = 10\log_{10}\left(\frac{\max(f^*)^2}{\mathrm{MSE}(f^*, f)}\right), \quad \text{where} \quad \mathrm{MSE}\big(f^*, f\big) = \frac{1}{|\mathcal{X}\times\mathcal{Y}|}\sum_{x,y}\big(f^*(x,y) - f(x,y)\big)^2 \tag{45}$$

while SNR is defined as

$$\mathrm{SNR}\big(f^*, f\big) = 20\log_{10}\left(\frac{||f^*||_2}{||f^* - f||_2}\right), \tag{46}$$

where f^* denotes the ground truth image, f the noisy/reconstructed image, and $||\cdot||_2$ the ℓ^2-norm. SNR is strictly less than PSNR, with higher SNR and PSNR corresponding to better image reconstruction. In the absence of any noise, f^* and f are identical,making SNR and PSNR infinite. For lossy images, PSNR is typically between 30-50dB, with values over 40dB considered very good, and values below 20dB considered unacceptable (Bull, 2014).

C.2. Types of Uncertainty

Bayesian modeling can address two distinct types of uncertainty: aleatoric and epistemic (Der Kiureghian & Ditlevsen, 2009; Kendall & Gal, 2017). **Aleatoric uncertainty** is due to *measurement noise*, such as X-ray detector noise. This type of uncertainty cannot be reduced, even if more measurements are taken, since it is inherent to the measurement. To see why, think of rolling an unbiased die. Irrespective of how many times you roll the die, you will always be uncertain of the outcome of the next roll, since each outcome has a $\frac{1}{6}^{\text{th}}$ probability. On the other hand, **epistemic or model uncertainty** accounts for *uncertainty in the model parameters*. This type of uncertainty can be reduced with more measurement data. To see why, imagine a model that aims to predict the outcome of a biased die roll, with no prior information about the bias. As more data is taken, the variance in the model parameters decreases, and the model output distribution better approximates the true biased die distribution. Even in the presence of aleatoric uncertainty, this work primarily focuses on quantifying epistemic/model uncertainty. Namely, we consider how well the model reconstructs the ground truth attenuation coefficient image from the sinogram data.

C.3. Calibration and Coverage

Calibration is a metric that assesses a model's ability to predict the probabilities of its outcomes, gauging the reliability of the model's confidence in its predictions. For example, a model performing class predictions is considered calibrated if it assigns a class 50% probability and that class actually appears 50% of the time in prediction. For further information on calibration of class prediction models, we refer the reader to (Guo et al., 2017a; Nixon et al., 2019). Since this work focuses on regression models, the remaining discussion is centered on calibrated regression (Kuleshov et al., 2018).

Let F be the cumulative distribution function (CDF) of model predictions $f(x,y)$, that seek to approximate ground truth image $f^* \in \mathcal{F}$, where $\mathcal{F}$ denotes the functional space of possible images. Letting $f_x := f(x,y)$ We denote the corresponding quantile function as

$$F_x^{-1}(\tilde{p}) = \inf\{f_x : \tilde{p} \leq F(f_x)\}, \tag{47}$$

where F^{-1} performs the mapping $F^{-1} : [0,1] \to \mathcal{F}$ and $\tilde{p}$ is a confidence interval. For calibrated regression, ground truth pixel $f(x,y)$ should fall in a, say, 90% confidence interval 90% of the time. Thus, the regression model is calibrated for confidence interval $\tilde{p}$ if

$$\lim_{X\to\infty}\frac{1}{X}\sum_{x=1}^{T}\mathbb{I}\big\{f_x \leq F_x^{-1}(\tilde{p})\big\} \;=\; \tilde{p}\;, \quad \forall\tilde{p}\in[0,1], \tag{48}$$

as the number of pixel samples approaches infinity, $X \to \infty$. If f_X^* denotes the ground truth value for i.i.d. random pixel $(x,y) \in X$, a sufficient condition for calibrated regression is

$$p\big(f_X^* \leq F_X^{-1}(\tilde{p})\big) = \tilde{p}\;, \;\forall\,\tilde{p}\in[0,1]. \tag{49}$$

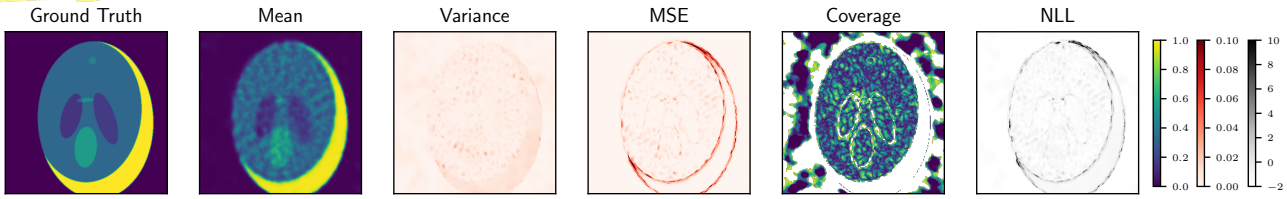

Figure 11. The ground truth image is plotted alongside different metrics for assessing the reconstructed predicted distribution generated by an UINR model. From left to right, we show the ground truth, predicted mean image, predicted variance image, mean squared error, coverage quantile of each pixel, and negative log-likelihood of each pixel. Note that PSNR is calculated using the ground truth and predicted mean image. In a real medical setting, the ground truth is unknown, the doctor would be given the predicted mean image and the predicted variance image could be provided as supplementary information to help the doctor reach a diagnosis. Further note that white regions in the coverage image denote that the ground truth pixel value did not fall in the range of the predicted distribution.

Since practical dataset sizes are finite, preventing perfect calibration, different metrics have been developed to assess empirical model calibration.

Reliability diagrams serve as a visual representation of model calibration, plotting expected sample accuracy as a function of average model confidence. Ideally these would be continuous plots, but, in practice, samples are binned into M bins according to their prediction confidence. Let B_m be the set of indices, i, of samples with prediction confidence, $\hat{p}_i$ in the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The expected accuracy (acc) and confidence (conf) are approximations to the terms of (49), namely

$$p\big(f_X^* \leq F_X^{-1}(\tilde{p})\big) \approx \mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{I}\big\{f_i^* \leq F_i^{-1}(\tilde{p})\big\} \tag{50}$$

$$\tilde{p} \approx \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i. \tag{51}$$

The **calibration error (CE)** is the discrepancy

$$\mathrm{CE}(\tilde{p}) = |\, p\big(f_X^* \leq F_X^{-1}(\tilde{p})\big) - \tilde{p} \,| \approx |\, \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \,| = \mathrm{CE}(B_m). \tag{52}$$

It can be measured on a reliability diagram as the difference between the expected accuracy curve and the ideal $\mathrm{acc}(B_m) = \mathrm{conf}(B_m)$ line. The **expected calibration error (ECE)** quantifies the calibration error of the full distribution as

$$\mathrm{ECE}(f^*, F_X^{-1}, \tilde{p}) = \frac{1}{M} \sum_{m=1}^{M} |\, \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \,|. \tag{53}$$

The model is considered calibrated if $\mathrm{ECE}(x, f) = 0$.

In practice, modifications were made to the previously described theory of reliability curves and ECE. You may notice in Figure 12 that, instead of plotting *accuracy* and *confidence*, we instead plot analogous *target coverage* and *achieved coverage*. Typically, a **coverage** value, $\bar{p}$, refers to a quantile of data points lying within $\pm\frac{\bar{q}}{2}\%$ of the median (50% quantile). Specifically, in our setup, we use different uncertainty quantification methods (BBB, MCD, and DE) to sample N different model weights for our INR, each set of weights corresponding to a different model output. Given that each output corresponds to an image, for each pixel, (x, y), we have a distribution of N predicted values, $F_N(x, y)$. Ideally, the median of the pixel distribution would be equivalent to the ground truth pixel value, $f^*(x, y)$. However, this is unrealistic to expect in practice. Thus, we check whether the ground truth pixel lies within the predicted pixel distribution quantile, Q_n, specified by coverage value, $\bar{p}$,

$$Q_{50-\frac{\bar{p}}{2}}\big(F_N(x, y)\big) \leq f^*(x, y) \leq Q_{50+\frac{\bar{p}}{2}}\big(F_N(x, y)\big). \tag{54}$$

If a model is perfectly calibrated, $\bar{p}\%$ of reconstructed pixels distributions will contain the ground truth pixel in their $\bar{p}\%$-th quantile, corresponding to the grey dashed line in Figure 12. Thus, model confidence can be seen as a pre-selected quantile for each pixel (target coverage), p, while accuracy is the percentage of pixel distributions

