Adjoint operators enable fast and amortized machine learning based Bayesian uncertainty quantification

Rafael Orozco[a], Ali Siahkoohi[b], Gabrio Rizzuti[c], Tristan van Leeuwen[d], and Felix Herrmann[a]

[a]Computational Science and Eng., Georgia Institute of Technology at Atlanta, GA, USA
[b]Rice University at Houston, TX, USA
[c]Department of Mathematics; Utrecht University, Utrecht, NL
[d]Centrum Wiskunde & Informatica; Amsterdam, NL

ABSTRACT

Machine learning algorithms are powerful tools in Bayesian uncertainty quantification (UQ) of inverse problems. Unfortunately, when using these algorithms medical imaging practitioners are faced with the challenging task of manually defining neural networks that can handle complicated inputs such as acoustic data. This task needs to be replicated for different receiver types or configurations since these change the dimensionality of the input. We propose to first transform the data using the adjoint operator —ex: time reversal in photoacoustic imaging (PAI) or back-projection in computer tomography (CT) imaging — then continue posterior inference using the adjoint data as an input now that it has been standardized to the size of the unknown model. This adjoint preprocessing technique has been used in previous works but with minimal discussion on if it is biased. In this work, we prove that conditioning on adjoint data is unbiased for a certain class of inverse problems. We then demonstrate with two medical imaging examples (PAI and CT) that adjoints enable two things: Firstly, adjoints partially undo the physics of the forward operator resulting in faster convergence of a learned Bayesian UQ technique. Secondly, the algorithm is now robust to changes in the observed data caused by different transducer subsampling in PAI and number of angles in CT. Our adjoint-based Bayesian inference method results in point estimates that are faster to compute than traditional baselines and show higher SSIM metrics, while also providing validated UQ.

Keywords: Uncertainty Quantification, Bayesian Inference, Amortized Inference, Normalizing Flows, Inverse Problems, Medical Imaging, Machine Learning, Deep Learning

1. DESCRIPTION OF PURPOSE

The power of machine learning methods bring accelerated and high fidelity solutions for inverse problems in a variety of fields.[1] On the downside, many machine learning methods are black boxes with failure cases that are difficult to predict and interpret. This is one of the reasons that their adoption in safety critical settings is hampered. Our purpose is to increase trustworthiness of machine learning (ML) for medical imaging by enabling uncertainty. This is important since applying ML under distribution shifts or poor training can cause instabilities and even hallucinations that could lead to incorrect diagnoses.[2,3] Uncertainty quantification (UQ) alleviates this problem by communicating to practitioners when a method is confident in its result versus when it should not be trusted.

We describe a practical Uncertainty Quantification framework based on adjoint operators and amortized variational inference (AVI). These two concepts marry powerful data-driven methods with physics knowledge allowing us to amortize over unseen data and also different imaging configurations. By amortizing, we mean that the framework trades an expensive pretraining phase for fast inference results on unseen observations. The particular class of algorithms we study, can be trained given only examples of the parameters of interest $\mathbf{x}$ and their corresponding simulated data $\mathbf{y}$. In medical imaging, data $\mathbf{y}$ can contain complex physical phenomena such as acoustic waves in photoacoustic imaging. Under the hood, an ML-based algorithm learns to undo the complex physical phenomena when providing an estimate of $\mathbf{x}$. This can lead to long training times and large training data quotas which we would like to ameliorate. We are also interested in creating an algorithm that is robust to

Send correspondence to Rafael Orozco. E-mail: rorozco@gatech.edu

changes in the dimensionality of the observations $\mathbf{y}$ such as when changing the number of transducers. To solve both problems, we propose preprocessing the data $\mathbf{y}$ with the adjoint operator.

We first show that the posterior given data is equivalent to the posterior given data preprocessed with the adjoint in the case of linear forward operators and Gaussian noise. Many medical imaging modalities fall in this category i.e. photoacoustic imaging, CT and MRI. Equipped with this theoretical result, we demonstrate that this method solves our two problems since it accelerates training convergence of AVI with conditional normalizing flows. This is a welcome result since normalizing flows are notoriously costly to train (40 GPU weeks for the seminal GLOW normalizing flow[4]). Second, the adjoint brings data to the model space and therefore standardizes data size, enabling us to learn a single amortized normalizing flow that can sample the posterior for a variety of imaging configurations. We demonstrate these results using two medical imaging applications.

2. METHODS

2.1 Bayesian Uncertainty

Given our quantity of interest $\mathbf{x} \in \mathcal{X}$ called the model, the forward problem is described by a linear operator $\mathbf{A} : \mathcal{X} \rightarrow \mathcal{Y}$ whose action on $\mathbf{x}$ gives observations $\mathbf{y} \in \mathcal{Y}$. Here, we consider linear problems with additive Gaussian noise term $\boldsymbol{\varepsilon}$—i.e.,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}. \tag{1}$$

Upon observing data $\mathbf{y}$, traditional methods in inverse problems create a single point estimate of the $\mathbf{x}$ that produced $\mathbf{y}$. In the absence of noise and for invertible operators this point estimate is enough to describe the solution of the inverse problem. However, This approach fails to fully characterize the solution space in ill-posed problems,[5] where there is no guarantee of a unique solution. In this scenario, it is important to be able to answer questions such as: Can we trust the solution or is this estimate affected by the null space of the operator? If there is more than one solution, what is the variability between these solution? These are questions that are answered by UQ.

At the forefront of uncertainty quantification (UQ) for inverse problems are Bayesian methods.[6] In a Bayesian framework, we try to find the set or *distribution* of solutions that all explain the data. This set of solutions is encoded by the conditional distribution $p(\mathbf{x} \mid \mathbf{y})$ called the posterior distribution and is the end goal of Bayesian inference . It contains all information needed to estimate $\mathbf{x}$ given $\mathbf{y}$ while providing UQ. Calculating the posterior distribution can be done with two main type of algorithms: first, sampling based algorithms such as Markov chain Monte Carlo (MCMC) algorithms[7] and second, optimization based algorithms such as variational inference (VI).[8] MCMC can be a costly method due to the amount of sampling required,[9] especially in high-dimensional problems. We instead explore variational inference (VI)[8] to sample the posterior since it allows for amortized training costs.

