Model-corrected learned primal-dual models for fast limited-view photoacoustic tomography

Andreas Hauptmann *Senior Member, IEEE*, and Jenni Poimala

***Abstract*—Learned iterative reconstructions hold great promise to accelerate tomographic imaging with empirical robustness to model perturbations. Nevertheless, an adoption for photoacoustic tomography is hindered by the need to repeatedly evaluate the computational expensive forward model. Computational feasibility can be obtained by the use of fast approximate models, but a need to compensate model errors arises. In this work we advance the methodological and theoretical basis for model corrections in learned image reconstructions by embedding the model correction in a learned primal-dual framework. Here, the model correction is jointly learned in data space coupled with a learned updating operator in image space within an unrolled end-to-end learned iterative reconstruction approach. The proposed formulation allows an extension to a primal-dual deep equilibrium model providing fixed-point convergence as well as reduced memory requirements for training. We provide theoretical and empirical insights into the proposed models with numerical validation in a realistic 2D limited-view setting. The model-corrected learned primal-dual methods show excellent reconstruction quality with fast inference times and thus providing a methodological basis for real-time capable and scalable iterative reconstructions in photoacoustic tomography.**

***Index Terms*—Learned image reconstruction, inverse problems, photoacoustic tomography, model correction, primal-dual methods, limited-view.**

I. Introduction

LEARNED iterative reconstructions are highly popular due to their capabilities to combine model information and learned components to achieve state-of-the-art results with robustness to model perturbations [1]–[4]. In particular, so-called end-to-end approaches that unroll an iterative scheme for a fixed number of iterations and train all iterates jointly provide superior reconstruction quality in many applications. Nevertheless, applicability of such end-to-end approaches is largely limited to two-dimensional scenarios with small image sizes where training and evaluation is feasible. For high resolution imaging, or even three-dimensional data, we encounter several key difficulties. First of all, memory demands increase linearly with number of unrolled iterates and secondly, computationally expensive evaluations of the model equations can severely limit reasonable training times. This is especially so in photoacoustic tomography, where the forward model is given by an acoustic wave equation. Additionally, many photoacoustic measurement configurations employ a limited-view geometry making the inverse problem severely ill-posed, necessitating advanced reconstruction approaches and strong priors in the reconstruction process.

One such class of suitable advanced reconstruction methods consists of primal-dual approaches that have been very successful in classic model-based imaging scenarios [5], [6] and in particular its learned version still provides state-of-the-art results [7]. The combination of a learned update in image (primal) and data (dual) space is particularly powerful, but also necessitates a joint training in an end-to-end fashion leading back to the main limitations of adaption to high dimensions. In particular, the memory demand is even further increased due to the inclusion of an updating network in the measurement space.

There have been several recent advances to overcome these computational limitations in different settings, roughly following two paradigms:

(i) Uncoupling of network training from the operator evaluation.
(ii) Methodological improvements in the learned iterative approach to negate impact of high computational demand.

In the first category are so-called plug-and-play approaches, where a denoiser is learned offline and subsequently used in an iterative scheme such as ADMM or proximal gradient descent, replacing the proximal operator by the learned denoiser [4], [8]–[11]. Another alternative is a greedy training that requires iterative wise optimality instead of training the unrolled iterates end-to-end [12], [13], enabling an application to 3D limited-view photoacoustic tomography [12]. Unfortunately, the joint nature of learned-primal dual methods prohibits to adapt a greedy training strategy.

The second category is more varied, as approaches may tackle only a specific subproblem. For instance, an approximate model has been used in [14]–[16] to speed-up training and deployment of the models. Deep equilibrium (DEQ) models [17], [18], aim to reduce memory consumption by reformulating the unrolled algorithm to a fixed-point iteration and differentiating only with respect to the fixed-point equation, allowing for constant memory consumption independent of the amount of iterates. Another work [19] aimed to reduce both memory consumption and operator evaluations by formulating the unrolled algorithm as a multiscale approach, reducing the computational cost of forward operator and memory consumption on the lower scales. Finally, invertible neural networks can be used [20]–[22], eliminating the need to store intermediate

This work has been supported by the Academy of Finland projects 338408, 346574, 353093. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme *Rich and Nonlinear Tomography* where work on this paper was undertaken, supported by EPSRC grant no EP/R014604/1.

JP and AH are with the Research Unit of Mathematical Sciences, University of Oulu, Finland.

AH is also with the Department of Computer Science, University College London, UK.

states for backpropagation, but the need to evaluate the forward operator remains in an unrolled algorithm.

In addition to the above computational advances, there are recently increased efforts to provide theoretical results for learned image reconstructions [3], [10], [23], [24], for instance in the form of convergence results for unrolling approaches. One such convergence guarantee is given for DEQ models to a fixed point by requiring the updating operator to be a contraction [17], [18]. Stronger notions of convergence can be satisfied, but generally require more restrictive constraints on the learned components, often reducing expressivity and performance of the networks. We refer to [3] for an overview of provably convergent approaches for learned image reconstruction.

Motivated by the above developments, this work aims to formulate a model-corrected learned primal-dual model, that can be trained end-to-end for high resolutions using fast approximate forward operators. Additionally, we provide theoretical insights and conditions on fixed-point convergence in a deep equilibrium formulation. To achieve this, we start with the classic model-based primal-dual hybrid gradient [5] and replace the expensive forward and adjoint operators with a fast, but approximate version [14]. The introduced errors are corrected with an operator correction in dual space, motivated by [15], and a learned updating operator in primal space. The resulting algorithm is called a model-corrected learned primal-dual (MC-PD). We then proceed to formulate the model-corrected learned primal-dual as a deep equilibrium model (PD-DEQ) and analyze conditions on parameters and networks to provide a contraction on the primal variable. Our main results state that fixed-point convergence can be achieved under a sufficient decrease assumption on the dual variable.

The developed methods are evaluated extensively for a 2D limited-view scenario. We compare our methods to a post-processing approach with U-Net and a version of the proposed MC-PD that does not use weight-sharing and hence is close to the original learned primal-dual. We evaluate scalability in terms of memory consumption and evaluation times to larger image sizes. Our results show that MC-PD provides the same excellent performance with weight-sharing as well as without. The DEQ models do not perform as well as the MC-PD with a drop in PSNR and SSIM, but providing good qualitative results. Additionally, we discuss challenges when training the DEQ models in the considered limited-view setting and computationally verify the contraction of the updating network. Finally, we will provide codes and implementation of our models.

This paper is structured as follows. In sec. II we introduce the main concepts in photoacoustic tomography and the variational formulation to reconstruction. We then review learned reconstruction approaches and introduce the fast approximate operators used in this study. In sec. III we introduce the model-corrected learned primal-dual and its deep equilibrium formulation. We provide some theoretical insights to proposed models and then proceed to derive conditions on the contraction property. Sec. IV discusses computational aspects and data simulation. We present the results in sec. V and discuss performance and challenges of the proposed models. Finally, sec. VI provides final conclusions.

II. Photoacoustic tomography and prior work

In photoacoustic tomography (PAT) a short pulse of near-infrared light is absorbed by chromophores in biological tissue [25], [26]. For a sufficiently short pulse, this will result in a spatially-varying pressure increase x inside the tissue, which initiates an ultrasound (US) pulse (*photoacoustic effect*), that propagates to the tissue surface. In this work we only consider the latter acoustic problem of photoacoustic tomography. That is, the forward model of acoustic propagation is modeled by an initial value problem for the acoustic wave equation [27], [28]

$$(\partial_{tt} - c^2\Delta)p(\mathbf{x},t) = 0,\ p(\mathbf{x},t_0) = x(\mathbf{x}),\ \partial_t p(\mathbf{x},t_0) = 0, \tag{1}$$

with $\mathbf{x} \in \mathbb{R}^2$ and $t_0 = 0$ denotes the initial time. We measure the pressure field on the boundary of the computational domain Ω for a finite time window. The measured data can then be modeled by a linear operator $\mathcal{M}$ acting on the pressure field $p(\mathbf{x},t)$:

$$y = \mathcal{M}\, p_{|\partial\Omega\times(0,T)}. \tag{2}$$

The acoustic propagation (1) and the measurement operator (2) define the linear forward model

$$Ax = y \tag{3}$$

from initial pressure $x \in X$ to the measured time series $y \in Y$. This represents the ideal accurate forward model and can be simulated, e.g., by a pseudo-spectral time-stepping model as outlined in [29], [30].

