Image Reconstruction via Deep Image Prior Subspaces

Riccardo Barbano *riccardo.barbano.19@ucl.ac.uk*
Department of Computer Science
University College London

Javier Antorán *ja666@cam.ac.uk*
Department of Engineering
University of Cambridge

Johannes Leuschner *jleuschn@uni-bremen.de*
University of Bremen

José Miguel Hernández-Lobato *jmh233@cam.ac.uk*
Department of Engineering
University of Cambridge

Bangti Jin *b.jin@cuhk.edu.hk*
The Chinese University of Hong Kong

Željko Kereta *z.kereta@ucl.ac.uk*
Department of Computer Science
University College London

Reviewed on OpenReview: *https://openreview.net/forum?id=torWsEui9N*

Abstract

Deep learning has been widely used for solving image reconstruction tasks but its deployability has been held back due to the shortage of high-quality paired training data. Unsupervised learning methods, e.g., deep image prior (DIP), naturally fill this gap, but bring a host of new issues: the susceptibility to overfitting due to a lack of robust early stopping strategies and unstable convergence. We present a novel approach to tackle these issues by restricting DIP optimisation to a sparse linear subspace of its parameters, employing a synergy of dimensionality reduction techniques and second order optimisation methods. The low-dimensionality of the subspace reduces DIP's tendency to fit noise and allows the use of stable second order optimisation methods, e.g., natural gradient descent or L-BFGS. Experiments across both image restoration and tomographic tasks of different geometry and ill-posedness show that second order optimisation within a low-dimensional subspace is favourable in terms of optimisation stability to reconstruction fidelity trade-off.

1 Introduction

Deep learning (DL) approaches have shown impressive results in a wide variety of linear inverse problems in imaging, e.g., denoising (Tian et al., 2020), super-resolution (Ledig et al., 2017; Ulyanov et al., 2020), magnetic resonance imaging (Zeng et al., 2021) and tomographic reconstruction (Wang et al., 2020). Mathematically, a linear inverse problem of recovering an unknown image $x \in \mathbb{R}^{d_x}$ from measurements $y \in \mathbb{R}^{d_y}$ is formulated as

$$y = Ax + \epsilon, \tag{1}$$

for an (ill-conditioned) matrix $A \in \mathbb{R}^{d_y \times d_x}$ and exogenous noise ϵ.

However, conventional supervised DL approaches are not ideally suited for practical inverse problems. Large quantities of clean paired data, typically needed for training, are not available in many problem domains, e.g., tomographic reconstruction. Moreover, the ill-posedness (due to the forward operator A and noise ϵ) and high-dimensionality of the images x pose significant challenges, and can be computationally very demanding. Whereas standard imaging tasks, e.g., denoising and deblurring, use high dimensional observations ($d_x \approx d_y$), tomographic imaging often requires reconstructing an image from observations that are of a much lower dimensionality compared to the sought-after images. For example, reconstructing a CT of a walnut may require reconstructing from observations that are only 3% of the size of the original image.

Unsupervised DL approaches do not require paired training data, and have received significant attention in the imaging community. Among these approaches, DIP (Ulyanov et al., 2018) has garnered the most traction. DIP parametrises the reconstructed image as the output of a convolutional neural network (CNN) with a fixed input. The reconstruction process learns low-frequency components before high-frequency ones (Chakrabarty & Maji, 2019; Shi et al., 2022), which can act as a form of regularisation.

Alas, the practicality of DIP is hindered by two key issues. Firstly, each DIP reconstruction requires training the entire network anew. This can take from several minutes up to a couple of hours for high-resolution images (Barbano et al., 2022b). Secondly, DIP requires careful early stopping (Wang et al., 2021) to avoid overfitting, which is often based on case-by-case heuristics. However, validation-based stopping criteria are often not viable in the unsupervised setting as they violate i.i.d. assumptions, see Wang et al. (2021).

This paper aims to address both of the existing issues inherent to DIP, and to illustrate the approach on image restoration and reconstruction. Building upon recent body of evidence which shows that neural network (NN) training often takes place in low-dimensional subspaces (Li et al., 2018; Frankle & Carbin, 2019; Li et al., 2023), we restrict DIP's optimisation to a sparse linear subspace of its parameters. This has a two-fold beneficial effect. First, subspace optimisation trades off some flexibility in capturing fine image structure details for robustness to overfitting. This is extraordinarily well suited in imaging problems belonging to inherently lower dimensional structures, but it is also shown to be competitive in restoring natural images. Moreover, it allows using stopping criteria based on the training loss, without sacrificing performance. Second, the low dimensionality induced by the subspace allows using second order optimisation methods (Amari, 2013; Martens & Grosse, 2015a). This greatly stabilises the reconstruction process, facilitating the use of a simple loss-based stopping criterion, and reduces the number of iterations to convergence.

Our contributions can be summarised as:

- We extract a principal subspace from DIP's parameter trajectory during a synthetic pre-training stage. To reduce the memory footprint of working with a non-axis-aligned subspace, we sparsify the extracted basis vectors using top-k leverage scoring.
- We use second order methods: natural gradient descent (NGD) and limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS), to optimise DIP's parameters in a low-dimensional subspace.
- We provide thorough experimental results across several linear inverse problems, comprising image restoration and tomographic tasks of different geometry and degree of ill-posedness, showing that subspace methods are favourable in terms of optimisation stability to reconstruction fidelity trade-off.

2 Related Work

The present work lies at the intersection of overfitting in unsupervised methods, subspace learning and second order optimisation. Below we discuss recent advances in the related fields.

Avoiding overfitting in DIP: Since the introduction of DIP (Ulyanov et al., 2018; 2020), stopping optimisation before overfitting to noise has been a necessity. The analysis in Chakrabarty & Maji (2019) and Shi et al. (2022) elucidates that U-Net is biased towards learning low-frequency components before high-frequency ones. The authors suggest a sharpness-based stopping criterion, which however requires a modified architecture. Jo et al. (2021) propose a criterion based on the Stein's unbiased risk estimate (Eldar, 2008) for denoising, which however performs poorly for ill-posed settings (Metzler et al., 2018). Wang et al. (2021) propose a running image-variance estimate as a proxy for the reconstruction error. Our experiments find this method somewhat unreliable for sparse CT reconstruction. Ding et al. (2022) and Yaman et al. (2021) propose to split the observation vector into training and validation sub-vectors, and use the loss functional on the latter as a stopping criteria. Unfortunately, this violates the i.i.d. data assumption that underpins validation-based early stopping (Shalev-Shwartz & Ben-David, 2014, Theorem 11.2). Independently, Liu et al. (2019) and Baguer et al. (2020) add a TV regulariser to the DIP objective (2). This only partially alleviates the need for early stopping and has seen widespread adoption. To the best of our knowledge, the present work is the first to successfully avoid overfitting without significant performance degradation.

Linear subspace estimation: Literature on low-rank matrix approximation is rich, with randomised SVD approaches being the most common (Halko et al., 2011; Martinsson & Tropp, 2020). However, in high-dimensions, even working with a small set of dense basis vectors can itself be prohibitively expensive. Matrix sketching methods (Drineas et al., 2012; Liberty, 2013) alleviate this through axis-aligned subspaces. To the best of our knowledge, our work is the first to combine these two method classes, producing non-axis-aligned but sparse approximations.

Optimising neural networks in subspaces: Closely related to the present work is that of Li et al. (2018) and Wortsman et al. (2021), who find that networks can be trained in low-dimensional subspaces of the parameters without loss of performance, and that more complex tasks need larger subspaces. Similarly to our methodology, Li et al. (2023) identify subspaces from training trajectories and observe the resulting robustness to label noise. Results reported in Frankle & Carbin (2019) show that pruning a very large number of parameters in fully-connected and convolutional feed-forward networks yields trainable sub-networks that can achieve performance comparable to that of the original network. This principle has yielded speedups in network evaluation (Wen et al., 2016; Daxberger et al., 2021). Shwartz-Ziv et al. (2022) use a low-rank estimate of the curvature around an optimum of a pre-training task to regularise subsequent supervised learning.

