Uncertainty Quantification in End-to-End Implicit Neural Representations for Medical Imaging

Francisca Vasconcelos[*]
francisca.vasconcelos@keble.ox.ac.uk

Bobby He[*]
bobby.he@stats.ox.ac.uk

Yee Whye Teh
y.w.teh@stats.ox.ac.uk
Department of Statistics
University of Oxford
Oxford, UK OX13LB

Abstract

Implicit neural representations (INRs) have recently achieved impressive results in image representation. This work explores the uncertainty quantification quality of INRs for medical imaging. We propose the first uncertainty aware, end-to-end INR architecture for computed tomography (CT) image reconstruction. Four established neural network uncertainty quantification techniques – deep ensembles, Monte Carlo dropout, Bayes-by-backpropagation, and Hamiltonian Monte Carlo – are implemented and assessed according to both image reconstruction quality and model calibration. We find that these INRs outperform traditional medical image reconstruction algorithms according to predictive accuracy; deep ensembles of Monte Carlo dropout base-learners achieve the best image reconstruction and model calibration among the techniques tested; activation function and random Fourier feature embedding frequency have large effects on model performance; and Bayes-by-backpropogation is ill-suited for sampling from the INR posterior distributions. Preliminary results further indicate that, with adequate tuning, Hamiltonian Monte Carlo may outperform Monte Carlo dropout deep ensembles.

1 Introduction

In computed tomography (CT), improving reconstructed image quality via increased measurement also increases patient exposure to harmful radiation [1, 2, 3]. As a result, there is interest in reconstruction techniques which achieve high image quality from few measurements. Machine learning approaches based on deep learning have proved promising in this regard. However, they require large training data sets, which are difficult to collect in the medical setting. A significant recent advance was the introduction of implicit neural representations (INRs), which represent complex coordinate-based signals as functions encoded by small neural networks. For example, an image can be represented as a function mapping (x, y) coordinates to (r, g, b) pixel intensities. INRs have taken the field of computer graphics by storm, achieving impressive results in novel view synthesis [4, 5, 6, 7], shape representation [8, 9, 10, 11, 12, 13], and texture synthesis [14, 15]. More recent work has also demonstrated the applicability of this technique to medical imaging [16, 17].

In all these applications, INRs were assessed on their predictive accuracy and reconstructed signal plausibility. However, in medical imaging, which affects doctor decisions and patient well-being, it is also important to understand model confidence in the reconstructed image. For example, a model

[*] Denotes equal author contribution. Order decided by rock-paper-scissors (best of 3).

Bayesian Deep Learning workshop, NeurIPS 2021.

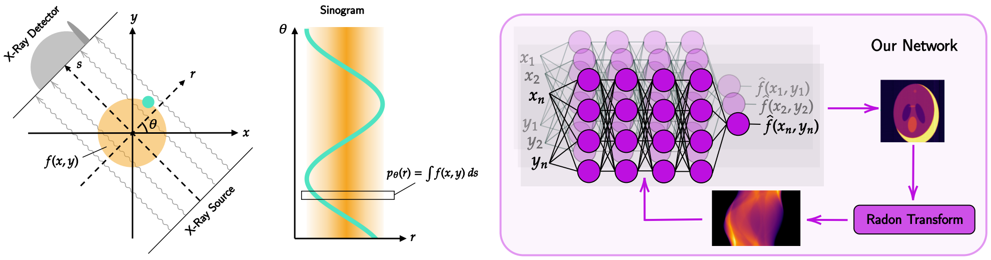

Figure 1: **Left)** An abstraction of the CT measurement process and resultant sinogram data. **Right)** An illustration of our end-to-end INR architecure, which directly outputs the desired image by embedding the sinogram measurement data in the network loss.

can quantify its uncertainty in each pixel by outputting a per-pixel variance. If this variance is large in critical image regions, such as the location of a potential tumor, a doctor should order additional measurements to ensure proper diagnosis. Uncertainty quantification can also be used to decrease healthcare cost via automated triage, e.g. by assigning images with varying degrees of uncertainty to healthcare providers of relevant expertise. Finally, understanding model uncertainty could inform more efficient measurement procedures, leveraging techniques such as active learning [18].

