Received: 9 August 2021 | Revised: 25 October 2021 | Accepted: 7 January 2022

DOI: 10.1002/mp.15595

RESEARCH ARTICLE

MEDICAL PHYSICS

INSIDEnet: Interpretable NonexpanSIve Data-Efficient network for denoising in grating interferometry breast CT

Stefano van Gogh[1,2] | **Zhentian Wang**[3,4] | **Michał Rawlik**[1,2] | **Christian Etmann**[5] | **Subhadip Mukherjee**[5] | **Carola-Bibiane Schönlieb**[5] | **Florian Angst**[6] | **Andreas Boss**[6] | **Marco Stampanoni**[1,2]

[1]Photon Science Division, X-Ray Tomography Group, Paul Scherrer Institute, Villigen PSI, Switzerland

[2]Department for Electrical Engineering and Information Technology, X-Ray Tomography Group, ETH Zürich, Zürich, Switzerland

[3]Department of Engineering Physics, Tsinghua University, Haidian District, Beijing, China

[4]Key Laboratory of Particle & Radiation Imaging (Tsinghua University) of Ministry of Education, Haidian District, Beijing, China

[5]Cambridge Image Analysis Group, Centre for Mathematical Sciences, University of Cambridge, Cambridge, United Kingdom

[6]Institute for Diagnostic and Interventional Radiology, University Hospital Zürich, Zürich, Switzerland

Correspondence
Stefano van Gogh, Paul Scherrer Institute, Photon Science Division, X-ray Tomography Group, Forschungsstrasse 111, 5232 Villigen PSI, Switzerland.
Email: stefano.van-gogh@psi.ch

Funding information
Promedica Stiftung; ETH-Research Commission, Grant/Award Number: ETH-12 20-2; SNF, Grant/Award Number: CRSII5 183568; Swisslos Lottery Fund of Kanton Aargau

Abstract

Purpose: Breast cancer is the most common malignancy in women. Unfortunately, current breast imaging techniques all suffer from certain limitations: they are either not fully three dimensional, have an insufficient resolution or low soft-tissue contrast. Grating interferometry breast computed tomography (GI-BCT) is a promising X-ray phase contrast modality that could overcome these limitations by offering high soft-tissue contrast and excellent three-dimensional resolution. To enable the transition of this technology to clinical practice, dedicated data-processing algorithms must be developed in order to effectively retrieve the signals of interest from the measured raw data.
Methods: This article proposes a novel denoising algorithm that can cope with the high-noise amplitudes and heteroscedasticity which arise in GI-BCT when operated in a low-dose regime to effectively regularize the ill-conditioned GI-BCT inverse problem. We present a data-driven algorithm called INSIDEnet, which combines different ideas such as multiscale image processing, transform-domain filtering, transform learning, and explicit orthogonality to build an Interpretable NonexpanSIve Data-Efficient network (INSIDEnet).
Results: We apply the method to simulated breast phantom datasets and to real data acquired on a GI-BCT prototype and show that the proposed algorithm outperforms traditional state-of-the-art filters and is competitive with deep neural networks. The strong inductive bias given by the proposed model's architecture allows to reliably train the algorithm with very limited data while providing high model interpretability, thus offering a great advantage over classical convolutional neural networks (CNNs).
Conclusions: The proposed INSIDEnet is highly data-efficient, interpretable, and outperforms state-of-the-art CNNs when trained on very limited training data. We expect the proposed method to become an important tool as part of a dedicated plug-and-play GI-BCT reconstruction framework, needed to translate this promising technology to the clinics.

KEYWORDS
image denoising, interpretable machine learning, breast CT

1 INTRODUCTION

Breast cancer is the most prevalent malignancy in women with one out of eight developing the disease in her lifetime.[1] To fight this public health burden, recent years have witnessed the introduction of many screening programs to early detect and consequently to better treat this disease.[2] Unfortunately, both false positive and false negative rates remain high, leading to unnecessary psychological distress and missed tumors, respectively.[2] The reason for this is attributable to the fact that current breast imaging techniques, most notably mammography, ultrasound, tomosynthesis, breast MRI, and absorption-based breast CT,[3,4] all suffer from some limitations. In fact, none of these techniques simultaneously yields fully three-dimensional (3D) data with sufficient soft-tissue contrast and spatial resolution to detect crucial imaging biomarkers (small soft tissue lesions and their margins, microcalcifications, architectural distortions, and tiny soft tissue density differences),[5] thereby making it difficult for radiologists to take confident decisions.

Consequently, there has been an ever-increasing effort to utilize X-ray phase contrast imaging which can potentially lead to higher soft-tissue contrast compared to absorption-based imaging, without sacrificing spatial resolution.[6] In fact, with synchrotron sources and in laboratory applications phase contrast already delivers far superior soft-tissue delineation.[7,8] X-ray grating interferometry (GI) is a phase contrast technique holding most of the prerequisites for clinical compatibility.[9] For this reason, our group has designed and is currently building a grating interferometry breast computed tomography (GI-BCT) prototype, that is, translating the technology into a first-of-its-kind compact medical device.

Obtaining high-quality phase-contrast images in clinically compatible settings remains a challenge. The theoretically achievable higher contrast in phase imaging compared to absorption cannot yet be fully exploited on compact X-ray sources because of the high-noise amplitudes and noise distributions.[10] Phase contrast images are in fact characterized by intrinsic low-frequency noise, especially in the low photon-count case.[11] Furthermore, those very same images contain a nonuniform noise distribution due to imperfect gratings. Finally, adhering to the severe constraints imposed by the clinical environment (such as radiation dose, scanning time, and patient comfort) demands for novel solutions to handle sparse sampling and photon starvation, thereby making it even more cumbersome to achieve high image quality.

It is important to mention here that this problem is specific to the phase contrast channel. In fact, in the absorption image the noise is predominantly present at the high frequencies, making it much easier to effectively denoise these data and for which a variety of established denoising methods[12,13] work well.

A powerful denoising algorithm is thus necessary to suppress the noise and let the higher intrinsic phase contrast emerge. In particular, a pipeline is needed to cope with the high-noise amplitudes and heteroscedasticity, while simultaneously offering high algorithm interpretability, reliability, and robustness.

The two major groups of denoising algorithms, which exist today, namely traditional methods and deep learning methods, have both significant limitations. Algorithms in the first category rely on hand-crafted priors[12,13] such as nonlocal similarity, low rankness, small gradient norms, nonnegative values, and sparsity in some transform domains, such as the wavelet domain. They form reliable pipelines which work extremely well on images with relatively little noise, but their performance drastically deteriorates when dealing with lower image quality, as they are unable to adapt to the data at hand. In contrast, deep learning methods have been shown to yield impressive results[14–16] by implicitly learning a prior in a data-driven fashion without the need of human intervention. A big downside of deep learning–based algorithms, however, is their high sensitivity to training data and their limited interpretability caused by the concatenation of (many) linear and nonlinear operations. It is well known that deep network–based denoisers might add structures which are not present and remove the ones which are present in the ground-truth data.[17] Such behavior is unacceptable in particular when it comes to denoising in medical applications. A third alternative is to combine the structure of handcrafted regularization with data-driven learning, thereby leveraging the power of data while maintaining interpretability.

To date, to the best of our knowledge, a single article has been published on the use of a data-driven method for denoising of GI, and in particular differential phase contrast (DPC), projection data.[18] However, the authors used a black-box model and restricted their analysis to radiography. In contrast, we focus on GI phase CT and propose a hybrid denoising algorithm, which we call "Interpretable NonexpanSIve Data-Efficient network" (INSIDEnet), that attempts to leverage the strengths of both worlds: the interpretability of classical filters and the flexibility of data-driven models. Importantly, the model has been parameterized to maximize interpretability and reliability, which are both imperative conditions for clinical applicability.

In this article, we demonstrate the performance of our hybrid algorithm on simulated breast phantoms and real data acquired on our GI-BCT prototype. Our INSIDEnet achieves better results compared to traditional, non-learning-based models, without paying in robustness. The INSIDEnet shows that it is possible to achieve an excellent denoising performance with superior interpretability and robustness compared to deep neural networks.

2 | MATERIALS AND METHODS

2.1 | Grating interferometry breast CT

Conventional X-ray imaging is based on photon absorption in the imaged tissues. Unfortunately, biological soft tissues have very similar attenuation coefficients as they are all mainly composed of carbon, oxygen, and water.[6] Therefore, there is limited contrast between different body constituents. For phase, in the absence of noise, the theoretically achievable contrast is higher because the real part of the index of refraction δ (related to phase) is orders of magnitude larger than the imaginary part β (related to absorption).[6] It still has to be explored though if such superior contrast can also be achieved in a clinically compatible setting. Contrary to attenuation-based imaging, which directly measures the intensity of the transmitted X-ray beam, it is not possible to directly detect the phase in a polychromatic setting, which therefore has to be measured indirectly. Several techniques have been proposed in this regard with the most notable ones being propagation based,[19] crystal interferometry,[20] analyzer based,[21] edge-illumination,[22] and GI.[9,23,24] While all methods can be applied at synchrotron light sources, GI has received special attention as it satisfies the prerequisites for clinical applicability: it has high mechanical robustness, can be scaled up to large fields of view (FOV) and only requires moderate spatial coherence and monochromaticity.[9]

GI encodes propagation-induced phase changes in the beam wavefront—when passing through a specimen—into an intensity modulation measured by a detector placed downstream. Its simplest configuration, Talbot interferometry, consists of two gratings placed in a partially coherent beam. The latter is usually provided by a third/fourth-generation synchrotron source or, with significantly less intensity, by a microfocus X-ray tube. The first grating is usually a phase grating, that is, it does not absorb the beam but imposes a phase-shift resulting in a controlled wavefront modulation at a specific distance downstream,[25] usually where the second, absorbing analyzer grating is placed. When the source does not provide a sufficiently high spatial coherence, like in the case of a conventional X-ray tube, a third grating can be introduced right after the source yielding to the so-called Talbot–Lau configuration,[23] as shown in Figure 1.

The intensity modulation of the resulting fringe pattern is characterized by its visibility, while retrieval of the absorption, phase, and dark-field signals can be done with various methods with phase stepping[26] and fringe scanning[27,28] being the most commonly used approaches. By combining the interferograms obtained with and without sample, one can retrieve the absorption signal, the differential phase signal, which is proportional to the larger-than-pixel-size refraction, and the dark-field signal, which is proportional to the incoherent refraction on a scale smaller than the pixel size. Finally, by rotating the sample or the X-ray source and detector, GI naturally extends to GI-CT.

High-quality phase-contrast tomograms of breast tissue have been demonstrated with GI on small FOV laboratory setups[7] and with propagation-based, large FOV imaging on a synchrotron.[8] In the first case, good image quality is obtained with long scanning times. The second case benefits from high sensitivity, thanks to a large propagation distance, and near-perfect coherent source. A clinically compatible device must on the other hand handle sparse sampling and photon starvation. As known from absorption-based X-ray imaging, low photon counts lead to lower signal-to-noise ratio (SNR) and thus lower image quality. Unfortunately, in GI this effect is even more detrimental. First, the characteristic imaging scheme based on the acquisition of an interferogram in GI leads to an amplification of the counting noise, with each of the three channels having a different noise propagation.[29] Second, noise amplification is not uniform across the FOV as, in fact, grating fabrication defects and grating misalignment cause strong local noise amplitude variations. Finally, during reconstruction, correlations between pixels are introduced and, crucially, integration of DPC leads to an amplification of low-frequency noise,[11] making the noise pattern even more difficult to deal with.

We would like to stress that since a clinically compatible one-shot acquisition method[30] does not allow to explicitly retrieve the sinograms for the three signals, the proposed INSIDEnet is ultimately envisioned to act as a denoising prior (or proximal operator) in a plug-and-play framework[31] by iteratively denoising image iterates within a gradient-based optimization scheme. This approach will be presented elsewhere and will not be further discussed in this article, as we would like to focus here on the introduction of the denoising engine itself.