Second order optimisation for neural networks: Despite their adoption in traditional optimisation (Liu & Nocedal, 1989), second order methods are rarely used with neural networks due to the high cost of dealing with curvature matrices for high-dimensional functions. Martens & Sutskever (2012) use truncated conjugate-gradient to approximately solve against a network's Hessian. However, a limitation of the Hessian is that it is not guaranteed to be positive semi-definite (PSD). This is one motivation for NGD (Foresee & Hagan, 1997; Amari, 2013; Martens, 2020), that uses the FIM (guaranteed PSD). Commonly, the KFAC approximation (Martens & Grosse, 2015a) is used to reduce the costs of FIM storage and inversion. Also, common deep-learning optimisers, e.g., Adam (Kingma & Ba, 2015) or RMSprop (Hinton et al., 2014) may be interpreted as computing online diagonal approximations to the Hessian.

3 Deep Image Prior

DIP expresses the reconstructed image $x = f(x_0, \theta)$ in terms of the parameters $\theta \in \mathbb{R}^{d_\theta}$ of a CNN $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_x}$, and fixed input $x_0 \in \mathbb{R}^{d_x}$. The parameters θ are learnt by minimising the loss

$$\mathcal{L}(\theta) = \|Af(x_0, \theta) - y\|_2^2 + \lambda \mathrm{TV}(f(x_0, \theta)), \tag{2}$$

composed of a data fidelity and total variation (TV), weighed by $\lambda > 0$. TV is the most popular regulariser for image reconstruction (Rudin et al., 1992; Chambolle et al., 2010). Its anisotropic version is given by

$$\mathrm{TV}(x) = \sum_{i,j} |X_{i,j} - X_{i+1,j}| + \sum_{i,j} |X_{i,j} - X_{i,j+1}|, \tag{3}$$

where $X \in \mathbb{R}^{h \times w}$ is a vector $x \in \mathbb{R}^{d_x}$ reshaped into an $h \times w$ image, and $d_x = h \cdot w$. In this work, f is a fully convolutional U-Net, see C.3 for more details. This implicitly regularises the reconstruction by preventing overfitting to noise, as long as the optimisation is stopped early enough.

DIP optimisation costs can be reduced by pre-training on synthetic data. E-DIP (Barbano et al., 2022b) first generates samples from a training data distribution P of random ellipses, and then applies the forward model A and adds white noise, following (1). The network input is set as the filtered back-projection (FBP) $x_0 = x^\dagger := A^\dagger y$, where $A^\dagger$ denotes the (approximate) pseudo-inverse of A. The pre-training loss is given by

$$\mathcal{L}_{\mathrm{pre}}(\theta) = \mathbb{E}_{x,y\sim P}\|f(x^\dagger, \theta) - x\|_2^2, \tag{4}$$

where P denotes the (empirical) joint distribution between the ground truth x and the corresponding noisy data. The pre-trained network can then be fine-tuned on any new observation y by optimising the objective (2) with FBP as the input. E-DIP decreases the DIP's training time, but can increase susceptibility to overfitting, making early stopping even more critical, cf. discussion in Section 5.

4 Methodology

We now describe our procedure for optimising DIP parameters in a subspace. We first describe how the E-DIP pre-training trajectory is used to extract a sparse basis for a low-dimensional subspace of the parameters. The objective is then reparametrised in terms of sparse basis coefficients. Finally, we describe how L-BFGS and NGD are used to update the sparsified subspace coefficients.

Step 1 — Identifying the sparse subspace: First, we find a subspace of parameters that is low-dimensional and *easy to work with*, but contains a rich enough set of parameters to fit to the observation y. We leverage E-DIP pre-training trajectory to acquire basis vectors by stacking d_{pre} parameter vectors, sampled at uniformly spaced checkpoints on the pre-training trajectory, into a matrix $\Theta^{\mathrm{pre}} \in \mathbb{R}^{d_\theta \times d_{\mathrm{pre}}}$. We then compute top-$d_{\mathrm{sub}}$ SVD of $\Theta^{\mathrm{pre}} \approx USV^\top$, and keep the left singular vectors $U \in \mathbb{R}^{d_\theta \times d_{\mathrm{sub}}}$, where $d_{\mathrm{sub}} \leq d_{\mathrm{pre}}$ is the dimensionality of the chosen subspace. We then sparsify the orthonormal basis U by computing leverage scores (Drineas et al., 2012), associated with each DIP parameter as

$$\ell_i = \sum_{k=1}^{d_{\mathrm{sub}}} [U]_{ik}^2, \quad i = 1, \ldots, d_\theta.$$

We keep only the basis vector entries corresponding to $d_{\mathrm{lev}} < d_\theta$ largest leverage scores. This can be achieved by applying a diagonal mask $M \in \{0,1\}^{d_\theta \times d_\theta}$ satisfying $[M]_{ii} = \mathbb{1}(i \in \arg\,\text{top-}d_{\mathrm{lev}}\,\ell)$, where $\ell = [\ell_1, \ell_2, \ldots, \ell_{d_\theta}]$. The sparse basis MU contains at most $d_{\mathrm{lev}} \cdot d_{\mathrm{sub}}$ non-zero entries.

Pre-training and sparse subspace selection are only performed once, and can be amortised across different reconstructions. We choose $d_{\mathrm{pre}} \approx 10^3$, resulting in a large memory footprint of the matrix Θ^{pre}, though this is stored in cpu memory. Alternatively, incremental SVD algorithms (Brand, 2002) can be used to further reduce the memory requirements (see Appendix A). Training DIP in the sparse subspace requires storing only the sparse basis vectors MU in accelerator (gpu / tpu) memory. Thus, sparsification allows training in relatively large subspaces $d_{\mathrm{sub}} > 10^3$ of large networks $d_\theta > 10^7$.

Step 2 — Network reparametrisation: We write the objective $\mathcal{L}(\theta)$ in terms of sparse basis coefficients $c \in \mathbb{R}^{d_{\mathrm{sub}}}$ as

$$\mathcal{L}_\gamma(c) := \mathcal{L}(\gamma(c)), \quad \text{with } \gamma(c) := \theta^{\mathrm{pre}} + MUc. \tag{5}$$

This restricts the DIP parameters θ^{pre} (from pre-training) to change only along the sparse subspace MU. The coefficient vector c is initialised as a sample from a uniform distribution on the unit sphere.

Step 3 — Second order optimisation: The subspace reparametrised objective $\mathcal{L}_\gamma$ in (5) differs from traditional DL loss functions in that the method-dependent local curvature matrix of $\mathcal{L}_\gamma$ can be computed and stored accurately and efficiently, without resorting to restrictive approximation of its structure, such as the often used diagonal or KFAC approximations (Martens & Grosse, 2015b). These facts open the door to second order optimisation of $\mathcal{L}_\gamma$, which may converge faster than first order methods. However, repeatedly evaluating second order derivatives of neural networks has a prohibitive cost, further compounded by the rapid change of the local curvature of the non-quadratic loss when traversing the parameter space. We mitigate this by performing online low-rank updates of a curvature estimate while only accessing loss Jacobians. In particular, we use L-BFGS (Liu & Nocedal, 1989) and stochastic NGD (Amari, 1998; Martens, 2020) in the experimental evaluation. The former estimates second directional derivatives by solving the secant equation. The latter keeps an exponentially moving average of stochastic estimates of the Fisher information matrix (FIM).

NGD for DIP in a subspace: The exact FIM is given as

$$\mathbb{E}_{v\sim\mathcal{N}(Af(x^{\dagger},\gamma(c)),\,I_{d_y})}[\nabla_c\mathcal{L}_\gamma(c)^\top\nabla_c\mathcal{L}_\gamma(c)], \tag{6}$$

where $\nabla_c\mathcal{L}_\gamma(c) = (Af(x^{\dagger},\gamma(c)) - v)^\top AJ_fMU \in \mathbb{R}^{1\times d_{\mathrm{sub}}}$ is the Jacobian of the subspace loss $\mathcal{L}_\gamma$ at the current coefficients (with $I_{d_y} \in \mathbb{R}^{d_y\times d_y}$ being the identity matrix), and $J_f \coloneqq \nabla_\theta f(x^{\dagger},\theta)|_{\theta=\gamma(c)} \in \mathbb{R}^{d_x\times d_\theta}$ is the neural network Jacobian at $\theta = \gamma(c)$. Note that TV's contribution is omitted since the FIM is defined only for terms that depend on the observations and not for the regulariser (Ly et al., 2017). At step t we estimate the FIM by the Monte-Carlo method as

$$\hat{F}_t = \frac{1}{n}\sum_{i=1}^{n}(z_i^\top AJ_fMU)^\top z_i^\top AJ_fMU, \quad \text{with } z_i \sim \mathcal{N}(0, I_{d_y}). \tag{7}$$

The moving average of the FIM is updated as

$$F_{t+1} = \beta F_t + (1-\beta)\hat{F}_t \quad \text{with } \beta \in (0,1). \tag{8}$$

Our implementation of NGD follows the approach of Martens & Grosse (2015a), with adaptive damping, and step size and momentum parameters chosen by locally minimising a quadratic model. See Appendix B for additional discussion.

