Published in final edited form as:
IEEE Trans Ultrason Ferroelectr Freq Control. 2019 May ; 66(5): 898–910. doi:10.1109/TUFFC.2019.2903795.

Beamforming and Speckle Reduction Using Neural Networks

Dongwoon Hyun,
Department of Radiology, Stanford University, Stanford, CA 94305 USA

Leandra L. Brickson,
Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA.

Kevin T. Looby,
Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA.

Jeremy J. Dahl
Department of Radiology, Stanford University, Stanford, CA 94305 USA

Abstract

With traditional beamforming methods, ultrasound B-mode images contain speckle noise caused by the random interference of subresolution scatterers. In this paper, we present a framework for using neural networks to beamform ultrasound channel signals into speckle-reduced B-mode images. We introduce log-domain normalization-independent loss functions that are appropriate for ultrasound imaging. A fully convolutional neural network was trained with simulated channel signals that were co-registered spatially to ground truth maps of echogenicity. Networks were designed to accept 16 beamformed subaperture radiofrequency signals. Training performance was compared as a function of training objective, network depth, and network width. The networks were then evaluated on simulation, phantom, and *in vivo* data and compared against existing speckle reduction techniques. The most effective configuration was found to be the deepest (16 layer) and widest (32 filter) networks, trained to minimize a normalization-independent mixture of the ℓ_1 and multi-scale structural similarity losses. The neural network significantly outperformed delay-and-sum and receive-only spatial compounding in speckle reduction while preserving resolution and exhibited improved detail preservation over a non-local means methods. This work demonstrates that ultrasound B-mode image reconstruction using machine-learned neural networks is feasible and establishes that networks trained solely *in silico* can be generalized to real-world imaging *in vivo* to produce images with significantly reduced speckle.

I. Introduction

In brightness mode (B-mode) ultrasound imaging, the echoes from an ultrasonic pulse are used to reconstruct images according to their magnitude (i.e. brightness). Thus, B-mode images are a map of the *echogenicity* of the insonified medium. The echo magnitudes are measured using an array of sensors via a process referred to as beamforming. The classical beamformer is delay-and-sum (DAS), which forms a point-wise estimate of echogenicity

Correspondence to: Dongwoon Hyun. dongwoon.hyun@stanford.edu.

based on the magnitude of the summed array signals. However, in medical ultrasound imaging, echoes are generated by scattering sources that are smaller than the resolution of the imaging system. These echoes interfere stochastically, producing a strong multiplicative noise in the measured DAS output referred to as *speckle*. Speckle manifests as a temporally stationary grainy texture in regions with homogeneous echogenicity. Speckle is commonly used to infer the scattering properties of tissue [1] and can be used for tracking blood flow and tissue displacements [2]. However, for the task of echogenicity estimation, speckle reduces the perceived resolution of the target [3] and is largely treated as an undesirable noise that degrades diagnostic B-mode imaging [4], [5].

Speckle reduction can be accomplished using beamforming methods that operate on the radiofrequency signals received by an array of transducer elements. Common beamforming techniques for speckle reduction include spatial and frequency compounding, in which the aperture or the bandwidth are subdivided, respectively. These subdivisions are used independently to reconstruct images that are subsequently averaged. The overall speckle is reduced because the speckle patterns observed by each subdivision are decorrelated from one another [6], [7]. However, beamforming has historically been viewed as a method of suppressing off-axis noises or improving resolution, rather than a means for speckle reduction. Examples of other beamformers include minimum variance beamforming [8], in which the channels are weighted to suppress off-axis noises, as well as a newly proposed machine learning method in which deep fully connected networks are used to filter signals arriving from outside of the main lobe [9]. In these cases, beamforming is used to preserve speckle, rather than to remove it. We propose that, in addition to suppressing noise from outside the intended focus, an ideal beamformer for B-mode imaging should be an estimate of the average echogenicity of the scatterers. In particular, the backscatter from a homogeneous region of tissue of constant echogenicity should produce a uniform response (as opposed to a speckle response) while preserving the structure and echogenicity of the medium.

Speckle reduction has been studied far more extensively in the context of post-processing filters, which are applied to images that have already been beamformed. Popular techniques include anisotropic diffusion [10], [11] and discrete wavelet transforms [12], [13]. A stochastic iterative technique to remove pixel outliers, called the squeeze box filter, has also been proposed [14]. More recently, non-local means methods have demonstrated excellent speckle reduction capabilities [15]–[17]. These techniques selectively smooth pixels originating from speckle while preserving other structures and details. However, a disadvantage to purely post-processing techniques is that they rely entirely on fully-formed images of demodulated and envelope detected data, and are thus unable to take advantage of channel and phase information that are irreversibly lost in the summation and image formation process, respectively.

Previous speckle-reduction methods were designed according to the underlying physics and statistics of speckle. By contrast, supervised learning with artificial neural networks has recently demonstrated extraordinary success in image recognition, segmentation, and denoising by utilizing a data-driven approach [18]. Supervised learning with neural networks is a class of machine learning techniques in which a cascade of transformations are applied to an input in order to eventually produce a desired output. The parameters of the transformation are "learned" via a gradient descent algorithm designed to minimize the error between the output of the network and the known ground truth. Neural networks can be trained to beamform speckle-reduced B-mode images by being presented with many instances of channel data and corresponding ground truth echogenicity, and are especially attractive for ultrasound imaging because they can easily be deployed in real-time [18]. Unfortunately, ground truth echogenicity is virtually unavailable *in vivo*. Ground truth can be obtained in simulations, but it has not yet been demonstrated whether a network trained entirely *in silico* can generalize to real-world imaging conditions such as *in vivo*, where ultrasound signals face additional challenges via image degradation such as phase aberration, acoustical noise, and electronic noise.

In this work, we use simple linear simulations of ultrasound imaging [19], [20] in conjunction with deep convolutional neural networks to empirically learn a speckle-reducing beamformer. We demonstrate how neural networks can be trained to transform channel data into B-mode images, and analyze performance as a function of network architecture and training objectives. Deep networks are trained using simulated ultrasound channel data that are co-registered with a reference ground truth map of echogenicity. The imaging performance of the trained networks are then compared against existing despeckling techniques in simulations, a calibrated phantom, and *in vivo* human liver and kidney.

II. Ultrasound Image Reconstruction

A. Problem Formulation

Consider a vectorized grid of $P \times Q$ field points (also referred to as "pixels") with true echogenicities $\boldsymbol{y} \in \mathbb{R}^{PQ}$. Let $\boldsymbol{X} \in \mathbb{C}^{PQ \times N}$ denote the demodulated analytic signals captured by the N elements of a transducer array after applying the appropriate time delays to focus the array at each of the field points. In B-mode image reconstruction, $\boldsymbol{y}$ is estimated from $\boldsymbol{X}$ using some function $f(\boldsymbol{X}) = \hat{\boldsymbol{y}}$. For instance, the traditional delay-and-sum (DAS) technique estimates $\boldsymbol{y}$ as the absolute value of the channel sum:

$$f_{\mathrm{DAS}}(\boldsymbol{X}) = |\boldsymbol{X}\mathbf{1}|, \tag{1}$$

where $\mathbf{1}$ is an N-vector of ones and $|\,|$ denotes an element-wise absolute value. Let us denote a beamforming neural network as $f_{\mathrm{NN}}(\boldsymbol{X};\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of the network. The goal of B-mode image reconstruction with neural networks is to find the optimal parameters $\boldsymbol{\theta}^{\star}$ that minimize the error between an estimated image $\hat{\boldsymbol{y}}$ and the true image $\boldsymbol{y}$, as quantified by some loss function $\mathscr{L}(y, \hat{y})$:

$$\theta^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathscr{L}\big(\boldsymbol{y}, f_{\mathrm{NN}}(\boldsymbol{X};\boldsymbol{\theta})\big)\,. \tag{2}$$

The minimization problem is typically solved using some form of gradient descent, an approach in which each of the parameters is iteratively updated to reduce the error:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} \mathcal{L}\left(\boldsymbol{y}, f_{\mathrm{NN}}(\boldsymbol{X}; \boldsymbol{\theta})\right), \tag{3}$$

where α denotes the step size, also called the “learning rate”.

