A Probabilistic Deep Image Prior for Computational Tomography

Riccardo Barbano

a joint work with

B. Jin, S. Arridge, J.M. Hernández-Lobato, J. Antoran, J. Leuschner

A brief primer on inverse problems

forward model

$$\mathbf{y}_\delta = \mathrm{A}\mathbf{x} + \boldsymbol{\eta}$$

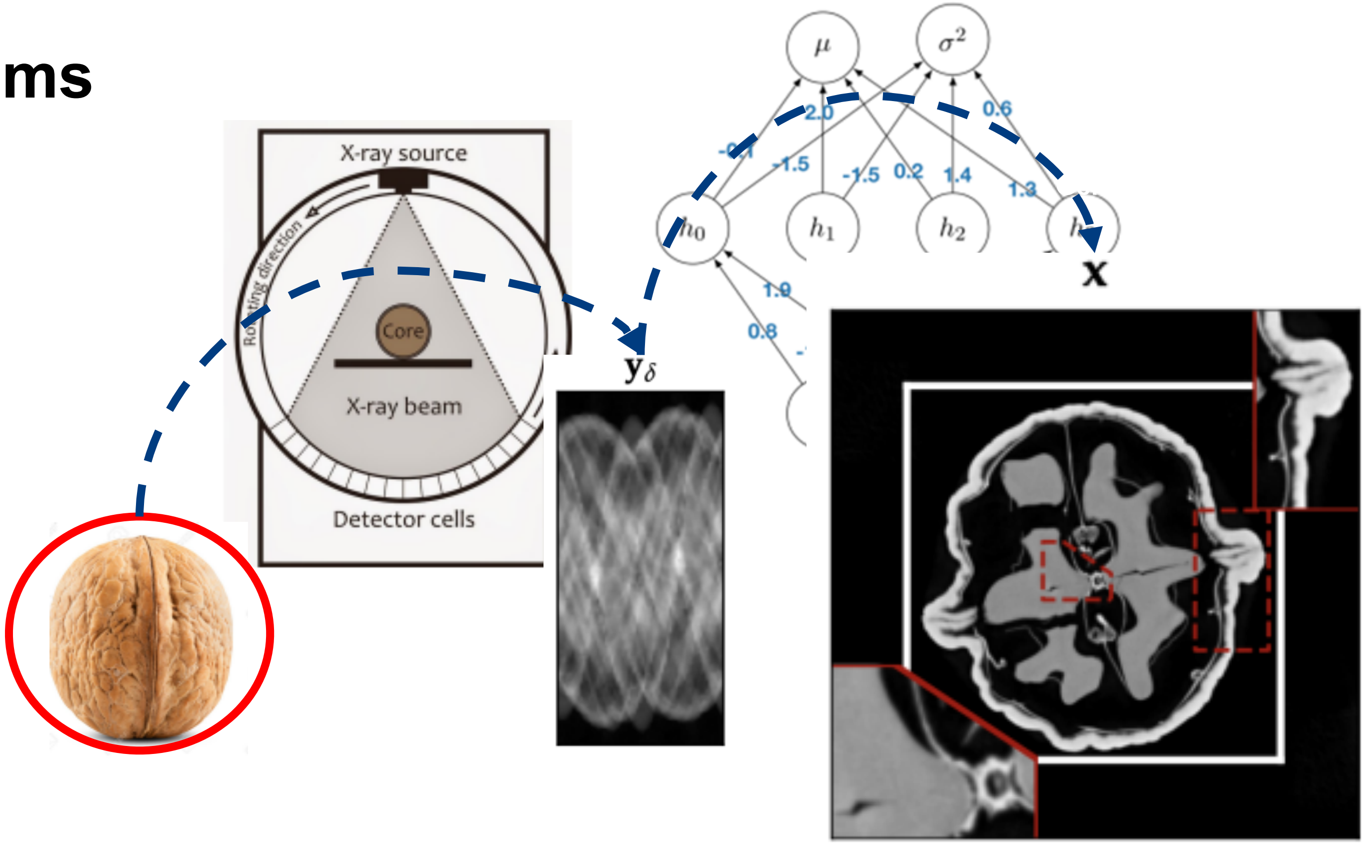
measurement data noise

retrieving the unknown

$$\min_{\mathbf{x}\in\mathbb{X}} \{\mathrm{l}(\mathbf{x}) := \mathrm{d}(\mathrm{A}\mathbf{x}, \mathbf{y}_\delta) + \lambda \mathrm{r}(\mathbf{x})\}$$

exploiting a priori knowledge

$$\mathrm{TV}(\mathbf{x}) = \sum_{i,j} |X_{i,j} - X_{i+1,j}| + \sum_{i,j} |X_{i,j} - X_{i,j+1}|,$$



Can we design a **neural Bayesian** prior to solve this task?

