Ultrasound transmission tomography image reconstruction with a fully convolutional neural network

Zhao, Wenzhao; Wang, Hongjian; Gemmeke, Hartmut; Van Dongen, Koen W.A.; Hopp, Torsten; Hesser, Jürgen

DOI
10.1088/1361-6560/abb5c3

Publication date
2020

Document Version
Accepted author manuscript

Published in
Physics in Medicine and Biology

Citation (APA)
Zhao, W., Wang, H., Gemmeke, H., Van Dongen, K. W. A., Hopp, T., & Hesser, J. (2020). Ultrasound transmission tomography image reconstruction with a fully convolutional neural network. *Physics in Medicine and Biology*, *65*(23), Article 235021. https://doi.org/10.1088/1361-6560/abb5c3

Ultrasound Transmission Tomography Image Reconstruction with Fully Convolutional Neural Network

Wenzhao Zhao[1], Hongjian Wang[2], Hartmut Gemmeke[3], Koen W. A. van Dongen[4], Torsten Hopp[3], and Jürgen Hesser[1]

[1] Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany

[2] School of Computer Science and Technology, Donghua University, 2999 North Renmin Road, 201620 Shanghai, China

[3] Institute for Data Processing and Electronics, Karlsruhe Institute of Technology (KIT), Campus Nord, P.O. Box 3640, 76021 Karlsruhe, Germany

[4] Department of Imaging Physics, Delft University of Technology, Delft, Netherlands

E-mail: hongjian.wang@dhu.edu.cn

May 2020

Abstract. Image reconstruction of ultrasound computed tomography based on wave equation is able to show much more structural details than simpler ray-based image reconstruction methods. However, to invert the wave-based forward model is computationally demanding. To address this problem, we develop an efficient fully learned image reconstruction method based on a convolutional neural network. The image is reconstructed via one forward propagation of the network given input sensor data, which is much faster than the reconstruction using conventional iterative optimization methods. To transform the ultrasound measured data in the sensor domain into the reconstructed image in the image domain, we apply multiple down-scaling and up-scaling convolutional units to efficiently increase the number of hidden layers with a large receptive and projective field that can cover all elements in inputs and outputs, respectively. For dataset generation, a paraxial approximation forward model is used to simulate ultrasound measurement data. The neural network is trained with a dataset derived from natural images in ImageNet and tested with a dataset derived from medical images in OA-Breast Phantom dataset. Test results show the superior efficiency of the proposed neural network to other reconstruction algorithms including popular neural networks. When compared with conventional iterative optimization algorithms, our neural network can reconstruct a 110×86 image more than 20 times faster on CPU and 1000 times faster on GPU with comparable image quality and is also more robust to noise.

Keywords: Breast cancer, Ultrasound transmission tomography, Image reconstruction, Paraxial approximation, Fully convolutional neural network.

1. Introduction

Breast cancer is one of the most commonly diagnosed cancers for females [1] [2]. Early breast cancer detection increases the chance of curative treatment [3]. Ultrasound computed tomography (USCT) is a promising diagnostic tool in this respect. The method for tomographic imaging with transmission ultrasound (i.e. ultrasound transmission tomography, UTT) has been intensively studied in recent years. UTT can record the speed of sound and attenuation simultaneously. The speed of sound is shown to be closely related to tissue density [4]. It has been proved that by combining the speed of sound and attenuation images with reflection images, we can discriminate healthy tissue from cancer masses better than the diagnosis only based on the speed of sound or attenuation [5].

Transmission tomography involves solving the wave equation (the Helmholtz equation), which is associated with a heavy computational burden [6] [7] [8]. To reduce the computational costs, approximation methods are used such as straight ray approximation, bent ray approximation [9], Born approximation [10], Rytov approximation [11], and paraxial approximation [9]. The straight ray approximation ignores refraction and diffraction, which leads to the worst image resolution [12]. Among the above mentioned approximation methods, the paraxial approximation achieves the highest precision that is similar to full-wave solutions with the computational complexity reduced effectively [7]. Recently, this approximation method has been combined with various optimization methods to accelerate the reconstruction [13]. However, this iterative optimization reconstruction strategy is sensitive to noise and needs regularization [14].

In recent years, deep learning has been demonstrated to improve the reconstruction of medical images. The state-of-the-art deep-learning-based medical image reconstruction falls into two categories: one is to combine deep learning with traditional algorithms to improve imaging quality, such as using deep learning as prior (or regularization) term [15]; or using neural networks as post-processing method for denoising, and artifact removal [16]. The other category is neural-network-based direct image reconstruction from measurement data [17] [18] [19]. One of the most successful algorithms in this category is Automap [17]. It combines fully connected layers with convolutional layers for MRI image reconstruction, where the fully connected layers are used for domain transform while the convolutional layers are for extracting high-level features from the data and forcing the image to be represented sparsely in the convolutional-feature space. However, the fully connected layer requires a huge number of parameters for normal-size images, which makes Automap difficult for practical applications. In the field of ultrasound imaging, there has been research works on applying neural network for improving and accelerating the image reconstruction [20] [21]. However, up to now, the research on deep-learning-based image reconstruction of transmission tomography is quite limited. The previous work in this respect yields a poor image quality with the neural network and involves fully connected layers to deal with small-size images only [22].

In this work, we propose a fully learned image reconstruction approach using a fully convolutional neural network for UTT. The contributions of this paper are embodied in four aspects:

- We designed a neural network that can efficiently reconstruct the ultrasound transmission tomography image. The proposed reconstruction method overcomes the deficiency of fully connected neural networks and can work on normal-size inputs with a reasonable number of model parameters.
- We show the importance of advanced down- and up-scaling (DUS) methods for efficient image reconstruction by neural networks, which allows a larger number of parameters with a less computaitonal burden.
- Compared with other state-of-the-art neural networks, the proposed neural network converges much faster in the training process and achieved a higher imaging quality.
- Compared with traditional algorithms, the proposed neural network is more robust to noise, at least 20 times faster on a CPU and 1000 times faster on a GPU. Its robustness to uncertainties in ultrasound transducer locations is also demonstrated.

2. Problem Formulation

The transmission tomography problem can be expressed as the minimization of the following objective function:

$$J(\eta) = \|\mathcal{T}(\eta) - p\|_2^2 \tag{1}$$

where $\eta \in X$ is the target image to be reconstructed and $p \in Y$ is the recorded data (frequency-dependent pressure field). X and Y are typically Hilbert Spaces, and the forward operator $\mathcal{T} : X \to Y$ models the relationship between the target image and the recorded data. In some conventional iterative algorithms, this inverse problem is often regularized by assuming that the reconstructed speed-of-sound (SoS) profle is smooth. The smooth constraint can be implemented by including the total variation (TV) of the reconstructed SoS vector $\|c\|_{TV}^2$ [23] [24]. Then we have

$$J(\eta) = \|\mathcal{T}(\eta) - p\|_2^2 + \lambda\|c\|_{TV}^2 \tag{2}$$

with weighting parameter λ.