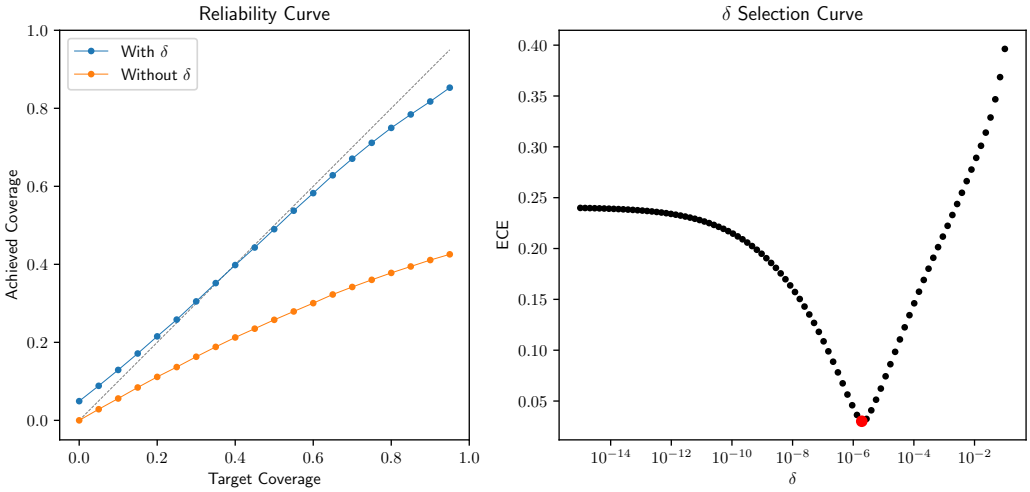

Figure 12. Both plots were made with an MCD UINR model trained on 20 views, achieving a PSNR of 19. **Left)** A plot of the model reliability curves, with the grey dashed line indicating a perfectly calibrated model. The blue curve is the empirical reliability curve of the model when a small δ term is added symmetrically to the target coverage, in order to slightly widen the quantile ranges. Although this δ term has nearly negligible magnitude, it significantly improves the model reliability curve, as illustrated by the orange curve of reliability without the added δ term. **Right)** The added δ term was not chosen arbitrarily, but selected to minimize ECE.

containing the ground truth that quantile (achieved coverage),

$$\mathrm{AC}(f^*, F_N, \bar{q}) = \frac{1}{|\mathcal{X}|}\frac{1}{|\mathcal{Y}|}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathbb{I}\Big\{Q_{50-\frac{\bar{q}}{2}}\big(F_N(x,y)\big) \le f^*(x,y) \le Q_{50+\frac{\bar{q}}{2}}\big(F_N(x,y)\big)\Big\}. \tag{55}$$

The ECE is thus implemented as,

$$\mathrm{ECE}(f^*, F_N) = \frac{1}{|\mathcal{P}|}\sum_{\bar{p}\in\mathcal{P}}|\mathrm{AC}(f^*, F_N, \bar{p}) - \bar{p}|, \tag{56}$$

where $\mathcal{P}$ is a finite set of percentages evenly spaced in $[0,1]$, separated by percentage interval $i << 1$. The fifth image of Fig. 11, plots the smallest quantile of each pixel containing the corresponding ground truth pixel, for an example UINR reconstruction with $N = 50$. Note that white regions indicate that the ground truth value does not fall within the minimum and maximum predicted pixel values.

There is one final modification made in implementing the reliability curves, in order to effectively assess model calibration. Since the final layer of all the NNs used for the INR have a sigmoid activation, ensuring that the model output is in the range $(0, 1)$. However, the sigmoid function only approaches 0 and 1 in the infinite limit, meaning that in practice our model will never output 0 or 1 exactly. However, our images contain a large percentage of pixels with exactly 0 value, especially for noiseless artificial data, which in the context of medical imaging is regions containing air and no tissue. This is problematic for calibration, since all of predicted pixel values will be near-zero, but will not actually contain the ground truth value of 0. This is illustrated by the orange reliability curve in Figure 12, for which only 40% of pixels contain the ground truth in their full range of predicted pixel values, for an MCD model trained on 20 views with $N = 50$. Our proposed solution to this issue is slightly widening the quantile range by adding a negligible δ term. Thus, for coverage value $\bar{p}$, we now check if the ground truth pixel lies in,

$$Q_{50-\frac{\bar{q}}{2}}\big(F_N(x,y)\big) - \delta \le f^*(x,y) \le Q_{50+\frac{\bar{q}}{2}}\big(F_N(x,y)\big) + \delta, \tag{57}$$

where $0 < \delta << 1$. In this case, if our predicted pixel values are slightly larger than 0, the δ offset can widen the quantile range to include 0, enabling these pixels to contribute to the achieved calibration (this also applies to pixels with exact value of 1). The improvement in using a delta offset is illustrated by the blue reliability curve in Figure 12, which is much closer to the ideal grey dashed line than the orange curve with δ. It should be noted

that the value of δ is not assigned arbitrarily, but instead optimized to minimize overall ECE. For too small a δ, the quantiles will not be widened sufficiently to capture ground truth 0 pixels. However, for too large of δ, the quantiles will be widened too much, reducing overall calibration, as achieved coverage is much higher than target coverage for low coverage values. Thus, ECE as a function of δ is expected to have a unique minima, as illustrated by the example in Figure 12.

C.4. Assessing Model Quality

Sec. 3.2 describes how PSNR and SNR quantify image reconstruction quality, coverage metrics (such as ECE) gauge the uncertainty calibration, and NLL encapsulates both. In this work, we aim to optimize both reconstruction and calibration quality, meaning the best metric would, naively, be NLL. However, there is often a trade-off between calibration and prediction quality. Specfically, NN overfitting to NLL manifests in probabilistic error rather than prediction error (Guo et al., 2017b). Furthermore, while this work focuses on uncertainty quantification of INRs for medical imaging, little prior work has addressed this problem. Most existing techniques only quantify reconstruction PSNR and SNR. Hence, for fair comparison, we assess and optimize our models primarily according to PSNR and SNR. However, for similarly performing models, we use NLL and ECE as secondary selectors for the best model. Note that initial attempts at optimizing models according to coverage metrics resulted in preference for the lowest capacity models possible. This indicates that optimal performance according to coverage favors blurry image reconstruction, with as little certainty as possible in the final image.

D. Bayesian Deep-Learning Implementation Notes

Building off of the descriptions of the different BDL methods discussed in Section 3.1, we provide more insight about implementation specifics.

D.1. Bayes-by-Backprop Implementation

In this work, the BBB variational posterior is treated as a Gaussian distribution, $\mathcal{N}(\mu_\psi, \sigma_\psi)$. The elements of σ_ψ comprise a diagonal covariance matrix, meaning weights are assumed to be uncorrelated. A Gaussian prior, $p(\theta) = \mathcal{N}(\theta|\sigma^2)$, with tunable σ is used to initialize the network. Training the network requires computing a forward-pass and backward-pass. Although the network is parameterized by a distribution of weights, in each forward pass a single sample is drawn from the variational posterior and propagated through the network to perform updates. A re-parameterization trick (Kingma & Welling, 2014), in which the sample ϵ is transformed by the function $\mu_\psi + \sigma_\psi \odot \epsilon$, is used to ensure a gradient can be calculated for backpropagation. Finally, to aid learning, it is common to modify the ELBO as

$$\tilde{\mathcal{L}}(y, \psi) = \mathbb{E}_{q(\theta|\psi)}[\log p(y|\theta)] - \xi \cdot \mathbb{KL}\left[\, q(\theta|\psi) \,||\, p(\theta) \,\right], \tag{58}$$

where $\xi{>}0$ is an added hyperparameter, known as the Kullback-Leibler (KL) factor. This is beneficial for training because it puts greater emphasis on the training data in the loss, through the $\mathbb{E}_{q(\theta|\psi)}[\log p(y|\theta)]$ In the context of medical imaging with INRs, this reweights the importance of obtaining a Radon transform of network outputs close to the sinogram measurement data.

D.2. Deep Ensembles Implementation

Typically, DEs induce randomness by training the same network several times with randomized initializations and data order. However, recent work (Zaidi et al., 2020) has shown that ensembling over architectures can outperform the more common single-architecture DEs for uncertainty estimation. For our large-scale hyperparameter study (on Shepp-Logan phantom data), in which thousands of BBB and MCD UncertaINR base learners were trained, we were able to easily implement architecture-ensembled DEs, by selecting the best-performing UncertaINRs as base learners. However, for the high-resolution AAPM-Mayo data, in which UncertaINR training times were extended significantly, we found it too computationally demanding to hypertune multiple architectures to ensemble. Thus, for the final results presented, DEs are created by ensembling the same network trained with randomized weight initialization.

D.3. Hardware and Software Notes

The experiments presented in this paper were computationally intensive, requiring hundreds of compute hours on parallelized GPUs. Specifically, they were run on a cluster of 4 GPU nodes consisting of 8 GPUs each, containing a mixture of GTX 1080, GTX 1080Ti, and GeForce RTX 2080 Ti cards. The project codebase was developed in Python, using Pytorch (Paszke et al., 2019), Hydra (Yadan, 2019), and Weights & Biases (Biewald, 2020) to implement the NN functionality and Blitz (Esposito, 2020) for BNN functionality.

E. Preliminary Ablation Study

Our preliminary ablation study presents a large-scale study of uncertainty quantification for INRs. Specifically, we test the relative ability of UncertaINR (using MCD, BBB, and/or DEs) to reconstruct artificial noiseless CT brain images. From this study, we present guiding principles for effective UncertaINR hyperparameter selection and compare to traditional CT reconstruction techniques.

E.1. Dataset

Given the long INR training times required to reconstruct large, high-frequency images (Yu et al., 2021), we opted for a dataset of simple images, enabling large-scale hyperparameter sweeps. Specifically, the Shepp-Logan phantom (Shepp & Logan, 1974) approach was used to generate (256×256 pixel) artificial brain images, with corresponding measurement sinograms generated via the Radon transform. No noise was added to the measurement sinograms. In all, 10 ground truth images, depicted in Figure 13, and 20 corresponding sinograms were generated: 5 validation and 5 test set sinograms each for the 5- and 20-view (ϕ) cases.