Note that he methodology presented in this section can be extended beyond the DIP framework, in the spirit of test-time fine-tuning strategies of learned reconstruction methods (Darestani et al., 2022). Here we offer an alternative for adapting reconstruction algorithms at test-time, by introducing a robust optimisation framework.

5 Experiments

Our experiments cover a wide range of image restoration and tomographic reconstruction tasks. In Section 5.1, we conduct an ablation study on CartoonSet (Royer et al., 2017), examining the impact of subspace dimension d_{sub}, subspace sparsity level d_{lev}, degree of problem ill-posedness, and choice of the optimiser on reconstruction speed and fidelity. The acquired insights are then applied to challenging tomography tasks and image restoration. We compare the reconstruction fidelity, propensity to overfitting, and convergence stability relative to vanilla DIP and E-DIP on a real-measured high-resolution CT scan of a walnut, in Section 5.2, and on medically realistic high-resolution abdomen scans from the Mayo Clinic, in Section 5.3.

We simulate observations using (1) with dynamically scaled Gaussian noise given by

$$\epsilon \sim \mathcal{N}(0,\,\sigma^2 I_{d_y}) \quad \text{with } \sigma = p/d_y\sum_{i=1}^{d_y}|y_i|, \tag{9}$$

with the noise scaling parameter set to $p = 0.05$, unless noted otherwise. We conclude with denoising and deblurring on a standard RGB natural image dataset (Set5) in Section 5.4. For studies in Sections 5.2, 5.3 and 5.4, we use a standard fully convolutional U-Net architecture with either $\sim$ 3M (for CT reconstruction tasks) or $\sim$ 1.9M (for natural images) parameters, see architectures in Appendix C.3. For the ablative analysis in Section 5.1, we use a shallower architecture ($\sim$ .5M) with only 64 channels and four scales, keeping the skip connections in lower layers. Following the literature, we train the vanilla DIP (Ulyanov et al., 2018) (labelled DIP) and E-DIP (Barbano et al., 2022b) with ADAM. We train subspace coefficients with Adam (Sub-DIP

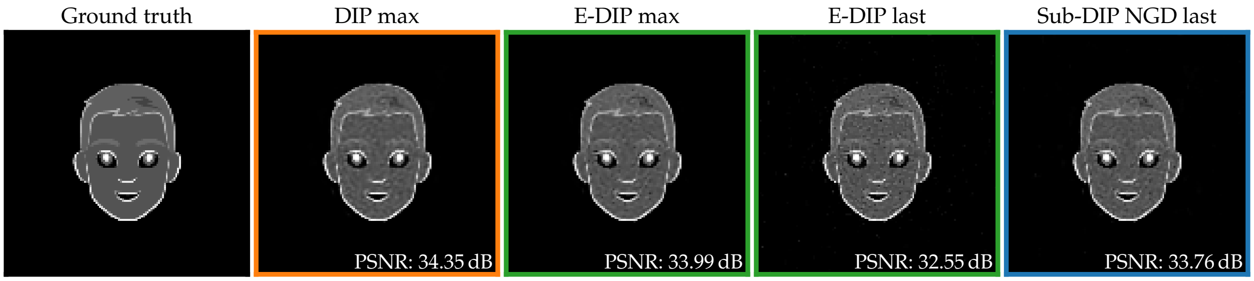

Figure 1: Reconstruction comparison for an example CartoonSet image from 45 angles. “max” indicates oracle (highest possible) PSNR, which we note is only available if a ground truth image exists. “last” denotes the final reconstruction.

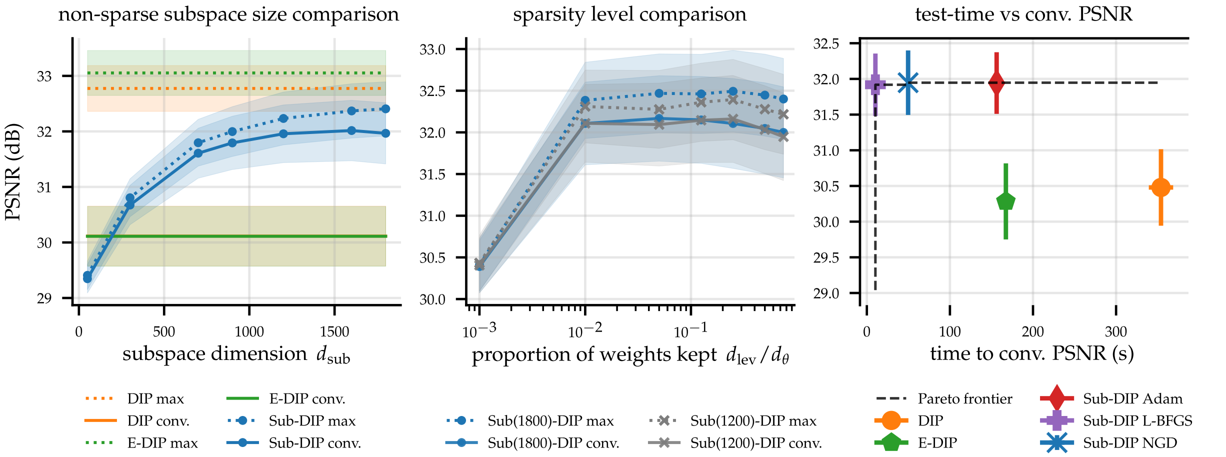

Figure 2: The influence of subspace dimension d_{sub} (left), sparsity level ${}^{d_{\mathrm{lev}}}/_{d_{\theta}}$ (middle) on PSNR, and the PSNR vs time Pareto-frontier (right) for CartoonSet. The dashed line defines a Pareto-frontier. “max” refers to the oracle stopping PSNR, while “conv.” refers to the PSNR at the stopping point found by applying Algorithm 1, cf. Appendix A, to (2), with δ=0.995 and patience of $\mathfrak{p}$=100. Left and middle plots use NGD. Results show mean and standard deviation over 25 reconstructed images from 95 angles. Runs are performed on A100 GPU. Note that the PSNR line for DIP conv., in the leftmost panel, is almost completely obscured by E-DIP conv. PSNR, since they have an almost identical PSNR value.

Adam), which serves as a baseline, L-BFGS (Sub-DIP L-BFGS) and NGD (Sub-DIP NGD). Image quality is assessed through peak signal-to-noise ratio (PSNR).

For tomographic tasks we use the same pre-training runs for E-DIP and all subspace methods: minimising (4) over $32k$ images of ellipses with random shape, location and intensity. Pre-training inputs are constructed from an FBP obtained with the same tomographic projection geometry as the dataset under consideration. Analogously, for image restoration tasks, we pre-train on ImageNet (Deng et al., 2009).

The method is built on top of the E-DIP library (`github.com/educating-dip`). The full implementation and data are available at `github.com/subspace-dip`.

5.1 Ablative analysis on CartoonSet (Royer et al., 2017)

We investigate Sub-DIP’s sensitivity to subspace dimension d_{sub}, sparsity level d_{lev}, and ill-posedness on 25 images of size $(128\,\mathrm{px})^2$ from CartoonSet (Royer et al., 2017, `google.github.io/cartoonset`). Example reconstructions are shown in Fig. 1. We simulate a parallel-beam geometry with 183 detectors and 45, 95 or 285 angles, corresponding to, respectively, a very sparse-view (d_y=8235), a moderately sparse-view (d_y=17385), and a fully sampled setting (d_y=52155). We construct subspaces by sampling d_{pre}=$2k$ parameters at uniformly spaced checkpoints during 100 pre-training epochs on ellipses as in Barbano et al. (2022b). We measure the degree of overfitting by comparing the highest PSNR obtained throughout optimisation (max) with that given by Algorithm 1, applied to the training loss (2) with δ=0.995 and $\mathfrak{p}$=100 steps (conv.).