As for the forward operator $\mathcal{T}$, we consider the wave equation in the frequency domain. The Helmholtz equation models the wave propagation of ultrasound through an acoustic background medium including refraction, diffraction, and multiple scattering as

$$\Delta p + k_0^2(1+\eta)^2 p = 0 \tag{3}$$

where p describes pressure field in the frequency domain (i.e. the Fourier transform of the raw waveform data), and the background wave number $k_0 = \omega/c_0$ with angular frequency ω and the SoS of the background medium c_0. The refractive index is $1+\eta$

and $\eta = a + i\frac{\mu}{k_0}$ accounts for the deviation of the inhomogeneity from the background medium. Specifically, $Re(\eta) = a = \frac{c_0}{c} - 1$ is related to SoS, where c and c_0 are the SoS in the soft tissue and the background medium respectively. $Im(\eta) = \frac{\mu}{k_0}$ depends on the parameter μ that accounts for frequency dependent attenuation with $i = \sqrt{-1}$.

The full solution of the Helmholtz equation poses a very high computational burden. In this paper we use hereby the paraxial approximation [7] [12] [25] [26] [27] which is faster to compute than the full-wave inversion.

According to [12], we consider that the wave sources (i.e. the emitters) are arranged around a circle and the receivers (i.e. transducers) are put in a line at the opposite side of the emitters, where the relative position of the emitters and transducers is fixed. For each source, the wave propagates from a slice to its neighboring slice (as shown in Fig. 1), where the average ray direction is denoted by z. The forward propagation from the k-th z slice to the $(k+1)$-th z slice on a 2D computational grid $[1, N_x] \times [0, N_z]$ with equidistant step width Δx and Δz can be calculated by the following equation:

$$p_{k+1} = e^{i\Delta z k_0 \eta_k} \cdot \mathcal{F}^{-1}\{e^{i\Delta z\sqrt{k_0^2 - \xi^2}} \cdot \mathcal{F}(p_k)\} \tag{4}$$

The index k at p and η represent the k-th z slice. The spectral variable $\xi = \frac{2\pi}{\Delta x(N_x - 1)}[-\frac{N_x}{2} + 1, \cdots, 0, \cdots, \frac{N_x}{2}]^T \in \mathbb{R}^{N_x}$. The 1D discrete Fourier transform with respect to the spatial coordinate x and the 1D inverse discrete Fourier transform are denoted by $\mathcal{F}$ and $\mathcal{F}^{-1}$, respectively. For the emitter at different position of the circle, we rotate the computational grid around the region of interest (ROI) accordingly. Supposing we have NE emitters and NT transducers at the opposite position of each emitter and their relative positions are fiexed, a full scan consists of $NE \times NT$ recorded waves.

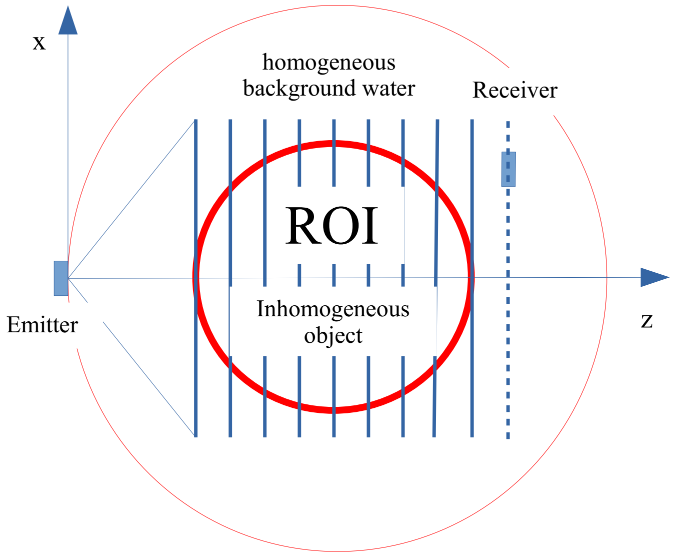


Figure 1: Steps of the paraxial approximation forward model. The ROI (indicated by the red circle) is covered by the computational grid of paraxial approximation. The ultrasound is emitted in an approximately spherical wave from the emitter in z-direction.

Our objective in this paper is to achieve fully learned direct reconstruction of image η from data p, i.e., $\mathcal{T}^{-1} : Y \to X$ with a convolutional neural network only. The structure of the subsequent paper is organized as follows. We will show the network architecture and training strategy in Section 3; the material and methods for experiments are illustrated in Section 4; and the results in Section 5; Finally, discussion and conclusion are put in Section 6.

3. Network Architecture and Training Strategy

3.1. Neural Network Architecture

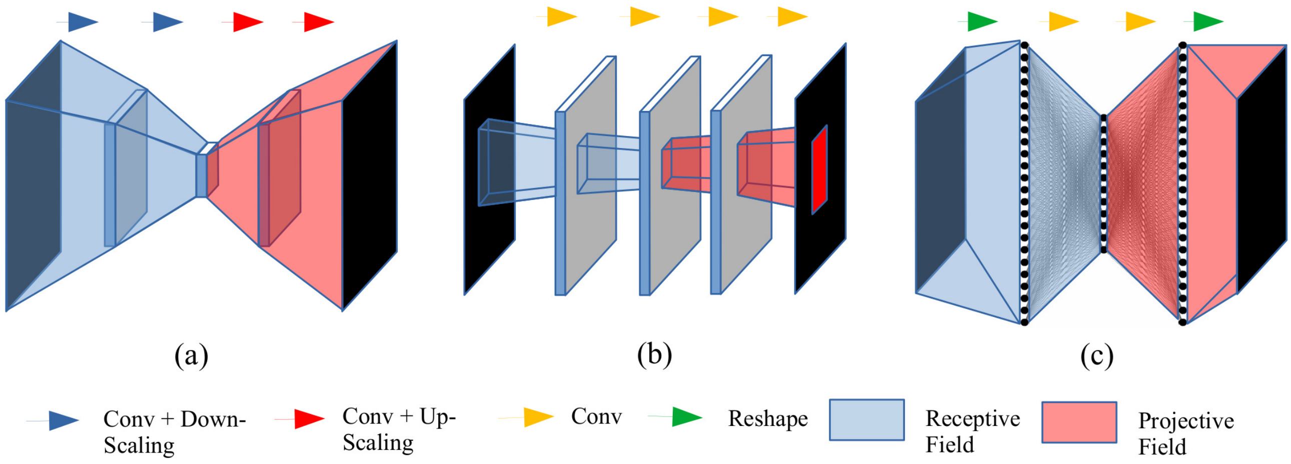
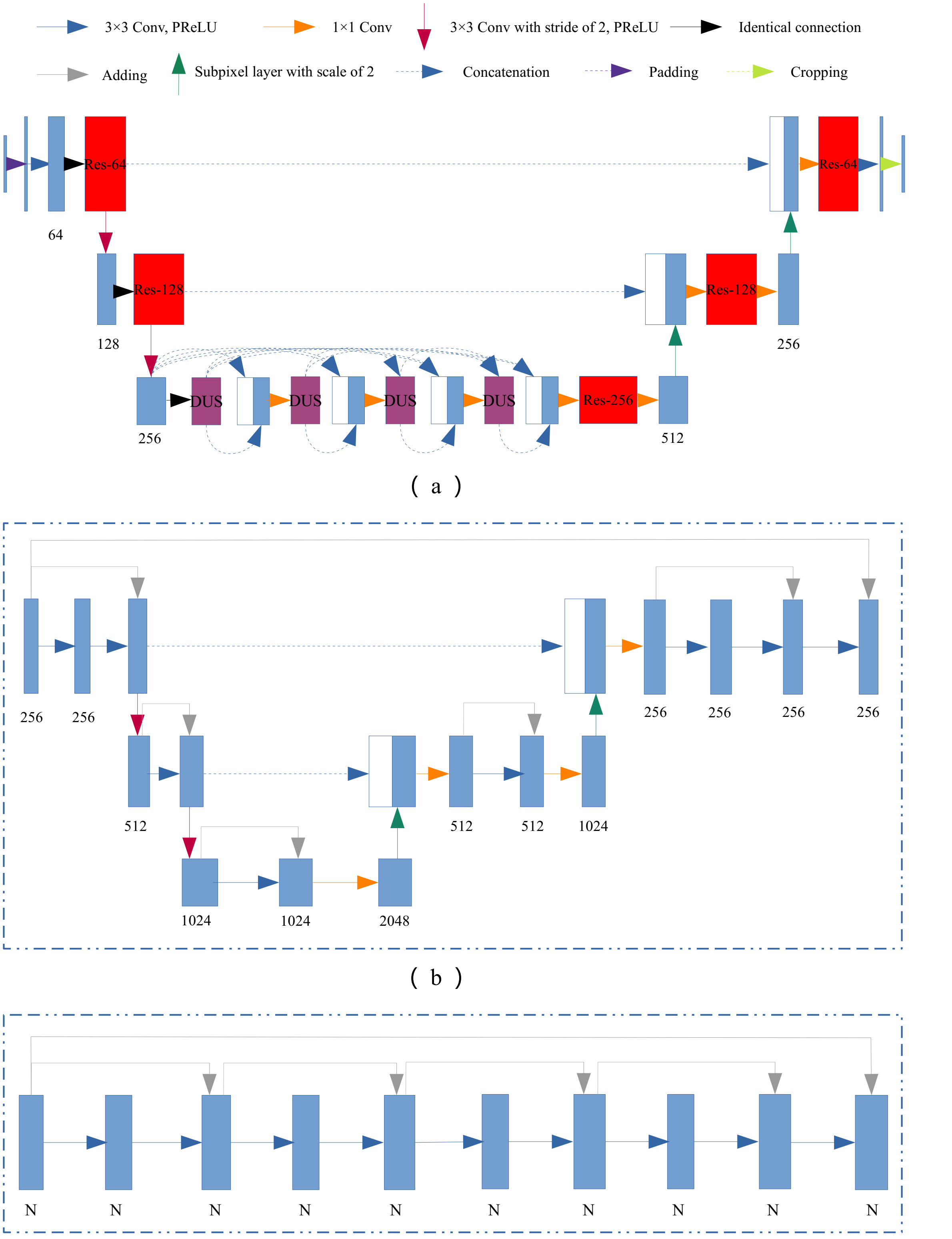


