Delft University of Technology

Simulation-to-real generalization for deep-learning-based refraction-corrected ultrasound tomography image reconstruction

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

DOI
10.1088/1361-6560/acaeed

Publication date
2023

Document Version
Final published version

Published in
Physics in medicine and biology

Citation (APA)
Zhao, W., Fan, Y., Wang, H., Gemmeke, H., van Dongen, K. W. A., Hopp, T., & Hesser, J. (2023). Simulation-to-real generalization for deep-learning-based refraction-corrected ultrasound tomography image reconstruction. *Physics in medicine and biology*, *68*(3), Article 035016. https://doi.org/10.1088/1361-6560/acaeed

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Physics in Medicine & Biology

IPEM
Institute of Physics and Engineering in Medicine

PAPER

Simulation-to-real generalization for deep-learning-based refraction-corrected ultrasound tomography image reconstruction

To cite this article: Wenzhao Zhao *et al* 2023 *Phys. Med. Biol.* **68** 035016

View the article online for updates and enhancements.

You may also like

- Cosmic Acceleration as the Solution to the Cosmological Constant Problem
  Philip D. Mannheim
- The origin of the highest energy cosmic rays
  Peter L Biermann
- Living without supersymmetry—the conformal alternative and a dynamical Higgs boson
  Philip D Mannheim

This content was downloaded from IP address 154.59.124.113 on 06/02/2023 at 13:27

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

RECEIVED
23 June 2022

REVISED
29 November 2022

ACCEPTED FOR PUBLICATION
28 December 2022

PUBLISHED
27 January 2023

PAPER

Simulation-to-real generalization for deep-learning-based refraction-corrected ultrasound tomography image reconstruction

Wenzhao Zhao[1], **Yuling Fan**[1], **Hongjian Wang**[2], **Hartmut Gemmeke**[3], **Koen W A van Dongen**[4], **Torsten Hopp**[3] **and Jürgen Hesser**[5]

[1] Interdisciplinary Center for Scientific Computing (IWR), Central Institute for Computer Engineering (ZITI), Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, D-68167 Mannheim, Germany

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

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

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

[5] Interdisciplinary Center for Scientific Computing (IWR), Central Institute for Computer Engineering (ZITI), CZS Heidelberg Center for Model-Based AI, Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, D-68167 Mannheim, Germany

E-mail: wenzhao.zhao@medma.uni-heidelberg.de

Keywords: deep learning, simulation-to-real generalization, measurement domain, Fourier transform, refraction-corrected ultrasound tomography

Supplementary material for this article is available online

Abstract

Objective. The image reconstruction of ultrasound computed tomography is computationally expensive with conventional iterative methods. The fully learned direct deep learning reconstruction is promising to speed up image reconstruction significantly. However, for direct reconstruction from measurement data, due to the lack of real labeled data, the neural network is usually trained on a simulation dataset and shows poor performance on real data because of the simulation-to-real gap. *Approach*. To improve the simulation-to-real generalization of neural networks, a series of strategies are developed including a Fourier-transform-integrated neural network, measurement-domain data augmentation methods, and a self-supervised-learning-based patch-wise preprocessing neural network. Our strategies are evaluated on both the simulation dataset and real measurement datasets from two different prototype machines. *Main results*. The experimental results show that our deep learning methods help to improve the neural networks' robustness against noise and the generalizability to real measurement data. *Significance*. Our methods prove that it is possible for neural networks to achieve superior performance to traditional iterative reconstruction algorithms in imaging quality and allow for real-time 2D-image reconstruction. This study helps pave the path for the application of deep learning methods to practical ultrasound tomography image reconstruction based on simulation datasets.

1. Introduction

Ultrasound computed tomography (USCT) is a promising tool for non-invasive and non-ionizing medical image diagnosis, especially for breast cancer detection in screening. It has been proven that sound-speed tomograms can help differentiate between different breast lesions and henceforth assess breast cancer risks (Li *et al* 2009).

A popular conventional approach for USCT image reconstruction is ray-based methods (Ozmen *et al* 2015). Ray-based methods reduce the wave propagation model into a ray-propagation problem which reduces the computation burden (Javaherian and Cox 2021). The ray-based methods achieve a good balance between

© 2023 Institute of Physics and Engineering in Medicine

imaging quality and computational cost, and allow handling larger problems with reduced computational efforts, which also helps accelerate the simulation for training data generation. The ray-based methods have been used for practical 3D USCT reconstruction for breast imaging (Birk *et al* 2014, Hopp *et al* 2014), where the TVAL3-based iterative reconstruction is combined with GPU acceleration to speed up the imaging process. More recently, the Bézier curve technique (Perez-Liva *et al* 2020, Zuch *et al* 2021) is introduced to further accelerate the bent-ray-based reconstruction.

With more computing power available, full-waveform inversion (FWI), an imaging method originally developed in the field of seismic exploration, has been intensively studied for ultrasound tomography. Requiring a heavy computational burden, FWI models both transmission and reflection and can make full use of the measurement data. It is believed that FWI has the potential for high resolution image reconstruction (Lucka *et al* 2021). However, to get a high fidelity reconstruction, FWI needs good initialization and low frequency information(a few hundreds of kHz) (Agudo *et al* 2018) to avoid cycle skipping (Robins *et al* 2021, Boehm *et al* 2022). This low frequency information is often unavailable in conventional ultrasound tomography machines.

Besides FWI, another seismic imaging technique, finite frequency traveltime tomography (Mercerat and Nolet 2013) has also been introduced for ultrasound tomography (Martiartu *et al* 2020). This method considers the frequency dependence and volumetric sensitivity of traveltime measurements and shows decent performance on 2D ultrasound tomography image reconstruction with real measurement data.

In recent years, there has been a growing interest in deep learning methods to allow for real-time USCT image reconstruction, which can be divided into two categories: hybrid approach and fully-learned approach. The hybrid approach aims at integrating deep learning methods with traditional iterative reconstruction methods to achieve faster reconstruction (Poudel *et al* 2019, Robins *et al* 2021, Stanziola *et al* 2021, Fan *et al* 2022). On the other hand, the fully learned approach tries to reconstruct images via end-to-end learning with deep learning methods from measurement data (Prasad and Almekkawy 2020, Zhao *et al* 2020). Due to the lack of real measured data with ground truth information, simulation data are typically used for training these neural networks. However, in practice, the neural networks trained with simulation datasets usually show poor performance on real data due to the subtle difference between simulation and real data, known as the simulation-to-real gap.

The discrepancy between simulation and real measurement data is unavoidable and originates from different reasons. One key part is systematic errors (Taylor 1997) including positioning errors of transducers, the time delay error between emitters and receivers, etc (Filipik 2008, Tan *et al* 2015), which can be reduced by calibration but cannot be eliminated completely. Random errors from such as temperature fluctuation can be another error source. In addition, the approximation methods used in simulation such as defining each sensor as a single point can also aggravate the gap between simulation and real data. Yet, in practice, a more realistic simulation usually requires a huge computational burden. For refraction-corrected transmission tomography, the estimation of time of flight (ToF) is also an important error source. In our simulation, given the distribution of emitters and sensors, and the speed of sound map, the ToF can be simply obtained via the ray-based forward model. Yet, the real measurement data are usually time series data recorded by receivers, where arrival time estimation algorithms should be applied in order to obtain the measurement data. In this study, we adopt a state-of-the-art arrival time estimation method, referred to as the sliding-window weighted Akaike information criterion method (Bao and Jia 2019).

The common problems of imperfect instrument calibration, arrival time estimation errors, and random perturbance in the real imaging process of USCT cause uncertainties in the measurement data. These uncertainties constitute a significant part of the simulation-to-real gap, and can pose a substantial source of errors for neural networks that use simulation data for model training. However, few research works have focused on this aspect of deep-learning-based USCT image reconstruction so far.

Deviations between simulation and measurement can be mitigated by making the simulation more accurate relative to the real setup (system identification), performing a domain adaptation, and domain randomization (Tobin *et al* 2017, Peng *et al* 2018). However, the system identification and calibration are expensive and error-prone (Tobin *et al* 2017). Thus, one focus of our work is to develop domain adaptation and randomization techniques for deep-learning-based USCT image reconstruction. In this paper, we do not attempt to improve the imaging process from the hardware aspect or to improve a specific simulation algorithm. Instead, we consider a common, realistic, and deep-learning-related scenario, where given the limited and imperfect measurement data obtained from the real imaging process and imprecise system parameters, we aim to achieve a decent data-driven image reconstruction with a neural network that is trained only on a simulation dataset generated by an efficient simulation algorithm (in this paper we adopt Eikonal-equation-based fast marching algorithm Hassouna and Farag 2007). In other words, we emphasize improving deep-learning methods' simulation-to-real generalizability. To achieve this goal, we consider the process of developing a deep learning model, and investigate three strategies in literature: simulation data generation, measurement domain generalization, and neural network architecture.

(1) In simulation data generation, a large dataset with image sources of high diversity is essential for training a deep learning model with high generalizability (Jush *et al* 2022). In this study, we use image sources from the ImageNet dataset to guarantee diversity. In contrast to increasing the diversity of datasets, there is a large volume of research works in computer vision that tries to develop image style transfer techniques to create a realistic dataset to the target image domain (Csurka 2017, Peng *et al* 2018, Yue *et al* 2019, Farahani *et al* 2021). This strategy is also adopted in recent works on ray-casting ultrasound simulation (Feng *et al* 2021) and seismic data simulation (Vitale *et al* 2020), where researchers try to improve the realism of simulation by applying generative deep learning models in the image domain. The image style transfer allows reducing the problem to a certain clinical application for a certain organ such as the human brain or lung, but may not improve the generalizability of the image reconstruction for objects of any kind of structure. Since deep learning methods are data-driven and the possible physical structure of real phantoms can be extremely diverse, the diversity of data sources helps avoid overfitting and improve the generalizability of neural networks from simulation data to real data (Jush *et al* 2022).