If the measured data is sparse or the detection geometry is limited to only a part of the boundary the inverse problem becomes severely ill-posed [31], such as in the limited-view scenario considered here, see Figure 1 for an illustration of the measurement setup. This necessitates regularization and advanced priors to compensate for the lost information, which can be formulated in a variational setting [32]–[34]. That is, given measured data y we obtain a reconstruction as the minimizer of

$$J(x) = \|Ax - y\|_2^2 + \lambda R(x), \tag{4}$$

where the first term measures the data-fit and the second term regularizes the problem and incorporates *a priori* knowledge one might have about the target, $\lambda > 0$ balances the influence of both terms. Solutions of eq. (4) can then be computed iteratively by a suitable optimization algorithm. For instance, using a proximal gradient scheme that allows for non-differentiable $R(x)$: given an initial reconstruction x_0 one can iteratively minimize eq. (4) by the updating equation

$$x_{k+1} = \operatorname{prox}_{R,\lambda\gamma_k}\left(x_k - \gamma_k A^*(Ax_k - y)\right), \tag{5}$$

with suitable step-length $\gamma_k > 0$. The proximal operator projects iterates to solutions admissible by the regularisation term R, by solving a corresponding denoising problem

$$\operatorname{prox}_{R,\alpha}(x) = \arg\min_u \left\{ R(u) + \frac{1}{2\alpha}\|u - x\|_2^2 \right\}. \tag{6}$$

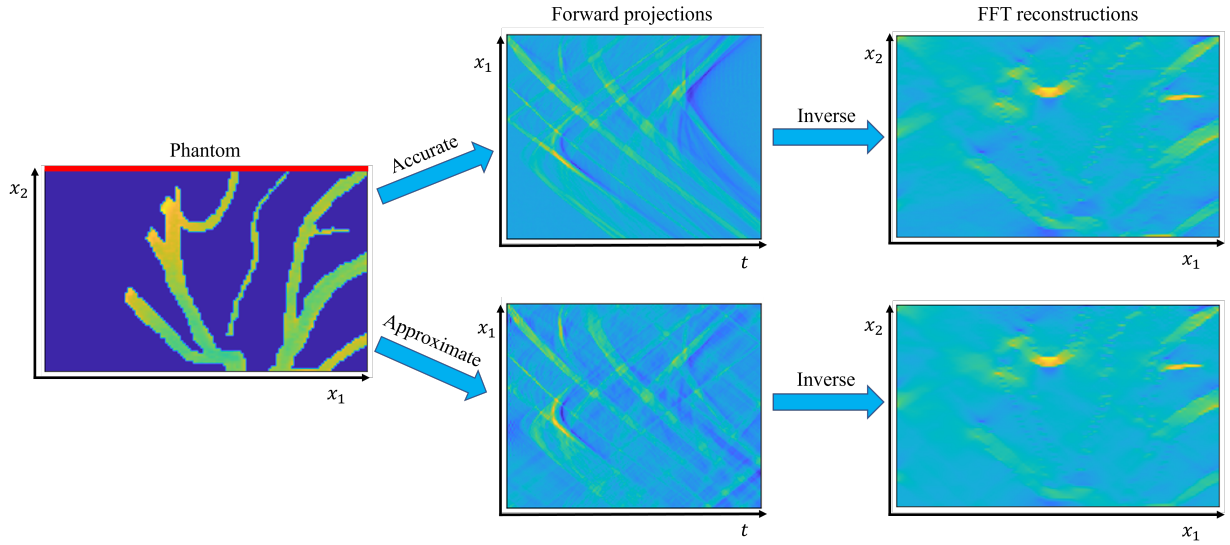

Fig. 1. Measurement setup and illustration of the fast models. The phantom (left) left is measured from a one-sided sensor on the top (red line). The ideal measurement is shown in the middle top. The fast approximate forward map is shown in the bottom and contains clear aliasing artifacts. Applying the inverse mapping provides reconstructions with strong limited-view artifacts. One can see on the bottom right, that the obtained reconstruction from the approximate model is smoother compared to the top.

Alternatively, utilizing the dual space we can formulate a min-max optimization problem in primal and dual space. This leads to the highly successful primal-dual hybrid gradient (PDHG) method, [5]. Here, the update equations for the primal variable $x \in X$ and dual variable $q \in Y$ to solve (4) are

$$q_{k+1} = \frac{q_k + \sigma(A\widetilde{x}_k - y)}{1+\sigma}, \tag{7}$$

$$x_{k+1} = \text{prox}_{R,\lambda\tau}\left(x_k - \tau A^* q_{k+1}\right), \tag{8}$$

$$\widetilde{x}_{k+1} = x_{k+1} + \alpha(x_{k+1} - x_k), \tag{9}$$

where $\sigma, \tau > 0$ and $\theta \in [0,1]$. Given the operator norm of the forward operator $L := \|A\| = \|A\|_{X\to Y}$ convergence to a saddle point for (q,x) is shown when $\sigma\tau L^2 < 1$ and $\theta = 1$ [5]. Let us point out, that eq. (7) corresponds to the proximal operator for the dual space, i.e., the least-squares data fidelity term in eq. (4). In the following we will base our method on the formulation in eqns. (7) and (8).

A. Learned model-based approaches to reconstruction

Learned reconstruction methods are increasingly popular in medical imaging, as they offer the possibility to combine strong data driven prior information with hand-crafted model components [1], while also providing a considerable speed-up of reconstruction times.

Most successful in terms of image quality and robustness are so-called model-based techniques that combine data-driven components with model components. That means, forward and adjoint operator are used within the network architectures [7], [12], [35], [36]. A prominent class are unrolled iterative methods that follow eq. (5) and, e.g., reformulate the updating equations to

$$x_{k+1} = G_{\theta_k}\left(x_k - \lambda_k A^*(Ax_k - y)\right). \tag{10}$$

That is, the proximal operator is replaced by a learned updating network G_{θ_k} with parameters θ, where the network weights are learned from supervised training data. Note, that in the above formulation the network weights are allowed to change in each iteration. We can also use weight-sharing and fix the parameters θ globally for all updates. Having a fixed updating operator allows for a more accessible analysis of the iterative reconstruction by constraining the network (weights), or the architecture of the updating operator [3].

A related approach are so-called plug-and-play methods [4] that use a separately learned denoiser G_θ, which is subsequently used in the iterative process. Note, also here the denoiser is fixed for all iterates. The nature of the network G_θ to act as a denoiser primarily limits this approach to inverse problems of denoising type, in contrast to the limited-view problem we encounter here. Nevertheless, recent advances show that if G_θ is given by a generative network one can obtain high-quality reconstructions and uncertainty maps in a related sampling framework [37].

Similarly to the classic model-based reconstruction in eqs. (5)-(8) we can extend the learned iterative approaches into a learned primal-dual method [7] by replacing the proximal operators in primal and dual space with learned operators. The learned updates are then given in a general form as

$$q_{k+1} = F_{\phi_k}\left(q_k, Ax_k, y\right)$$
$$x_{k+1} = G_{\theta_k}\left(x_k, A^* q_{k+1}\right).$$

The learned primal-dual scheme has shown superior performance for X-ray CT, but training is necessary to be performed in an end-to-end manner for all iterates jointly, due to the coupling of primal and dual updates, which can pose a major computational challenge. Some advances have been made by using invertible networks [38], [39] to reduce memory costs.

Nevertheless, in photoacoustic tomography the application of learned model-based techniques has been very limited [12],

[14], [40], even if their performance is shown to be quantitatively and qualitatively superior [41]. This is primarily due to the extensive computational cost of evaluating the forward model, which limits reconstruction speeds and often makes training prohibitive. Thus, the majority of learned approaches in PAT consider either one or two-step approaches, that is a direct mapping from data to reconstruction is learned or a post-processing network after an initial reconstruction is obtained, respectively. We refer to [42]–[44] for a comprehensive overview of learning based reconstruction methods for PAT.

As previously discussed, a possibility to overcome the limitations of the expensive forward model in the training procedure for PAT has been proposed in [12], by training the update operators in a greedy approach only requiring iterate wise optimality. The use of faster approximate models for planar (or linear) sensors [45], [46] further speeds up training and deployment times [12]. An alternative route considers to learn the forward model entirely, for instance by Fourier Neural Operators [47], [48] to embed these into an efficient iterative reconstruction algorithm.

B. A fast approximate forward and inverse model

Recent studies have investigated the effect of using approximate or imperfect operators in the reconstruction [49]–[51] and the possibility to improve results with a learned correction [15], [16]. Most importantly, it is shown that under a suitable learned correction, one can accurately solve the classic variational problem. We will first discuss the approximate model here and then introduce the learned reconstruction later in sec. III.

A fast forward and backward model is needed to enable feasible training of a learned iterative reconstruction. Here we will use an analytical approximation based on the Fast Fourier Transform (FFT), that is applicable for line and planar sensors [45], [46]. We note, that an FFT based algorithm also exists for circular domains [52] and thus our method extends naturally. We shortly review our approximate model as previously used in [14].

The approximate model exploits the fact that the measurement points lie on a line in 2D at $\mathbf{x}_2 = 0$ (or plane in 3D), and assuming constant sound-speed c. In two dimensions the pressure on the sensor can be related to x by [45], [46]:

$$p(\mathbf{x}_1, t) = \frac{1}{c^2}\mathcal{F}_{k_1}\left\{\mathcal{C}_\omega\left\{B(k_1,\omega)\tilde{x}(k_1,\omega)\right\}\right\}, \qquad (11)$$

where $\tilde{x}(k_1,\omega)$ is obtained from $\hat{x}(\mathbf{k})$ via the dispersion relation $(\omega/c)^2 = k_1^2 + k_2^2$ and $\hat{x}(\mathbf{k}) = \mathcal{F}_{\mathbf{x}}\{x(\mathbf{x})\}$ is the 2D Fourier transform of $x(\mathbf{x})$. $\mathcal{C}_\omega$ is a cosine transform from ω to t, and $\mathcal{F}_{k_1}$ is the 1D inverse Fourier Transform on the detector line. Finally, the weighting factor

$$B(k_1,\omega) = \omega / \left(\operatorname{sgn}(\omega)\sqrt{(\omega/c)^2 - k_1^2}\right), \qquad (12)$$

contains an integrable singularity which means that if (11) is evaluated by discretization on a rectangular grid to enable the application of the FFT for efficient calculations, then aliasing in the measured data $p(\mathbf{x}_1, t)$ results. Consequently, evaluating (11) using FFT leads to a fast but approximate forward model. One could control the strength of aliasing artifacts by thresholding certain incident angles as in [14], but this leads to a loss of signal strength. Thus, here we will not use a thresholding and deal with the aliasing by the learned model correction.

