Data-driven gradient regularization for quasi-Newton optimization in iterative grating interferometry CT reconstruction

Stefano van Gogh, Subhadip Mukherjee, Michał Rawlik, Alexandre Pereira, Simon Spindler, Marie-Christine Zdora, Martin Stauber, Zsuzsanna Varga and Marco Stampanoni

***Abstract*— Grating interferometry CT (GI-CT) is a promising technology that could play an important role in future breast cancer imaging. Thanks to its sensitivity to refraction and small-angle scattering, GI-CT could augment the diagnostic content of conventional absorption-based CT. However, reconstructing GI-CT tomographies is a complex task because of ill problem conditioning and high noise amplitudes. It has previously been shown that combining data-driven regularization with iterative reconstruction is promising for tackling challenging inverse problems in medical imaging. In this work, we present an algorithm that allows seamless combination of data-driven regularization with quasi-Newton solvers, which can better deal with ill-conditioned problems compared to gradient descent-based optimization algorithms. Contrary to most available algorithms, our method applies regularization in the gradient domain rather than in the image domain. This comes with a crucial advantage when applied in conjunction with quasi-Newton solvers: the Hessian is approximated solely based on denoised data. We apply the proposed method, which we call GradReg, to both conventional breast CT and GI-CT and show that both significantly benefit from our approach in terms of dose efficiency. Moreover, our results suggest that thanks to its sharper gradients that carry more high spatial-frequency content, GI-CT can benefit more from GradReg compared to conventional breast CT. Crucially, GradReg can be applied to any image reconstruction task which relies on gradient-based updates.**

***Index Terms*— Grating Interferometry, Iterative Reconstruction, Machine Learning, Regularization, Tomography**

I. Introduction

Over the last decades there has been an increasing interest in X-ray phase contrast CT due to the higher soft-tissue contrast that can be attained compared to absorption-based CT [1]. Among the many techniques that have been developed [2]–[4], grating interferometry (GI) [5]–[7] arguably holds the biggest potential to undergo a successful transition from synchrotron facilities to the clinics. GI yields three signals which arise from distinct physical phenomena, i.e. absorption, refraction and small-angle scattering, which could provide complementary diagnostic information. Several *ex-vivo* studies have shown promising results of GI in the context of planar breast imaging [8], [9]. Consequently, the logical next step is to extend the technology to three-dimensional CT. To this end, we have built a large field-of-view (FOV) prototype to assess the technical feasibility of scanning large specimens with grating interferometry CT (GI-CT) [10].

Given that GI-CT has been extensively described in recent articles [10]–[12] and since this paper revolves around algorithmic concepts, which are not limited to this particular use case but are applicable to virtually any imaging modality, we will only give a brief overview of GI-CT and focus instead on the main problems our approach seeks to solve. When used together with conventional X-ray tubes, GI requires three gratings placed between the source and the detector. The first grating, commonly called G0, is placed behind the source to increase the coherence of the beam. The G1 grating is a phase grating that creates a spatial interference pattern called Talbot carpet. Finally, a third grating called G2 is placed before the detector to analyze shifts and amplitude reductions in this pattern, which in turn allows one to measure the X-ray refraction and small-angle scattering, respectively. In GI-CT the source, the detector, and the gratings jointly rotate around the scanned subject to acquire sufficient angular projections for a tomographic reconstruction.

There are two main challenges associated with the tomographic reconstruction of these data: high noise amplitudes [13], [14] and ill problem conditioning [11]. The noise amplitudes in GI are more significant compared to conventional CT because the G2 grating absorbs half of the flux. Therefore, for a given dose, half of the photons will hit the detector in GI-CT compared to conventional CT. Moreover, the uncertainty in the measured data is further increased by the modest visibility, i.e. the amplitude of the interference pattern, that can be achieved with currently available gratings [10]. In fact, the latter plays a crucial role in the uncertainty propagation in GI [13]. In particular, low visibilities lead to higher uncertainties in the phase- and small-angle scattering images. On the other hand, the ill problem conditioning mainly affects the phase-contrast channel and manifests itself by very slow convergence during iterative reconstruction, which consequently exacerbates the noise accumulation [11]. It is intrinsically linked to the physics of the signal acquisition in GI-CT, in particular to the

Stefano van Gogh (e-mail:stefano.van-gogh@psi.ch), Michał Rawlik, Alexandre Pereira, Simon Spindler, Marie-Christine Zdora, and Marco Stampanoni are with the ETH Zürich and the Paul Scherrer Insitute. Subhadip Mukherjee is with the Indian Institute of Technology (IIT) Kharagpur. Martin Stauber is with GratXray. Zsuzsanna Varga is with the University Hospital Zürich.

differential nature of the refraction measurement as well as to the periodicity created by the gratings. Note that the bad conditioning of GI-CT's phase contrast operator is not directly related to the ill-conditioning that often arises in CT due to angular under-sampling. In fact, GI-CT acquisitions can be ill-conditioned despite dense angular sampling.

In light of these challenges, solving the inverse problem in GI-CT is highly complex. As we show in [12], the ill-conditioning can be circumvented by assuming a fixed non-linear dependence between the phase and the absorption channel which enables fusing the strengths of both channels into one. This allows to reconstruct an image that takes the low spatial frequency information from absorption while taking the high frequencies from phase. This strategy enabled us to reduce the dose necessary to achieve a given spatial resolution compared to conventional CT. However, the problem of high noise amplitudes persists and becomes particularly cumbersome at low-dose acquisitions. To address the noise problem, promising reconstruction results were obtained in [11] by combining physics-based likelihoods with data-driven priors. In particular, data-driven Plug-and-Play regularization, i.e. the use of an off-the-shelf denoising algorithm that acts as a regularizer in iterative reconstruction, was combined with a quasi-Newton solver to iteratively reconstruct phase-contrast tomographies based on differential phase contrast sinograms. Specifically, in our previous work, we applied the denoising step in the image space every k iterations and restarted the quasi-Newton updates after each denoising step. This strategy was proposed because, as we will explain in detail below, it is not possible to efficiently denoise after each iteration in the image space when using quasi-Newton solvers. However, this has the disadvantage that noise accumulates over a larger number of iterations, which makes the regularization more cumbersome.

In this paper, we build upon these previous works and present an algorithm that seamlessly combines data-driven denoising with a quasi-Newton solver to reconstruct virtually noise-free fused absorption-phase contrast tomograms by applying the regularization step in the gradient space. In particular, contrary to the previously published work [11], the regularization step takes place after each likelihood update, thus greatly facilitating denoising, and consequently improving the performance. This article is organized as follows. We will first briefly explain the fused iterative reconstruction algorithm developed in [12]. We will then highlight a feature of this reconstruction algorithm that motivated us to develop a novel regularization strategy which, contrary to most regularization approaches, operates on the image gradients, rather than in image space. Next, we will show on both *in-silico* and real data that the proposed regularization strategy, which we call GradReg, significantly improves the reconstruction quality of both conventional CT and GI-CT. Finally, we demonstrate that GI-CT reconstruction benefits more from the GradReg compared to conventional CT.

II. Methods

A. Fused intensity-based iterative reconstruction

In a recent article [12] we proposed a new reconstruction algorithm for GI-CT called fused intensity-based iterative reconstruction (FIBIR) in which a single fused absorption-phase contrast image is reconstructed, alongside the dark-field image. The method is an extension of two previously published algorithms which seek to jointly reconstruct the three image channels [15], [16], which however fail to achieve robust convergence under clinically compatible acquisition conditions. The FIBIR algorithm is based on a quasi-Newton solver, i.e. the L-BFGS algorithm [17], which optimizes the following loss function:

$$\begin{aligned}\operatorname{argmin}_{\mu,\epsilon} \frac{1}{2} \| \log\left(I/I_0\right) + A_\mu \mu \\ - \log(1 + V_0 \exp\left[-A_\epsilon \epsilon\right] \cos(\varphi_0 - A_\delta m_\theta[\mu]))\|_2^2. \quad (1)\end{aligned}$$

I_0, V_0 and φ_0 are the flat-field intensity-, visibility- and phase-map. μ represents the fused absorption-phase image and ϵ is the dark-field image. m_θ is an image-to-image mapping function defined as

$$m_\theta(\mu) = \begin{cases} \frac{\delta_a}{\mu_a}\mu, & \mu < \frac{\mu_a+\mu_f}{2} \\ \frac{\delta_f}{\mu_f}\mu, & \frac{\mu_a+\mu_f}{2} \leq \mu \leq \frac{\mu_f+\mu_s}{2} \\ \frac{\delta_s}{\mu_s}\mu, & \frac{\mu_f+\mu_s}{2} < \mu \end{cases} \quad (2)$$

Here, μ_a, μ_f, μ_s are the linear attenuation coefficients for adipose, fibroglandular and skin tissue at E_d, i.e. the design energy of our prototype, which are related to the imaginary part of the index of refraction via $\mu_i = 4\pi\beta_i/\lambda$. $\delta_a, \delta_f, \delta_s$ are the real parts of the indices of refraction of adipose, fibroglandular and skin tissue at E_d.