A Bayesian approach to quantify uncertainty

in the **Bayesian framework**, instead of finding a **single** best image, the posterior distribution,

model evidence

prior beliefs

$$p(\mathbf{x}|\mathbf{y}_\delta) = p(\mathbf{y}_\delta)^{-1} p(\mathbf{y}_\delta|\mathbf{x}) p(\mathbf{x}),$$

objective hyperparameters optimization

scores every image according to their **agreement** with the **observation** and the **prior belief,** and the **discrepancy** among these solutions acts as an **estimate** of **uncertainty**

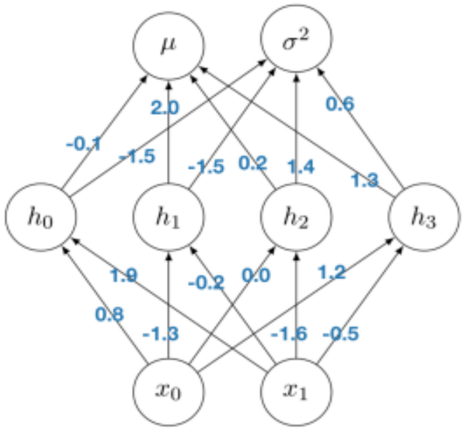
Tools for modelling uncertainty in deep learning

1. **Place** a prior distribution over NN parameters

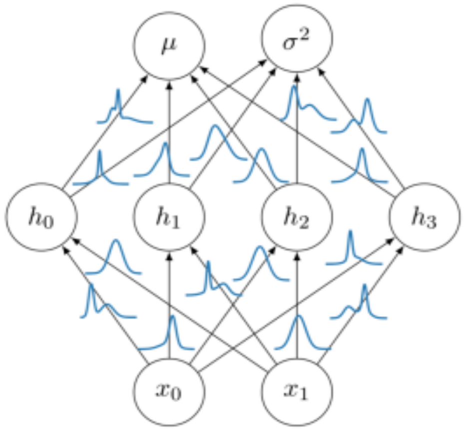
2. **Define** some likelihood function to characterize the agreement of the NN function with the observations

3. **Update** the weight distribution using Bayes' rule

NN



Bayesian NN



The Laplace approximation

1. **Train** the neural network (find a mode)

$$-\log p(\mathbf{x}^i|\mathbf{f}(\mathrm{A}^\dagger\mathbf{y}^i_\delta;\boldsymbol{\theta})) \quad -\log p(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}^\star \in \underset{\boldsymbol{\theta}\in\mathbb{R}^{d_\theta}}{\operatorname{argmin}} \left\{ \mathtt{l}(\mathcal{D};\boldsymbol{\theta}) := \sum_{i=1}^{N} \underline{\mathrm{d}(\mathbf{x}^i, \mathbf{x}_*; \boldsymbol{\theta})} + \underline{\mathrm{r}(\boldsymbol{\theta})} \right\}$$

2. **Approximate** the (intractable) posterior distribution over the NN parameters

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}; \underline{\boldsymbol{\theta}^\star, \Sigma_{\boldsymbol{\theta}}}), \quad \text{with} \quad \Sigma_{\boldsymbol{\theta}} = -\left[\nabla^2_{\boldsymbol{\theta}}\mathtt{l}(\mathcal{D};\boldsymbol{\theta})_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right]^{-1}$$

3. **Further…** approximate

$$q_{\mathrm{GGN}}(\boldsymbol{\theta}) := \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, \left(\mathrm{J}(\mathrm{A}^\dagger\mathbf{y}_\delta;\boldsymbol{\theta})^\top \boldsymbol{\Lambda}(\mathrm{A}^\dagger\mathbf{y}_\delta;\mathbf{f}(\boldsymbol{\theta}))\mathrm{J}(\mathrm{A}^\dagger\mathbf{y}_\delta;\boldsymbol{\theta}) + \mathrm{S}_0^{-1}\right)^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right)$$

… further approximate …

(MacKay, 1992)

The linearized Laplace method

A Gaussian can be a **very poor approximation** to the NN's posterior distribution … yet **experimentally** it is a **very good** for a linear model