The relation in (11) then defines the approximate forward model $\widetilde{A} : X \to Y$ for this study. Similarly, we can invert eq. (11) to obtain a mapping from measured time-series to initial pressure as inverse mapping or reconstruction operator $A^\dagger : Y \to X$. This leads to two mappings between image X and data Y space:

$$\widetilde{A}x = \tilde{y}, \quad A^\dagger y = \tilde{x}. \qquad (13)$$

We note, that the fast inverse mapping does not introduce any aliasing, but information is naturally limited by the information present from the limited-view geometry, see Figure 1 for an illustration. In the following we will use the above fast mappings in the learned reconstruction.

Additionally, we have implemented the mappings in eq. (13) with PyTorch to enable flexible deployment in the training of the unrolled algorithms. This way, we have full GPU and automatic differentiation support of the models, enabling efficient training and evaluation. As part of this study, we will make our implementation of the models available to researchers.

III. Model corrected primal-dual and deep equilibrium models

We can now combine the above pieces to formulate a learned iterative scheme starting from the classic PDHG formulation (7) and (8). We will first include the model correction term for the forward operator and include a learnable update operator in the primal space. Additionally, our formulation allows to formulate the proposed algorithm as a deep equilibrium model.

A. Model corrected learned primal dual

The plug-and-play framework suggests to substitute the proximal operator in eq. (5) with a learned network. This is known to work well, if restricted to the primal space. The same approach is usually not applied to learning the proximal operator for the dual space, as we can not simply learn a denoising network in the dual space. Additionally, the proximal operator in dual space has a clear connection to the data-fidelity term and the noise model. Thus, for our method we will keep the explicit form of the proximal in dual space as given by eq. (7) and only replace the proximal operator for the primal update in image space with a network G_θ.

We replace the accurate but expensive model A with a fast approximate version $\widetilde{A}$, which would lead to errors in the data and hence the dual update. As solution, we propose to learn a correction of the forward operator by a network F_ϕ. Instead of using the (expensive) adjoint in the primal space, we will make use of the fast inverse mapping $A^\dagger$. Together, the update

steps with the model correction and learned proximal operator are then given by

$$q_{k+1} = \frac{q_k + \sigma(F_\phi(\widetilde{A}x_k) - y)}{1+\sigma} \tag{14}$$

$$x_{k+1} = G_\theta\left(x_k - \tau A^\dagger q_{k+1}\right). \tag{15}$$

Note, that we use weight sharing of the network parameters θ, ϕ for all iterates, and we set a maximum number of iterations $K > 0$. We refer to this unconstrained version as model-corrected learned primal-dual (MC-PD) as outlined in Algorithm 1.

Algorithm 1 Model-Corrected Primal-Dual (MC-PD)

Given $y, \widetilde{A}, A^\dagger$
Set σ, τ, K
$x_0 \leftarrow A^\dagger y$
$q_0 \leftarrow 0$
$k \leftarrow 0$
for $0 < k \leq K-1$ **do**
  $q_{k+1} \leftarrow \frac{q_k + \sigma(F_\phi(\widetilde{A}x_k) - y)}{1+\sigma}$
  $x_{k+1} \leftarrow G_\theta\left(x_k - \tau A^\dagger q_{k+1}\right)$
end for

B. A Primal-dual deep equilibrium model

In recent years convergent learned image reconstruction methods have gained considerable interest. Instead of simply training the learned reconstruction in a supervised method, we can constrain weights to obtain theoretical guarantees on the behavior of the unrolled iterates. One such option is to require that the learned iterates converge to a fixed point, this is the underlying paradigm of deep equilibrium models. In its simplest form, we can ensure that the learned proximal gradient in eq. (10) satisfies a fixed-point equation, that is

$$x^* = G(x^*; y) = G_\theta(x^* - \lambda A^*(Ax^* - y)). \tag{16}$$

First we note, that weights are shared between iterates and no maximum number of iterations is assumed, but rather a convergence in the limit is considered by ensuring, e.g, that the updates are a contraction. In deep equilibrium models the forward solver applies a fixed-point iteration, either a vanilla version or more advanced choices such as Anderson acceleration, see [17] for details and an extension to inverse problems in [18]. Training is then performed by differentiating the fixed-point formulation, eliminating the need to perform backpropagation through all iterates and hence reducing memory consumption.

In our case, we are given the two variables x, q for which we would like to formulate the deep equilibrium model. That means, we are now looking for a pair of fixed points $z^* = (x^*, q^*) \in X \times Y$, such that the fixed-point equation is satisfied with

$$(x^*, q^*) = \mathrm{PD}_{(\phi,\theta)}(x^*, q^*; y), \tag{17}$$

where the corresponding update equations for the mapping PD are given by eq. (14) and eq. (15). In the following we will first discuss the forward pass, followed by a discussion on the computation of backpropagation.

1) Forward pass for the deep equlibrium formulation: Here, we will be applying the Anderson accelerated fixed-point method to both, primal and dual, variables. More precisely, given the primal-dual update as above for iterate k:

$$(x_{k+1}, q_{k+1}) = \mathrm{PD}_{(\phi,\theta)}(x_k, q_k; y).$$

We denote with a superscript which update we compute, i.e., PD^q and PD^x, that is we get the formulation

$$q_{k+1} = \sum_{i=0}^{m-1} \beta_i \mathrm{PD}^q_{(\phi,\theta)}(x_{k-i}, q_{k-i}; y), \tag{18}$$

$$x_{k+1} = \sum_{i=0}^{m-1} \alpha_i \mathrm{PD}^x_{(\phi,\theta)}(x_{k-i}, q_{k-i}; y), \tag{19}$$

with memory $m \geq 1$, $\sum_i \alpha_i = 1$ and $\sum_i \beta_i = 1$. The parameter vectors α and β are computed by solving

$$\min_\alpha \|H\alpha\|_2^2, \text{ subject to } \sum_i \alpha_i = 1,$$

where the matrix H consists of the vectorized residuals

$$H = \Big[\mathrm{PD}^x_{(\phi,\theta)}(x_k, q_k; y) - x_k, \cdots, \mathrm{PD}^x_{(\phi,\theta)}(x_{k-m+1}, q_{k-m+1}; y) - x_{k-m+1}\Big],$$

and similarly we perform the same for β and the dual variable q. The forward pass of the primal-dual deep equilibrium model (PD-DEQ) is then given similarly as in Algorithm 1, except that new iterates are computed by the Anderson accelerated version in eqs. (18) and (19). Additionally, one could not only consider a maximum number of iterations but also check if the fixed-point iteration has converged, for instance, by setting a limit on the relative residual of the primal variable

$$r = \frac{\|PD^x_{(\phi,\theta)}(x_{k+1}, q_{k+1}; y) - x_k\|_2}{\|PD^x_{(\phi,\theta)}(x_{k+1}, q_{k+1}; y)\|_2}.$$

2) Backward pass: For training, even though we perform the update on both variables, we only optimize the network parameters with respect to a loss on the primal variable x, where training data is easily available. That means, we formulate the fixed-point equation in (17) only with respect to x. We write for a fixed point x^* under dual variable q and measurement y, with a slight abuse of notation,

$$x^* = \mathrm{PD}_{(\phi,\theta)}(x^*; q, y). \tag{20}$$

To simplify notation we denote the network variables as one $\zeta = \{\phi, \theta\}$. The networks are then trained supervised to minimize the mean-squared error (MSE) with respect to the groundtruth x by

$$\ell(\zeta) = \|\mathrm{PD}_{(\zeta)}(x^*; q, y) - x\|_2^2 = \|x^* - x\|_2^2.$$

The backward pass is computed by implicit differentiation of the fixed-point equation (20). We denote the Jacobian at the fixed point with respect to the parameters ζ by $J_{\mathrm{PD}^{x^*}}(\zeta)$ and the Jacobian with respect to x^* by $J_{\mathrm{PD}_{(\zeta)}}(x^*)$, then we obtain for the gradients the implicit equation

$$\nabla\ell(\zeta) = \left(J_{\mathrm{PD}^{x^*}}(\zeta)\right)^T \left(\mathrm{Id} - J_{\mathrm{PD}_{(\zeta)}}(x^*)\right)^{-T} \left(x^* - x\right). \tag{21}$$

Note, that we do not need to compute the backwards pass through all iterates and thus memory consumption is independent of the number of unrolled iterates. Nevertheless, eq. (21) requires a solver for the implicit equation, which can add some computational overhead. We will consider here the full formulation, but further reductions in the computation can be obtained by considering approximations of $J_{\mathrm{PD}_{(\zeta)}}(x^*)$, most notably a Jacobian free variant [53].

While our primary interest lies in the convergence of the primal x it is now reasonable to ask what do we need to assume on q. Let us discuss this in the following.

C. Convergence analysis

Here we will shortly examine some theoretical properties of the proposed methods. We investigate how the model correction can be interpreted in the MC-PD formulation. Followed by a discussion on the contraction and fixed-point properties of the deep equilibrium model.