2.2 | Simulated breast phantoms

2.2.1 | Clean breast phantoms

We generated 30 in silico breast phantoms of 44 × 1536 × 1536 voxels with a voxel size of 100 µm (real phantom size of 0.44 × 15.36 × 15.36 cm) containing three main tissue types, namely adipose, glandular, and skin.

To start with, we randomly generated 10 000 ellipsoids of different sizes, shapes, positions, and orientations, followed by two thresholds. We then multiplied this preliminary phantom with the central part of two randomly rotated binary masks (one for the whole breast and one for the skin) obtained from a 3D mask of a real breast that has been acquired with a breast CT scanner at the University Hospital Zürich. These data were part of a retrospective analysis of patient data approved by the local ethics committee. All patients gave their writ-

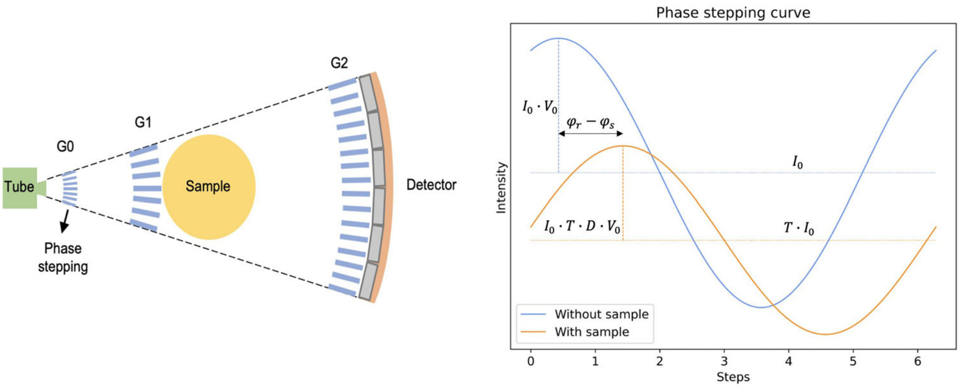

FIGURE 1 Left: Schematic drawing of GI setup in Talbot–Lau configuration. Right: phase stepping curves, with (orange) and without a sample (blue). The logarithm of the ratio of the average of the blue and orange curve gives the absorption signal, the reduction in its amplitude gives the dark-field signal, whereas the relative shift of the curve with respect to its reference gives the differential phase signal

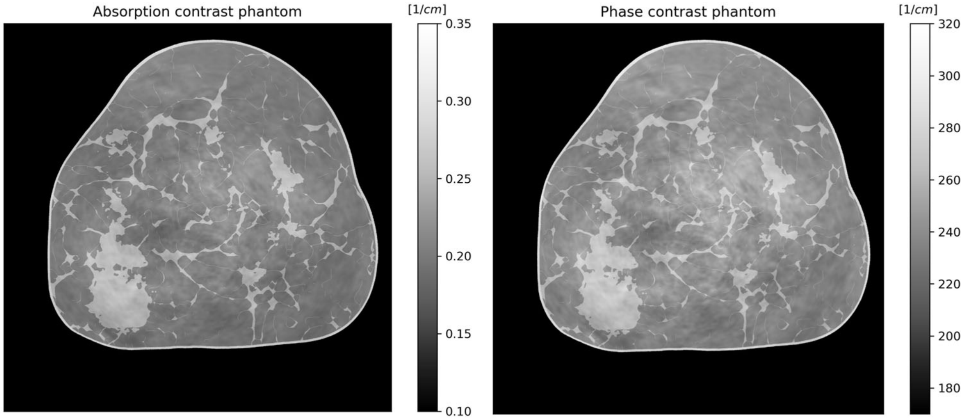

FIGURE 2 Absorption and phase contrast breast phantoms. The units for absorption are attenuation coefficients, while for phase it is phase shift coefficients

ten informed consent. The edges of the ellipsoids were used to simulate duct-like structures radiating out of the glandular tissue.

Realistic attenuation coefficients $\mu\,[\mathrm{cm}^{-1}]$ and phase shift coefficients $\phi\,[\mathrm{cm}^{-1}]$ for the absorption and the phase image, respectively, were assigned to the different regions representing adipose tissue, glandular tissue, and skin. Phase shift values were calculated starting from decrements in the real part of the index of refraction δ and using the known relation $\phi = 2\pi\delta/\lambda$,[10] where λ is the X-ray wavelength corresponding to the design energy of our prototype. δ and β values have been calculated using NIST XCOM[32] and f1f2 Kissel.dat of the DABAX library,[33] respectively, based on the tissue definition in ICRU 46.[34] Finally, we added anatomical noise to model more realistic slight β and δ inhomogeneities in the tissues as well as breast lesions of different sizes and contrasts to investigate the algorithms' performance in terms of lesion detectability. One slice of a phantom pair for absorption and phase is shown in Figure 2. It is important to mention here that these values represent an ideal setting. In reality, it will not be possible to reconstruct these μ and δ values precisely due to X-ray polychromaticity, Compton scattering, and limited phase sensitivity.

We would like to highlight that in this simulation, apart from differing contrasts between tissues, phase, and absorption images contain exactly the same structures, that is, the same information. While research into X-ray phase contrast imaging is evidently being carried out in the hope that extra diagnostic information becomes available, to date we do not know how such extra information would look like at a macroscopic scale in mammary tissue at clinically compatible doses. We are therefore not able to realistically simulate such structures. Since absorption and phase contrast data will be processed independently in this paper, the lack of extra information in phase contrast images will not affect the validity of the proposed method.

2.2.2 | Noisy breast phantoms

Each of the simulated breast phantom pairs was then used to generate noisy counterparts. First, we used the ASTRA toolbox projector[35] to obtain differential phase and transmission sinograms as follows[10]:

$$\varphi_s = \frac{\lambda d_2}{g_2} \frac{\partial}{\partial x} \int \phi(x, y, z) \mathrm{d}z, \tag{1}$$

$$T = \exp\left[- \int \mu(x, y, z) \mathrm{d}z \right], \tag{2}$$

where d_2 is the sample-G2 distance, g_2 is the pitch of G2. The sample-G2 distance was 73 cm, the pitch of G2 was 4.2 µm, and the system's design energy was 46 keV. The source-to-sample distance was 103 cm, and the sample-to-detector distance was 74 cm. More detailed information about our prototype is provided in Section 2.3.

Flat-field data (intensity map I_0, visibility map V_0, and phase map φ_r) were obtained from our scanner to provide simulations as realistic as possible. Due to grating imperfections, I_0, V_0 φ_r were all highly inhomogeneous, as shown in Figure 3. Combining these data, we then simulated phase-stepping curves with the sample in place:

$$I_{s,k} = I_0 T \cdot [1 + V_0 D \cdot \cos(k + \varphi_r - \varphi_s)]. \tag{3}$$

Here, k is the kth phase step, uniformly spaced between 0 and 2π and $I_{s,k}$ is the intensity value measured at the kth phase step. We neglected visibility reduction in the sample, that is, $D = 1$, as no significant small-angle scattering is to be expected in breast tissue apart from microcalcifications, which, however, were not simulated here.

Likewise, the background phase stepping curves were simulated as follows:

$$I_{r,k} = I_0 \cdot [1 + V_0 \cdot \cos(k + \varphi_r)]. \tag{4}$$

We then simulated detector quantum noise by sampling from the Poisson distribution with mean $I_{s,k}$. We empirically determined the necessary photons to match the image quality of our real data. We thus simulated 40 000 photons leaving the source at every exposure, which resulted in the flat-field data displayed in the first row of Figure 3.

We would like to mention here that this corresponds to a much higher dose than what is allowed in clinical practice. Such a higher dose allows to compensate for the yet insufficient grating quality which severely impacts the visibility. Once grating quality, and especially flat-field visibility, will improve, the same image quality will be achievable with significantly less dose.

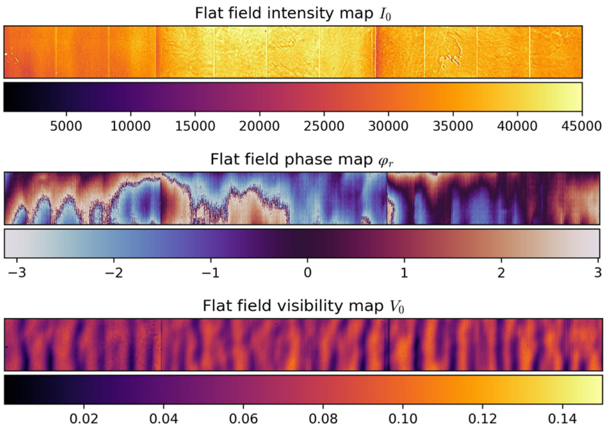

FIGURE 3 Flat-field data of our GI-BCT prototype: intensity map I_0, phase map φ_r, and visibility map V_0

By combining the two-phase stepping curves $I_{s,k}$ and $I_{r,k}$, we then retrieved both the differential phase and the attenuation signal with simple Fourier analysis. Finally, absorption and phase images were obtained by reconstructing the retrieved signals with analytical reconstruction algorithms available in the ASTRA toolbox.[35] Importantly, the Hilbert filter had to be applied for reconstructing the phase contrast image.[36]

All denoising algorithms were deployed on two-dimensional (2D) slices of the reconstructed data. While the algorithm could be easily extended to work on 3D data, we did not pursue this because of GPU memory constraints.

While denoising could also have been performed in the projection domain, we found that due to strong local noise amplitudes in the measurements, this yielded far inferior results than denoising the reconstructions in which these strong local noise amplitudes have previously been attenuated by backprojection. Moreover, since the algorithm is envisioned to be used as a proximal operator in the image space, we wanted to test its ability to denoise the latter rather than the sinogram data.

10 volumes were used for training, 10 for validation during training, and 10 for testing. All reported metrics in the simulations study have been obtained on the 10 testing volumes.

2.3 | Real data

To demonstrate the performance of our algorithm on real data we applied our algorithm trained on simulated data to a tomogram acquired on our GI-BCT prototype of a chunk of meat with roughly the same volume as a human breast.

We used a fifth Talbot-order symmetric interferometer with 4.2 µm pitch, with a single G0, single G1, and three G2 gratings. All gratings are bent. As X-ray source we

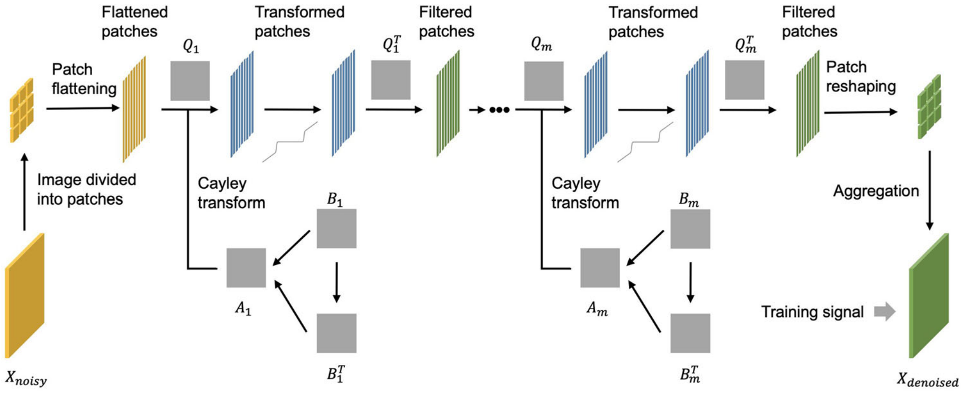

FIGURE 4 Denoising pipeline applied at each image scale. The trainable parameters are displayed in light gray. The learning signal enters the pipeline after the aggregation step. In yellow: noisy data, in blue: transformed data, and in green: denoised data

used a Comet MXR-225HP/11 tube operated at 70 kVp and 10 mA and as a detector a CdTe photon counting Dectris prototype with 75 µm pixel size operated at 10 Hz. We acquired five scans in continuous rotation, each consisting of 600 projections and then averaged the tomograms to obtain the input for our denoising algorithm. This averaging of five scans has been performed to compensate for the (yet) insufficient flat field visibility of our scanner.