2.2 Variational Inference for Posterior Distribution Learning

To reduce the overall costs of sampling, VI methods reduce the sampling problem into an optimization problem by finding a approximate distribution that best fits the desired distribution.[8] The goodness of fit is typically measured by the Kullback-Leibler (KL) divergence. Among the various VI methods, normalizing flows[10] have been shown to be flexible, efficient and powerful while working on a variety of distributions including conditional distributions.[4, 11] For our case, the goal is to find the normalizing flow f_θ parameterized by θ that makes a learned conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{y})$ approximate the desired posterior distribution $p(\mathbf{x} \mid \mathbf{y})$. We measure the "closeness" with the KL-divergence making the optimization objective

$$\hat{\theta} = \underset{\theta}{\arg\min}\, \mathbb{KL}\left(p_\theta(\mathbf{x} \mid \mathbf{y}) \mid\mid p(\mathbf{x} \mid \mathbf{y})\right). \tag{2}$$

Previous work has been put into VI with normalizing flows that involves costly optimization for each incoming observed data $\mathbf{y}$.[12–15] The optimization is costly because learned parameters of f_θ typically parameterize neural networks that are costly to optimize. On top of that, these VI objectives requires online use of the forward operator $\mathbf{A}$ and its adjoint $\mathbf{A}^*$ during optimization. With this formulation, VI is not efficient enough to enable quick inference. Quick results are particularly important in medical imaging settings since they extend the abilities of a given modality for example by enabling the use of hand-held probes.[16,17] In general, minimizing turnover time makes the difference in providing a timely diagnosis.[18]

A different formulation of VI called amortized variational inference (AVI)[19–23] aims to obtain fast inference on test data without having to re-optimize an objective. This is accomplished by an intensive pretraining phase that optimizes a (KL) divergence based objective averaged over different data $\mathbf{y}$ sampled from the distribution $p(\mathbf{y})$: $\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})} \left[\mathbb{KL}\left(p(\mathbf{x} \mid \mathbf{y}) \,||\, p_{\theta}(\mathbf{x} \mid \mathbf{y})\right)\right]$. One can optimize a simplified objective that only requires samples from the joint distribution $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$.[24,25] For a conditional normalizing flow (CNF) f_{θ}, our objective becomes

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left(\|f_{\theta}(\mathbf{x}^{(n)}; \mathbf{y}^{(n)})\|_2^2 - \log|\det \mathbf{J}_{f_{\theta}}| \right) \tag{3}$$

where N is the size of a training dataset $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ and $\mathbf{J}_{f_{\theta}}$ is the Jacobian of the CNF. This objective is particularly simple to implement with conditional normalizing flows since they allow for tractable computation of the determinant Jacobian term $\det \mathbf{J}_{f_{\theta}}$ by design.

Our goal is to generalize our CNF for a variety of imaging configurations. Then we must consider that a different imaging configuration $\mathbf{A}_i$ can change the size of $\mathbf{y}_i$. These different data sizes can arise from changing receiver settings such as number of receivers or view angles. To handle a set of imaging configurations $\{i = 1 : M\}$ (where M is the number configurations), one would need to define a network f_{θ_i}, for all M configurations i.e. manually defining downsampling layers. Standardizing observations to a single size would allow practitioners to only need to define a single network f_{θ} since now all inputs to the network are the same size regardless of their imaging configurations. The adjoint $\mathbf{A}_i^*$ offers a physics informed way of standardizing data inputs to a single size, namely the size of the model.

Standardizing the size of incoming data is related to the concept of a summary statistic[24,26] so can we interpret the adjoint as a physics-informed summary statistic. We will show that on top of being robust over different imaging configurations, adjoints also accelerate the convergence of the training objective in Equation 3. Before demonstrating these two practical advantages of using the adjoint, we present our main contribution: a theoretical discussion showing when preprocessing data with the adjoint will not affect the result of posterior inference.

2.3 Adjoint Data is Bayesian Sufficient

As noted in the previous section, there are good reasons to use the adjoint operator as a preprocessor. This leads to an important question that we phrase using the language of:[27] can we condition on adjoint data without introducing bias into the inference procedure? We will answer this question in the affirmative by using Proposition 1 from[28] and specifying a class of inverse problems that satisfies the proposition with adjoint preprocessing.

Proposition 1:[28] If $\mathcal{B}$ is injective on the range of Π then $p(\mathbf{x} \mid \mathcal{B}\,\Pi\,\mathbf{y})$ will be equal to $p(\mathbf{x} \mid \mathbf{y})$ if and only if the information lost by observing $\Pi\mathbf{y}$ instead of $\mathbf{y}$ is conditionally independent of $\mathbf{x}$ given $\Pi\mathbf{y}$:

$$p(\mathbf{x} \mid \mathcal{B}\,\Pi\,\mathbf{y}) = p(\mathbf{x} \mid \mathbf{y}) \iff \mathbf{x} \perp\!\!\!\perp \mathbf{y} - \Pi\mathbf{y} \mid \Pi\mathbf{y}. \tag{4}$$

Proposition 1a: Given data $\mathbf{y}$ (created as in Equation (1)), the posterior conditioned on adjoint-preprocessed data $p(\mathbf{x} \mid \mathbf{A}^*\mathbf{y})$ will be equal to the original posterior $p(\mathbf{x} \mid \mathbf{y})$ if the additive noise $\boldsymbol{\varepsilon}$ is Gaussian.

Proof: We use Proposition 1 from[28] and set $\mathcal{B} = \mathbf{A}^*$; and $\Pi = \mathbf{A}\mathbf{A}^+$ where $\mathbf{A}^+$ is the Moore-Penrose inverse. Since $p(\mathbf{x} \mid \mathcal{B}\Pi\mathbf{y}) = p(\mathbf{x} \mid \mathbf{A}^*\mathbf{A}\mathbf{A}^+\mathbf{y}) = p(\mathbf{x} \mid \mathbf{A}^*\mathbf{y})$ then our proof is complete if we show that the following conditional independence is true: $\mathbf{x} \perp\!\!\!\perp \mathbf{y} - \mathbf{A}\mathbf{A}^+\mathbf{y} \mid \mathbf{A}\mathbf{A}^+\mathbf{y}$.