(2) For deep-learning-based end-to-end image reconstruction, the measurement data as input to neural networks has a direct influence over the neural network parameters after training henceforth to the output. Yet, few research works focus on this aspect. Meanwhile, in computer vision, data augmentation (Volpi *et al* 2018, Hendrycks *et al* 2019, Zeng *et al* 2020, Li *et al* 2021) and data preprocessing (Qiu and Qiu 2020, Haque *et al* 2021) techniques are frequently used to improve the neural networks' generalizability and robustness. However, these techniques are all targeted toward the image domain. In the measurement domain, the uncertainties in the real world due to imperfect calibration and random errors need to be considered. In this paper, we focus on data augmentation strategies and deep-learning-based data preprocessing techniques for measurement domain generalization.

(3) The architecture of neural networks can affect neural networks' robustness (Devaguptapu *et al* 2021). It has been shown that increasing the depth and the number of parameters can help improve the robustness (Madry *et al* 2017, Xie and Yuille 2019). However, it can lead to a higher computation burden of neural networks. In this paper, we propose to develop a Fourier-transform-integrated neural network to help improve the robustness and generalizability of the neural network without increasing the number of parameters.

The rest of the paper is organized as follows. We introduce the background of the reconstruction problem and our method in the Materials and Methods section. In the Experiments and Results section, we describe the experiment setting and show our experimental results on both simulation and real data. The further discussion and final conclusion are in the Discussions section and the Conclusions section.

2. Materials and methods

2.1. Problem formulation and forward model

Our approach adopts ray-based wave propagation forward model for quick simulation data generation. The acoustic wavefront propagation with inhomogeneous sound-speed distributions v can be modeled by Eikonal equation (Hassouna and Farag 2007):

$$\|\nabla t(\mathbf{x})\| = \frac{1}{v(\mathbf{x})}, \tag{1}$$

where $t \in \mathbb{R}_+$ is travel time, ∇denotes the gradient, $\|\cdot\|$ is the Euclidean norm, and $v\colon \Omega \to \mathbb{R}_+$ represents the sound speed at location $\mathbf{x}$ in the considered domain Ω. Based on this equation and the sound-speed distribution v of the refractive medium, we trace rays across the medium efficiently by fast marching methods (FMM) (Hassouna and Farag 2007), and get the ToF T_{er} from wave sources e to receivers r.

As in a USCT imaging system, we consider the location of each sender-receiver pair to be fixed, generally, we can denote the Eikonal-equation-based forward operator as $\mathcal{T}\colon X \to Y, v \mapsto T$; with $v \in X, T \in Y; X, Y \in \mathcal{H}$ (Hilbert Space). Thus the USCT reconstruction problem can be formulated as the minimization of the following objective functional:

$$\mathcal{J}(v) = \|\mathcal{T}(v) - T_{\text{obs}}\|, \tag{2}$$

where T_{obs} is the observed travel time at receivers. In traditional iterative reconstruction algorithms, smoothness regularization terms are often introduced to improve imaging quality. In this case, the objective function can be rewritten as

$$\mathcal{J}(v) = \|\mathcal{T}(v) - T_{obs}\| + \lambda\mathcal{R}(v), \tag{3}$$

where $\lambda \in \mathbb{R}_+$ is the regularization parameter, and the regularization functional $\mathcal{R}(\cdot)$: $X \to \mathbb{R}_+$ can be the Laplace operator (Ali *et al* 2019), or, alternatively one could choose the Total Variation (Li *et al* 2013) for smoothing the velocity field (Ozmen *et al* 2015).

In our fully learned approach, we aim at training a deep neural network to achieve a direct mapping from observed measurements T_{obs} to the sound-speed distribution v, i.e. $v = \mathcal{T}^{-1}(T_{obs})$.

2.2. Fourier-transform-integrated convolutional neural network

Recent publications in computer vision have demonstrated that Fourier transform and frequency information are useful for improving machine learning model's robustness and domain adaptation (Yin *et al* 2019, Yang and Soatto 2020). Yet, they are simply using Fourier transform as a data preprocesssing or post-processing method. In this paper, we propose to integrate the Fourier transform directly between layers of neural network to improve convolutional neural networks' robustness and generalizability for image reconstruction. We present a Fourier integrated convolutional neural network named as 'split-step Fourier convolutional network', i.e. SSFnet by inserting the fast Fourier transform (FFT) layers into the residual layers of a U-shaped convolutional neural network. As shown in figure 1(c), in the implementation, we separately insert the 2D FFT and inverse 2D FFT in two positions of a residual neural network block, where only the real part of outputs of FFT and inverse FFT are used in the following layers. Discrete cosine transform can be another option, which can give a similar performance. Yet for efficient implementation, in this paper, we only use FFT to exploit the global frequency information. Within the residual block, the skip connection between the frequency domain and the non-frequency domain helps the neural network to combine information from both domains. In figure 1(b)–(c), N represents the number of channels. Empirically, we set $N = 256$ for the reconstruction network and $N = 128$ for the patch-wise preprocessing network.

The overall architecture of our neural network is shown in figure 1. The basic convolutional unit is a convolution layer with a kernel size of 3 × 3 followed by PReLU activation function (He *et al* 2015, Zhao *et al* 2020). The down-sampling layer is implemented by changing the stride size to 2. We adopt the sub-pixel convolutional unit (Shi *et al* 2016) for the up-sampling layer because of its efficiency. In general, the neural network has three parts: the encoding part, the remapping part in the latent space, and the decoding part. The encoding part uses multiple down-sampling layers to reduce the feature size, which allows the following reconstruction to be realized more efficiently. The remapping part contains most of the parameters and achieves the mapping between the measurement domain and the image domain. Finally, the decoding part up-samples the remapped features and reconstructs the images. The number of parameters of the neural network in total is about 27.1×10^6. In the following parts of the paper, we denote the reconstruction neural network (ReconNN) with FFT integrated as SSFnet, and the corresponding version without FFT as CNN.

2.3. Measurement data augmentation and preprocessing for simulation-to-real domain generalization

There are various factors that can lead to a gap between real data and simulation data. If considering water-only measurement data, we can find that there is a difference between the estimated ToF t_w via ToF estimation based on time series data, and the ToF calculated by $t_c = d/c$ with d the distance between emitters and receivers, c the speed of sound in water. This can be caused by various factors: random noise, the error in sensor positions, the temperature change of the environment, and outliers by various anomalous causes.

In this paper, we apply specific data augmentation to improve the generalization of the neural network. Considering the above-mentioned factors that affect the generalization of the neural network, we perform the following data augmentation strategies.

(1) Random noise (RN): To improve the robustness to random noise, we add random Gaussian noise $\mathcal{N}(\mu, \sigma^2)$ with variance σ^2 and a mean value μ that follows uniform distribution in the range of $[-u_n, u_n]$. Here all the measurement data has been normalized to a range of [0, 1] beforehand. Both σ and u_n should be small because heavy noise will result in a performance drop of neural networks. We empirically set $\sigma = 0.020$, and u_n=0.035.

(2) Sensor position perturbance (SPP): To improve the robustness against the error of sensor geometry, we further generate another dataset by adding random uniform noise to the sensor position in the range of $[-u_b, u_b]$ mm. In practice, u_b should be close to the physical size of the transducer's element pitch. In this study, the target machine has d_{pitch} mm pitch. As we also consider the displacement error of transducers, we set a larger sensor position perturbance with $u_b = d_{\text{spp}}$ mm.

(3) Semi-random bias (RB): We also consider the calibration based on the water-only measurement data. We add adjust bias $\delta t = (t_c - t_w) \cdot \alpha$ with α a uniform random variable in the range of [0.7, 1.3]. It should be