2.4 | INSIDEnet: Interpretable NonexpanSIve Data-Efficient network

The proposed method combines different image-processing paradigms, namely transform domain thresholding, transform learning, multiscale processing, explicit orthogonality (thus nonexpansiveness), and deep learning.

With the assumption that the signal can be expressed as a linear combination of few basis elements, transform domain thresholding transforms image patches into another domain such as the wavelet domain or the discrete cosine transform (DCT) domain.[12] Owing to the high levels of cross-correlation within patches, this representation of the data will be highly sparse.[12] Filtering can be performed by thresholding or shrinking the coefficients. Finally, the filtered patches are transformed back to the image domain and the patches are aggregated to form the final denoised image.

Data-adaptive bases have been shown to be superior compared to handcrafted ones. Transform learning in particular builds upon the idea that it might be beneficial to replace a fixed hand-designed operator such as the wavelet transform or the DCT with an operator adapted to the data at hand. It thus aims to learn a transform matrix which moves the images to a space in which they have a highly sparse representation.[37] This transform operator can then be used to regularize ill-posed inverse problems like denoising.

We propose to combine these two ideas with deep learning and multiscale image processing to efficiently denoise highly corrupted images. Crucially, an explicit orthogonality constraint[38] has been used to achieve high model robustness and interpretability. To exclude channel cross-talk and be able to inspect the quality of the two signals of interest independently, we applied the INSIDEnet to absorption and phase data separately. It should be noted though that the proposed method can be applied to multiple channels if correlations between the two signals should be leveraged.

The fixed transform is replaced with learnable matrices, and the images are processed across four different scales. This ensures that our model can remove the noise across a broad frequency band. The entire pipeline is end-to-end trainable, thus allowing to not only learn the transforms but also the thresholds, thereby leaving very few tunable hyperparameters.

In the simulation study, the models have been trained in a supervised manner to map noisy images to their clean counterparts. For quantitative and qualitative evaluation, the trained models have been applied to simulated testing data and real data, respectively.

We will first introduce the stacked orthogonal transform learning denoising pipeline which is applied across each image scale (see Figure 4), before explaining how the multiple scales are generated and combined (see Figure 5).

2.4.1 | Image preprocessing

Before entering the denoising pipeline, the data are scaled to be within [0,1]. The same scaling is then applied to their noisy counterparts. During training, the images are randomly shuffled to ensure unbiased learning.

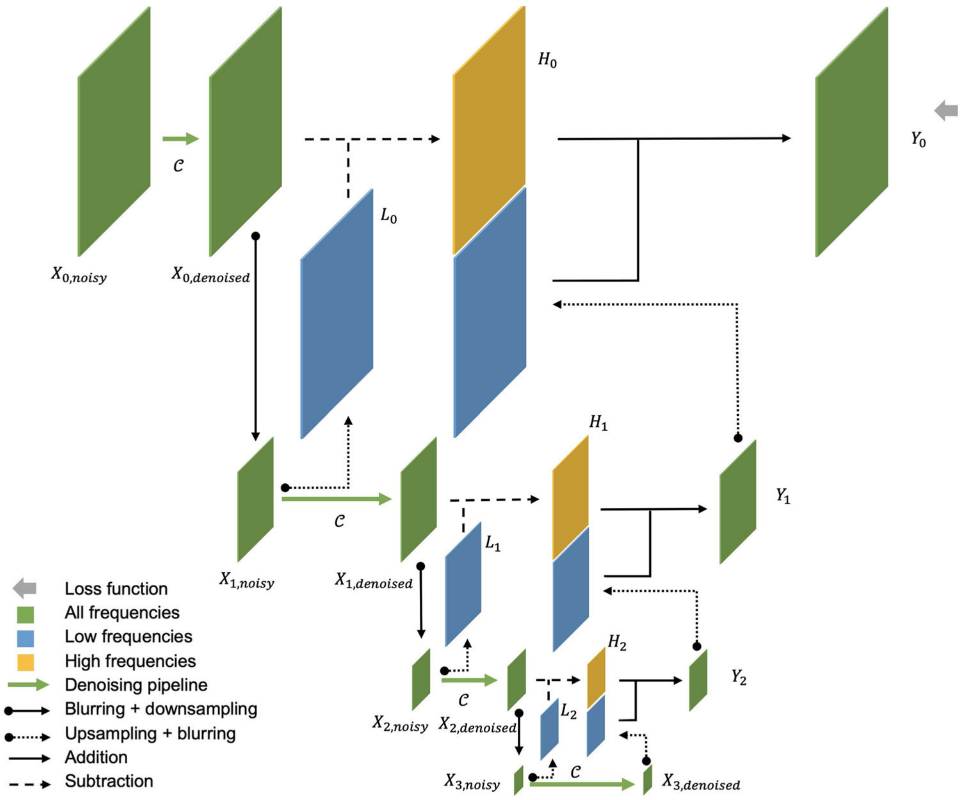

FIGURE 5 INSIDEnet overview. Image decomposition: the noisy input image gets first denoised at full resolution, with the aim to remove the highest frequencies. Next, downsampling leads to a lower resolution image with still all lower frequency noise present. Applying the denoising pipeline again removes the noise corresponding to high frequencies at this scale. This process is repeated three times, effectively removing noise at both high and low frequencies. Image reconstruction: high-frequency components are obtained by subtracting the low frequencies at each scale. These denoised high-frequency components are then iteratively added to the lowest frequency components in a Laplacian pyramid fashion

2.4.2 | Stacked orthogonal transform learning

Let $X_{\text{noisy}} \in \mathbb{R}^{N_X \times N_X}$ be the noisy input image and $X_{\text{target}} \in \mathbb{R}^{N_X \times N_X}$ the corresponding target image.

X_{noisy} is divided into overlapping patches (stride 2 was empirically determined to be optimal), here represented by tensor $P \in \mathbb{R}^{(4N_X/N_P-1)\times(4N_X/N_P-1)\times N_P \times N_P}$, where N_P is the patch size in horizontal and vertical directions. Next, the patches are flattened, thereby effectively reshaping P, which now is in $\mathbb{R}^{(4N_X/N_P-1)\times(4N_X/N_P-1)\times N_P^2}$. This tensor gets then multiplied through an Einstein sum with the orthogonal transform matrix $Q \in \mathbb{R}^{N_P^2 \times N_P^2}$ yielding the transformed image patches $\hat{P} \in \mathbb{R}^{(4N_X/N_P-1)\times(4N_X/N_P-1)\times N_P^2}$

$$\hat{P} = QP. \tag{5}$$

We enforce Q to be orthogonal by employing the Cayley transform.[38,39] An orthogonal matrix can in fact be obtained by computing $Q = (A - I)(A + I)^{-1}$, where A is a skew-symmetric matrix which can in turn be obtained by an arbitrary matrix B:

$$A = B - B^{\top}. \tag{6}$$

We thus let B be our trainable matrix, with random initialization, which is explicitly transformed into an orthogonal matrix.

The filtering step itself is very simple and draws inspiration from the proximal operator of the l_0 norm, that is, a hard threshold on the coefficient magnitudes. To allow this threshold to be trainable, we approximate the hard threshold with a steep sigmoid function:

$$T = \frac{1}{1 + \exp\left(-(|\hat{P}| - \Gamma)\cdot \mu\right)}, \tag{7}$$

where $\mu = 100$ is the of the sigmoid and is not trainable. It was selected by visually plotting the resulting activation function and exploiting the fact that $\hat{P}$ is approximately in the same range (0–1) as P. Γ, which defines the threshold, was initialized to 10^{-6} at all scales to ensure that at the start of the training all coefficients are kept, thus mapping the input image to itself. Element-wise multiplication of T with $\hat{P}$ yields the filtered representation of our image:

$$\hat{P}_{\text{denoised}} = \hat{P} \odot T. \tag{8}$$

The data are then transformed back to the image domain through an Einstein sum

$$P_{\text{denoised}} = Q^{\mathrm{T}} \hat{P}_{\text{denoised}}. \tag{9}$$

The steps above are repeated m times. This effectively coincides with iterative denoising since every iteration of transform domain thresholding improves the quality of the image by removing some noise. Alternatively, it can be interpreted as a layer of a neural network. However, where in convolutional neural networks (CNNs) each layer is just an unconstrained forward operator followed by an arbitrary nonlinearity, here each layer has a clear mathematical meaning and is stabilized by projecting back the filtered coefficients to the image space. Our network is in a way also more general than a convolutional network in the sense that the transform operators are not necessarily convolutional. This can compensate for a possible loss of transform expressiveness imposed by the orthogonality restriction.

Finally, the patches are rearranged back to their original position in the image to obtain X_{denoised}.

We would like to stress that, while many algorithms have been proposed to learn sparsifying transforms,[40] this end-to-end approach allows (1) to learn the “regularization”parameters (i.e., the thresholds) as well, which would otherwise have to be tuned with an expensive grid search and (2) to jointly learn stacked transform matrices which would otherwise have to be optimized separately in a greedy fashion.

2.4.3 | Image decomposition

Since the noise in our phase data corrupts the entire frequency spectrum, applying the steps above solely to the input image will only attenuate the high-frequency noise, leaving all low-frequency noise intact. This is because the lowest frequency value accessible to the pipeline above is N_{P} pixels (the patch size). Therefore, in our particular case, it is important to process the images at multiple scales. After applying our proposed pipeline to the input image $X_{\text{noisy}} \in \mathbb{R}^{N_{\mathrm{X}} \times N_{\mathrm{X}}}$, we thus blur the denoised output image with a 2D Gaussian kernel of radius (or size) $r = 3$ and standard deviation $\sigma = 0.667$ and then downsample it by a factor of 2. σ has been chosen so that 99% of the downsampled area is covered by the kernel, $r = 3$ was sufficiently large coverage for the kernel. The resulting image is again denoised and then downsampled. This process is iteratively repeated three times, thereby generating n downscaled and denoised versions $X_{\text{n,denoised}} \in \mathbb{R}^{\frac{N_{\mathrm{X}}}{2^n} \times \frac{N_{\mathrm{X}}}{2^n}}$ of the input image X_{noisy} for $n \in [0, 1, 2, 3]$ (see Figure 5). We stopped at $n = 3$ because successive layers did not improve denoising performance, that is, the mean squared error (MSE) did not further improve (see Figure 7).

ALGORITHM 1 Image decomposition

```
$n = 0$;
while $n \leq 3$ do
    $X_{\mathrm{n,denoised}} = \mathcal{C}[X_{\mathrm{n,noisy}}]$;
    $X_{\mathrm{n+1,noisy}} = d(\mathcal{G}[X_{\mathrm{n,denoised}}])$;
    $n = n + 1$
end
```

Let $\mathcal{C}$ be the denoising pipeline illustrated in the previous section. d and u (used below) are 2 × 2 downsampling and upsampling operators, respectively, and $\mathcal{G}$ is a zero-centered Gaussian smoothing filter with $\sigma = 0.667$. The image decomposition of our model is then summarized by Algorithm 1.

It is important to note here that the denoising at multiple scales happens sequentially. While it could also be applied in parallel, this led to less accurate results. This is also more intuitive, since it seems redundant to denoise the same frequencies at more than one scale. As shown in Figure 5, this sequential process resembles the encoder of the U-net.[41] In the U-net, however, the filtering is not as interpretable since the filtering steps are not based on an established image filtering technique such as transform domain thresholding, but instead apply convolutions followed by a rather arbitrary activation function. The highly nonlinear and complicated decoder of the U-net is replaced in our model by a simple Laplacian pyramid assembling.