To show this, we note that the noise $\boldsymbol{\varepsilon}$ can be decomposed as the sum of two independent components – one that lives in $\mathrm{ran}(\mathbf{A})$ (the range of $\mathbf{A}$) and another that lives in $\mathrm{ran}(\mathbf{A})^{\perp}$: $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{\mathbf{ran}} + \boldsymbol{\varepsilon}_{\perp}$. Assuming this structure, the observed data is $\mathbf{y} = \mathbf{Ax} + \boldsymbol{\varepsilon}_{\perp} + \boldsymbol{\varepsilon}_{\mathbf{ran}}$. Since $\mathbf{AA}^{+}$ is the orthogonal projector onto $\operatorname{ran}(\mathbf{A})$ then whenever $\mathbf{AA}^{+}$ interacts with $\mathbf{y}$ it will make the contribution from $\boldsymbol{\varepsilon}_{\perp}$ vanish:

$$\begin{aligned}
&\mathbf{x} \perp\!\!\!\perp \mathbf{y} - \mathbf{AA}^{+}\mathbf{y} \mid \mathbf{AA}^{+}\mathbf{y} \\
=\ &\mathbf{x} \perp\!\!\!\perp (\mathbf{Ax} + \boldsymbol{\varepsilon}_{\perp} + \boldsymbol{\varepsilon}_{\mathbf{ran}}) - \mathbf{AA}^{+}(\mathbf{Ax} + \boldsymbol{\varepsilon}_{\perp} + \boldsymbol{\varepsilon}_{\mathbf{ran}}) \mid \mathbf{AA}^{+}(\mathbf{Ax} + \boldsymbol{\varepsilon}_{\perp} + \boldsymbol{\varepsilon}_{\mathbf{ran}}) \\
=\ &\mathbf{x} \perp\!\!\!\perp \mathbf{Ax} + \boldsymbol{\varepsilon}_{\perp} + \boldsymbol{\varepsilon}_{\mathbf{ran}} - \mathbf{AA}^{+}\mathbf{Ax} - \boldsymbol{\varepsilon}_{\mathbf{ran}} \mid \mathbf{AA}^{+}\mathbf{Ax} + \boldsymbol{\varepsilon}_{\mathbf{ran}} \\
=\ &\mathbf{x} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{\perp} \mid \mathbf{Ax} + \boldsymbol{\varepsilon}_{\mathbf{ran}}. \qquad (5)
\end{aligned}$$

By the d-separation criterion,[29] Equation 5 is true because $\boldsymbol{\varepsilon}_{\perp}$ is independent of all other elements, including $\boldsymbol{\varepsilon}_{\mathbf{ran}}$ as per the assumption. Thus we prove that for linear problems with Gaussian additive noise $p(\mathbf{x} \mid \mathbf{A}^{*}\mathbf{y}) = p(\mathbf{x} \mid \mathbf{y})$ meaning adjoint preprocessing does not change the posterior inference.

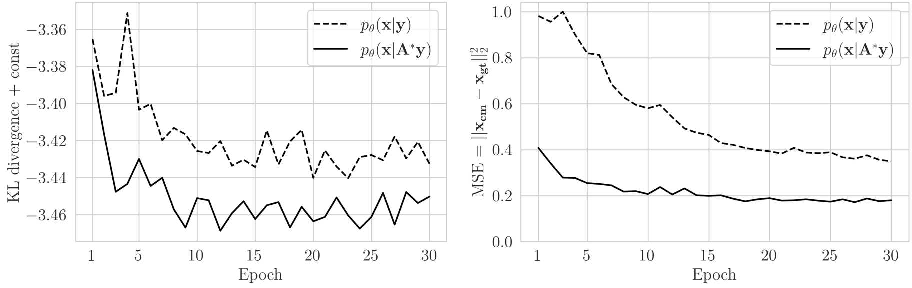
Corruption noise is often approximated as Gaussian additive in medical modalities including our applications in photoacoustic imaging and CT.[30] Our proof does not cover non-Gaussian noise, multiplicative noise or noise correlated with $\mathbf{A}$ or $\mathbf{x}$. We leave those for future work.

3. RESULTS

We show three main results. First, that adjoint preprocessing accelerates the convergence of a conditional normalizing flow training to sample from the posterior distribution. Secondly, we demonstrate that the adjoint operator enables amortization over varying imaging configurations while using the same underling neural network. Finally, we validate the learned UQ by demonstrating posterior consistency through three tests.

3.1 Adjoint Accelerates Convergence:

We design a photoacoustic simulation and CNF architecture to compare two scenarios, namely learning $p_{\theta}(\mathbf{x} \mid \mathbf{y})$ by training on pairs $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ or adjoint preprocessed learning of $p_{\theta}(\mathbf{x} \mid \mathbf{A}^{*}\mathbf{y})$ with $\mathcal{D}^{*} = \{(\mathbf{x}^{(n)}, \mathbf{A}^{*}\mathbf{y}^{(n)})\}_{n=1}^{N}$. As noted in our motivations, creating a CNF f_{θ} that can accept raw data $\mathbf{y}$ as an input is a laborious task but we do this for one imaging configuration to provide a fair comparison. This task involves manually defining downsampling layers in the CNF that bring $\mathbf{y}$ to the appropriate dimensionality. Then we can proceed to train two CNF's where both underlying networks have the same architectures (with addition of the downsampling layer) and are trained using the same hyperparameters.