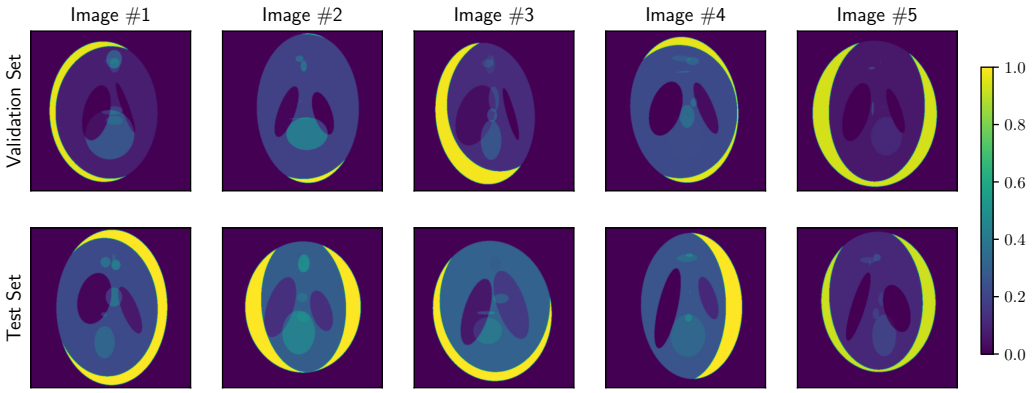

Figure 13. The ground truth images used in tuning and assessing our preliminary Shepp-Logan hyperparamter exploration. The five validation set images were used to optimized model hyperparameters, while test set images were used to assess the finalized models.

E.2. Baselines

In this study, we compared UncertaINR to the classical medical image reconstruction algorithms – FBP, CGLS, EM, SART, and SIRT – described in Section 2.2 and Appendix B, implemented using the TomoPy Astra wrapper (Pelt et al., 2016). A detailed analysis of the classical reconstruction methods on the Shepp-Logan validation set is presented in Appendix B.7.

E.3. Tuneable INR Model Parameters

There are several degrees of freedom in designing an INR, including its size, embeddings, activation functions, and optimizer. These design choices are critical in determining model performance, but there is little theoretical understanding of how to best select most of these model parameters. In this section, we lay out the different parameters we considered in designing our INRs and provide any known insights as to how they affect model performance. These insights informed the hyperparameter sweeps described in the following section.

E.3.1. Width and Depth

The size of an NN is determined by both its width (number of nodes per layer) and depth (number of layers). Universal approximation theorems (Hornik et al., 1989) have been derived in both the arbitrary-width (Cybenko, 1989; Hornik, 1991) and arbitrary-depth (Lu et al., 2017a; Hanin, 2019; Kidger & Lyons, 2020) cases, demonstrating that NNs are theoretically guaranteed universal function approximators in the infinite limit. In practice, however, neural networks have finite width and depth. Recent work has empirically demonstrated and theoretically suggested that, in this regime, increased-depth networks generally perform better than increased-width networks (Lu et al.,

2017b; Raghu et al., 2017). It is also known that, while neural networks are overparametrized relative to the amount of training data, this overparametrization is key for their generalization ability (Neyshabur et al., 2019; Jacot et al., 2021). However, for INRs specifically, it has been shown that relatively small networks can typically be used to learn decent functional image encodings (Dupont et al., 2021). Thus, we tend to sweep over smaller widths and depths than standard deep-learning networks.

E.3.2. Fourier Feature Mappings

Random Fourier features (RFF) were first introduced in 2007 by (Rahimi & Recht, 2007) as a means of accelerating kernel methods. The key idea is to map the input data to a randomized low-dimensional feature space, while maintaining the kernel of the original data. Given input $\vec{x} \in \mathbb{R}^n$, the RFF mapping takes the form

$$\gamma_{\text{RFF}}(\vec{x}) = [\cos(2\pi B\vec{x}), \sin(2\pi B\vec{x})]^T, \tag{59}$$

where B is an $m \times n$ matrix, with each entry sampled from $\mathcal{N}(0, \Omega_0^2)$. The standard deviation, Ω_0, is a tuneable hyperparameter, but remains static after initialization – i.e. it is not modified with NN weights during the MLP training. There exist other types of Fourier feature mappings, such as positional encodings, in which

$$\gamma_{\text{PE}}(\vec{x}) = [..., \cos(2\pi\Omega_0^{j/m}\vec{x}), \sin(2\pi\Omega_0^{j/m}\vec{x}), ...]^T, \tag{60}$$

for $j = 0, ..., m-1$.

In 2018, it was theoretically demonstrated that NNs can be approximated by kernel regression via the **neural tangent kernel (NTK)** (Jacot et al., 2021). Using this intuition, in 2020, it was argued that applying a simple Fourier feature mapping to input data enables MLPs to learn high-dimensional functions rapidly, even in low-dimensional problem domains (Tancik et al., 2020), making the technique particularly well-suited for INRs. In fact, positional encodings have been shown to have key importance in the success of NeRF (Mildenhall et al., 2020) and Fourier feature mappings have been shown to boost the performance of the CoIL network for medical image reconstruction (Sun et al., 2021). In this work, all networks apply an RFF mapping, γ_{RFF}, to the input data. The standard deviation, Ω_0, is tuned among other hyperparameters.

E.3.3. Activation Functions

Activation functions are key to the success of neural networks, transforming what would otherwise be simple linear systems into complex, non-linear universal function representers. We performed hyperparameter sweeps with five activations widely used in the MLP and INR literature – ReLU, SiLU, Sine, SoftPlus, and Tanh. We now briefly review these activation functions, as well as their use in deep learning.

The **rectified linear unit (ReLU)**, plotted in blue in Figure 14, was introduced as early as the 1960s for visual feature extraction in hierarchical NNs (Hara et al., 2015). The ReLU is defined as

$$\text{ReLU}(x) = \max\{0, x\}\ , \tag{61}$$

returning its input if greater than zero and otherwise returning zero. Despite its hard non-linearity at zero, non-differentiability at zero, and vanishing gradient challenge (Lu et al., 2019), the ReLU was shown in 2011 to enable better training than previously used activation functions, such as Sigmoid and Tanh, by inducing sparse representations (Glorot et al., 2011). As of 2017, the ReLU was the most popular activation function for deep NNs (Ramachandran et al., 2017).

The **sigmoid-weighted linear unit (SiLU)**, plotted in orange in Figure 14, is a specific instance of the Swish activation function family and was proposed in 2017 as a continuous, 'undershooting' version of the ReLU (Ramachandran et al., 2017). The Swish family, parameterized by β, is defined as

$$\text{Swish}_\beta(x) = x \cdot \sigma(\beta x) = \frac{x}{1 + e^{-\beta x}}\ , \tag{62}$$

where $\sigma(x)$ is the sigmoid function. By setting β to different values in $[0, \infty)$, Swish_β non-linearly interpolates smooth functions between the linear function and ReLU. In 2017, Swish was empirically shown to outperform ReLU, a result theoretically attributed to its bounded, smooth, and non-monotonic nature (Ramachandran et al.,

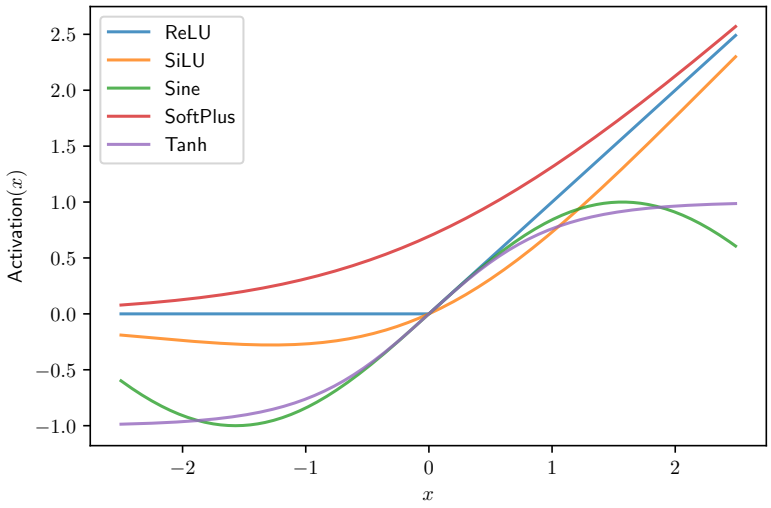

Figure 14. The five different activation functions – ReLU, SiLU, Sine, SoftPlus, and Tanh – tested in our preliminary hypertuning experiment.

2017). More recently, Swish has been shown to outperform both ReLU and Sine in the context of CT image reconstruction via Automatic Integration (AutoInt) (Lindell et al., 2021). The SiLU is the specific instance of Swish where $\beta = 1$,

$$\text{SiLU}(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}} \,. \tag{63}$$

The **Sine** activation function, plotted in green in Figure 14, is the sinusoid

$$\text{Sine}_{\omega_0} = \sin(\omega_0 \cdot x). \tag{64}$$

In the 2020 SIREN paper (Sitzmann et al., 2020), INRs with sinusoidal activation functions and random Fourier features were empirically demonstrated to outperform ReLU-based INRs. Theoretically, it was argued that these periodic activations are better suited to capturing naturally complex signals and their derivatives. However, the performance of these activations depends strongly on the choice of frequency, ω_0, which needs to be tuned.