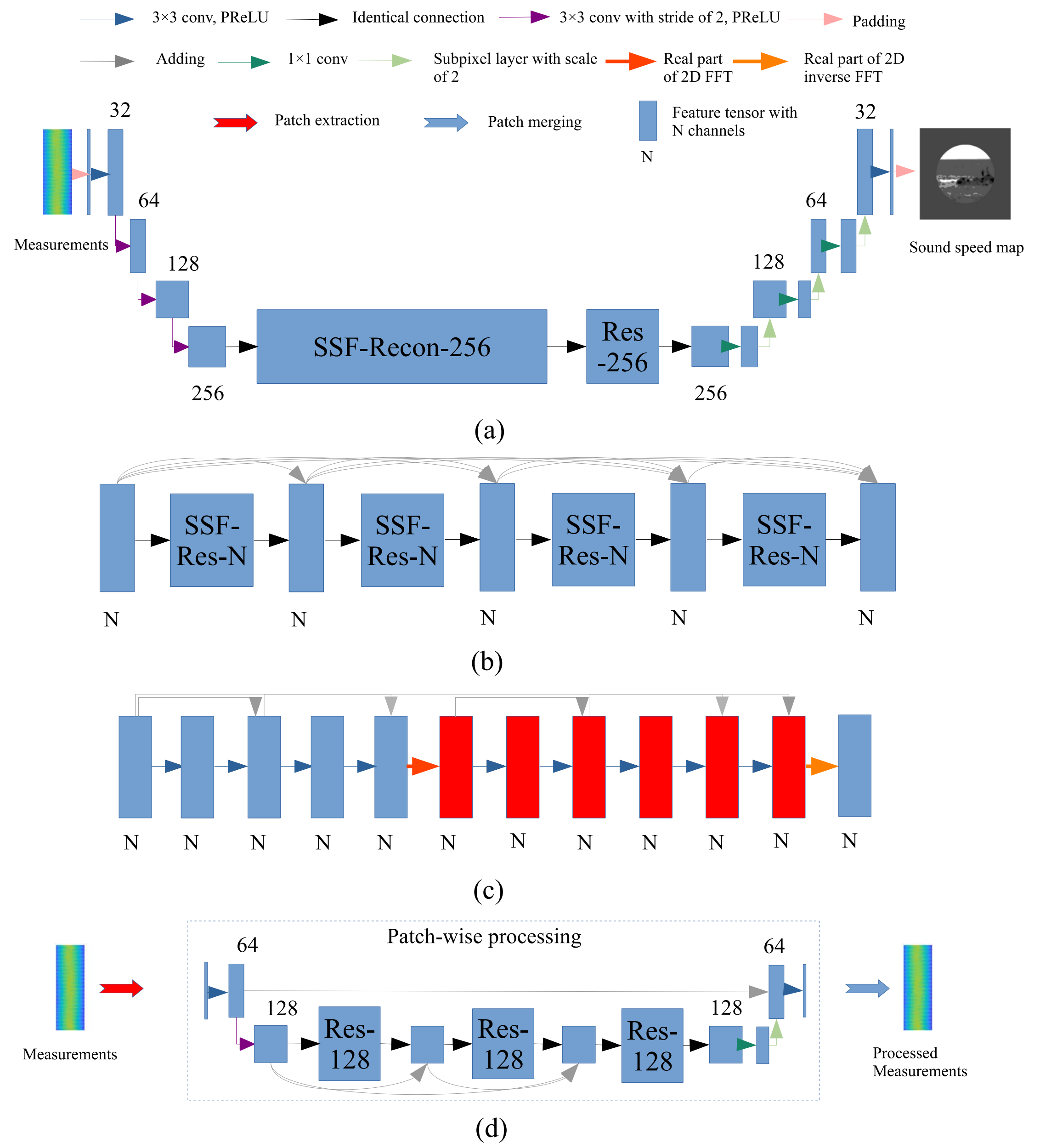

Figure 1. The architecture of the proposed network. (a) The overall architecture of SSFnet (The version without FFT is denoted as CNN); (b) The SSF-Recon-N block with N the number of channels; (c) The SSF-Res-N residual block (Replacing 2D FFT layers with identical connections yields a plain residual block Res-N); (d) The preprocessing neural network PNN for outliers removal.

noted that we do not include the water-only measurement data into the training dataset directly. We only use the water-only measurement for the data augmentation with RB.

(4) Besides data augmentation, data preprocessing is also a promising direction for improving the generalization of neural networks. Inspired by the recent work in domain adaptation based on data reconstruction (Ghifary *et al* 2015), we try to use deep-learning-based preprocessing model to reconstruct the measurement data with learned robust features and thus also help remove the outliers. On the other hand, recent research works in computer vision also show that the random mask or pixel deletion training procedure helps improve neural networks' robustness against adversarial noise (Globerson and Roweis 2006, Naveed 2021, He *et al* 2022, Xu *et al* 2022). Based on the above consideration, in this paper, we design a patch-wise preprocessing neural network (PNN) as shown in figure 1 (d). A self-supervised learning approach is applied based on random pixel deletion (RPD), which can help the neural network give more focus on the global structure instead of the local changes and thus help to learn robust features. Specifically, we first split the ToF measurement data I into small patches of size 16×16, and then we randomly remove some areas by assigning null values to relevant pixels and get I_{RPD}. We train the PNN to recover the area with self-supervised loss function $L_s = |I - \hat{I}|$ with $\hat{I}$ the output of PNN, and $| \cdot |$ the l_1 loss. After pretraining, the PNN is connected to the

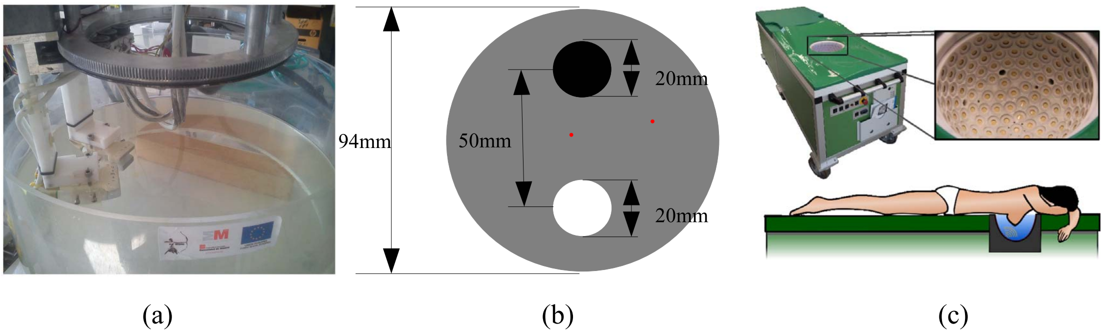

Figure 2. Measurement data acquisition from two different prototype machines. (a) The MUBI system (Ozmen *et al* 2015, Camacho *et al* 2018, Perez-Liva *et al* 2020, Usct data 2022); (b) schematic superior View of the gelatin breast tissue-mimicking phantom; (c) The KIT 3D USCT system (Ruiter *et al* 2012, 2017). Reproduced with permission from 2d CSIC/UCM USCT. The data is freely available dual licensed under the 3-clause BSD-license and the "Open Data Commons Attribution License", in accordance with the "Berlin Declaration".

reconstruction network by using the output of PNN as the input of the reconstruction network, and we finetune the whole network for 6 epochs with combined loss function $L_c = 0.3 \cdot L_s + 0.7 \cdot L_r$, where $L_r = |O - \hat{O}|$ is the reconstruction loss with O the ground truth image, and $\hat{O}$ the output of the reconstruction network.

All considered neural networks are trained using Adam optimizer with a constant learning rate 1×10^{-4} and batch size 16. The training has two stages: (1) Pretraining stage: PNN self-supervised training with RPD for 3 epochs; ReconNN training for 21 epochs (ReconNN). (2) Robust training stage: PNN+ReconNN training with the combinations of RN, RB, and SPP for N_{train} epochs (ReconNN+PNN+RN+RB+SPP), where considering each training sample, we have a 70 percent chance that a data augmentation like RN or RB is applied. It should be noted that all the parameters including the number of epochs are determined empirically. For different machines with different geometric designs, the parameters need to be changed accordingly to achieve the best performance.

2.4. Prototype machines

Our experiments consider two different prototype machines: the Multimodal Ultrasound Breast Imaging (MUBI) system (Medina-Valdés *et al* *2015*, Ruiter *et al* 2017, Camacho *et al* 2018, Perez-Liva *et al* 2020) and the KIT 3D USCT system (Ruiter *et al* 2012).

The MUBI system is shown in figure 2. The system performs circular scans by two 3.5 MHz and 128 elements moving arrays (0.22 mm pitch, P2-4/30EP, Prosonic, Korea). These arrays move around in a water tank of 95 mm radius with an angular resolution of 0.1°. A total of 23 fan beams are obtained. For each fan beam, the emitter array is fixed at a certain position around the circle, while the receiver moves around the tank circle to receive wave signals at 11 different positions. For both emitter and receiver arrays, only 1 out of every 8 array elements was used. In this way, we have 16 × 11 × 16 A-scans for each fan beam.

To estimate the ToF for each A-scan, we adopt the sliding-window-based arrival time estimation method as in (Bao and Jia 2019). Since the waveform in water-only measurement data differs from the waveform in the measurement data for gelatine phantom, we did not use the cross-correlation phase correction in the final step of arrival time estimation. We set the size of the sliding window $w = 40$ empirically.

KIT 3D USCT II system has a semi-ellipsoidal 3D aperture with a diameter of 26 cm and a height of 16 cm. Wavefronts are generated by each emitter at 2.5 MHz (bandwidth 1.5 MHz). The transducers (emitters or receivers) have opening angles of 38.2 deg(standard deviation 1.5 deg). There are 628 emitters and 1413 receivers in total, which are divided into 157 transducer array systems (TAS) with each TAS consisting of four emitters and nine receivers. Virtual positions of the ultrasound transducers can be created by rotational and translational movement of the complete sensor system. For the selected columnar gelatine phantom object, the measurements for 10 different movements of the sensor systems are available. For 2D imaging experiments with the columnar gelatine phantom, we select 24 TAS at the upper part of the 3D aperture that are approximately arranged around a horizontal circle. Therefore, 96 emitter and 216 receivers are used.

We use the same arrival time estimation methods as the experiment with MUBI system. However, due to the strong reflection waves observed, to eliminate their influence and to focus on transmission waves as much as possible, we set the time range of interest (t_s, t_e) for the signal at receivers according to the distance d_{RE} between a pair of receiver and emitter. Since the average sound speed of breast tissue is in the range of (1400, 1700) m s^{-1},

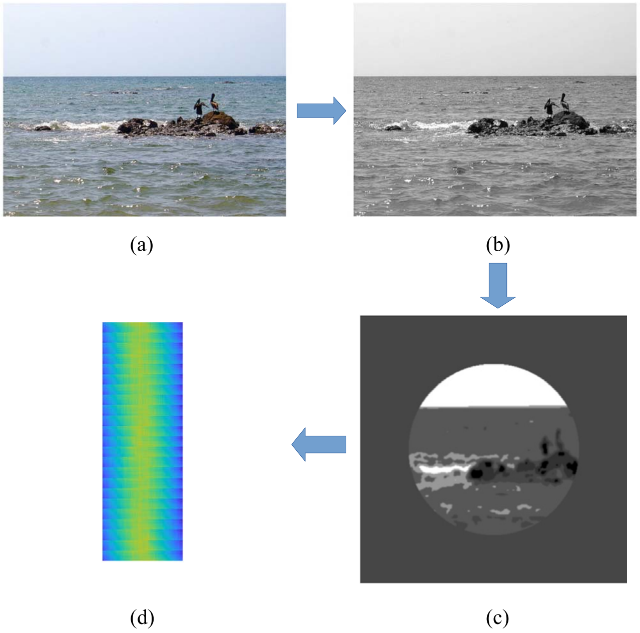

Figure 3. The procedure of simulation data generation. (a) The source RGB image from ImageNet dataset, Reproduced with permission from Image-net.org.; (b) The Y-channel luminance extracted from the RGB image, Adapted with permission from Image-net.org; (c) The sound speed map generated based on the luminance image; (d) The ToF measurements generated by Eikonal-equation-based forward operator.

we empirically set $t_s = d_{RE}/1700$, and $t_e = d_{RE}/1400 + 400dt$ with $dt = 5.0 \times 10^{-8}$ s the sampling interval of signal. The arrival time estimated based on water-only measurements is used for calibration. The final arrival time estimation is the average of estimation results for measurements at all the 10 different movements.