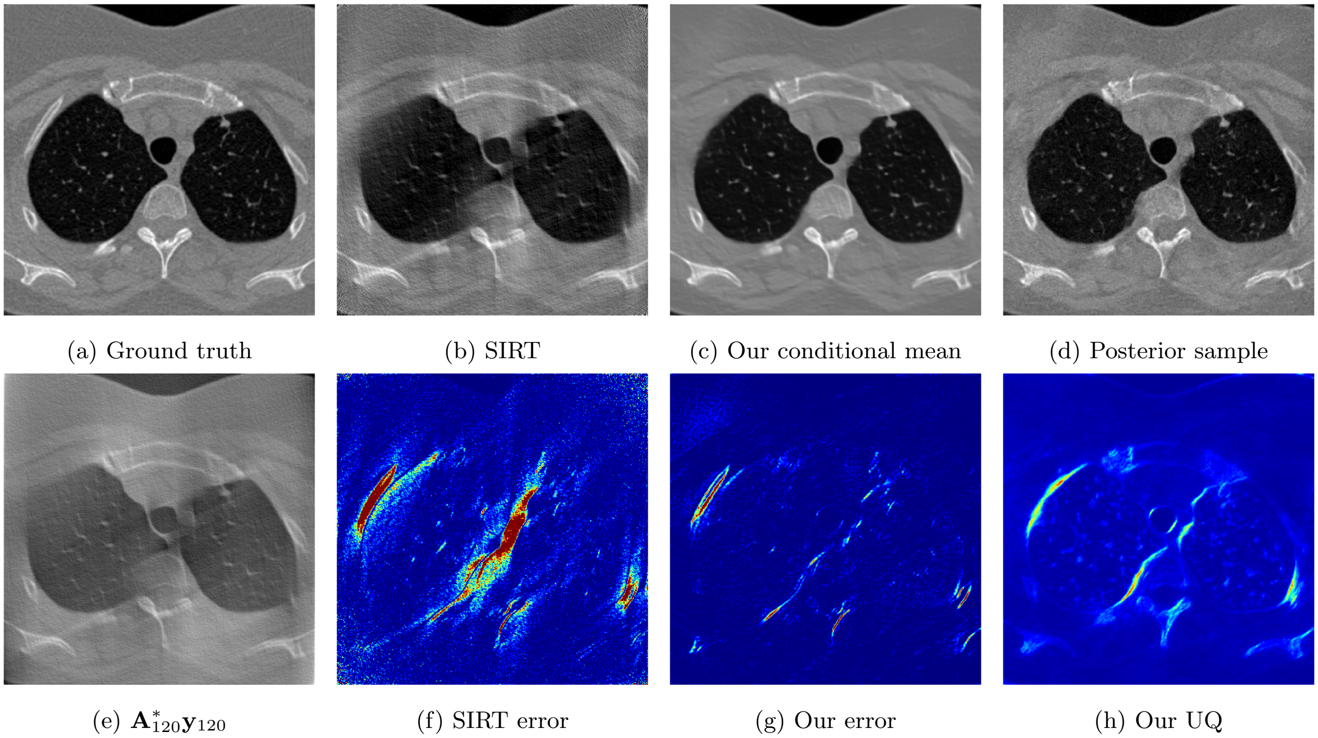


(a) (b)

Figure 1: Convergence plots. (a) Posterior learning objective for data without (dashed) and with preprocessing with the adjoint $\mathbf{A}^{*}$. The adjoint accelerates convergence. (b) MSE of the conditional mean yielding improved Bayesian inference with less training time (juxtapose solid and dashed line for raw and preprocessed data).

Figure 1 shows that for equivalent base architectures and training, learning $p_{\theta}(\mathbf{x}|\mathbf{A}^{*}\mathbf{y})$ is accelerated compared to learning $p_{\theta}(\mathbf{x}|\mathbf{y})$. We also plot the mean squared error (MSE) between the calculated conditional mean $\mathbf{x_{cm}}$ and the ground truth $\mathbf{x_{gt}}$. While MSE is not the training objective, it is still an important proxy since the conditional mean of the true posterior is the one that gives the lowest expected error.[31,32] Here and throughout, the conditional mean $\mathbf{x_{cm}}$ and the variance (our UQ) is estimated using an average of 64 generated samples from the posterior. The training logs in Figure 1 are averages of an unseen validation set N_{val}=192 created by a 10% split from the training dataset of N=2048 pairs.

We emphasize that our posterior sampling after training is fast. After applying the adjoint, the CNF is conditioned on $\mathbf{A}^*\mathbf{y}$ then the user generates the desired quantity of posterior samples (10 millisec/sample on our GPU). The time to create a point estimate with the conditional mean $\mathbf{x}_{\mathbf{cm}}$ (around 2 seconds with 64 posterior samples) is favorable compared to traditional least-squares approaches that require several forward and adjoint evaluations.


(a) Ground truth (b) SIRT (c) Our conditional mean (d) Posterior sample

(e) $\mathbf{A}^*_{120}\mathbf{y}_{120}$ (f) SIRT error (g) Our error (h) Our UQ

Figure 2: Computer tomography 360×360 images with uncertainty quantification. (a) Ground truth image; (b) Reconstructed image using SIRT baseline with 300 iterations; (c) Our image reconstruction made by averaging 64 samples from the learned posterior; (d) A single posterior sample from our method; (e) The adjoint data that has been brought to image space; (f) Error made by the SIRT baseline; (g) Error made by our conditional mean; (h) Our UQ calculated from the variance of posterior samples; Note: error plots and the UQ plot have the same colorbar limits $[0, 0.014]$

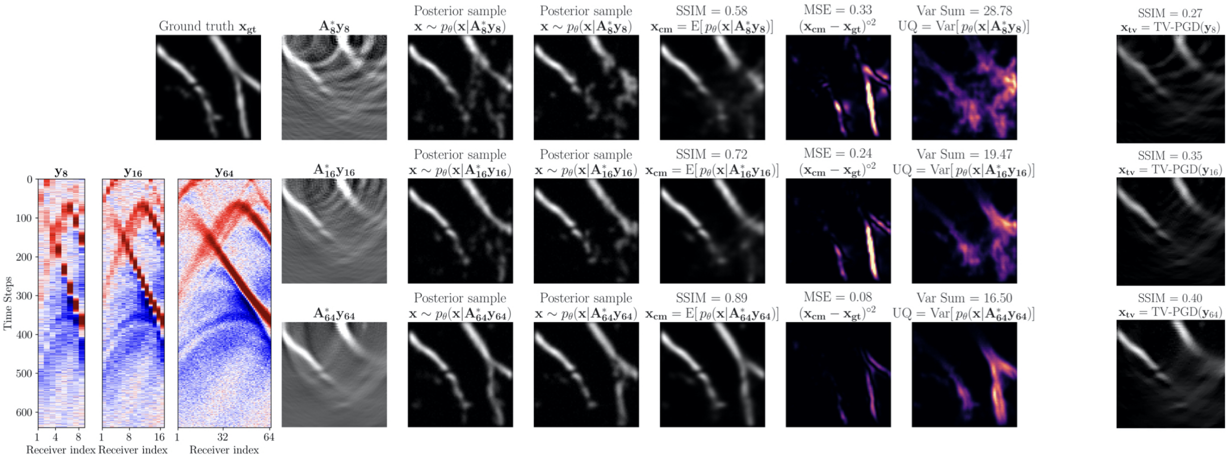
In Figure 2, we show images of results from posterior sampling on limited-view CT after training with adjoint preprocessing. For training and testing, we used the lodopab-ct dataset in the original resolution of 360×360.[33] For CT forward and adjoint simulations we use.[34] Our experimental setup, follows[35] for limited-view CT with SNR = 40dB additive Gaussian noise. It has been said that bijective methods were not viable for a resolution of 256×256.[35] We did not find this to be the case, our conditional normalizing flow is completely bijective and the implementation from InvertibleNetworks.jl[36] we did not see any out-of-memory problems training this method on the original 360×360 resolution.