U-Nets are state-of-the-art models for image-to-image problems, and our approach opens up the possibility of being competitive with U-Nets using an interpretable algorithm.

2.4.4 | Image reconstruction

Once all n images $X_{\text{n,denoised}}$ have been denoised at different scales, they are used to reconstruct the final image. The approach has been inspired by Burger and Harmeling[42] and consists of the steps displayed in Algorithm 2 and Figure 5.

At scale n, we separate the low-frequency L_{n} (blue in Figure 5) and high-frequency components H_{n} (yellow in Figure 5). L_{n} is obtained by first smoothing and downsampling the image, thus removing the high-frequency details, followed by upsampling and smoothing. The high frequencies in contrast are simply obtained by subtracting the low frequencies from the image. Com-

ALGORITHM 2 Image reconstruction

```
$n = 2$;
while $n \geq 0$ do
    $L_{\mathrm{n}} = \mathcal{G}[u(d(\mathcal{G}[X_{\mathrm{n,denoised}}]))]$;
    $H_{\mathrm{n}} = X_{\mathrm{n,denoised}} - L_{\mathrm{n}}$;
    $Y_{\mathrm{n}} = H_{\mathrm{n}} + \mathcal{G}[u(X_{\mathrm{n+1,denoised}})]$;
    $X_{\mathrm{n,denoised}} = Y_{\mathrm{n}}$;
    $n = n - 1$
end
```

bining the best of two adjacent scales, that is, the low frequencies from the lower scale and the high frequencies from the upper scale, we obtain Y_{n}, a better version of $X_{\mathrm{n,denoised}}$. By starting with the coarsest two scales and iteratively applying these steps, the final image is assembled.

2.4.5 | Loss function and optimization

We used the MSE on the full-resolution image as a loss function and propagated back all gradients with respect to both the transform matrices and the thresholds:

$$\mathcal{L} = ||X_{\mathrm{denoised}} - X_{\mathrm{target}}||_2^2. \tag{10}$$

We deployed the Adam optimization algorithm[43] with an exponentially decaying learning rate (initial learning rate of 0.0001) and trained all models with a batch size of 1 (because of GPU memory constraints) until convergence, that is until the MSE on the validation set was not improving anymore.

2.4.6 | Algorithm interpretability and nonexpansiveness

Our algorithm is essentially composed of linear filters and thresholding. By reshaping the learned transforms into $N_{\mathrm{P}} \times N_{\mathrm{P}}$ filter kernels, it is possible to get insights about what types of features are being used to build sparse data representations. Likewise, the learned threshold values indicate how strongly the images get filtered.

While the inspection of filter kernels is also possible in CNNs, and the bias terms can give some hints about filtering strength, CNNs do not allow to easily inspect how the images are being denoised within the network. This is because of the immense number of filter channels that are applied at each layer, and which are not projected back into the image space after each filtering step. In contrast, our model allows us to easily visualize the results in the intermediate layers of the network as we iteratively move back and forth between image space and transform space. By looking at Figures 6 and 7, we observe that our model gradually improves image quality, both visually and quantitatively. From this plot, it also emerges that four scales and 10 filtering steps are good hyperparameters for this application.

Moreover, by design, our model starts training with an excellent parameter initialization. In fact, owing to its peculiar architecture and by setting the starting thresholds to 10^{-6}, prior to training, a forward pass through the network will leave the image unchanged. This is in strong contrast with conventional models such as the U-net in which a first forward pass will yield a very different output than the image target. This implies that smaller parameter adjustments need to be made during training in the INSIDEnet as compared to the U-net.

What further sets our model apart from most state-of-the-art denoising networks is that the linear transformation is explicitly constrained to be orthogonal, thereby increasing the stability of the network. In fact, it has been shown that orthogonality is sufficient for a 1-Lipschitz and nonexpansiveness property which in turn makes networks more robust to adversarial attacks.[38] Moreover, nonexpansiveness is an important property when it comes to proximal operators in plug-and-play frameworks, where the presented model is envisioned to be used.

All in all, our model resembles the U-net architecture. However, each part in our architecture has a clear mathematical rationale which comes with desirable properties and the possibility to better look inside the denoising process, thus making our approach both more interpretable and more reliable compared to standard CNNs.

2.4.7 | Computational aspects

The algorithm has been implemented in Tensorflow 2.1,[44] and all computations have been carried out on an NVIDIA Titan RTX GPU with 24 GB of memory. Processing of a single 1536 × 1536 slice takes an average of 0.15 s, depending on the hyperparameters N_{P} and m. This is three orders of magnitude faster than the BM3D filter and approximately as fast as a deep CNN.

3 | RESULTS

3.1 | Simulated data

To assess the effectiveness of the INSIDEnet, we performed a comparative study on the simulated breast phantoms (absorption and phase). We compared the INSIDEnet with a classical state-of-the-art algorithm, namely the BM3D filter, as well as with a deep CNN. We performed various experiments by changing the hyperparameters N_{P} and m and found $N_{\mathrm{P}} = 8$ and $m = 10$ to work well. We would like to mention that by using

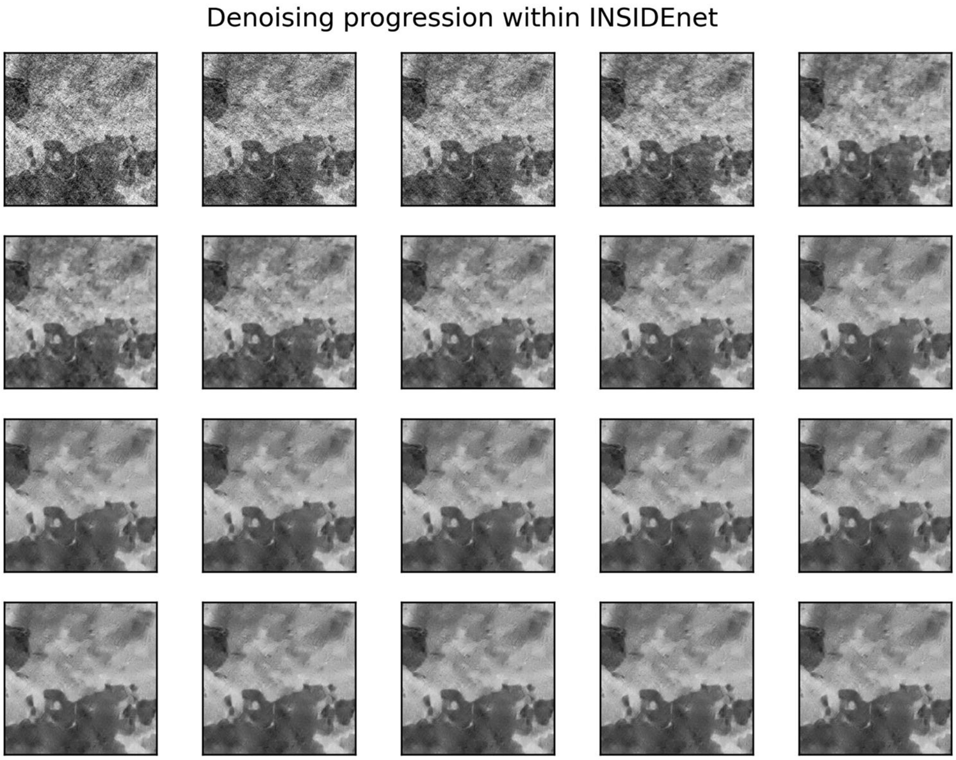

FIGURE 6 Denoising progression within INSIDEnet. The outputs after every second filter across the four scales are shown. Each row corresponds to one scale. From left to right, subsequent filters are shown. Only every second filter is shown for better visualization. Starting at the top left and moving to the bottom right, we observe that the image quality steadily improves. This is also confirmed by the results in Figure 7. As expected, the filters in the upper row only remove the high-frequency noise. The lower frequency noise is removed at the lower scales (subsequent rows)

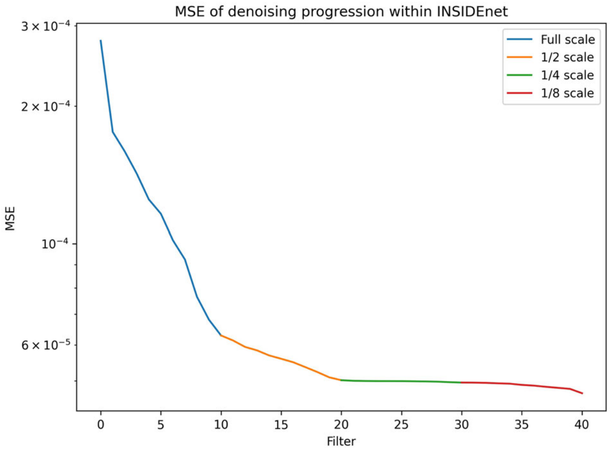

FIGURE 7 MSE of the images in Figure 6. Different scales (rows in Figure 6) are plotted in different colors for clarity. The biggest improvement in MSE happens at the finest scales, with less and less improvement when going to coarser scales

a larger N_P, less image scales would be necessary as large patches allow to denoise lower frequencies compared to small patches. However, this would go at the expense of a much higher parameter number needed to transform large patches. Therefore, to keep a smaller parameter number for higher generalizability, it is recommended to use a small patch size with more image scales. All performance metrics have been computed on 440 testing images and are provided in Table 1. To be able to compute SNR and contrast-to-noise ratio (CNR) efficiently over all 440 slices, we calculated mean and standard deviations in regions of interest where the gray level values in the ground truth image were approximately constant.

We used the BM3D filter of the bm3d.py software package.[12] The sigma value was obtained by computing the standard deviation in a uniform region of the noisy input image. Processing of a single 1536 × 1536 pixels slice took 76 s on a CPU.

As a deep CNN, we implemented a U-net[41] with 269 176 parameters trained with am MSE loss in Tensorflow 2.1[44] which, as for the INSIDEnet, separately processes either phase or absorption images. The parameter number has been kept relatively small to be approximately in the same range as the 166 403 parameters of the INSIDEnet model and to avoid overfitting. We used an initial learning rate of 0.0001 with exponential decay, along with the Adam optimization algorithm ($\beta_1 = 0.9, \beta_2 = 0.999$).[43] Before entering the network, the input images were brought to zero mean and unit variance. As for our proposed algorithm, the training set consisted of 440 pairs of noisy and clean simulated