We found that with FIBIR, GI-CT can outperform conventional CT in a clinical-dose regime. More specifically, we showed that it is possible to either increase the spatial resolution at a given dose level, or to decrease the dose for a given spatial resolution. To allow for a fair comparison on equal image quality, the reconstructed images were compared at a contrast-to-noise ratio (CNR) of 5, which was obtained by filtering the raw tomograms with a Gaussian denoising kernel [12].

Examination of the raw tomograms however revealed that while GI-CT images contain more information compared to conventional CT, and in particular more high-frequency information, they are also affected by more noise since the data acquired in GI-CT has half the counts compared to conventional CT. This can also be observed when looking at the gradient images of the loss with respect to the two signals during reconstruction (see Fig. 1). Note that here the term gradient image does not refer to an image which has been processed with a gradient (or finite-difference) filter but to the gradients of the loss function we seek to optimize during reconstruction. Since regularization can remove unwanted noise but cannot artificially create signal, we thus hypothesized that GI-CT could benefit more from regularization compared to conventional CT.

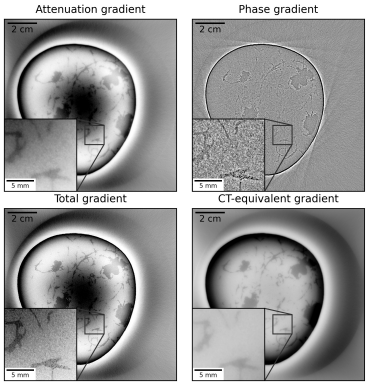

Fig. 1. FIBIR-enabled GI-CT yields sharper but noisier gradients compared to conventional CT. On the top, from left to right: attenuation and phase gradient in FIBIR. On the bottom, from left to right: the total FIBIR gradient, i.e. the combination of the two above, and the CT-equivalent gradient. The latter refers to iterative reconstruction of conventional CT, i.e. absorption-only. All gradients were taken at iteration 7. As in [12], this iteration number was chosen because in the first iterations the gradients carry only coarse signals while at later iterations, when the image is close to convergence, the gradients do not contain much signal anymore.

B. Gradient denoising for quasi-Newton optimization

Before explaining the concepts behind GradReg, let us consider a prototypical inverse problem in tomographic reconstruction where we wish to reconstruct the tomogram x based on the acquired projections y. With a (possibly non-linear) forward operator A modelling the imaging physics and assuming Gaussian noise we seek to minimize the likelihood function

$$\mathcal{L} = \frac{1}{2}\|Ax - y\|_2^2. \tag{3}$$

Usually, gradient-based optimization schemes are employed to solve this problem with the following update rule:

$$x_{k+1} = x_k - \gamma \nabla_x \mathcal{L}, \tag{4}$$

where γ is the step size and $\nabla_x \mathcal{L}$ is the gradient of the loss function with respect to the image x_k at iteration k. In these schemes the image is thus obtained by iteratively refining it based on the available gradient information. When the loss function is convex, which is the case for linear inverse problems, these types of optimization schemes are guaranteed to converge [18].

In the real world, the measurement y is always noisy and, consequently, the gradient $\nabla_x \mathcal{L}$ as well. Therefore, more noise will be added to the image iterate x_k at every iteration. To avoid reconstructing a corrupted image, prior knowledge can be introduced into the inverse problem in the form of regularization. Most existing regularization strategies act in the image space as they represent some prior knowledge about the images we seek to reconstruct. This can be achieved by two main approaches. In the first one a regularization term $R(x)$ is explicitly added to the reconstruction loss

$$\mathcal{L} = \frac{1}{2}\|Ax - y\|_2^2 + \lambda R(x). \tag{5}$$

Here, λ is a scalar which determines how strongly the prior knowledge shall be weighted during reconstruction. A second type of update rule that can be used, especially if $R(x)$ is not continuously differentiable or if it is computationally infeasible to compute $\nabla_x R(x)$, is

$$x_{k+1} = \mathcal{T}\left(x_k - \gamma A^T(Ax_k - y)\right), \tag{6}$$

where $\mathcal{T}$ is a projection operator which effectively acts as an image denoising/artefact removal step. The latter approach can be viewed as an implicit type of regularization which is implemented by alternating a gradient-based data update and a regularization or denoising step in the image space.

In this article, we are interested in the latter type of regularization. A multitude of classical approaches that fit into this scheme has been proposed over the last decades, with most projection operators involving either a thresholding or a shrinkage step [19], [20]. In recent years the concept of data-driven regularization has been gaining popularity and, in the context of alternating schemes, Plug-and-Play-like denoising has received a lot of attention due to its astonishing performance [21]–[24]. In these approaches, the projector $\mathcal{T}$ is parameterized by a neural network f_θ trained a-priori on a representative set of data. After training this then leads to the following update rule:

$$x_{k+1} = f_\theta\left(x_k - \gamma \nabla_x \mathcal{L}\right). \tag{7}$$

Instead of regularizing in the image space, we propose to regularize, i.e. denoise, the gradients $g_k(x) = \nabla_x \mathcal{L}$ at every iteration k:

$$g_{denoised,k}(x_k) = f_\theta\left(g_k(x_k)\right). \tag{8}$$

One might ask why do we need to denoise the gradients if image-space Plug-and-Play works so well? In fact, they do perform extremely well, but only when first-order algorithms are used. In such cases, each iterate x_{k+1} is projected to the manifold of clean images and will thus be noise-free. Sometimes, however, plain gradient descent-based algorithms suffer from slow convergence, especially when dealing with highly ill-conditioned problems. Therefore, in those cases one might want to use quasi-Newton methods which can significantly increase convergence speed, and possibly limit noise accumulation.

Quasi-Newton methods aim at leveraging information not only about the steepness of the loss function, i.e. the gradient, but also about its curvature, i.e. the Hessian. Unfortunately, computing H explicitly is computationally infeasible for large-scale imaging problems. Luckily, it can be approximated at every iteration by employing the secant equation which relates the difference in iterates $(x_{k+1} - x_k)$ and the difference in gradients $(g_{k+1} - g_k)$ with the Hessian:

$$(g_{k+1} - g_k) \approx H(x_{k+1})(x_{k+1} - x_k). \tag{9}$$

This then allows to compute an approximate Hessian B_{k+1} at iteration $k+1$ with

$$B_{k+1}(x_{k+1} - x_k) = g_{k+1} - g_k. \tag{10}$$

The approximated Hessian can then be used to compute the image update. Different quasi-Newton methods have been proposed [25], but most update the image iterate x_{k+1} as a function of a history of approximate Hessians B_i at previous iterations i. An example of such an update rule is the L-BFGS method [17] which stores the difference in gradients $y_k = g_{k+1} - g_k$ and the difference in iterates $s_k = x_{k+1} - x_k$ at every iteration. In the first iteration, the update direction p is given by the gradient, i.e. $p = g_0$. For later iterations the descent direction is computed as follows [17]:

$$
\begin{aligned}
&p = g_k \\
&\rho = 1/(y_k^T s_k) \\
&\text{for } i = k-1, k-2, \ldots, k-m \\
&\qquad \alpha_i = \rho_i s_i^\top p \\
&\qquad p = p - \alpha_i y_i \\
&\gamma_k = \frac{s_{k-1}^\top y_{k-1}}{y_{k-1}^\top y_{k-1}} \\
&H^0 = \gamma_k I \\
&p = H^0 p \\
&\text{for } i = k-m, k-m+1, \ldots, k-1 \\
&\qquad \beta_i = \rho_i y_i^\top p \\
&\qquad p = p + s_i(\alpha_i - \beta_i) \\
&p = -p
\end{aligned}
$$

Applying classical Plug-and-Play methods in a naive way to quasi-Newton optimization schemes thus yields suboptimal results. In fact, the update rule would be based on data coming from two different functions, i.e. gradients originating from a noisy function g_k and clean iterates x_k coming from a denoising step. Consequently, it would yield noisy approximations of the Hessian B_k. On the contrary, in our approach g_k will be noise free and thus x_k will also be noise free, thus yielding an update rule which is based solely on denoised data. Therefore a noise-free approximation of the Hessian B_k will be obtained. Note that for linear inverse problems, gradient- and image-space denoising yield equivalent secant equations as $g_{k+1} - g_k$ is independent of the noise in those cases. However, since quasi-Newton methods compute Hessian approximations with recursive algorithms which require the current gradient, gradient-space denoising offers superior results even on linear inverse problems.