2.5. Simulation data generation

To train our neural network to reconstruct the real measurement data from the prototype machines, we generate simulation data with the same geometrical structure parameters as those of the corresponding prototype machines. We adopt the ray-based forward model as described in section 2.1: Problem formulation and forward model. We consider each used element in the emitter and receiver array as a single point.

The synthetic sound-speed maps used for our simulation are derived from natural images. We collect 49 998 natural RGB images from ImageNet dataset (Deng *et al* 2009) as source images for simulation, where 47 998 images are used for the training set, and 2000 images are for the validation set. The RGB color images from ImageNet are converted to grayscale images with pixel value $x \in [0, 255]$ by extracting the Y-channel luminance. To further enlarge the source images for the training set, we perform two augmentation operations: grayscale-value reversing and 90-degree rotation, which ends up yielding 47 998 × 4 source images. We scale the source images to a size of 128 × 128, and use a Gaussian filter of size 3 × 3 to smooth the source images, which are then scaled to a size of $H \times W$.

It has been suggested that in the breast, fat and some glands often have slower sound speed than water, while blood, skin, and tumors often have higher sound speed values (Hendee and Ritenour 2003, Tissue properties–speed of sound 2022). Thus, we split a natural image into six areas with different sound speed values accordingly. Specifically, the pixel values of all the source images are scaled to a range of [0, 6]. For each source image, two sound-speed values x_1 and x_2 in the range of $[c_L, c_w]$ $\mathrm{m \cdot s^{-1}}$ and three other sound-speed values x_4, x_5 and x_6 in the range of $[c_w, c_U]$ $\mathrm{m \cdot s^{-1}}$ are generated randomly following the uniform distribution in the respective range. We set $x_3 = c_w$ $\mathrm{m \cdot s^{-1}}$ as sound speed in water. We assign sound speed x_i in $\mathrm{m \cdot s^{-1}}$ to the image pixels in the range of $(i - 1, i]$, $i = 1, 2, \ldots, 6$, and 0-value pixel has sound speed x_1 $\mathrm{m \cdot s^{-1}}$. In this way, we obtain a sound-speed map for each source image. Let both the horizontal and vertical distance between adjacent pixels be d_x m. We consider a circle area of radius $r = R$ m to be the region of interest (ROI) and the area outside of the ROI is considered to be water. The ROI size is chosen based on both the radius of sensor arrangement and the size of the phantom. Empirically, we chose the value between them but closer to the radius of the sensor arrangement. Figure 3 shows an example of the whole procedure for generating simulation measurement data from natural images.

For two different prototype machines, empirically, we set different parameters (as shown in table 1) to yield optimal performance. According to the literature (Hendee and Ritenour 2003, Tissue properties–speed of

Table 1. The parameter settings for two different prototype machines.

Machine	H	W	c_w	c_L	c_U	d_x	R	d_{pitch}	d_{spp}	N_{train}
MUBI	417	417	1479.7	1430.0	1600.0	4.80×10^{-04}	0.064	0.22	0.9	6
KIT 3D USCT II	403	403	1483.0	1430.0	1600.0	7.20×10^{-04}	0.120	NULL	1.4	21

sound 2022), we set the range of the sound speed in breast tissues as (1430.0, 1600.0) m s^{-1}. It should be noted that for KIT 3D USCT II, since it needs more epochs, to reduce the training time, we further scale down the size of images to $H = 256$ and $W = 256$ as the target output of neural networks.

For testing purposes, two additional synthetic phantoms A and B are used. Phantom A contains regular geometric patterns. Phantom B is based on a breast model derived from an MRI scan of a cancerous breast containing cancerous tissue (Bakker *et al* 2009). Different tissues are assigned with different sound speeds.

3. Experiments and results

Two neural networks DeepPet (Häggström *et al* 2019) and mWnet (Zhao *et al* 2020) are included in our experiments. Both of these two neural networks are originally designed for end-to-end fully learned image reconstruction from measurement data, and have achieved decent performance for PET image reconstruction, and paraxial-approximation-based ultrasound tomography image reconstruction. Since the mWnet is originally designed for inputs of size $c \times 128 \times 128$ with c the channel number, we do size adaptation by first padding the border of the input measurement data of size $1 \times m \times n$ into the size of $1 \times (128ceil(m/128)) \times (128ceil(n/128))$ with $ceil(\cdot)$ the ceil function, and then reshape it into the size of $(128ceil(m/128))(128ceil(n/128)) \times 128 \times 128$. The output size is scaled back to the target size via bilinear interpolation.

We compare deep learning methods with the TV-based iterative algorithm TVAL3 (Li *et al* 2013), which is the state-of-the-art algorithm for ray-based USCT image reconstruction (Dapp 2013, Birk *et al* 2014). We set the optimal parameters for TVAL3 by grid search, where the tests on the simulation dataset and real dataset use different parameters for TVAL3's optimal performance. The result of the finite-frequency method (Martiartu *et al* 2020) on image reconstruction for the gelatin phantom with MUBI system is also considered for visual comparison.

We implement all the compared deep learning methods with Pytorch and follow the same training routine as described in the section of Measurement data augmentation and preprocessing for simulation-to-real domain generalization. We use the root-mean-square error (RMSE) and structure similarity (SSIM) (Wang *et al* 2004) to measure the imaging quality of algorithms. Compared with RMSE, SSIM is more consistent with the visual perception of human eyes in general.

3.1. Ablation experiments on Fourier-transform-integrated neural network

To investigate the optimal use of Fourier transform in the network layers, we perform an ablation study on different Fourier transform setups. Specifically, we have 'Full': Fourier transform is applied at all the four residual blocks as shown in figure 1; 'Forward only': only forward Fourier transform is kept; 'Inverse only': only inverse Fourier transform is kept; 'Double forwards': we replace the inverse Fourier transform as forward transform; 'P1': from the left side, only the first SSF residual block is kept; 'P1-2': from the left side, only the first two SSF residual blocks are kept; 'P1-3': from the left side, only the first three SSF residual blocks are kept. The quantitative results are shown in table 2. And the visual results are shown in figures 4–6. On simulation data, 'P1-2' and 'Inverse only' give the worst performance. However, on real measurement data, 'P1-2' gives the best results visually, which implies that a good result on noise-free simulation data does not guarantee a good result on real data. In the following experiments, we test further the 'Full' (SSFnet v1) and 'P1-2' (SSFnet v2) setups to see how other different simulation-to-real generalization strategies affect the neural networks' performance.

3.2. Results on simulation data

The results on noise-free simulation test set are shown in figures S1-S2 for SSFnet v1 and figures S3-S4 for SSFnet v2 in the supplementary document. The corresponding quantitative results are shown in table 3. We see that the proposed deep learning methods (SSFnet and CNN) have similar performance. All the deep learning methods are significantly better than the TVAL3 method. Deep learning methods are able to learn deep priors from natural image phantom data from ImageNet. This helps the neural networks achieve superior performance. The best RMSE and SSIM results are obtained by RB, which is probably because RB is a relatively weak way to add noise to the dataset and thus allows the neural network training converging faster than RN, SPP, and PNN with RPD.

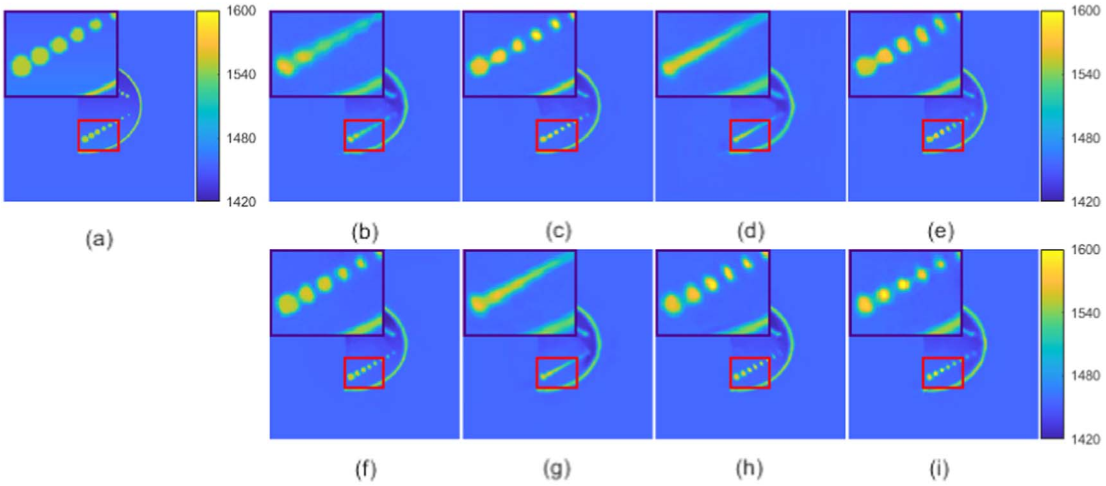

Figure 4. The sound-speed image reconstruction results ($m \cdot s^{-1}$) on noise-free simulation phantom A. The simulation configuration is based on MUBI. (a) Ground truth; (b) Full; (c) P1; (d) P1-2; (e) P1-3; (f) Forward only; (g) Inverse only; (h) Double forwards; (i) CNN.