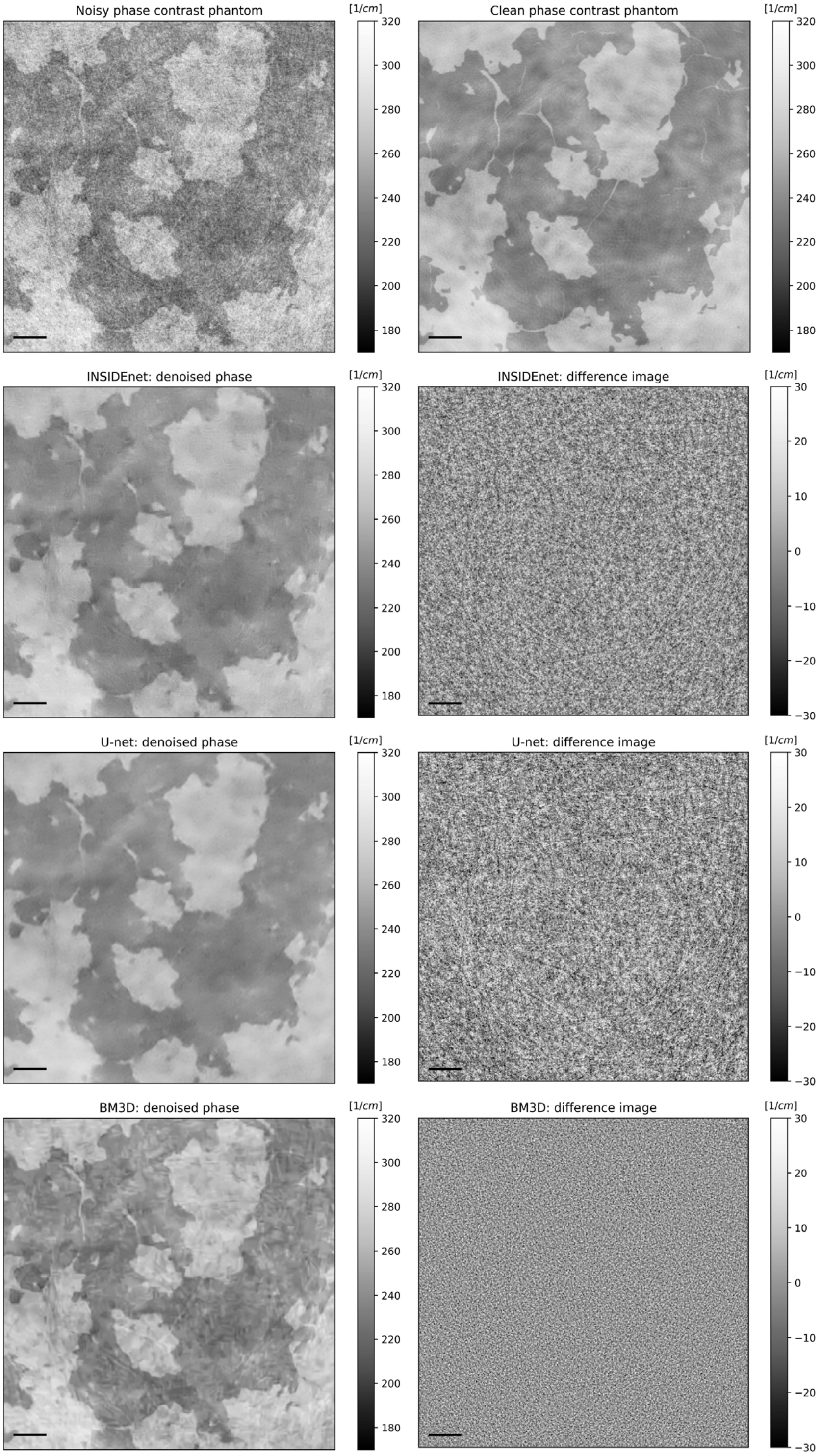

FIGURE 8 Denoising result on the phase contrast phantom obtained with the INSIDEnet, the U-net, and the BM3D filter. Top left: noisy input image, top right: clean image. In subsequent rows left: denoised image, on the right: difference image between noisy input and denoised image. Gray value units are phase shift coefficients [cm^{-1}]. The black scale bar is 5 mm

breast phantom slices. We would like to point out that we purposely chose a rather small training set, as this will be our real-world scenario once we will start acquiring real data. Therefore, we wanted to investigate the algorithm's capability to generalize well from little data. Early stopping has been used to arrest the training when the validation loss was not improving anymore. The processing time for a single slice during prediction was 0.06 s.

The denoising results of the three algorithms on phase contrast phantoms are shown in Figure 8. We display the results on a zoomed-in part of the image for better visualization. In the top row, a noisy image is shown along with its clean counterpart. In the other rows, the performance of the three algorithms is shown, along with the difference image of the denoised and the input data. We see that the INSIDEnet is able to

TABLE 1 Denoising results

Model	nMAE	SNR	CNR	SSIM
Phase				
Input	0.030 (0.002)	29.776 (1.998)	2.905 (0.193)	0.861 (0.023)
INSIDEnet	0.017 (0.003)	64.305 (5.944)	6.179 (0.470)	0.962 (0.012)
BM3D	0.017 (0.002)	53.420 (4.380)	5.208 (0.417)	0.965 (0.010)
U-net	0.015 (0.002)	72.919 (7.095)	6.934 (0.480)	0.973 (0.007)
Absorption				
Input	0.059 (0.004)	15.143 (1.111)	2.975 (0.207)	0.694 (0.014)
INSIDEnet	0.017 (0.002)	62.463 (6.473)	11.663 (0.995)	0.926 (0.011)
BM3D	0.018 (0.002)	63.103 (6.515)	11.742 (0.948)	0.924 (0.010)
U-net	0.016 (0.002)	64.541 (6.378)	11.945 (0.958)	0.939 (0.008)

Note: Standard deviations of the metrics across all 440 slices are displayed in parentheses.
Abbreviations: CNR, contrast-to-noise ratio; Interpretable NonexpanSIve Data-Efficient network; INSIDEnet, nMAE, normalised mean absolute error; SNR, signal-to-noise ratio; SSIM, structural similarity index.

effectively remove noise across the frequency spectrum while keeping sharp edges. This is also supported by the difference image in which no signal, but a large range of frequencies are present. The third row shows that, as our proposed model, also the U-net is able to satisfactorily denoise the data. The performance of the BM3D filter in contrast is much inferior: it is unsurprisingly unable to remove low-frequency noise owing to its small patch-based strategy. In fact, the difference image shows only high-frequency noise. A quantitative comparison between the three models (see the upper part of Table 1) reveals that the proposed model is only slightly inferior compared to the U-net which, in the absence of any architectural constraints, achieves the best results. In agreement with a visual inspection, the BM3D model performs significantly worse in terms of SNR and CNR. Interestingly, in terms of normalised mean absolute error (nMAE) and structural similarity index (SSIM), the BM3D algorithm is competitive with the two data-driven pipelines. This shows the limitations of such metrics as visually the superior performance of the latter methods is evident.

By looking at the denoising performance of the three algorithms on absorption data in Figure 9, we see a different pattern than in phase. In particular, owing to the different noise spectrum, concentrated in the high frequencies, all three filters achieve a satisfying result. The metrics in the lower part of Table 1 confirm that the three algorithms achieve a very similar performance.

As expected, a comparison between denoising results on phase and absorption data reveals that it is easier to denoise the latter data. In fact, all image quality metrics improve more significantly in absorption than in phase.

3.1.1 | Data efficiency

To assess the data-driven models' generalization performance, we performed an experiment in which we trained both architectures on very limited data, that is, a single image and tested them on 440 slices as in the previous experiment. The results in Table 2 show that under such conditions the INSIDEnet quantitatively outperforms the U-net. A visual inspection of the results in Figure 10 confirms these findings. In fact, a close look at the images reveals that the U-net blurs the phase image and introduces dark artifacts around some edges in the absorption image.

These results suggest that (1) the INSIDEnet architecture imposes a strong inductive bias on the denoising problem, thereby enabling it to fit its parameters with very limited data; and (2) that the INSIDEnet's efficient initialization strategy indeed helps to more easily fit the model. These two aspects clearly set our model apart from classical CNNs. This has a great practical significance as training data is always scarce in a clinical setting and especially in the development phase of a clinical prototype. Therefore, having a model which can effectively be trained with small amounts of clean images is crucial.

3.1.2 | Lesion detectability task

To assess our algorithm's performance in terms of lesion detectability, we added identical lesions with different contrasts at random locations to our simulated testing data. In Figure 11, five lesions with varying contrast (1.02, 1.04, 1.06, 1.08, and 1.10) are shown for the clean phase contrast phantom data, for the corresponding noisy data as well as for all algorithms considered in this paper. We see that none of the algorithms is able to recover the lesion below a contrast of 1.06. The BM3D filter leads to relatively good lesion delineation, despite its inability to remove the low-frequency noise. The INSIDEnet and the U-net achieve a comparable performance when trained on 440 images. A close look reveals that the INSIDEnet seems more robust

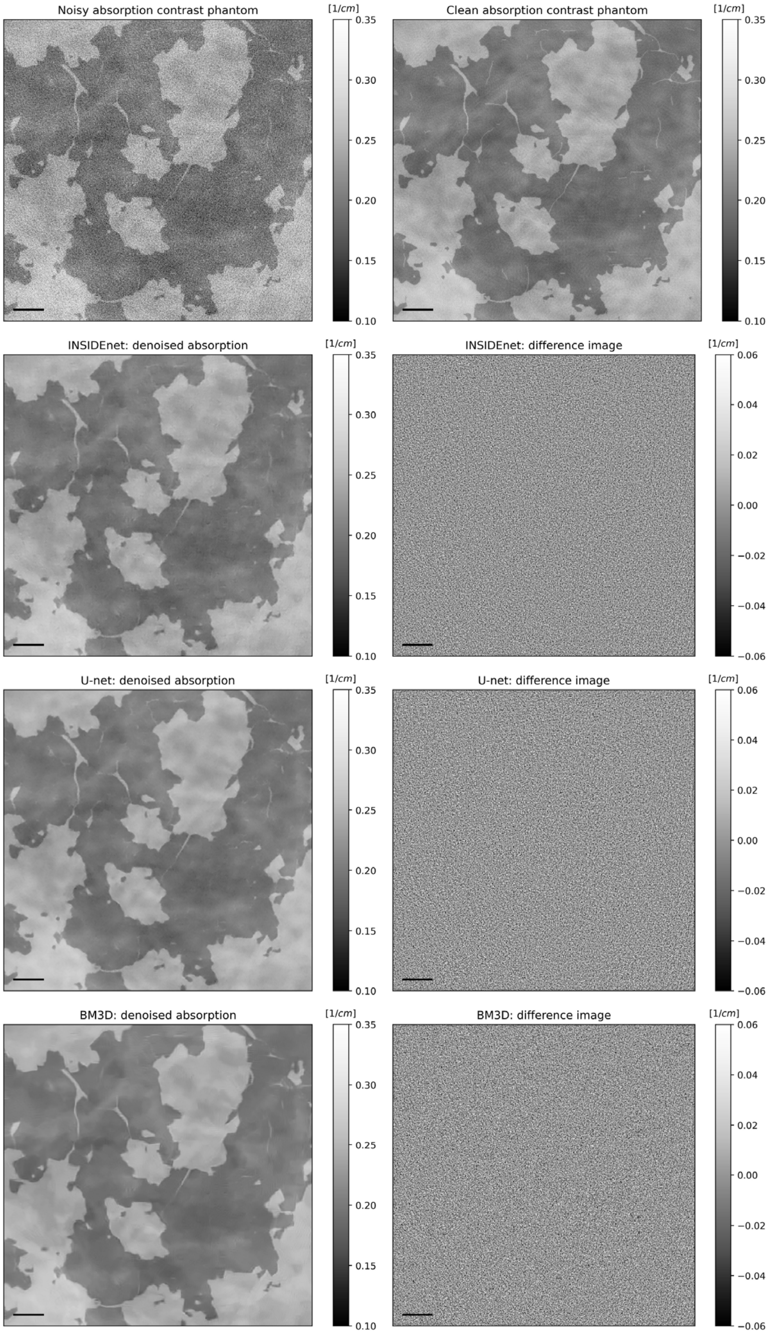
Noisy absorption contrast phantom
Clean absorption contrast phantom
INSIDEnet: denoised absorption
INSIDEnet: difference image
U-net: denoised absorption
U-net: difference image
BM3D: denoised absorption
BM3D: difference image
[1/cm]
0.35
0.30
0.25
0.20
0.15
0.10
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06

FIGURE 9 Denoising result on the absorption contrast phantom obtained with the INSIDEnet, the U-net, and the BM3D filter. Top left: noisy input image, top right: clean image. In subsequent rows left: denoised image, on the right: difference image between noisy input and denoised image. Gray value units are attenuation coefficients [cm^{-1}]. The black scale bar is 5 mm

TABLE 2 Denoising results with a model trained on a single image pair

	nMAE	SNR	CNR	SSIM
Phase				
INSIDEnet	0.018 (0.003)	60.347 (5.480)	5.832 (0.413)	0.956 (0.012)
U-net	0.021 (0.002)	52.369 (5.241)	4.928 (0.333)	0.961 (0.009)
Absorption				
INSIDEnet	0.017 (0.002)	61.459 (6.313)	11.515 (0.973)	0.924 (0.011)
U-net	0.020 (0.002)	53.149 (8.013)	9.265 (0.874)	0.924 (0.010)

Note: Standard deviations of the metrics across all 440 slices are displayed in parentheses.
Abbreviations: CNR, contrast-to-noise ratio; Interpretable NonexpanSIve Data-Efficient network; INSIDEnet, nMAE, normalised mean absolute error; SNR, signal-to-noise ratio; SSIM, structural similarity index.