The **SoftPlus** activation function, plotted in red in Figure 14, has continuous and differentiable form

$$\text{Softplus}(x) = \ln(1 + e^{x}). \tag{65}$$

It was introduced in 2001 (Dugas et al., 2001) as the primitive of the sigmoid function. It is primarily used as a smooth approximation to the ReLU activation and to constrain to positive outputs, since $\text{Softplus}(x) \in (0, \infty)$.

The hyperbolic tangent **Tanh**, plotted in purple in Figure 14, has form

$$\text{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \tag{66}$$

and is both differentiable and monotonic. It has a form similar to the sigmoid function, with $\text{Tanh}(x) = 2\sigma(2x) - 1$, but lies in the range (-1,1) instead of (0,1), meaning it does not constrain to positive values. Before the ReLU became popular, the sigmoid and Tanh were two of the most common activation functions. Tanh was easier to train and typically outperformed the sigmoid as an activation function. However, because these sigmoidal activation functions saturate for large inputs, their derivatives vanish for these inputs, leading to slow convergence of learning algorithms. This has motivated the increased use of ReLU-like activation functions (Goodfellow et al., 2016), which ameliorate the vanishing derivative problem.

E.4. Experimental Design

The goal of this hyperparameter study was to understand the relative performance of UncertaINR with BBB, MCD, and DEs. In all these cases, well-tuned hyperparameters were needed to achieve decent model reconstruction

accuracy and uncertainty calibration. Large hyperparameter sweeps were used to find the optimal parameters for BBB and MCD. The best performing MCD UncertaINRs were used as base learners for the DEs.

E.4.1. BBB & MCD

For both BBB and MCD, MLP design choices had a large effect on INR reconstruction performance. Carefully designed hyperparameter sweeps were thus run to strategically search the MLP parameter space for four different settings: (1) MCD 5-view, (2) MCD 20-view, (3) BBB 5-view, and (4) BBB 20-view.

The model selection process began with a coarse grid search to efficiently prune across the wide range of possible parameter combinations. Specifically, we considered model activation type, depth, width, RFF embedding frequency Ω_0, and dropout probability. Among these, activation type was the only categorical parameter, using the five activation types plotted in Fig. 14: Tanh, SoftPlus, Sine, SiLU, and ReLU. For the remaining parameters, this initial coarse grid search was used to get a sense of orders of magnitude, for the sake of computational feasibility. We swept over model depths of 3, 6, and 9; widths of 16, 64, 256, and 1024; and RFF Ω_0's of 1, 5, 10, and 15. Three values - 0.2, 0.5, and 0.8 - were considered for the final parameter, dropout probability, which is specific to MCD. For BBB, we instead swept over the Gaussian prior standard deviation (values 10, 100, and 1000)[2] and KL factor (values 1e-10, 1e-5, and 1e-1)[3]. For these coarse grid searches, all networks were trained using the Adam optimizer with no weight decay and the default learning rate of 3e-4. For each set of parameters, three individual INRs were trained, one for each of the three validation images, Image #1-#3, shown in Figure 13[4]. All reported metrics are averaged across the three test images, in an effort to ensure model generalization and prevent overfitting to a particular image. For both the 5- and 20-view experiments, 2,160 (2,512) models were trained and tested for the MCD (BBB) coarse grid sweeps. This resulted in a total of 9,344 models trained and tested during these initial grid searches. Since the performance metrics are calculated from a distribution of predictions, sampled according to the uncertainty method, all hyperparameter sweeps used 50 prediction samples to enable uncertainty quantification, mean output prediction, and metric calculation.

For now, we conclude our discussion of methodology, with the second hyperparameter sweep – a fine Bayesian search used to generate the final models. This Bayesian sweep leveraged a reduced search space, informed by the first coarse grid search. It was found that Bayesian sweeps do not perform well with categorical variables, so independent sweeps of ∼200 runs were performed for each of the three best performing activation functions from the grid search. Since only 600 models were trained in total in each uncertainty-view setting, all five validation images of Fig. 13 were used as the validation set, to improve model generalization ability. Further, AdamW was used to optimize the models, with a weight decay hyperparameter added to the sweep. In the case of MCD with 20-views, three sweeps were performed, one for each of the three activation functions: Sine, SiLU, and Tanh. A log uniform distribution, in the range 1e-16 to 1e-1, was swept for the weight decay, while uniform distributions $U(\text{min, max, } q)$, where q is the discrete interval, were generated and swept over for the remaining numerical parameters: depth $\in U(2, 12, 1)$, width $\in U(200, 1000, 100)$, $\Omega_0 \in U(3, 15, 1)$, and $p(\text{dropout}) \in U(0.1, 0.6, 0.1)$. The top performing model according to PSNR, across all three Bayesian sweeps, was selected as the final model.

E.4.2. Deep Ensembles

Given the robust and computationally intensive nature of DEs, we did not perform large-scale hyperparameter sweeps, as was the case for BBB and MCD. DEs combine the outputs of multiple base learner models to improve uncertainty calibration. Assuming each base learner makes a reasonable prediction, even if not optimized, adding more base learners should only maintain or improve uncertainty calibration. In order to create a DE of size M (DE-M), the top-M performing models identified by the MCD hyperparameter sweeps were used as base learners. If N total samples were desired for uncertainty quantification, each of the MCD baselearners was sampled repeatedly to generate $\frac{N}{M}$ predictions. These predictions were pooled together to create a sample of size N, from which uncertainty was quantified, the mean prediction generated, and model metrics calculated. In order to remain consistent with the BBB and MCD experiments, we used $N = 50$.

[2] We originally swept over BBB standard deviation (values 0.2, 0.5, and 0.8), which are more on par with theoretical expectations. However, we found that increasing prior standard deviation significantly improved final model performance.

[3] Given the added BBB uncertainty hyperparameter, we reduced relative number of search values for the remaining sweep parameters.

[4] Again, for the sake of computationally efficiency, only validation Images #1 and #2 were used for BBB.

E.5. Uncertainty Quantification Method Performance Analysis

E.5.1. MCD Hyperparameter Analysis

For MCD hyperparameter sweeps, metrics were averaged across the INR model reconstruction of the first three validation set images of Figure 13. However, we also recorded the PSNR of the INRs trained on each individual image. Box plots for the individual distributions of PSNR for validation Images #1, #2, and #3 are shown in Figure 15. Ideally, PSNR should be consistent across the three images. This is roughly the case (barring a few outlier points) for models using the Tanh, SoftPlus, Sine, and Silu activation functions. Notice, however, that in all cases the distribution is broadest for Image #3. This effect is exacerbated for models using the ReLU activation function, for which PSNR of Image #3 ranges all the way from approximately 0dB to nearly the maximum achieved PSNR, in both the 5- and 20-view cases. This indicates that ReLU in particular, but all the all activation functions to some extent, struggle to capture features of Image #3.

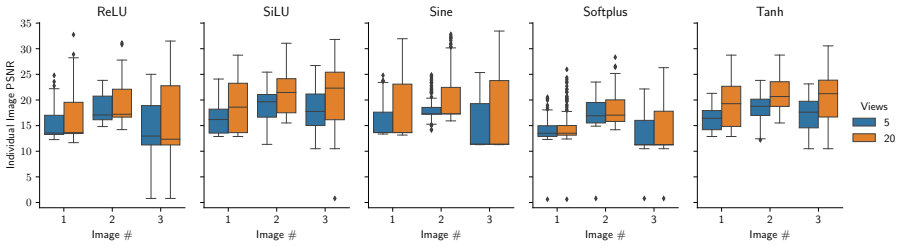

Figure 15. Reconstruction PSNR for Images #1-3 of Figure 15, for MCD INRs with different activations and different numbers of views.

To understand what is unique about Image #3, the image reconstructions of the best overall performing model are shown, for each activation function and 20-views, in Fig. 16. It is apparent that Sine and SiLU produce smoother images, while Tanh, Softplus, and ReLU produce spottier images. Except for SoftPlus, all activation functions manage to capture low-frequency image information and strong edges. However, all activations struggle to capture the small, low-intensity ellipses in the center of Image #3. Overall, it is clear that, irrespective of activation function, 20-views are insufficient to robustly capture fine CT image details. This individual image analysis also suggests that the Sine activation performs the best for MCD, achieving the highest overall PSNR.

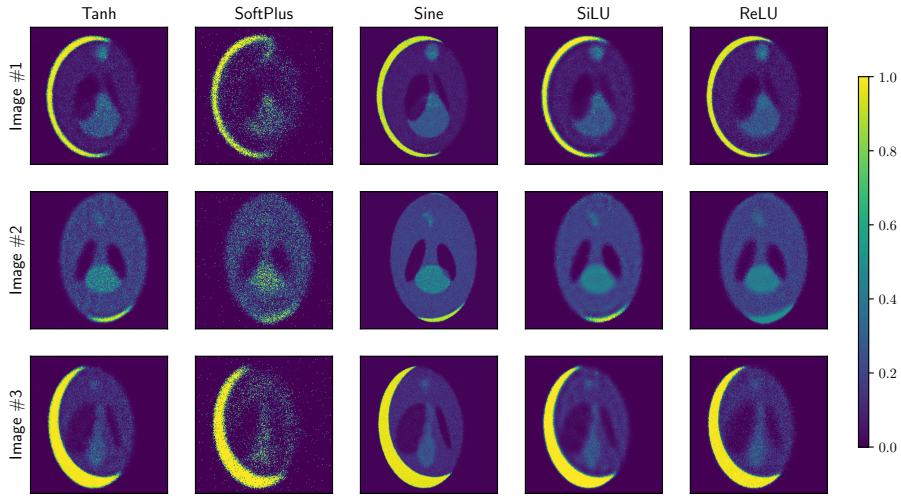

Figure 16. Best image reconstructions obtained with each activation function, for 20-view MCD.

