Deep Image Prior for Inverse Problems: Acceleration and Probabilistic Treatment

Bangti Jin (btjin@math.cuhk.edu.hk)

Chinese University of Hong Kong

Workshop on Imaging with Uncertainty Quantification
September 27–29, 2022, Denmark

Outline

joint work with

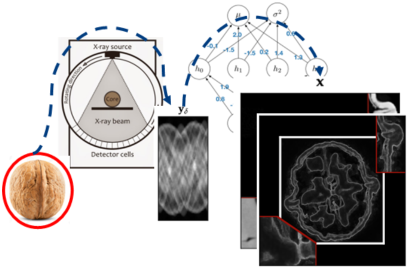
- **Riccardo Barbano**, Simon Arridge (UCL)
- Jose Miguel Hernández-Lobato, **Javier Antorán** (Cambridge)
- **Johannes Leuschner**, Maximilian Schmidt, Alexander Denker, Peter Maass (Bremen)
- Andreas Hauptmann (Oulu)

model setting: linear inverse problem

$$y^\delta = Ax + \eta$$

- x: unknown image / signal
- y^δ: noisy measurements
- η: data noise
- A: linear forward operator

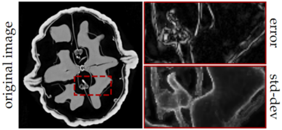
The problem is often ill-posed



Traditionally, it is estimated through **regularized reconstruction.**
Can we design a **neural solver** with quantified uncertainty?

Real μCT measurement data

- Cone-beam measurements using 3 source positions
- 1200 equidistant angles over $[0, 360^\circ)$
- 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 a given geometry
- Sparse-view task: reconstruct from 120 (or 60) angles (10x/ 20x subs.)
- Ground truth publicly available



H. Sarkissian, F. Lucka, M. van Eijnatten, G. Colacicco, S. Coban, K. J. Batenburg. (2019). Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 1-8

- there are many successful supervised approaches using deep learning for solving inverse problems.

$$\theta^* \in \arg\min \sum_i \|f_\theta(y^i) - x^i\|^2, \quad x^q = f_\theta(y^q)$$

  - unrolling (ISTA, ADMM, GD, PDHG, EM, ...)
  - learned prior / likelihood
  - ...

features

- cheap at deployment (passing through the network)
- require (abundant) paired training data
- performance may **degrade significantly** when the test data distributes differently from the training data! V Antun, F Renna, C Poon, B Adcock, A C Hansen. PNSA 2020

Deep image prior

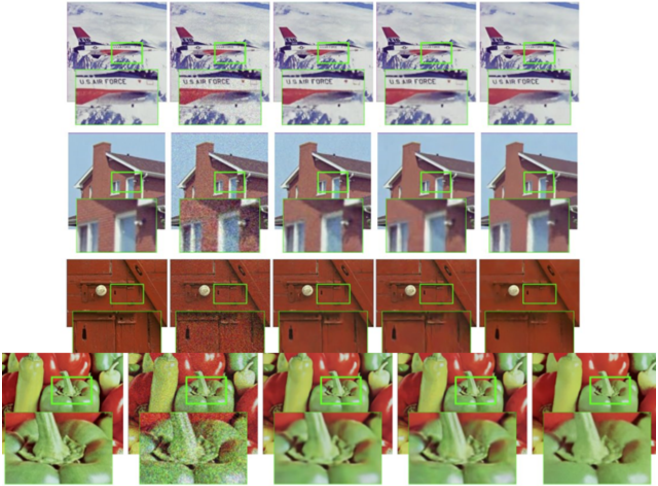
unsupervised approach:

- deep image prior: D Ulyanov, A Vedaldi, V Lempitsky CVPR 2018, 9446–9454

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d_\theta}} \|A f_\theta(z) - y^\delta\|^2, \quad x^* = f_{\theta^*}(z)$$

z: random input, θ: DNN parameters
architecture alone can regularize inverse problems in imaging ...