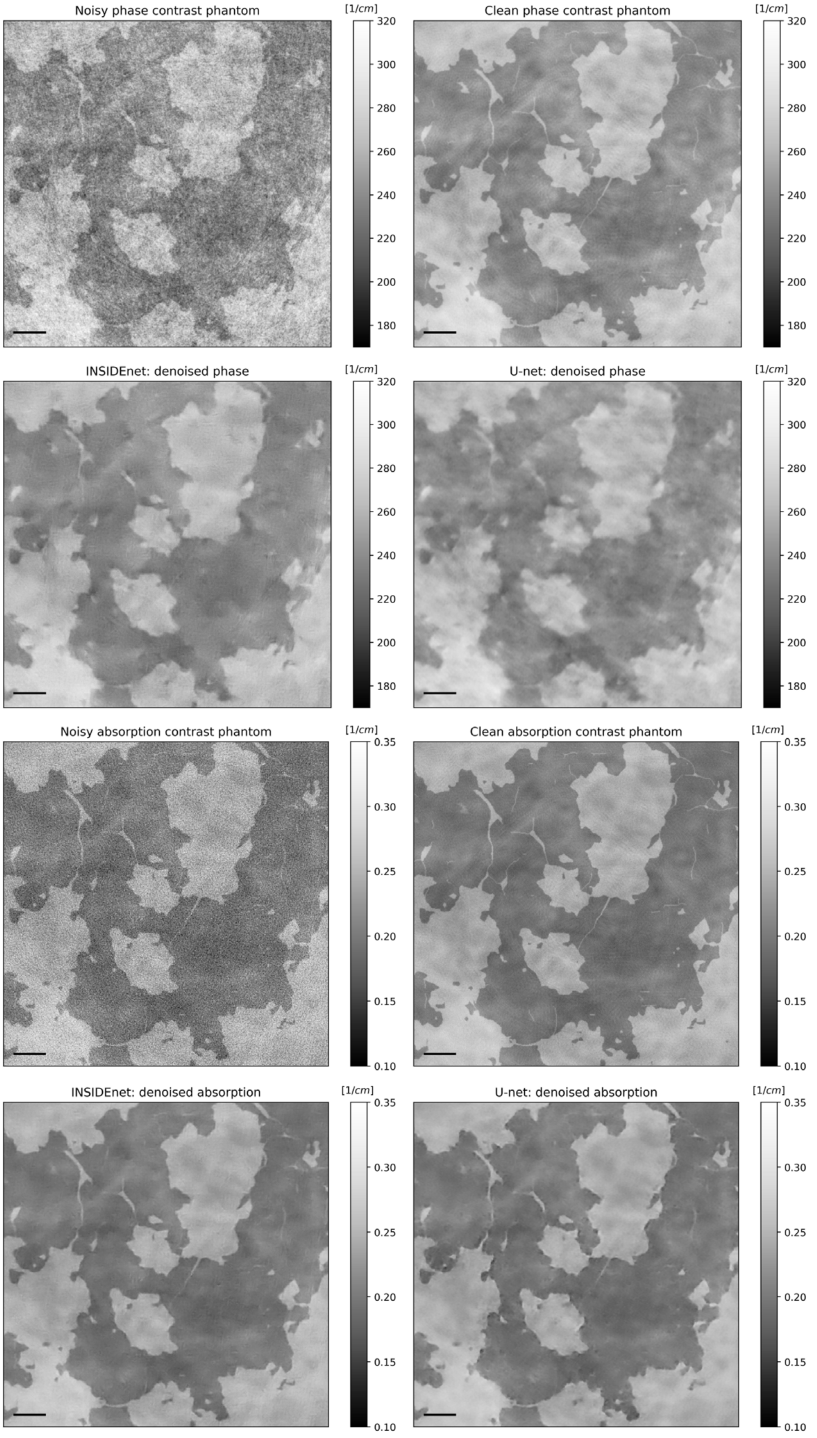
Noisy phase contrast phantom
Clean phase contrast phantom
INSIDEnet: denoised phase
U-net: denoised phase
Noisy absorption contrast phantom
Clean absorption contrast phantom
INSIDEnet: denoised absorption
U-net: denoised absorption
[1/cm]
320
300
280
260
240
220
200
180
0.35
0.30
0.25
0.20
0.15
0.10

FIGURE 10 Illustration of higher data efficiency of the INSIDEnet compared to the U-net. The models have been trained on a single training image. The black scale bar is 5 mm

compared to the U-net as it leads to slightly better lesion delineation than the U-net, when trained on a single image.

By looking at Figure 12 for detectability on absorption, we see a slightly different result. The INSIDEnet and the BM3D filter enable lesion delineation down to the contrast of 1.06, with the latter achieving the best overall performance. Somewhat surprisingly, the U-net is unable to retrieve the lesion at contrast 1.06. As for the phase, we see that the INSIDEnet is more robust compared to the U-net when trained on a single image. In fact, the former show very little loss in performance when trained on a single image.

3.2 | Real data

Figure 13 shows the meat sample scanned on our GI-BCT prototype in phase and absorption contrast. It is easy to see that indeed in absorption the noise

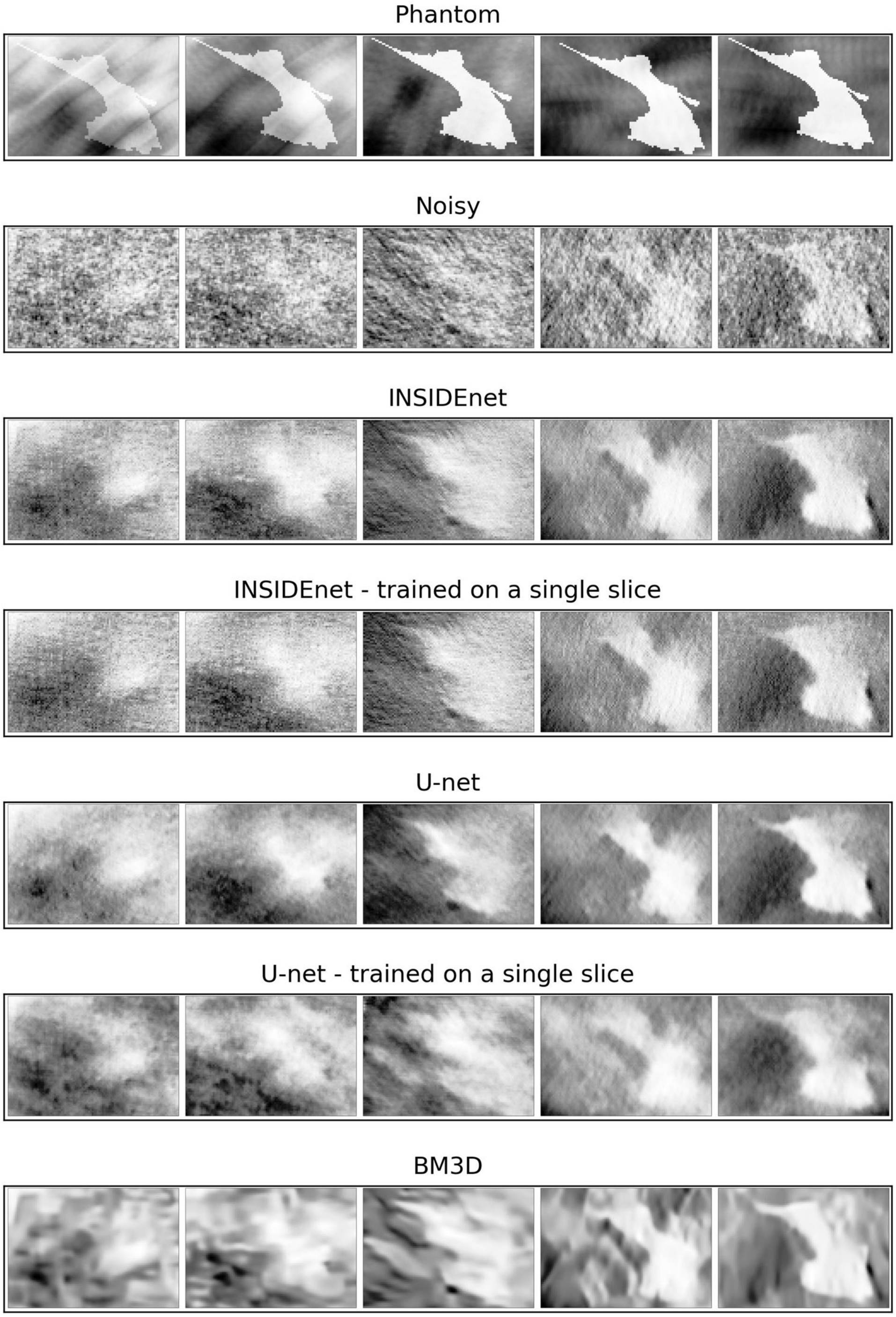
Phantom
Noisy
INSIDEnet
INSIDEnet - trained on a single slice
U-net
U-net - trained on a single slice
BM3D

FIGURE 11 Lesion detectability task on phase contrast data. First row: phantom data, second row: noisy data, third row: denoised with INSIDEnet, fourth row: denoised with INSIDEnet trained on a single image, fifth row: denoised with U-net, sixth row: denoised with U-net trained on a single image, seventh row: denoised with BM3D. From left to right the lesion contrast increases (1.02, 1.04, 1.06, 1.08, 1.10). The displayed area is 7 mm × 10 mm

is concentrated in the high frequencies, whereas in phase also lower frequencies are affected, thus making denoising more challenging. However, as it emerges from Table 3, phase contrast data actually have both a higher SNR as well as a higher CNR than in absorption contrast, thus highlighting why there is such a high interest in bringing phase contrast to clinical practice.

However, on this particular sample there does not appear to be a qualitative advantage of phase contrast data over absorption as no extra information appears to be present compared to absorption.