The benefits of GradReg can be analyzed by looking at the simple experiment displayed in Fig. 2. We simulated a noisy quadratic loss function, minimized it with gradient descent and saved both the iterates and the gradients. Assuming perfect denoising, we then plotted the approximated Hessians computed with (10) for each of the iterations, once for gradient-space denoising (GradReg) and once for image space denoising. The second plot in Fig. 2 shows that, if we assume perfect gradient denoising, this leads to near perfect approximations of the Hessian in GradReg. On the contrary, perfect image space denoising leads to highly unstable approximations of the Hessian due to the noise that flows into (10) from the corresponding gradients. If we now used the Hessian approximations obtained with image space denoising together with a quasi-Newton solver, optimization would become highly unstable. On the contrary, a stable solution could be found with GradReg.

Since in this work we made use of the L-BFGS algorithm, we will focus our further analysis on this particular update rule. However, the proposed concept holds for all methods which aim at approximating the Hessian based on a history of gradients and image iterates [26].

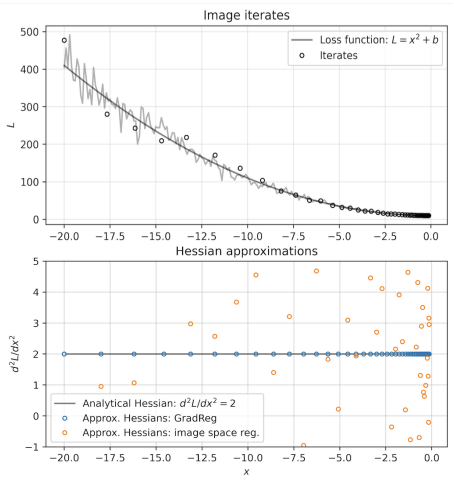

Fig. 2. The effect of gradient space denosing on the approximation of the Hessian. In the first plot a noisy quadratic loss is shown together with the iterations steps needed to minimize it with gradient descent. In the second plot the approximations for the Hessian are shown for GradReg (blue) and image-space denoising (orange). A perfect denoiser was assumed in this experiment.

C. Data generation pipeline

In order to regularize the gradients with a data-driven algorithm we had to provide the means to train the denoising model f_θ. We decided to opt for a supervised training strategy in which the model learns to map noisy gradients to their corresponding clean ground truths.

We simulated noisy and clean sinograms using the *in-silico* absorption and phase breast phantoms introduced in [27]. For the dark-field channel, we simulated simple dots representing microcalcifications. We projected the absorption, phase and dark-field phantoms with the ASTRA toolbox [28] into the sinogram space, and then combined simulated phase stepping with the GI forward model

$$I_0 \exp\left[-A_\mu \mu\right](1 + V_0 \exp\left[-A_\epsilon \epsilon\right] \cos(\varphi_0 - A_\delta \delta)) \qquad (11)$$

to generate clean phase stepping data I_{clean}. Finally, we simulated Poisson noise to obtain realistic phase stepping data I_{noisy}. While in this work we used phase stepping data, the developed methods are also applicable to stepping-free scans with fringe-based phase maps as discussed in [12]. We would like to point out that we used *in-silico* data in this article as we do not possess sufficient amounts of real data yet to train a data-driven algorithm on actual measurements. However, the presented approach is compatible with real data as well, in which case the clean ground truth would be replaced by high-quality scans. If no high-quality scans are available, unsupervised or self-supervised approaches [29]–[32] could also be envisioned.

A straight-forward way to generate the gradient images used as training data is to take noisy and clean sinograms,

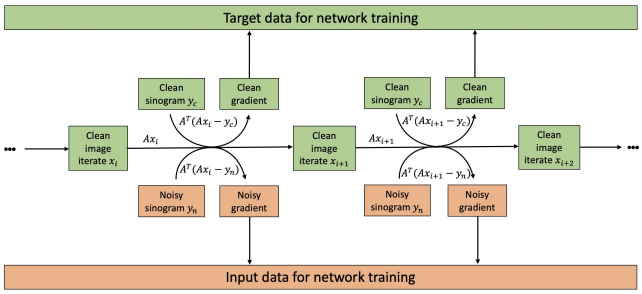

Fig. 3. Data generation pipeline. The clean data is displayed in green, the noisy data in orange. The pipeline can be applied for an arbitrary number of iterations. This figure displays the data generation for a linear inverse problem (see (3)) with a gradient descent-based update rule. The forward operator is therefore indicated by $\boldsymbol{A}$, the backward operator by $\boldsymbol{A^T}$, the image iterate by $\boldsymbol{x_i}$ the noisy and clean measurement data by $\boldsymbol{y_n}$ and $\boldsymbol{y_c}$, respectively. In this work the operator was based on (13) and the L-BFGS update rule was used instead of gradient descent. The terms *input* and *target* refer to the data intended for training the regularization network. Specifically, the former refers to the inputs for the denoiser, whereas the latter to the labels, i.e. the training signals. On the contrary, the inputs for the data generation pipeline are the noisy and clean sinograms.

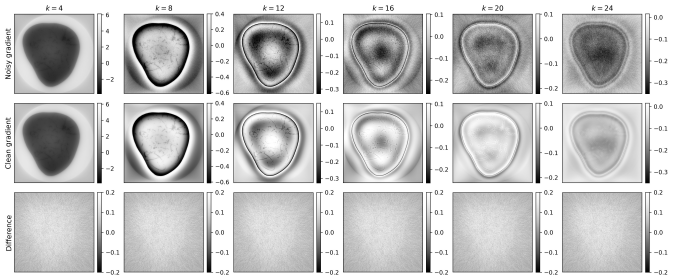

Fig. 4. Noisy and clean FIBIR gradients as inputs for the regularization network. From top to bottom: noisy gradients, clean gradients and their difference images. From left to right: increasing iteration numbers $\boldsymbol{k = 4, 8, 12, 16, 20, 24}$.

iteratively reconstruct the tomograms and save the gradients at each iteration. Unfortunately, this does not generate the data we need to train our model. In fact, the data we simulate must allow the denoiser to learn what it shall do at inference time during reconstruction. Since our denoiser has to remove the noise which is introduced by a single likelihood-based update step, we must generate training data in which the corrupted gradient images contain the noise backprojected from one single update step. On the contrary, storing the gradients that arise in unregularized iterative reconstruction leads to increasing amounts of noise with increasing iterations. The regularizer would therefore face an increasingly arduous denoising task.

This can be avoided by running an (unregularized) reconstruction pipeline which takes both the noisy and the clean sinograms as input. More specifically, we start with an empty image x_0, use the noisy sinogram to compute the first noisy gradient and save it. We then use the clean sinogram and repeat the same process. However, this time we also update our image iterate with the clean gradient, which will leave us with a clean image iterate that can be used to compute the subsequent noisy and clean image gradients. By repeating the process a dataset that contains noisy and clean gradients can be built in which the noise arises from a single update step. The entire data-generation process is displayed in Fig. 3 for a generic image reconstruction task, i.e. not specifically for GI-CT.

To train a network f_θ, we ran ten simulations to generate the training data and ten to generate the validation data. The testing data consisted of a single simulation experiment. Each simulated experiment had an image size of $16 \times 1340 \times 1340$ voxels and a sinogram size of $16 \times 1200 \times 1340$ pixels. We saved 30 gradients per experiment as we observed that after 30 iterations the reconstructions had always converged, i.e. only noise was being added to the reconstructions. The total number of 2D gradient images in both the training and validation sets was therefore $10 \times 16 \times 30 = 4800$, yielding a total dataset size of roughly 140 GB for every network trained in this study.