Figure 2: The receptive and projective field of the hidden layers in different neural networks. (a) The convolutional network with multiple down- and up-scaling operations; (b) The convolutional neural network without scaling operations; (c) The fully connected neural network.

For direct image reconstruction from recorded data, the value of each pixel is related to the measurement data from all sensors. Thus we need a neural network with a large receptive and projective field [28] that can cover the whole input sensor data and output image data. Recent work normally employs a fully connected neural network to obtain the maximum receptive and projective field (as shown in Fig. 2c). However, the fully connected neural network is limited to reconstructing small 2D images due to its large memory requirements. Apart from fully connected neural networks, a UNet [29] can be another option, which uses pooling units (for down-scaling) and unpooling units (for up-scaling) to gain a large receptive and projective field. However, the normal UNet uses low-pass filters such as max-pooling or average pooling methods to finish the downscaling operation. As demonstrated in [16] and [30], even though the UNet has by-pass connection to compensate for the loss of high frequency signal, it still emphasizes too much on the low-frequency signal because of the duplication of the low frequency branch. In [31], wavelet-based scaling methods are used to give more focus on high-frequency signals. However, as the wavelet transform is a special case of a convolutional layer, using the wavelet for down-scaling may limit the performance when compared with the case of using a trainable convolutional layer [32]. In this paper, we adopt a convolutional layer with a stride of 2 for tensor down-scaling as in [32] and [33]. The sub-pixel convolutional unit [34] is used for up-scaling because of its low computational cost.

The overall architecture of the proposed neural network is shown in Figure 3. The whole neural network is like a big U-shaped residual neural network containing 4 densely connected small DUS units that each forms a small U-shaped residual neural network. We hereby denorminate the neural network as multiple W-net (mWnet for short). As a whole, the neural network comprises 3 parts: feature extraction, domain transform, reconstruction.

- **Initial feature extraction**: The convolutional layers with a stride of 2 are used to down-scale the feature map. At each scaling level, a residual block with nine convolutional layers (as shown in Fig. 3) is used to encode the feature map.
- **Domain transform**: Since the domain transform is mainly done by the 'high level' layers (i.e. the hidden layers processing highly down-scaled feature maps) with a large receptive and projective field, it is necessary to put more parameters to the 'high level' layers. Inspired by the deep-learning-based work in image denoising [32] [35] and super-resolution [33], we add multiple down- and up-scaling (DUS) units to gain the number of 'high level' layers. Since these 'high level' layers are used to process the highly scaled feature maps only, the filters from these layers are more computationally efficient than those filters in the 'low level' layers (i.e. the hidden layers processing feature maps with less downscaling). Inspired by the work of DenseNet [36], we give a dense connection between these down- and up-scaling (DUS) units to further boost the performance. We use 1×1 convolutional layer for feature pooling and dimension reduction.
- **Reconstruction**: Immediately after the domain-transform part, we put a residual block after every up-scaling operation to reconstruct the image. We use sub-pixel convolutional layers to up-scale the features without degradation of features. With skip connection from the feature extraction part to the reconstruction part, we can reuse the extracted feature at different scaling levels to enhance the accuracy of reconstruction.



(a)

(b)

(c)

Figure 3: The architecture of the proposed network. (a) The overall architecture of mWnet with 4 DUS blocks; (b) The DUS block; (c) Res-N block of N channels.

We refer to the network with 1 DUS unit as mWnet_1 and the network with 4 DUS units as mWnet_4. The size of input tensor is $2 \times 110 \times 128$ (110 transducers and 128 emitters), and the size of output tensor is $2 \times 110 \times 86$ (containing the real and imaginary part of η that are related to the SoS and attenuation, respectively). Before processed by convolutional layers, each input tensor is first padded into a $2 \times 128 \times 128$ tensor. After the processing of convolutional layers, each output tensor is obtained by cropping a $2 \times 128 \times 128$ tensor. The basic convolutional unit consists of one convolutional layer followed by PReLU [37] activation function. We only use convolutional kernels with size $= 3 \times 3$ or 1×1. The total number of parameters for mWnet_1 and mWnet_4 are about 34.5 million and 113.6 million, respectively.

3.2. Training Strategy

The model is trained on the simulated data derived from natural images from ImageNet dataset [38] using paraxial approximation forward model. In the training of the neural network, to augment the data and speed up the convergence [39], random Gaussian noise is added to the input tensor with a probability of 0.7. As for the noise level, it should be noted that too high a noise level may affect the accuracy of the reconstruction, while an excessively low noise level (close to zero) cannot lead to a decent boost of convergence. Empirically, we set the SNR (Signal to Noise Ratio) range as 112 to 142 dB. Adam optimizer and l_1 loss are applied. For the implementation of l_1 loss, the real part and imaginary part of η are multiplied by a coefficient of τ and $1 - \tau$, respectively. We empirically set $\tau = 0.9$ for optimal training performance. To reduce the training time, we adopt the training strategy described in [40] by fixing the learning rate $lr = 1.0 \cdot 10^{-4}$ and increasing the batch size gradually. The model is first trained with a batch size of 16 for 49 epochs, then a batch size of 32 for 8 epochs, a batch size of 64 for 8 epochs, a batch size of 128 for 8 epochs, a batch size of 256 for 8 epochs, and finally a batch size of 512 for 8 epochs. To implement the training with a large batch size, we split a large batch of samples into a few mini-batches of size 16, and accumulate the gradients of these mini-batches before updating the variable.