For the baseline CT, we use the simultaneous iterative reconstruction technique (SIRT) as described in.[37, 38] Compared to the baseline Figure 2b, our method Figure 2c produces cleaner images that can deal with the substantial null-spaces due to limited-views. Importantly, structures in the null-space are illuminated by our UQ Figure 2h pointing to reduced confidence in those areas.

3.2 Adjoint Generalizes Different Imaging Configurations:

Data that is preprocessed with the adjoint $\mathbf{A}^*\mathbf{y}$ will always live in the space of the image. Thus we can use a single network to learn the posterior for different imaging configurations by augmenting our training dataset with examples from the desired configurations $(\mathbf{x}^{(n)}, \mathbf{A}_i^*\mathbf{y}_i^{(n)})\}_{n=1}^{N}$.

Receivers in photoacoustic imaging: We generalize over four photoacoustic imaging configurations consisting of data collected with 8, 16, 32, and 64 receivers. These different configurations are encoded by forward and adjoint operators $\mathbf{A}_i, \mathbf{A}_i^*$ with $i \in \{8, 16, 32, 64\}$. We train our CNF with N=2048 examples for each imaging configuration. Training took 14 hours on P1000 4GB GPU.



Ground truth $\mathbf{x}_{\mathrm{gt}}$
$\mathbf{A}_8^*\mathbf{y}_8$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)$
SSIM = 0.58 $\mathbf{x}_{\mathrm{cm}} = \mathrm{E}[p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)]$
MSE = 0.33 $(\mathbf{x}_{\mathrm{cm}} - \mathbf{x}_{\mathrm{gt}})^{\circ 2}$
Var Sum = 28.78 UQ = $\mathrm{Var}[p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)]$
SSIM = 0.27 $\mathbf{x}_{\mathrm{tv}}$ = TV-PGD($\mathbf{y}_8$)
$\mathbf{y}_8$
$\mathbf{y}_{16}$
$\mathbf{y}_{64}$
$\mathbf{A}_{16}^*\mathbf{y}_{16}$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})$
SSIM = 0.72 $\mathbf{x}_{\mathrm{cm}} = \mathrm{E}[p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})]$
MSE = 0.24 $(\mathbf{x}_{\mathrm{cm}} - \mathbf{x}_{\mathrm{gt}})^{\circ 2}$
Var Sum = 19.47 UQ = $\mathrm{Var}[p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})]$
SSIM = 0.35 $\mathbf{x}_{\mathrm{tv}}$ = TV-PGD($\mathbf{y}_{16}$)
Time Steps
0
100
200
300
400
500
600
$\mathbf{A}_{64}^*\mathbf{y}_{64}$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})$
SSIM = 0.89 $\mathbf{x}_{\mathrm{cm}} = \mathrm{E}[p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})]$
MSE = 0.08 $(\mathbf{x}_{\mathrm{cm}} - \mathbf{x}_{\mathrm{gt}})^{\circ 2}$
Var Sum = 16.50 UQ = $\mathrm{Var}[p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})]$
SSIM = 0.40 $\mathbf{x}_{\mathrm{tv}}$ = TV-PGD($\mathbf{y}_{64}$)
1 4 8
Receiver index
1 8 16
Receiver index
1 32 64
Receiver index

(a) (b)

Figure 3: (a) Amortized training of neural networks capable of sampling from posterior distributions for differently sized observations $\mathbf{y}_i$. As the number of receivers is increased, the samples show posterior contraction, a Bayesian phenomenon[39] that says increasing the amount of data should decrease the width of the posterior. (b) The baseline method (TV-projected gradient descent) fails to image vertical vessels.