Fig. 4 shows an example of the generated training data. There are two interesting observations to be made. First, the image content significantly varies among iterations. Specifically, the first iterations mostly introduce trends into the tomograms, i.e. large features, whereas the later ones add details, i.e. small features. Second, with increasing iterations the noise intensity stays constant (see difference images in Fig. 4), whereas the signal intensity decreases (see the ranges of the colorbars). As we will see below, this fact can be leveraged to enhance the denoising performance.

D. Residual network architecture for gradient denoising

Since the image gradients vary significantly over the course of the iterations, we initially hypothesized that training a single network to denoise them all would be challenging because the underlying data distribution is spread out over a large space. In fact, training a convolutional neural network (CNN) to map all noisy gradients to their clean counterparts proved to be difficult. However, since there exists a common feature across all iterations, i.e. the noise, we reformulated the problem to leverage this prior knowledge. In particular, we used a CNN with a residual loss which learns to fit the difference between the input and the target image, i.e. that learns to extract the noise rather than to fit the signal. This led to considerably better results. It is important to note that classical filters could have been used instead of the more complex data-driven denoiser. However, this led to considerably worse results. We attributed this to two main reasons. First, strong noise amplitudes cannot be easily removed with simple filters. Second, since gradient images vary significantly from iteration to iteration, it is not possible to apply a classical filter with the same hyperparameters to each iteration, thus requiring careful hyperparameter tuning for each iteration. On the contrary, a neural network-based denoiser automatically learns to process all iterations.

The network f_θ is thus trained with the following loss function

$$\arg\min_\theta \frac{1}{S}\frac{1}{k_{\max}} \sum_s \sum_{k=1}^{k_{\max}} \left\| g^s_{k,\mu,noisy} - f_\theta(g^s_{k,\mu,noisy}) - g^s_{k,\mu} \right\|_2^2. \tag{12}$$

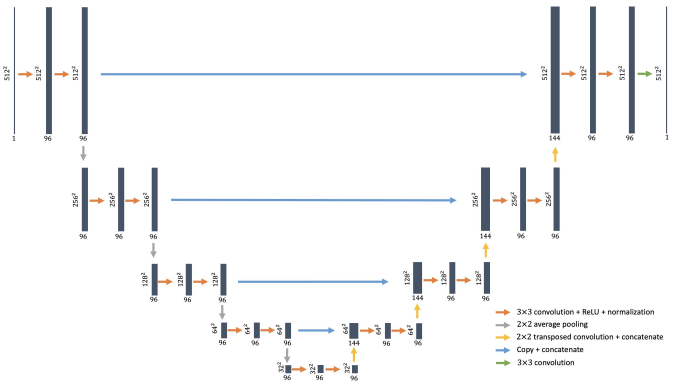

Fig. 5. Architecture of the network $f_{\boldsymbol{\theta}}$ used to denoise image gradients. The data that flows through the network is represented by dark-blue rectangles. The channel dimension is given below each rectangle, the spatial dimension on their left. Different operations are highlighted by arrows of different colors. The network is trained to fit the noise in the input gradient.

$g^s_{k,\mu,noisy}$ refers to the noisy 2D gradient at iteration k coming from sample s and $g^s_{k,\mu}$ to its clean counterpart. S is the dataset size and $k_{\max}$ is the highest iterations number, which was set to 30. f_θ was parameterized with a U-net-like architecture with 5 scales and 96 channels per layer, implemented in Tensorflow [33] (see Fig. 5). Downsampling was performed with average pooling and a feature-wise trainable normalization was applied to each layer. We used ReLU activation functions and the kernel size in each channel was set to 3. The model had a total of 1656769 parameters. The Adam algorithm [34] was used to optimize the loss with a decaying learning rate starting at 10^{-4} on a NVIDIA Titan GPU with 24GB of memory. At each iteration, we randomly chose a sample s and an iteration number k corresponding to one particular gradient image in our dataset. The images were then randomly cropped to 512×512 to fit into the GPU memory. All models were trained until convergence, i.e. until the validation loss saturated. Consequently, the training time varied considerably from experiment to experiment, with an average of around 63 hours.

E. The reconstruction algorithm

The full reconstruction algorithm, which combines the L-BFGS optimizer and the data-driven gradient regularizer, is depicted in Algorithm 1. In what follows the method will be described for GI-CT reconstruction. However, the GI-CT forward model can be replaced with the forward model associated with any other image reconstruction task, be this in tomography, medical imaging or beyond.

In its GI-CT implementation the algorithm requires the measurement data I, the flat-field data (I_0, V_0, φ_0), empty starting images (μ_0, ϵ_0) for the fused absorption-phase channel and for the dark-field channel, the forward operators $(A_\mu, A_\delta, A_\epsilon)$ associated with the three channels, the trained gradient regularizer f_θ, the number of L-BFGS iterates to consider m, the maximal number of iterations $k_{\max}$ and the mapping function m_θ. The forward operators are based on spherically symmetric blob functions [11], [35], [36] and A_μ is equal to A_ϵ in the current implementation of the algorithm.

The algorithm starts by forward projecting the image channels based on

$$I_0 \exp\left[-A_\mu \mu\right](1 + V_0 \exp\left[-A_\epsilon \epsilon\right] \cos(\varphi_0 - A_\delta m_\theta[\mu])). \quad (13)$$

Next, it computes the gradient with respect to μ and ϵ. Subsequently, the gradient with respect to μ is fed into the denoising network. We did not apply gradient denoising to the dark-field channel in this work as we were only interested in the fused channel here. However, we expect the method to be applicable also to the dark-field channel by training a separate denoiser for the latter. Once the fused gradient is denoised, it is concatenated with the dark-field gradient. The same holds for the two image iterates. Together they are then used to compute the image update based on the L-BFGS update rule [17]. Finally, this update is subtracted from the current iterate to yield the next one, and the two images and gradients are decatenated, thereby concluding one iteration of our algorithm.

The reconstruction algorithm is very similar to the data-generation pipeline. The main difference lies in the fact that instead of using a parallel stream to generate the clean gradients needed to train the denoising network, it uses the latter to regularize the gradients during reconstruction.

One iteration of our algorithm takes around 45 seconds for an image volume of $1340 \times 1340 \times 16$ with sinograms of size $5 \times 1200 \times 1340 \times 16$ on a NVIDIA Titan GPU with 24GB of memory. Once the network has been trained, the regularization step is very fast. In fact, it has a negligible computation time in the order of tens of milliseconds per slice. A large fraction of the computation time is thus used by the forward and backward operators.

F. Regularization in GI-CT vs. CT-equivalent

In [12] we compared the performance of GI-CT and conventional CT at different radiation doses. The results suggested that the former could either increase the resolution or lower the dose compared to conventional CT. Here, we want to explore how GradReg influences this comparison.

We therefore performed a simulation study at 3 different radiation doses: 4 mGy, 8 mGy and 16 mGy. As in [12] we matched the photon counts in the projection space of our ASTRA toolbox-based [28] ray-tracing simulations to the counts of CT scans for which dose calculations had been performed by the means of Monte Carlo simulations, where the simulation geometry and source parameters were validated through dosimetric measurements [10]. We used a visibility of 13% which corresponds to the current mean value we can reach on our setup. At each dose, we generated training data as explained in section II-C, i.e. at each dose 10 image volumes of $1340 \times 1340 \times 16$ voxels were generated for training and 10 for validation. For each of them 30 noisy and clean gradient iterations were saved. We then trained separate networks for every dose level, one for FIBIR and one for the CT-equivalent, thus leading to a total of 6 trained networks.

