Mixed scale dense convolutional networks for x-ray phase contrast imaging

Kannara Mom, Bruno Sixou, Max Langer

Research Article Vol. 61, No. 9 / 20 March 2022 / Applied Optics 1

applied optics

Mixed scale dense convolutional networks for x-ray phase contrast imaging

KANNARA MOM,[1] BRUNO SIXOU,[1] AND MAX LANGER[1,2,*]

[1]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621 Villeurbanne, France
[2]Current address: Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, F-38000 Grenoble, France
*Corresponding author: max.langer@univ-grenoble-alpes.fr

Received 15 September 2021; revised 16 December 2021; accepted 25 February 2022; posted 28 February 2022; published 0 MONTH 0000

X-ray in-line phase contrast imaging relies on the measurement of Fresnel diffraction intensity patterns due to the phase shift and the attenuation induced by the object. The recovery of phase and attenuation from one or several diffraction patterns is a nonlinear ill-posed inverse problem. In this work, we propose supervised learning approaches using mixed scale dense (MS-D) convolutional neural networks to simultaneously retrieve the phase and the attenuation from x-ray phase contrast images. This network architecture uses dilated convolutions to capture features at different image scales and densely connects all feature maps. The long range information in images becomes quickly available, and greater receptive field size can be obtained without losing resolution. This network architecture seems to account for the effect of the Fresnel operator very efficiently. We train the networks using simulated data of objects consisting of either homogeneous components, characterized by a fixed ratio of the induced refractive phase shifts and attenuation, or heterogeneous components, consisting of various materials. We also train the networks in the image domain by applying a simple initial reconstruction using the adjoint of the Fréchet derivative. We compare the results obtained with the MS-D network to reconstructions using U-Net, another popular network architecture, as well as to reconstructions using the contrast transfer function method, a direct phase and attenuation retrieval method based on linearization of the direct problem. The networks are evaluated using simulated noisy data as well as images acquired at NanoMAX (MAX IV, Lund, Sweden). In all cases, large improvements of the reconstruction errors are obtained on simulated data compared to the linearized method. Moreover, on experimental data, the networks improve the reconstruction quantitatively, improving the low-frequency behavior and the resolution.

https://doi.org/10.1364/AO.443330

1. INTRODUCTION

Phase sensitive x-ray imaging techniques with partially coherent radiation permit a significant increase of sensitivity with respect to x-ray tomography on the micro and nano scale. Phase contrast imaging is now widely used in material science and biomedical imaging with coherent x rays. The intensity can be measured at one or several propagation distances after the sample. The nonlinear relationship between the attenuation and phase shift induced by a sample and the measured intensity relies on the Fresnel diffraction theory and sets a nonlinear ill-posed inverse problem.

To obtain an approximate solution, direct inversion approaches based on linearization of the forward problem have been proposed. The contrast transfer function (CTF) [1] is one such method that allows to retrieve both attenuation and phase, while others rather rely on assumptions on relationships between phase and attenuation [2,3]. Iterative methods are not limited by these constraints; of those are techniques that retrieve the object by alternating projections on constraints between the detector and object space [4]. These also include variational approaches based on the Fréchet derivative of the forward operator [5] in conjunction with the Landweber algorithm. This kind of algorithm permits a flexible inclusion of priors, based on nonnegativity or total variation, for example, but consider the attenuation and the phase as independent unknowns to retrieve. It has recently been extended to other iteration schemes such as iteratively regularized Gauss–Newton (IRGN) [6]. Deep learning methods have been much developed in recent years for signal processing tasks [7]. Recent approaches based on deep learning have yielded promising results for reducing the reconstruction error for several inverse problems [8,9]. Some approaches optimize a reconstruction network trained to map the measured data and the reconstructed image [10]. Several iterative schemes have been proposed using deep learning methods to improve the results obtained with classical iterative approaches for inverse problems [11,12]. Some deep learning architectures applied to the phase problem have been proposed, for instance, to learn a regularization into a CTF-based optimization algorithm [13].

Others such as PhaseGAN [14] were able to recover both attenuation and phase from a single measured intensity by including explicitly the Fresnel propagator in the training. But few proposed to retrieve both attenuation and phase directly from the diffraction patterns. Recently, a new network architecture—the mixed scale dense (MS-D) network—has been proposed [15]. This network has been used to improve the reconstruction quality and produce more accurate results over traditional methods and other convolutional neural networks for tomographic reconstruction problems [16]. Therefore, the goal of this work is to develop an end-to-end deep learning approach for phase and attenuation retrieval from x-ray phase contrast images using MS-D neural networks. We compare the reconstruction results obtained with those given by the model-based CTF linear approach. The network was trained on simulated data of objects consisting of combinations of one or several different homogeneous materials at several signal to noise levels, involving a single or several propagation distances.

The remainder of the paper is organized as follows: in Section 2, we discuss the physical model of propagation based x-ray phase contrast imaging, as well as direct reconstruction methods, and we present the architecture of the network used. In Section 3, we give a detailed description of the implementation, including the data used for training. We also discuss the reconstruction quality as well as post-processing of direct reconstructions. Comparison of the results obtained with the various methods for simulated and experimental data is shown. In Section 4, we give conclusions about the results obtained and perspectives for future work.

2. MATERIALS AND METHODS

In this section, we describe the direct problem for phase contrast imaging and present the CTF approach as well as the MS-D convolutional neural networks. Additionally, we detail the synthetic datasets used for training our networks.