To compare the neural networks with different numbers of DUS units, we implemented two different models in Pytorch [41]: mWnet_1 with one DUS unit only, and mWnet_4 with 4 DUS units. The training of all these two models follows the same strategy. The training was performed on a server with GPU of NVIDIA TITAN XP, where the training of neural network mWnet_1 and mWnet_4 needed about 4.5 and 6.3 days, respectively.

For comparison, we further trained three other neural networks with the same training strategy: Automap, UNet, and FC-DenseNet103 [42]. Both of these three neural networks have been used in the reconstruction of MRI images successfully [43].

For the tests on uncertainties in transducer locations, the mWnet_4 trained using the above training strategy is further trained on the dataset simulated with perturbed settings and follows the same training strategy.

4. Material and Methods

4.1. Data Preparation

The image size for all phantoms is 110 × 86 with each pixel of size 1.88 mm, where the radius of the measuring device is 130 mm and the radius of the phantom is 79.7 mm, and 110 transducers and 128 emitters are simulated at the frequency of 0.5 MHz.

The natural images from ImageNet are used to generate training set and validation set. We obtain grayscale images with pixel value $x \in [0, 255]$ by extracting Y-channel luminance from the RGB color images. In total, 49,998 natural grayscale images are accumulated from ImageNet. Specifically, 47,998 images are used as training set. The size of the training set is further quadrupled by combining two data augmentation operations: grayscale-value reversing and 90 degree rotation. Considering any pixel with grayscale value x, the grayscale value is reversed as $255 - x$. Each image is then rotated through 90 degrees to double the size of the training set. After data augmentation, all the 47,998 × 4 images are scaled to a size of 110 × 86. Then, for validation set, the remaining 2000 images are used and scaled to the same size. For any grayscale image derived from ImageNet set with value in the range of $[0, 255]$, its grayscale value is scaled and discretized into six integers i.e. 0, 1, 2, 3, 4 and 5 which represents water, skin, fat, gland, tumor, and calcification, respectively.

The test dataset comprises four standard test phantoms (as shown in Figure 4) and nine medical images that are randomly selected from the OA-Breast Phantom dataset [44] (as shown in Figure 5). As for the nine medical images, the pixels for different tissues in the breast are labeled as: 0 for background, 2 for fibro-glandular tissue, 3 for fat, 4 for skin layer, and 5 for blood vessel. All the images are scaled to the size of 110 × 86.

Given the image set with each tissue labeled, we assign the value of η for each pixel according to the property of SoS and attenuation for each tissue. Specifically, for water, skin, fat, gland, tumor, and calcification, the values simulated for SoS are 1485, 1570, 1450, 1490, 1560, and 6420 m/s, respectively; and the values simulated for attenuation are 0, 2.08, 1.26, 0.88, 1.60, and 8.0 $dB/cm/MHz$, respectively. The η image is then smoothed by a Gaussian filter so as to ensure that the area between different tissues has a smooth gradient of SoS and attenuation and thus becomes more realistic.

The measurement data collected by receivers (i.e. frequency-dependent sound pressure p in the frequency domain) is then calculated based on the η image and the paraxial approximation method for wave equation as described in [12]. The complex-valued measurement data (of size 110 × 128) is turned into a 2 × 110 × 128 real input tensor. Finally, the element values of all the input tensors and target images are scaled to the range of $(0, 1)$.

To test the algorithms' robustness to uncertainties in transducer locations, we generate another set of measurement data based on all the above-mentioned η images and the paraxial approximation method by adding random additive white Gaussian noise to transducers' location parameters. Specifically, we add zero-mean white Gaussian

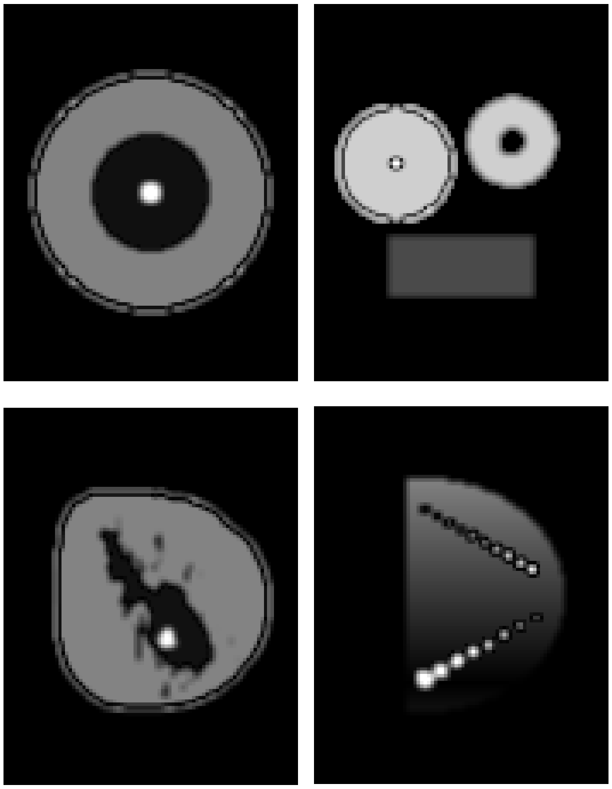


Figure 4: Four standard test samples.

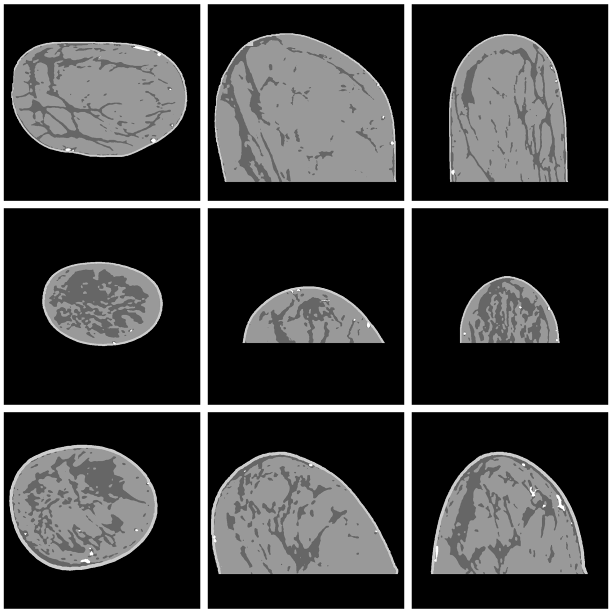


Figure 5: Nine medical image samples from the OA-Breast Phantom dataset.

noise of standard variance 0.02° to the rotation angle, and we add zero-mean white Gaussian noise of standard variance $0.01mm$ to both the x value and the z value (as shown in Fig. 1) of each receiver.