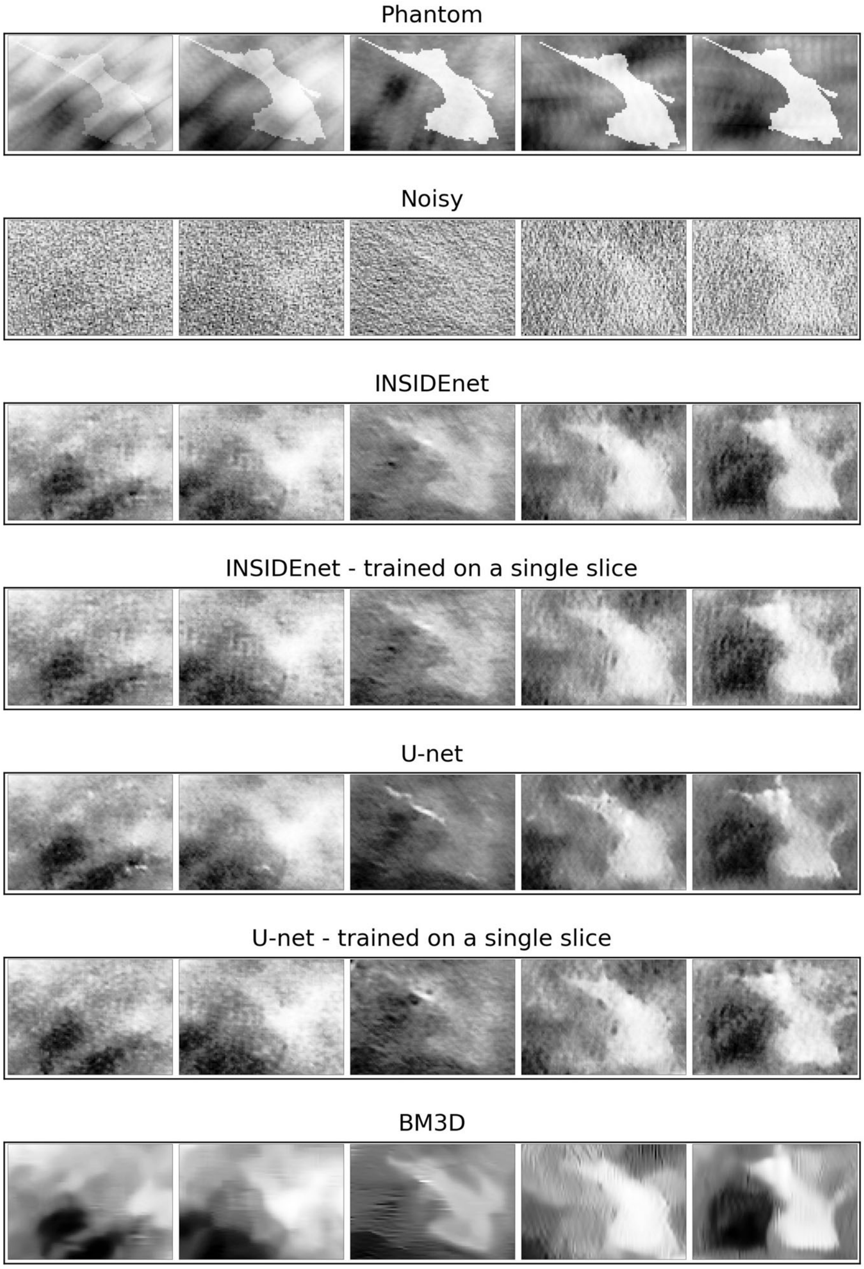
Phantom
Noisy
INSIDEnet
INSIDEnet - trained on a single slice
U-net
U-net - trained on a single slice
BM3D

FIGURE 12 Lesion detectability task on absorption contrast data. First row: phantom data, second row: noisy data, third row: denoised with INSIDEnet, fourth row: denoised with INSIDEnet trained on a single image, fifth row: denoised with U-net, sixth row: denoised with the U-net trained on a single image, seventh row: denoised with BM3D. From left to right, the lesion contrast increases (1.02, 1.04, 1.06, 1.08, 1.10). The displayed area is 7 mm × 10 mm

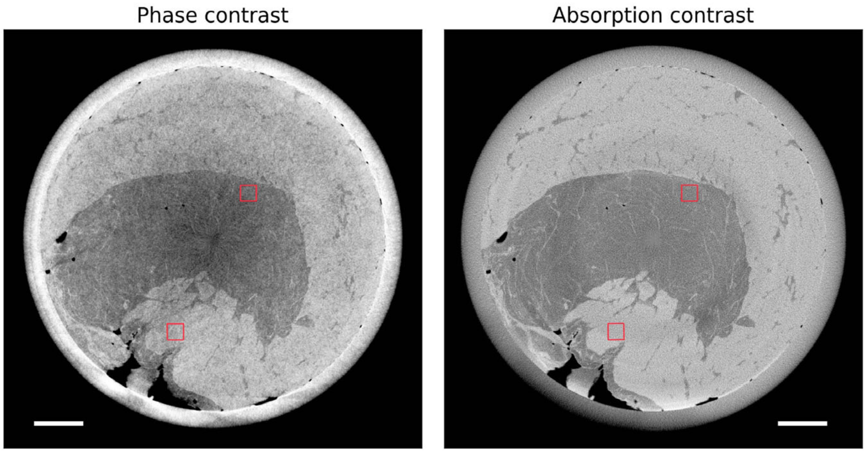

FIGURE 13 Phase contrast (left) and absorption (right) contrast of meat scan acquired on our GI-BCT prototype. The red boxes have been used to compute SNR and CNR values in Table 3. The white scale bars are 15 mm

TABLE 3 Denoising results on real data

	Phase		Absorption	
Model	SNR	CNR	SNR	CNR
Input	25.607	3.618	13.996	1.869
INSIDEnet	57.967	8.332	82.777	13.540
BM3D	38.161	5.391	79.956	12.321
U-net	63.949	8.785	86.270	14.278

Abbreviations: CNR, contrast-to-noise ratio; Interpretable NonexpanSIve Data-Efficient network; INSIDEnet, nMAE, normalised mean absolute error; SNR, signal-to-noise ratio; SSIM, structural similarity index.

Figure 14 shows the denoising results of the two data-driven models as well as BM3D. As on the simulations, on phase contrast both data-driven models achieve a satisfactory performance, whereas BM3D fails to effectively denoise the tomograms. A close look suggests though that the U-net slightly blurs the image as it was the case when we trained the model on a single slice. This might suggest that the U-net tends to blur images when it has to generalize to unseen data. Quantitatively, the U-net achieves the highest performance. BM3D performs significantly worse on both metrics. On absorption, all methods achieve a satisfying performance. The U-net achieves the highest SNR and CNR, closely followed by both the INSIDEnet and the BM3D filter. The line profiles in Figure 15 indicate that all models are able to maintain sharp edges, except for the U-net on phase contrast data.

The fact that the models could be trained on simulations and applied to real measurements suggests that both image and noise statistics of the simulations match well to real measurements. This is an important finding as it allows us to efficiently train models in the absence of ground truth.

4 | DISCUSSION

This paper is part of a larger effort to translate GI-BCT to the clinics and provides physicians and patients with a technology which can offer higher tissue contrast and thus increased chances of spotting potential malignancies. To achieve this, data-processing algorithms are needed which can handle the low raw data quality as it is currently acquired in GI-BCT. The goal of this paper was to investigate denoising of reconstructed tomograms. We used two established methods (BM3D and U-net), proposed a new dedicated method (INSIDEnet), and tested their ability to reliably deal with the complex noise pattern in GI's phase contrast channel.

Where the BM3D filter was unable to satisfactorily denoise phase contrast images, we could show that it is possible to significantly increase phase contrast image quality with the two data-driven methods. The U-net yielded the best quantitative results when trained on relatively large amounts of data, whereas the proposed model has proven to be very data efficient. In fact, a single image sufficed to obtain a satisfactory denoising performance. The power of combining data- and model-based methods has thus been highlighted by the INSIDEnet in outperforming the U-net when trained on very limited data.

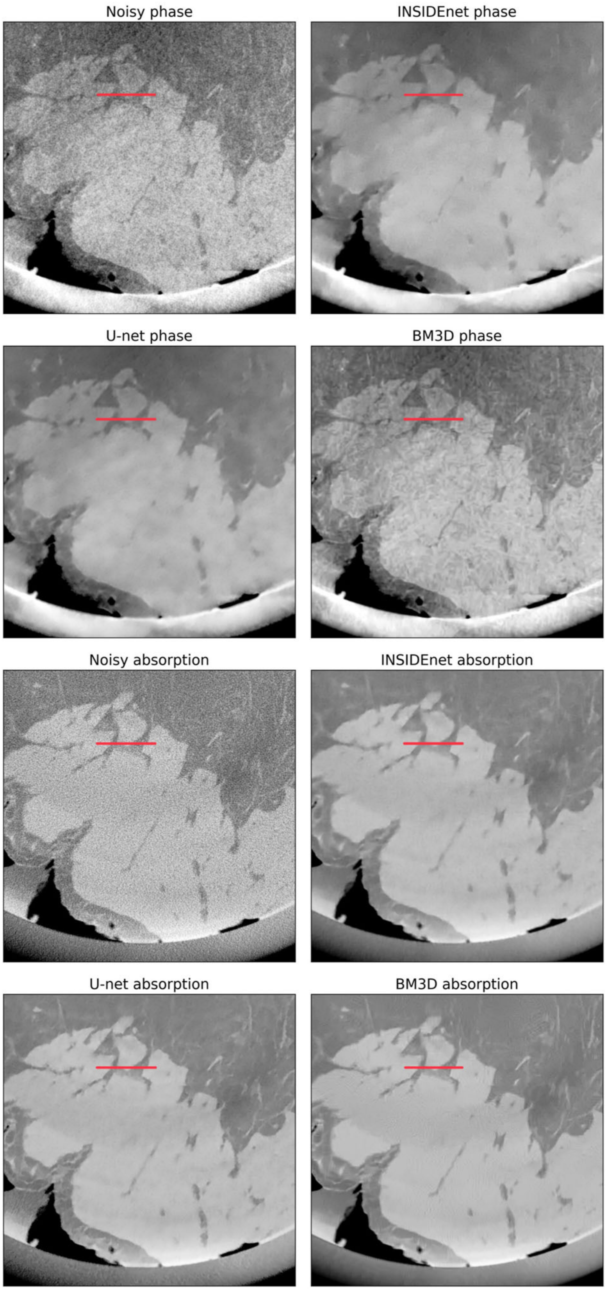

FIGURE 14 Denoising results on a scan of a chunk of meat. The red lines (10 mm) show the region at which the line profiles in Figure 15 have been computed

The INSIDEnet has been designed to find a good trade-off between performance and interpretability/robustness, crucial in the medical field. It is possible to inspect all of its trainable weights which are interpretable as they (1) indicate how strong the sparsity

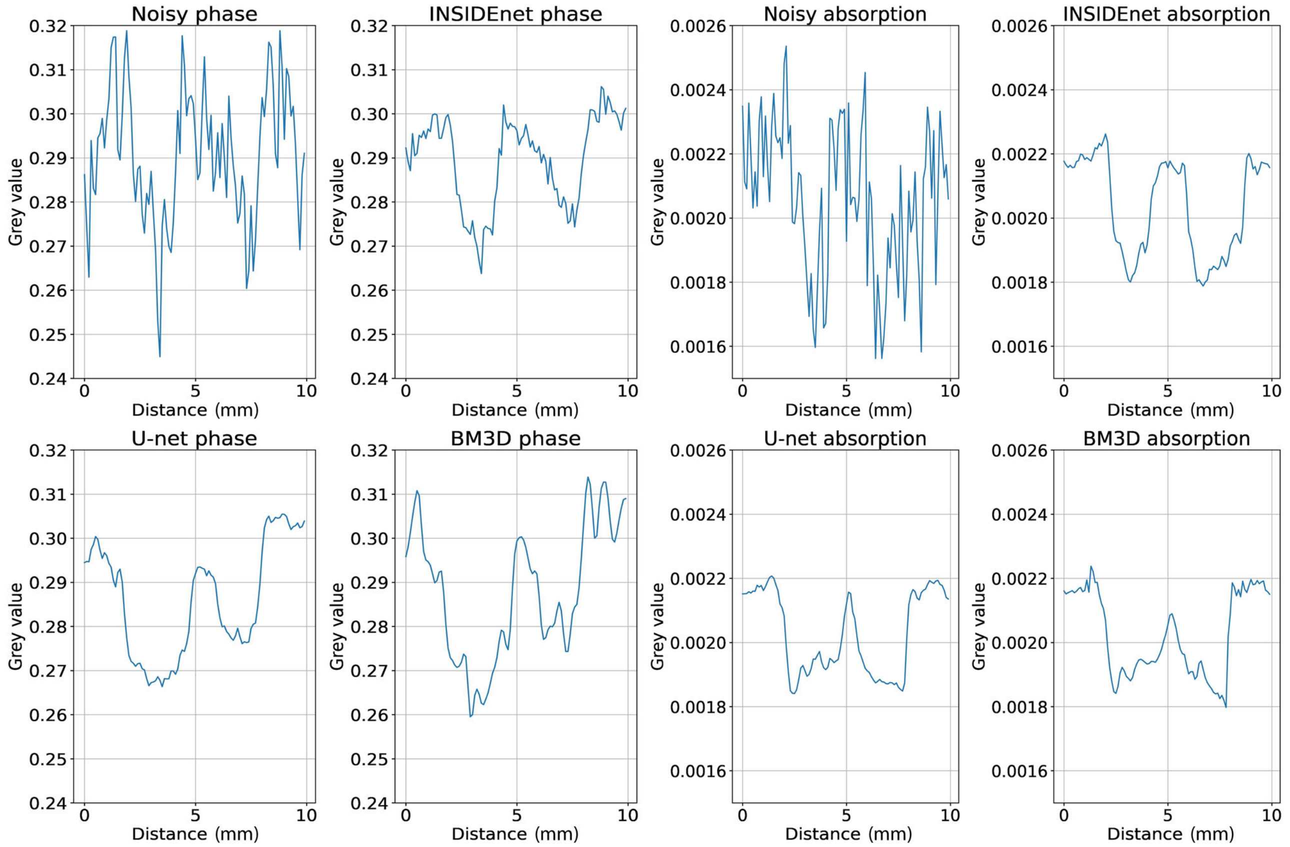

FIGURE 15 Line profiles taken at the locations displayed in Figure 14

enforcing regularization is or (2) show the filter kernels used to achieve sparse data representation. Furthermore, as shown in Figure 6, one can visualize the progress the model makes in its deeper layers, thus offering superior interpretability compared to CNNs. The steady image quality improvement that emerges in this regard supports our claim of high model reliability.