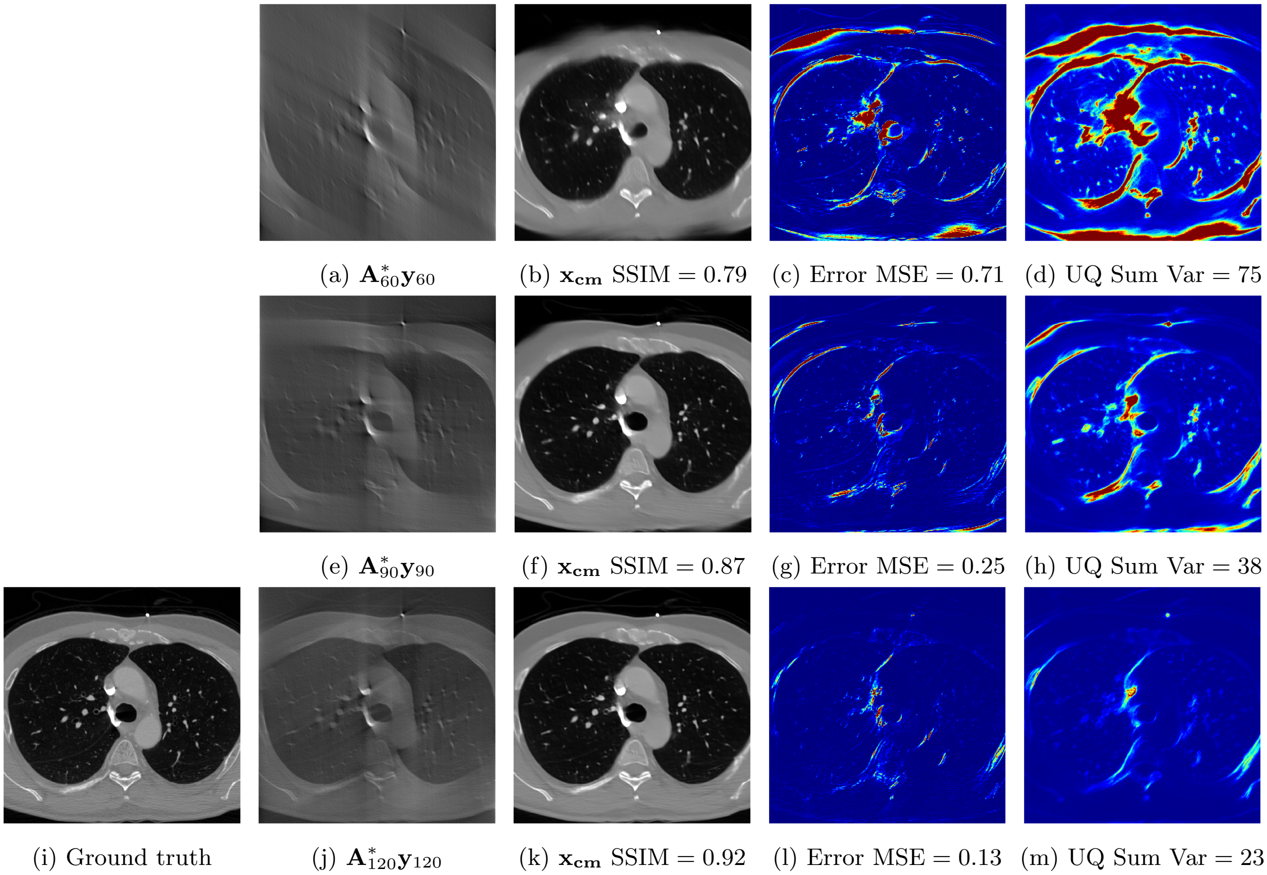
After training, we sample from the learned posterior for unseen test data examples, $\mathbf{y}_i$ for varying numbers of receivers. The results demonstrate proper posterior contraction.[39] Visually, this is confirmed in Figure 3a since uncertainty (quantified via pointwise variance) goes down when we increase receivers from 16 to 64. To quantitatively capture the global variation of these UQ images, we use the sum of pointwise variances: Var Sum = $\|\mathrm{Var}\|_1$ or equivalently the trace of the covariance matrix.[40] As the uncertainty reduces, the quality of the estimates increase, thus the posterior is contracting on the ground truth. We quantitatively verify this statement in the next section. Importantly, uncertainty is high where we expect, namely for vessels that are close to being vertical. These vertical events are in the null-space of the forward operator because the receivers are only located at the top of the model.

We compare our point estimate $\mathbf{x}_{\mathbf{cm}}$ with TV-projected gradient descent (TV-PGD)[41,42] $\mathbf{x}_{\mathbf{tv}}$. Timing and quality metrics averaged over a test set of 96 samples are in Table 1. Across all receiver configurations, we show better Structural Similar Index Measure (SSIM). Also our method, produces the image reconstruction in less time.

Photoacoustic imaging	Timing (Seconds)	Quality metric (SSIM)			
	$N_{\mathrm{rec}} = 64$	$N_{\mathrm{rec}} = 8$	$N_{\mathrm{rec}} = 16$	$N_{\mathrm{rec}} = 32$	$N_{\mathrm{rec}} = 64$
TV-proj GD $\mathbf{x}_{\mathbf{tv}}$	16.78	0.27	0.36	0.42	0.44
Our conditional mean $\mathbf{x}_{\mathbf{cm}}$	**1.72**	**0.62**	**0.74**	**0.80**	**0.81**

Table 1: Photoacoustic image reconstruction timing and quality metric comparison.

View angles in computer tomography: To demonstrate generalization in CT, we train a single normalizing flow to sample from the posterior of three different quantities of views 60, 90 and 120 degrees. We add 40dB Gaussian noise to measurements. The raw data measurements $\mathbf{y}_{60}, \mathbf{y}_{90}, \mathbf{y}_{120}$ are shown in Appendix Figure 6. To train we use 4000 images for each of the three configurations for a total of 12000 training images.


(a) $\mathbf{A}_{60}^{*}\mathbf{y}_{60}$ (b) $\mathbf{x_{cm}}$ SSIM = 0.79 (c) Error MSE = 0.71 (d) UQ Sum Var = 75

(e) $\mathbf{A}_{90}^{*}\mathbf{y}_{90}$ (f) $\mathbf{x_{cm}}$ SSIM = 0.87 (g) Error MSE = 0.25 (h) UQ Sum Var = 38

(i) Ground truth (j) $\mathbf{A}_{120}^{*}\mathbf{y}_{120}$ (k) $\mathbf{x_{cm}}$ SSIM = 0.92 (l) Error MSE = 0.13 (m) UQ Sum Var = 23

Figure 4: Generalizing computer tomography over different view angles. (a,b,c,d) The first row shows results of our method for 60 view angles. (e,f,g,h) The second row shows results of our method for 90 view angles. (j,k,l,m) The third row shows results of our method for 120 view angles.

The results of the CT application show similar behaviour to the photoacoustic case. Bayesian contraction is clearly shown in Figure 4 since the uncertainty is reduced as more angles are observed. Also the uncertainty is physically consistent with our understanding of the imaging system since higher uncertainty is placed on the view angles that are unseen. We compare the quality of our CT generalized normalizing flow by comparing with the SIRT baseline. In Table 2, we show timing results for the maximum amount of angles and SSIM metrics for all tested angles. SSIM metrics are averages over a test set of 50 images. Our method is faster since the SIRT needs various applications of the forward and adjoint CT simulation while our method uses a single adjoint.

Computed tomography	Timing (Seconds)	Quality metric (SSIM)		
	$N_{\text{ang}} = 120$	$N_{\text{ang}} = 60$	$N_{\text{ang}} = 90$	$N_{\text{ang}} = 120$
Simultaneous iterative reconstruction $\mathbf{x}_{\text{SIRT}}$	41.32	0.59	0.69	0.75
Our conditional mean $\mathbf{x_{cm}}$	**1.21**	**0.78**	**0.84**	**0.88**

Table 2: Limited-view computer tomography image reconstruction timing and quality metric comparison.