1) Functional convergence under forward correction: We start by considering the analytic PDHG formulation in eqs. (7)-(9) using the approximate operator $\widetilde{A}$, its corresponding adjoint $\widetilde{A}^*$, and the proximal operator corresponding to some regularization R. Given the operator norm $\widetilde{L} = \|\widetilde{A}\|$, we can choose the hyperparameters as $\sigma = \tau < 1/\widetilde{L}$ and $\alpha \in [0, 1]$. Then, applying PDHG will minimize the cost functional

$$\|\widetilde{A}x - y\|_2^2 + \lambda R(x).$$

Minimizing the above functional with the approximate forward operator instead of the correction does result in suboptimal results as demonstrated in [15].

Consequently, let us now consider including the forward operator correction in the dual update, i.e., we replace A with $F_\phi(\widetilde{A})$ in eqn. (7), while using the correct adjoint in the primal update (8). We denote the operator norm of the correction F_ϕ by $\varepsilon_\phi = \|F_\phi\|$. Then, under a perfect training assumption on the adjoint we can recover convergence of PDHG as summarized in the following result.

Proposition 1. *Given corrected forward operator $F_\phi(\widetilde{A})$ with norm $\varepsilon_\phi\widetilde{L}$, accurate adjoint A^* with norm L. Assume that for $k \geq 0$ we have*

$$\big(F_\phi(\widetilde{A}x_k), y_k\big)_Y = \big(x_k, A^* y_k\big)_X. \tag{22}$$

Then for the choice $\tau < 1/L$ and $\sigma < 1/(\varepsilon_\phi\widetilde{L})$ we recover the iterations of PDHG in eqs. (7)-(9) *for the accurate functional* (4)*.*

Proof. The condition (22) requires that the forward correction does satisfy the adjoint equation. By uniqueness of the adjoint this is equivalent to a perfect training condition

$$\big(F_\phi(\widetilde{A}x_k) - Ax_k, y_k\big)_Y = 0 \quad \forall k \geq 0.$$

Thus, all iterates in (7)-(9) are identical to the accurate operators. By the choice of hyper parameters σ and τ we have that $\tau\sigma L\widetilde{L}\varepsilon_\phi < 1$ and thus the the convergence condition of the classic PDHG is satisfied. □

The condition (22) could be used for training without the need to access the accurate forward operator. That is one could add an additional adjoint loss to the training

$$L(x;\phi) = (F_\phi(\tilde{A}x), h)_Y - (x, A^*h)_X \text{ for } h \in Y. \tag{23}$$

Clearly, a perfect training can not be achieved in practice. In this case, it can be treated as a perturbation in the adjoint, or, in fact, in the forward. We refer to [54], [55] for a discussion on convergence under mismatched adjoints in primal-dual algorithms.

Note, that in the above adjoint loss (23) the accurate adjoint is still needed though. Thus, we now consider the case, where both forward and adjoint are replaced by their respective fast operators. Then, the corresponding adjoint loss becomes

$$L(x;\phi) = (F_\phi(\tilde{A}x), h)_Y - (x, A^\dagger h)_X \text{ for } h \in Y, \tag{24}$$

where adjointness of the forward correction to the fast inverse is enforced. Clearly, this does not minimize the accurate functional, but rather the corresponding functional

$$\|F_\phi(\widetilde{A})x - y\|_2^2 + \lambda R(x),$$

where closeness to the true model is governed by the closeness of the fast inverse $A^\dagger$ to the adjoint A^*. If convergence with respect to the accurate functional is desired, an adjoint correction is required [15].

2) fixed-point convergence for the deep equilibrium model: Let us now replace the proximal operator with a learned network G_θ and corresponding norm ε_θ. First we note, that if we do not enforce any further conditions on G_θ, we can not expect functional convergence anymore, but we can obtain a fixed-point convergence [3]. We remind first, that training of the PD-DEQ model is performed only with respect to the primal variable x. Thus, we would like to verify two conditions

(i) Under which parameter choices and conditions on the networks F_ϕ and G_θ do we obtain a contraction on the primal variable?
(ii) Can we also ensure fixed-point convergence?

As we will see, both conditions are closely connected as condition (ii) will require a contraction, but we will see that a dependence on q will not provide fixed-point convergence without further assumptions. To begin with, let us first combine the update equations (14) and (15) by

$$x_{k+1} = G_\theta\Big(x_k - \tau A^\dagger\big(q_k + \sigma\big(F_\phi(\widetilde{A}x_k) - y\big)\big)/(1+\sigma)\Big).$$

We can see here the dependence on q_k in the updates. Let us isolate the update for x by introducing the operator $T : X \to X$ such that

$$T_{\mathrm{Id}}(x) := (\mathrm{Id} - T)(x) = x - \tau A^\dagger\sigma F_\phi(\widetilde{A}x)/(1+\sigma). \tag{25}$$

Thus, T is a scaled version of the normal operator A^*A. We note, the above update rule on x (25) convergences to a fixed-point if T is firmly nonexpansive, which implies that $\mathrm{Id} - T$ is firmly nonexpansive [56]. To show firmly nonexpansiveness we need an adjoint loss condition. Rather than assuming a perfect fit, we consider an approximate adjoint loss condition, such that

$$(F_\phi(\tilde{A}x), h)_Y \leq (1+\varepsilon)(x, A^\dagger h)_X \text{ for } h \in Y. \tag{26}$$

That means we only fulfill the adjoint condition with an $\varepsilon > 0$ error depending on the value of the right inner product. This is a more realistic case, as we can not expect a perfect adjoint fit. We note, that the assumption (26) requires $(x, A^\dagger h)_X$ to be positive. However, since all samples x and relevant h are positive we can assume this to hold. Now we can show the firmly nonexpansiveness of T.

Theorem 1. *The operator $T : X \to X$ defined in* (25) *is firmly nonexpansive, under condition* (26) *with $\varepsilon > 0$, $L^\dagger = \|A^\dagger\|$, and $\sigma, \tau > 0$ chosen such that*

$$(1+\varepsilon)\frac{\tau\sigma}{1+\sigma}(L^\dagger)^2 \leq 1. \tag{27}$$

Proof. We need to show that

$$\|T(x) - T(v)\|_2^2 \leq (T(x) - T(v), x - v)_X, \tag{28}$$

where $T : X \to X$ is given as in (25). We start by bounding the left hand side of the above inequality, using linearity of $A^\dagger$ and $\|A^\dagger\| = L^\dagger$, we get

$$\begin{aligned}
&\Big\| \frac{\tau\sigma}{1+\sigma} A^\dagger F_\phi(\widetilde{A}x) - \frac{\tau\sigma}{1+\sigma} A^\dagger F_\phi(\widetilde{A}v) \Big\|_2^2 \\
&\quad \leq \Big| \frac{\tau\sigma}{1+\sigma} \Big|^2 \Big| L^\dagger \Big|^2 \Big\| F_\phi(\widetilde{A}x) - A^\dagger F_\phi(\widetilde{A}v) \Big\|_2^2 \\
&\quad = \Big| \frac{\tau\sigma L^\dagger}{1+\sigma} \Big|^2 \Big(F_\phi(\widetilde{A}x) - F_\phi(\widetilde{A}v), F_\phi(\widetilde{A}x) - F_\phi(\widetilde{A}v) \Big)_X .
\end{aligned}$$

We can now separate the inner product, using semi-linearity and symmetry. Then we estimate each term using the approximate adjoint condition in (26) and recombine, to obtain

$$\begin{aligned}
&\Big(F_\phi(\widetilde{A}x) - F_\phi(\widetilde{A}v), F_\phi(\widetilde{A}x) - F_\phi(\widetilde{A}v) \Big)_X \\
&\quad \leq (1+\varepsilon) \Big(A^\dagger F_\phi(\widetilde{A}x) - A^\dagger F_\phi(\widetilde{A}v), x - v \Big)_X .
\end{aligned}$$

Now under the condition (27) we get the required estimate. □

We continue by examining the iterations of the learned primal-dual updates and obtain a contraction for fixed k under boundedness of the dual updates.

Theorem 2. *Assume $\|q_k - q_{k-1}\|_2 < C_q$ with $k > 0$ fixed and $\|G_\theta\| = \varepsilon_\theta < 1$. Under the assumptions in Theorem 1, there exists a choice for $\tau > 0$, such that the iteration for x_{k+1} is contracting, i.e.,*

$$\|x_{k+1} - x_k\|_2 < \|x_k - x_{k-1}\|_2.$$

Proof. We start by reformulating the update for x_{k+1} using T_{Id} as

$$x_{k+1} = G_\theta \Big(T_{\mathrm{Id}}(x_k) - \frac{\tau}{(1+\sigma)} A^\dagger (q_k - \sigma y) \Big).$$

With the above update rule, using the norm of the network $\|G_\theta\| = \varepsilon_\theta$ and $\|A^\dagger\| = L^\dagger$ we can obtain by triangle inequality a first bound of iterates as

$$\begin{aligned}
&\|x_{k+1} - x_k\|_2 \leq \\
&\quad \varepsilon_\theta \Big(\big\| T_{\mathrm{Id}}(x_k) - T_{\mathrm{Id}}(x_{k-1}) \big\|_2 + \frac{\tau L^\dagger}{1+\sigma} \big\| q_k - q_{k-1} \big\|_2 \Big).
\end{aligned} \tag{29}$$

Set $0 < \delta = 1 - \varepsilon_\theta < 1$, with $\|q_k - q_{k-1}\|_2 < C_q$ and since $\|x_k - x_{k-1}\|_2$ is finite, there exists $\tau > 0$ small enough such that

$$\frac{\tau L^\dagger}{1+\sigma} C_q < \delta/2 \|x_k - x_{k-1}\|_2. \tag{30}$$

We then obtain the final estimate for eq. (29), using firmly nonexpansiveness of T_{Id} in the first term and (30) for the second term, by

$$\begin{aligned}
\|x_{k+1} - x_k\|_2 &\leq \varepsilon_\theta (1 + \delta/2) \|x_k - x_{k-1}\|_2 \\
&< \|x_k - x_{k-1}\|_2,
\end{aligned}$$

since by assumption

$$\varepsilon_\theta (1 + \delta/2) < \varepsilon_\theta + \delta/2 = (1 + \varepsilon_\theta)/2 < 1.$$

□

The above theorem only provides an iterate-wise convergence under a uniform bound on the dual q. That means, if q_k is not converging τ may become arbitrarily small. We note, that in practice we only consider finite iterates and hence a uniform bound exists. For the limit case we could assume that q_k converges at least as fast as x_k, then we can obtain a uniform choice for $\tau > 0$. For that, let us assume that both variables have a $1/k$ convergence.