Image denoising D Ulyanov, A Vedaldi, V Lempitsky CVPR 2018, 9446–9454



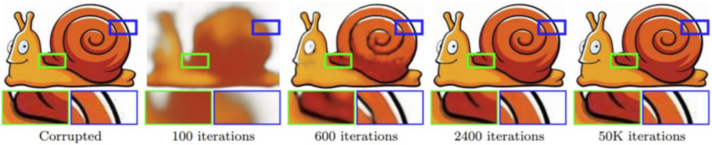
(a) GT (b) Input (c) Ours (d) CBM3D (e) NLM short history of the method

- first proposed by D Ulyanov, A Vedaldi, V Lempitsky CVPR 2018
  image denoising, super-resolution, inpainting, delurring
- deep decoder R Heckel, P Hand ICLR 2019
- theory: DIP for compressed sensing with convolution type architecture, smooth components converge faster than noise
  R Heckel, M Soltanolkotabi ICML 2020
- theory: connection with Tikhonov regularization (analytic DIP)
  S Dittmer, T Kluth, P Maass, D Otero Baguer JMIV 2020; Clemens Arndt 2022, Inverse Problems
- many applications: CT, PET, MRI, MPI, ...

pros and cons of DIP

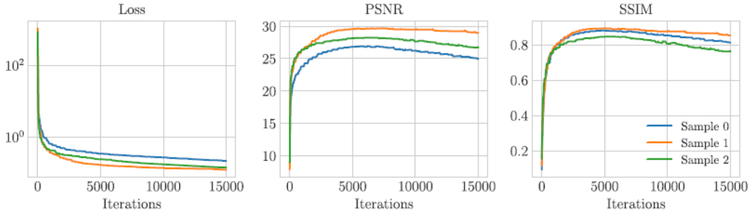
- free from any training data, robust w.r.t. distributional shift
- the need of early stopping
- fresh training for each new data $\Rightarrow$ expensive to deploy
- no uncertainty estimate

overfitting without regularization D Ulyanov, A Vedaldi, V Lempitsky CVPR 2018



message: early stopping is needed !

overfitting on CT reconstruction



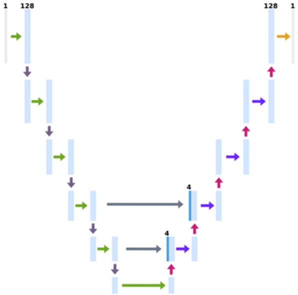
Take-home message: total variation (TV) regularization can improve image quality and relaxes the need for early stopping when using a suitable regularization parameter.

D. Otero Baguer, J. Leuschner, M. Schmidt. Inverse Problems 2020

deep image prior for inverse problems

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d_\theta}} \{\ell(\theta) := \|Af_\theta(z) - y^\delta\|^2 + \lambda |f_\theta(z)|_{\mathrm{TV}}\}, \quad x^* = f_{\theta^*}(z)$$

- **Unsupervised** learning
- ~~Ordered pairs of ground truths images and real measurement data~~?
- Robust to out-of-distribution samples?
- Suitable for imaging applications with scarce data availability??



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

Can DIP benefit from **pretraining** for accelerating subseq. reconstructive tasks? If so, can we easily construct an informative dataset to warm–start DIP? How do inductive biases of pretraining impact the reconstructive task?

R Barbano, J Leuschner, M Schmidt, A Denker, A Hauptmann, P Maass, B Jin. 2021. arXiv:2111.11926

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

- Step 1: Supervised pretraining on synthetic training data $(x^i, y_i^\delta)_{i=1}^N$

$$\theta_s^* = \arg\min_{\theta \in \mathbb{R}^{d_\theta}} \frac{1}{N} \sum_{i=1}^{N} \| f_\theta(A^\dagger y_i^\delta) - x^i \|^2$$

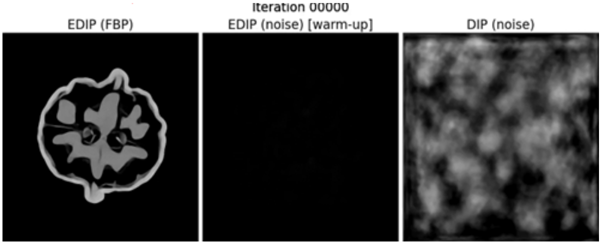
- Step 2: Unsupervised fine-tuning on real data

$$\theta_t^* \in \arg\min_{\theta \in \mathbb{R}^{d_\theta}} \|A f_\theta(z) - y^\delta\|^2 + \lambda |f_\theta(z)|_{\mathrm{TV}}$$

$$x^* = f_{\theta_t^*}(z)$$

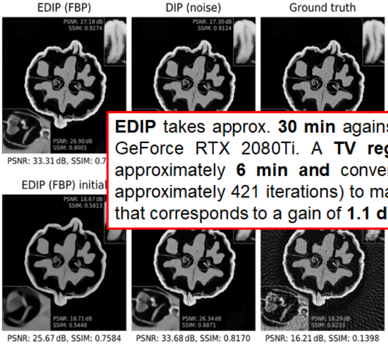
Can be interpreted as a MAP objective given a prior that constrains reconstructions to be the output of a U-net and have low TV.