In Appendix A we report results of additional ablation experiments on the extracted subspaces, examining their adaptability to changes in the forward operator, and comparing them to random subspace bases.

Subspace dimension: Fig. 2 (left) explores the trade-off between subspace dimension d_{sub} and reconstruction quality. We use 95 angles, subspace methods with no sparsity (d_{lev}=d_θ) and the NGD optimiser. Both standard DIP and E-DIP overfit, showing a $\approx$ 3 dB gap between max and conv. PSNR values, while subspace methods exhibit only a $\approx$ 0.5 dB gap. Both max and conv. PSNR present a monotonically increasing trend with subspace dimension, while the gap at $d_{\text{sub}} > 1k$ stays roughly constant at $\sim$0.25 dB. Thus, these subspace dimensions are too small for significant overfitting to occur. In spite of this, d_{sub}=100 is enough to obtain better conv. PSNR than DIP and E-DIP.

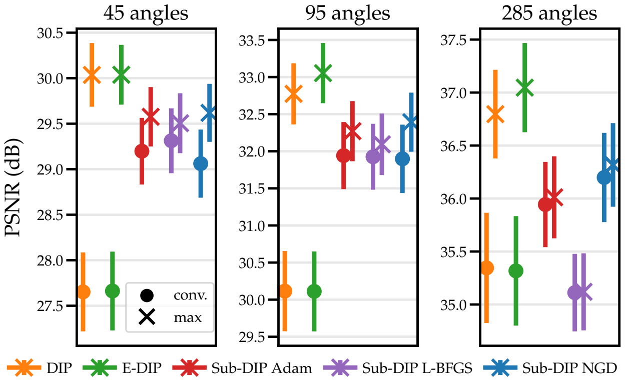

Figure 3: PSNR mean and standard deviation over 50 CartoonSet images from 45, 95, and 285 angles.

Subspace sparsity level: Fig. 2 (middle) shows that for ${}^{d_{\text{lev}}}/_{d_\theta} > 0.01$, the reconstruction fidelity is largely insensitive to the sparsity level. This is true for both d_{sub}=1200 and d_{sub}=1800. This effect is also somewhat independent of the subspace dimensionality d_{sub}. Hence, *sparse subspaces should be constructed by choosing a large* d_{sub} *and then sparsifying its basis vectors to ensure computational tractability.*

Problem ill-posedness: We report the performance with varying numbers of view angles in Fig. 3, with non-sparse subspaces of dimension d_{sub}=1800. The smaller is the number of view angles, the more ill-posed the reconstruction problem becomes. Subspace methods present a smaller gap between max and conv. PSNR ($\sim$ 0.25 dB) than DIP and E-DIP ($\sim$ 2.5 dB). Subspace methods also present better fidelity at convergence for all the studied number of view angles. Notably, the improvement is larger for more ill-posed settings ($\sim$ 2.5 dB at 45 and 95 vs $\approx$ 1 dB at 285 angles), despite the worsening performances for all methods in sparser settings. This is expected as there is a reduced risk of overfitting in data-rich regimes and more flexible models can do better. E-DIP's max reconstruction fidelity is consistently above that of other methods by at least 0.5 dB. This may be attributed to full-parameter flexibility with benign inductive biases from pre-training. However, obtaining max PSNR performance requires oracle stopping and is not achievable in practice. In Fig. 4, we show PSNR trajectories of the studied methods (DIP, E-DIP, Sub-DIP Adam, Sub-DIP LBFGS, and Sub-DIP NGD) for 45, 95, and 285 angles, averaged over 50 reconstructed images. This figure provides additional evidence supporting previous observations. While DIP and E-DIP tend to overfit to noise once they reach their maximum PSNR values, subspace methods consistently maintain stable reconstruction performance with no noticeable performance degradation, due to the inherent regularising effect obtained by restricting to a subspace.

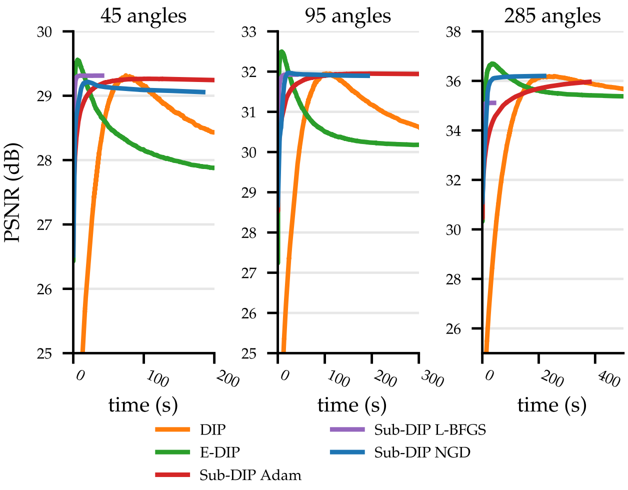

Figure 4: Average PSNR trajectories for 45, 95, and 285 angles over 50 reconstructions.

First vs second order optimisation: We compare optimisers' conv. PSNR vs their time to convergence. Fig. 2 (right) shows that Sub-DIP L-BFGS and NGD converge in less than 50 seconds. These methods are optimal along the known Pareto-frontier, reaching $\sim$ 1.5 dB higher reconstruction fidelity than DIP and E-DIP. LBFGS converges in only $\sim$20 seconds. Sub-DIP Adam retains protection against overfitting but converges at a rate similar to non-subspace first order methods (in $\sim$180 seconds). These trends hold across studied degrees of ill-posedness; see Appendix A.1.

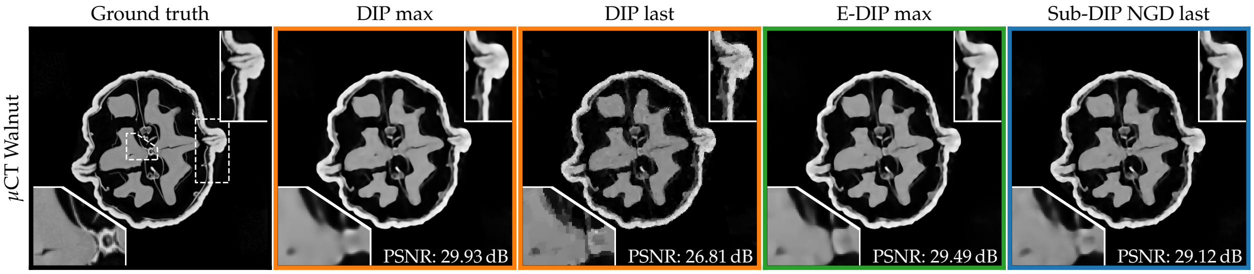

Figure 5: Reconstructions of the Walnut using 60 angles. Sub-DIP reconstructions do not capture any noise, but present slightly increased ringing around very thin structures. "max" indicates oracle (highest possible) PSNR. "last" denotes the final reconstruction.

Algorithm 1 Early stop criterion

```
Inputs: metric $g(\cdot)$, patience $\mathfrak{p}$, decrease proportion $\delta$
$g_{\min} \leftarrow \infty$, $i \leftarrow 0$, $i_{\min} \leftarrow \infty$
while $i \leq i_{\min} + \mathfrak{p}$ do
    if $g(i) < \delta \cdot g_{\min}$ then
        $g_{\min} \leftarrow g(i)$ and $i_{\min} \leftarrow i$
    end
    $i \leftarrow i + 1$
end
Output: $i_{\min}$
```

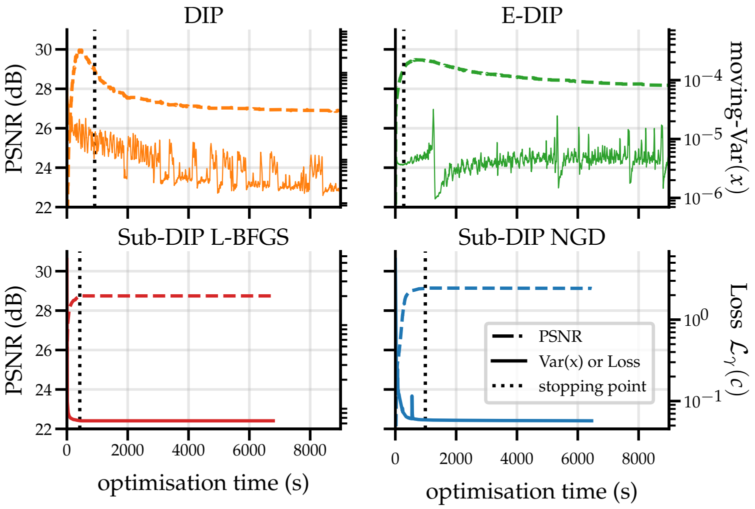

Figure 6: The evolution of PSNR and stopping metrics (variance-based for DIP and E-DIP, and loss-based for Sub-DIP NGD and L-BFGS) vs optimisation time on Walnut with one seed.