A. Direct Problem Definition

The interaction of a coherent and parallel x-ray beam with an object is related to its complex refractive index:

$$n(x, y, z) = 1 - \delta_r(x, y, z) + i\beta(x, y, z), \tag{1}$$

where δ_r is the refractive index decrement, and β is the absorption index for the spatial coordinate (x, y, z). Both δ_r and β depend on the material as well as the x-ray wavelength λ. For thin objects and straight-line propagation of the beam along the propagation direction z, this interaction can be described by a transmittance function T of the coordinates $\boldsymbol{x} = (x, y)$:

$$T(\boldsymbol{x}) = \exp[-B(\boldsymbol{x}) + i\varphi(\boldsymbol{x})] = a(\boldsymbol{x})\exp[i\varphi(\boldsymbol{x})]. \tag{2}$$

$B(\boldsymbol{x})$ is the absorption and $\varphi(\boldsymbol{x})$ the phase shift induced by the object. The phase shift and the absorption are projections of the absorption and refraction index, respectively, defined with the following line integrals:

$$B(\boldsymbol{x}) = \frac{2\pi}{\lambda}\int \beta(\boldsymbol{x}, z)\mathrm{d}z, \tag{3}$$

$$\varphi(\boldsymbol{x}) = \frac{2\pi}{\lambda}\int (1 - \delta_r(\boldsymbol{x}, z))\mathrm{d}z. \tag{4}$$

In the framework of the Fresnel diffraction theory, letting the beam propagate in free space over a relatively short distance D after interaction with the object can be described as a 2D convolution of the transmittance and of the Fresnel propagator for a distance D:

$$u_D(\boldsymbol{x}) = T(\boldsymbol{x}) * P_D(\boldsymbol{x}), \tag{5}$$

where

$$P_D(\boldsymbol{x}) = \frac{1}{i\lambda D}\exp\left(i\frac{\pi}{\lambda D}|\boldsymbol{x}|^2\right). \tag{6}$$

The intensity measured at a distance D downstream of the object is thus given by

$$I_D(\boldsymbol{x}) = |u_D(\boldsymbol{x})|^2 + \varepsilon(\boldsymbol{x}), \tag{7}$$

where takes into account noise and artifacts due to the acquisition conditions. Estimating the phase shift from these intensities, or diffraction patterns, is called phase retrieval. The retrieved phase shift can be used in conjunction with tomography to reconstruct the 3D refractive index. This process is called 3D phase tomography or holotomography. Our aim is to estimate both the phase and the attenuation from one or several intensity measurements whether the object involves one or several materials.

B. Contrast Transfer Function

The CTF method is based on an assumption of weak absorption and slowly varying phase shift:

$$B(\boldsymbol{x}) \ll 1, \quad |\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{x} + \lambda D\boldsymbol{f})| \ll 1. \tag{8}$$

The forward model is linearized by Taylor expanding the transmittance function to the first order:

$$T(\boldsymbol{x}) \approx 1 - B(\boldsymbol{x}) + i\varphi(\boldsymbol{x}). \tag{9}$$

Substituting into (7) and again keeping only first order terms gives

$$\tilde{I}_D(\boldsymbol{f}) = \delta(\boldsymbol{f}) - 2\cos(\pi\lambda D|\boldsymbol{f}|^2)\tilde{B}(\boldsymbol{f}) + 2\sin(\pi\lambda D|f|s^2)\tilde{\varphi}(\boldsymbol{f}), \tag{10}$$

where $\boldsymbol{f}$ is the variable in the Fourier domain, $\delta(\boldsymbol{f})$ is the unit impulse function, $\tilde{B}(\boldsymbol{f})$ is the Fourier transform of the absorption, and $\tilde{\varphi}(\boldsymbol{f})$ is the Fourier transform of the phase. Although this expression is obtained by assuming weak object interaction, it can be shown to be valid for weak absorption and slowly varying phase. Since the phase contrast factor before $\tilde{\varphi}(\boldsymbol{f})$ in (10) has zero crossings, several distances have to be used to cover as much of the Fourier domain as possible. Then, a linear least squares optimization problem is considered, taking the different distances into account with the minimization of the sum

$$\sum_D \left|2\sin(\pi\lambda D|\boldsymbol{f}|^2)\tilde{\varphi}(\boldsymbol{f}) - 2\cos(\pi\lambda D|\boldsymbol{f}|^2)\tilde{B}(\boldsymbol{f}) - \tilde{I}_D(\boldsymbol{f})\right|^2. \tag{11}$$

This can be solved simultaneously for $\tilde{B}(\boldsymbol{f})$ and $\tilde{\varphi}(\boldsymbol{f})$ by linear least squares optimization. We then have two retrieval formulas for both phase and absorption:

$$\tilde{B}(f) = \frac{1}{2\Delta + \alpha}\left[A\sum_D \tilde{I}_D(\boldsymbol{f})\sin(\pi\lambda D|\boldsymbol{f}|^2) - B\sum_D \tilde{I}_D(\boldsymbol{f})\cos(\pi\lambda D|\boldsymbol{f}|^2)\right], \quad (12)$$

$$\tilde{\varphi}(f) = \frac{1}{2\Delta + \alpha}\left[C\sum_D \tilde{I}_D(\boldsymbol{f})\sin(\pi\lambda D|\boldsymbol{f}|^2) - A\sum_D \tilde{I}_D(\boldsymbol{f})\cos(\pi\lambda D|\boldsymbol{f}|^2)\right], \quad (13)$$

with the following coefficients:

$$A = \sum_D \sin(\pi\lambda D|\boldsymbol{f}|^2)\cos(\pi\lambda D|\boldsymbol{f}|^2)$$

$$B = \sum_D \sin^2(\pi\lambda D|\boldsymbol{f}|^2), \quad (14)$$

$$C = \sum_D \cos^2(\pi\lambda D|\boldsymbol{f}|^2) \quad \Delta = BC - A^2, \quad (15)$$

where the parameter α is a Tikhonov regularization parameter.