B. Standard Loss Functions

The choice of loss function significantly affects the training process. The ℓ_1 and ℓ_2 norms are often used to quantify the reconstruction error [21]:

$$\mathcal{L}_{\ell_1}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{PQ} \sum_{p=1}^{PQ} \left| y_p - \hat{y}_p \right| \tag{4}$$

$$\mathcal{L}_{\ell_2}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \left(\frac{1}{PQ} \sum_{p=1}^{PQ} \left(y_p - \hat{y}_p \right)^2 \right)^{\frac{1}{2}}, \tag{5}$$

where the p-th pixels of $\boldsymbol{y}$ and $\hat{y}$ are respectively denoted as y_p and $\hat{y}_p$. Alternative metrics such as structural similarity (SSIM) and multi-scale structural similarity (MS-SSIM) have also been proposed [21]–[23]. The SSIM between pixels y_p and $\hat{y}_p$ is computed as

$$\mathrm{SSIM}\left(y_p, \hat{y}_p\right) = \left(\frac{2\mu_{y_p}\mu_{\hat{y}_p} + C_1}{\mu_{y_p}^2 + \mu_{\hat{y}_p}^2 + C_1} \right) \left(\frac{2\sigma_{y_p \hat{y}_p} + C_2}{\sigma_{y_p}^2 + \sigma_{\hat{y}_p}^2 + C_2} \right), \tag{6}$$

where C_1 and C_2 are empirically selected scalar parameters to enhance numerical stability, μ_{y_p} and $\mu_{\hat{y}_p}$ are the mean values of a neighborhood around y_p and $\hat{y}_p$, respectively, $\sigma_{y_p}^2$ and $\sigma_{\hat{y}_p}^2$ are their variances, and $\sigma_{y_p \hat{y}_p}$ is their covariance. The means, variances, and covariance are obtained using Gaussian filters [22], [23]. SSIM values range from −1 to 1, where 1 indicates perfect correspondence between the images. Therefore, a loss function can be defined as

$$\mathcal{L}_{\mathrm{SSIM}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{1}{PQ} \sum_{p=1}^{PQ} \mathrm{SSIM}\left(y_p, \hat{y}_p\right). \tag{7}$$

The $\mathcal{L}_{\text{MS-SSIM}}$ metric extends SSIM by combining $\mathcal{L}_{\mathrm{SSIM}}$ measurements using several neighborhood sizes in order to compare the images at multiple resolution scales. The different scales can be achieved either by downsampling the image [22] or by changing the standard deviation parameter of the Gaussian filter [21]. In this work, we adopt the latter approach.

C. Loss Functions for Ultrasound B-mode Imaging

In B-mode imaging, it is important to accurately reconstruct hypoechoic targets such as blood vessels and heart chambers, whose signals can be more than 40 dB (100 times) weaker than the background tissue. B-mode images are typically compressed prior to being viewed in order to visualize the wide dynamic range. However, the large discrepancy in signal strengths may cause standard loss functions to over-emphasize errors in strongly echogenic targets and to under-emphasize errors in hypoechoic targets. Therefore, we propose to compute losses using logarithmically compressed images, i.e., $\mathscr{L}(\log y, \log \hat{y})$. This allows the errors to be measured in the same domain that the images are viewed in.

Another challenge is that B-mode images and their ground truth echogenicity maps are both defined in arbitrary units, making it unclear how to compare the two images. One approach is to normalize each respective image, (e.g., by their maximum values). Unfortunately, the standard loss functions are highly sensitive to the normalization, whereas an ideal loss function for images with arbitrary units should be independent of their normalization.

We propose a new loss function that is intrinsically independent of normalization and is suitable for comparing two images with arbitrary units. Let $\mathscr{L}^{\star}$ define the minimum achievable loss $\mathscr{L}$ when $\hat{y}$ is scaled by some positive weight parameter w:

$$\mathscr{L}^{\star}(\log \boldsymbol{y}, \log \hat{\boldsymbol{y}}) = \min_{w > 0} \mathscr{L}(\log \boldsymbol{y}, \log w\hat{\boldsymbol{y}})\,. \tag{8}$$

Closed form expressions for the $\mathscr{L}^{\star}_{\ell_1}$, $\mathscr{L}^{\star}_{\ell_2}$, and $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ loss functions are provided in the Appendix. These loss functions apply precise normalization to each image such that the $\mathscr{L}_{\ell_1}$, $\mathscr{L}_{\ell_2}$, and $\mathscr{L}_{\text{MS-SSIM}}$ losses are minimized, respectively, allowing the images to be compared according to their relative contrasts and structures rather than their absolute magnitudes.

III. Methods

A. Field II Simulation Training Dataset

The Field II Pro simulation package [19], [20], [24] was used to simulate ultrasound channel data from 128 elements of a Verasonics L12–3v linear array transducer. A full synthetic aperture data set was simulated using single-element transmits at 8 MHz with 60% bandwidth. Speckle was simulated with random uniformly distributed scatterers in a 10 mm×10 mm×3 mm phantom centered at the elevation focus of 2 cm. The scatterer density was selected to be 60 scatterers per resolution cell, and the scattering amplitudes were normally distributed and weighted according to ground truth echogenicity maps.

To provide the network with a wide range of features and contrasts, real photographic images were used as the ground truth echogenicity. The images were taken from publicly available image databases: 512 images from a Places2 [25] test set and 512 images from an ImageNet [26] validation set were selected for a total of 1024 images. These images were used solely for their patterns and contrasts; their corresponding classification labels were not used. The images were converted to grayscale and cropped into a 224 pixel ×224 pixel square patch. The patch was then mapped onto the lateral and axial extent of the phantom (10 mm×10 mm) to serve as the ground truth echogenicity map, and the pixel intensities were used to weight the scattering amplitudes of the simulated scatterers according to their positions via bilinear interpolation. This approach was chosen as a convenient alternative to custom-designing a wide variety of ground truth echogenicity maps.

For each of the 1024 images, an independent set of random scatterers was weighted and used in a full synthetic aperture simulation. For each simulation, the received radiofrequency (RF) channel signals were demodulated and focused into the same 224 pixel×224 pixel grid as the ground truth echogenicity map, with dynamic focusing applied on both transmit and receive. The dimensions of the resulting data cube were 224 pixels×224 pixels×128 channels.

The set of 1024 simulations was then resampled to generate an augmented training dataset comprised of 5000 co-registered pairs of focused channel data and reference ground-truth echogenicity. The resampled data/reference pairs were selected to have a smaller patch size of 64 pixels×64 pixels and each was drawn randomly from one of the 1024 simulations. While the number of pixels was held constant, the lateral and axial positions and sizes of the patches were allowed to vary independently of one another, resulting in rectangular patches. The resulting 64×64 patches were then treated as though they were square patches with stretched speckle patterns, enabling the emulation of a wide variety of imaging configurations from a limited number of simulations. Each channel dataset was corrupted by a random amount of white thermal noise and band-limited acoustical reverberation noise, specified in decibels (dB) relative to the RMS of the channel signals. Each attribute was selected randomly from a uniform distribution over the range of values listed in Table I.