Figure 17 shows boxplots of the average PSNR distributions for MCD models trained in the 5- and 20-view cases. Each column contains all the models trained with one of the five activation functions, while each row shows how the PSNR distribution changes as a function of hyperparameter – depth, width, probability of dropout, or RFF frequency Ω_0. Ideally, PSNR would be consistently large across hyperparameter values, indicating that the architecture is robust and does not require much tuning. In practice, however, we find that the activation functions

are either consistent or high-performing, but not both. As previously mentioned, Sine achieves the best overall PSNR. However, it is also the least consistent activation function, with its highest performing models typically being outliers (indicated by diamonds in Figure 17). Softplus, on the other extreme, is very consistent across hyperparameter values, but performs consistently poorly. Tanh, Silu, and ReLU have less extreme variations. Their top models perform slightly worse than the best Sine models, but they perform much more consistently across hyperparameter values (ReLU is a bit inconsistent in width and probability of dropout). This suggests that the Sine network is potentially the best reconstruction network, but significant tuning (in terms of hyperparameter search) effort may be needed to achieve that solution. For practitioners inclined to perform less tuning, the Tanh and SiLU networks may be a preferred solution, due to their robustness and competitive top performance.

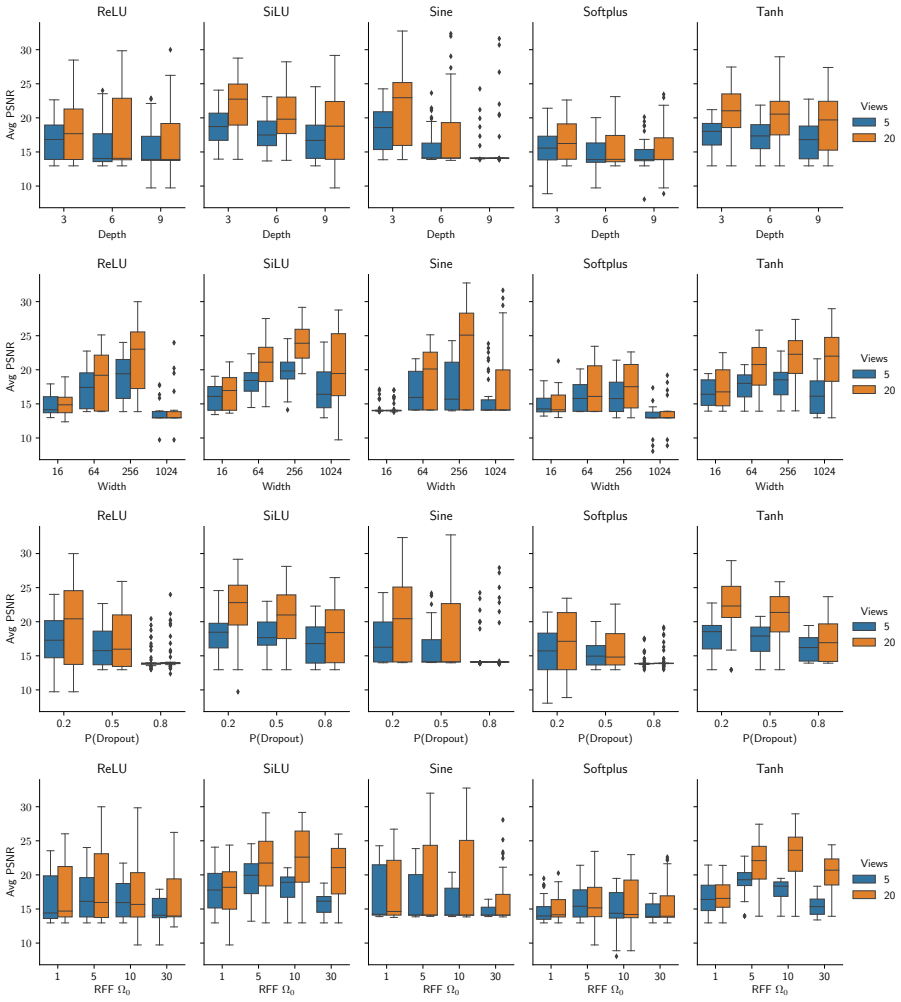

Figure 17. Boxplots of the average PSNR of MCD models trained in the coarse grid search hyperparameter sweep, for both 5 and 20 views. Each column corresponds to a different activation function and each row to a sweep over one of the remaining hyperparameters - depth, width, probability of dropout, and RFF frequency Ω_0. Individual diamond points are outliers.

For ReLU, Tanh, SiLU, and Sine, it should also be noted that the PSNR distributions behave similarly in the 5- and 20-view cases (with 5 views performing consistently worse than 20 views, as expected) for all hyperparameters except RFF Ω_0. This is consistent with recent results (Tancik et al., 2020) suggesting that RFF embeddings enable NNs to learn higher frequency image information. In the 5-view case, where the available data is insufficient for the INR to confidently learn high-frequency image features, performance is poor for increasing Ω_0. In the 20-view case, the increased data enables the network to learn higher frequency image components. However, because a higher-frequency Ω_0 is required to ensure the network can actually learn those frequencies, best performance tends to occur for larger Ω_0. This effect can also be observed in Figure 18. For 5 views, the reconstruction is blurry for $\Omega_0 = 1$ but the network starts to learn smaller ellipses for $\Omega_0 = 5$. By $\Omega_0 = 10$, however, the network is

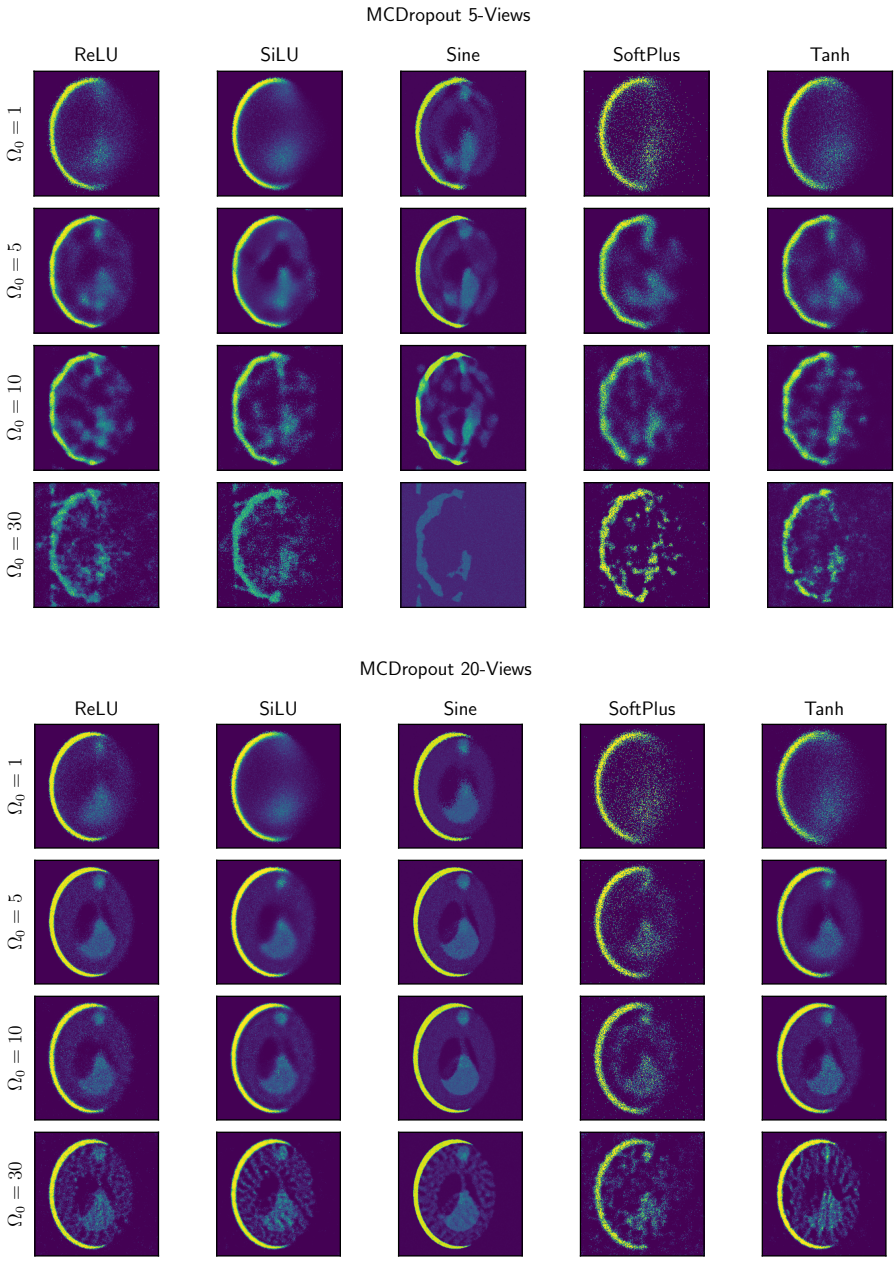

Figure 18. MCD image reconstruction, in both the 5- and 20-view cases, for each activation function and RFF frequency Ω_0 value. Note that in the 5-view case, $\Omega_0 = 5$ enables the network to learn low-frequency image features, without many artifacts. Smaller Ω_0 causes the network to produce overly simple output, whereas larger Ω_0 induces high-frequency artifacts. In the 20-view case, similar observations are made, but the optimal $\Omega_0 = 10$. In this case, the reconstruction has higher frequencies without significant artifacts, for most activation functions.

trying to learn higher-frequencies than the data cannot specify, resulting in artifacts, which are exacerbated for $\Omega_0 = 30$. In the 20-view case, a similar trend of reduced blurriness and increasingly sharper images can be seen between $\Omega_0 = 1$ an $\Omega_0 = 10$. However, the reconstructed images have much more recognizable details and less artifacts than those obtained with 5 views. It is only for $\Omega_0 = 30$ that artifacts start to appear.