2 Methods

Encoding 2D CT image reconstruction in INRs As shown in Figure 1, CT measurement data comes in the form of a sinogram, $p_\theta(r)$, where r is the X-ray detector location and θ is the measurement angle. However, the goal of reconstruction is to generate a photo of the 2D cross-section of attenuation coefficients, $f(x, y)$. This is achieved using an *end-to-end* approach to image reconstruction, in which our model predicts the final cross-section attenuation coefficient function, as illustrated in Figure 1. The model input is pixel coordinate (x, y) and its output is the corresponding predicted attenuation coefficient value $\hat{f}(x, y)$. The sinogram data is incorporated in the model via the training loss function. Given a ground truth sinogram $p_\theta(r)$, the loss of the INR output $\hat{f}(x, y)$ is defined as

$$\mathcal{L}\big(p_\theta(r), \hat{f}(x,y)\big) = \frac{1}{2|\Theta \times \mathcal{R}|} \sum_{\theta \in \Theta} \sum_{r \in \mathcal{R}} \left(p_\theta(r) - \int_{\mathcal{Y}} \int_{\mathcal{X}} \hat{f}(x,y)\ \delta(x \cos\theta + y \sin\theta - r)\ dxdy \right)^2, \tag{1}$$

where $\Theta = \{\theta_1, ..., \theta_n\}$ are the view angles, $\mathcal{R} = \{r_1, ..., r_n\}$ the X-ray detector locations, and $\mathcal{X} \times \mathcal{Y}$ the image pixels (x, y). The integral surrounding $\hat{f}$ is the Radon transform. Since the loss is calculated directly on the desired output, end-to-end training minimizes the propagation of error, but has a training complexity cost. Each training iteration requires sampling the model $|\mathcal{X} \times \mathcal{Y}|$ times, once per image pixel, and a Radon transform must be calculated for all $|\Theta \times \mathcal{R}|$ points in the sinogram. However, given the relatively small nature of INRs by deep learning standards, we did not find this computationally barring, with networks taking no more than a few minutes to train.[2]

Uncertainty quantification of INRs We implemented and compared multiple methods for uncertainty quantification of INR parameters and predictions. Experimental details are reported in Appendix A. Bayes-by-backpropagation (BBB) [19], Monte Carlo dropout (MCD) [20], and Hamiltonian Monte Carlo (HMC) [21, 22, 23] were used for approximate Bayesian neural network (BNN) inference, while deep ensembles of size N (DE-N) [24] were used to aggregate the results of N MCD base learners. Besides HMC, which is regarded as the gold standard inference scheme for BNNs [25], all these approximate inference schemes are computationally efficient and common choices within the uncertainty deep learning community [26]. Model performance was assessed according to peak signal-to-noise ratio (PSNR), negative log likelihood (NLL), and expected calibration error (ECE).

Baseline As a baseline, we compared our INRs to standard medical image reconstruction algorithms: filtered back-projection (FBP), conjugate-gradient least squares (CGLS), expectation maxi-

[2] Experiments were run on a cluster of four graphical processing unit (GPU) nodes of 8 GPUs each, consisting of a mixture of GTX 1080, GTX 1080Ti, and GeForce RTX 2080 Ti cards, all with under 12GB VRAM.

# View Angles	Recon Type	Validation Set			Test Set		
		PSNR	NLL	ECE	PSNR	NLL	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	–	–
	HMC*	–	–	–	24.87*	-1.616*	0.090*
	BBB	23.26	-1.190	0.152	22.52	0.138	0.203
	MCD	26.15	-1.473	0.111	24.45	-1.572	0.083
	DE-2	26.31	-1.730	0.091	24.49	-1.774	0.069
	DE-5	**26.44**	-1.737	0.085	**24.88**	-1.751	0.067
	DE-10	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	–	–
	HMC*	–	–	–	29.12*	-1.676*	0.009*
	BBB	28.25	1.650	0.121	28.16	0.562	0.119
	MCD	33.74	0.701	0.135	33.08	1.093	0.113
	DE-2	33.96	0.005	0.136	33.44	-0.372	0.102
	DE-5	34.31	-0.364	0.134	**34.02**	-0.625	0.101
	DE-10	**34.38**	**-0.529**	**0.131**	33.86	**-0.774**	**0.096**