1. **Linearization** of the underlying BNN

Jacobian acts as **basis expansion**

$$\mathrm{h}(\boldsymbol{\theta}^{\star}) := \mathrm{f}(\mathrm{A}^{\dagger}\mathbf{y}_{\delta};\boldsymbol{\theta}^{\star}) + \underline{\mathrm{J}(\mathrm{A}^{\dagger}\mathbf{y}_{\delta};\boldsymbol{\theta}^{\star})}(\boldsymbol{\theta}-\boldsymbol{\theta}^{\star})$$

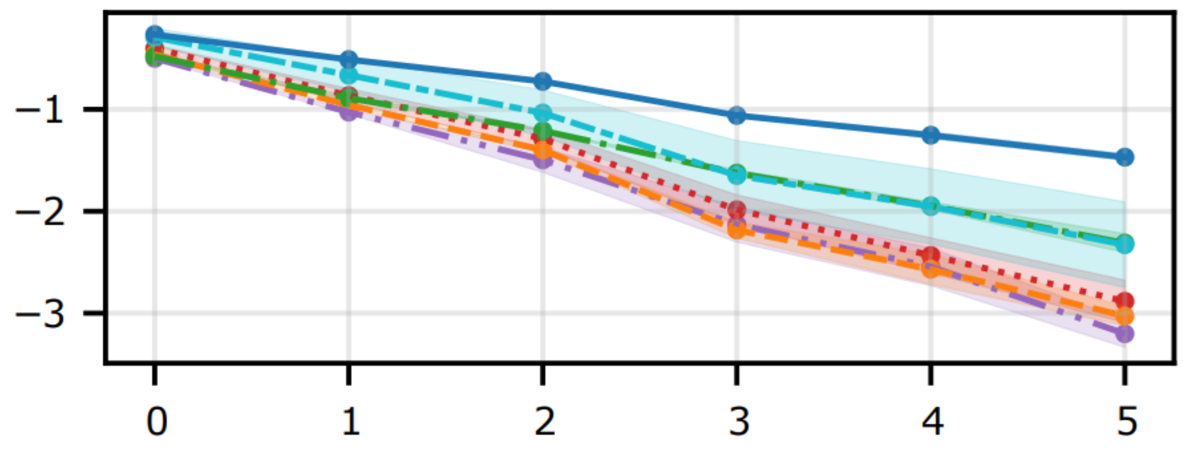
inducing a (GP) deep image prior

$$\mathbf{x} \sim \mathcal{N}(\mathrm{h}(\mathbf{0}),\ \underline{\mathrm{J}\Sigma_{\boldsymbol{\theta}}(\ell,\sigma^2_{\boldsymbol{\theta}})\mathrm{J}^{\top}})$$

we can now perform approximate inference in the **GP model or** solve it in **closed-form** for regression!

(Immer, 2021)