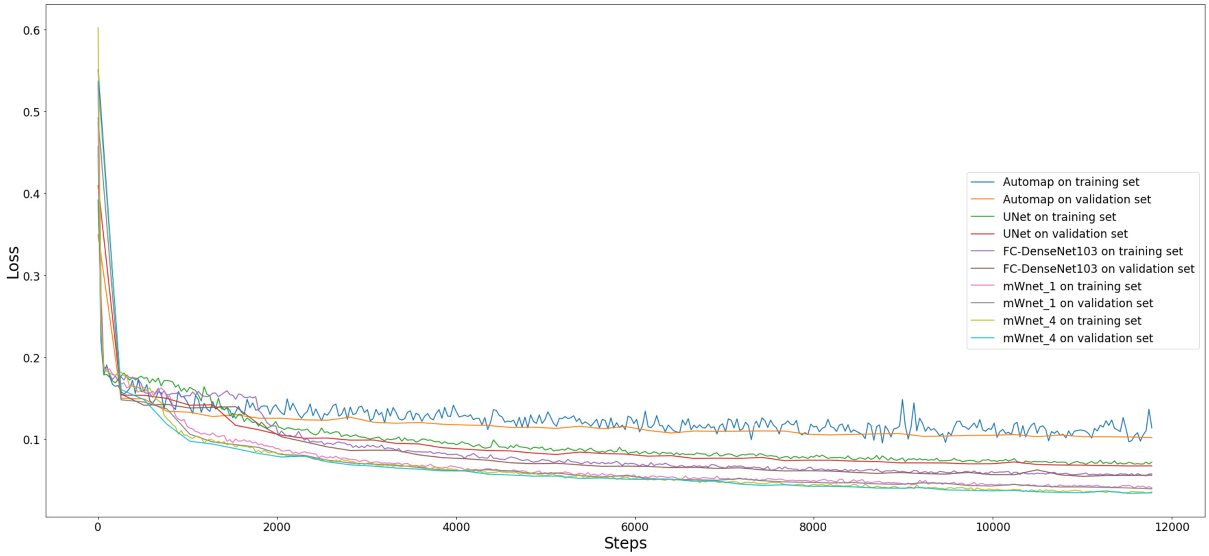


Figure 6: The learning curves for different neural networks with horizontal axis the training steps and vertical axis the l_1 loss.

Table 1: The training time and number of parameters for different neural networks

	Automap	UNet	FC-DenseNet103	mWnet_1	mWnet_4
Training Time (days)	4.6	2.1	4.7	4.5	6.3
Number of Parameters (million)	356.0	7.8	9.3	34.5	113.6

Table 2: The average runtime per image for different algorithms

	Newton CG	L-BFGS	Automap	UNet	FC-DenseNet103	mWnet_1		mWnet_4	
Runtime	49.1min(CPU)	24.9s(CPU)	0.011s(GPU)	0.008s(GPU)	0.014s(GPU)	0.778s(CPU)	0.012s(GPU)	1.056s(CPU)	0.018s(GPU)

4.2. Performance Evaluation

Imaging quality is quantified using two standard metrics: normalized root mean square error (NRMSE) and structure similarity (SSIM). The NRMSE is defined as:

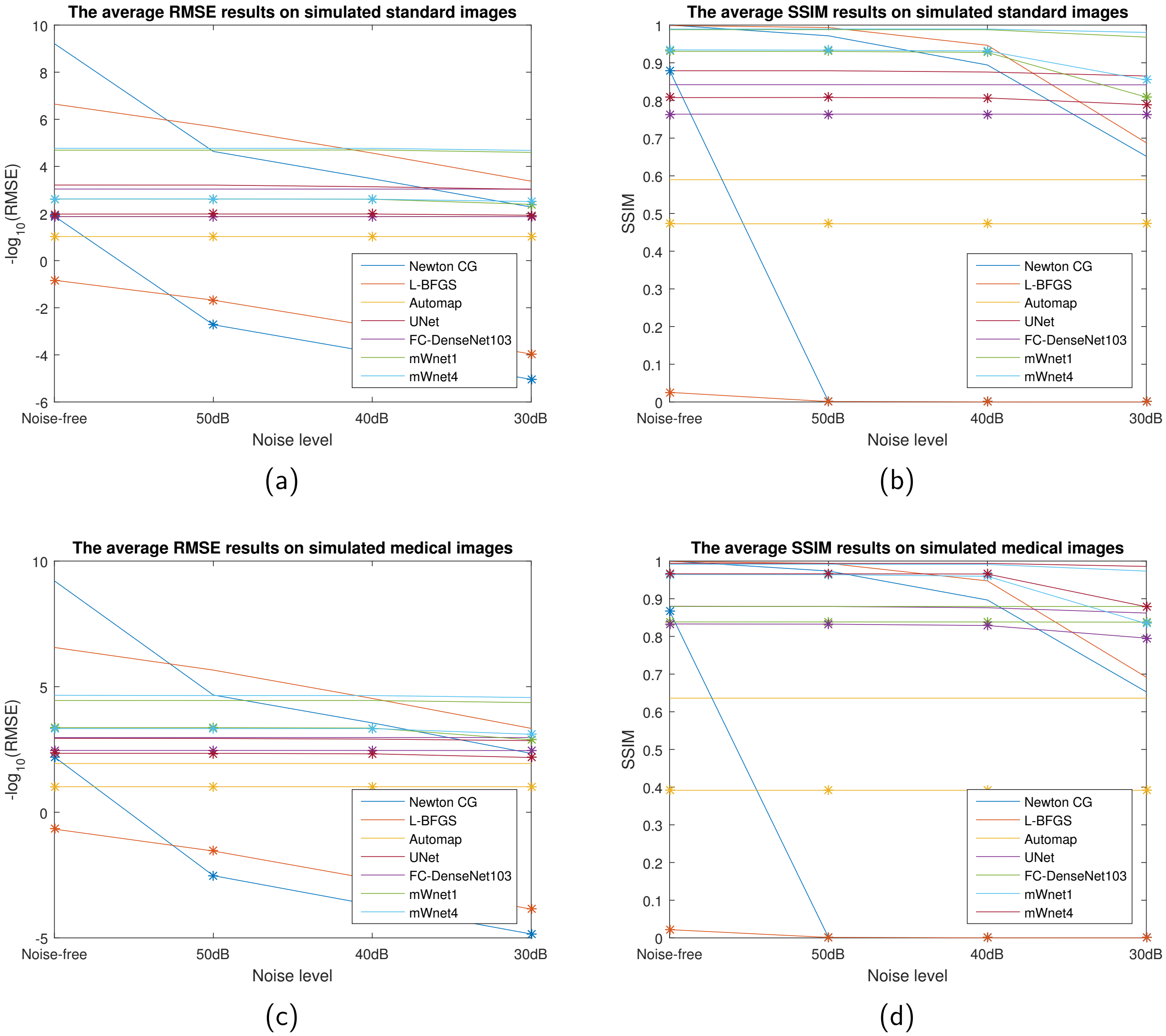
$$NRMSE = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[x(i,j)-y(i,j)]^2}{MN(x_{max}-x_{min})^2}} \tag{5}$$

where x and y denote the ground truth and the reconstructed image, respectively. M and N are the number of pixels for row and column, respectively. x_{max} and x_{min} are the maximal and minimal pixel value of the ground truth image, respectively.