5.2 µCT Walnut (Der Sarkissian et al., 2019)

In this section we study the methods on a real-measured high-resolution µCT problem. We reconstruct a $(501\,\mathrm{px})^2$ slice from a very sparse, real-measured cone-beam scan of a walnut, using 60 angles and 128 detector pixels ($d_y = 7680$). We compare reconstructions against ground-truth (Der Sarkissian et al., 2019), which uses classical reconstruction from full data acquired at 3 source positions with 1200 angles and 768 detectors. This task involves fitting both broad-stroke and fine-grained image details (see Fig. 5), making it a good proxy for µCT of industrial context. We pre-train the 3M parameter U-Net for 20 epochs and save d_{pre}=$5k$ parameter checkpoints. Following Section 5.1, we construct a d_{sub}=$4k$ dimensional subspace and sparsify it down to ${}^{d_{\text{lev}}}/{}_{d_\theta}$=0.5 of the parameters.

Fig. 7 (left) shows the optimisation curves for all optimisers, averaged across 3 seeds. Qualitatively, the results are similar to the cartoon data. Second order subspace methods converge to their highest reconstruction fidelity within the first 500 seconds and do not overfit. Vanilla DIP and E-DIP also converge quickly but suffer from overfitting leading to ∼3 dB and ∼ 1.8 dB of performance degradation, respectively. Sub-DIP Adam takes over 3000 seconds to converge and does not overfit.

Stopping criterion challenges: Capturing the oracle performance of DIP and E-DIP would require a robust stopping criterion. Note that we cannot base a stopping criterion on (2). We instead turn to the method from Wang et al. (2021), which minimises a rolling estimate of the reconstruction variance across optimisation steps, a proxy for the squared error. Following Wang et al. (2021), we compute variances with a 100 step window and apply Algorithm 1 with a patience of $\mathfrak{p}$=1000 steps and δ=1. Figure 6 shows this metric to be very noisy when applied to DIP and E-DIP. This breaks the smoothness assumption implicit in the stopping criterion (Algorithm 1), leading to stopping more than 1 dB before/after reaching the max PSNR. This phenomenon is attributed to the severe ill-posedness of the tasks causing variance across a large subspace of reconstructions that fit our observations well. For tomographic reconstruction, the variance curve becomes more non-convex, and its minimum does not correspond to the optimal PSNR. Since subspace methods do not overfit and the loss is smooth, we can use it as our stopping metric (δ=0.995, $\mathfrak{p}$ = 100).

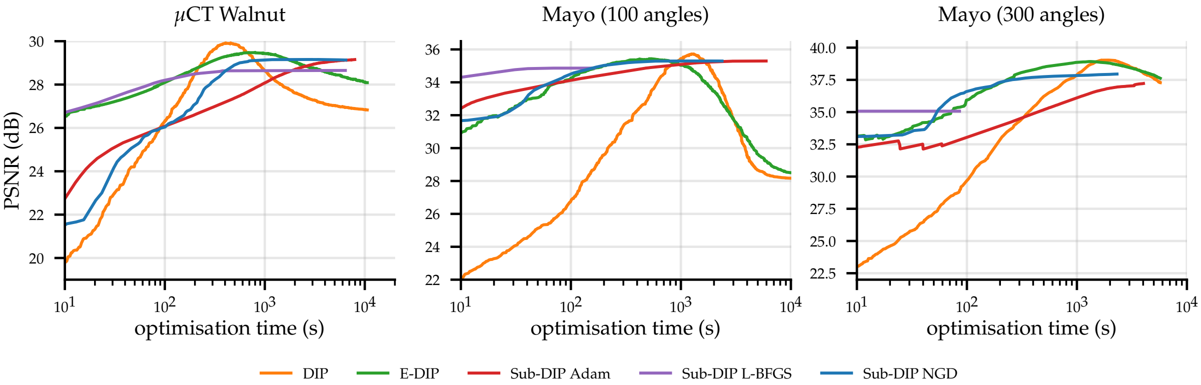

Figure 7: Optimisation curves for the μCT Walnut data (left), and Mayo using 100 angles (middle) and Mayo using 300 angles (right) averaged over 10 images.

5.3 Mayo Clinic dataset (Moen et al., 2021)

To investigate a medical setting, we use 10 clinical CT images of the human abdomen released by Mayo Clinic (Moen et al., 2021) and study the behaviour of subspace optimisation. We reconstruct from a simulated fan-beam projection, with 300 angles and white noise with the noise scaling parameter $p = 0.025$, cf. (9). As a more ill-posed reference setting, we use 100 angles (comparable sparsity to the Walnut setting) and Gaussian noise with p=0.05. For both tasks, we use the 3M parameter U-Net, as in Section 5.2, pre-trained on $32k$ ellipse samples. For the sparse setting (100 angles), we use a $d_{\text{sub}} = 4k$ dimensional subspace constructed from d_{pre}=$5k$ checkpoints, but with sparsity ratio ${}^{d_{\text{lev}}}/_{d_\theta}$=0.25. For the more data-rich setting (300 angles), we use $d_{\text{sub}} = 8k$, sampled from d_{pre}=$10k$ checkpoints, and similarly, we sparsify it down to ${}^{d_{\text{lev}}}/_{d_\theta}$=0.25.

In the 300 angle setting, Sub-DIP NGD reaches 37 dB within 200 seconds. PSNR the continutes to increas, albeit slowly, without performance saturation, cf. Fig. 7 (right). In contrast, for the sparser Walnut and Mayo data, Sub-DIP NGD maintains a steep PSNR increase until reaching max PSNR. Interestingly, L-BFGS does not perform well in the 300 angle setting, obtaining < 36 dB PSNR. This might be due to L-BFGS's tendency to stop the iterations too early in high dimensions.

Note that the observed time efficiency of the Sub-DIP reported in Fig. 2, is not preserved in high-dimensional CT tasks; see Fig. 7. This behaviour is expected as the efficiency of second order optimisation methods is affected both by the high-dimensionality of the measurements—resulting in costlier forward operator evaluations—and more importantly, by the depth of the U-Net—resulting in costlier vector-Jacobian and Jacobian-vector products. As discussed, second order methods converge in fewer iterations (see e.g., Fig. 19 in Appendix A) but require a higher per-iteration cost.

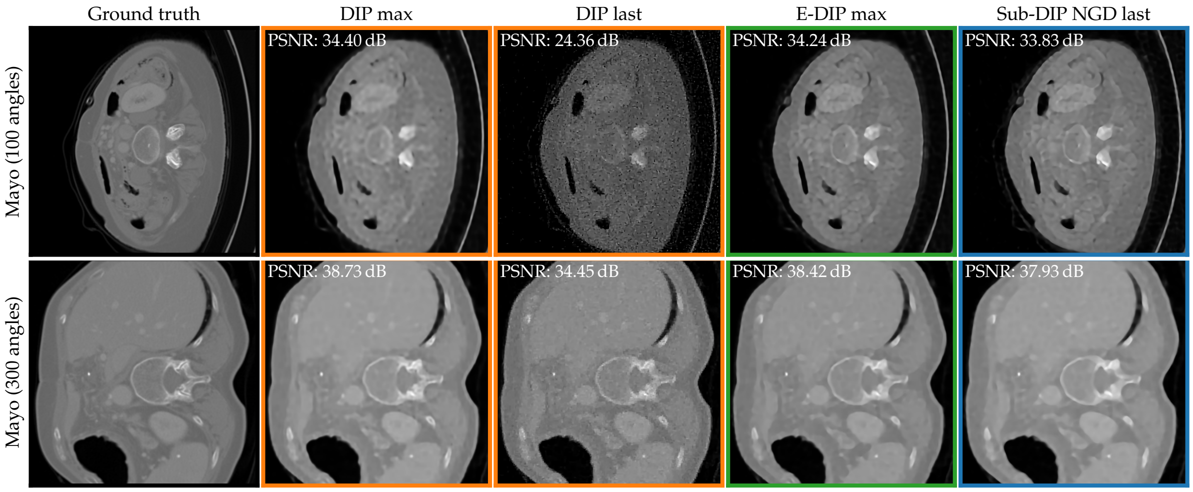

Figure 8: Example reconstructions on the Mayo dataset in 100 (top) and 300 (bottom) angle settings.