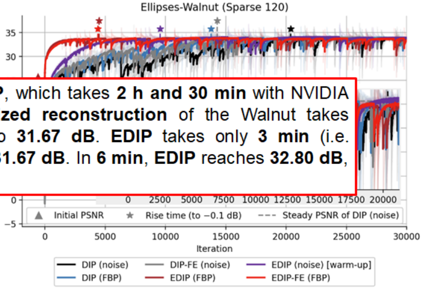
Barbano, R., Leuschner, J., Schmidt, M., Denker, A., Hauptmann, A., Maaß, P. and Jin, B., 2021. arXiv:2111.11926



https://educateddip.github.io/docs.educated_deep_image_prior/

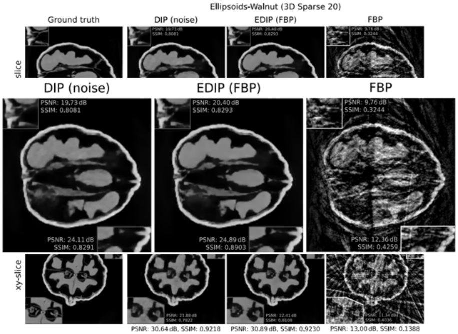
Reconstruction with walnut dataset

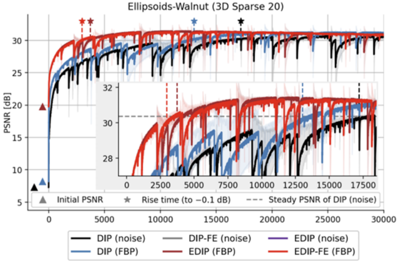




EDIP takes approx. **30 min** against **DIP**, which takes **2 h and 30 min** with NVIDIA GeForce RTX 2080Ti. A **TV regularized reconstruction** of the Walnut takes approximately **6 min and** converge to **31.67 dB**. **EDIP** takes only **3 min** (i.e. approximately 421 iterations) to match **31.67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

Reconstruction with **3D** walnut dataset





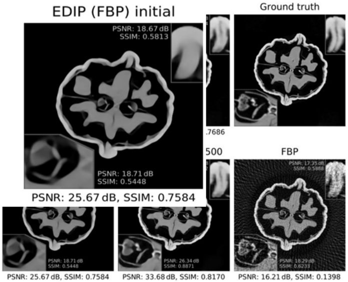
Shedding insights into pretraining

Feature reuse plays a very crucial role

- The deployment of the pretrained model on an out-of-distribution task shows high input-robustness

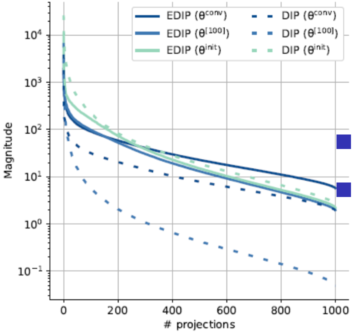
The feature reuse mechanism leads to hallucinatory behaviors

- Amending the knowledge acquired via pretraining protects from instabilities due to distributional shifts



spectral evaluation of pretraining

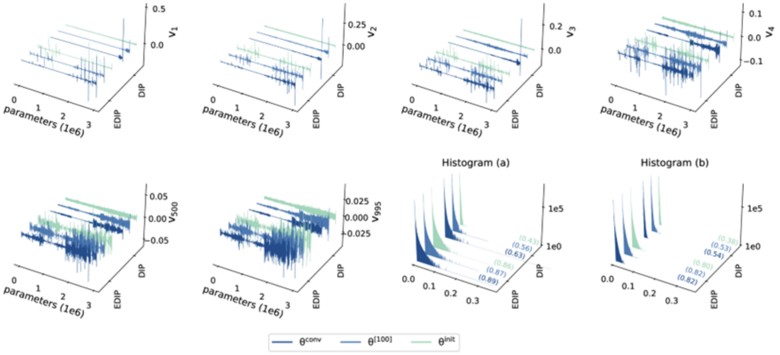
construct an approx SVD of the Jacobian via randomization



- SVs of DIP Jacobian shifts greatly
- SVs of EDIP Jacobian are stable

spectral evaluation of pretraining

construct an approx SVD of the Jacobian via randomization



Uncertainty quantification?