B. Image Reconstruction Quality Metrics

The quality of image reconstruction was measured using several metrics. The $\mathscr{L}_{\ell_1}^{\star}$, $\mathscr{L}_{\ell_2}^{\star}$, and $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ errors were computed between the reconstructed log-image and the ground truth echogenicity log-image when available. The $\mathscr{L}_{\ell_2}^{\star}$ and $\mathscr{L}_{\ell_1}^{\star}$ errors were obtained in units of dB, whereas the $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ error was bounded from 0 to 2, with 0 being achieved when $y = \hat{y}$. The reconstruction quality was also measured using the contrast and contrast-to-noise ratio (CNR) of cyst targets and using the signal-to-noise ratio (SNR) of the tissue background:

$$\text{Contrast} = 20\log_{10}\left(\frac{\mu_t}{\mu_b}\right) \tag{9}$$

$$\text{CNR} = \frac{\mu_t - \mu_b}{\sqrt{\sigma_t^2 + \sigma_b^2}} \tag{10}$$

$$\mathrm{SNR} = \frac{\mu_b}{\sigma_b}, \tag{11}$$

where μ_t and μ_b denote the means and σ_t and σ_b the standard deviations of the target and background, respectively. Contrast, CNR, and SNR were computed on the linear scale images, prior to log-compression. The CNR essentially combines the contrast and SNR metrics into a single measure of lesion detectability [27]. In the simulations and the phantom, the cyst contrasts and CNRs were measured in concentric regions of interest (ROIs). The target ROI was selected as a circle with a radius of 0.8 times the cyst radius, and the background ROI as a ring with inner and outer radii of 1.1 and 1.5 times the cyst radius, respectively. The background SNR was measured in a homogeneous region of speckle. For DAS beamforming, the measured pixel values in a region of homogeneous echogenicity are classically expected to follow the Rayleigh distribution, resulting in an SNR of 1.91.

Image resolution was measured using a methodology similar to Dahl et al. [28]. First, the system response to a transition between two regions of different echogenicities was measured, referred to as the edge spread function (ESF). Field II Pro was used to simulate a target wherein half the azimuthal field of view was a speckle region of constant echogenicity and the other half was an anechoic region. The resulting images were averaged axially and over 16 independent scatterer realizations to reduce noise. Next, the ESF was differentiated in the azimuthal dimension to produce a line spread function (LSF). Finally, the FWHM of the LSF was measured to obtain the image resolution.

C. Deep Convolutional Neural Networks

Convolutional neural networks were used to estimate echogenicity from the focused and demodulated channel signals. Analysis was restricted to a maximum of 16 channels per pixel due to computational and memory constraints. To achieve this, the 128 element array was subdivided into 16 equal subapertures of 8 elements each, and the signals from each subaperture were beamformed into a single signal, yielding a total of 16 complex signals per pixel. The real and imaginary components were then concatenated in the channel dimension prior to being input into the neural network, resulting in 32 distinct channels per pixel, i.e., for a pixel grid of size $P \times Q$, the resulting data was $P \times Q \times 32$.

The network architecture is illustrated in Fig. 1. The network consisted of repeated convolution “blocks”, with each block applying a 2D convolution, batch normalization [29], and rectified linear unit activation [30]. This motif was repeated for M blocks. Each 2D convolution layer was composed of F machine-learned filters, and the size of each filter was $7 \times 7 \times 32$ for the first layer, $5 \times 5 \times F$ for the second layer, and $3 \times 3 \times F$ for subsequent layers, with the convolution occurring over the first two dimensions of each. Each convolution was zero-padded in the first two dimensions such that the input and output were the same size, making the size of the output of each convolutional layer $P \times Q \times F$.

An uncompressed B-mode image was separately formed from the input data using equation (1) and concatenated to the output of the M-th block, yielding a data array of size $P \times Q \times (F + 1)$. One final convolution filter of size $1 \times 1 \times (F + 1)$ was used to produce an output image of size $P \times Q \times 1$ that was subsequently squared on a pixel-wise basis to yield positive values of echogenicity. The fully convolutional nature of the network allowed echogenicity estimation on a pixel-wise basis. Unless otherwise specified, the default parameters for the networks were $M = 16$ convolution blocks with $F = 32$ filters per layer. Pooling layers and dropout were not utilized.

The neural networks were implemented in Python using TensorFlow [31] using the Adam [32] optimizer with a single NVIDIA GeForce GTX 1080 Ti GPU, which has 11 GB of memory. We performed hyperparameter tuning of the learning rates, batch size, and both ℓ_1 and ℓ_2 regularization of filter weights to achieve optimal performance using an independent validation dataset that was separate from all other datasets. The final hyperparameters used are tabulated in Table II.

D. Neural Network Training and Analysis

Neural network training performance was analyzed as a function of training objective. A network with 16 convolution blocks and 32 filters per layer was trained to minimize either $\mathscr{L}^{\star}_{\ell_1}$, $\mathscr{L}^{\star}_{\ell_2}$, $\mathscr{L}^{\star}_{\text{MS-SSIM}}$, or a mixture of ℓ_1 and MS-SSIM, defined as:

$$\mathscr{L}^{\star}_{\text{Mix}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \mathscr{L}^{\star}_{\ell_1}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \beta \mathscr{L}^{\star}_{\text{MS}-\text{SSIM}}(\boldsymbol{y}, \hat{\boldsymbol{y}}), \tag{12}$$

where β was set heuristically to 200 to equalize the contributions of each loss function. (This formulation is adapted from the mixed loss function given in [21].) Speckle reduction was evaluated on a validation dataset consisting of 32 new Field II Pro simulations, also based on photographs. Performance was quantified using the $\mathscr{L}^{\star}_{\ell_1}$, $\mathscr{L}^{\star}_{\ell_2}$, and $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ loss functions.

Similarly, speckle reduction performance was also analyzed as a function of depth using networks with 2, 4, 8, or 16 blocks and 32 filters per layer, as well as networks with 16 blocks and 4, 8, 16, or 32 filters per layer, where all of the networks were trained to minimize $\mathscr{L}^{\star}_{\text{Mix}}$. Finally, performance was compared with and without concatenating a B-mode image after the M-th convolution block.

E. Testing of Speckle Reduction Methods

All testing was performed using a neural network with a depth of 16 blocks, a width of 32 filters, and trained to minimize the $\mathscr{L}^{\star}_{\text{Mix}}$ loss function for 30 training epochs unless otherwise indicated. The speckle reduction performance of the network was compared against those of receive spatial compounding [6] and optimized Bayesian nonlocal means (OBNLM) [15], [16]. Spatial compounding was implemented on receive by performing DAS beamforming and envelope detection independently on four non-overlapping subapertures and summing the result. The OBNLM algorithm was applied to images that were beamformed according to equation (1) using the publicly available MATLAB implementation provided by the authors of this algorithm [15], [16]. The default parameter values provided by the authors were used ($M = 7$, $\alpha = 3$, $h = 0.7$).

Simulation, phantom, and *in vivo* test datasets were used to evaluate beamforming performance. A simulation test dataset was obtained with Field II Pro with the same imaging configuration as the training dataset. Hypoechoic cylindrical cysts with diameters of 1 mm and 3 mm and contrasts of −20 dB and −6 dB were centered at the elevation focus of 2 cm depth. The full synthetic aperture set of channel signals was retrospectively focused into a 1 cm×1 cm pixel region also centered at the elevation focus, and the channels were delay-and-summed into 16 subaperture IQ signals. The simulated results were assessed using the $\mathscr{L}^{\star}_{\ell_1}$, $\mathscr{L}^{\star}_{\ell_2}$, and $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ loss functions, as well as contrast, CNR, and SNR.

Generalizability from simulations to real-world ultrasound data was tested on a CIRS Model 040GSE calibrated phantom imaged using a Verasonics Vantage 256 research scanner with an L12–3v linear array transducer. Single element transmissions at 6 MHz sampled at 24 MHz were used to obtain a full synthetic transmit aperture for dynamic focusing on both transmit and receive. The channel data were focused into a pixel grid of 4 cm in depth and 2 cm in width, with a pixel spacing of λ/2 in both dimensions. Transmit and receive aperture growth were applied to achieve an *f*-number of 2. The phantom images were evaluated using contrast, CNR, and SNR.