However, it is worth noting that when working with a limited runtime budget, Sub-DIP (L-BFGS and NGD) methods may take more time to reach their maximum PSNR, but they consistently deliver the highest image quality within a 100-second runtime, as illustrated in Fig. 7. At the same time, Sub-DIP methods do not suffer performance degradation due to overfitting to noise. A comparison in terms of the number of iterations, instead of optimisation time, can be found in the rightmost panel in Figs. 19, 20 and 21 in the Appendix.

Fig. 8 shows example reconstructions using 100 and 300 angles. If stopped within a narrow max PSNR window, DIP and E-DIP can deliver reconstructions that better capture high-frequency details than Sub-DIP methods as expected. However, while DIP and E-DIP reconstructions become noisy once they start to overfit, Sub-DIP methods do not exhibit any noise. We deem the *increased robustness vs reduced flexibility* trade-off provided by the Sub-DIP to be favourable, even in the well-determined setting.

In Appendix A.3 we investigate the effect of the dataset used to extract the subspace, in the 300 angle case. Namely, we compare the task-agnostic dataset of random ellipses with a task-specific Mayo dataset. As expected, the results show that using a task-specific dataset, matching the target image manifold, improves reconstruction performance. However, the generality of the approach is compromised, since a suitable dataset of images specific enough for a task and modality at hand might not always be available and the method starts to resemble supervised learning. In the data-poor settings, using a synthetic dataset provides a good performance compromise.

5.4 Image restoration of Set5

We conduct denoising and deblurring on five widely used RGB natural images (“baboon”, “jet F16”, “house”, “Lena”, “peppers”) of size $(256\,\mathrm{px})^2$. The pre-training is done on ImageNet (Deng et al., 2009), a dataset of natural images, which we use to extract the basis for the subspace methods.

For denoising we consider four noise settings with the noise scaling parameter $p \in \{0.10, 0.15, 0.25, 0.5\}$, cf. (9), We extract a single subspace for all the noise levels. To this end, during the pre-training stage, we construct a dataset by adding noise to each training datum, with a randomly selected noise scaling parameter $p \sim \mathrm{Uni}(0.05, 0.5)$. Then a $d_{\mathrm{sub}} = 8k$ subspace with sparsity level $d_{\mathrm{lev}}/d_{\theta} = 0.25$ is extracted from $d_{\mathrm{pre}} = 10k$ samples. In the high noise case ($p = 0.5$), we sub-extract to a smaller $d_{\mathrm{sub}} = 1k$ subspace. This sub-extraction is necessary to minimise overfitting in high noise level scenarios. Further comments on this are at the bottom of the page. For deblurring we consider two settings, using a Gaussian kernel with std of $\kappa \in \{0.8, 1.6\}$ pixels and $p = 0.05$ Gaussian noise. We then follow an analogous procedure; adding $p = 0.05$ Gaussian noise and applying $\kappa \sim \mathrm{Uni}(0.4, 2)$ blur to each training datum, and then constructing a single subspace.

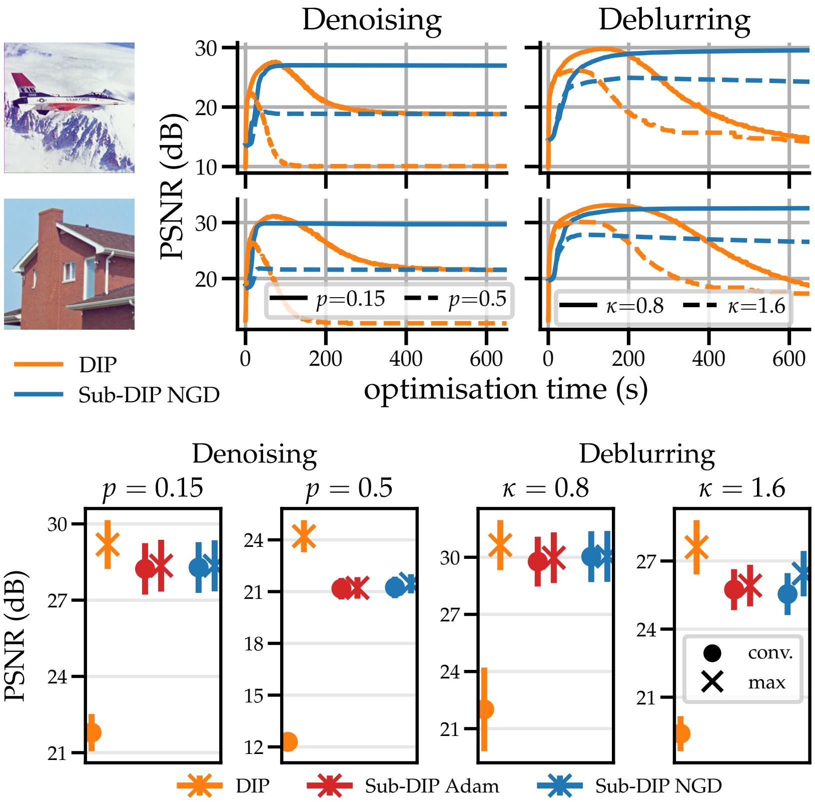

Figure 9: Denoising ($p = \{0.15, 0.5\}$) and deblurring ($\kappa = \{0.8, 1.6\}$) on the "jet F16" and "house" images. On the top we show PSNR trajectories for each of the tasks, comparing DIP and Sub-DIP NGD, and on the bottom we show mean and std of max and conv. PSNR over the studied 5 images. Results for the other two noise levels are reported in Appendix A.

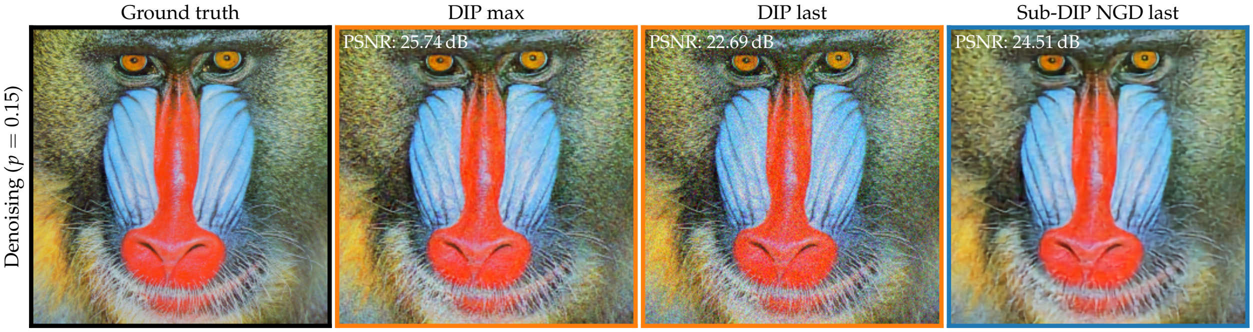

Figure 10: Restoration of the noisy (with noise scaling parameter $p = 0.15$) "baboon" image.