The SSIM is defined as:

$$SSIM = \frac{(2\mu_y\mu_x + c_1)(2\sigma_{yx} + c_2)}{(\mu_y^2 + \mu_x^2 + c_1)(\sigma_y^2 + \sigma_x^2 + c_2)} \tag{6}$$

where μ_y is an average of y, σ_y^2 is a variance of y, and σ_{yx} is a covariance of y and x. There are two variables to stabilize the division such as $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$. L



(a)

(b)

(c)

(d)

Figure 7: The average RMSE and SSIM results on simulated images. The lines with different colors represent the quantitative results for different algorithms. The lines marked with asterisks are results for the imaginary part of η and lines without asterisks are results for the real part of η.

is a dynamic range of the pixel intensities. k_1 and k_2 are constants with $k_1 = 0.01$ and $k_2 = 0.03$ by default.

5. Results

We compare mWnet_1 and mWnet_4 with three neural networks: Automap, UNet, and FC-DenseNet, and two other traditional reconstruction algorithms that use different optimization methods: Gauss Newton CG [14] and L-BFGS [13] on a laptop with CPU

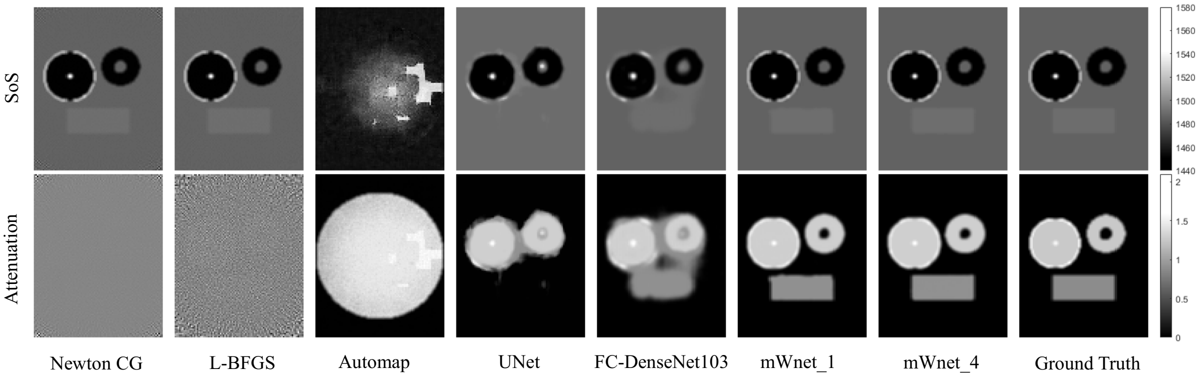


Figure 8: The reconstructed SoS (m/s) and attenuation $(dB/cm/MHz)$ results on Phan 5 with $SNR = 50dB$.

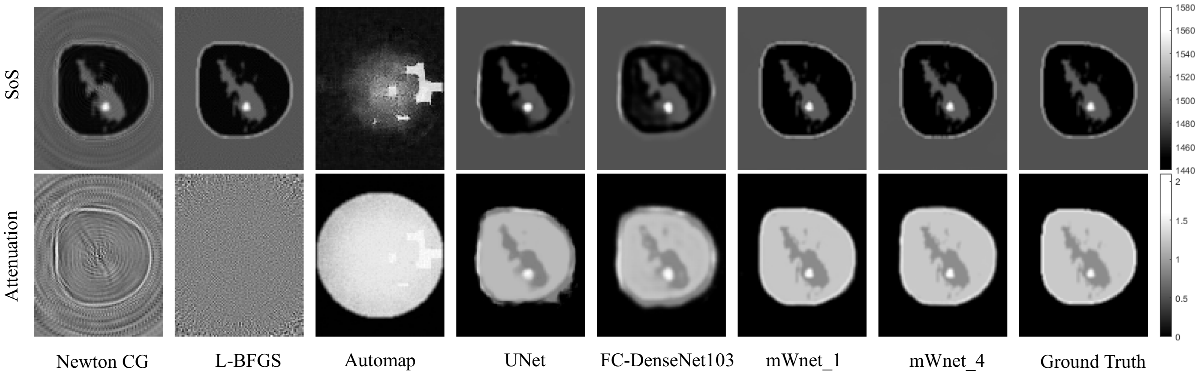


Figure 9: The reconstructed SoS (m/s) and attenuation $(dB/cm/MHz)$ results on Phan 6 with $SNR = 40dB$.

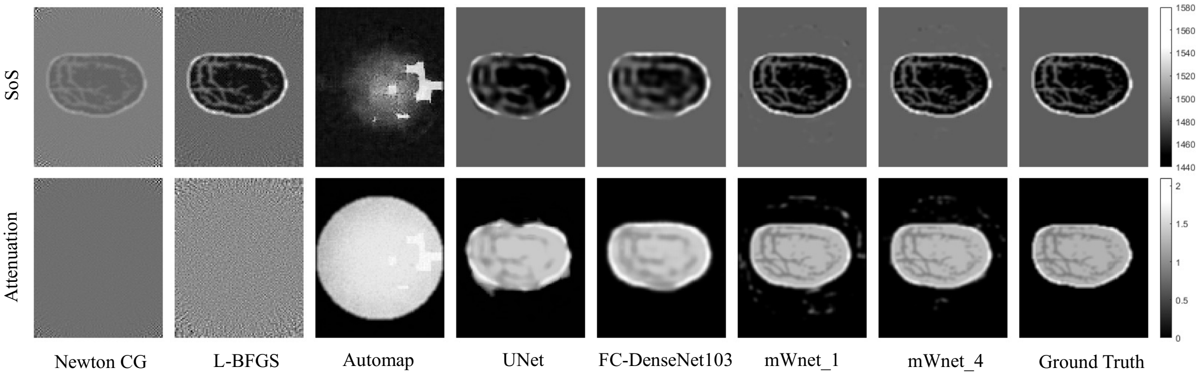


Figure 10: The reconstructed SoS (m/s) and attenuation $(dB/cm/MHz)$ results on img07bx00328 from the OA-Breast Phantom dataset with $SNR = 30dB$.

Intel Core i5 8400 2.80GHz and GPU Nvidia GeForce RTX 2070. All the algorithms are tested with their optimal default settings, where the maximum iteration numbers for Gauss Newton CG and L-BFGS are 500 and 100, respectively.

The learning curves for different neural networks are displayed in Fig. 6. We see that the proposed neural networks converge much faster than other algorithms. Table.

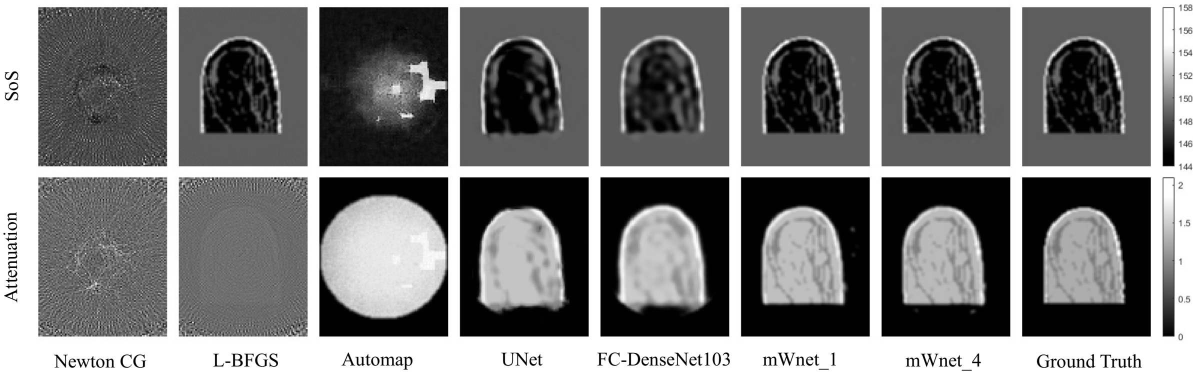


Figure 11: The reconstructed SoS (m/s) and attenuation ($dB/cm/MHz$) results on img07bz00347 from the OA-Breast Phantom dataset without noise.