Once the networks were trained we used them to regularize the gradients within iterative reconstruction as described in Algorithm 1. The reconstructions were stopped at iteration

Algorithm 1 Reconstruction algorithm with L-BFGS optimizer and data-driven gradient regularization

Require: $I, (I_0, V_0, \varphi_0), (\mu_0, \epsilon_0), (A_\mu, A_\epsilon, A_\delta), f_\theta, m, k_{\max}, m_\theta$
  $k \leftarrow 0$
  while $k < k_{\max}$ **do**

$\mathcal{L}_k = \frac{1}{2}\|\log(I/I_0) + A_\mu\mu_k - \log(1 + V_0 \exp[-A_\epsilon\epsilon_k]\cos(\varphi_0 - A_\delta m_\theta[\mu_k]))\|_2^2$ ▷ Compute loss
$g_{k,\mu} \leftarrow \nabla_\mu \mathcal{L}_k$ ▷ Compute gradient of fused image
$g_{k,\epsilon} \leftarrow \nabla_\epsilon \mathcal{L}_k$ ▷ Compute dark-field gradient
$g_{k,\mu} \leftarrow g_{k,\mu} - f_\theta(g_{k,\mu}, k)$ ▷ Residual denoising of gradient of fused image
$g_k \leftarrow [g_{k,\mu}, g_{k,\epsilon}]$ ▷ Concatenation
$x_k \leftarrow [\mu_k, \epsilon_k]$ ▷ Concatenation

if $k > 1$ **then** ▷ Start L-BFGS update calculation
  $y_k \leftarrow g_{k+1} - g_k$
  $\rho_k \leftarrow 1/(y_k^T s_k)$
end if
$p \leftarrow g_k$
if $k > 0$ **then**
  $\text{counter}_{\max} \leftarrow \max(-1, k - m - 1)$
  $i \leftarrow k - 1$
  while $i > \text{counter}_{\max}$ **do**
    $\alpha_i \leftarrow \rho_i s_i^T p$
    $p \leftarrow p - \alpha_i y_i$
    $i \leftarrow i - 1$
  end while
  $\nu_k \leftarrow s_{k-1}^T y_{k-1} / y_{k-1}^T y_{k-1}$
  $H_0 \leftarrow \nu_k$
  $p \leftarrow H_0 p$
  $\text{counter}_{\min} \leftarrow \max(0, k - m)$
  $i \leftarrow \text{counter}_{\min}$
  while $i < k$ **do**
    $\beta_i \leftarrow \rho_i y_i^T p$
    $p \leftarrow p + s_i(\alpha_i - \beta_i)$
    $i \leftarrow i + 1$
  end while
end if ▷ End L-BFGS update calculation

$\gamma \leftarrow$ Line search
$x_{k+1} \leftarrow x_k - \gamma p$ ▷ Image update
$s_k \leftarrow x_{k+1} - x_k$
$[g_{k,\mu}, g_{k,\epsilon}] \leftarrow g_k$ ▷ Decatenation
$[\mu_k, \epsilon_k] \leftarrow x_k$ ▷ Decatenation
$k \leftarrow k + 1$
  end while

30, which was after the norm of the difference between two reconstruction iterates did not significantly change anymore. In parallel, we reconstructed all simulated data also without regularization. To compare all these experiments we then analyzed both spatial resolution and reconstruction error (MSE) of the final tomographies.

III. Results

A. Gradient regularization performance

As a first step, we inspected the performance of the regularization network in denoising the image gradients. Fig. 6 shows an example of the denoising results for different iterations at a dose of 16 mGy. In the first iterations the denoising performance is nearly perfect. With increasing iterations, the quality naturally slightly deteriorates since the signal intensity decreases while the noise intensity stays constant. A look at iteration 16 and 20 shows that the denoiser is still able to extract a significant amount of signal from the noisy gradient, even though the task becomes highly challenging for the human eye. Finally, in the last iterations the denoiser is not able anymore to extract any further signal. Note though that in these last iterations the output of the network is almost an empty image (see iteration 24), which suggests that after convergence the algorithm does not start overfitting to the noise since practically nothing is added anymore to the image iterates at this stage. Importantly, while the network is unable to extract the little remaining signal towards the later iterations, we observed that the network is not hallucinating any structures which are not present in the data.

The results in Fig. 6 thus confirm that, even though the image content varies significantly between iterations, one single network trained with a residual architecture can satisfactorily process them all. There is therefore no need to condition the architecture on the gradient iteration.

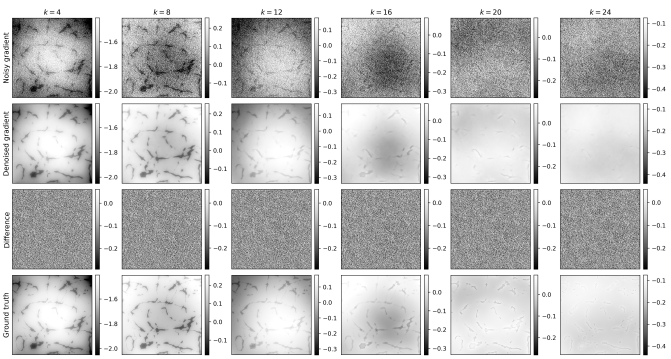

Fig. 6. Gradient denoising. From top to bottom: noisy gradients, denoised gradients, difference between noisy and denoised gradients, and ground truth (clean) gradients. From left to right, different iterations are plotted at $\boldsymbol{k = 4, 8, 12, 16, 20, 24}$. Zoomed-in sections of the full gradient images are plotted with a FOV of 30 × 30 mm and with a pixel size of 75 µm.

B. Regularized reconstruction

After having verified the successful gradient denoising, we will now analyze the reconstruction results. The focus will initially be on the 16 mGy simulation to show some important characteristics of GradReg. Next, a comparison across different radiation dose levels will be provided. Fig. 7 shows the reconstruction error of the image iterates, in terms of the MSE, compared to the clean ground truth. It compares the performance of the regularized and non-regularized algorithm, for both FIBIR-enabled GI-CT and conventional CT. As it was shown in [12], we observe that the unregularized FIBIR-based reconstruction leads to faster convergence compared to its CT-equivalent counterpart since it reaches its lowest MSE within less iterations, i.e. 13 vs. 23 in this case. Moreover, we see that the regularized dashed lines steadily decrease whereas the unregularized solid lines increase once the algorithms start fitting the noise, thus confirming the excellent regularizing effect of GradReg. This is also consistent with the fact that at later iterations the regularized algorithm proposes nearly empty image updates instead of adding noise (see Fig. 6), thus avoiding the increase in the reconstruction error that occurs in the unregularized case. Further, we note that the regularization-induced improvement in FIBIR is significantly larger than in CT. This confirms the hypothesis that thanks to its sharper but noisier gradients, GI-CT benefits more from regularization compared to classical CT, in which the gradients are less noisy but also less sharp. Finally, although conventional CT operates on data with twice the amount of photons, because of the absence of the G2 grating, the final MSE in regularized classical CT and in regularized FIBIR-enabled GI-CT is very similar.

The favourable quantitative findings of Fig. 7 are confirmed by the excellent qualitative results in Fig. 8 which plots the image iterates for the unregularized and regularized reconstruction of CT-equivalent and FIBIR-enabled GI-CT. For the

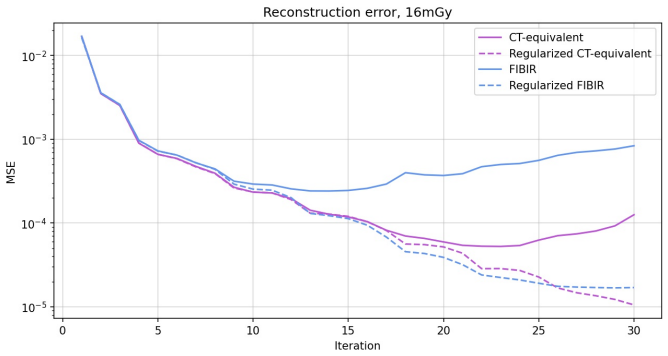

Fig. 7. Regularization effect on the reconstruction error. The reconstruction error is plotted as a function of the iteration. The regularized reconstructions are plotted as dashed lines, the unregularized counterparts as solid lines. The CT-equivalent reconstruction is displayed in purple, the FIBIR-enabled GI-CT in blue. The MSE is plotted in log-scale for clarity.