Considering the case of a homogeneous object, characterized by a fixed known δ_r/β ratio of the induced refractive phase shifts and attenuation, both the absorption and phase can be retrieved from a single diffraction pattern [17,18].

C. Mixed Scale Dense Convolutional Neural Networks

The MS-D neural network has recently been proposed [15]. This network requires fewer trainable parameters and intermediate images than encoder–decoder networks to obtain accurate reconstruction results. The application to large images is possible, and the amount of training images needed is reduced. The MS-D convolutional neural network architecture has been used to improve tomographic reconstruction from limited data. This network gives more accurate reconstruction results than traditional methods or others convolution neural networks [16]. MS-D networks are densely connected [19]: to compute an image of a certain layer, all previous layer images are used as input instead of only those of the previous layers. MS-D networks use dilated convolutions to retain image features at various scales. The scales are mixed by choosing adapted dilation factor distributions to avoid the gridding effect [20]. With dilated convolutional filters, the long range information in images becomes quickly available in early layers of network. Greater receptive field size can be obtained earlier, making it possible to use this information to improve the results of deeper layers. This feature seems in line with the action of the Fresnel operator.

Each feature map is the result of applying the same set of operations to all previous feature maps: (1) dilated convolutions with 3×3 filters with a dilation rate selected from the list $[d_{\min}, d_{\min}+1, \ldots, d_{\max}]$, (2) summing resulting images, (3) adding a constant bias, and (4) applying a rectified linear unit (ReLU) activation function defined as $\mathrm{ReLu}(x) = \max(0, x)$. Finally, the output of the network consists of a linear combination of all feature maps generated and input channels, after the application of the ReLU activation function. The weight of each feature map, including input channels, is learned according to the receptive field in the generated images that is in line with the desired output. It means feature maps or input channels whose receptive field is more in line with the desired output would be given more weight in the final output formation in processing by pointwise convolution.

D. U-Net

The U-Net architecture was originally designed to solve segmentation problems [21]. It has been successfully used in image reconstruction as a post-processing tool of direct reconstruction in computed tomography [10]. U-Net is based on: (1) multilevel decomposition by dyadic scale decomposition based on max pooling, so that the effective filter size in the middle layers is larger than that of the early and late layers, and (2) multichannel filtering, such that there are multiple feature maps at each layer. More precisely, the U-Net architecture consists of downscaling and upscaling parts that give it the U-shaped network structure. The downscaling follows the typical architecture of a convolutional neural network. It consists of the repeated application of convolutions with 3×3 filters, each followed by a ReLU activation function, batch normalization layer, and then a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. On the other side, the upscaling part consists of an upsampling of the feature map with a 3×3 up-convolution that halves the number of feature channels and a concatenation with the correspondingly cropped feature map from the downscaling path, from which we apply two convolutions with 3×3 filters, each followed by ReLU and batch normalization. At the final layer, a 1×1 convolution is used to learn a linear combination of all feature maps to reach the desired output. The two main parameters that influence performance are the number of downscaling (and subsequent upscaling) operations and the number of channels per feature map.

E. Datasets

To compare the performance of the MS-D network and the CTF, we generated synthetic x-ray phase contrast images. The x-ray energy was set to 13 keV for a wavelength of $\lambda = 0.095$ nm, and the pixel size in object space was set to 6 nm. We created projection datasets from 3D objects created from random combinations of one to 10 shapes, consisting of either one homogeneous material (to create homogenous objects) or three different materials (to create heterogeneous objects). The refractive indices (1) used for the materials are given in Table 1.

The shapes used were ellipsoids and paraboloids with random positions and orientations. 2D analytical tomographic projections of the real and imaginary parts of the refractive

Table 1. Complex Refractive Index Materials at 13 keV

Material	Symbol	$\mu(\mathrm{cm}^{-1})$	$\frac{2\pi}{\lambda}\delta_r(\mathrm{cm}^{-1})$	δ_r/β
Gold	Au	2 790	11 395	8.16
Palladium	Pd	615	8 251	26.83
Zinc	Zn	859	5 270	12.27

index, corresponding to the phase (4) and the attenuation (3), respectively, were obtained from the 3D objects for an image size of 2048 × 2048 pixels. Objects and projections were generated using the software *TomoPhantom* [22]. Phase contrast images were generated from the projection images according to (7) at propagation distances $D = [10.1, 15.5, 17.8, 19, 20.3]$mm and downsampled to 512 × 512 to avoid aliasing in the calculation of the diffraction patterns. The datasets were generated using different levels of white Gaussian noise to yield a certain peak to peak signal to noise ratio (PPSNR) in the longest distance. The noise level was kept the same at all distances, corresponding to usual experimental conditions. We generated two datasets. The first consisted of only homogeneous objects, and the material used was gold, The second consisted of heterogeneous objects using the three materials given in Table 1. Each dataset consisted of 12,000 pairs of five input images (phase contrast images at different propagation distances) and two output images (attenuation and phase). From each dataset, 10,000 images were used for training, 1000 for validation during training, and 1000 for evaluation. An augmentation of the training data was performed by random 90 deg rotations or flipping, to a factor of two, yielding a total of 20,000 training images.

3. EXPERIMENTS

For the simulations, we trained different MS-D networks corresponding to different inputs: (1) a single diffraction pattern (the position closest to the focus among the five available, i.e., $D = 10.1$ mm), (2) all five diffraction patterns, and (3) an initialization with the adjoint of the Fréchet derivative (see Section 3.C). This was done for each dataset, i.e., for both homogeneous and heterogeneous objects. Thus, a total six MS-D networks were trained. Overall, we used the same MS-D network architecture composed of 100 layers and 3 × 3 dilated convolutional kernel. The dilation rates were selected in the list $[1, 2, \ldots, 10, 1, 2, \ldots, 10, 1, 2 \ldots]$. The networks were trained using the ADAM optimizer with l_2 norm between labels and predictions as a loss function. An independent set of image pairs was used as a validation set to monitor the network quality during training and provide a stopping criterion. The network parameters that yielded the lowest validation error were saved as output from the training procedure.