E.5.2. BBB Hyperparameter Analysis

The BBB model selection analysis is presented in a similar fashion to that of MCD. The main differences to MCD are that, in the BBB grid search, metrics are averaged over only the first two validation set images; the width and RFF search spaces are reduced; and the uncertainty parameters are the KL factor and Gaussian prior standard deviation (not probability of dropout). Figure 19 shows boxplots of average PSNR as a function of different hyperparameter values. Unlike MCD, model performance is extremely consistent across activation functions for every hyperparameter, except for width. We note that Tanh has some variation for RFF Ω_0, following trends similar to those discussed in the case of MCD. Overall, SiLU performs the best, with ReLU and Tanh following closely behind. Interestingly, Sine performs the worst, failing to reach even 20db. However, greater performance consistency across hyperparameter configurations comes at the price of weaker top-performing models, which achieve lower PSNR values than those of MCD. Because the image sets are not identical, these comparisons across approaches should be taken with some reservation.

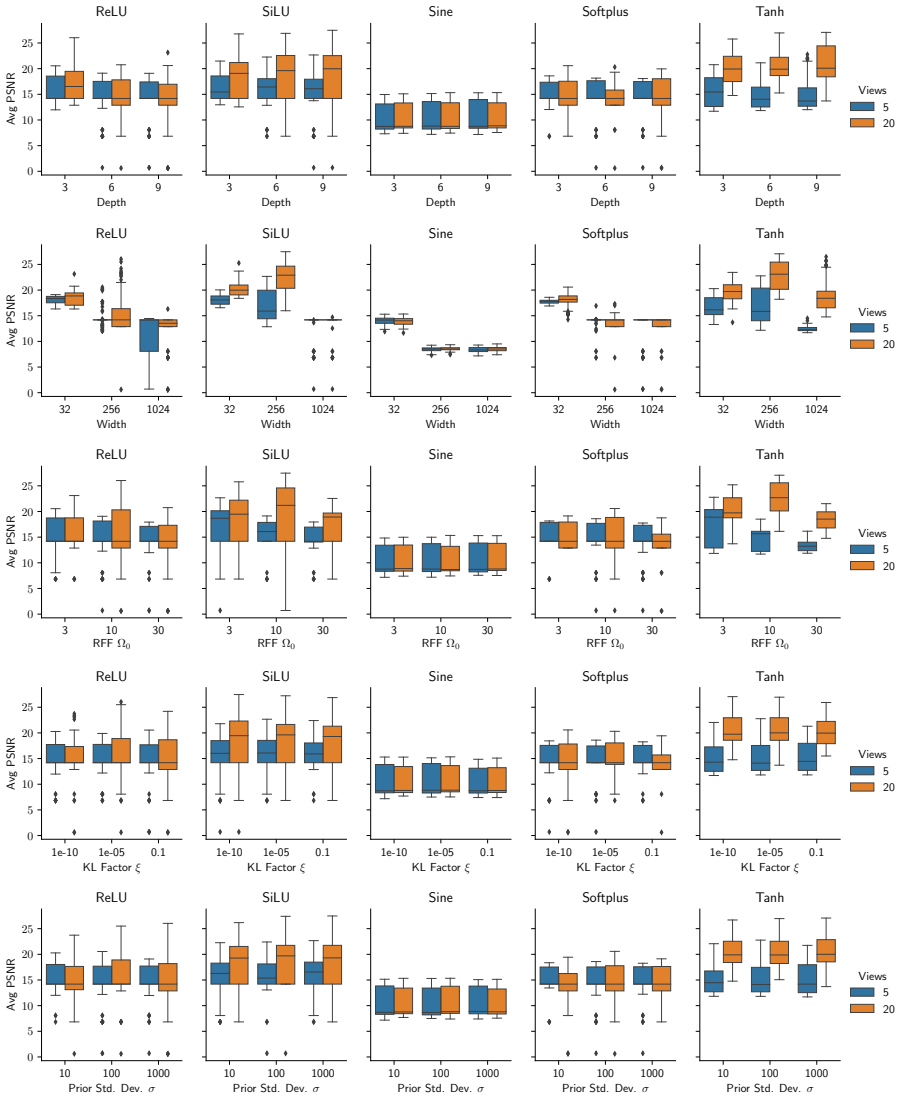

Figure 19. Boxplots of the average PSNR of BBB models trained in the coarse grid search hyperparameter sweep, for both 5 and 20 views. Each column corresponds to a different activation function and each row to a sweep over one of the remaining hyperparameters - depth, width, RFF Ω_0, KL factor ξ, and prior standard deviation σ.

To understand why there is so much variation in BBB performance as a function of width, consider Figure 20, where the average PSNR obtained for each width is plotted as a function of prior standard deviation. It can be seen

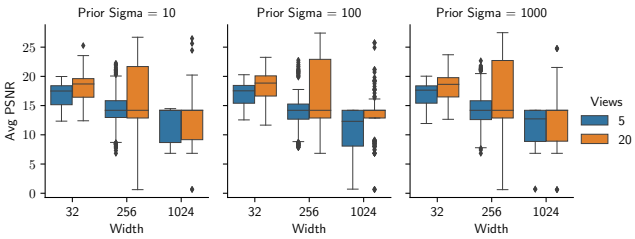

Figure 20. Boxplots of average PSNR of BBB models of varying width for each value of prior standard deviation σ, for both 5- and 20-view cases.

that the PSNR values are extremely consistent for each network width, across prior standard deviations. From a Bayesian perspective, the fact that the prior does not affect inference suggests that the latter is dominated by the model likelihood. However, mean performance decreases as a function of model width, indicating that the model becomes increasingly misspecified for larger widths. Given the Gaussian assumptions made in the variational inference specifications of BBB, this suggests that the true posterior distribution becomes less Gaussian as network width increases, and the variational approximation deteriorates. When combined with the low performance of BBB relative to other uncertainty quantification methods, reported in Table 1, this indicates that BBB may not be well suited for uncertainty quantification of INRs in the medical imaging context.

E.5.3. DE Performance Analysis

As described in Appendix E.4.2, DEs of M base learners were created by combining the top-M performing NNs, according to average PSNR. As reported in Table 1, the best MCD model outperforms the best BBB model significantly, with at least a 2dB increase in PSNR, as well as reduced NLL and ECE, for both the 5- and 20-view cases. Thus, we ensembled the top MCD models produced by the second fine Bayesian hyperparameter sweeps of Section E.5.2. The parameterizations and performance of the 10 best performing models for both the 5- and 20-view cases are listed in Table 4. Note that, in both cases, the model architectures vary greatly across models. While the Sine activation function is fairly consistent, the remaining parameters vary greatly. For example, in the 5-view case, model depths range from 3 to 8, widths from 300 to 800, RFF Ω_0 from 2 to 8, and probability of dropout (PD) from 0.2 to 0.7, all with a variety of weight decays. These variations increase base learner diversity well beyond different weight values.

DE combines varied base learners to improve uncertainty calibration. In principle, if all base learners achieved the

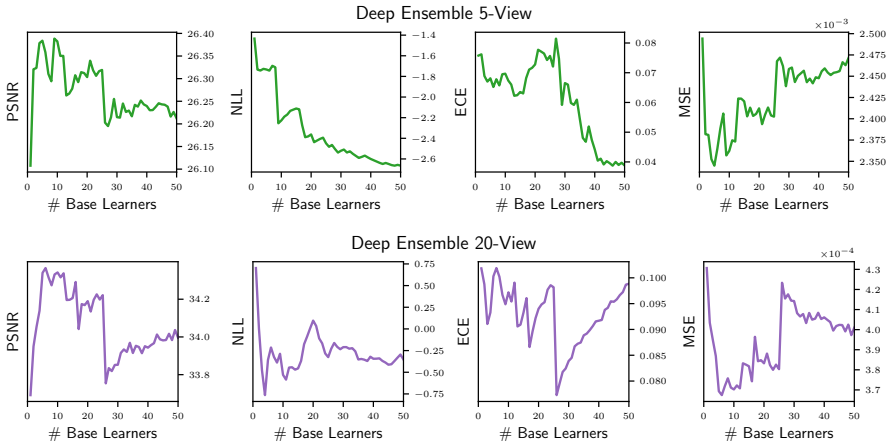

Figure 21. PSNR, NLL, ECE, and MSE plotted as a function of the number of base-learners used in DEs, for both the 5- and 20-view cases. Note that each added base learner performs slightly worse in terms of image reconstruction quality than the network preceding it.

same PSNR, model performance should only increase (or plateau) and variance should only decrease (or plateau) as more base-learners are added. In our case, however, each new added base learner had a slightly lower PSNR on the validation set. To verify how this affected model performance and calibration, we generated plots of each metric as a function of # of DE base learners. As shown in Figure 21, both PSNR and MSE improve significantly as the first base learners are added to the ensemble, but begin to worsen for larger ensembles. Considering that models of worse PSNR are being added with each increase in # of base learners, it is unsurprising that performance eventually degrades. However, ensembling never reduces performance below that of using the single best model. NLL and ECE are more sensitive to the number of base learners. NLL demonstrates the overall best performance gain as a function of baselearners. ECE, however, does not change consistently with DE size, with large ensembles sometimes even harming performance relative to the single best model. In all, it is clear that ensembling improves model performance and calibration. However, larger ensembles have no gains over smaller ensembles and are far more computationally expensive. Thus, for the final results, presented in Table 1, only ensembles of sizes 2, 5, and 10 are considered.