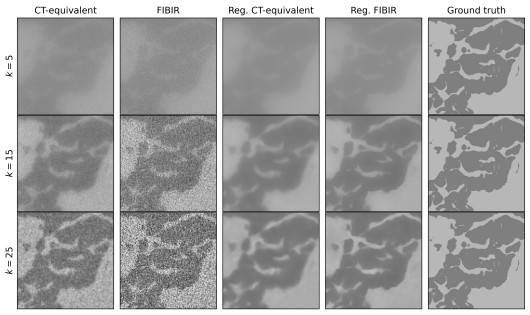

Fig. 8. Regularization effect during iterative reconstruction. From left to right: image iterates of unregularized CT-equivalent reconstruction, unregularized FIBIR-enabled GI-CT reconstruction, regularized CT-equivalent reconstruction, regularized FIBIR-enabled GI-CT reconstruction and ground truth. From top to bottom: different iterations for $\boldsymbol{k} = \mathbf{5, 15, 25}$. Zoomed-in sections of the full images are plotted with a FOV of of 15 × 15 mm and with a pixel size of 75 µm. Quantitative results for these images are shown in Fig. 7.

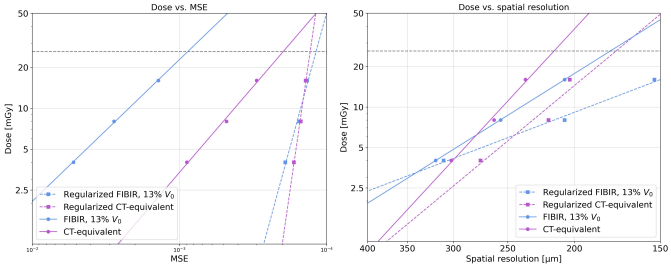

Fig. 9. Left panel: reconstruction errors at different radiation doses. Right panel: dose vs. resolution for FIBIR-enabled GI-CT and CT-equivalent, with and without regularization. FIBIR results are plotted in blue, CT-equivalent in purple. Regularized results are shown as dashed lines, unregularized results as solid lines. The dashed horizontal grey line displays the dose limit for today's breast CT [38].

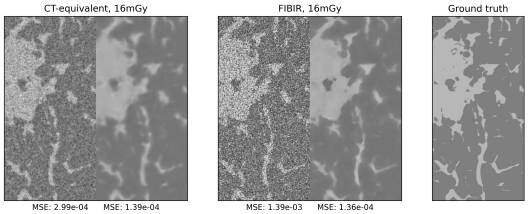

Fig. 10. Reconstruction at 16 mGy. From left to right: CT-equivalent, FIBIR and ground truth. In the left and central images, the left panel of the images is unregularized, the right panel is regularized. All results are shown for iteration $\boldsymbol{k} = \mathbf{30}$, i.e. the iteration closest to convergence as seen in Fig. 7. Zoomed-in sections of the full images are plotted.

unregularized case, FIBIR-enabled GI-CT yields noisier but sharper images compared to the CT-equivalent. The higher noise is caused by half the photon counts in the former case, whereas the sharper images are enabled by being sensitive not only to the Radon transform of the image but also to its 1-dimensional *derivative* [12]. The two regularized results reveal that the FIBIR-enabled GI-CT reconstruction is significantly sharper compared to the CT-equivalent image. In both cases the noise is removed practically entirely. Finally, as for the unregularized case, also *regularized* FIBIR allows to speed up convergence compared to the CT-equivalent counterpart.

Without regularization the FIBIR-enabled GI-CT reconstruction has a better spatial resolution, computed with Fourier ring correlation (FRC) [37], compared to the CT-equivalent counterpart (203 µm vs. 231 µm). With gradient-based regularization the resolution of the FIBIR reconstruction improves to 155 µm, while for the CT-equivalent reconstruction it improves to 200 µm. Gradient-based regularization thus allows to increase the spatial resolution in both cases, with the former achieving a slightly bigger improvement in this experiment.

We will now discuss general trends of GradReg across different radiation doses. The left panel in Fig. 9 shows the reconstruction errors of the CT-equivalent and FIBIR, with and without regularization for different doses, at $k_{\max}$. In the unregularized case GI-CT has significantly higher MSEs since it operates on half the number of counts. As expected, regularization significantly improves both results. In particular, *with* regularization both methods achieve similar results across different doses. This again shows that FIBIR benefits more from regularization. In the unregularized case the two lines have only slightly different slopes, which makes them cross only at extremely high doses, whereas with regularization they cross at a dose of 10 mGy. Consequently, for doses higher than 10 mGy, regularized FIBIR is expected to yield lower reconstruction errors compared to conventional CT despite the higher noise in the measured data.

Fig. 10 shows the reconstructed images which correspond to the quantitative results in Fig. 9 at 16 mGy. Unregularized results are plotted in the left panels of each image, whereas regularizated results in the right panels. If we first consider the unregularized results, we see again that the CT-equivalent reconstructions are affected by less noise due to the higher photon counts, which results in lower MSE values. Looking at the regularized reconstructions, we observe that the noise is strongly suppressed. The results are of excellent quality with high sharpness.

The right panel in Fig. 9 shows the spatial resolution of the reconstruction results at different doses measured with Fourier ring correlation. Note that here the resolutions were calculated directly on the reconstructed images. Therefore, contrary to [12], no smoothing was involved to keep the comparison to the unregularized case fair. First, we see that regularization

improves the resolution for both FIBIR and CT-equivalent across all investigated doses. Only at doses below 3.5 mGy our results suggest that for FIBIR this is not the case anymore because of prohibitively high noise amplitudes. Second, we observe that with regularization the fitted CT-equivalent and FIBIR lines cross at a comparable dose (around 6 mGy) but at a better spatial resolution (278 µm vs. 244 µm).

These results suggest that regularization allows to increase dose efficiency. In fact, if we want to achieve a resolution of e.g. 155 µm (the resolution FIBIR reached in the experiment at 16 mGy), without regularization we would need 156 mGy for conventional CT and 40 mGy for FIBIR-enabled GI-CT. With regularization this improves to 43 mGy and 15 mGy, respectively. GradReg thus allows to significantly decrease the dose necessary to achieve this particular spatial resolution, and enables FIBIR-based GI-CT to reach a clinically-compatible dose regime. Furthermore, the results in Fig. 9 argue that from 10 mGy upwards, regularized FIBIR-enabled GI-CT is superior to conventional CT both in terms of spatial resolution and in terms of reconstruction error for a visibility of 13%. While this dose is already within a range that is considered to be clinically acceptable [38], the break-even point between GI-CT and conventional CT is expected to decrease with increasing grating quality in the future [10].

To investigate the algorithm's effect on the spectral characteristics of the noise, in Fig. 11 we plot the noise power spectrum (NPS) for unregularized and regularized reconstructions for both conventional CT and FIBIR. In the unregularized case we confirm the presence of more high-frequency noise in FIBIR compared to CT, which is due to 1) GI-CT's sensitivity to a differential signal and 2) the presence of the G2 grating which kills half of the X-ray flux. Moreover, we see that regularization allows to significantly reduce the noise power across the entire studied frequency range. Like in previous experiments, we observe that the noise suppression is more significant in FIBIR compared to conventional CT. In our experiments the ground truth image had no texture and therefore the network learned to predict a piece-wise linear image. This is also visible in the NPS results since the regularized curves show a very smooth behaviour. Real samples however contain texture, which in fact can be critical for physicians in detecting suspicious lesions. Therefore, future work will have to investigate how the NPS is modified when the network is trained on real measurement data.