Generalizability to clinical imaging conditions was assessed *in vivo* in the kidney of a healthy 58-year-old male volunteer and in the liver of a 68-year-old female who had a focal lesion with a surrounding fluid capsule. Both subjects provided written consent and imaging was performed under an IRB approved protocol. Channel datasets were acquired with a modified Siemens S2000 scanner using a Siemens 4C1 transducer. Pulse-inversion harmonic imaging was performed using focused transmissions at 1.82 MHz. Due to technical limitations, the channel signals were obtained for only 64 (out of 192) elements in a sector of 54 (out of 192) beams. The remaining beams were acquired using the full aperture using conventional DAS. The 64 receive channel signals were beamformed into 16 subaperture signals prior to being input into the neural network. For this particular dataset only, a partially-trained network (15 epochs rather than 30) was used to evaluate speckle reduction to avoid overfitting. Image quality was assessed using contrast, CNR, and SNR of the liver lesion vs. the surrounding fluid, with the ROIs selected to obtain a large region of speckle while avoiding obvious and significant changes in underlying echogenicity.

IV. Results

A. Neural Network Training and Analysis

Figure 2 plots the (a) $\mathscr{L}^{\star}_{\ell_1}$, (b) $\mathscr{L}^{\star}_{\ell_2}$, and (c) $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ validation losses of neural networks over 50 training epochs for several training objectives, as computed on the validation dataset. Neural networks were trained to minimize either $\mathscr{L}^{\star}_{\text{MS-SSIM}}$, $\mathscr{L}^{\star}_{\ell_1}$,$\mathscr{L}^{\star}_{\ell_2}$, or $\mathscr{L}^{\star}_{\text{Mix}}$. The networks trained to minimize $\mathscr{L}^{\star}_{\ell_1}$ and $\mathscr{L}^{\star}_{\text{Mix}}$ were found to yield the lowest $\mathscr{L}^{\star}_{\ell_1}$ losses. The networks trained to minimize $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ and $\mathscr{L}^{\star}_{\text{Mix}}$ yielded the lowest $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ and $\mathscr{L}^{\star}_{\ell_2}$

losses. Minimizing $\mathscr{L}_{\ell_2}^{\star}$ resulted in unstable performance, with erratic increases in all three losses occurring partway through training.

Figure 3 shows the validation losses as a function of network depth and width for a network trained to minimize $\mathscr{L}_{\mathrm{Mix}}^{\star}$ on the validation dataset. Given a fixed width of 32 filters, deeper networks converged to lower $\mathscr{L}_{\ell_1}^{\star}$, $\mathscr{L}_{\ell_2}^{\star}$, and $\mathscr{L}_{\mathrm{MS\text{-}SSIM}}^{\star}$ losses after 50 epochs, and approached these values more rapidly as well. Similar results were observed when fixing the network depth to 16 layers and increasing the width. With the exception of the network with a width of 4 filters, all networks outperformed DAS beamforming by the $\mathscr{L}_{\ell_1}^{\star}$, $\mathscr{L}_{\ell_2}^{\star}$, and $\mathscr{L}_{\mathrm{MS\text{-}SSIM}}^{\star}$ metrics.

Figure 4 plots the same validation losses for networks trained with and without a concatenated B-mode image after the *M*-th convolutional block. The network without B-mode had significantly higher $\mathscr{L}_{\ell_1}^{\star}$ and $\mathscr{L}_{\ell_2}^{\star}$ losses than both DAS and the network with DAS.

Though not plotted here, the training errors closely followed the trends of the validation errors in all cases in Figs. 2, 3, and 4, indicating that the networks were not overfitting to the training data.

B. Image Resolution

The FWHM of the LSF was used to measure the azimuthal resolution of each imaging method. The resolutions were found to be: $\mathrm{DAS}_{\mathrm{FWHM}} = 0.132$ mm; $\mathrm{SC}_{\mathrm{FWHM}} = 0.156$ mm; $\mathrm{OBNLM}_{\mathrm{FWHM}} = 0.186$ mm; and $\mathrm{NN}_{\mathrm{FWHM}} = 0.129$ mm, where the neural network was evaluated after being trained to minimize $\mathscr{L}_{\mathrm{Mix}}^{\star}$ for 30 epochs.

C. Cyst Simulation Results

Images reconstructed using DAS, spatial compounding, OBNLM, and the neural network are pictured in Fig. 5, along with the reference images. Image quality metrics for each method in the test dataset of simulated cysts are presented in Table III. Overall, the $\mathscr{L}_{\ell_1}^{\star}$ and $\mathscr{L}_{\ell_2}^{\star}$ losses in the relatively simple test set were lower than those found in the more complex validation set used in Figs. 2 and 3, which was composed of photographic images. DAS accurately reproduced contrast in the 3 mm lesions, and was within 3 dB for the 1 mm lesions.

Qualitatively, spatial compounding exhibited moderate speckle reduction while preserving contrast and resolution throughout the images. Spatial compounding approximately halved the test loss values. The contrast measurements of the −20 dB cyst were slightly worse than DAS, but the SNR was improved by 34%, resulting in overall improvements in CNR. OBNLM applied considerable smoothing to the speckle (497% SNR increase) and produced the lowest $\mathscr{L}^{\star}_{\ell_1}$ losses of all methods. However, OBNLM exhibited a visible loss in resolution, resulting in significantly degraded contrasts in the 1 mm cysts. Note that speckle artifacts are still visible in the background texture of the images, indicating that the smoothing parameter of OBNLM was set conservatively. We observed that relaxing the smoothing parameter resulted in worsened speckle texture without a gain in resolution, while more aggressive smoothing eliminated these artifacts but led to a further loss in resolution (not pictured). Both OBNLM and the neural network resulted in a 20-fold reduction in $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ losses from DAS. The neural network additionally had the lowest $\mathscr{L}^{\star}_{\ell_2}$ loss of all methods, and increased the speckle SNR by 488%. The neural network resulted in higher CNR than DAS, SC, and OBNLM in all cysts except for the −6 dB 3 mm cyst.

D. Phantom Imaging

Images of the tissue mimicking phantom are shown in Fig. 6. The top images show a −12 dB hypoechoic cyst and a smaller anechoic cyst, while the bottom images show a +12 dB hyperechoic cyst and point targets at the top. Contrasts and CNRs for −12 dB, −6 dB (not pictured), +6 dB (not pictured), and +12 dB 8 mm cysts are included in Table IV, along with the speckle SNR. In the DAS images, some mild signal attenuation was visible in the lower half of each image resulting in darker textures. Additionally, both the +6 dB and +12 dB hyperechoic cysts exhibited unexpectedly low contrasts of +1 dB and +5 dB in the DAS image. The DAS speckle SNR of 1.90 matched the classical prediction. Spatial compounding preserved cyst contrasts while smoothing the speckle texture, improving the SNR by 35%. Over the four cysts, spatial compounding improved the CNR on average by 67%. The OBNLM technique significantly smoothed the speckle and preserved the point targets, though the lateral edges of the hypoechoic and anechoic cysts were blurred. Overall, OBNLM improved SNR by 511% and CNR by 211%.

The neural network demonstrated excellent speckle reduction, improving the SNR by 306% over DAS. The contrasts were also preserved, resulting in higher CNRs than both DAS and spatial compounding in all cases. The neural network images also presented sharp high-resolution outlines around the −12 dB and anechoic cysts, while the outlines around the +12 dB cyst were less sharp. The images were subject to similar attenuation effects as DAS. Some dark textures and structures were visible around the deep cysts, and appeared to correspond to the speckle textures in the DAS image. The three point targets near the transducer surface were also preserved, but were slightly enlarged.