Corollary 1. *Under the assumptions of Theorem 2, assume additionally that both $\|q_k - q_{k-1}\| = \mathcal{O}(1/k)$ and $\|x_k - x_{k-1}\| = \mathcal{O}(1/k)$. Then there exists a uniform $\tau > 0$ for all $k > 0$ such that the x_k are contracting.*

Proof. By assumption both sides of the estimate eq. (30) can be divided by k (with possible additional constants on each side). Then there exists a small enough $\tau > 0$ that holds for all k. □

First of all, the above analysis reveals that it is advisable to choose τ and σ conservatively small to ensure that the iterations contract. As we will see, this is in fact important to have stability in the DEQ formulation. Secondly, we would like small norm of the residuals in the dual q and ideally fast convergence. Third, it is interesting to note that in fact only the dual network G_θ is required to be contractive, while the model correction network F_ϕ is connected to the inverse $A^\dagger$ by the approximate adjoint error. In the experiments, we will see that q does indeed converge fast and the residuals have significantly smaller norm than for x. Thus, the experiments confirm that we are able to recover a fixed-point iteration on both primal and dual variables, even if the theory assumes dependence on the contraction of q.

IV. Photoacoustic data simulation and computational aspects

In this study we will evaluate the performance of the proposed algorithms with a simulated two-dimensional scenario. For this, we will first discuss the training data generation. We will then continue to discuss some details on the implementation of forward and inverse models, followed by the employed training procedures.

A. Photoacoustic simulation in 2D and training data

For simulating training data we use k-wave, which is based on a pseudo-spectral method, for accurate simulation of the acoustic wave equation (1) in two dimensions. We consider a rectangular computational domain of size 80×128 with isotropic spatial pixel size of $dx = 106\mu$m. The photoacoustic sensor is located at the top edge of the domain, see Figure 1 for an illustration. The sound speed is set to be constant at 1500 m/s and temporal sampling rate at $dt = 50$ ns, the measurement data is of dimension 160×128, with the spacing $dt \times dx$. We have added 1% Gaussian random noise to the simulated measurements, which corresponds to a SNR of roughly 18dB.

For training data generation we used the *DRIVE: Digital Retinal Images for Vessel Extraction*[1], consisting of retinal photographs and segmentation masks for the training. We have used 20 images and corresponding masks to create phantoms of vessel structures. First, by converting the images to grayscale and normalization, then by multiplying the mask with the images. Each resulting image has dimension 584×565, we have then extracted smaller patches of size 80×128 from the images and only saved those with enough features by checking the sum over all pixels, i.e., if greater than 150. We have repeated the same procedure for each image transposed for data augmentation, this resulted in a total of 893 images, we have further augmented this data set by flipping the image vertically, since this will provide different limited view artifacts. In total we obtained 1786 images with a split into $N = 1600$ for training and 93 for validation and testing each.

B. Implementing the fast models

For this study, we have implemented both models, fast approximate forward $\widetilde{A}$ and the fast inverse $A^\dagger$, fully in PyTorch, to allow for efficient evaluation. This also enables to use automatic differentiation within the model-based learned reconstructions. The implementation follows largely the description in sec. II-B using the inbuilt FFT with PyTorch. For the mapping between image k-space, $\mathbf{k} = (k_1, k_2)$, and measurement k-space, (k_1, ω), we have implemented our own grid interpolator within PyTorch, as no inbuilt functions were available.

For efficient computation during training, all necessary static variables and k-grids are pre-computed and passed to the fast model. This follows a similar paradigm as in the k-wave toolbox [29]. As part of this study we publish the fast models and network codes. [2]

C. Implementation and training of the model-corrected primal dual models

The networks F_ϕ and G_θ for the proposed models are both chosen as U-Nets, with 3 scales, i.e., 2 downsampling layers, 64 channels in the first scale, followed by 128 and 256 in the next scales. We set a maximum of 10 iterations for both algorithms, the Anderson acceleartion of PD-DEQ uses a memory of $m = 5$. For the hyperparameters τ and σ we have computed an approximate operator norm of $\widetilde{L} \approx 0.7$ and set $\tau = \sigma = 1/(10\widetilde{L})$. The overestimation by a factor 10 is due to the presented theory and provided improved stability in practice.

As mentioned earlier, we only train the networks with respect to the primal variable. We are given ground-truth images x^i and corresponding measurements y^i and the corresponding reconstruction operator $\mathcal{R}_\zeta : Y \to X$ (either MC-PD or PD-DEQ) with parameters $\zeta = \{\phi, \theta\}$. We then minimize the ℓ^2-loss (MSE) to find the optimal set of parameters ζ^* as

$$\zeta^* = \arg\min_{\zeta} \sum_{i=1}^{N} \|\mathcal{R}_\theta(y^i) - x^i\|_2^2. \tag{31}$$

We only use 1 sample per training iteration due to memory constraints and at each iteration the training sample is drawn from the uniform distribution over all training pairs. We use the Adam optimizer in PyTorch with a cosine decay on the learning rate initialized by $2{\cdot}10^{-4}$ and a total of 25000 training iterations. Training of the MC-PD models took 100 minutes and 180 minutes for the PD-DEQ models on a Nvidia Quadro RTX 6000 with 24 GB memory.

For the training of the PD-DEQ model, we use automatic differentiation at the fixed point formulation to compute the gradients given by eq. (21). We note that solving eq. (21) stably requires the network to be contractive. Thus, we have tested to use spectral normalization to explicitly constrain the Lipschitz constant of G_θ, but better results have been obtained without, we refer to Section V-C2 for a discussion. This suggests that the contraction is implicitly enforced by the fixed-point formulation. Nevertheless, we remind that a Lipschitz constraint is only needed for the update network G_θ in primal space, while the dual network F_ϕ can be unconstrained. We also point out that supervised training with respect to the dual variable is not straight-forward as the ideal reference should only be residual noise.

1) A combined hybrid approach: We will see that the PD-DEQ model has some difficulties to achieve high values of PSNR in the presented limited-view setting, despite providing good image quality. Thus, we will additionally consider a hybrid combination of both approaches, MC-PD and PD-DEQ. That means we first perform a fixed amount of MC-PD iterations, followed by PD-DEQ. The order is chosen such that first MC-PD will help to alleviate the impact of the limited-view setting, then the final PD-DEQ iterations provide the fixed-point formulation for to the final reconstructions. Here, we need to find a trade-off between the MC-PD performance and the memory reduction provided by PD-DEQ, as we will discuss later. We call this approach in the following PD-Hybrid

For the training objective, we train the network in an alternating way, that is in each training iteration we first evaluate MC-PD, denoted by $\mathcal{R}_{\zeta_1}^{MC}$ and minimize (31) for sample i. Then given the output $(x_k^i, q_k^i) = \mathcal{R}_{\zeta_1}^{MC}(y^i)$ we evaluate the PD-DEQ, denoted by $\mathcal{R}_{\zeta_2}^{DEQ}$, and minimize the loss function only for the parameters of the PD-DEQ model,

[1] https://drive.grand-challenge.org/

[2] Link will be added after revision

i.e.,

$$\|\mathcal{R}^{DEQ}_{\zeta_2}(x^i_k, q^i_k, y^i) - x^i\|^2_2.$$

We will use 5 iterations of MC-PD followed by 5 iterations of PD-DEQ. Training of the PD-Hybrid models takes about 160 minutes.

D. Comparison methods

We will further compare the performance of both model-corrected primal-dual models with two well-established methods. First we use a post-processing approach, where an initial reconstruction is obtained with the inverse $x_0 = A^\dagger(y)$, then a post-processing network is trained to remove noise and limited-view artefacts, i.e.,

$$x_{\text{rec}} = G_\theta(x_0) = G_\theta(A^\dagger(y)).$$

We will use here the same U-Net architecture for the network G_θ as in the MC-PD and PD-DEQ models.

Secondly, we will compare our methods to a version of MC-PD that does not use weight-sharing and thus is close to the original learned primal dual (LPD) [7], we will also refer to this model as LPD in the following. This network clearly has the most expressive power as each iteration has its own network weights in primal and dual space, and hence we will consider this here as the state-of-the-art to compare our method to. We note, that also here the networks are chosen with the same U-Net architecture. Training of the U-Net takes only 12 minutes and training the LPD takes 120 minutes.

V. Results and discussion

We will evaluate the proposed methods for their quantitative and qualitative performance, but also the ability to provide scalable and robust reconstructions. Followed by a discussion on the DEQ models.