A. Simulation Results

In this section, we evaluate the different trained MS-D networks on synthetic data. We compare results of trained MS-D networks with the U-Net architecture and to CTF using the normalized mean square error (NMSE) defined by

$$\mathrm{NMSE}(x) = \frac{\|x - x_{\mathrm{true}}\|_2}{\|x_{\mathrm{true}}\|_2}. \tag{16}$$

Table 2. Normalized Mean Square Error and Standard Deviation (in %) for 1000 Test Images, Heterogeneous Objects

	# Distances	# Parameters	Attenuation	Phase
CTF	5	–	42.4 (19.7)	30.3 (8.99)
U-Net	5	31×10^6	11.1 (12.3)	7.65 (9.35)
MS-D Net	5	49×10^3	7.67 (10.6)	5.33 (6.74)
MS-D Net	1	45×10^3	11.8 (9.05)	7.76 (6.36)

Table 3. Normalized Mean Square Error and Standard Deviation (in %) for 1000 Test Images, Homogeneous Objects

	# Distances	# Parameters	Attenuation	Phase
CTFHomo	5	–	13.5 (3.92)	13.5 (3.92)
CTFHomo	1	–	21.4 (14.0)	21.4 (14.0)
U-Net	5	31×10^6	4.29 (4.82)	4.29 (4.82)
MS-D Net	5	49×10^3	3.95 (4.41)	3.95 (4.41)
MS-D Net	1	45×10^3	4.37 (5.50)	4.37 (5.50)

As quantitative measures of reconstruction quality, we computed the average NMSE on 1000 images that were not used in training or validation. In all cases, the regularization parameter for the CTF method was optimized upstream.

The results obtained on homogeneous objects are summarized in Table 3. The MS-D network correctly reconstructs the attenuation and phase as identical up to a constant factor, as can be seen in the identical reconstruction error in attenuation and phase. On the contrary, U-Net was not able to reconstruct both the attenuation and phase simultaneously. It retrieves only the phase while putting attenuation to zero. This is the reason that we trained only U-Net to output a single channel (phase), and consider the attenuation proportional to the phase, which explains the identical reconstruction error.

The results obtained on heterogeneous objects are summarized in Table 2. The MS-D network performs somewhat worse on heterogeneous objects because of the diversity of the dataset, but the results remain very good. We find that the network retrieves the phase somewhat better than the attenuation. For qualitative evaluation, some examples of reconstructed phase projections are displayed in Fig. 1.

B. Experimental Results

In this section, we apply the MS-D networks to experimental data [23] acquired at beamline NanoMAX at the MAX IV synchrotron (Lund, Sweden) [24]. The different diffraction patterns were magnified to have the same pixel size, which was measured to be 6 nm. The x-ray energy was set to 13 keV. The sample was placed at different positions relative to the focus and detector positions for different amounts of magnification and consequently different effective propagation distances corresponding to $D = [10.1, 15.5, 17.8, 19, 20.3]$ mm. Those phase contrast images were not directly used as input; they were magnified to have the same pixel size (Fig. 2). The object in question represents a stack of gold, palladium, and zinc with thicknesses of 163 nm, 32 nm, and 10 nm, respectively; thus the expected values for attenuation and phase are 0.0483 and 0.217, respectively.

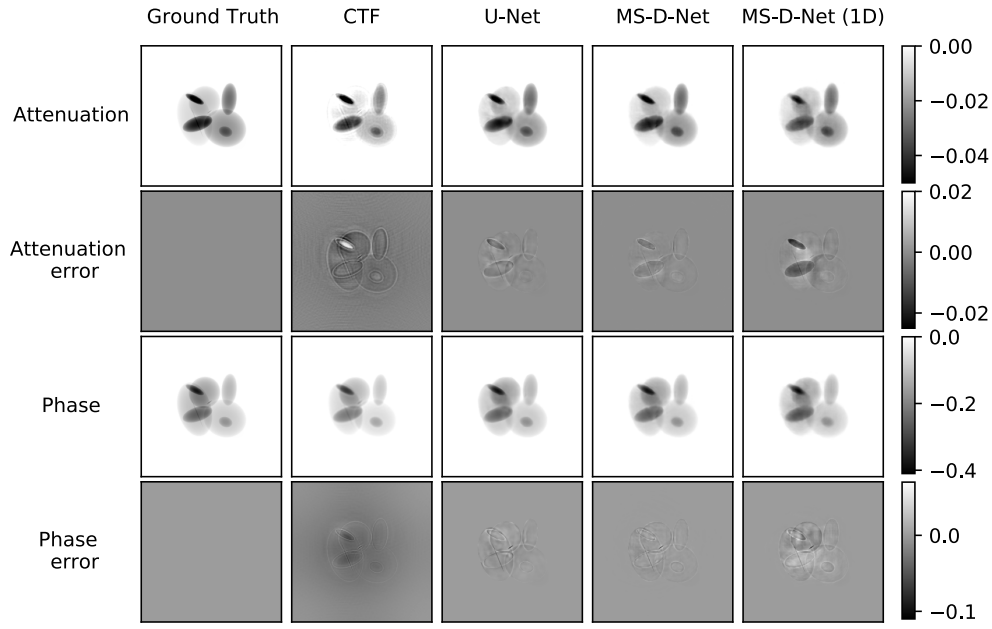

Fig. 1. Comparison of the different approaches on simulated heterogeneous objects.