E. In Vivo Imaging

In vivo images of a kidney and a focal liver lesion are shown in Fig. 7 and image quality metrics are included in Table V. Both spatial compounding images exhibited significantly reduced speckle with marginal losses in resolution in both images. However, there was a visible loss in contrast throughout. The effects combined for a net increase in SNR of 60% and CNR of 56%. OBNLM preserved bright granular structures throughout the brighter regions of tissue but applied aggressive smoothing in darker image regions with gradual changes in echogenicity. Overall, OBNLM increased SNR and CNR over DAS by 12% and by 10%, respectively. The neural network significantly reduced speckle while improving contrast. The structures within the kidney and the lobes of the liver lesion were well-preserved, and the clutter within the surrounding fluid of the liver lesion was reduced. Shadowing effects visible in DAS were preserved in the neural network images. Overall, the SNR and CNR was increased by 90% and 93% over DAS, respectively.

V. Discussion

Beamforming with neural networks represents a fundamentally different paradigm for speckle reduction as compared to traditional techniques. Spatial compounding is a classical array processing technique that utilizes independently-beamformed subapertures to observe the target from multiple angles, reducing speckle at the cost of lateral resolution. However, improvements in speckle SNR are limited to $\sqrt{N}$ when spatial compounding with N uncorrelated images, with lesser improvements when the compounded images are correlated [6]. These effects were observed in the simulation, phantom, and *in vivo* results. OBNLM is a post-processing technique that excels at preserving sharp discontinuities and point targets. However, OBNLM struggled to maintain resolution and target structure for broader targets and gradual changes in echogenicity. OBNLM also utilizes three parameters which must be tuned precisely to achieve the desired speckle reduction. The default parameters provided by the authors of the method [16] were observed to be inadequate for the *in vivo* imaging cases shown here. Moreover, as a post-processing method, OBNLM is fundamentally subject to any noise artifacts that are present in the envelope-detected DAS image, such as clutter. By contrast, the proposed neural network is an array processing technique that performs beamforming and speckle reduction in tandem using parameters that are *learned*, nonlinearly transforming complex channel data into echogenicity estimates.

This study establishes the feasibility of using neural networks to perform ultrasound beamforming and speckle reduction. In simulations, the neural network was able to estimate the true echogenicity more accurately than DAS and spatial compounding, as measured by the $\mathscr{L}_{\ell_1}^{\star}$, $\mathscr{L}_{\ell_2}^{\star}$, and $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ losses. The neural network beamformer outputted a homogeneous response in regions of constant echogenicity while mostly preserving the shapes of the cysts.

The results also demonstrate that a beamforming neural network trained entirely *in silico* can generalize to real-world imaging, including *in vivo*. The neural network reduced speckle more effectively than receive-only spatial compounding in both the phantom and *in vivo*. The *in vivo* results were particularly remarkable considering that the training and test datasets were acquired with a different transducer (L12–3v at 8 MHz vs. 4C1 at 3.6 MHz), imaging configuration (linear vs. curvilinear), and transmit focusing scheme (full synthetic aperture vs. focused transmits), and additionally contained reverberation clutter and inhomogeneities in sound speed leading to focusing errors. The robust performance may be attributed to the wide variety of speckle patterns observed in the randomly generated training dataset. By randomly selecting the physical width, height, and position of each 64×64 pixel training patch, the networks were provided with diverse examples of speckle shapes, which was effectively equivalent to providing the networks with diverse examples of point spread functions [33]. Although the emphasis of this work was placed on the task of estimating echogenicity, the training dataset also contained simulated examples of both white electronic and band-limited acoustic noise, which may have further aided the networks in generalizing to real data.

However, the beamforming networks also exhibited several imaging artifacts. For example, the neural network resulted in slightly enlarged point targets throughout the phantom. Although this symptom implies a loss in resolution, the sharply preserved edges in the simulation, phantom, and *in vivo* indicate otherwise. The resolution measurements reported in Section IV-B further corroborate that the neural network preserves resolution in speckle targets, unlike spatial compounding and OBNLM. A potential explanation for these artifacts is that the network was trained entirely with diffuse scatterers and had never previously encountered a sharp point target. We hypothesize that providing the neural network with exposure to these sorts of targets in the training set would reduce the artifacts. Additionally, we observed that performance suffered when the neural network was provided with previously unseen speckle-to-pixel size ratios. For example, using a pixel spacing outside of the training range resulted in little to no speckle reduction. This suggests that the training data should be tailored to match the specific imaging parameters (e.g., transducer geometry, frequency, imaging depth, and focusing configuration) of the anticipated applications. Performance may also be improved by explicitly incorporating prior knowledge about the physics of ultrasound beamforming. For example, full spatial compounding on both transmit and receive is known to improve the edge definition of specular targets by interrogating the targets from multiple angles [6]; a hybridized approach with spatial compounding of neural network-beamformed images could potentially yield better visibility of specular boundaries while reducing speckle.

The networks trained to minimize $\mathscr{L}_{\ell_1}^{\star}$ achieved the best $\mathscr{L}_{\ell_1}^{\star}$ losses but also the worst $\mathscr{L}_{\ell_2}^{\star}$ and $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ losses (Fig. 2). Conversely, the network trained to minimize $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ achieved the best $\mathscr{L}_{\ell_2}^{\star}$ and $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ losses, but gave the worst $\mathscr{L}_{\ell_1}^{\star}$ loss. By minimizing both $\mathscr{L}_{\ell_1}^{\star}$ and $\mathscr{L}_{\text{MS-SSIM}}^{\star}$ simultaneously, excellent performance was achieved across all three metrics. Notably, the $\mathscr{L}_{\ell_2}^{\star}$ training objective led to erratic training. These results imply that the $\mathscr{L}_{\ell_2}^{\star}$ loss function, which is more sensitive to outliers than $\mathscr{L}_{\ell_1}^{\star}$, does not provide an efficient route for the gradient descent algorithm in its search for $\boldsymbol{\theta}^{\star}$. A similar conclusion was reached in [21], where a mixture of $\mathscr{L}_{\ell_1}$ and $\mathscr{L}_{\text{MS-SSIM}}$ was also found to outperform $\mathscr{L}_{\ell_2}$. Additionally, we observed in Fig. 4 that concatenating a B-mode image after the M-th convolutional block significantly enhanced training. Although the input channel data theoretically contain all of the raw information necessary to reconstruct a B-mode image, it appears that the concatenated B-mode image served as a good initial estimate of echogenicity for the network to improve upon, rather than learning the reconstruction from scratch.

Overall, the neural networks trained quickly and robustly. All of the presented networks were trained in under 30 minutes on a workstation equipped with a single NVIDIA GeForce GTX 1080 Ti GPU, and could process a single image frame in under 30 ms. The networks converged to nearly identically performing states for random filter weight initializations, indicating good repeatability. The networks were also observed to be insensitive to hyperparameters such as convolution filter weight regularization, which was set to values ranging from 10^{-4} to 10^{-1} with marginal differences in output.

Deeper and wider networks can represent a broader range of functions, a property called expressive power [18], [34], [35]. The benefits of expressive power were shown in Fig. 3, where the deeper and wider networks were able to lower the $\mathscr{L}^{\star}_{\ell_1}$, $\mathscr{L}^{\star}_{\ell_2}$, and $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ losses in the validation dataset more rapidly and to a lower overall value. However, expressive networks are also more prone to overfitting the training data and generalizing poorly to new data. Overfitting can be detected by observing an increase in validation loss during training, and is often caused by a dataset that is too small relative to the expressive power of the network. This type of overfitting was not exhibited in Fig. 3, which showed stable improvements in validation loss with more training epochs, suggesting that our training dataset was large enough.