As common practice when deploying DIP on restoration tasks, we do not include the TV ($\lambda = 0$) in (2) for neither denoising nor deblurring. Instead, *the regularising property of the reconstruction stems exclusively from restricting the DIP optimisation to a low-dimensional subspace of its parameters.*

The top row in Fig. 9 shows PSNR trajectories for denoising and deblurring on "jet F16" and "house" images. These images contain large regions of continuous colour intensity with clearly defined edges. Hence, we expect subspace methods to perform well. This is confirmed by the results: Sub-DIP NGD and DIP have a comparable max PSNR. The bottom row in Fig. 9 shows the mean and std of max and conv. PSNR over 5 studied natural images for low and high noise and blur. The conv. reconstructions are computed by Algorithm 1, applied to the training loss (2) with δ=0.995 and $\mathfrak{p}$=100 steps. We notice that DIP mildly

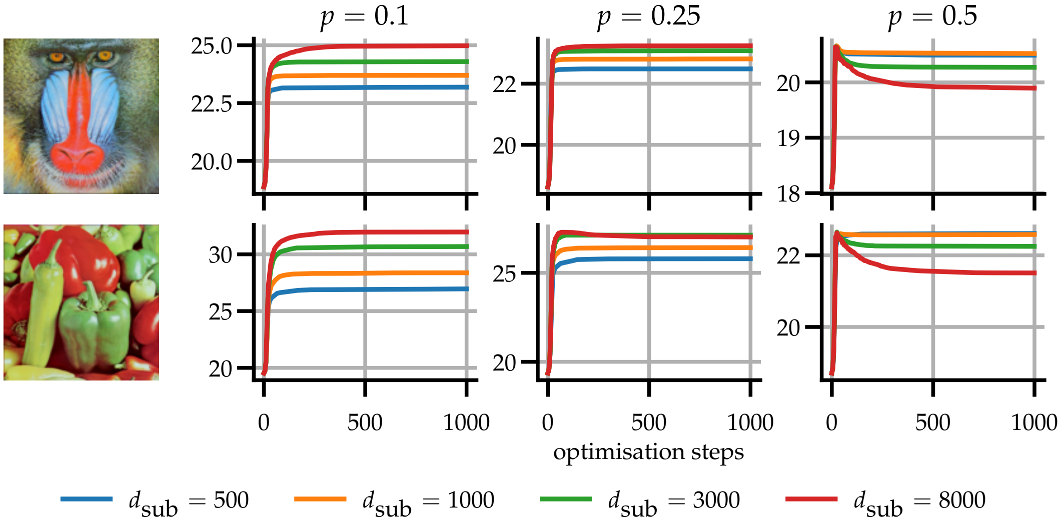

Figure 11: Investigation of the regularising effect of the dimensionality of the chosen subspace on Set5. We report PSNR trajectories for two specific images, the "baboon" and the "peppers" using Sub-DIP NGD. Our analysis encompasses three distinct noise levels and four dimensions of the subspace.

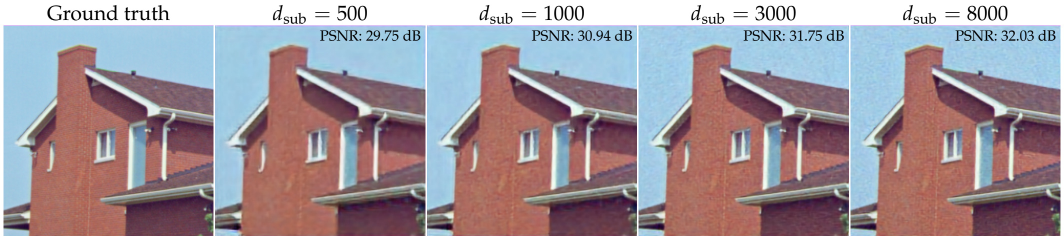

Figure 12: Restoration of the noisy (with noise scaling parameter $p = 0.1$) "house" image in Set5 using different subspace dimensions.

outperforms Sub-DIP in terms of time to max PSNR. However, DIP is also significantly more prone to overfit to noise and exhibits a sharp decline in image quality after reaching max PSNR. Thus, selecting a good stopping criterion is not only critical but also challenging.

Fig. 10 shows the denoising results for the "baboon" image with a moderate noise level ($p = 0.15$). The "baboon" image contains fine-grained details and sharp transitions throughout the image. Hence, we expect a somewhat comparatively worse performance of subspace methods on this task, as higher frequency information can be lost when using low-dimensional approaches. The results confirm this intuition: the max PSNR for the "baboon" with E-DIP is at least 1 dB higher than that for subspace methods. Moreover, while standard DIP experiences a sharp decline after achieving max PSNR, subspace methods retain the performance.

To gain deeper insight into the interplay between the dimension of the subspace and the noise level corrupting the data, we sweep over $d_{\text{sub}} \in \{500, 1000, 3000, 8000\}$ and $p \in \{0.1, 0.15, 0.25, 0.5\}$. While we maintain the ratio ${}^{d_{\text{lev}}}/_{d_\theta}$ at 1 for subspaces up to $3k$, for computational feasibility we reduce it to 0.25 for the $8k$-dimensional case. As shown in Fig. 11, and in Fig. 23 and Fig. 22 in Appendix A.4, in high-noise scenarios (i.e., $p = 0.5$), using a larger subspace leads to overfitting. Conversely, in low-noise settings, a lower-dimensional subspace limits the network's capacity to fit higher frequencies and adapt, resulting in overly smooth reconstructions, cf. Fig. 12. We can thus conclude that *noisier problems require more regularisation, thus a lower-dimensional subspace may be beneficial.*

6 Conclusion

In this work, we develop a novel approach that constrains DIP optimisation to a sparse principal low-dimensional subspace, extracted from pre-training trajectories. This greatly mitigates, if not completely eliminates, overfitting. Our approach may be understood from the perspective of the bias-variance trade-off. At initialisation, the vanilla DIP presents a useful bias towards learning low-frequency image components; but the over-parameterisation of the neural network leads to overfitting. Pre-training only partially removes DIP's low pass bias, allowing the E-DIP to often fit images quickly. This comes at the cost of increased variance. Our approach seeks to efficiently navigate the bias-variance trade-off. Constraining the optimisation to a low-dimensional subspace greatly reduces the variance. By extracting the subspace from principal directions of pre-training trajectories, and through the use of leverage scoring, we limit the bias introduced into the model. Furthermore, optimising in lower dimensional subspaces allows using fast and stable second order optimisation methods. In future work, alternative approaches for the identification of the subspace and their implicit regularising properties will be investigated.

In our experiments on several image restoration and tomographic tasks, subspace DIP methods deliver reconstructions on par with DIP's max performance, and result in better reconstruction quality than the overfit DIP reconstructions.

Acknowledgements

R.B. acknowledges the support from the i4health PhD studentship (UK EPSRC EP/S021930/1), and from The Alan Turing Institute (UK EPSRC EP/N510129/1). J.A. acknowledges the support from Microsoft Research, through its PhD Scholarship Programme, and from the EPSRC. J.L. was funded by the German Research Foundation (DFG; GRK 2224/1) and by the Federal Ministry of Education and Research via the DELETO project (BMBF, project number 05M20LBB). J.M.H.-L. acknowledges support from a Turing AI Fellowship EP/V023756/1 and an EPSRC Prosperity Partnership EP/T005386/1. The work of B.J. is partially supported by UK EPSRC grant EP/V026259/1, Hong Kong RGC General Research Fund (Project 14306423), and a start-up fund from The Chinese University of Hong Kong. Z.K. acknowledges the support from the UK EPSRC grant EP/X010740/1. This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (http://www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/T022159/1.

References

Jonas Adler, Holger Kohr, and Ozan Öktem. Operator discretization library (ODL). *Software available from https://github.com/odlgroup/odl*, 2017.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Shun-ichi Amari. Information geometry and its applications: Survey. In Frank Nielsen and Frédéric Barbaresco (eds.), *Geometric Science of Information - First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings*, volume 8085 of *Lecture Notes in Computer Science*, pp. 3. Springer, 2013. doi: 10.1007/978-3-642-40020-9_1. URL https://doi.org/10.1007/978-3-642-40020-9_1.

Javier Antorán. Understanding uncertainty in bayesian neural networks. Mphil in Machine Learning and Machine Intelligence Thesis, University of Cambridge, 2019.

Javier Antorán, Riccardo Barbano, Johannes Leuschner, José Miguel Hernández-Lobato, and Bangti Jin. A probabilistic deep image prior for computational tomography. Preprint, arXiv:2203.00479, 2022.

Javier Antorán, David Janz, James Urquhart Allingham, Erik A. Daxberger, Riccardo Barbano, Eric T. Nalisnick, and José Miguel Hernández-Lobato. Adapting the linearised Laplace model evidence for modern deep learning. *CoRR*, abs/2206.08900, 2022. doi: 10.48550/arXiv.2206.08900. URL https://doi.org/10.48550/arXiv.2206.08900.