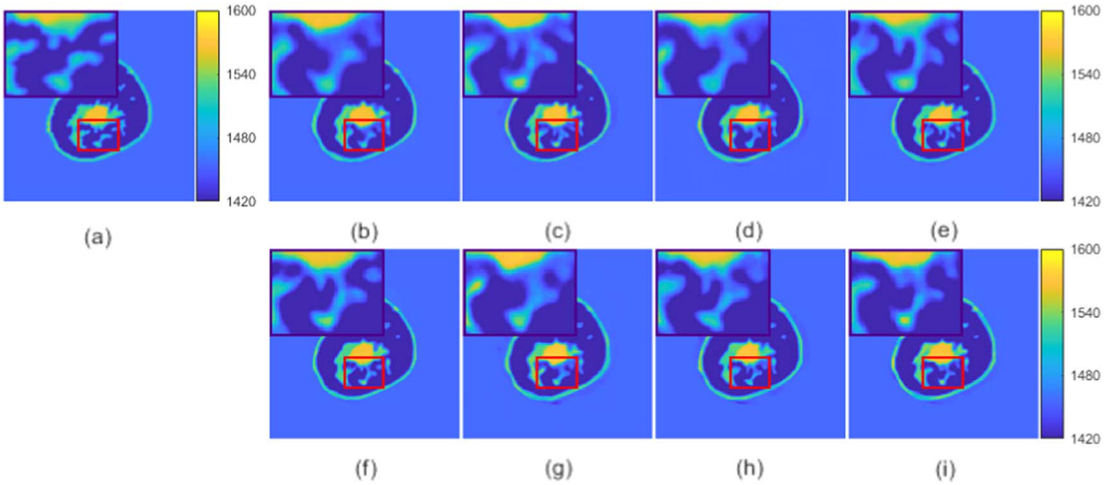

Figure 5. The sound-speed image reconstruction results ($m \cdot s^{-1}$) on noise-free simulation phantom B. The simulation configuration is based on MUBI. (a) Ground truth; (b) Full; (c) P1; (d) P1-2; (e) P1-3; (f) Forward only; (g) Inverse only; (h) Double forwards; (i) CNN.

Table 2. The RMSE and SSIM results on noise-free simulation test images with Fourier transform setups. The simulation configuration is based on MUBI. The SSIM scores are shown in brackets.

	Phantom A	Phantom B
Full	5.258(0.9617)	4.712(0.9488)
Forward only	4.415(0.9586)	4.018(0.9581)
Inverse only	5.470(0.9506)	5.431(0.9386)
Double forwards	5.136(0.9541)	4.105(0.9550)
P1	4.286(0.9715)	4.569(0.9541)
P1-2	6.245(0.9456)	4.983(0.9470)
P1-3	4.334(0.9735)	4.561(0.9540)

The results on simulation test set with sensor position perturbation are shown in figures S5-S6 for SSFnet v1 and figures S7-S8 for SSFnet v2 in the supplementary document, where uniform noise in the range of [−1.8, 1.8] mm is added to sensor positions. The corresponding quantitative results are shown in table 4. It should be

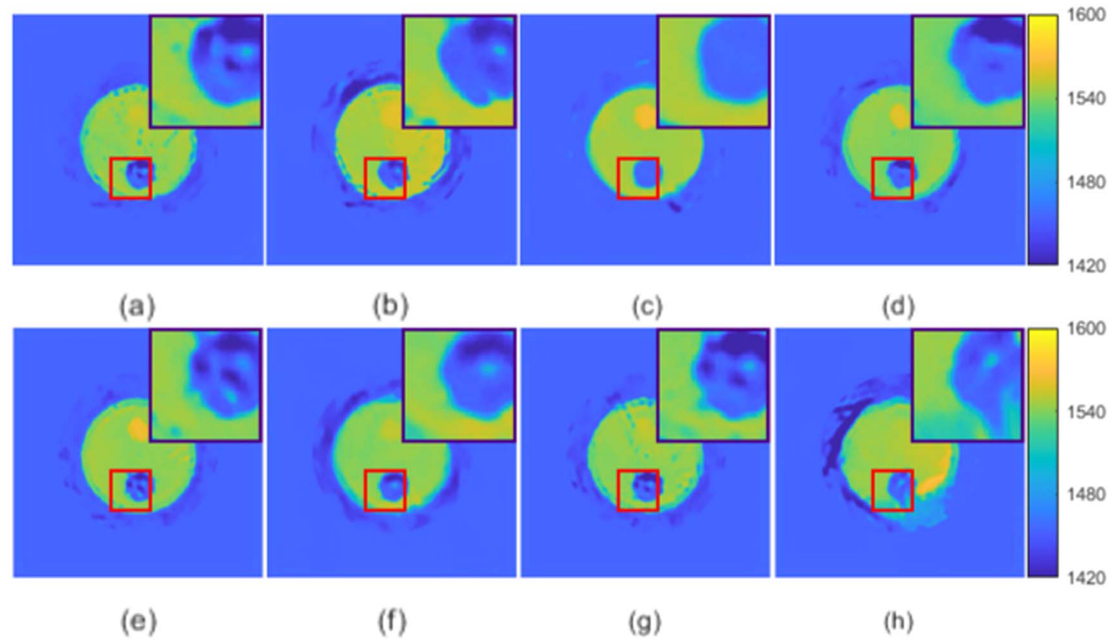

Figure 6. The sound-speed image reconstruction results (m · s^{-1}) on real measurement data with different Fourier transform setups. The simulation configuration is based on MUBI. (a) Full; (b) P1; (c) P1-2; (d) P1-3; (e) Forward only; (f) Inverse only; (g) Double forwards; (h) CNN.

Table 3. The RMSE and SSIM results on noise-free simulation test images with different algorithms and training strategies. The simulation configuration is based on MUBI. The SSIM scores are shown in brackets. Bold font indicates the best results for each phantom.

TVAL3	Phantom A			Phantom B		
	8.327(0.9179)			10.38(0.8143)		
	CNN	SSFnetv1	SSFnetv2	CNN	SSFnetv1	SSFnetv2
Baseline	5.190(0.9607)	4.728(0.9658)	6.245(0.9456)	4.337(0.9519)	4.311(0.9531)	4.983(0.9470)
RN	5.295(0.9521)	5.016(0.9582)	5.681(0.9575)	4.127(0.9551)	4.750(0.9527)	5.542(0.9422)
RB	4.846(0.9590)	**4.356(0.9732)**	5.273(0.9606)	**3.897**(0.9560)	4.263(0.9564)	4.209(**0.9572**)
SPP	5.381(0.9516)	4.499(0.9648)	4.555(0.9701)	4.642(0.9481)	4.275(0.9531)	4.743(0.9478)
RN+RB+SPP	5.791(0.9539)	5.050(0.9523)	6.020(0.9547)	5.960(0.9234)	4.891(0.9481)	5.893(0.9320)
PNN	5.251(0.9559)	4.931(0.9568)	5.007(0.9635)	4.040(0.9539)	4.116(0.9570)	4.426(0.9540)
PNN+RN	5.948(0.9479)	5.648(0.9445)	6.012(0.9340)	5.198(0.9426)	4.657(0.9493)	5.640(0.9414)
PNN+RB	5.045(0.9649)	4.573(0.9659)	5.588(0.9570)	4.528(0.9504)	4.586(0.9506)	5.086(0.9474)
PNN+SPP	4.960(0.9606)	4.863(0.9557)	5.413(0.9657)	4.756(0.9493)	5.276(0.9435)	5.146(0.9456)
PNN+SPP+RN	6.087(0.9424)	5.620(0.9379)	6.538(0.9338)	5.777(0.9305)	5.596(0.9262)	6.239(0.9299)
PNN+SPP+RB	5.236(0.9463)	4.934(0.9578)	5.902(0.9546)	4.808(0.9501)	5.047(0.9488)	4.660(0.9502)
PNN+SPP+RN+RB	5.883(0.9492)	5.657(0.9549)	6.216(0.9373)	6.265(0.9235)	5.622(0.9366)	6.242(0.9278)

noted that our SPP only uses uniform noise in a range of $[-0.9, 0.9]$ mm, which is much smaller than that of the test data. Yet, we see that SPP with small sensor position perturbation enables the neural network to handle larger sensor perturbation significantly better.

The baseline SSFnet v2 is superior to both plain CNN and SSFnet v1 with respect to SSIM, which proves its superior robustness versus other models. However, after adding data augmentation and preprocessing strategies, the performance difference between the different neural networks is narrowed.

3.3. Efficiency comparison

The average runtime per image of TVAL3 on CPU Intel(R) Core(TM) i5-8400 CPU @ 2.80 GHz is about 15min51s. To compare the efficiency of the neural networks, the number of parameters and inference time on GPU Nvidia RTX3090 are shown in table 5. Since our data augmentation techniques do not affect the reconstruction time, we do not include them in the tables. We see that introducing Fourier transform in SSFnet leads to a tiny increase in inference time. The computation burden by PNN is acceptable in general.

Table 4. The RMSE and SSIM results on noisy simulation test images with different algorithms and training strategies. The simulation configuration is based on MUBI. The SSIM scores are shown in brackets. Bold font indicates the best results for each phantom.