Image classification under **distribution shift**



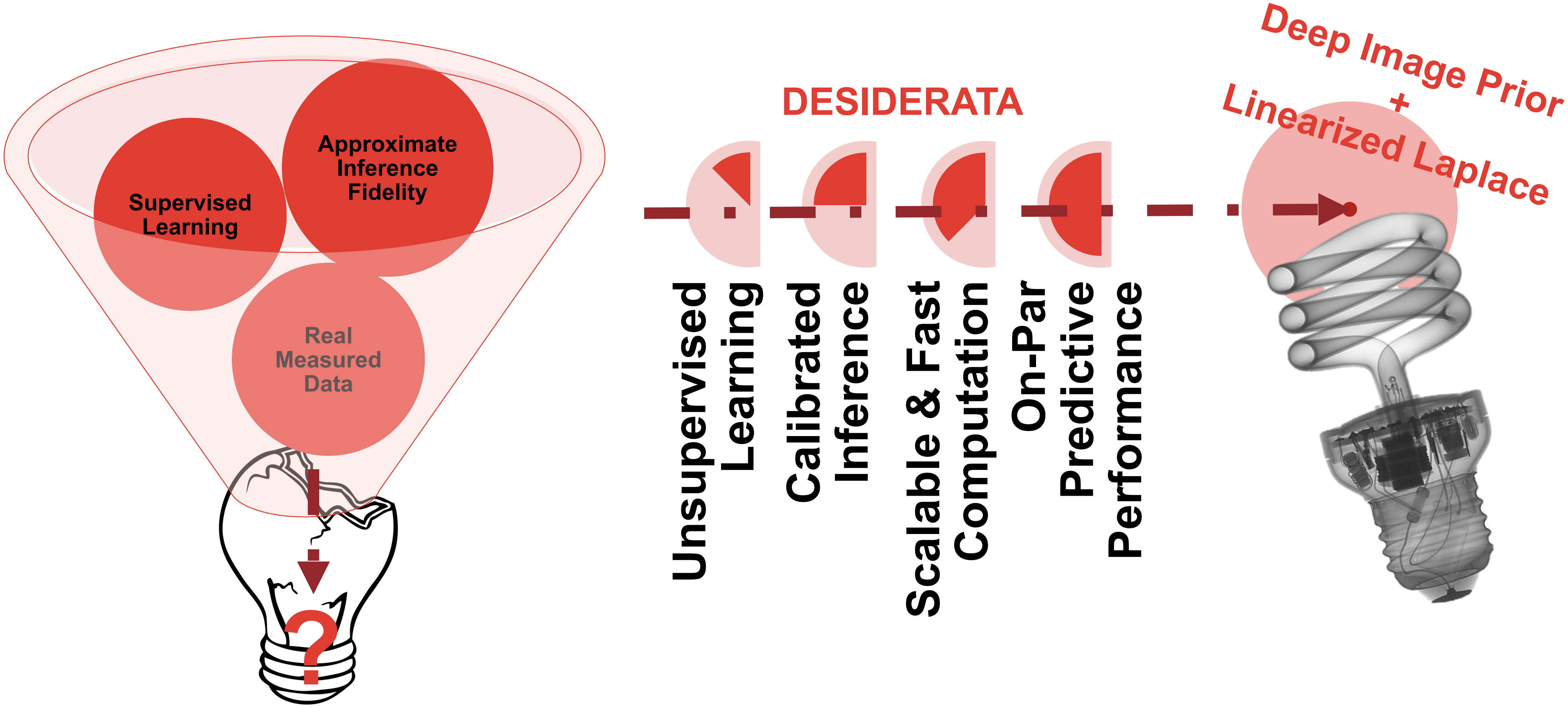
Model:
ResNet-18 with **11M** weights

Inference:
Lin Laplace Subnetwork
(Daxberger et. al. 2021)

Baselines:
- MAP
- Diagonal Laplace
- MC Dropout (Gal 2016)
- Deep Ensembles (Lakshminarayanan 2017)
- SWAG (Maddox 2019)

"Bayesian Deep Learning via Subnetwork Inference"

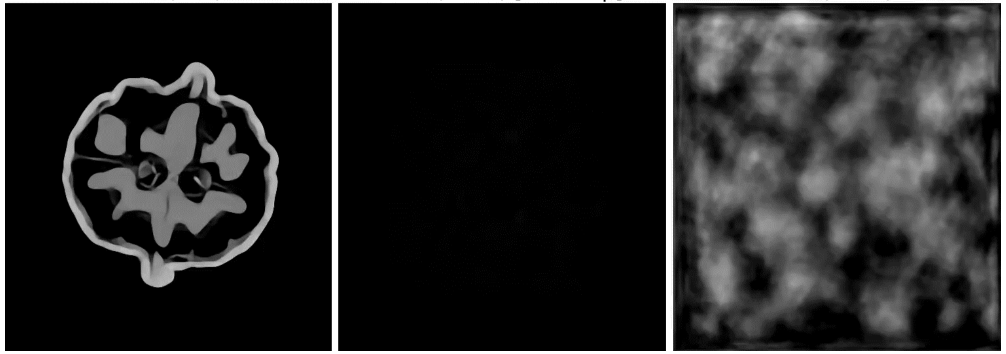
The deep image prior for inverse problems



Rethinking DIP – is DIP in need of a good education ?

Our Method

Baseline



Stage 1: Supervised pretraining on synthetic training data

$$\boldsymbol{\theta}_{\mathrm{s}}^{\star} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}}{\operatorname{argmin}} \left\{ \mathrm{l}_{\mathrm{s}}(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{f}(\mathrm{A}^{\dagger} \mathbf{y}_{\delta}^{i}; \boldsymbol{\theta}) - \mathbf{x}^{i} \|_2^2 \right\}$$