To confirm the promising results obtained on *in-silico* data we scanned a formalin-fixed breast specimen received at the Department for Pathology and Molecular Pathology at the University Hospital Zürich without any significant pathological findings. The specimen has been obtained with written informed consent under the ethical approval KEK- 2012_554 obtained from the Cantonal Ethics Commission of canton Zürich. The results in Fig. 12 show reconstructions of a fixed mastectomy scan, with and without regularization, in FIBIR and CT-equivalent, at a dose of 22 mGy. The results indicate that data-driven gradient regularization is able to produce stable reconstruction results also on real data, even though the network was trained solely on simulations. Specifically, we used the network trained on 16 mGy data for the reconstruction

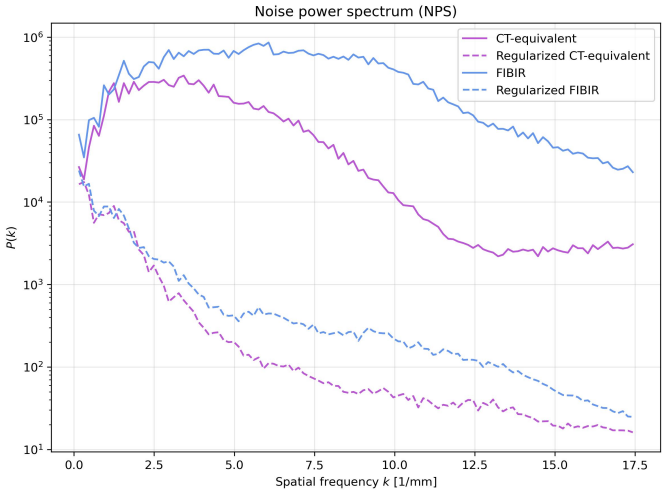

Fig. 11. Noise power spectrum of regularized and unregularized reconstructions for conventional CT and FIBIR. FIBIR results are plotted in blue, CT-equivalent in purple. Regularized results are shown as dashed lines, unregularized results as solid lines. The NPS was computed on a homogeneous region in the *in-silico* reconstructed images.

of this 22 mGy measurement data. This mismatch in the dose was necessary because the dose level in simulations was slightly inferior compared to the dose-equivalent of real measurements, probably due to non-simulated effects such as polychromaticity, a source size of roughly 400 µm which blurs the image, vibrations, etc. Fig. 12 shows that also on real data the regularized FIBIR images are of higher quality compared to the CT-equivalent ones, despite the higher noise in the former. Specifically, the signal-to-noise ratio (SNR) of the unregularized CT-equivalent and FIBIR was 4.79 and 5.07, whereas with regularization it became 88.44 vs. 195.54, confirming a larger improvement in FIBIR also in terms of SNR. Unfortunately, the MSE could not be computed as no clean ground truth image was available.

While the overall reconstruction performance is of high quality, we observe that some small details are missing. We attribute this shortcoming to the domain shift that occurs between training (on simulations) and testing (on real measurements). This limitation will be addressed once large numbers of samples will have been scanned.

Since no ground truth was available for this data, it is difficult to determine whether these details constitute real structures in the fixed mastectomy or hallucinated structures. Considering however that the same features are present both in the CT-equivalent as well as in FIBIR, it seems probable that the details indeed represent real structures in the mastectomy. In fact, the probability of two distinct regularization networks hallucinating the same features appears small.

We could, however, compute the spatial resolution of the reconstructions. As in the *in-silico* study regularization is able to increase the spatial resolution in both conventional CT (from 536 µm to 471 µm) and in FIBIR (from 453 µm to 380 µm), with the latter achieving the best metric. The resolution on real data is lower by a factor of roughly two compared to the resolution on *in-silico* data. This is likely due to polychromaticity, finite source size, scattering and other

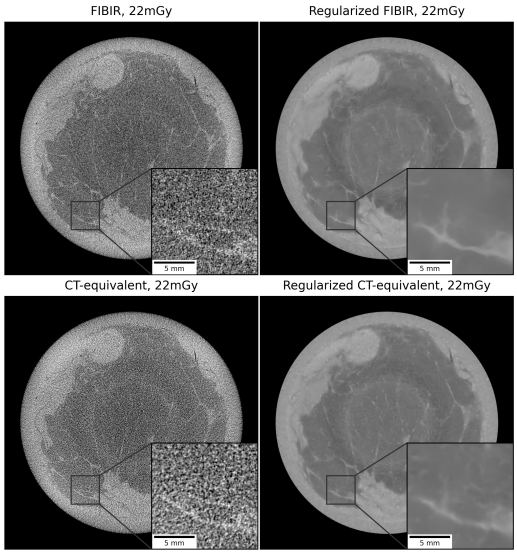

Fig. 12. Reconstruction results of fixed mastectomy scanned at a dose of 22 mGy. On the top: unregularized FIBIR and regularized FIBIR. On the bottom: unregularized CT-equivalent and regularized CT-equivalent. Ring-removal was applied after reconstruction [39].

physical phenomena which are not modelled in the forward operators of the reconstruction algorithm, and which were not simulated in the *in-silico* study.

IV. Conclusion

In this article we have proposed a new type of data-driven regularization strategy called GradReg which, instead of operating in the image space, acts on the image gradients. This method is more flexible than conventional Plug-and-Play algorithms that act in the image space since it can in principle be applied to any type of gradient-based optimization scheme, both to first-order methods as well as to quasi-Newton algorithms. GradReg in particular offers important advantages in the context of quasi-Newton solvers. In fact, in these cases the data updates are computed based on a history of past image iterates *and* image gradients. Consequently, combining non-denoised gradients and denoised image iterates, as it is done in classical image-space Plug-and-Play regularization, will result in an invalid secant equation and thus to noisy approximations of the inverse Hessian. On the contrary, our strategy ensures that both the image iterates and the gradients will be denoised at each iteration. The secant equation will thus be based solely on denoised data which yields better estimates of the inverse Hessian used to compute the next data update.

The motivation for developing this algorithm came from the observation that the fused intensity-based iterative reconstruction (FIBIR) algorithm in GI-CT yields gradients which are noisier but at the same time also sharper compared to the gradients obtained in classical iterative CT reconstruction. We thus hypothesized that especially the former could gain a large benefit from gradient-space denoising. This hypothesis was confirmed in the results in Figs. 7 - 10, which clearly demonstrate that the improvements in both reconstruction error and spatial resolution are larger in GI-CT compared to conventional CT.

Importantly, all experiments in this work were performed with a mean visibility of 13%. Since the visibility of a GI setup strongly influences the noise propagation in the acquired data [13], improvements in grating fabrication and consequently in visibility are expected to increase the reconstructed image quality in GI-CT, thereby reducing the break-even point in terms of radiation dose between the latter and conventional CT.

We did not have access to sufficient amounts of real data to train the data-driven regularizer and to validate the reconstruction performance of the proposed hybrid algorithm. Therefore, we trained all our models on simulated data and applied the hybrid reconstruction algorithm to both *in-silico* and to real data. When we applied the algorithm to real data there was thus a distribution mismatch between the training data, which was based on simulated data, and the testing data, which came from real measurements. Despite this mismatch, the results in Fig. 12 showed very promising results. We thus expect that training all models on real data will improve the performance on real measurements even further.

A potential weakness of data-driven denoising algorithms concerns their generalization to different scanned specimens and image acquisitions. Considering the variability in specimens, it will be critical to build a representative training set that contains samples which cover the whole space of mammographic tomograms. This will allow the network to effectively approximate the posterior distribution $p(g_{denoised,k}|g_k)$ across the entire relevant domain. Regarding different image acquisitions, it would be best to train one network for each image acquisition scheme since the noise and artifacts in the gradients strongly depend on factors such as photon count, angular sampling, pitch etc. Considering that our scanner will operate with a limited number of different acquisition schemes, training a small number of different models will not pose a problem. A possible strategy to introduce robustness to distribution shift could be test-time training/instance adaptation as suggested in [40].

To train the data-driven regularizer on real data, two strategies could be envisioned. If high-dose acquisitions are possible, e.g. in *ex-vivo* cases, high-quality target gradients could be obtained by scanning samples with high photon count statistics. If this is not possible, e.g. when radiation exposure must be minimized in *in-vivo* applications, the models could be trained in a unsupervised or self-supervised manner [29]–[32] without the need for high-quality targets. In any case, training the regularization networks on large amounts of real data will likely increase the reconstruction performance and reduce the chances of hallucinating or smoothing out features.

A strength of the proposed regularization method is that there is no need to fine tune the regularization strength as it is e.g. the case in variational optimization where the regularization function is usually weighted by a scalar λ. In fact, in GradReg the network is trained to remove exactly the correct amount of noise, which consequently determines the perfect regularization strength. The way the training data is generated, and in particular how much noise it contains, thus implicitly determines the regularization strength of the method.

Even though in this work we applied the presented method

to GI-CT and conventional CT reconstruction, we would like to highlight that GradReg is applicable to any image reconstruction task that relies on gradient-based methods. Also, the denoising engine could in principle be replaced by a classical denoiser such as the non-local means (NLM) algorithm [41] or block matching and 3D denoising (BM3D) algorithm [42]. In this work we used a data-driven denoiser because we wanted to investigate the feasibility of processing all iterates with a single data-driven denoising engine.

In this article we have not performed theoretical convergence analyses, nor did we study necessary conditions for convergence. Similarly to the recently published analysis for Plug-and-Play quasi-Newton method convergence in [43] future work should thus try to develop theoretical guarantees for this new type of reconstruction algorithms.