	Phantom A			Phantom B		
TVAL3	1.230×10^4 (0.2277)			6.337×10^3 (0.3092)		
	CNN	SSFnet v1	SSFnet v2	CNN	SSFnet v1	SSFnet v2
Baseline	38.99(0.6748)	35.65(0.6100)	39.93(0.7426)	46.57(0.6160)	35.47(0.5805)	38.66(0.6714)
RN	14.80(0.8281)	44.21(0.6847)	44.67(0.7009)	28.70(0.7164)	24.33(0.7011)	26.90(0.7023)
RB	32.48(0.6470)	42.41(0.4671)	93.84(0.5183)	39.69(0.6057)	46.08(0.4523)	85.26(0.5385)
SPP	10.91(0.9155)	5.507(0.9516)	**4.947(0.9669)**	6.847(0.8867)	6.344(0.9032)	5.997(0.9046)
RN+RB+SPP	12.09(0.9154)	5.851(0.9565)	6.922(0.9493)	9.985(0.8156)	6.972(0.8841)	7.357(0.9040)
PNN	56.46(0.6796)	75.34(0.6798)	61.85(0.6951)	50.22(0.6602)	61.52(0.6650)	54.37(0.6813)
PNN+RN	25.12(0.7985)	9.398(0.8973)	9.763(0.9131)	22.74(0.7373)	13.60(0.8172)	12.40(0.8405)
PNN+RB	42.22(0.6891)	47.26(0.7083)	39.23(0.7526)	40.56(0.6733)	36.41(0.6748)	35.83(0.6781)
PNN+SPP	7.769(0.9460)	5.320(0.9523)	6.423(0.9561)	7.121(0.8835)	**4.934(0.9454)**	7.118(0.8858)
PNN+SPP+RN	16.13(0.8831)	7.127(0.9380)	7.518(0.9458)	15.50(0.7305)	9.330(0.8367)	7.455(0.8884)
PNN+SPP+RB	6.753(0.9544)	5.363(0.9578)	7.002(0.9485)	6.115(0.9173)	5.583(0.9198)	7.269(0.8665)
PNN+SPP+RN+RB	24.60(0.8424)	6.345(0.9525)	7.637(0.9397)	14.98(0.7319)	6.956(0.8917)	8.632(0.8602)

Table 5. The efficiency comparison of different neural networks. The simulation configuration is based on MUBI.

	DeepPet	mWnet	CNN	SSFnet	PNN+CNN	PNN+SSFnet v1
Number of parameters (million)	11.0	113.6	27.1	27.1	31.2	31.2
Inference time for 2000 samples (s)	63	61	49	51	73	76

3.4. Results on real data

Figures S9 and S10 show the results of neural networks trained with different generalization techniques on real measurement data from MUBI using the gelatin phantom shown in figure 2(b). We see TVAL3 has more blurred results than our deep learning methods. Generally, for both SSFnet v1 and SSFnet v2, both RN and RB alone can help improve the results of neural networks slightly. SPP can significantly improve networks' imaging performance. The combination of RN, RB and SPP together leads to further improvement, and the white hollow becomes clearer and more observable. Meanwhile, PNN can help remove outliers and reduce singular dots in the reconstructed images. The combination of SSFnet+PNN+RN is able to give the smoothest results compared to other combinations, but the white hollow area is less observable. Generally, SSFnet shows much better imaging quality on real measurement data, especially in the cases without SPP, which demonstrates that SSFnet has higher robustness and better generalizability than the SSFnet baseline, the plain CNN.

The comparison results of the state-of-the-art imaging methods on real measurement data from MUBI are shown in figure 7. It is noted that among the results by neural networks, SSFnet v1 combined with the proposed PNN and data augmentation methods achieved the best results by using far less number of parameters than mWnet. We also see that the low-sound-speed circle area in the result by the finite-frequency method is smaller than the high-sound-speed circle area, which indicates that the refraction correction by the multiresolution method may be incorrect. On the other hand, the results by TVAL3 and neural networks have a slightly larger low-sound-speed circle area, which is consistent with each other and is probably because they share the same arrival time estimation method.

Figure S11 in the supplementary document and figure 8 show the results on real measurement data from the prototype machine KIT 3D USCT II using a columnar gelatin phantom (different from the phantom in figure 2(b)). From the result by TVAL3, we see that a part of the edge of the columnar gelatin phantom is well reconstructed. This is probably because of the strong reflection wave from the surface of the gelatine phantom while the transmission wave is too weak to be detected. This could explain why the neural networks get terrible results. However, with the proposed data augmentation (RN,RB, SPP) and preprocessing network (PNN), the neural network is still able to catch some edge information.SSFnet v1+SPP and SSFnet v1+PNN give the best results among the neural network models, which also proves that SSFnet has better generalizability than plain CNN.

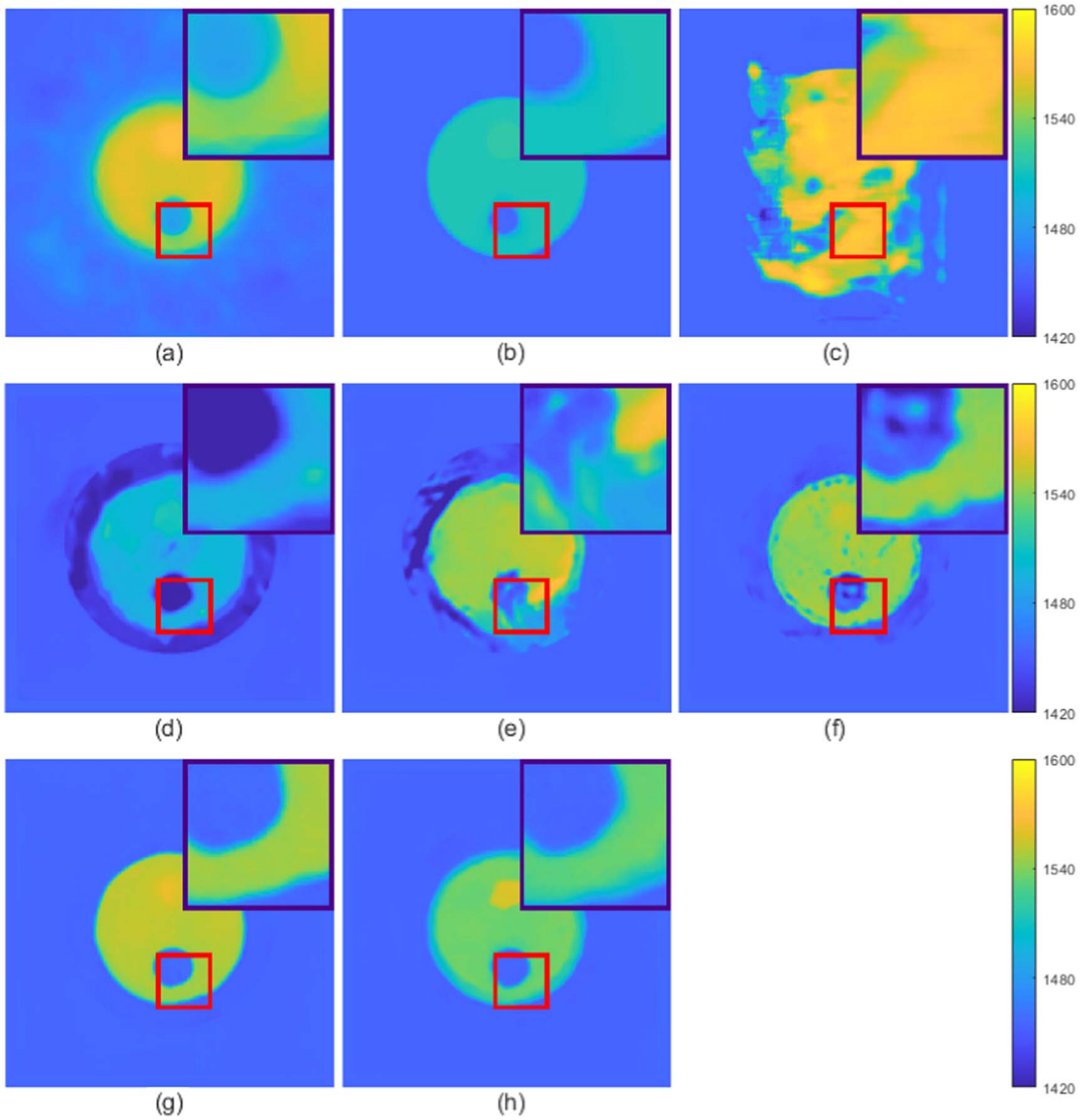

Figure 7. The image reconstruction results ($\mathrm{m}\cdot\mathrm{s}^{-1}$) of the state-of-the-art imaging methods on real measurement data from MUBI. (a) TVAL3; (b) Finite-frequency method; (c) DeepPet; (d) mWnet; (e) CNN; (f) SSFnet v1; (g) SSFnet v1+PNN+RN; (h) SSFnet v1 +PNN+RN+RB+SPP.

4. Discussions

4.1. The comparison with the state-of-the-art methods

In this work, we focus on improving neural networks' generalization against uncertainties due to imperfect system calibration and random errors in the imaging process of USCT. To our best knowledge, few works have addressed this problem. In addition, we emphasize the measurement domain generalization instead of the popular image domain generalization in computer vision. The techniques used in the image domain for data augmentation and data preprocessing may not be suitable for measurement domain generalization due to the difference between measurement data and image data. The characteristic of the targeted machine is a key aspect to be considered. We believe our work paves a new research front for simulation-to-real generalization in the measurement domain.