A. Results for 2D simulated data

We first evaluate the performance on the test data as described in sec. IV-A, that is consistent with the training data, i.e., in a resolution of 80×128 with 1% noise in the measurements. For each algorithm we have trained 3 instances and chose the one with the highest validation error to compute the quantitative measures PSNR and SSIM on the test set of $N = 93$ samples, see Table I for the obtained values.

TABLE I
Quantitative measures (PSNR and SSIM) for all considered methods on the test data ($N = 93$).

	U-Net	LPD	MC-PD	PD-DEQ	Hybrid
PSNR	21.95	29.06	**29.34**	22.1	24.27
SSIM	0.943	**0.991**	**0.991**	0.959	0.974

There are two interesting observations that we can make. First of all, LPD and MC-PD perform very similar even though LPD does not use weight-sharing as the MC-PD does. This is largely consistent with the literature, that the weight-shared version performs similarly well as its counterpart. Interestingly, in this instance MC-PD was able to achieve even a slightly higher PSNR, but SSIM values are the same. Secondly, we notice that the constrained version PD-DEQ has a significant drop in PSNR. This is an unfortunate result, especially when compared to U-Net we only see a slight improvement. Nevertheless, we can observe a clear improvement in SSIM over post-processing with U-Net. The hybrid approach that combines MC-PD and PD-DEQ is able to improve both quantitative values, but can not reach the performance of the unconstrained version. This is in contrast to previous literature, which largely report that DEQ models can achieve nearly as good (or better) results as their unconstrained versions [18]. We point out here, that in our case we are in a severely ill-posed limited-view setting. Thus, we attribute this drop in performance to the added complications of the limited-view problem in PAT, as we will discuss further below.

Let us now examine the qualitative performance for one sample from the test set shown in Figure 2. We can immediately notice that the most striking differences are in the recovered quantitative values of each method. None, of the methods manages to recover the contrast very well in the left upper vessel, LPD and MC-PD perform best in this regard. The two constrained DEQ methods primarily show a loss of contrast in the center, whereas U-net has a strong overestimation as well as stark jumps in the contrast. In terms of limited-view artefacts, which are primarily at the bottom corners, all methods perform very well, only U-net and PD-DEQ show some slight blurring in the lower vessels. Generally, the visual appearance is remarkably good given the limited-view setting. In particular with respect to vessel shapes, the other three LPD, MC-PD, and PD-Hybrid perform very similarly and provide overall visually good results.

B. Scalability

The primary motivation to use approximate models and the DEQ formulation is to enable a scalable learned iterative method, which means applicability to large image sizes and short computation times to enable reasonable training times. Consequently, we will examine next how the different algorithms scale with respect to image size and computation times.

For this we have progressively increased the considered image size from the initial $80 \times 128 \approx 10^4$ pixels to $640 \times 1024 \approx 6 \cdot 10^5$ pixels, while keeping the network architectures fixed. We note, that the measurement data for the largest case is consequently of size 1280×1024. In Figure 3 we can see how the different methods scale on increasing data size. First of all, we see that the memory consumption of the PD-DEQ model and U-Net scale well with image size and can be applied on the final image size. While MC-PD runs out of memory already at a factor 3 of the initial image size. PD-Hybrid can leverage the reduced consumption and scales slightly better. We have omitted LPD here, as it would behave similar to MC-PD. We note, that better scalability with respect to memory consumption can be easily achieved by considering smaller networks (especially fewer channels) and for MC-PD and PD-Hybrid fewer iteration, wheres the DEQ memory consumption only depends on network size and is iterate independent. In terms of runtime, U-Net clearly outperforms

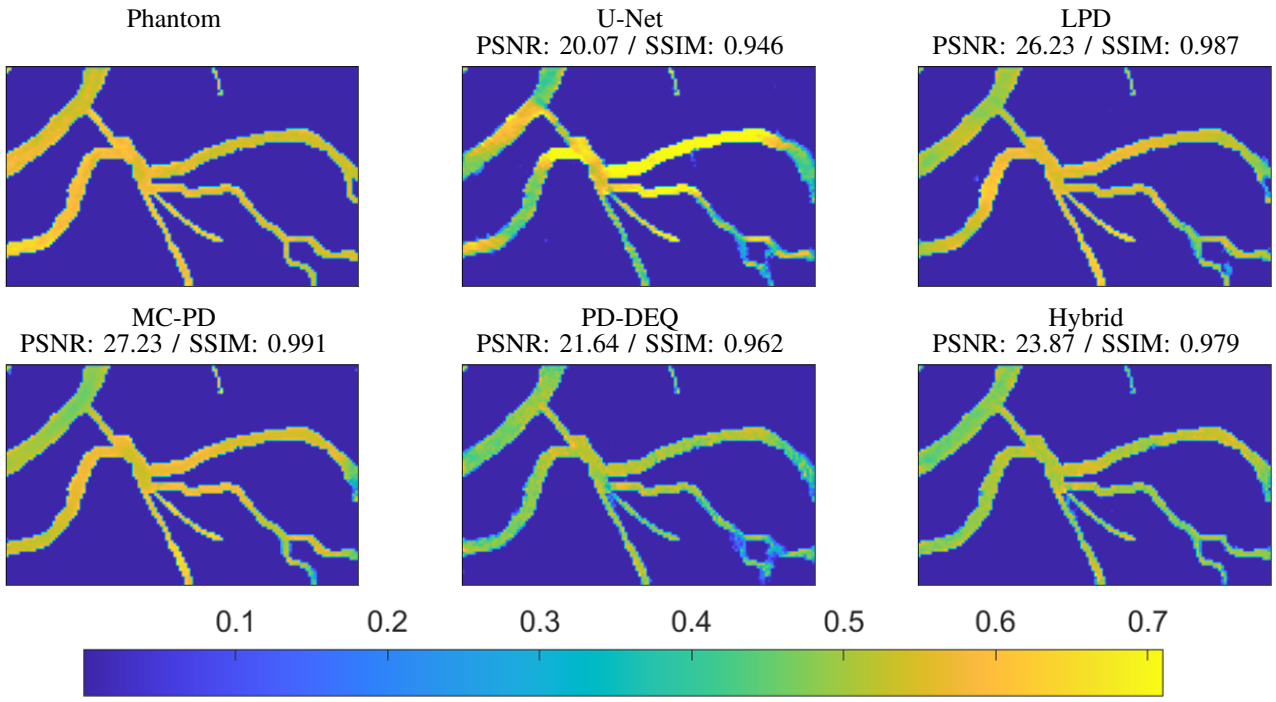

Fig. 2. Obtained reconstructions for the considered methods on a sample from the test set with quantitative measures. All images are on the same color scale. Measurements are taken on the top edge, with 1% Gaussian noise. The resolution of reconstructions is 80×128.

all other methods as it only requires one reconstruction with the fast inverse and one network evaluation. The iterative approaches all perform similar, with a slight but not significant overhead for the PD-DEQ model compared to PD-Hybrid and MC-PD. Up to a factor 4, i.e., image size of 320×512, we can achieve a runtime of less than 1 sec. for the PD-DEQ model. Here, a further speed up can be easily achieved by considering fewer unrolled iterations. We note that 1 second for the forward pass translates to roughly 7 hours for the training time with 25000 training iterates. Thus, even the largest image size for PD-DEQ would have still acceptable training times with $\sim$35 hours.

Finally, we examine how the models perform when applied to larger image size than trained on. For this test we used the full image size in the data generation and created samples of size 160×512 for testing. The results are shown in Figure 4. We can see that the networks work well when applied to a larger image size than trained on. While all iterative methods provide good image quality consistent with the results on the test set, we can see that U-Net post-processing creates some residual artifacts in the empty areas. Similarly as in the previous case, the constrained methods can not fully recover the contrast in the high absorbing region and some limited-view artifacts can be seen by slight blurred out features in the bottom corners.

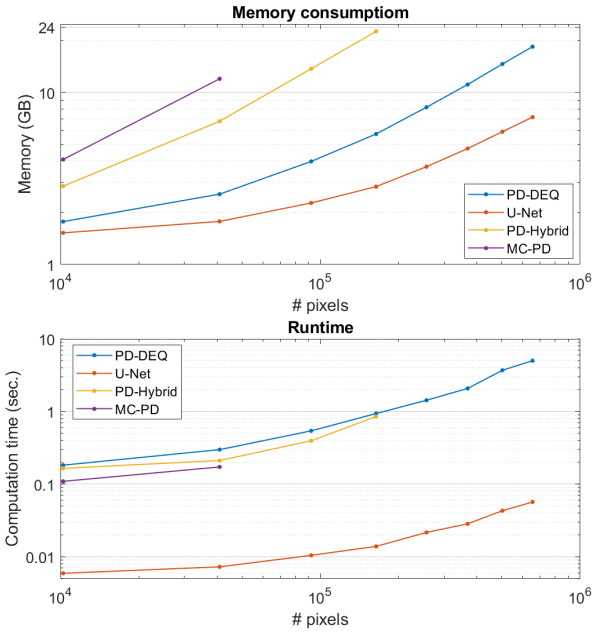

Fig. 3. Scalability tests for considered methods. The top shows memory consumption and the bottom plot shows runtime of the forward pass.