A second form of overfitting can occur when the training dataset distribution differs from the testing dataset distribution, as was the case in this study. The simulated training dataset was obtained using significantly different configurations and noise conditions from the phantom and *in vivo* test data. Unfortunately, the validation loss could not be used to observe this type of overfitting because the ground truth was unavailable in the phantoms and *in vivo*. Instead, we qualitatively observed that a network trained for 30 epochs led to less speckle reduction and an over-emphasis of small speckle troughs *in vivo* as compared to a network trained for just 15 epochs. This suggests that the second form of overfitting occurred, with the network learning to recognize features in the simulated environment that were not present in the *in vivo* data. We hypothesize that generalizability to *in vivo* imaging can be improved by training the network with more realistic simulations that model full wave propagation [36] to include the effects of phase aberration, reverberation, and attenuation, as well as diffuse, specular, and point reflectors.

In addition to using a new dataset and reducing the number of training epochs, overfitting can be mitigated by employing regularization. We used the ℓ_1- and ℓ_2-norms of the filter weights to enforce smaller weights. Furthermore, we utilized *a priori* knowledge that the backscatter from diffuse scatterers has high correlation coefficients between neighboring array elements [37] to reduce the full 128 element array into 16 beamformed subapertures. Although motivated by computational constraints, subaperture beamforming effectively applied a regularization by reducing the expressivity of the network. More sophisticated forms of regularization, such as total variation [38], can be included to further reduce the impact of overfitting.

VI. Conclusion

We have presented a framework for speckle reduction in ultrasound B-mode imaging using artificial neural networks. Deep fully convolutional neural networks were trained to accept channel signals and to output speckle-reduced estimates of echogenicity. We introduced two ultrasound-appropriate modifications to the ℓ_1, ℓ_2, and MS-SSIM loss functions. First, the loss functions were measured on logarithmically compressed images to account for the large dynamic range of ultrasound signals and to match the log-domain in which the images were displayed. Second, we utilized a normalization-independent formulation to compare two images with arbitrary units. Speckle-reduction performance was analyzed as a function of training objective, network depth, and network width, with the best performance being achieved using the deepest (16 convolution layers) and widest (32 filters per layer) network, trained to minimize the $\mathscr{L}^{\star}_{\mathrm{Mix}}$ loss. After training the neural network using 5000 simulations of channel data over 30 epochs, the network was evaluated on simulations, a phantom, and *in vivo* kidney and liver. The network outperformed spatial compounding and provided comparable speckle-reduction to OBNLM with improved detail preservation while preserving resolution. In particular, the neural network improved the speckle SNR to values of 7.73 and 3.09 in a phantom and *in vivo*, respectively.

Acknowledgment

The authors would like to thank Dr. Matthew Lungren for sharing computational resources and Siemens Healthineers for their in-kind and technical support.

This research was funded through the Human Placenta Project, an initiative supported by the Eunice Kennedy Shriver National Institute Of Child Health and Human Development of the National Institutes of Health under award number R01-HD086252, and by the National Institute of Biomedical Imaging and Bioengineering grant R01-EB015506.

APPENDIX

A. The $\mathscr{L}^{\star}_{\ell_2}$ loss function

The optimal weight for the ℓ_2-norm is found as:

$$w^{\star}_{\ell_2} = \underset{w > 0}{\operatorname{argmin}} \left\| \boldsymbol{y} - w\hat{\boldsymbol{y}} \right\|_2 \tag{13}$$

$$= \underset{w > 0}{\operatorname{argmin}}\, w^2 \hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}} - 2w\hat{\boldsymbol{y}}^T \boldsymbol{y}\,. \tag{14}$$

The optimum value can be found by taking the derivative with respect to *w* and solving for the zero point, giving

$$w^{\star}_{\ell_2} = \frac{\hat{\boldsymbol{y}}^T \boldsymbol{y}}{\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}}}\,. \tag{15}$$

B. The $\mathscr{L}^{\star}_{\ell_1}$ loss function

The optimal weight for the ℓ_1-norm is obtained as:

$$w^{\star}_{\ell_1} = \underset{w>0}{\operatorname{argmin}} \sum_{p=1}^{P} \left| w - \frac{y_p}{\hat{y}_p} \right|. \tag{16}$$

The minimum, where the derivative with respect to w is zero, can be found by using the relation

$$\frac{\partial}{\partial w} \left| w - u \right| = \operatorname{sign}(w - u), \tag{17}$$

yielding an optimal solution that is the median value of $y_p/\hat{y}_p$:

$$0 = \sum_{p=1}^{P} \operatorname{sign}\left(w^{\star}_{\ell_1} - \frac{y_p}{\hat{y}_p} \right) \tag{18}$$

$$w^{\star}_{\ell_1} = \operatorname{median} \left\{ \frac{y_p}{\hat{y}_p} \right\}_{p=1}^{P}. \tag{19}$$

C. The $\mathscr{L}^{\star}_{MS-SSIM}$ loss function

The SSIM metric is the product of the differences in luminance, contrast, and structure [23], which correspond to differences in the mean, standard deviation, and the normalized signals, respectively. When computed on log-images, the optimal weight is additive. In the case of the SSIM, we find that the derivative with respect to w for one pixel p is

$$\frac{\partial}{\partial w} \operatorname{SSIM}\left(\log y_p, w + \log \hat{y}_p \right) = \tag{20}$$

$$\frac{4\mu_{\log y_p} \sigma^2_{\log y_p \log \hat{y}_p} \left(\mu^2_{\log y_p} - \left(w + \mu_{\log \hat{y}_p} \right)^2 \right)}{\left(\sigma^2_{\log y_p} + \sigma^2_{\log \hat{y}_p} \right) \left(\mu^2_{\log y_p} + \left(w + \mu_{\log \hat{y}_p} \right)^2 \right)^2}, \tag{21}$$

where the stabilizing constants C_1 and C_2 are omitted. The optimal weight is obtained by solving for w over all pixels:

$$0 = \sum_{p=1}^{P} \frac{\partial}{\partial w} \operatorname{SSIM}\left(\log y_p, w + \log \hat{y}_p \right). \tag{22}$$

Unfortunately, this equation is intractable for large P.

We instead utilize a (potentially suboptimal) weight:

$$w^{\star} = \frac{1}{PQ}\sum_{p=1}^{PQ}\log y_p - \frac{1}{PQ}\sum_{p=1}^{PQ}\log \hat{y}_p. \tag{23}$$

This is equivalent to computing the SSIM without the luminance term:

$$\mathrm{SSIM}^{\star}(\log \boldsymbol{y}, \log \widehat{\boldsymbol{y}}) = \frac{1}{PQ}\sum_{p=1}^{PQ}\frac{2\sigma_{\log y_p \log \hat{y}_p} + C_2}{\sigma^2_{\log y_p} + \sigma^2_{\log \hat{y}_p} + C_2}. \tag{24}$$

Similarly, the multi-scale luminance-independent MS-SSIM loss function is then defined as

$$\mathscr{L}^{\star}_{\mathrm{MS-SSIM}}(\log \boldsymbol{y}, \log \hat{\boldsymbol{y}}) = 1 - \prod_{j}\mathrm{SSIM}^{\star}_{j}(\log \boldsymbol{y}, \log \hat{\boldsymbol{y}}), \tag{25}$$

where j indexes the scales over which the SSIM* is computed.

Biographies

Dongwoon Hyun received the B.S.E. and Ph.D. degrees in biomedical engineering from Duke University, Durham, NC, USA, in 2010 and 2017.

He is currently a Research Engineer in the Department of Radiology at Stanford University, Stanford, CA, USA. His current research interests include machine learning methods for beamforming, molecular ultrasound imaging, and real-time software beamforming.

Leandra L. Brickson received the B.S. degree in Electrical Engineering from the University of Minnesota in 2012, and the M.S. degree in Electrical Engineering from North Carolina State University in 2015. She is currently a Ph.D. candidate in Electrical Engineering at Stanford University. Leandra's background includes optics and fabrication, and her current research focuses on image processing and deep learning of ultrasound images. Her research interests include biomedical applications of machine learning using deep learning.

Kevin T. Looby received the B.S. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign in 2013. He received the M.S. degree in Electrical Engineering from Stanford University in 2018. His graduate research focused on signal and image processing, tissue modeling, and phase aberration correction. He now works on wearable medical devices for continuous vitals monitoring.