The state-of-the-art ray-based reconstruction algorithms that work on real data i.e. TVAL3 and the finite-frequency method (Martiartu *et al* 2020) are considered for comparison with our deep learning scheme. Even though the finite frequency method gives the clearest imaging result, it still shows a slight distortion with respect to the geometric shape of the reconstruction phantom image. The difference between the results of finite frequency method and TVAL3 is probably partly due to the different ways in which the measurement data are preprocessed. When compared with TVAL3, the proposed deep-learning-based approach achieves superior visual and quantitative performance on both simulation and real datasets. With robust training, we see that the

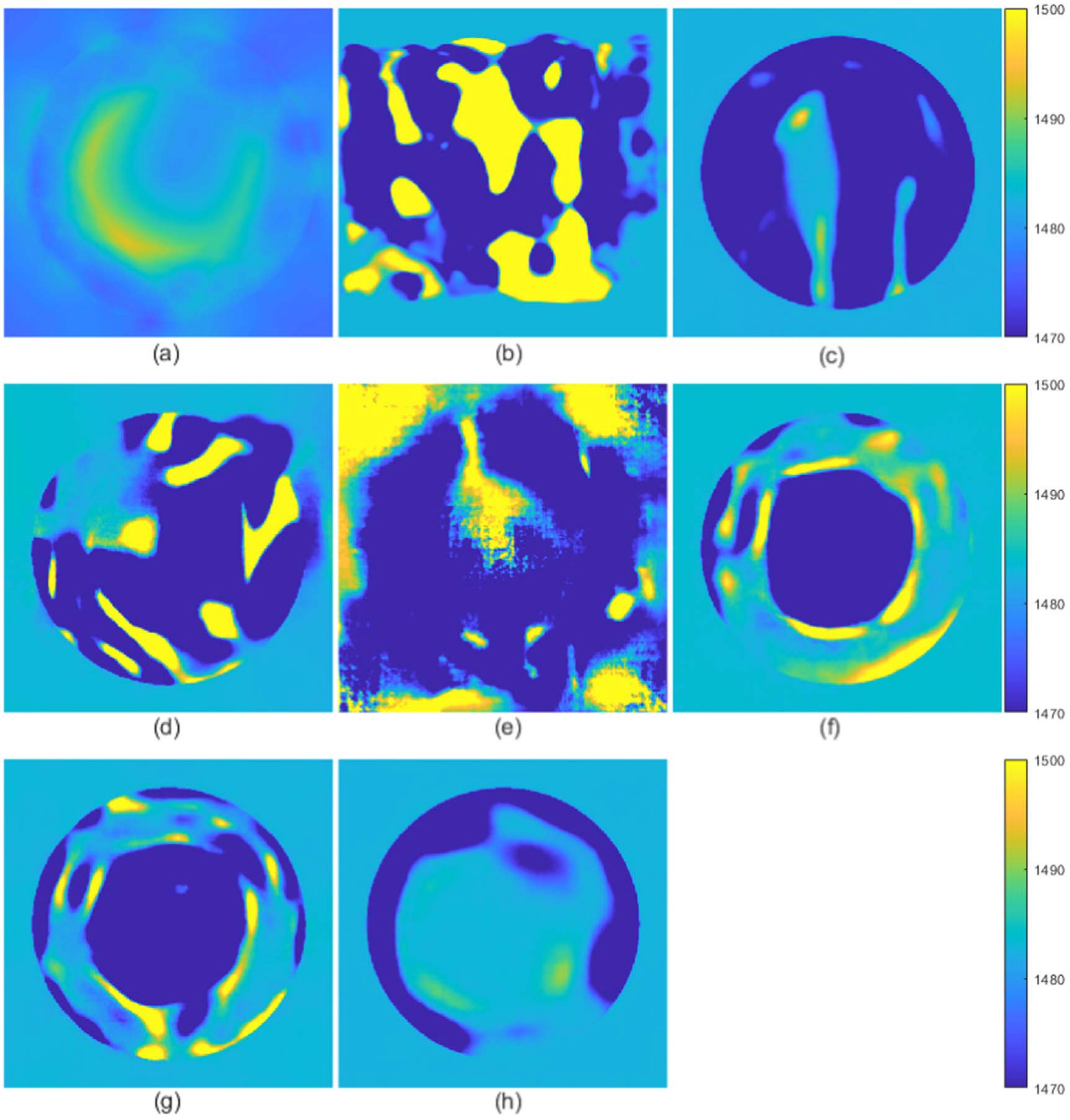

Figure 8. The image reconstruction results ($\mathrm{m \cdot s^{-1}}$) on real measurement data from KIT 3D USCT II. (a) TVAL3; (b) DeepPet; (c) mWnet; (d) CNN; (e) SSFnet v1; (f) SSFnet v1+SPP; (g) SSFnet v1+PNN; (h) SSFnet v1+PNN+RN+RB+SPP.

neural networks gain superior robustness to perturbance of sensor position. However, as a black-box model, the deep learning network itself lacks interpretability and the risk exists that unexpected artifacts may happen on measurement data with a certain adversarial noise. Adversarial defense is a promising direction to mitigate this issue. In addition, for a certain deep learning model, once trained, it can only apply to the machine of the same geometry and setup, which shows a lack of flexibility. A more flexible deep learning approach can be a future research direction.

4.2. The role of FFT in neural network layers

The ablation study on positions of FFT layers shows that putting more FFT layers close to the encoder side (the left side in figure 1) is beneficial to the generalization on real data. As we know the left side focuses on dealing with the measurement domain data, and transforming from the measurement domain to the image domain requires a global view of input data. FFT itself is essentially a form of convolution with global kernel size and thus is helpful in handling the mapping from measurement domain to image domain. In other words, FFT provides global frequency information that helps capture the useful global pattern in the feature map encoded from measurement data. This probably explains the superior performance of integrating FFT into the neural networks.

4.3. The role of different simulation-to-real generalization strategies

Our experiments show that both RN and RB are helpful in improving the neural networks' robustness. SPP gives the most significant improvement of neural networks' robustness to sensor displacement error. The combination of RN, RB and SPP can improve the robustness further, yet make it more difficult for convergence during the network training on the simulation dataset, which leads to a decrease in imaging quality on noise-free data quantitatively. In this case, SSFnet shows a much better performance than plain CNN. PNN can help remove outliers in the measurement data, and its combination with RN gives a highly smoothed result on real measurement data.

We observe that RN and RB can sometimes improve neural networks' performance on noise-free data. A similar phenomenon has been observed in other applications (Audhkhasi *et al* 2016), where suitable noise addition is helpful for performance improvement. On the other hand, heavy noise reduces their quantitative performance on noisy simulation data. However, it improves its visual performance on real data. It is because heavy noise sacrifices the prediction accuracy of the neural networks and, at the same time, allows the model to be more robust to handle the simulation-to-real gap.

The embedding of FFT into neural network layers greatly improves neural networks' performance on both simulated noisy data and real data. However, when we add more generalization training strategies (including data augmentation and preprocessing techniques), the performance difference between different neural networks is reduced. It is observed that a combination of improvements in both architecture and training strategies helps yield the best results on real data.

4.4. The limitations and future work

The ray-based approach requires a good estimation of ToF. In this paper, we adopt the ToF estimation methods proposed in (Bao and Jia 2019). We also note that recently there have been deep-learning-based ToF estimation methods being proposed, which claim to have superior performance to traditional methods. However, the robustness and generalizability of this deep learning approach remain to be tested and evaluated, which is beyond the focus of our paper. We hereby safely take the state-of-the-art traditional ToF estimation method we have for experiments.

Another limitation of this work is that we only test the real dataset for two different prototype machines with two different tissue-mimicking phantoms. In the future, we will test our methods on more machines with different geometric designs and more complex phantoms, which will also include 3D USCT image reconstruction.

5. Conclusions

In this paper, we present a deep learning scheme for fully learned USCT image reconstruction with real data. We show that integrating Fourier transform into a neural network helps achieve better robustness and generalizability. We develop and evaluate a series of simulation-to-real measurement data augmentation and preprocessing strategies on both simulation and real measurement data. Our approach can consistently improve the neural networks' performance on real data, and achieve a decent performance when compared to the state-of-the-art ray-based reconstruction algorithm.

ORCID iDs

Wenzhao Zhao https://orcid.org/0000-0001-5150-3781
Koen W A van Dongen https://orcid.org/0000-0001-6711-5898
Jürgen Hesser https://orcid.org/0000-0002-4001-1164

References

Agudo O C, Guasch L, Huthwaite P and Warner M 2018 3d imaging of the breast using full-waveform inversion *Proc. Int. Workshop Med. Ultrasound Tomogr* pp 99–110

Ali R, Hsieh S and Dahl J 2019 Open-source gauss-newton-based methods for refraction-corrected ultrasound computed tomography *Medical Imaging 2019: Ultrasonic Imaging and Tomography* vol 10955 (SPIE) pp 39–52

Audhkhasi K, Osoba O and Kosko B 2016 Noise-enhanced convolutional neural networks *Neural Netw.* **78** 15–23

Bakker J, Paulides M, Obdeijn I-M, Van Rhoon G and Van Dongen K W A 2009 An ultrasound cylindrical phased array for deep heating in the breast: theoretical design using heterogeneous models *Phys. Med. Biol.* **54** 3201 10

Bao Y and Jia J 2019 Improved time-of-flight estimation method for acoustic tomography system *IEEE Trans. Instrum. Meas.* **69** 974–84

Birk M, Dapp R, Ruiter N V and Becker J 2014 Gpu-based iterative transmission reconstruction in 3d ultrasound computer tomography *J. Parallel Distrib. Comput.* **74** 1730–43

Boehm C, Krischer L, Ulrich I, Marty P, Afanasiev M and Fichtner A 2022 Using optimal transport to mitigate cycle-skipping in ultrasound computed tomography *Medical Imaging 2022: Ultrasonic Imaging and Tomography* vol 12038 (SPIE) pp 48–57

Camacho J, Cruza J, González-Salido N, Fritsch C, Pérez-Liva M, Herraiz J and Udías J 2018 A multi-modal ultrasound breast imaging system *Proceedings of the International Workshop on Medical Ultrasound Tomography (Speyer, Germany, 1.-3. Nov. 2017)* (KIT Scientific Publishing) p 119

Medina-Valdés L., Pérez-Liva M., Camacho J., Udías J.M., Herraiz J.L. and González-Salido N. 2015 Multi-modal Ultrasound Imaging for Breast Cancer Detection *Physics Procedia* **63** 134-140134-140

Csurka G 2017 *arXiv* 1702.0537Mon, 13 Aug 2018 16:46:05 +0200 Domain adaptation for visual applications: A comprehensive survey

Dapp R 2013 *Abbildungsmethoden für die brust mit einem 3d-ultraschall-computertomographen* Karlsruher Institut für Technologie (KIT)