C. Challenges in the limited-view setting

In our experiments we have observed that training the DEQ models in the limited-view setting comes with some challenges. First of all, as we can see in the results the DEQ networks do not perform as well as their unconstrained counterparts. This can be in parts attributed to the fact that the constrained networks are limited in their expressivity, similar behavior can be observed for other constrained networks in inverse problems, such as convexity constraints [57]. Furthermore, the limited-view setting causes the initial reconstruction to have wrong contrast, due to a loss in energy from only one-sided measurements. To counteract this, we have scaled

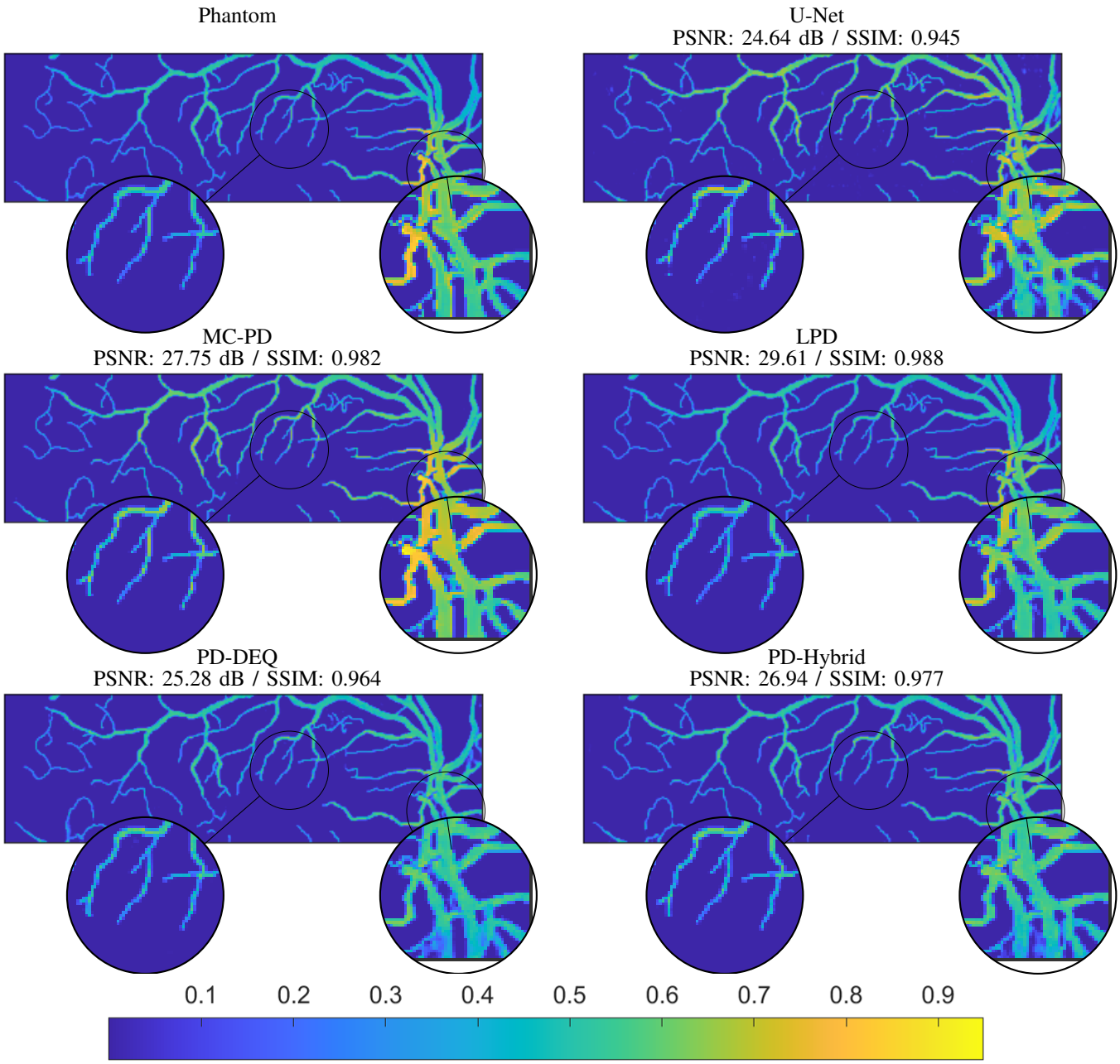

Fig. 4. Obtained reconstructions for the considered methods on larger test images of size 160×512, shown with quantitative measures. All images are on the same color scale. Measurements are taken on the top edge, with 1% Gaussian noise.

the input values (by factor 4) to have approximately the correct contrast between in $[0, 1]$. Nevertheless, the areas close to the detector are better resolved and have higher contrast than the bottom area. We postulate, that the contractive nature of the DEQ models is not ideally suited to correct in this scenario for this shortcoming and hence we can generally see a loss of contrast in the PD-DEQ and PD-Hybrid reconstructions compared to MC-PD as well as U-Net. We believe this is the main cause for the drop in PSNR.

1) Influence of algorithm parameters: The theory in Section III-C predicts that small values of τ, σ are needed to provide convergence. Indeed, we have observed that with too large $\tau, \sigma > 1/(5L)$ the iterations are unstable and training may not be successful. While too small values $< 1/(20L)$ will diminish the influence of the updates. Our experiments have shown that between those values good stability of the training and convergence is observed.

For the DEQ models, the number of iterations plays an important role as well. Here a similar behavior has been observed, too few iterates < 5 did not provide a sufficient improvement, while too many iterates > 20 often failed to converge and lead to unstable training. Thus, we have chosen 10 iterates for this study.

2) Spectral normalization: We have trained the primal update networks G_θ with and without spectral normalization. While spectral normalization stabilizes the training and larger

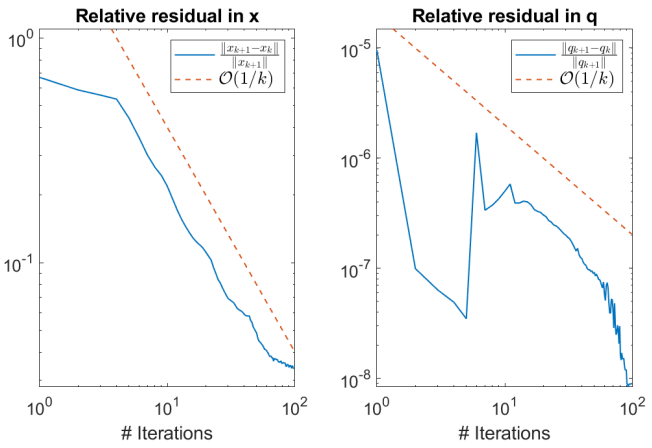

Fig. 5. Convergence of the residuals in the PD-DEQ formulation compared to a rate of $1/k$. Networks are trained for 10 iterates and tested for 100.

values of τ, σ were possible to choose, we have also observed that the training got more often stuck in bad local minima during training. Without spectral normalization early iterates may be unstable, but with small enough τ, σ we were able to reliably train the networks and better performance was observed for most training runs. As discussed in sec. IV-C, even without spectral normalization it has been observed that the training implicitly enforces the contractiveness and a stable training is possible. Additionally, spectral normalization adds a computational overhead. Consequently, we have performed the final training and tests without spectral normalization.

3) Training without the model correction: An ablation study for the DEQ models has been performed by training without the model correction network F_ϕ, which is closer to the original DEQ model. Unfortunately, we were not able to obtain sufficiently satisfying results. While the instability issues were exacerbated, the networks failed to find good local minima. Smaller values for $\tau, \sigma <$ helped to stabilize the training, but prevented good results. In our experiments we have not been able to exceed a PSNR of 15 dB in the validation set. This suggests that the model correction network in the dual space is indeed necessary in this setting.

D. Convergence of the PD-DEQ model

We have numerically tested how the PD-DEQ model behaves when we continue to run the iterates after the maximum of 10 training iterates has been hit. The theory would predict that the iterates are contractive in the primal variable, while the dual variable is uniformly bounded. This can be observed in Figure 5 for the test example shown in Figure 2. The residual in x does indeed contract with a rate of $1/k$, while the dual does not contract in the beginning, but has considerably smaller relative norm, after the initial 10 iterations also the dual variable contracts with a rate higher than $1/k$ as required for the fixed-point convergence. Nevertheless, the best reconstruction quality is obtained at the initial 10 iterations, while after those a further loss of contrast occurs.

VI. Conclusions

This study aimed to formulate a joint model-correction and learned iterative reconstruction that can be scaled to larger image sizes. We have formulated a model-corrected learned primal dual (MC-PD) and a corresponding deep equilibrium formulation (PD-DEQ). MC-PD provides excellent reconstruction quality, but requires higher memory demand due to the linear growth of the computational graph with number of iterates. PD-DEQ provides excellent scalability, but with quantitatively worse results. Additionally, we have proposed a PD-Hybrid method that balances both, improved reconstruction quality under reduced memory requirements. Further modifications on the hybrid approach are expected to improve quantitative results.

Secondly, this study aimed to provide further methodological and theoretical insights into learned iterative reconstructions and model corrections as well as the use of DEQ models in a limited-view setting. While our PD-DEQ implementation did not perform as well as the unconstrained version we also observed that a DEQ model without the model correction component could not be trained successfully at all. This suggests the importance of including a model correction and provides promising directions for future research.

References

[1] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[2] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[3] Subhadip Mukherjee, Andreas Hauptmann, Ozan Öktem, Marcelo Pereyra, and Carola-Bibiane Schönlieb. Learned reconstruction methods with convergence guarantees: A survey of concepts and applications. *IEEE Signal Processing Magazine*, 40(1):164–182, 2023.