Stage 2: Unsupervised fine-tuning on real measurement

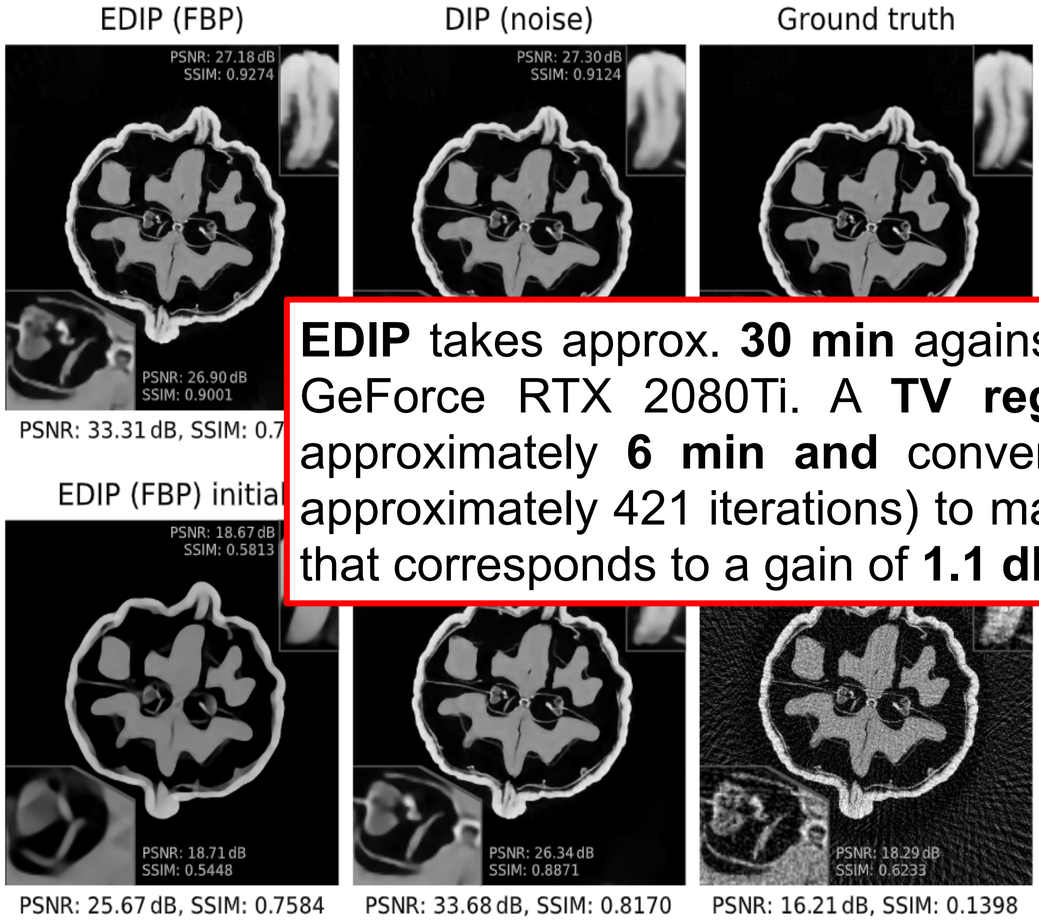
$$\boldsymbol{\theta}_{\mathrm{t}}^{\star} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}}{\operatorname{argmin}} \left\{ \mathrm{l}_{\mathrm{t}}(\boldsymbol{\theta}) := \| \mathrm{A}\, \mathbf{f}(\mathbf{z}; \boldsymbol{\theta}) - \mathbf{y}_{\delta} \|_2^2 + \lambda \, \mathrm{TV}\left(\mathbf{f}(\mathbf{z}; \boldsymbol{\theta}) \right) \right\},$$