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 Conf. on Computer Vision and Pattern Recognition, Ieee* pp 248–55

Devaguptapu C, Agarwal D, Mittal G, Gopalani P and Balasubramanian V N 2021 On adversarial robustness: A neural architecture search perspective *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 152–61

Fan Y, Wang H, Gemmeke H, Hopp T and Hesser J 2022 Model-data-driven image reconstruction with neural networks for ultrasound computed tomography breast imaging *Neurocomputing* **467** 10–21

Farahani A, Voghoei S, Rasheed K and Arabnia H R 2021 A brief review of domain adaptation *Adv. Data Sci. Inf. Eng.* (Springer International Publishing) pp 877–94

Feng S, Lin Y and Wohlberg B 2021 Multiscale data-driven seismic full-waveform inversion with field data study *IEEE Trans. Geosci. Remote Sens.* **60** 1–14

Filipik A, Jan J, Peterlik I, Hemzal D, Ruiter N and Jirik R 2008 Modified time-of-flight based calibration approach for ultrasonic computed tomography *2008 XXX Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, IEEE* pp 2181–4

Ghifary M, Kleijn W B, Zhang M and Balduzzi D 2015 Domain generalization for object recognition with multi-task autoencoders *Proc. of the IEEE Int. Conf. on Computer Vision* pp 2551–9

Globerson A and Roweis S 2006 Nightmare at test time: robust learning by feature deletion *Proc. of the XXIII Int. Conf. on Machine Learning* pp 353–60

Häggström I, Schmidtlein C R, Campanella G and Fuchs T J 2019 Deeppet: A deep encoder-decoder network for directly solving the pet image reconstruction inverse problem *Med. Image Anal.* **54** 253–62

Haque S, Liu A W, Liu S and Chan J H 2021 Improving the robustness of a convolutional neural network with out-of-distribution data fine-tuning and image preprocessing *The XII Int. Conf. on Advances in Information Technology* pp 1–7

Hassouna M S and Farag A A 2007 Multistencils fast marching methods: A highly accurate solution to the eikonal equation on cartesian domains *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 1563–74

He K, Chen X, Xie S, Li Y, Dollár P and Girshick R 2022 Masked autoencoders are scalable vision learners *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 16000–9

He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: Surpassing human-level performance on imagenet classification *Proc. of the IEEE Int. Conf. on Computer Vision* pp 1026–34

Hendee W R and Ritenour E R 2003 *Medical Imaging Physics* (New York: John Wiley & Sons)

Hendrycks D, Mu N, Cubuk E D, Zoph B, Gilmer J and Lakshminarayanan B 2019 *arXiv* 1912.0278Mon, 17 Feb 2020 06:16:13 UTC Augmix: A simple data processing method to improve robustness and uncertainty

Hopp T, Šroba L, Zapf M, Dapp R, Kretzek E, Gemmeke H and Ruiter N V 2014 Breast imaging with 3d ultrasound computer tomography: results of a first in-vivo study in comparison to mri images *International Workshop on Digital Mammography* (Berlin: Springer) pp 72–9

Javaherian A and Cox B 2021 Ray-based inversion accounting for scattering for biomedical ultrasound tomography *Inverse Prob.* **37** 115003

Jush F K, Biele M, Dueppenbecker P M and Maier A 2022 *arXiv* 2202.01208Tue, 1 Feb 2022 11:09:35 UTC Deep learning for ultrasound speed-of-sound reconstruction: Impacts of training data diversity on stability and robustness

Li C, Duric N, Littrup P and Huang L 2009 In vivo breast sound-speed imaging with ultrasound tomography *Ultrasound Med. Biol.* **35** 1615–28

Li C, Yin W, Jiang H and Zhang Y 2013 An efficient augmented lagrangian method with applications to total variation minimization *Comput. Optim. Appl.* **56** 507–30

Li P, Li D, Li W, Gong S, Fu Y and Hospedales T M 2021 A simple feature augmentation for domain generalization *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 8886–95

Lucka F, Pérez-Liva M, Treeby B E and Cox B T 2021 High resolution 3d ultrasonic breast imaging by time-domain full waveform inversion *Inverse Prob.* **38** 025008

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A 2017 *arXiv* 1706.06083Wed, 4 Sep 2019 18:53:10 UTC Towards deep learning models resistant to adversarial attacks

Martiartu N K, Boehm C and Fichtner A 2020 3-d wave-equation-based finite-frequency tomography for ultrasound computed tomography *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **67** 1332–43

Mercerat E D and Nolet G 2013 On the linearity of cross-correlation delay times in finite-frequency tomography *Geophys. J. Int.* **192** 681–7

Naveed H 2021 *arXiv* 2106.07085Wed, 16 Jun 2021 10:42:19 +0200 Survey: Image mixing and deleting for data augmentation

Ozmen N, Dapp R, Zapf M, Gemmeke H, Ruiter N V and van Dongen K W A 2015 Comparing different ultrasound imaging methods for breast cancer detection *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **62** 637–46

Peng X, Usman B, Saito K, Kaushik N, Hoffman J and Saenko K 2018 *arXiv* 1806.09755Mon, 13 Aug 2018 16:48:21 +0200 Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation

Peng X B, Andrychowicz M, Zaremba W and Abbeel P 2018 Sim-to-real transfer of robotic control with dynamics randomization *2018 IEEE Int. Conf. on Robotics and Automation (ICRA), IEEE (Brisbane, QLD, Australia, 21-25 May 2018)* pp 3803–10

Perez-Liva M, Udías J M, Camacho J, Merčep E, Deán-Ben X L, Razansky D and Herraiz J L 2020 Speed of sound ultrasound transmission tomography image reconstruction based on bézier curves *Ultrasonics* **103** 106097

Poudel J, Forte L A and Anastasio M A 2019 Compensation of 3d-2d model mismatch in ultrasound computed tomography with the aid of convolutional neural networks (conference presentation) *Medical Imaging 2019: Ultrasonic Imaging and Tomography* vol 109551095507(SPIE)

Prasad S and Almekkawy M 2020 A fast and efficient ultrasound tomography using deep learning *J. Acoust. Soc. Am.* **148** 2450–2450

Qiu M and Qiu H 2020 Review on image processing based adversarial example defenses in computer vision *2020 IEEE VI Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), IEEE* pp 94–9

Robins T, Camacho J, Agudo O C, Herraiz J L and Guasch L 2021 Deep-learning-driven full-waveform inversion for ultrasound breast imaging *Sensors* **21** 457013

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

Ruiter N V, Zapf M, Hopp T, Gemmeke H and Van Dongen K W A 2017 USCT data challenge *USCT data challenge SPIE Medical Imaging 2017: Ultrasonic Imaging and Tomography* vol 10139 (SPIE) pp 412–9

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 *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 1874–83

Stanziola A, Arridge S R, Cox B T and Treeby B E 2021 A helmholtz equation solver using unsupervised learning: Application to transcranial ultrasound *J. Comput. Phys.* **441** 110430

Tan W Y, Steiner T and Ruiter N V 2015 Newton's method based self calibration for a 3d ultrasound tomography system *2015 IEEE Int. Ultrasonics Symposium (IUS), IEEE* pp 1–4

Taylor J 1997 *Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements, 2nd Edition* (648 Broadway, Suite 902, New York, NY 10012: University Science Books)

Tissue properties–speed of sound 2022 -11-24https://itis.swiss/virtual-population/tissue-properties/database/acoustic-properties/speed-of-sound/

Tobin J, Fong R, Ray A, Schneider J, Zaremba W and Abbeel P 2017 Domain randomization for transferring deep neural networks from simulation to the real world *2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE* pp 23–30

Usct data: 2d csic/ucm usct 2022 -02-28http://ipeusctdb1.ipe.kit.edu/~usct/challenge/?page_id=183

Vitale S, Orlando J I, Iarussi E and Larrabide I 2020 Improving realism in patient-specific abdominal ultrasound simulation using cyclegans *Int. J. Comput. Assist. Radiol. Surg.* **15** 183–92

Volpi R, Namkoong H, Sener O, Duchi J C, Murino V and Savarese S 2018 Generalizing to unseen domains via adversarial data augmentation *Adv. Neural Inf. Process. Sys.* 31

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12

Xie C and Yuille A 2019 *arXiv* 1906.03787Fri, 14 Jun 2019 09:38:24 +0200 Intriguing properties of adversarial training at scale

Xu H, Ding S, Zhang X, Xiong H and Tian Q 2022 *arXiv* 2206.04846Fri, 10 Jun 2022 02:41:48 UTC Masked autoencoders are robust data augmentors

Yang Y and Soatto S 2020 Fda: Fourier domain adaptation for semantic segmentation *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 4085–95

Yin D, Gontijo Lopes R, Shlens J, Cubuk E D and Gilmer J 2019 A Fourier perspective on model robustness in computer vision *Adv. Neural Inf. Process. Syst.* vol 32

Yue X, Zhang Y, Zhao S, Sangiovanni-Vincentelli A, Keutzer K and Gong B 2019 Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data, in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 2100–10

Zeng Y, Qiu H, Memmi G and Qiu M 2020 A data augmentation-based defense method against adversarial attacks in neural networks *Int. Conf. on Algorithms and Architectures for Parallel ProcessingBerlin* pp 274–89

Zhao W, Wang H, Gemmeke H, Van Dongen K W A, Hopp T and Hesser J 2020 Ultrasound transmission tomography image reconstruction with a fully convolutional neural network *Phys. Med. Biol.* **65** 235021

Zuch F, Hopp T, Zapf M and Ruiter N 2021 Refraction corrected transmission imaging based on bézier curves: first results with kit 3d usct *Proc. of the Int. Workshop on Medical Ultrasound Tomography* ed C Böhm *et al* p 235 https://publikationen.bibliothek.kit.edu/1000128316