Question: can we put an **error bar** on the DIP reconstruction, on realistic 3D dataset ?

J Antorán, R Barbano, J Leuschner, JM Hernández-Lobato, B Jin. A probabilistic deep image prior for computational tomography (2022) arXiv:2203.00479

A Bayesian approach to quantify uncertainty

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

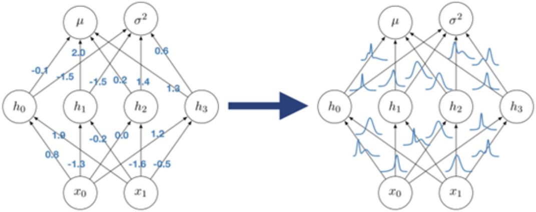
$$p(x|y^\delta) = p(y^\delta|x) \overbrace{p(x)}^{\text{prior belief}} \;/\; \underbrace{p(y^\delta)}_{\text{model evidence}}$$

ranks every x according to their agreement with y^δ and the prior belief

The normalization constant $p(y^\delta)$ provides an objective for hyperparameter selection without the need for validation data

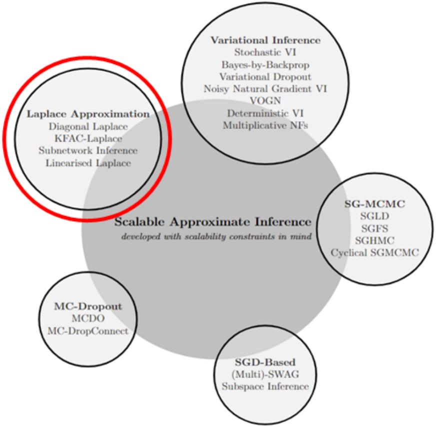
Tools for modelling uncertainty in deep learning

- Place a prior distribution over NN parameters
- Define a likelihood function to characterize the agreement of the NN function with the observations
- Update the weight distribution using Bayes' rule



intractable posterior distribution over weights

towards scalable approximation ...



Barbano,..., 2022. Uncertainty quantification in medical image synthesis. In Biomedical Image Synthesis and Simulation (pp. 601-641). Academic Press.

Laplace approximation

- train the neural network (find a mode)

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d_\theta}} \sum_{i=1}^{N} \underbrace{\ell(x^i, x^{i*}; \theta)}_{-\log p(x^i | f_\theta(A^\dagger y_i^\delta))} + \underbrace{R(\theta)}_{-\log p(\theta)}$$

- Approximate the posterior distribution over the NN parameters

McKay 1992

$$p(\theta|\mathcal{D}) \approx q(\theta) := \mathcal{N}(\theta; \theta^*, \Sigma_\theta), \quad \text{with } \Sigma_\theta = -[\nabla_\theta^2 J(\mathcal{D}; \theta)|_{\theta=\theta^*}]^{-1}$$

- Further ... approximate (with input $A^\dagger y^\delta$)

$$q_{\mathrm{ggn}}(\theta) := \mathcal{N}(\theta; \theta^*, (J^\top \Lambda(\theta) J + S_0^{-1})^{-1}|_{\theta=\theta^*}$$

A. Immer, M. Korzepa, M. Bauer. ICML 2021; E. Daxberger, E. Nalisnick, J. Allingham, J. Antorán J. M. Hernández-Lobato ICML 2021

linearized Laplace method

- A Gaussian can be a very poor approximation to the NN's posterior distribution ... yet experimentally it is a very good posterior for a linear model (in some cases exact!)
- linearized the underlying BNN

$$h(\theta) = f_{\theta^*}(A^\dagger y^\delta) + \underbrace{J_{\theta^*}(A^\dagger y^\delta)}_{\text{basis expansion!}} (\theta - \theta^*)$$