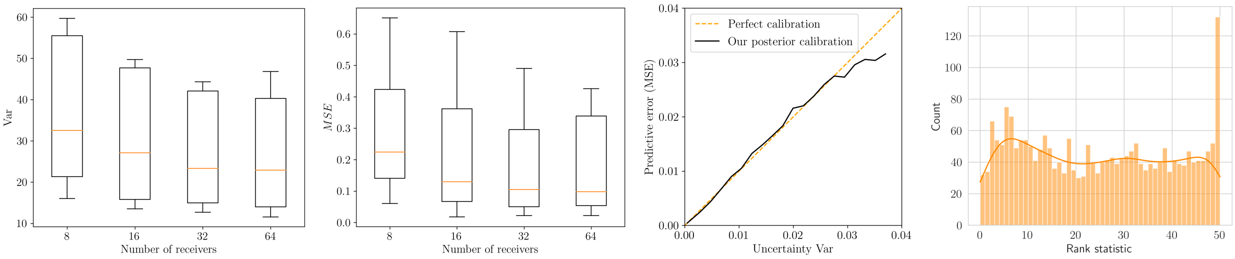
3.3 Validation of Uncertainty Quantification:

Since our problem does not have Gaussian priors, we can not analytically verify that our converged posteriors have reached the ground truth posterior. We can use three tests on the photoacoustic application to check that our posteriors are consistent and useful.

(i) posterior contraction by testing the CNF for different amount of data to check whether the posterior demonstrates contraction to the ground truth when increasing the amount of observed data. This contraction ultimately points to posterior consistency;[39]

(ii) posterior calibration by checking whether our UQ correlates with regions in the image with large errors. This important check of UQ is called calibration and is established qualitatively by visually juxtaposing errors with UQ in Figure 3a. For a more quantitative test, we plot a calibration line by using σ-scaling;[43, 44]

(iii) simulation-based calibration (SBC) by testing for uniformity in the rank statistic when comparing various samples drawn from the proposed posterior with the known prior.[24, 45]
The results of these three tests are included in Figure 5 and show our learned posterior is consistent and approximates the true posterior. For this reason, we argue that a practitioner is justified in using our posterior for uncertainty quantification.



(a) (b) (c) (d)

Figure 5: Validation of uncertainty quantification (a) Posterior contraction when increasing the amount of data. (b) Posterior contraction towards the ground truth as measured by MSE. (c) Our posterior calibration is close to perfect calibration showing that our UQ is correlated with error made. (d) The uniformity of the SBC test shows that our marginalized posterior samples recover the prior distribution.

4. CONCLUSIONS

For linear operators and Gaussian noise, we prove that adjoint preprocessing posterior is equivalent to the original posterior $p(\mathbf{x}|\mathbf{A}^*\mathbf{y}) = p(\mathbf{x}|\mathbf{y})$. Although the distributions are ultimately the same, learning them poses different computational burdens for ML training. We showed that the adjoint accelerates convergence of CNFs for AVI. We also demonstrate that the adjoint allows us to train a single network to handle many different imaging configurations thus saving costs associated with designing network architectures for individual configurations. Our amortized posterior gives physically meaningful uncertainties that we also validate.

REFERENCES

[1] Ongie, G., Jalal, A., Metzler, C. A., Baraniuk, R. G., Dimakis, A. G., and Willett, R., "Deep learning techniques for inverse problems in imaging," *IEEE Journal on Selected Areas in Information Theory* **1**(1), 39–56 (2020).

[2] Antun, V., Renna, F., Poon, C., Adcock, B., and Hansen, A. C., "On instabilities of deep learning in image reconstruction and the potential costs of ai," *Proceedings of the National Academy of Sciences* **117**(48), 30088–30095 (2020).

[3] Bhadra, S., Kelkar, V. A., Brooks, F. J., and Anastasio, M. A., "On hallucinations in tomographic image reconstruction," *IEEE transactions on medical imaging* **40**(11), 3249–3260 (2021).

[4] Kingma, D. P. and Dhariwal, P., "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems* **31** (2018).

[5] Hadamard, J., "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton university bulletin* , 49–52 (1902).

[6] Tarantola, A., [*Inverse problem theory and methods for model parameter estimation*], SIAM (2005).

[7] Robert, C. P., Casella, G., and Casella, G., [*Monte Carlo statistical methods*], vol. 2, Springer (1999).

[8] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D., "Variational inference: A review for statisticians," *Journal of the American statistical Association* **112**(518), 859–877 (2017).

[9] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., [*Bayesian data analysis*], Chapman and Hall/CRC (1995).

[10] Dinh, L., Sohl-Dickstein, J., and Bengio, S., "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803* (2016).

[11] Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U., "Analyzing inverse problems with invertible neural networks," *arXiv preprint arXiv:1808.04730* (2018).

[12] Whang, J., Lindgren, E., and Dimakis, A., "Composing normalizing flows for inverse problems," in [*International Conference on Machine Learning*], 11158–11169, PMLR (2021).

[13] Sun, H. and Bouman, K. L., "Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging," *arXiv preprint arXiv:2010.14462* **9** (2020).

[14] Rizzuti, G., Siahkoohi, A., Witte, P. A., and Herrmann, F. J., "Parameterizing uncertainty by deep invertible networks: An application to reservoir characterization," in [*Seg technical program expanded abstracts 2020*], 1541–1545, Society of Exploration Geophysicists (2020).

[15] Zhao, X., Curtis, A., and Zhang, X., "Bayesian seismic tomography using normalizing flows," *Geophysical Journal International* **228**(1), 213–239 (2022).

[16] Dima, A. and Ntziachristos, V., "Non-invasive carotid imaging using optoacoustic tomography," *Optics express* **20**(22), 25044–25057 (2012).

[17] Buehler, A., Kacprowicz, M., Taruttis, A., and Ntziachristos, V., "Real-time handheld multispectral optoacoustic imaging," *Optics letters* **38**(9), 1404–1406 (2013).

[18] Bauer, S., Seitel, A., Hofmann, H., Blum, T., Wasza, J., Balda, M., Meinzer, H.-P., Navab, N., Hornegger, J., and Maier-Hein, L., "Real-time range imaging in health care: a survey," in [*Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*], 228–254, Springer (2013).

[19] Kovachki, N., Baptista, R., Hosseini, B., and Marzouk, Y., "Conditional sampling with monotone gans," *arXiv preprint arXiv:2006.06755* (2020).