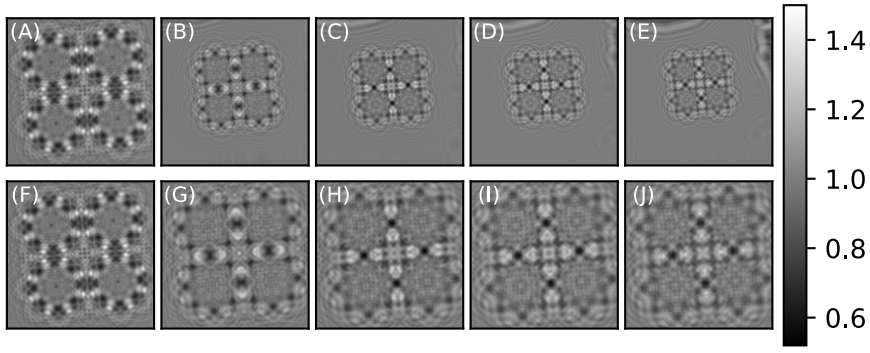

Fig. 2. (A)–(E) Phase contrast images acquired at sample positions progressively further from the focus (and thus closer to the detector) showing the varying degree of magnification and phase contrast. (F)–(J) Phase contrast images magnified to have the same pixel size (6 nm).

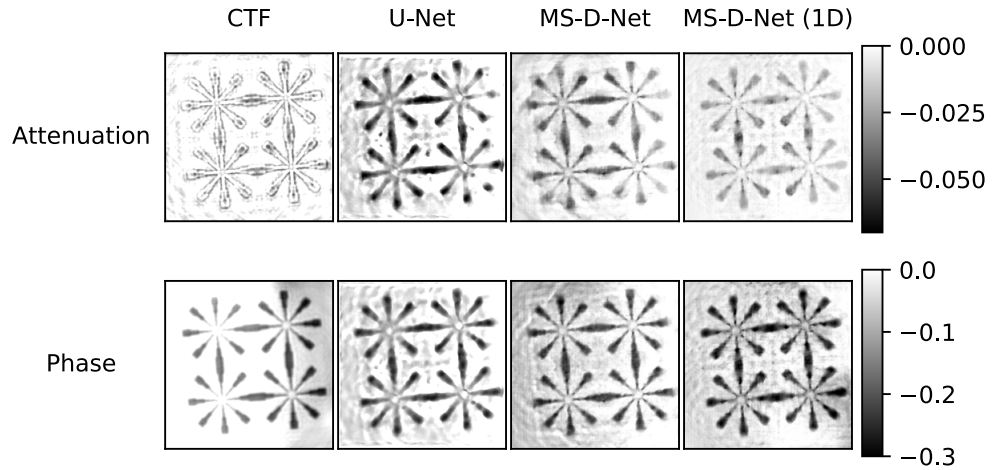

Fig. 3. Comparison of the different approaches on experimental data when trained on heterogeneous objects.

The differents results in the case of heterogenous assumption are displayed in Fig. 3. We see that the CTF method retrieved well the shape of the object but left artifacts on the low-frequency range. On the other hand, U-Net seems to reduce those artifacts but recovers the shape of the object somewhat roughly. The MS-D network reduces the artifacts while

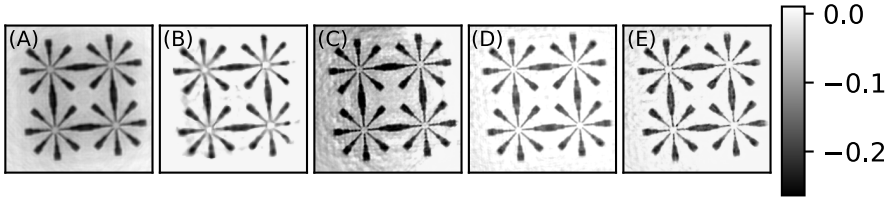

Fig. 4. Comparison of phase reconstructions using the different approaches on experimental data when trained on homogeneous objects. (A) CTFHomo (five distances). (B) U-Net. (C) MS-D-Net (five distances). (D) CTFHomo (one distance). (E) MS-D-Net (one distance).

Table 4. Reconstruction Quality for the Different Algorithms for Experimental Data when Trained on Homogeneous/Heterogeneous Data

		Heterogeneous			
		Attenuation		Phase	
	# Distances	NE (RSD) in %	Resolution (nm)	NE (RSD) in %	Resolution (nm)
CTF	5	81.3 (177)	102	21.6 (30.0)	213
U-Net	5	6.83 (35.5)	96	2.30 (16.0)	159
MS-D Net	5	33.7 (40.6)	98	3.22 (14.2)	202
MS-D Net	1	48.2 (32.0)	92	−11.5 (15.2)	208
		Homogeneous			
		Attenuation		Phase	
	# Distances	NE (RSD) in %	Resolution (nm)	NE (RSD) in %	Resolution (nm)
CTFHomo	5	4.14 (11.5)	197	4.14 (11.5)	197
CTFHomo	1	25.3 (16.0)	128	25.3 (16.0)	128
U-Net	5	14.7 (27.5)	140	14.7 (27.5)	140
MS-D Net	5	2.03 (11.6)	165	1.84 (11.2)	165
MS-D Net	1	2.96 (12.5)	93	2.76 (12.3)	93

reconstructing well the shape of the object, even when a single distance is given as input.

Assuming homogeneous object composition, we compare the homogeneous version of CTF with the networks. We see in Fig. 4 that both CTF and MS-D networks outperformed the U-Net approach, whether we use one or several distances. The MS-D network reconstructs with fewer artifacts in the low-frequency range than the linearized method, and both are able to recover well the shape of the object compared to U-Net.