$$x \sim \mathcal{N}(h(0), J\Sigma_\theta(\ell, \sigma_\theta^2)J^\top)$$

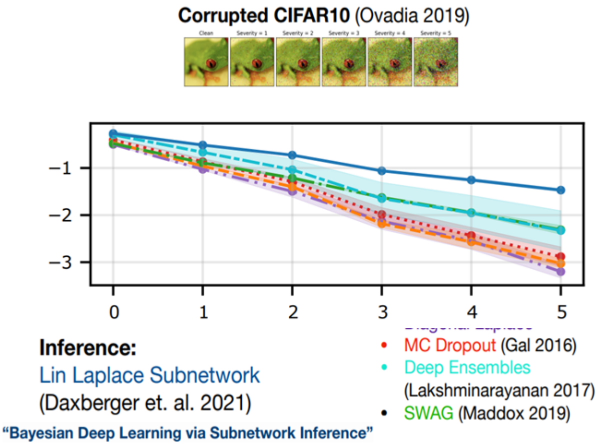
linearization induces a **GP deep image prior**

- perform approx. inference in the GP model or solve it in closed form for regression $\Rightarrow$ conjugate Gaussian-linear model has:

Closed form **predictive posterior** and **marginal likelihood**

The revival of linearized neural networks

- David Mackay (1992, Bayesian methods for adaptive models) first introduced NN linearisation as an approximation to yield a closed form error-bars for Laplace approximated posteriors
- Neil D. Lawrence (2000, Variational inference in probabilistic models) finds the Laplace approximation to underperform without the linearisation step. This was also noted by Ritter (2018, Online structured Laplace approximations for overcoming catastrophic forgetting) in their efforts to scale-up the Laplace approximation using Kronecker factorisation
- Khan (2019, Approximate Inference Turns Deep Networks into Gaussian Processes) and Alexander Immer (2021, Improving predictions of Bayesian neural nets via local linearization) re-popularised the linearisation step by showing that it improves the quality of uncertainty estimates



... retains better uncertainty calibration and robustness to distribution shift E. Daxberger, E. Nalisnick, J. Allingham, J. Antorán J. M. Hernández-Lobato ICML 2021

Deep image prior + linearized Laplace $\Rightarrow$

Probabilistic inference with linearized deep image prior for CT image reconstruction?

- TV is not smooth enough ! $\Rightarrow$ linearized Laplace does not apply
- tractable hierarchical construction via predictive complexity prior

...

Method: Compute a Gaussian-linear model type error-bars using a local linearization of the DIP around its optimal reconstruction hierarchical modeling: Designing a tractable Bayesian prior over reconstructed images mimicking TV

- Bayesian construction

$$y^\delta|\theta \sim \mathcal{N}(y^\delta; Ah(\theta), \sigma^2 I),$$
$$\theta|\ell \sim \mathcal{N}(\theta; 0, \Sigma_\theta(\ell, \sigma_\theta^2)), \quad \ell \sim p(\ell)$$

- Matern kernel

$$[\Sigma_\theta(\ell, \sigma_\theta^2)]_{kij,k'i'j'} = \sigma_d^2 e^{-\frac{-d(i-i',j-j')}{\ell_d}} \delta_{kk'}$$

with $k \sim$ convolutional filters, $i, j \sim$ spatial location within the filter

- connecting TV via predictive complexity prior

$$p(\ell) = \prod_d p(\ell_d) = \prod_d \mathrm{Exp}(\lambda\kappa_d)\left|\frac{\partial \kappa_d}{\partial \ell_d}\right|$$
$$\kappa_d := \mathbb{E}_{p(\theta_d|\ell_d,\sigma_d^2)\prod_{i\neq d}\delta_{\theta_i}}[|h(\theta)|_{\mathrm{TV}}]$$

optimizing $(\ell, \sigma_y^2, \sigma_d^2)$ via marginal likelihood (Type-II evidence)

$$
\begin{aligned}
&\log p(\mathbf{y}_\delta | \boldsymbol{\ell}; \sigma_y^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) + \log p(\boldsymbol{\ell}; \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) = \\
&-\frac{1}{2}\sigma_y^{-2} ||\mathbf{y}_\delta - \mathrm{Af}(\boldsymbol{\theta}^\star)||_2^2 - \frac{1}{2}||\boldsymbol{\theta}^\star||^2_{\Sigma_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)} \\
&-\frac{1}{2}\log |\Sigma_{\mathbf{y}_\delta \mathbf{y}_\delta}(\sigma_y^2, \boldsymbol{\ell}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)| - \lambda \sum_{d=1}^{D} \kappa_d + \log \left| \frac{\partial \kappa_d}{\partial \ell_d} \right| .
\end{aligned}
$$

many computational tricks

- randomized estimators of determinants / gradients
- PCG with half precision
- low rank + diagonal preconditioner
- ...

public implementation available via software package (soon !)