[20] Kruse, J., Detommaso, G., Scheichl, R., and Köthe, U., "Hint: Hierarchical invertible neural transport for density estimation and Bayesian inference," *arXiv preprint arXiv:1905.10687* (2019).

[21] Orozco, R., Siahkoohi, A., Rizzuti, G., van Leeuwen, T., and Herrmann, F. J., "Photoacoustic imaging with conditional priors from normalizing flows," in [*NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*], (2021).

[22] Siahkoohi, A., Orozco, R., Rizzuti, G., and Herrmann, F. J., "Wave-equation-based inversion with amortized variational Bayesian inference," *arXiv preprint arXiv:2203.15881* (2022).

[23] Gershman, S. and Goodman, N., "Amortized inference in probabilistic reasoning," in [*Proceedings of the annual meeting of the cognitive science society*], **36**(36) (2014).

[24] Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U., "Bayesflow: Learning complex stochastic models with invertible neural networks," *IEEE transactions on neural networks and learning systems* (2020).

[25] Siahkoohi, A., Rizzuti, G., Orozco, R., and Herrmann, F. J., "Reliable amortized variational inference with physics-based latent distribution correction," *arXiv preprint arXiv:2207.11640* (2022).

[26] Deans, M. C., "Maximally informative statistics for localization and mapping," in [*Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*], **2**, 1824–1829, IEEE (2002).

[27] Baptista, R., Cao, L., Chen, J., Ghattas, O., Li, F., Marzouk, Y. M., and Oden, J. T., "Bayesian model calibration for block copolymer self-assembly: Likelihood-free inference and expected information gain computation via measure transport," *arXiv preprint arXiv:2206.11343* (2022).

[28] Adler, J., Lunz, S., Verdier, O., Schönlieb, C.-B., and Öktem, O., "Task adapted reconstruction for inverse problems," *Inverse Problems* **38**(7), 075006 (2022).

[29] Pearl, J., [*Probabilistic reasoning in intelligent systems: networks of plausible inference*], Morgan kaufmann (1988).

[30] Gravel, P., Beaudoin, G., and De Guise, J. A., "A method for modeling noise in medical images," *IEEE Transactions on medical imaging* **23**(10), 1221–1232 (2004).

[31] Banerjee, A., Guo, X., and Wang, H., "On the optimality of conditional expectation as a bregman predictor," *IEEE Transactions on Information Theory* **51**(7), 2664–2669 (2005).

[32] Adler, J. and Öktem, O., "Deep Bayesian inversion," *arXiv preprint arXiv:1811.05910* (2018).

[33] Leuschner, J., Schmidt, M., Baguer, D. O., and Maaß, P., "The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods," *arXiv preprint arXiv:1910.01113* (2019).

[34] Van Aarle, W., Palenstijn, W. J., De Beenhouwer, J., Altantzis, T., Bals, S., Batenburg, K. J., and Sijbers, J., "The astra toolbox: A platform for advanced algorithm development in electron tomography," *Ultramicroscopy* **157**, 35–47 (2015).

[35] Khorashadizadeh, A., Kothari, K., Salsi, L., Harandi, A. A., de Hoop, M., and Dokmani'c, I., "Conditional injective flows for bayesian imaging," *arXiv preprint arXiv:2204.07664* (2022).

[36] Philipp Witte, Gabrio Rizzuti, M. L. A. S. and Herrmann, F., "A julia framework for invertible neural networks," (2020).

[37] 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).

[38] Geyer, L. L., Schoepf, U. J., Meinel, F. G., Nance Jr, J. W., Bastarrika, G., Leipsic, J. A., Paul, N. S., Rengo, M., Laghi, A., and De Cecco, C. N., "State of the art: iterative ct reconstruction techniques," *Radiology* **276**(2), 339–357 (2015).

[39] Ghosal, S. and Van der Vaart, A., [*Fundamentals of nonparametric Bayesian inference*], vol. 44, Cambridge University Press (2017).

[40] Oja, H., "Affine invariant multivariate sign and rank tests and corresponding estimates: a review," *Scandinavian Journal of Statistics* **26**(3), 319–343 (1999).

[41] Zhang, Y., Wang, Y., and Zhang, C., "Total variation based gradient descent algorithm for sparse-view photoacoustic image reconstruction," *Ultrasonics* **52**(8), 1046–1055 (2012).

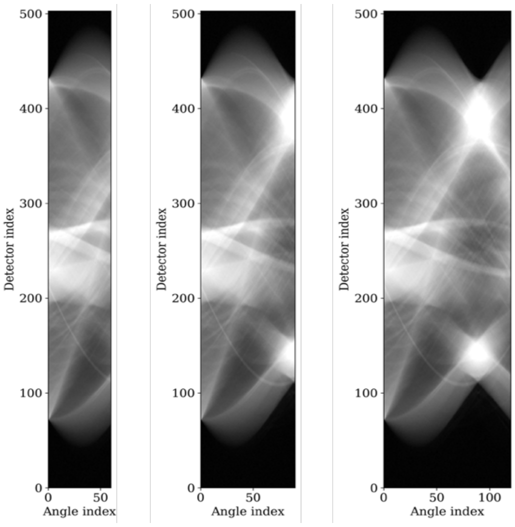
[42] Schwab, J., Antholzer, S., Nuster, R., and Haltmeier, M., "Real-time photoacoustic projection imaging using deep learning," *arXiv preprint arXiv:1801.06693* (2018).

[43] Laves, M.-H., Ihler, S., Fast, J. F., Kahrs, L. A., and Ortmaier, T., "Well-calibrated regression uncertainty in medical imaging with deep learning," in [*Medical Imaging with Deep Learning*], 393–412, PMLR (2020).

[44] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., "On calibration of modern neural networks," in [*International Conference on Machine Learning*], 1321–1330, PMLR (2017).

[45] Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A., "Validating Bayesian inference algorithms with simulation-based calibration," *arXiv preprint arXiv:1804.06788* (2018).

APPENDIX A. RAW OBSERVATIONS FOR LIMITED-VIEW COMPUTER TOMOGRAPHY



(a) Data recorded with increasing view angles from left to right: $\mathbf{y}_{60}$, $\mathbf{y}_{90}$ and $\mathbf{y}_{120}$

Figure 6: Experimental setup used to show generalization to computed tomography angles. (a) Sinogram data with increasing angles;