Jeremy J. Dahl (M'11) received the B.S. degree in electrical engineering from the University of Cincinnati, Cincinnati, OH, USA, in 1999, and the Ph.D. degree in biomedical engineering from Duke University, Durham, NC, USA, in 2004.

He is currently an Associate Professor with the Department of Radiology at Stanford University, Stanford, CA, USA. His current research interests include beamforming, coherence and noise in ultrasonic imaging, speed of sound estimation, and ultrasound radiation force imaging technology.

References

[1]. Shankar PM, "A general statistical model for ultrasonic backscattering from tissues." IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 47, no. 3, pp. 727–736, 2000.

[2]. Pinton GF, Dahl JJ, and Trahey GE, "Rapid tracking of small displacements using ultrasound," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 4, pp. 2062–2065, 2006.

[3]. Abbott JG and Thurstone FL, "Acoustic speckle: Theory and experimental analysis," Ultrasonic Imaging, vol. 1, no. 4, pp. 303–324, 1979. [PubMed: 575829]

[4]. Goodman JW, Speckle phenomena in optics: theory and applications. Roberts & Company, 2007.

[5]. Wagner RF, Smith SW, Sandrik JM, and Lopez H, "Statistics of Speckle in Ultrasound B-Scans," IEEE Transactions on Sonics and Ultrasonics, vol. 30, no. 3, pp. 156–163, 1983.

[6]. Trahey GE, Smith SW, and von Ramm OT, "Speckle Pattern Correlation with Lateral Aperture Translation: Experimental Results and Implications for Spatial Compounding," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 33, no. 3, pp. 257–264, 1986.

[7]. Trahey GE, Allison JW, Smith SW, and von Ramm OT, "A quantitative approach to speckle reduction via frequency compounding," Ultrasonic Imaging, vol. 164, no. 8, pp. 151–164, 1986.

[8]. Synnevag J-F, Austeng A, and Holm S, "Adaptive beamforming applied to medical ultrasound imaging," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 54, no. 8, pp. 1606–1613, 2007.

[9]. Luchies AC and Byram BC, "Deep neural networks for ultrasound beamforming," IEEE Transactions on Medical Imaging, vol. 37, no. 9, pp. 2010–2021, 2018. [PubMed: 29994441]

[10]. Yu Y and Acton ST, “Speckle Reducing Anisotropic Diffusion,” IEEE Transactions on Image Processing, vol. 11, no. 11, pp. 1260–1270, 2002. [PubMed: 18249696]

[11]. Aja-Fernández S and Alberola-López C, “On the estimation of the coefficient of variation for anisotropic diffusion speckle filtering,” IEEE Transactions on Image Processing, vol. 15, no. 9, pp. 2694–2701, 2006. [PubMed: 16948314]

[12]. Yue Y, Croitoru MM, Bidani A, Zwischenberger JB, and Clark JW, “Nonlinear multiscale wavelet diffusion for speckle suppression and edge enhancement in ultrasound images,” IEEE Transactions on Medical Imaging, vol. 25, no. 3, pp. 297–311, 2006. [PubMed: 16524086]

[13]. Khare A, Khare M, Jeong Y, Kim H, and Jeon M, “Despeckling of medical ultrasound images using Daubechies complex wavelet transform,” Signal Processing, vol. 90, no. 2, pp. 428–439, 2010.

[14]. Tay PC, Garson CD, Acton ST, and Hossack JA, “Ultrasound despeckling for contrast enhancement,” IEEE Transactions on Image Processing, vol. 19, no. 7, pp. 1847–1860, 2010. [PubMed: 20227984]

[15]. Kervrann C, Boulanger J, and Coupé P, “Bayesian non-local means filter, image redundancy and adaptive dictionaries for noise removal,” Proc. Conf. Scale-Space and Variational Meth. (SSVM’ 07), pp. 520–532, 2007.

[16]. Coupé P, Hellier P, Kervrann C, and Barillot C, “Nonlocal means-based speckle filtering for ultrasound images,” IEEE Transactions on Image Processing, vol. 18, no. 10, pp. 2221–2229, 2009. [PubMed: 19482578]

[17]. Breivik LH, Snare SR, Steen EN, and Solberg AHS, “Real-Time Nonlocal Means-Based Despeckling,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 64, no. 6, pp. 959–977, 2017.

[18]. LeCun Y, Bengio Y, and Hinton G, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015. [PubMed: 26017442]

[19]. Jensen JA and Svendsen NB, “Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 39, no. 2, pp. 262–267, 1992.

[20]. Jensen JA, “Field: A Program for Simulating Ultrasound Systems,” Medical & Biological Engineering & Computing, vol. 34, no. 1, pp. 351–353, 1996. [PubMed: 8945858]

[21]. Zhao H, Gallo O, Frosio I, and Kautz J, “Loss Functions for Neural Networks for Image Processing,” IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47–57, 2017.

[22]. Wang Z, Simoncelli EP, and Bovik AC, “Multiscale structural similarity for image quality assessment,” in The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2 IEEE, 2003, pp. 1398–1402.

[23]. Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP, “Image quality assessment: From error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004. [PubMed: 15376593]

[24]. Jensen JA, “A multi-threaded version of Field II,” IEEE International Ultrasonics Symposium, IUS, pp. 2229–2232, 2014.

[25]. Zhou B, Lapedriza A, Khosla A, Oliva A, and Torralba A, “Places: A 10 million image database for scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452–1464, 2018. [PubMed: 28692961]

[26]. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, and Fei-Fei L, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[27]. Smith SW, Wagner RF, Sandrik JM, and Lopez H, “Low contrast detectability and contrast/detail analysis in medical ultrasound,” IEEE Transactions on Sonics and Ultrasonics, vol. 3, no. 3, pp. 164–173, 1983.

[28]. Dahl JJ, Guenther DA, and Trahey GE, “Adaptive imaging and spatial compounding in the presence of aberration,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency control, vol. 52, no. 7, pp. 1131–1144, 2005.

[29]. Ioffe S and Szegedy C, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[30]. Nair V and Hinton GE, “Rectified linear units improve restricted boltz-mann machines,” in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[31]. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X, and Brain G, “TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning,” in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ‘16), 2016.

[32]. Kingma DP and Ba J, “Adam: A method for stochastic optimization,” arXiv preprint arXiv: 1412.6980, 2014.

[33]. Wagner RF, Insana MF, and Smith SW, “Fundamental correlation lengths of coherent speckle in medical ultrasonic images.” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 35, no. 1, pp. 34–44, 1988.

[34]. Bengio Y and Delalleau O, “On the expressive power of deep architectures,” in International Conference on Algorithmic Learning Theory. Springer, 2011, pp. 18–36.

[35]. Raghu M, Poole B, Kleinberg J, Ganguli S, and Sohl-Dickstein J, “On the expressive power of deep neural networks,” arXiv preprint arXiv:1606.05336, 2016.

[36]. Pinton GF, Dahl J, Rosenzweig S, and Trahey GE, “A heterogeneous nonlinear attenuating full-wave model of ultrasound.” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 56, no. 3, pp. 474–88, 2009.

[37]. Mallart R and Fink M, “The van Cittert-Zernike theorem in pulse echo measurements,” The Journal of the Acoustical Society of America, vol. 90, no. 5, pp. 2718–2727, 1991.

[38]. Strong D and Chan T, “Edge-preserving and scale-dependent properties of total variation regularization,” Inverse problems, vol. 19, no. 6, p. S165, 2003.