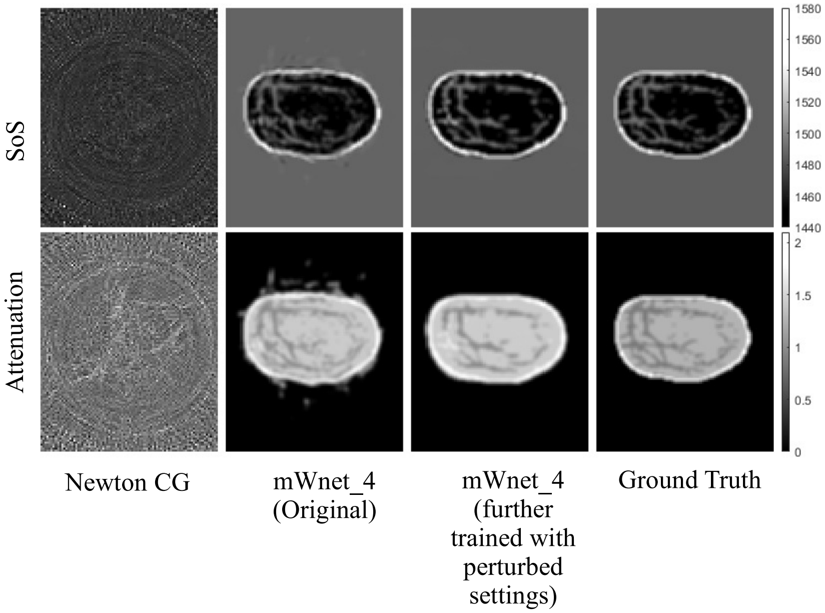


Figure 12: The reconstructed SoS (m/s) and attenuation ($dB/cm/MHz$) results on img07bx00300 from the OA-Breast Phantom dataset simulated with perturbed transducer locations.

1 shows the general training time and number of parameters. The average runtime per image for different algorithms is shown in Table 2. We see that compared with traditional algorithms, deep-learning-based algorithms are much faster on CPU. Running on GPU can further speed up the deep-learning-based reconstruction significantly. Even though the number of parameters of mWnet_4 is about 3 times as large as that of mWnet_1, the increase of runtime of mWnet_4 is less than a factor of two. This is because the additional parameters in mWnet_4 lie only in the DUS units that process the down-scaled tensors.

The quantitative results are presented in Fig. 7 for both the four standard test images and the medical images derived from the OA-Breast Phantom dataset. We have the input data corrupted by Additive white Gaussian noise at 4 different noise levels with the signal-noise-ratio SNR=30dB, 40dB, 50dB and noise-free, respectively. As a whole, the neural network is more robust to noise and can reconstruct both the real part and imaginary part well, while the traditional algorithms can only reconstruct noisy real part and the imaging quality decreases significantly with the increase of noise. In addition, it can be noted that mWnet_4 is more robust to noise than mWnet_1, especially at a high noise level such as $SNR = 30dB$.

The visual results for standard test samples and the medical images at different noise levels are shown in Fig. 8-11, which verify the proposed neural network's superior imaging quality for noisy inputs. Specifically, on the one hand, for the reconstruction of the real part of η, image quality by mWnet is comparable to that by traditional algorithms for noise-free cases, and mWnet is superior to the traditional algorithms at a high noise level (i.e., $SNR = 30dB$ in Fig. 10). On the other hand, for the reconstruction of the imaginary part, mWnet can always achieve a decent visual performance, while traditional algorithms normally fail to reconstruct a meaningful image in the presence of noise.

In addition, mWnet_4 has much better visual performance than mWnet_1 for noisy cases. For example, in Fig.8, the sides of the rectangular are straight in the result for mWnet_4 but are distorted in the result for mWnet_1; in Fig. 10, mWnet_1 has more artifacts than mWnet_4.

Fig.12 shows mWnet_4's robustness to uncertainties in transducer locations compared with Gauss Newton CG. With more training on the dataset simulated with perturbed settings, the mWnet_4's imaging quality is improved further.

6. Discussion and Conclusion

The proposed neural networks show superior imaging quality to any other neural networks including the Automap, FC-DenseNet, and the classical UNet. Among these three neural networks, the Automap is the most inefficient one with the worst imaging quality and the highest number of parameters due to the use of fully connected layers. On the other hand, the fully convolutional neural networks FC-DenseNet and UNet have the smallest number of parameters but only yield a blurred result. Meanwhile, the proposed neural networks show the highest imaging quality and maintain an acceptable inference speed. Compared with the other two popular convolutional neural networks FC-DenseNet and UNet, the factors that lead to mWnet's superior performance in both imaging quality and efficiency are: 1) As demonstrated in the section of Network Architecture and Training Strategy, the proposed neural networks use advanced down-scaling and up-scaling operators, which give more emphasis on the high-frequency part of the data; 2) The proposed neural networks iteratively implement multiple down- and up- scaling operator to gain the number of layers with large receptive and projective field, which also allows the efficient implementation of large number of parameters.

The results show that with more DUS units in the hidden layers, the neural network obtains a higher imaging quality and becomes more robust to noise, which is also confirmed by the works in [32] and [33]. Although the increase of the number of DUS units leads to a significant increase in model size, it is still much smaller than the neural network using fully connected layers, and the computational burden is controlled in an acceptable range. The multiple down-up scaling strategy allows the neural networks to implement more parameters efficiently with much lower computational burden. It is also possible to extend the mWnet to dealing with larger-size 2D or 3D images by putting more scaling operations to both the initial feature extraction part and the reconstruction part. This kind of extension ensures that the size of the feature maps processed by DUS units is within the receptive field and projective field of hidden layers in DUS units. Adding more scaling operations will increase a tiny number of parameters and help control the computational burden within a reasonable range. Apart from the number of scaling operations in the initial feature extraction part and the reconstruction part, the number of DUS units in mWnet should also be changed based on experiments.

The traditional algorithm Gauss Newton CG gives a result much better than the result by L-BFGS for the noise-free test cases. It is because of their different settings on initialization, step length, and iterative number, which make the implementation of Gauss Newton CG more suitable for noise-free cases. Meanwhile, we also see that these conventional algorithms all perform better than mWnet on noise-free cases. One reason for this phenomenon is that Gauss Newton CG and L-BFGS use exactly the same forward model for simulation to iteratively optimize the solution, while neural network only learns the solution indirectly via the dataset generated based on the forward model and ends up yielding an approximate solution. When training dataset getting larger, this gap between the traditional algorithms and neural networks can be narrowed further.

It should be noted that the traditional iterative optimization algorithms (Gauss Newton CG and L-BFGS) are often stuck into various local optimal solutions in the presence of noise or uncertainties in transducer location, while the neural network can obtain an approximation solution that is closer to the global solution easily. Meanwhile, the fully learned neural network approach is also much faster than the iterative optimization approaches. However, in the case of low noise level, the neural network is inferior to traditional algorithms in terms of imaging accuracy. For higher imaging accuracy of the neural network, one solution is to enlarge the training set or to involve medical images into the training set for finetuning. In addition, the imaging quality can be further improved by using the neural network to get a good initialization for traditional optimization.