Table 1: INR accuracy and calibration results of all four uncertainty quantification approaches are presented. Classical approaches do not produce uncertainty estimates. For BBB, MCD, and DEs results are averaged over all validation and all test set images. HMC results are preliminary, with a star demarcating that HMC was trained and tested on a single image outside the validation and test sets. The best result for each metric – PSNR, NLL, and ECE – is bolded in each subcategory.

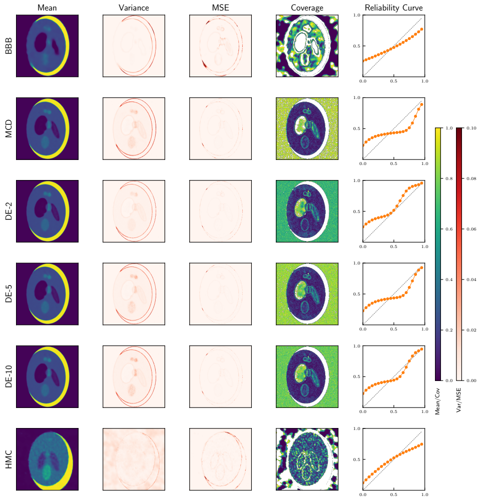

Figure 2: The pixel-wise mean, pixel-wise variance, pixel-wise mean squared error (MSE), pixel-wise coverage, and image reliability curves are shown for BBB, MCD, and DE image reconstructions of a test set 20-view sinogram. The same is shown for HMC on a 20-view sinogram, outside the test set.

mization (EM), simultaneous algebraic reconstruction technique (SART), and simultaneous iterative reconstruction technique (SIRT). These were implemented using the TomoPy Astra wrapper [27].

Data The Shepp-Logan phantom [28] approach was used to generate artificial brain images, with corresponding sinograms generated via the Radon transform. In this preliminary work, no noise was added to the images. In all, 10 ground truth images and 20 corresponding sinograms were generated: 5 validation and 5 test set sinograms each for the 5- and 20-view (θ) cases.

3 Results and discussion

Although restricted to noiseless phantom data, this work presents the first large-scale study of model parameterization for INRs with uncertainty quantification. Experimental results are presented in Table 1; sample reconstructed images, variances, coverage plots, and calibration curves in the 20-view case are presented in Figure 2; and boxplots of MCD and BBB model performance by hyperparameter are presented in Appendix A. MCD significantly outperformed BBB, with activation function substantially affecting reconstruction quality. We found the Sine activation produced top-performing MCD models, as expected from the recent SIREN work [29]. Silu, Tanh, and Relu achieved slightly lower performance, but greater consistency. We also confirmed previous findings that random Fourier feature (RFF) embeddings enable neural networks to learn high-frequency image components better [30]. However, we found that the RFF frequency must be consistent with the amount of data used in training the INR. Too low a frequency prohibits higher frequency learning, while too high a frequency results in high-frequency image artifacts. Consistent with previous work [31], we further found that ensembling MCD architectures can improve image reconstruction quality and model calibration, achieving significant improvements even for small numbers of base learners. In all, DEs of the top 5 or 10 performing MCD models achieved the best results in terms of image reconstruction and calibration. In future work, to better separate the influence of inference method from INR prior choice, we will further explore the performance of tuned HMC for INRs.