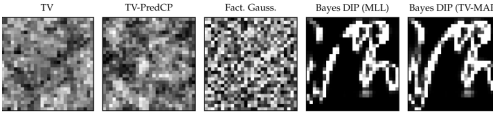
Five steps from regularized reconstruction to Bayesian inference

- Train (educated) U-net with “standard objective”
- Linearize around some acceptable parameter setting
- Build Bayesian hierarchical model: Build surrogate prior with a kernel that enforces TV smoothness
- Optimize hyperparameters with marginal likelihood
- Make prediction (UQ) through sampling

Outcome: computational pipeline providing **pixel-wise uncertainty** estimates and a **marginal likelihood** objective

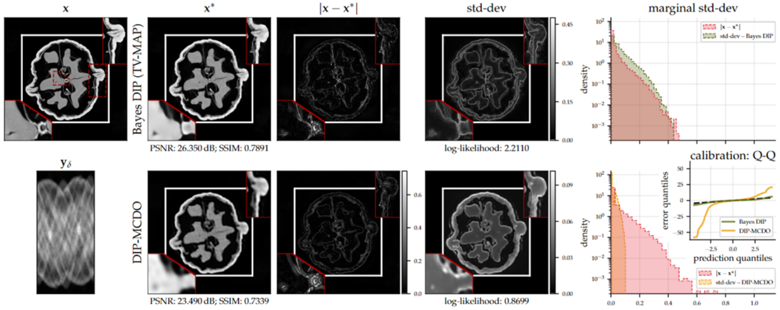
NOTE: Estimate the uncertainty associated with a specific reconstruction, instead of trying to characterise a full posterior distribution over reconstructions.

samples from different priors:

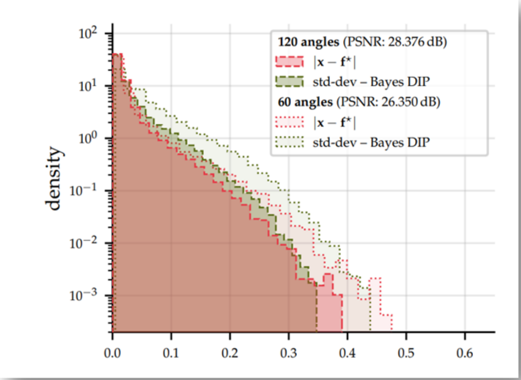


hierarchical construction: very powerful task-specific kernels

probabilistic DIP for high-resolution CT reconstruction



probabilistic DIP yields **more calibrated uncertainty estimates** than existing DIP approaches (MC dropout) apply the method to a problem twice as large



Summary: Take-home message

- Pretraining is effective for accelerating DIP
- Probabilistic DIP provides calibrated uncertainty estimates for DIP reconstructions
- Provide an efficient implementation of the method (as a library)
- the approach can be extended for Bayesian experimental design
  R Barbano, J Leuschner, J Antorán, B Jin, JM Hernández-Lobato arXiv:2207.05714
- other imaging modalities ...

THANK YOU FOR YOUR ATTENTION!

(very incomplete) references

- **D Ulyanov, A Vedaldi, V Lempitsky. Deep image prior. CVPR 2018, pp. 9446-9454**
- R Heckel, M Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. ICML 2020
- D Otero Baguer, J Leuschner, M Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. Inverse Problems 2020; 36(9): 094004
- R Barbano, J Leuschner, M Schmidt, A Denker, A Hauptmann, P Maass. Is deep image prior in need of a good education? (2021) arXiv:2111.11926
- J Antorán, R Barbano, J Leuschner, JM Hernández-Lobato, B Jin. A probabilistic deep image prior for computational tomography. (2022) arXiv:2203.00479
- R Barbano, J Leuschner, J Antorán, B Jin, JM Hernández-Lobato. Bayesian experimental design for computed tomography with the linearised deep image prior. (2022) arXiv:2207.05714
- C Arndt. Regularization theory of the analytic deep prior approach. Inverse Problems 2022; 38(11) : 115005
- ....