For the experimental data, we used as quantitative evaluation the normalized error (NE) and relative standard deviation (RSD) calculated as

$$\mathrm{NE} = \frac{l_t - l_m}{l_t} \quad \text{and} \quad \mathrm{RSD} = \frac{s_m}{l_m}, \tag{17}$$

where l_t is the expected value, l_m the measured mean value, and s_m the standard deviation in the corresponding material. Calculation of l_m and s_m was done inside the object to avoid the influence of blur at the edges. We also measured the resolution by fitting an error function to a line profile across an object edge, and then calculating the corresponding Gaussian full width at half maximum (FWHM) based on the error function fitting parameters [3]. The result for experimental data are presented in Table 4.

For heterogeneous data, the CTF yields a reconstruction with strong low-frequency noise. All the networks achieve a quantitatively more accurate reconstruction than the CTF. Note that although U-net achieves the best reconstruction quantitatively in terms of reconstructed values, the reconstruction is qualitatively not as good: the shapes of the stars are not correctly reconstructed with some disconnections, rounded corners, and wavy contours. U-net also yields a better resolution since the reconstructions tend to approach a piecewise constant and thus work well for this particular sample.

For homogeneous data, the CTFHomo algorithm yields a quantitatively very good reconstruction with some remaining low-frequency noise, and the edges are more blurred than in the reconstructions from the networks. It can be noted that this algorithm is perfectly adapted to the imaged object and that we explicitly give the δ_r/β ratio as input, whereas the networks learn this parameter implicitly. Surprisingly, U-net does somewhat worse than for heterogeneous objects, again with some parts of the star incorrectly reconstructed, despite yielding a very clean background. The MS-D networks on the other hand yield very good reconstructions, using both five distances and a single distance.

Overall, the MS-D network when trained using a single distance propagation achieves better resolution than using several distances. This may be due to uncertainty in the measurement of the physical propagation distances and difficulty to exactly align this kind of phase contrast image. The same remark applies if we compare the results obtained by CTFHomo with one or five distances, albeit with greater improvement in quantitative results when using several distances.

Table 5. Normalized Mean Square Error and Standard Deviation for 1000 Test Images (in %), Initialized with (19)

	Homogeneous		Heterogeneous	
	Attenuation	Phase	Attenuation	Phase
U-Net	9.60 (13.8)	4.74 (8.85)	16.2 (12.3)	5.77 (7.94)
MS-D Net	1.96 (3.05)	1.96 (3.05)	8.19 (6.68)	4.83 (5.17)

C. MS-D Network as Post-Processing

Finally, we trained the networks in the reconstruction domain by performing a preliminary reconstruction by applying the adjoint of the Fréchet derivative [5] of the forward operator (7)

$$\left[I'_D(B,\varphi)\right]^*(u) = \left\{\left[\left(-u\mathrm{e}^{B-i\varphi} * P_D\right) * \overline{P_D}\right]\mathrm{e}^{-B+i\varphi}, \times \left[\left(u\mathrm{e}^{B-i\varphi} * P_D\right) * \overline{P_D}\right] i\mathrm{e}^{-B+i\varphi}\right\} \quad (18)$$

directly on the data as initialization.

The inputs to the networks then consist of the average of (18) over all distances at the point $(B, \varphi) = (0, 0)$:

$$(B_0, \varphi_0) = \sum_{i=1}^{N_D} \left[I'_{D_i}(0,0)\right]^* (I^{\mathrm{obs}}_{D_i}). \quad (19)$$

Performing this operation before training allows us to take into account prior knowledge on the physics of the inverse problem, and transforms the input data to the same domain as the output.

On simulated data, we see that initializing with this direct reconstruction using the adjoint of the derivative improves the reconstruction quality of the phase, but makes the reconstruction quality of the attenuation somewhat worse (Table 5). Initialization given by (19), as well as the reconstruction for the networks are displayed in Fig. 5.

Quantitative evaluation of the networks on the experimental data, trained by initializing with the adjoint of the derivative, are given in Table 6. When trained on heterogeneous data, the MS-D network yields better reconstruction quantitatively and in terms of resolution for the phase, while U-Net performs better for the attenuation reconstruction, and the reconstructed images no longer show the artifacts mentioned in Section 3.B. On the other hand, on homogeneous data, we see that the MS-D network yields a very good reconstruction quantitatively and achieves better resolution than U-Net.

Table 6. Reconstruction Quality for the Different Algorithms for Experimental Data when Trained on Objects Initialized with the Adjoint of Fréchet Derivative

Heterogeneous				
	Attenuation		Phase	
	NE (RSD) in %	Resolution (nm)	NE (RSD) in %	Resolution (nm)
Initialization	31.6(93.9)	371	22.5(19.6)	192
U-Net	15.1(24.3)	72	7.37(12.9)	135
MS-D Net	−27.5(34.2)	107	−4.14(11.5)	116
Homogeneous				
	Attenuation		Phase	
	NE (RSD) in %	Resolution (nm)	NE (RSD) in %	Resolution (nm)
Initialization	31.6(93.9)	371	22.5(19.6)	192
U-Net	7.04(11.6)	133	8.34(8.64)	133
MS-D Net	1.45(7.18)	122	1.38(7.15)	122

4. DISCUSSION

We used MS-D networks to perform attenuation and phase retrieval from x-ray in-line near-field phase contrast images. The network was trained and evaluated on simulated data with noise using relatively simple objects consisting of combinations of ellipsoids and paraboloids, involving one or several materials. None of the parameters of the physical model such as the energy of the x ray, propagation distance, or pixel size was given explicitly to the network. They were implicitly captured in the intensity images. The aim of the network was to learn an inverse without such information.