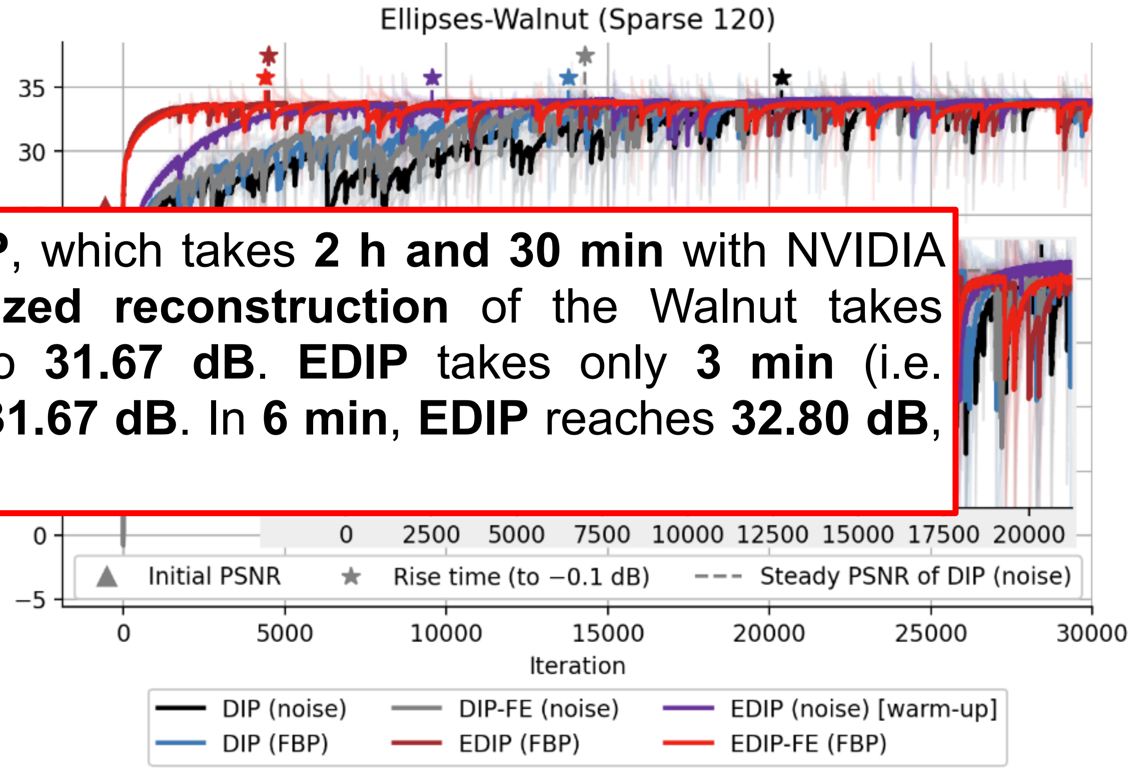
$$\mathbf{x}^{\star} = \mathbf{f}(\mathbf{z}; \boldsymbol{\theta}_{\mathrm{t}}^{\star}),$$

Can be interpreted as a MAP objective given a prior that constrains reconstructions to be the output of a U-net and have low TV

Is Deep Image Prior in Need of a Good Education?
R Barbano, J Leuschner, M Schmidt, A Denker, A Hauptmann, P Maaß, B Jin

Reconstructing the Walnut data





EDIP takes approx. **30 min** against **DIP**, which takes **2 h and 30 min** with NVIDIA GeForce RTX 2080Ti. A **TV regularized reconstruction** of the Walnut takes approximately **6 min and** converge to **31.67 dB**. **EDIP** takes only **3 min** (i.e. approximately 421 iterations) to match **31.67 dB**. In **6 min**, **EDIP** reaches **32.80 dB**, that corresponds to a gain of **1.1 dB**.

Building a probabilistic deep image prior

1. **Train** (educated) U-net with “standard objective”

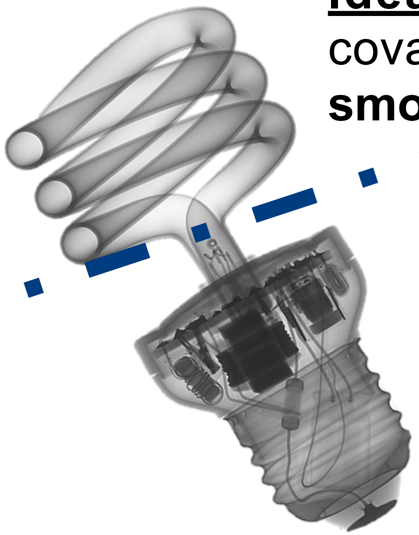
2. **Build** Bayesian hierarchical model

$$\mathbf{y}_\delta|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{y}_\delta; \mathrm{A}\mathbf{h}(\boldsymbol{\theta}), \sigma_y^2 \mathrm{I}),$$
$$\boldsymbol{\theta}|\boldsymbol{\ell} \sim \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \Sigma_{\boldsymbol{\theta}}(\boldsymbol{\ell}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}})), \quad \boldsymbol{\ell} \sim p(\boldsymbol{\ell})$$

$$p(\boldsymbol{\ell}) = \prod_{d=1}^{D} p(\ell_d) = \prod_{d=1}^{D} \mathrm{Exp}(\lambda \kappa_d) \left| \frac{\partial \kappa_d}{\partial \ell_d} \right|,$$
with $\kappa_d := \mathbb{E}_{p(\boldsymbol{\theta}_d|\ell_d;\sigma_d^2) \prod_{i=1, i\neq d}^{D} \delta(\boldsymbol{\theta}_i)} [\mathrm{TV}(\mathbf{h}(\boldsymbol{\theta}))]$