Table 4. Top 10 performing MCD models and their performances, for the 5-view case (**Top**) and 20-view case (**Bottom**). Models are ranked by PSNR, but NLL and ECE are also reported.

Rank	Activation	Depth	Width	RFF Ω_0	PD	W. Decay	PSNR	NLL	ECE
1	Sine	4	800	2	0.4	0.001	26.15	-1.437	0.122
2	Sine	3	600	4	0.4	0.157	25.79	-1.799	0.008
3	Sine	3	600	8	0.7	0.366	25.76	-1.579	0.009
4	Sine	3	700	2	0.4	4.46e-4	25.75	-1.533	0.117
5	Sine	4	700	3	0.5	9.36e-4	25.72	-1.390	0.011
6	Sine	3	500	5	0.7	0.077	25.63	2.622	0.161
7	Sine	5	700	3	0.6	0.012	25.62	-1.149	0.123
8	Sine	3	500	5	0.7	0.068	25.62	-1.643	0.007
9	SiLU	8	300	3	0.2	2.68e-7	25.62	0.219	0.389
10	Sine	3	500	4	0.7	0.170	25.59	-1.79	0.083

Rank	Activation	Depth	Width	RFF Ω_0	PD	W. Decay	PSNR	NLL	ECE
1	Sine	3	400	9	0.4	2.06e-5	33.74	0.701	0.134
2	Sine	4	300	12	0.4	2.63e-7	33.41	1.076	0.149
3	Sine	3	500	8	0.5	0.006	33.37	-0.502	0.137
4	Sine	5	500	11	0.4	3.74e-6	33.29	-0.288	0.137
5	Sine	6	400	10	0.2	5.85e-6	33.28	4.654	0.152
6	Sine	6	500	8	0.2	1.117e-4	33.24	2.622	0.161
7	Sine	4	400	9	0.5	0.001	33.21	-0.798	0.143
8	Sine	3	500	10	0.5	2.53e-4	33.20	-0.716	0.129
9	Sine	5	500	11	0.2	7.87e-7	33.18	1.973	0.163
10	Sine	4	500	8	0.5	4.56e-5	33.17	-1.257	0.114

Figure 22 illustrates image reconstruction performance for the different ensemble types on test set Image #3. In the 5-view case, DEs achieve impressive performance improvements, with a PSNR increase of over 1.5dB for DE-5 and an ECE reduction of 0.011 for DE-2. In the 20-view case the baseline is much better performing.

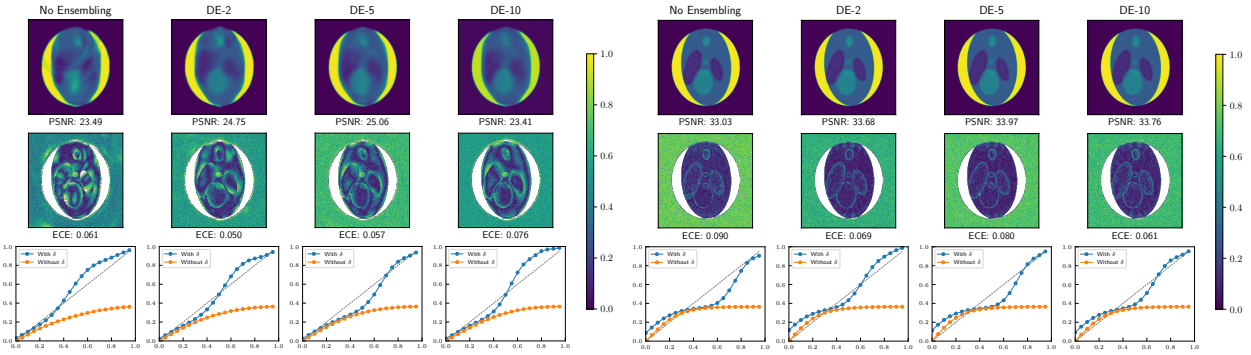

Figure 22. Image reconstruction and calibration performance of the final DE models on test set Image #3 for 5-views (**Left**) and 20-views (**Right**). Different columns show the results of different ensemble sizes, ranging from 1 to 10. Top row shows the reconstructed image, middle row the pixelwise coverage, and bottom row the reliability curve.

Hence, although there are gains in PSNR, these are not very noticeable in the reconstructed image. However, the improvements in calibration are larger, with an ECE drop of 0.2 for DE-2 and a clear improvement in the image reliability curve.

E.6. Final Results Conclusions

The primary goal of this study was to understand the relative performance of different uncertainty quantification techniques for UncertaINR, on the Shepp-Logan dataset. Table 1, in the main text, presents metrics assessing the best-performing MCD, DE-2 MCD, DE-5 MCD, and DE-10 MCD UncertaINRs. These methods consistently outperformed the classical reconstruction techniques in terms of image quality, while producing reasonably well calibrated uncertainty estimates. BBB was found to be the worst performing uncertainty quantification approach, generally producing the poorest calibrated uncertainty estimates and worse image reconstruction than classical techniques in the 20-view regime. MCD consistently outperformed classical approaches and was generally better calibrated than BBB. Ensembling over MCD base learners, however, was the most successful approach, outperforming the best classical approach by $\sim$4dB in the 5-view case and $\sim$3dB in the 20-view case, as well as achieving the lowest overall NLL and ECE values.

One further remark regarding to Table 1 is that although the classical reconstruction procedures did not use the validation set images as a validation set (since no hyperparameters were tuned), image reconstruction quality still deteriorated in the test set. This indicates that the test set data is actually more challenging than the validation set. Given that UncertaINR performance did not significantly decline on the test set, we have strong reason to believe that none of the UncertaINR approaches over-fitted to the validation set.

Figure 23 visually illustrates the difference in the uncertainty quantification of the different methods, for the 5- and 20-view cases respectively. Beginning with the 5-view case, the mean predicted image was fairly blurry for all methods, only capturing low-frequency image components and exhibiting high uncertainty surrounding edges. Furthermore, note that all reliability curves are very well calibrated in this 5-view case (after a δ-selection adjustment), except for BBB. Similar trends are present in the 20-view case, but it is visually harder to distinguish differences in the image output due to the higher quality of all reconstructions.

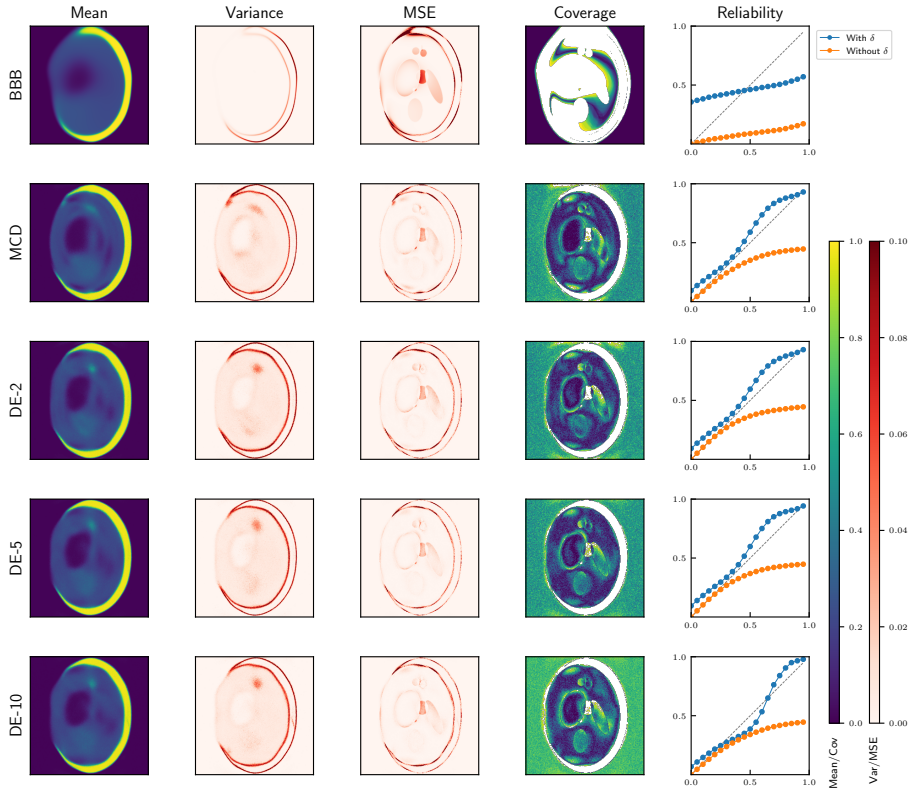

Figure 23. Validation results of all approaches for 5-view Shepp-Logan data. From left to right: mean image reconstruction, variance, MSE, coverage, and reliability diagram.