Since the wave-based transmission tomography has high degrees of scattering due to the long wave length at the scale of the objects, the UTT image reconstruction has higher complexity than Radon inversion in the straight-ray-based tomography (such as X-ray CT) and is of high non-linearity, the favorable performance of the proposed neural network on UTT image reconstruction proves its potential to tackle other image reconstruction problems such as CT and MRI image reconstruction.

In the future work, we will continue investigating how to impove the efficiency of the neural network further to deal with large-size images. We will also test the proposed neural networks on real data from different imaging tasks.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants no. HE 3011/37-1 and HO 5565/2-1, the Fundamental Research Funds for the Central Universities from Donghua University under grants no. 20D111205, and the Young Teacher Research Startup Fund from Donghua University under grants no. 112-07-0053079.

References

[1] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin D M, Forman D and Bray F 2013 *Lyon, France: International agency for research on cancer* **2016**

[2] Fitzmaurice C, Akinyemiju T F, Al Lami F H, Alam T, Alizadeh-Navaei R, Allen C, Alsharif U, Alvis-Guzman N, Amini E, Anderson B O *et al.* 2018 *JAMA oncology* **4** 1553–1568

[3] Ruiter N, Zapf M, Dapp R, Hopp T and Gemmeke H 2012 First in vivo results with 3d ultrasound computer tomography *2012 IEEE International Ultrasonics Symposium* (IEEE) pp 1–4

[4] Glide C, Duric N and Littrup P 2007 *Medical physics* **34** 744–753

[5] Johnson S, Abbott T, Bell R, Berggren M, Borup D, Robinson D, Wiskin J, Olsen S and Hanover B 2007 Non-invasive breast tissue characterization using ultrasound speed and attenuation *Acoustical Imaging* (Springer) pp 147–154

[6] van Dongen K W and Wright W M 2006 *The Journal of the Acoustical Society of America* **120** 2086–2095

[7] Taskin U, Ozmen N, Gemmeke H and van Dongen K W 2018 *Archives of Acoustics* **43**

[8] Kak A C, Slaney M and Wang G 2002 *Medical Physics* **29** 107–107

[9] Dapp R 2013 *Abbildungsmethoden für die Brust mit einem 3D-Ultraschall-Computertomographen* Ph.D. thesis KIT-Bibliothek

[10] Duric N, Li C, Roy O and Schmidt S 2011 Acoustic tomography: promise versus reality *2011 IEEE International Ultrasonics Symposium* (IEEE) pp 2033–2041

[11] Simonetti F, Huang L and Duric N 2009 *Applied Physics Letters* **95** 061904

[12] Althaus L 2016 *MS thesis*

[13] Wang H, Gemmeke H, Hopp T and Hesser J 2019 Accelerating image reconstruction in ultrasound transmission tomography using l-bfgs algorithm *Medical Imaging 2019: Ultrasonic Imaging and Tomography* vol 10955 (International Society for Optics and Photonics) p 109550B

[14] Gemmeke H, Althaus L, Van Dongen K W, Egger H, Hesser J, Mayer J, Ruiter N V, Zapf M and Hopp T 2016 Wave equation based transmission tomography *2016 IEEE International Ultrasonics Symposium (IUS)* (IEEE) pp 1–4

[15] Jin K H, McCann M T, Froustey E and Unser M 2017 *IEEE Transactions on Image Processing* **26** 4509–4522

[16] Han Y and Ye J C 2018 *IEEE transactions on medical imaging* **37** 1418–1429

[17] Zhu B, Liu J Z, Cauley S F, Rosen B R and Rosen M S 2018 *Nature* **555** 487

[18] Häggström I, Schmidtlein C R, Campanella G and Fuchs T J 2019 *Medical image analysis* **54** 253–262

[19] Li Y, Li K, Zhang C, Montoya J and Chen G H 2019 *IEEE transactions on medical imaging*

[20] Yoon Y H, Khan S, Huh J and Ye J C 2018 *IEEE transactions on medical imaging* **38** 325–336

[21] Gao Z, Wu S, Liu Z, Luo J, Zhang H, Gong M and Li S 2019 *Medical image analysis* **58** 101534

[22] Cheng A, Kim Y, Anas E M, Rahmim A, Boctor E M, Seifabadi R and Wood B J 2019 Deep learning image reconstruction method for limited-angle ultrasound tomography in prostate cancer *Medical Imaging 2019: Ultrasonic Imaging and Tomography* vol 10955 (International Society for Optics and Photonics) p 1095516

[23] Ramirez A B and van Dongen K W 2016 *The Journal of the Acoustical Society of America* **140** 1749–1757

[24] Ozmen N, Dapp R, Zapf M, Gemmeke H, Ruiter N V and van Dongen K W 2015 *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* **62** 637–646

[25] Thomson D J and Chapman N 1983 *The Journal of the Acoustical Society of America* **74** 1848–1854

[26] Levy M 2000 *Parabolic equation methods for electromagnetic wave propagation* 45 (IET)

[27] Saad Y and Lee D 1986 *A new algorithm for solving the wide angle wave equation* (Yale University. Department of Computer Science)

[28] Le H and Borji A 2017 *arXiv preprint arXiv:1705.07049*

[29] Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *International Conference on Medical image computing and computer-assisted intervention* (Springer) pp 234–241

[30] Ye J C, Han Y and Cha E 2018 *SIAM Journal on Imaging Sciences* **11** 991–1048

[31] Liu P, Zhang H, Zhang K, Lin L and Zuo W 2018 Multi-level wavelet-cnn for image restoration *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* pp 773–782

[32] Yu S, Park B and Jeong J 2019 Deep iterative down-up cnn for image denoising *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* pp 0–0

[33] Haris M, Shakhnarovich G and Ukita N 2018 Deep back-projection networks for super-resolution *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 1664–1673

[34] Shi W, Caballero J, Huszár F, Totz J, Aitken A P, Bishop R, Rueckert D and Wang Z 2016 Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 1874–1883

[35] Abdelhamed A, Timofte R and Brown M S 2019 Ntire 2019 challenge on real image denoising: Methods and results *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* pp 0–0

[36] Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 4700–4708

[37] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: Surpassing human-level performance on imagenet classification *Proceedings of the IEEE international conference on computer vision* pp 1026–1034

[38] Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L 2009 Imagenet: A large-scale hierarchical image database *2009 IEEE conference on computer vision and pattern recognition* (Ieee) pp 248–255

[39] Audhkhasi K, Osoba O and Kosko B 2016 *Neural Networks* **78** 15–23

[40] Smith S L, Kindermans P J, Ying C and Le Q V 2017 *arXiv preprint arXiv:1711.00489*

[41] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L and Lerer A 2017

[42] Jégou S, Drozdzal M, Vazquez D, Romero A and Bengio Y 2017 The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* pp 11–19

[43] Chen Y, Shaw J L, Xie Y, Li D and Christodoulou A G 2019 Deep learning within a priori temporal feature spaces for large-scale dynamic mr image reconstruction: Application to 5-d cardiac mr multitasking *International Conference on Medical Image Computing and Computer-Assisted*

Intervention (Springer) pp 495–504

[44] Lou Y, Zhou W, Matthews T P, Appleton C M and Anastasio M A 2017 *Journal of biomedical optics* **22** 041015