[4] Ulugbek S Kamilov, Charles A Bouman, Gregery T Buzzard, and Brendt Wohlberg. Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023.

[5] Antonin Chambolle, Stacey E. Levine, and Bradley J. Lucier. An upwind finite-difference method for total variation-based image smoothing. *SIAM J. Imaging Sci.*, 4(1):277–299, 2011.

[6] Emil Y Sidky, Jakob H Jørgensen, and Xiaochuan Pan. Convex optimization problem prototyping for image reconstruction in computed tomography with the chambolle–pock algorithm. *Physics in Medicine & Biology*, 57(10):3065, 2012.

[7] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 2018.

[8] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013.

[9] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

[10] Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pages 5546–5557. PMLR, 2019.

[11] Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play priors: when langevin meets tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022.

[12] A. Hauptmann, F. Lucka, M. Betcke, N. Huynh, J. Adler, B. Cox, P. Beard, S. Ourselin, and S. Arridge. Model based learning for accelerated, limited-view 3d photoacoustic tomography. *IEEE Transactions on Medical Imaging*, 2018.

[13] Amir Aghabiglou, Matthieu Terris, Adrian Jackson, and Yves Wiaux. Deep network series for large-scale high-dynamic range imaging. *arXiv preprint arXiv:2210.16060*, 2022.

[14] Andreas Hauptmann, Ben Cox, Felix Lucka, Nam Huynh, Marta Betcke, Paul Beard, and Simon Arridge. Approximate k-space models and deep learning for fast photoacoustic reconstruction. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 103–111. Springer, 2018.

[15] Sebastian Lunz, Andreas Hauptmann, Tanja Tarvainen, Carola-Bibiane Schonlieb, and Simon Arridge. On learned operator correction in inverse problems. *SIAM Journal on Imaging Sciences*, 14(1):92–127, 2021.

[16] Danny Smyl, Tyler N Tallman, Jonathan A Black, Andreas Hauptmann, and Dong Liu. Learning and correcting non-gaussian model errors. *Journal of Computational Physics*, 432:110152, 2021.

[17] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Davis Gilton, Gregory Ongie, and Rebecca Willett. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:1123–1133, 2021.

[19] Andreas Hauptmann, Jonas Adler, Simon Arridge, and Ozan Öktem. Multi-scale learned iterative reconstruction. *IEEE transactions on computational imaging*, 6:843–856, 2020.

[20] Patrick Putzky, Dimitrios Karkalousos, Jonas Teuwen, Nikita Miriakov, Bart Bakker, Matthan Caan, and Max Welling. i-rim applied to the fastmri challenge. *arXiv preprint arXiv:1910.08952*, 2019.

[21] Jevgenija Rudzusika, Buda Bajic, Ozan Öktem, Carola-Bibiane Schönlieb, and Christian Etmann. Invertible learned primal-dual. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021.

[22] Rafael Orozco, Mathias Louboutin, and Felix J Herrmann. Memory efficient invertible neural networks for 3d photoacoustic imaging. *arXiv preprint arXiv:2204.11850*, 2022.

[23] Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, and Yves Wiaux. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, 2021.

[24] Housen Li, Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. Nett: Solving inverse problems with deep neural networks. *Inverse Problems*, 36(6):065005, 2020.

[25] P. Beard. Biomedical photoacoustic imaging. *Interface Focus*, 1(4):602–631, 2011.

[26] Ben Cox, Jan G Laufer, Simon R Arridge, and Paul C Beard. Quantitative spectroscopic photoacoustic imaging: a review. *Journal of biomedical optics*, 17(6):061202–061202, 2012.

[27] Peter Kuchment and Leonid Kunyansky. Mathematics of photoacoustic and thermoacoustic tomography. In *Handbook of mathematical methods in imaging*, pages 817–865. Springer, 2011.

[28] Kun Wang and Mark A Anastasio. Photoacoustic and thermoacoustic tomography: image formation principles. In *Handbook of Mathematical Methods in Imaging*. 2015.

[29] Bradley E Treeby and Benjamin T Cox. k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields. *Journal of Biomedical Optics*, 15(2):021314, 2010.

[30] Bradley E Treeby, Jiri Jaros, Alistair P Rendell, and B T Cox. Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method. *The Journal of the Acoustical Society of America*, 131(6):4324–36, 2012.

[31] Experimental evaluation of reconstruction algorithms for limited view photoacoustic tomography with line detectors. *Inverse Problems*, 23(6):S81, 2007.

[32] Simon Arridge, Paul Beard, Marta Betcke, Ben Cox, Nam Huynh, Felix Lucka, Olumide Ogunlade, and Edward Zhang. Accelerated high-resolution photoacoustic tomography via compressed sensing. *Physics in Medicine & Biology*, 61(24):8908, 2016.

[33] A novel compressed sensing scheme for photoacoustic tomography. *SIAM Journal on Applied Mathematics*, 75(6):2475–2494, 2015.

[34] Yoeri E Boink, Marinus J Lagerwerf, Wiendelt Steenbergen, Stephan A van Gils, Srirang Manohar, and Christoph Brune. A framework for directional and higher-order reconstruction in photoacoustic tomography. *Physics in Medicine & Biology*, 63(4):045018, 2018.

[35] Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[36] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[37] Jiaming Liu, Rushil Anirudh, Jayaraman J Thiagarajan, Stewart He, K Aditya Mohan, Ulugbek S Kamilov, and Hyojin Kim. Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. *arXiv preprint arXiv:2211.12340*, 2022.

[38] Buda Bajić, Ozan Öktem, and Jevgenija Rudzusika. 3d helical ct reconstruction with memory efficient invertible learned primal-dual method. *arXiv e-prints*, pages arXiv–2205, 2022.

[39] Nikita Moriakov, Jan-Jakob Sonke, and Jonas Teuwen. Lire: Learned invertible reconstruction for cone beam ct. *arXiv preprint arXiv:2205.07358*, 2022.

[40] Yoeri E Boink, Srirang Manohar, and Christoph Brune. A partially-learned algorithm for joint photo-acoustic reconstruction and segmentation. *IEEE transactions on medical imaging*, 39(1):129–139, 2019.

[41] Ko-Tsung Hsu, Steven Guan, and Parag V Chitnis. Comparing deep learning frameworks for photoacoustic tomography image reconstruction. *Photoacoustics*, 23:100271, 2021.

[42] Andreas Hauptmann and Ben T Cox. Deep learning in photoacoustic tomography: Current approaches and future directions. *Journal of Biomedical Optics*, 25(11):112903, 2020.

[43] Janek Gröhl, Melanie Schellenberg, Kris Dreher, and Lena Maier-Hein. Deep learning for biomedical photoacoustic imaging: A review. *Photoacoustics*, 22:100241, 2021.

[44] Changchun Yang, Hengrong Lan, Feng Gao, and Fei Gao. Review of deep learning for photoacoustic imaging. *Photoacoustics*, 21:100215, 2021.

[45] K P Koestli, M Frenz, H Bebie, and H P Weber. Temporal backward projection of optoacoustic pressure transients using Fourier transform methods. *Phys Med Biol*, 46(7):1863–1872, 2001.

[46] B. Cox and P. Beard. Fast calculation of pulsed photoacoustic fields in fluids using k-space methods. *J. Acoust. Soc. Am.*, 117(6):3616–3627, 2005.

[47] Ko-Tsung Hsu, Steven Guan, and Parag V Chitnis. Fast iterative reconstruction for photoacoustic tomography using learned physical model: Theoretical validation. *Photoacoustics*, page 100452, 2023.

[48] Steven Guan, Ko-Tsung Hsu, and Parag V Chitnis. Fourier neural operator network for fast photoacoustic wave simulations. *Algorithms*, 16(2):124, 2023.

[49] Leon Bungert, Martin Burger, Yury Korolev, and Carola-Bibiane Schönlieb. Variational regularisation for inverse problems with imperfect forward operators and general noise models. *Inverse Problems*, 36(12):125014, 2020.

[50] Arttu Arjas, Mikko J Sillanpää, and Andreas Hauptmann. Sequential model correction for nonlinear inverse problems. *arXiv preprint arXiv:2301.10094*, 2023.

[51] Ruanui Nicholson, Noemi Petra, Umberto Villa, and Jari P Kaipio. On global normal linear approximations for nonlinear bayesian inverse problems. *Inverse Problems*, 2023.

[52] Leonid Kunyansky. Fast reconstruction algorithms for the thermoacoustic tomography in certain domains with cylindrical or spherical symmetries. *Inverse Problems and Imaging*, 6(1):111–131, 2012.

[53] Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. Jfb: Jacobian-free backpropagation for implicit networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6648–6656, 2022.

[54] Dirk A Lorenz and Felix Schneppe. Chambolle–pock's primal-dual method with mismatched adjoint. *Applied Mathematics & Optimization*, 87(2):22, 2023.

[55] Emilie Chouzenoux, Andrés Contreras, Jean-Christophe Pesquet, and Marion Savanier. Convergence results for primal-dual algorithms in the presence of adjoint mismatch. *SIAM Journal on Imaging Sciences*, 16(1):1–34, 2023.

[56] Heinz H Bauschke, Sarah M Moffat, and Xianfu Wang. Firmly nonexpansive mappings and maximally monotone operators: correspondence and duality. *Set-Valued and Variational Analysis*, 20:131–153, 2012.

[57] Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Learned convex regularizers for inverse problems. *arXiv e-prints*, pages arXiv–2008, 2020.