In conclusion, given the promising results obtained in this article, the proposed hybrid reconstruction algorithm with data-driven gradient denoising could become an important tool in the field of GI-CT reconstruction and beyond.

REFERENCES

[1] S.-A. Zhou and A. Brahme, "Development of phase-contrast x-ray imaging techniques and potential medical applications," *Physica Medica*, vol. 24, no. 3, pp. 129–148, 2008.

[2] U. Bonse and M. Hart, "An x-ray interferometer," *Applied Physics Letters*, vol. 6, no. 8, pp. 155–156, 1965.

[3] A. Snigirev, I. Snigireva, V. Kohn, S. Kuznetsov, and I. Schelokov, "On the possibilities of x-ray phase contrast microimaging by coherent high-energy synchrotron radiation," *Review of scientific instruments*, vol. 66, no. 12, pp. 5486–5492, 1995.

[4] T. Davis and A. Stevenson, "Direct measure of the phase shift of an x-ray beam," *JOSA A*, vol. 13, no. 6, pp. 1193–1198, 1996.

[5] A. Momose, S. Kawamoto, I. Koyama, Y. Hamaishi, K. Takai, and Y. Suzuki, "Demonstration of x-ray talbot interferometry," *Japanese journal of applied physics*, vol. 42, no. 7B, p. L866, 2003.

[6] T. Weitkamp, A. Diaz, C. David, F. Pfeiffer, M. Stampanoni, P. Cloetens, and E. Ziegler, "X-ray phase imaging with a grating interferometer," *Optics express*, vol. 13, no. 16, pp. 6296–6304, 2005.

[7] F. Pfeiffer, T. Weitkamp, O. Bunk, and C. David, "Phase retrieval and differential phase-contrast imaging with low-brilliance x-ray sources," *Nature physics*, vol. 2, no. 4, pp. 258–261, 2006.

[8] M. Stampanoni, Z. Wang, T. Thüring, C. David, E. Roessl, M. Trippel, R. A. Kubik-Huch, G. Singer, M. K. Hohl, and N. Hauser, "The first analysis and clinical evaluation of native breast tissue using differential phase-contrast mammography," *Investigative radiology*, vol. 46, no. 12, pp. 801–806, 2011.

[9] Z. Wang, N. Hauser, G. Singer, M. Trippel, R. A. Kubik-Huch, C. W. Schneider, and M. Stampanoni, "Non-invasive classification of microcalcifications with phase-contrast x-ray mammography," *Nature communications*, vol. 5, no. 1, p. 3797, 2014.

[10] M. Rawlik, A. Pereira, S. Spindler, Z. Wang, L. Romano, K. Jefimovs, Z. Shi, M. Polikarpov, J. Xu, M.-C. Zdora, *et al.*, "Refraction beats attenuation in breast ct," *arXiv preprint arXiv:2301.00455*, 2023.

[11] S. van Gogh, S. Mukherjee, J. Xu, Z. Wang, M. Rawlik, Z. Varga, R. Alaifari, C.-B. Schönlieb, and M. Stampanoni, "Iterative phase contrast ct reconstruction with novel tomographic operator and data-driven prior," *Plos one*, vol. 17, no. 9, p. e0272963, 2022.

[12] S. Van Gogh, M. Rawlik, A. Pereira, S. Spindler, S. Mukherjee, M.-C. Zdora, M. Stauber, R. Alaifari, Z. Varga, and M. Stampanoni, "Towards clinical-dose grating interferometry breast ct with fused intensity-based iterative reconstruction," *Optics Express*, vol. 31, no. 5, pp. 9052–9071, 2023.

[13] V. Revol, C. Kottler, R. Kaufmann, U. Straumann, and C. Urban, "Noise analysis of grating-based x-ray differential phase contrast imaging," *Review of Scientific Instruments*, vol. 81, no. 7, p. 073709, 2010.

[14] R. Raupach and T. G. Flohr, "Analytical evaluation of the signal and noise propagation in x-ray differential phase-contrast computed tomography," *Physics in Medicine & Biology*, vol. 56, no. 7, p. 2219, 2011.

[15] B. Brendel, M. von Teuffenbach, P. B. Noël, F. Pfeiffer, and T. Koehler, "Penalized maximum likelihood reconstruction for x-ray differential phase-contrast tomography," *Medical physics*, vol. 43, no. 1, pp. 188–194, 2016.

[16] M. v. Teuffenbach, T. Koehler, A. Fehringer, M. Viermetz, B. Brendel, J. Herzen, R. Proksa, E. J. Rummeny, F. Pfeiffer, and P. B. Noël, "Grating-based phase-contrast and dark-field computed tomography: a single-shot method," *Scientific reports*, vol. 7, no. 1, p. 7476, 2017.

[17] J. Nocedal, "Updating quasi-newton matrices with limited storage," *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.

[18] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[20] S. Ravishankar and Y. Bresler, "Learning doubly sparse transforms for images," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4598–4612, 2013.

[21] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948, IEEE, 2013.

[22] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[23] R. Cohen, M. Elad, and P. Milanfar, "Regularization by denoising via fixed-point projection (red-pro)," *SIAM Journal on Imaging Sciences*, vol. 14, no. 3, pp. 1374–1406, 2021.

[24] J. Hertrich, S. Neumayer, and G. Steidl, "Convolutional proximal neural networks and plug-and-play algorithms," *Linear Algebra and its Applications*, vol. 631, pp. 203–234, 2021.

[25] P. E. Gill and W. Murray, "Quasi-newton methods for unconstrained optimization," *IMA Journal of Applied Mathematics*, vol. 9, no. 1, pp. 91–108, 1972.

[26] J. E. Dennis, Jr and J. J. Moré, "Quasi-newton methods, motivation and theory," *SIAM review*, vol. 19, no. 1, pp. 46–89, 1977.

[27] S. van Gogh, Z. Wang, M. Rawlik, C. Etmann, S. Mukherjee, C.-B. Schönlieb, F. Angst, A. Boss, and M. Stampanoni, "Insidenet: Interpretable nonexpansive data-efficient network for denoising in grating interferometry breast ct," *Medical physics*, vol. 49, no. 6, pp. 3729–3748, 2022.

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

[29] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.

[30] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," in *International Conference on Machine Learning*, pp. 524–533, PMLR, 2019.

[31] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2129–2137, 2019.

[32] A. Krull, T. Vičar, M. Prakash, M. Lalit, and F. Jug, "Probabilistic noise2void: Unsupervised content-aware denoising," *Frontiers in Computer Science*, vol. 2, p. 5, 2020.

[33] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] T. Köhler, B. Brendel, and E. Roessl, "Iterative reconstruction for differential phase contrast imaging using spherically symmetric basis functions," *Medical physics*, vol. 38, no. 8, pp. 4542–4545, 2011.

[36] R.-D. Bippus, T. Köhler, F. Bergner, B. Brendel, E. Hansis, and R. Proksa, "Projector and backprojector for iterative ct reconstruction with blobs using cuda," in *Fully 3D 2011: 11th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, Potsdam, Germany, 11-15 July 2011*, Citeseer, 2011.

[37] R. P. Nieuwenhuizen, K. A. Lidke, M. Bates, D. L. Puig, D. Grünwald, S. Stallinga, and B. Rieger, "Measuring image resolution in optical nanoscopy," *Nature methods*, vol. 10, no. 6, pp. 557–562, 2013.

[38] Y. Zhu, A. M. O'Connell, Y. Ma, A. Liu, H. Li, Y. Zhang, X. Zhang, and Z. Ye, "Dedicated breast ct: State of the art—part ii. clinical application and future outlook," *European radiology*, vol. 32, no. 4, pp. 2286–2300, 2022.

[39] D. Gürsoy, F. De Carlo, X. Xiao, and C. Jacobsen, "Tomopy: a framework for the analysis of synchrotron tomographic data," *Journal of synchrotron radiation*, vol. 21, no. 5, pp. 1188–1193, 2014.

[40] M. Z. Darestani, J. Liu, and R. 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.

[41] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2, pp. 60–65, Ieee, 2005.

[42] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[43] H. Y. Tan, S. Mukherjee, J. Tang, and C.-B. Schönlieb, "Provably convergent plug-and-play quasi-newton methods," *arXiv preprint arXiv:2303.07271*, 2023.