3. **Optimize** hyperparameters with **marginal likelihood**

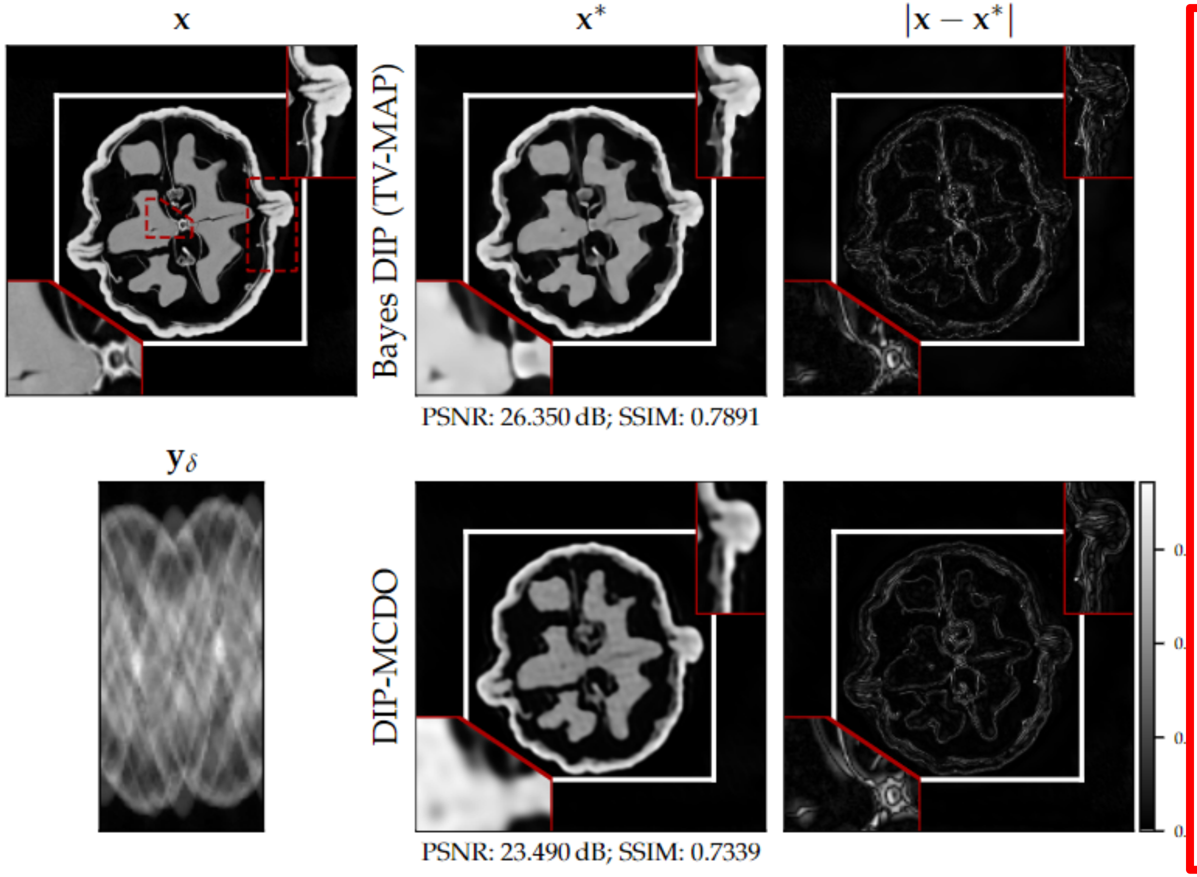
4. **Predict (UQ)**!



<u>Idea</u>: build surrogate prior with a covariance kernel that **enforces TV smoothness**