We illustrated the MS-D network's potential on synthetic data generated with *TomoPhantom* software and compared the results with the linearized CTF method and another neural network implementation using U-Net. The results are reported in Tables 2 and 3 and in Fig. 1. Both the MS-D network and U-Net performed better than the CTF method, both quantitatively and qualitatively. The MS-D network was able to retrieve both attenuation and phase from a single diffraction pattern with similar quality as U-Net when using five measured intensities. On homogeneous objects, the MS-D network retrieves phase and attenuation as identical up to a constant factor, which means that it learned the constant δ_r/β ratio while U-Net did not. Since U-net yielded zeros in the attenuation channel, we

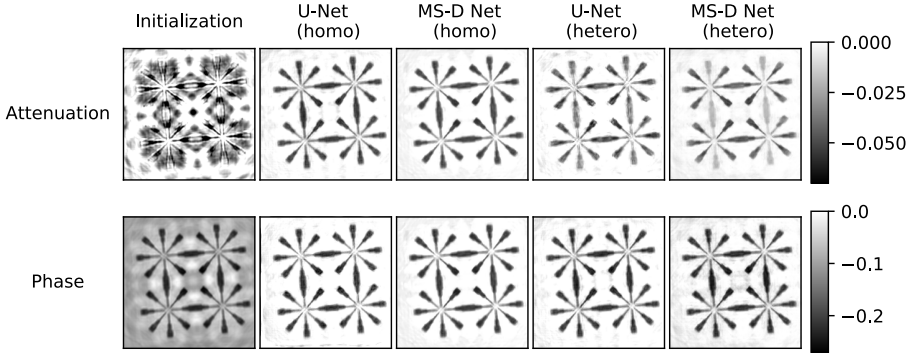

Fig. 5. Comparison of the different approaches on experimental data when initialized with the adjoint of Fréchet derivative.

implemented a single output channel for this network and calculated the corresponding attenuation image by multiplying with the δ_r/β ratio. Both networks performed better than the homogeneous version of CTF whose δ_r/β ratio was given explicitly. On heterogeneous objects, both networks performed better than the CTF, and recovered the phase better than the attenuation.

The different networks were also applied on experimental data (Figs. 3 and 4, Tables 2 and 3). Although U-Net performed very well quantitatively—the reconstructed values very close to the expected ones in the interior region of the sample—qualitatively, it reconstructed the shape of the stars in an unsatisfactory manner with rounded corners, some disconnections, and wavy contours. The low-frequency noise compared to CTF reconstruction was substantially improved, however. On the other hand, the reconstructions from the MS-D network, while yielding a quantitatively less accurate reconstruction with some ringing artifacts in the background, qualitatively reconstructed correctly the shape of the sample and also improved substantially the low-frequency artifacts compared to the CFT. The MS-D network correctly reconstructed the shape of the sample despite this kind of shape not being explicitly present in the training data. Note that U-net used approximately a factor 600 more coefficients than the MS-D networks.

Additionally, we showed using the methods as post-processing of images from a simple direct reconstruction using directly the adjoint of the Fréchet derivative. The results (Fig. 5, Table 5) show that the MS-D network performed better than U-Net in this case. For U-net, the qualitative aspect of the reconstruction was substantially improved, but quantitatively, it performed somewhat worse. The MS-D network performance was improved in all aspects using this initialization. For the presented data, the reconstruction using the MS-D network trained on homogeneous objects and initialized as described yielded the best reconstruction in quantitative terms.

5. CONCLUSION

We used MS-D networks, an architecture of convolutional neural networks, to perform phase retrieval from x-ray in-line phase contrast images. The MS-D network combines short and long range information, which seems to be appropriate with respect to the action of the Fresnel propagator. We compared the reconstruction results obtained with the MS-D network to a classical linearized algorithm, the CTF method, as well as to another deep learning method based on the U-net architecture. The MS-D network performed better than the CTF and U-net on simulated data.

On simulated data, the MS-D networks give better reconstructions than the CTF method and U-Net, despite U-net using a factor of 600 more parameters. Moreover, the MS-D networks were able to simultaneously reconstruct the phase and attenuation of heterogeneous objects from a single distance, with similar reconstruction errors to U-Net using five distances. The MS-D network was able to learn the correct δ_r/β ratio when trained on homogeneous data and performed better than U-Net, despite that U-net was trained with a single output and the correct δ_r/β ratio was applied to this reconstruction.

On experimental data, all three methods performed well, with a trade-off between the high-frequency artifacts of neural networks and the low-frequency artifacts of CTF. When using networks trained on heterogeneous objects, the MS-D network and the CTF method give the qualitatively best reconstructions with the morphology of the object well preserved. The U-Net reconstruction, while quantitatively very accurate, was qualitatively less acceptable with some disconnections and less accurate reconstruction of the object shape. When using networks trained on homogeneous objects, the MS-D network yields qualitatively and quantitatively the best reconstruction, while U-Net performed somewhat worse, with some parts of the object not reconstructed and values in the interior reconstructed less accurately.

Finally, we compared networks trained in the image domain by making a simple initialization reconstruction using the adjoint of the Fréchet derivative. The MS-D network showed excellent performance for phase and attenuation retrieval on both experimental and simulated data and improved the reconstruction compared to both the CTF and U-net.

As with all learning approaches, the reconstruction quality is limited by the quality and precision of the training data. Here, the networks are dependent on the training data to learn different physical parameters such as energy, propagation distance, and pixel size. The implementation of the simulation might also introduce its own artifacts, for example, from implementation issues such as sampling and numerical precision, and from incomplete modeling of the physics, for example, not taking into account scattering in the sample [25]. In future work, we will consider including networks in iterative schemes that incorporate some knowledge of the direct model into a data driven model. This will be done by unrolling a knowledge-driven iterative scheme and replacing the iterations with CNNs taking into account the forward operator. We will also investigate the case of a partially known forward operator by including some of the physical parameters in the optimization scheme.