F. UncertaINR AAPM-Mayo Performance Assessment

F.1. Dataset

In order to assess the performance of UncertaINR in a realistic setting, we used data from the American Association of Physicists in Medicine (AAPM) and Mayo Clinic 2016 Low-Dose CT grand challenge (McCollough et al., 2017). Specifically, we use the 8 reconstructed images used as a test set in the CoIL work (Sun et al., 2021), shown in Figure 25, so as to directly compare to their results. These ground truth images were reconstructed from real CT scan measurement data.

In order to make the dataset more realistic and compare to the results presented in CoIL, noise was added to the image sinograms before being input to the models. This artificial data generation pipeline is illustrated in Figure 26. More specifically, uniformly distributed Gaussian white noise was added to the sinogram, so as to achieve a desired SNR relative to the original, noise-less sinogram. In this work, we chose to present results for a noise level achieving sinograms of 40dB SNR.

F.2. Final Model Hyperparameters

Given the increased computational costs in running UncertaINR on the AAPM-Mayo dataset, it was not feasible to perform large-scale hyperparameter sweeps, like those performed in the ablation study presented in Appendix E. However, coarse searches were performed for each hyperparameter and new training approaches were used in order to achieve the competitively performing models presented in Table 2. For reproducibility, we provide the final hyperparameters and training procedures.

Beyond training insights learned from the hyperparameter study, further improvements were made to UncertaINR training in order to achieve competitive reconstruction accuracy on the higher-resolution, higher-frequency, and noisy Mayo-AAPM data. Specifically, images were zero-padded to ensure that all image content was contained in

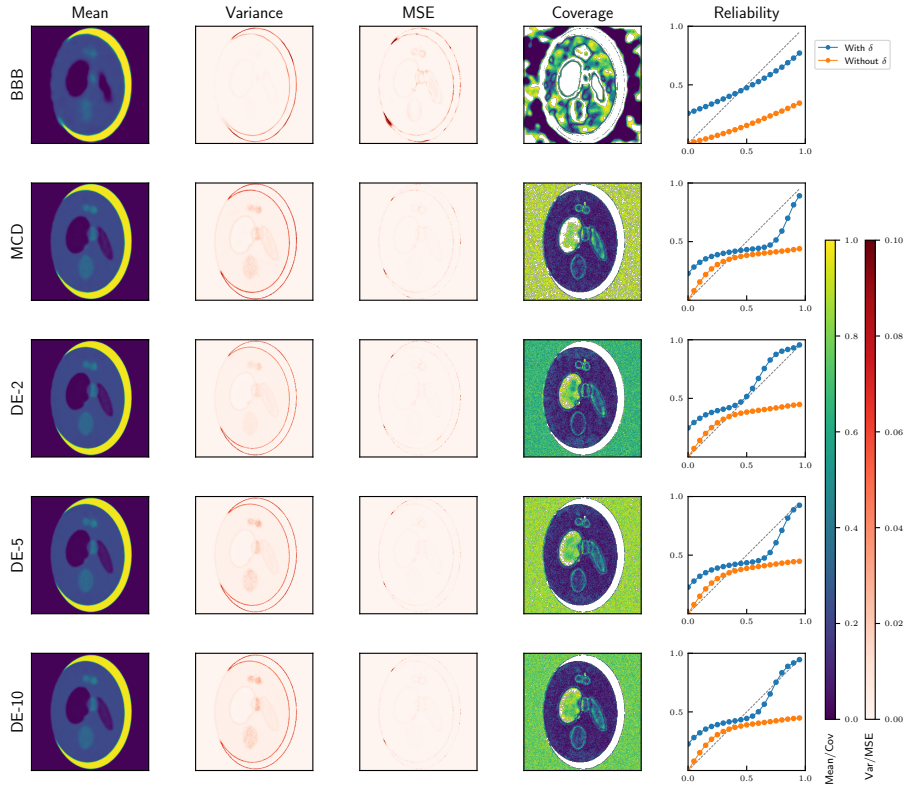

Figure 24. Validation results of all approaches for 20-view Shepp-Logan data. From left to right: mean image reconstruction, variance, MSE, coverage, and reliability diagram.

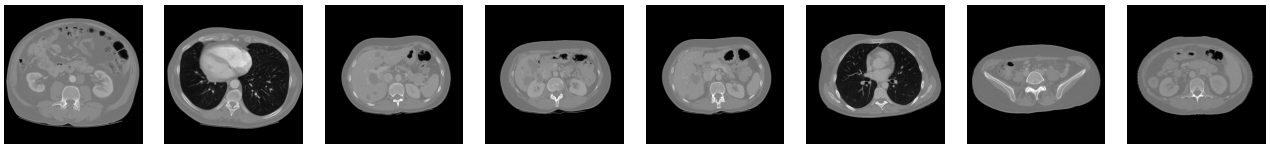

Figure 25. The 8 AAPM-Mayo test set images used to assess UncertaINR performance.

the projections when calculating the Radon transform. All UncertaINR were optimized using Adam (not AdamW, to eliminate the computational costs of tuning the weight decay parameter) and we found that longer training times were needed (on the order of 15,000 epochs). If that many training epochs cannot be used for some reason, we found that stochastic weight averaging (Izmailov et al., 2018) can be used to improve reconstruction accuracy by a few decibel for UncertaINRs trained with few epochs. With the increased number of training iterations, the Sine activation proved too unstable and we found SiLU to be the best performing activation function, with Tanh a close second. Noting the significance of RFFs from the ablation study, we found it crucial that the width of the RFF layer be at least as large as the padded image width. Finally, before adding a regularization term to the UncertaINR loss, our models were not competitive with state-of-the-art reconstruction techniques. While adding an isotropic TV regularization,

$$\tilde{T}_{\text{ISO}}(f) = \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} (f(x+1,y) - f(x,y))^2 + (f(x,y+1) - f(x,y))^2,$$

boosted performance by a few decibel, the anisotropic approximation of the TV regularizer,

$$\tilde{T}_{\text{ANISO}}(f) = \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} |f(x+1,y) - f(x,y)| + |f(x,y+1) - f(x,y)|,$$

achieved better reconstruction accuracy, especially in the low-measurement (60-view) regime. (at the slight cost of uncertainty calibration). The relative performance of these two regularizers for the MCD UINR are presented

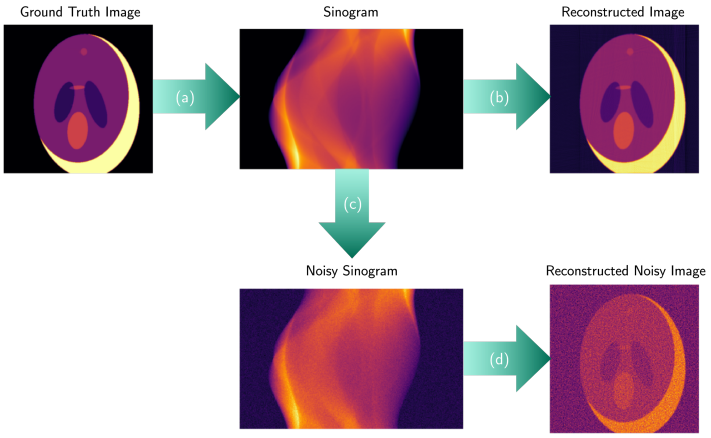

Figure 26. A flowchart of the artificial data generation pipeline used in this work. **(a)** The Radon transform is used to generate a sinogram, corresponding to measurement data, from the ground truth data. **(b)** In the noiseless case, this sinogram would be directly used to train our model. Note that even though the sinogram is noiseless, the reconstruction is not perfect. **(c)** In a more realistic scenario, Gaussian white noise is added to the sinogram. **(d)** This noisy sinogram data is then fed into the model, which produces a reconstructed image of the ground truth, of lower quality that produced in the noiseless case.

in Table 5. While we do not have a justification for why the anisotropic approximation performs better, we believe that regularization/prior exploration would be interesting future work.

Table 5. The affect of using different approximations to the TV regularizer for MCD UINR performance.

Reconstruction	Regularization	60 Views			120 Views		
Method	Type	SNR	NLL	ECE	SNR	NLL	ECE
MCD UINR	Isotropic	25.78	-3.815	0.066	28.06	-4.118	0.060
	Anisotropic	27.38	-3.447	0.078	28.65	-3.759	0.071

Given all these training insights, the hyperparameters of the top-performing MCD UINRs, reported in Table 2, are presented in Table 6.

Table 6. Hyperparameters of the top-performing MCD UINRs and INRs reported in Table 2.

Model	# Views	Act. Func.	Depth	Width	RFF Ω_0	Reg Type	Reg Coeff	p(Dropout)	# Epochs
INR	60	SiLU	5	280	48	Aniso	0.05	0	15,000
MCD UINR	60	SiLU	5	280	35	Aniso	0.03	0.2	15,000
INR	120	SiLU	5	280	60	Aniso	0.05	0	15,000
MCD UINR	120	SiLU	5	280	40	Aniso	0.008	0.2	15,000