We started developing this algorithm by leveraging classical ideas in image denoising. Curiously, the final architecture resembles the U-net in many ways. First of all, the multiscale processing is very similar to the "U" structure of the U-net. Second, our model also consists of linear matrix multiplications, followed by nonlinear activations. In our case, however, the activation function is not a simple Rectified linear unit (ReLU) function but is rather mathematically motivated by the proximal operator of the l_0 norm. Finally, the concatenation of the transpose of the current transform $Q_n^\top$ and the next transform Q_{n+1} could be parameterized with a single matrix without losing expressive power. However, we believe that the inductive bias realized by going back to image space after each thresholding operation, by leveraging explicit orthogonality, instead of applying solely feed-forward matrix multiplications, is the key feature that gives our model higher robustness compared to the U-net.

A crucial part of our algorithm is thus given by the explicit orthogonality constraint enforced with the Cayley transform.[38,39] In fact, with no such constraint, model training and performance were much more unstable. While such explicit constraints could also be implemented in a convolutional layer,[38] this would require significantly higher computational costs, which are avoided by enforcing kernel orthogonality on transform matrices operating on image patches.

We would like to further point out that, besides providing more than satisfactory results, our algorithm is fast, which is critical in our project as we must process large image volumes. Finally, while the algorithm has been developed to cope with phase contrast CT, it can be applied to a variety of image data.

Finally, we could show that real data could be efficiently denoised by our model, even if trained on simulated data, thus confirming the validity of our simulation study. Unfortunately, both simulations and real data have shown that with the current setup we are unable to generate phase contrast data which are superior to absorption, neither prior to denoising nor after. Future efforts will thus be oriented towards improving grating quality to reduce the noise propagation in the phase contrast images.

5 | CONCLUSION

In conclusion, we have shown that it is possible to efficiently denoise both simulated and real noisy phase contrast images with a data-efficient, fast, and interpretable data-driven algorithm. We expect the INSIDEnet to become an important tool as part of a dedicated plug-and-play iterative reconstruction framework, which is currently under development. We hope that in future, together with improved grating quality, this will allow us to effectively deal with the low-frequency noise of the phase contrast channel and let its higher intrinsic contrast emerge.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Joan Vila Comamala and Simon Spindler for their help in designing the simulation. This work has been funded by the Promedica Stiftung Chur, ETH-Research Commission Grant Nr. ETH-12 20-2, the SNF Sinergia Grant Nr. CRSII5 183568 as well as the Swisslos Lottery Fund of Kanton Aargau.

Open access funding provided by Eidgenossische Technische Hochschule Zurich.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

The data and source code that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Harbeck N, Gnant M. Breast cancer. *Lancet*. 2017;389:1134-1150.
2. Løberg M, Lousdal ML, Bretthauer M, Kalager M. Benefits and harms of mammography screening. *Breast Cancer Res*. 2015;17:1-12.
3. Kalender WA, Kolditz D, Steiding C, et al. Technical feasibility proof for high-resolution low-dose photon-counting CT of the breast. *Eur Radiol*. 2017;27:1081-1086.
4. Shim S, Saltybaeva N, Berger N, Marcon M, Alkadhi H, Boss A. Lesion detectability and radiation dose in spiral breast CT with photon-counting detector technology: a phantom study. *Invest Radiol*. 2020;55:515-523.
5. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, et al. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*; 2013. Reston, VA: American College of Radiology.
6. Zhou SA, Brahme A. Development of phase-contrast X-ray imaging techniques and potential medical applications. *Phys Med*. 2008;24:129-148.
7. Vila-Comamala J, Romano L, Jefimovs K, et al. High sensitivity X-ray phase contrast imaging by laboratory grating-based interferometry at high Talbot order geometry. *Opt Express*. 2021;29(2):2049-2064.
8. Longo R, Arfelli F, Bonazza D, et al. Advancements towards the implementation of clinical phase-contrast breast computed tomography at Elettra. *J Synchrotron Radiat*. 2019;26:1343-1353.
9. Weitkamp T, Diaz A, David C, et al. X-ray phase imaging with a grating interferometer. *Opt Express*. 2005;13:6296-6304.
10. Raupach R, Flohr T. Performance evaluation of x-ray differential phase contrast computed tomography (PCT) with respect to medical imaging. *Med Phys*. 2012;39:4761-4774.
11. Raupach R, Flohr TG. Analytical evaluation of the signal and noise propagation in x-ray differential phase-contrast computed tomography. *Phys Med Biol*. 2011;56:2219-2244.
12. Dabov K, Foi A, Egiazarian K. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans Image Process*. 2007;16:2080-2095.
13. Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Piscataway, NJ: IEEE Press; 2005:Vol. 2;60-65.
14. Lempitsky V, Vedaldi A, Ulyanov D. Deep image prior. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE Press; 2018:9446-9454.
15. Lehtinen J, Munkberg J, Hasselgrem J, et al. Noise2Noise: Learning image restoration without clean data. In: *35th International Conference on Machine Learning, ICML 2018*. Red Hook, NY: Curran Associates; 2018:Vol. 80; 2965-2974.
16. Batson J & Royer L Noise2Self: Blind denoising by self-supervision. In: 36th International Conference on Machine Learning. Red Hook, NY Curran Associates; 2019:524-533.
17. Abdelhamed A, Timofte R, Brown MS, et al. NTIRE 2019 challenge on real image denoising: methods and results. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA: IEEE Computer Society; 2019:2197-2210.
18. Ge Y, Liu P, Ni Y, et al. Enhancing the X-Ray differential phase contrast image quality with deep learning technique. *IEEE Trans Biomed Eng*. 2021;68:1751-1758.
19. Snigirev A, Snigireva I, Kohn V, Kuznetsov S, Schelokov I. On the possibilities of x-ray phase contrast microimaging by coherent high-energy synchrotron radiation. *Rev Sci Instrum*. 1995;66:5486-5492.
20. Bonse U, Hart M. An x-ray interferometer. *Appl Phys Lett*. 1965;6:155-156.
21. Davis TJ, Stevenson AW. Direct measure of the phase shift of an X-ray beam. *J Opt Soc Amer A*. 1996;13:1193-1998.
22. Diemoz PC, Endrizzi M, Hagen CK, et al. Edge illumination X-ray phase-contrast imaging: nanoradian sensitivity at synchrotrons and translation to conventional sources. *J Phys Conf Ser*. 2014;499:012006.
23. Pfeiffer F, Weitkamp T, Bunk O, David C. Phase retrieval and differential phase-contrast imaging with low-brilliance X-ray sources. *Nat Phys*. 2006;2:258-261.
24. Olivo A, Speller R. A coded-aperture technique allowing X-ray phase contrast imaging with conventional sources. *Appl Phys Lett*. 2007;91:074106.
25. Talbot H. LXXVI. Facts relating to optical science. No. IV. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1836;9:401-407.
26. Momose A, Kawamoto S, Koyama I, Hamaishi Y, Takai K, Suzuki Y. Demonstration of X-ray Talbot interferometry. *Japan J Appl Phys*. 2003;42:L866-L868.
27. Arboleda C, Wang Z, Stampanoni M. Tilted-grating approach for scanning-mode X-ray phase contrast imaging. *Opt Express*. 2014;22:15447-15458.
28. Kottler C, Pfeiffer F, Bunk O, Grünzweig C, David C. Grating interferometer based scanning setup for hard X-ray phase contrast imaging. *Rev Sci Instrum*. 2007;78:043710.
29. Revol V, Kottler C, Kaufmann R, Straumann U, Urban C. Noise analysis of grating-based X-ray differential phase contrast imaging. *Rev Sci Instrum*. 2010;81:073709.
30. Teuffenbach MV, Koehler T, Fehringer A, et al. Grating-based phase-contrast and dark-field computed tomography: a single-shot method. *Sci Rep*. 2017;7:1-8.

31. Reehorst ET, Schniter P. Regularization by denoising: clarifications and new interpretations. *CoRR*. 2018;abs/1806.02296.
32. Berger M, Hubbell JH, Seltzer SM, et al. NIST Standard Reference Database 8 (XGAM). NIST, PML, Radiation Physics Division; 2010.
33. DABAX library. http://ftp.esrf.eu/pub/scisoft/xop2.3/DabaxFiles.
34. White DR, Griffith RV, Wilson IJ. ICRU Report 46. *J Int Comm Radiat Units Meas*. 1992;os24.
35. van Aarle W, Palenstijn WJ, De Beenhouwer J, et al. The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*. 2015;157:35-47.
36. Huang ZF, Kang K-J, Li Z, et al. Direct computed tomographic reconstruction for directional-derivative projections of computed tomography of diffraction enhanced imaging. *Appl Phys Lett*. 2006;89:041124.
37. Ravishankar S, Bresler Y. Learning sparsifying transforms. *IEEE Trans Signal Process*. 2013;61:1072-1086.
38. Trockman A, Kolter JZ. Orthogonalizing convolutional layers with the Cayley transform. Paper presented at *Ninth International Conference on Learning Representations*, 2021.
39. Chang JR, Li C-L, Póczos B, Vijaya Kumar B, Sankaranarayanan AC. One network to solve them all – solving linear inverse problems using deep projection models. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE Press; 2017:5889-5898.
40. Ravishankar S, Bresler Y. Learning doubly sparse transforms for images. *IEEE Trans Image Process*. 2013;22:4598-4612.
41. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI*. Lecture Notes in Computer Science, vol 9351. Cham, Switzerland: Springer; 2015;234-241.
42. Burger HC, Harmeling S. Improving denoising algorithms via a multi-scale meta-procedure. In: *Pattern Recognition. DAGM*, vol. 6835. Berlin, Germany: Springer; 2011:206-215.
43. Kingma DP, Ba JL. Adam: a method for stochastic optimization. Proceedings of the Third International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings; ICLR 2015, San Diego California. 2015:1-15.
44. Abadi M, Barham P, Chen J, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Software available from tensorflow.org; 2015.

How to cite this article: van Gogh S, Wang Z, Rawlik M, Etmann C, Mukherjee S, Schönlieb C-B, Angst F, Boss A, Stampanoni M. INSIDEnet: Interpretable NonexpanSIve Data-Efficient network for denoising in grating interferometry breast CT. *Med Phys*. 2022;1-20. https://doi.org/10.1002/mp.15595