While this work is the first use of uncertainty quantification for INRs, it is not the first proposal of INRs for CT image reconstruction. The CoIL architecture [16] was recently demonstrated to improve image reconstruction pipelines by learning a functional form of the measurement sinogram. However, this necessitates use of classical image reconstruction to generate the final, desired image cross-section. It also lacks support for uncertainty estimation, since the relation between the uncertainty of sinogram values and image pixels is not immediately evident. Instead, we propose an end-to-end image reconstruction pipeline where the network output is the desired image cross section, which, as shown in this work, provides seamless support for uncertainty estimation. In future work, we aim to compare the performance of our end-to-end approach to that of the CoIL architecture & consider approaches to increase the training efficiency of the proposed end-to-end solution, as well as extending our approach to other medical imaging settings, such as 3D CT & MRI.

Potential Negative Societal Impact

We do not foresee many potential negative societal impacts to our work. Given the importance of robust and reliable predictions in the medical imaging domain, we believe calibrated uncertainty quantification is an important quality for any model that is to be deployed in practice, including potentially Implicit Neural Representations.

Acknowledgements

We would like to acknowledge Nalini Singh for her useful discussions in this work.

References

[1] Eugene C. Lin. Radiation Risk From Medical Imaging. *Mayo Clinic Proceedings*, 85(12):1142–1146, 2010.

[2] Eugenio Picano. Sustainability of Medical Imaging. *Bmj*, 328(7439):578–580, 2004.

[3] Amy Berrington de Gonzalez and Sarah Darby. Risk of Cancer from Diagnostic X-Rays: Estimates for the UK and 14 Other Countries. *The Lancet*, 363(9406):345–351, 2004.

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

[5] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020.

[6] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019.

[7] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. *Advances in Neural Information Processing Systems*, 32:1121–1132, 2019.

[8] Zhiqin Chen and Hao Zhang. Learning Implicit Fields for Generative Shape Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.

[9] Boyang Deng, John P. Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA: Neural Articulated Shape Approximation. *arXiv preprint arXiv:1912.03207*, 2019.

[10] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T. Freeman, and Thomas Funkhouser. Learning Shape Templates with Structured Implicit Functions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7154–7164, 2019.

[11] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local Deep Implicit Functions for 3D Shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020.

[12] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local Implicit Grid Representations for 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020.

[13] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[14] Philipp Henzler, Niloy J. Mitra, and Tobias Ritschel. Learning a Neural 3d Texture Space from 2D Exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8356–8364, 2020.

[15] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture Fields: Learning Texture Representations in Function Space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019.

[16] Yu Sun, Jiaming Liu, Mingyang Xie, Brendt Wohlberg, and Ulugbek S. Kamilov. CoIL: Coordinate-based Internal Learning for Imaging Inverse Problems. *arXiv preprint arXiv:2102.05181*, 2021.

[17] Albert W. Reed, Hyojin Kim, Rushil Anirudh, K. Aditya Mohan, Kyle Champley, Jingu Kang, and Suren Jayasuriya. Dynamic CT Reconstruction from Limited Views with Implicit Neural Representations and Parametric Motion Fields. *arXiv preprint arXiv:2104.11745*, 2021.

[18] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active Learning with Statistical Models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[19] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[21] Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[22] Radford M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[23] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[25] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? *arXiv preprint arXiv:2104.14421*, 2021.

[26] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32:13991–14002, 2019.

[27] Daniël M Pelt, Doga Gürsoy, Willem Jan Palenstijn, Jan Sijbers, Francesco De Carlo, and Kees Joost Batenburg. 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.

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

[29] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. *Advances in Neural Information Processing Systems*, 33, 2020.

[30] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547. Curran Associates, Inc., 2020.

[31] Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris Holmes, Frank Hutter, and Yee Whye Teh. Neural Ensemble Search for Uncertainty Estimation and Dataset Shift. *arXiv preprint arXiv:2006.08573*, 2020.

[32] Adam D Cobb and Brian Jalaian. Scaling Hamiltonian Monte Carlo Inference for Bayesian Neural Networks with Symmetric Splitting. *Uncertainty in Artificial Intelligence*, 2021.