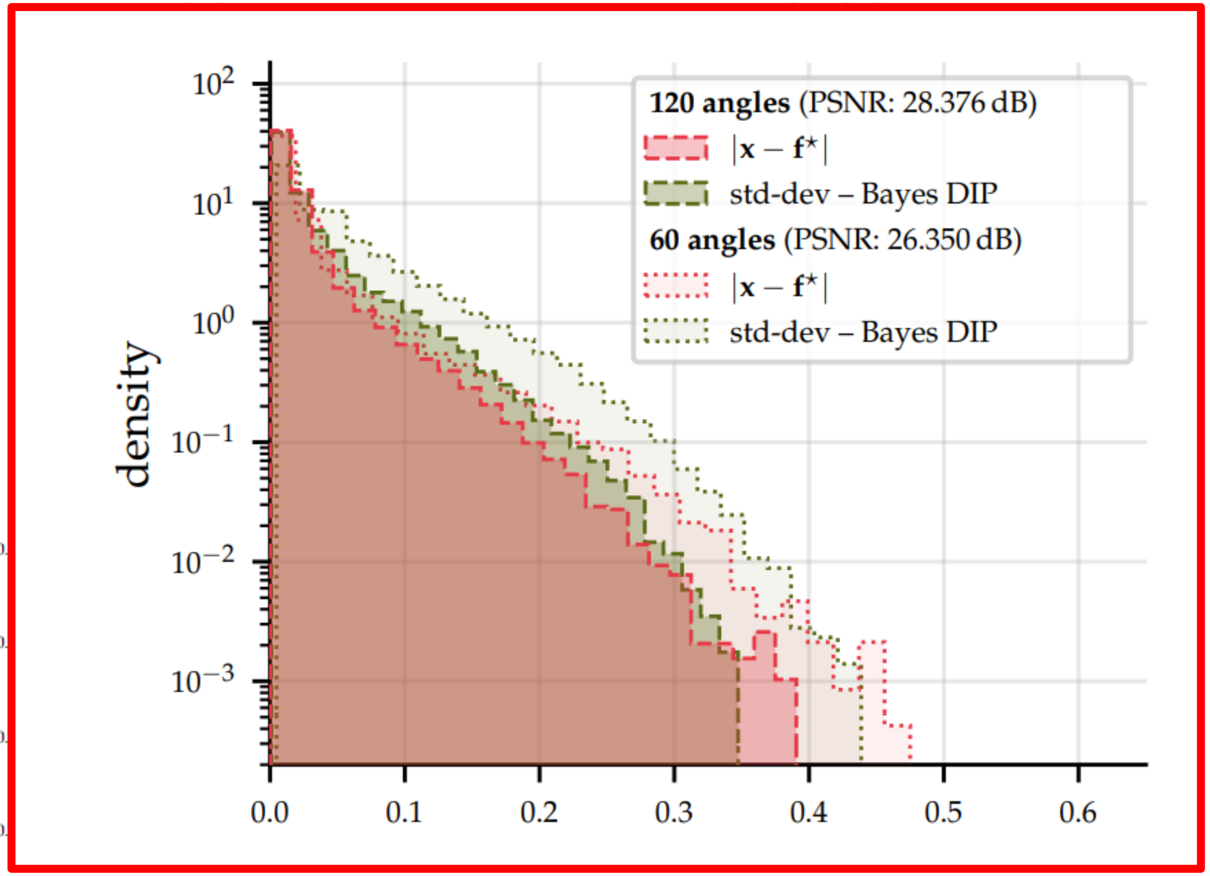
	wall-clock time
DIP optim. w/ pretraining	7m
Hyperparam. optim. with CG	7h:50m
Assemble $\Sigma_{\mathbf{yy}}$	2h:30m
Draw k posterior samples (without CG)	16m ($k = 2^{10}$)

Probabilistic DIP for high resolution CT



A Probabilistic Deep Image Prior for Computational Tomography
R Barbano, J Antoran, J Leuschner, MH Lobato, B Jin (under submission).

Our method remains well-calibrated in this setting!



we apply our method to a problem twice as large (i.e., 120×128=15360)

Summary/Contributions

1. **Designing** a tractable Bayesian prior over reconstructed images mimicking the TV

2. **Combining** such a prior with the linearized Laplace method to obtain more calibrated uncertainty estimates than existing DIP approaches

3. **Proposing** an efficient implementation of the method.

Relevant literature

[1] Artemev et al., “Tighter Bounds on the Log...”, 2021. [2] David J.C. MacKay, “Bayesian Methods for Adaptive Models...”, 1992. [3] C. E. Rasmussen & C. K. I. Williams, “Gaussian Processes for Machine Learning”, 2006. [4] Matthias W. Seeger, “Bayesian Inference and Optimal Design...”, 2008. [5] Helin et al., “Edge-promoting Adaptive Bayesian...”, 2021. [6] Immer et al., “Improving Predictions of Bayesian...”, 2021. [7] C. Bishop, “Pattern Recognition and Machine Learning”, 2006. [8] Barbano et al., “Is Deep Image Prior in Need... ”, 2021. [9] Kendal & Gal, “What Uncertainties Do We Need in Bayesian...”, 2017.

Thank you for listening

Appendix

μCT Measurement Data

- Cone-beam measurements using 3 source positions
- 1200 equidistant angles over [0, 360°)
- Reduce geometry to 2D volume slice, selecting a subset of measurement pixels
- Assemble forward operator as a sparse matrix for image resolution $(501\ \text{px})^2$ from given geometry
- Sparse-view task: reconstruct from 120 (or 60) angles (10x/ 20x subs.)
- Ground truth publicly available