Javoe; Antorán and Antonio Miguel. Disentangling and learning robust representations with natural clustering. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 694–699, 2019.

Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9):094004, 2020.

Riccardo Barbano, Johannes Leuschner, Javier Antorán, Bangti Jin, and José Miguel Hernández-Lobato. Bayesian experimental design for computed tomography with the linearised deep image prior. *CoRR*, abs/2207.05714, 2022a. doi: 10.48550/arXiv.2207.05714. URL `https://doi.org/10.48550/arXiv.2207.05714`.

Riccardo Barbano, Johannes Leuschner, Maximilian Schmidt, Alexander Denker, Andreas Hauptmann, Peter Maaß, and Bangti Jin. An educated warm start for deep image prior-based micro CT reconstruction. *IEEE Transactions on Computational Imaging*, 8:1210–1222, 2022b. doi: 10.1109/TCI.2022.3233188.

Matthew Brand. Incremental singular value decomposition of uncertain data with missing values. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*, pp. 707–720. Springer, 2002.

Prithvijit Chakrabarty and Subhransu Maji. The spectral bias of the deep image prior. *CoRR*, abs/1912.08905, 2019. URL `http://arxiv.org/abs/1912.08905`.

Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. In *Theoretical foundations and numerical methods for sparse recovery*, pp. 263–340. de Gruyter, 2010.

Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *International Conference on Machine Learning*, pp. 4754–4776. PMLR, 2022.

Erik A. Daxberger, Eric T. Nalisnick, James Urquhart Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2510–2521. PMLR, 2021. URL `http://proceedings.mlr.press/v139/daxberger21a.html`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Henri Der Sarkissian, Felix Lucka, Maureen van Eijnatten, Giulia Colacicco, Sophia Bethany Coban, and K. Joost Batenburg. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 1-8, 2019. URL `https://doi.org/10.5281/zenodo.2686726`. *Zenodo.*

Lijun Ding, Zhen Qin, Liwei Jiang, Jinxin Zhou, and Zhihui Zhu. A validation approach to over-parameterized matrix and image recovery. *CoRR*, abs/2209.10675, 2022. doi: 10.48550/arXiv.2209.10675. URL `https://doi.org/10.48550/arXiv.2209.10675`.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13:3475–3506, 2012. doi: 10.5555/2503308.2503352. URL `https://dl.acm.org/doi/10.5555/2503308.2503352`.

Yonina Eldar. Generalized sure for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2008.

F. Dan Foresee and Martin T. Hagan. Gauss-Newton approximation to Bayesian learning. In *Proceedings of International Conference on Neural Networks (ICNN'97), Houston, TX, USA, June 9-12, 1997*, pp. 1930–1935. IEEE, 1997. doi: 10.1109/ICNN.1997.614194. URL `https://doi.org/10.1109/ICNN.1997.614194`.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011. doi: 10.1137/090771806. URL `https://doi.org/10.1137/090771806`.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning course slides. available at `https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`, 2014.

Yeonsik Jo, Se Young Chun, and Jonghyun Choi. Rethinking deep image prior for denoising. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5067–5076. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00504. URL `https://doi.org/10.1109/ICCV48922.2021.00504`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015. Available at arxiv:1412.6980., 2015. URL `http://arxiv.org/abs/1412.6980`.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4158–4169, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/46a558d97954d0692411c861cf78ef79-Abstract.html`.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 105–114. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.19. URL `https://doi.org/10.1109/CVPR.2017.19`.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1):1–12, 2021.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=ryup8-WCW`.

Tao Li, Lei Tan, Zhehao Huang, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional trajectory hypothesis is true: Dnns can be trained in tiny subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3411–3420, 2023. doi: 10.1109/TPAMI.2022.3178101.

Edo Liberty. Simple and deterministic matrix sketching. In Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy (eds.), *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pp. 581–588. ACM, 2013. doi: 10.1145/2487575.2487623. URL `https://doi.org/10.1145/2487575.2487623`.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989. doi: 10.1007/BF01589116. URL `https://doi.org/10.1007/BF01589116`.

Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S Kamilov. Image restoration using total variation regularized deep image prior. In *ICASSP 2019*, 2019. doi: 10.1109/ICASSP.2019.8682856.

Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers. A tutorial on fisher information, 2017. URL https://arxiv.org/abs/1705.01064.

James Martens. New insights and perspectives on the natural gradient method. *J. Mach. Learn. Res.*, 21: 146:1–146:76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2408–2417. JMLR.org, 2015a. URL http://proceedings.mlr.press/v37/martens15.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015b.

James Martens and Ilya Sutskever. Training deep and recurrent networks with hessian-free optimization. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller (eds.), *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pp. 479–535. Springer, 2012. doi: 10.1007/978-3-642-35289-8_27. URL https://doi.org/10.1007/978-3-642-35289-8_27.

Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: foundations and algorithms. *Acta Numer.*, 29:403–572, 2020. ISSN 0962-4929. doi: 10.1017/s0962492920000021. URL https://doi.org/10.1017/s0962492920000021.

Christopher A. Metzler, Ali Mousavi, Reinhard Heckel, and Richard G. Baraniuk. Unsupervised learning with stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

Taylor R. Moen, Baiyu Chen, David R. Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G. Fletcher, and Cynthia H. McCollough. Low-dose ct image and projection dataset. *Medical Physics*, 48(2): 902–911, 2021. doi: https://doi.org/10.1002/mp.14594. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14594.

Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings, 2017. URL https://arxiv.org/abs/1711.05139.

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

Silvia Scarpetta, Magnus Rattray, and David Saad. Matrix momentum for practical natural gradient learning. *Journal of Physics A: Mathematical and General*, 32(22):4047, 1999.

Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738, 2002.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5. URL http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

Zenglin Shi, Pascal Mettes, Subhransu Maji, and Cees G. M. Snoek. On measuring and controlling the spectral bias of the deep image prior. *International Journal of Computer Vision*, 130:885–908, 2022.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew Gordon Wilson. Pre-train your loss: Easy bayesian transfer learning with informative priors. *CoRR*, abs/2205.10279, 2022. doi: 10.48550/arXiv.2205.10279. URL https://doi.org/10.48550/arXiv.2205.10279.

Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020. doi: 10.1016/j.neunet.2020.07.025. URL `https://doi.org/10.1016/j.neunet.2020.07.025`.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9446–9454, 2018.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *Int. J. Comput. Vis.*, 128(7): 1867–1888, 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01303-4.

Wim van Aarle, Willem Jan Palenstijn, Jan De Beenhouwer, Thomas Altantzis, Sara Bals, K. Joost Batenburg, and Jan Sijbers. The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*, 157:35–47, 2015. doi: https://doi.org/10.1016/j.ultramic.2015.05.002.

Ge Wang, Jong Chul Ye, and Bruno De Man. Deep learning for tomographic image reconstruction. *Nature Machine Intelligence*, 2(12):737–748, 2020.

Hengkang Wang, Taihui Li, Zhong Zhuang, Tiancong Chen, Hengyue Liang, and Ju Sun. Early stopping for deep image prior. *CoRR*, abs/2112.06074, 2021. URL `https://arxiv.org/abs/2112.06074`.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2074–2082, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/41bfd20a38bb1b0bec75acf0845530a7-Abstract.html`.

Mitchell Wortsman, Maxwell Horton, Carlos Guestrin, Ali Farhadi, and Mohammad Rastegari. Learning neural network subspaces. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11217–11227. PMLR, 2021. URL `http://proceedings.mlr.press/v139/wortsman21a.html`.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, and Mehmet Akcakaya. Zero-shot physics-guided deep learning for subject-specific MRI reconstruction. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021. URL `https://openreview.net/forum?id=Nzv2jICkWV7`.

Gushan Zeng, Yi Guo, Jiaying Zhan, Zi Wang, Zongying Lai, Xiaofeng Du, Xiaobo Qu, and Di Guo. A review on deep learning MRI reconstruction without fully sampled k-space. *BMC Medical Imaging*, 21(1): 195, 2021. doi: 10.1186/s12880-021-00727-9. URL `https://doi.org/10.1186/s12880-021-00727-9`.