A Appendix - Uncertainty Experiments

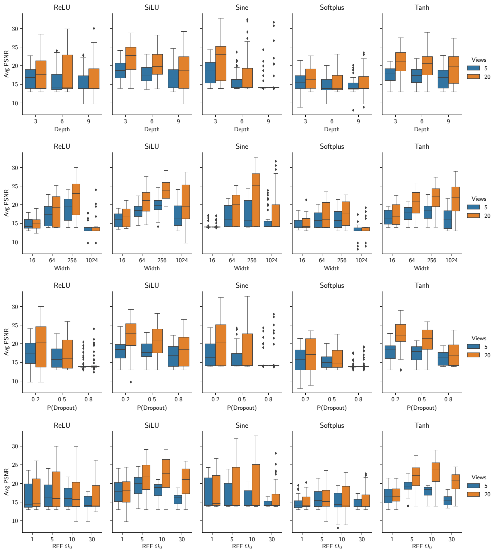

Figure 3: Boxplots of the average PSNR of MCD models trained in a coarse grid hyperparameter sweep, for both 5- and 20-views. Each column corresponds to a different activation function and each row to a sweep over one of the remaining hyperparameters - depth, width, probability of dropout, and RFF frequency (Ω_0). Individual diamond points are outliers.

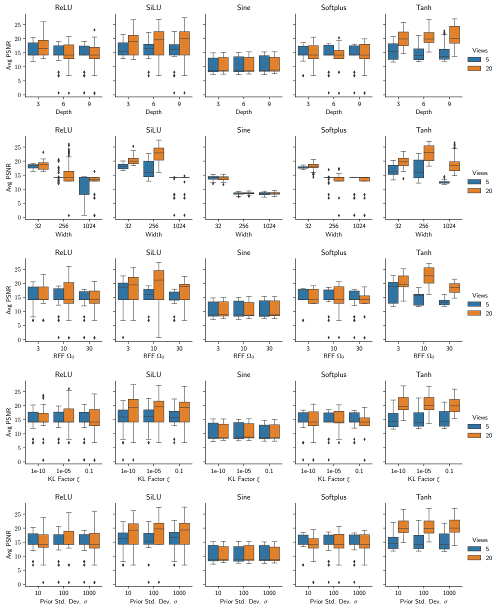

Figure 4: Boxplots of the average PSNR of BBB models trained in the coarse grid search hyperparameter sweep, for both 5 and 20 views. Each column corresponds to a different activation function and each row to a sweep over one of the remaining hyperparameters - depth, width, RFF frequency (Ω_0), KL factor (ξ), and prior standard deviation (σ).

For MCD and BBB, large-scale hyperparameter sweeps were performed to find optimal model parameters for: activation function (Relu, SiLU, Sine, Softplus, Tanh), model depth (3, 6, 9), model width (16, 64, 256, 1024), and random fourier feature (RFF) frequency (1, 5, 10, 15). For MCD probability of dropout (0.2, 0.5, 0.8) and for BBB prior standard deviation (10, 100, 1000) and KL factor (1e-10, 1e-5, 1e-1) were additionally swept over. Hyperparameter sweep models were assessed according to average PSNR, negative log likelihood (NLL), and expected calibration error (ECE) on the validation set. These experiments were run in both 5- and 20-view cases. Boxplots of the model performance for MCD and BBB according to each hyperparameter are presented in Figures 3 and 4. The best performing model in each case was further assessed on the test set.

Since MCD outperformed BBB, shown in Table 1, MCD networks were used as base learners for the DEs. DEs of size N (DE-N) were created by ensembling the N top-performing unique MCD model architectures from the hyper-parameter sweeps, in both the 5- and 20-view case. The DEs were tested on the same validation and test sets as MCD and BBB.

HMC was implemented via Hamiltorch [32], sampling from an INR of width 256, depth 3, ReLU activation, and RFF frequency 10. In this preliminary work, no hyper parameter tuning was performed, so no validation set was used. Further, given the computationally intensive nature of HMC, the network was tested only on a single Shepp-Logan phantom, outside the previously described validation and test sets. In future work, we will perform a hyperparameter search, similar to those of MCD and BBB, and report scores for the full validation and test sets.