Acknowledgment. We acknowledge Pablo Villanueva-Perez (Lund University, Lund, Sweden) and Sebastian Kalbfleich (MAX IV Laboratory, Lund, Sweden) for acquisition of the experimental data, and Jesper Wallentin and Lert Chayanun (NanoLund and Lund University, Lund, Sweden) for the sample.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the simulated results presented in this paper can be generated using *TomoPhantom* software [22]. Data underlying the experimental results presented are available through the *PyPhase* package [26].

REFERENCES

1. D. Paganin, *Coherent X-ray Optics*, Oxford Series on Synchrotron Radiation, 2006.
2. D. Paganin, S. C. Mayo, T. E. Gureyev, P. R. Miller, and S. W. Wilkins, "Simultaneous phase and amplitude extraction from a single defocused image of a homogeneous object," J. Microsc. **206**, 33–40 (2002).
3. M. Langer, P. Cloetens, B. Hesse, H. Suhonen, A. Pacureanu, K. Raum, and F. Peyrin, "Priors for x-ray in-line phase tomography of heterogeneous objects," Philos. Trans. R. Soc. A **372**, 20130129 (2014).
4. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).

5. V. Davidoiu, B. Sixou, M. Langer, and F. Peyrin, "Absorption and phase retrieval with Tikhonov and joint sparsity regularization," Inverse Prob. Imaging **7**, 267–282 (2013).
6. S. Maretzke, M. Bartels, M. Krenkel, T. Salditt, and T. Hohage, "Regularized newton methods for x-ray phase contrast and general imaging problems," Opt. Express **24**, 6490–6506 (2016).
7. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
8. S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," Acta Numer. **28**, 1–174 (2019).
9. J. Adler and S. Lunz, "Banach wasserstein GAN," in *Advances in Neural Information Processing Systems (NIPS)* (2018), pp. 6754–6763.
10. K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," IEEE Trans. Image Process. **26**, 4509–4522 (2017).
11. J. Adler and O. Oktem, "Solving ill-posed inverse problems using iterative deep neural networks," Inverse Prob. **33**, 124007 (2017).
12. A. Hauptmann, F. Lucka, M. Betcke, N. Huynh, J. Adler, C. Ben, B. Paul, S. Ourselin, and A. Simon, "Model-based learning for accelerated, limited-view 3-D photoacoustic tomography," IEEE Trans. Med. Imaging **37**, 1382–1393 (2018).
13. C. Bai, M. Zhou, J. Min, S. Dand, X. Yu, P. Zhang, T. Peng, and B. Yao, "Robust contrast-transfer-function phase retrieval via flexible deep learning networks," Opt. Lett. **44**, 5141–5144 (2019).
14. Y. Zhang, M. A. Noack, P. Vagovic, K. Fezzaa, F. Garcia-Moreno, T. Ritschel, and P. Villanueva-Perez, "PhaseGAN: a deep-learning phase-retrieval approach for unpaired datasets," Opt. Express **29**, 19593–19604 (2021).
15. D. M. Pelt and J. A. Sethian, "A mixed-scale dense convolutional neural network for image analysis," Proc. Natl. Acad. Sci. USA **115**, 254–259 (2018).
16. D. M. Pelt, K. J. Batenburg, and J. A. Sethian, "Improving tomographic reconstrucion from limited data using mixed-scale dense convolutional neural networks," J. Imaging **4**, 128 (2018).
17. L. Turner, B. B. Dhal, J. P. Hayes, A. P. Mancuso, K. Nugent, D. Paterson, R. Scholten, C. Tran, and A. G. Peele, "X-ray phase imaging: demonstration of extended conditions with homogeneous objects," Opt. Express **12**, 2960–2965 (2004).
18. B. Yu, L. Weber, A. Pacureanu, M. Langer, C. Olivier, P. Cloetens, and F. Peyrin, "Evaluation of phase retrieval approaches in magnified x-ray phase nano computerized tomography applied to bone tissue," Opt. Express **26**, 11110–11124 (2018).
19. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
20. Z. Wang and S. Ji, "Smoothed dilated convolutions for improved dense prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM, 2018), pp. 2486–2495.
21. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds. (Springer International Publishing, 2015), pp. 234–241.
22. D. Kazantsev, V. Pickalov, S. Nagella, P. Edoardo, and P. J. Withers, "TomoPhantom, a software package to generate 2D–4D analytical phantoms for CT image reconstruction algorithm benchmarks," SoftwareX **7**, 150–155 (2018).
23. M. Langer, Y. Zhang, D. Figueirinhas, J.-B. Forien, K. Mom, C. Mouton, R. Mokso, and P. Villanueva-Perez, "PyPhase—a Python package for x-ray phase imaging," J. Synchrotron Radiat. **28**, 1261–1266 (2021).
24. S. Kalbfleisch, Y. Zhang, M. Kahnt, K. Buakor, M. Langer, T. Dreier, H. Dierks, P. Stjärneblad, E. Larsson, K. Gordeyeva, L. Chayanun, D. Soederberg, J. Wallentin, M. Bech, and P. Villanueva-Perez, "X-ray in-line holography and holotomography at the NanoMAX beamline," J. Synchrotron Radiat., submitted for publication.
25. M. Langer, Z. Cen, S. Rit, and J. M. Létang, "Towards Monte Carlo simulation of x-ray phase contrast using GATE," Opt. Express **28**, 14522–14535 (2020).
26. M. Langer, "PyPhase-1.0.1 : An open source phase retrieval code," 2021, https://doi.org/10.5281/zenodo.4623696.