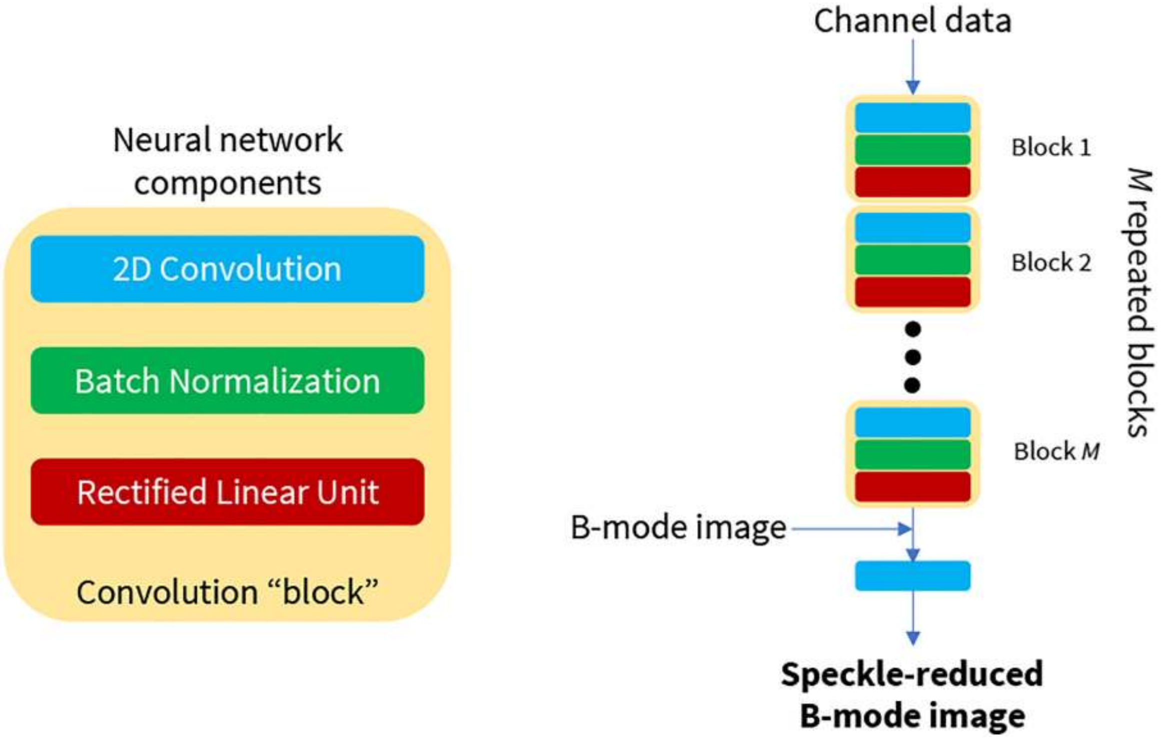
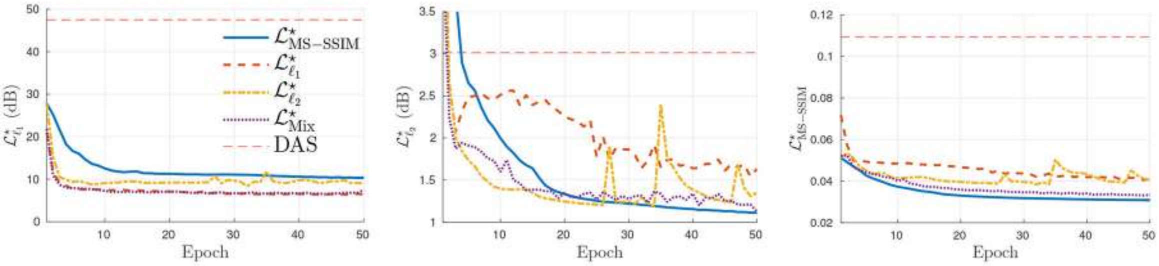


Fig. 1.

The fully convolutional neural network architecture, consisting of *M* convolution "blocks". Each block consisted of a 2D convolution, batch normalization, and a rectified linear unit. The input to the network was a 64×64×32 dataset, where the last dimension corresponded to real and imaginary components of 16 subaperture signals. A conventional envelope-detected image was also concatenated to the output of the *M*-th block. The speckle-reduced output was obtained by applying one final 2D convolution and squaring the result.



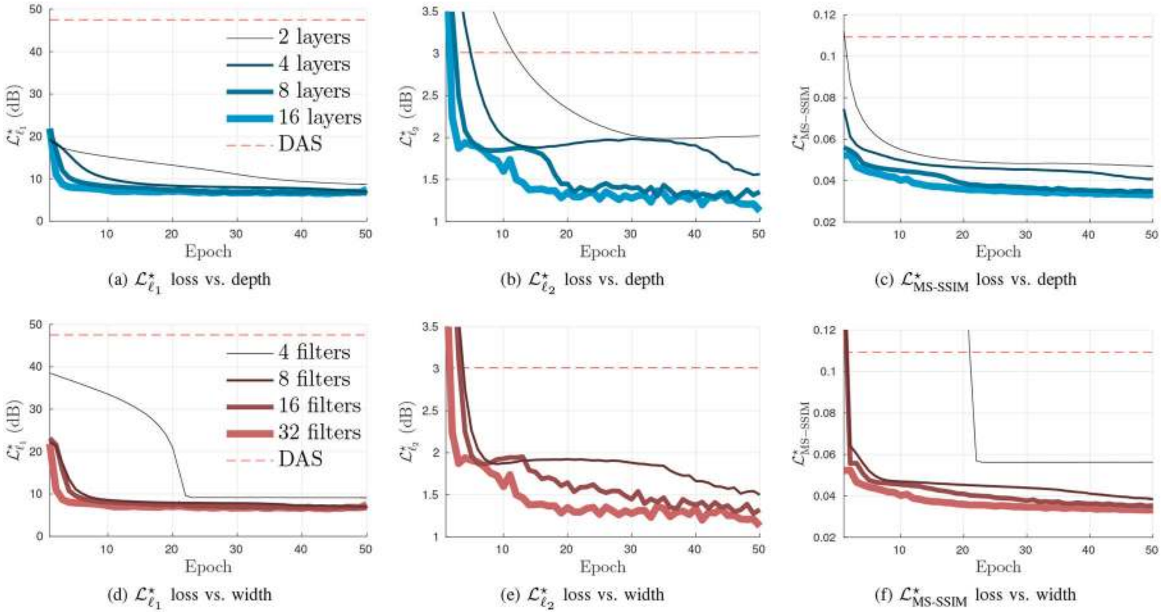
(a) $\mathcal{L}^{\star}_{\ell_1}$ loss vs. training objective

(b) $\mathcal{L}^{\star}_{\ell_2}$ loss vs. training objective

(c) $\mathcal{L}^{\star}_{\text{MS-SSIM}}$ loss vs. training objective

Fig. 2.

Validation losses versus training objectives. The neural network was trained to minimize $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ (solid), $\mathscr{L}^{\star}_{\ell_1}$. (dashed), $\mathscr{L}^{\star}_{\ell_2}$ (dot-dashed), or $\mathscr{L}^{\star}_{\text{Mix}}$ (dotted). Performance was measured using the (a) $\mathscr{L}^{\star}_{\ell_1}$, (b) $\mathscr{L}^{\star}_{\ell_2}$, and (c) $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ losses after each training epoch, as measured on the validation dataset. The losses of the DAS reconstruction (thin line) are shown for reference. Minimizing $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ and $\mathscr{L}^{\star}_{\ell_1}$ resulted in the lowest $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ and $\mathscr{L}^{\star}_{\ell_1}$ losses, respectively, but minimizing $\mathscr{L}^{\star}_{\ell_2}$ led to unstable performance. Minimizing $\mathscr{L}^{\star}_{\text{Mix}}$ resulted in fast and smooth reduction in all three losses.



(a) $\mathcal{L}^{\star}_{\ell_1}$ loss vs. depth

(b) $\mathcal{L}^{\star}_{\ell_2}$ loss vs. depth

(c) $\mathcal{L}^{\star}_{\text{MS-SSIM}}$ loss vs. depth

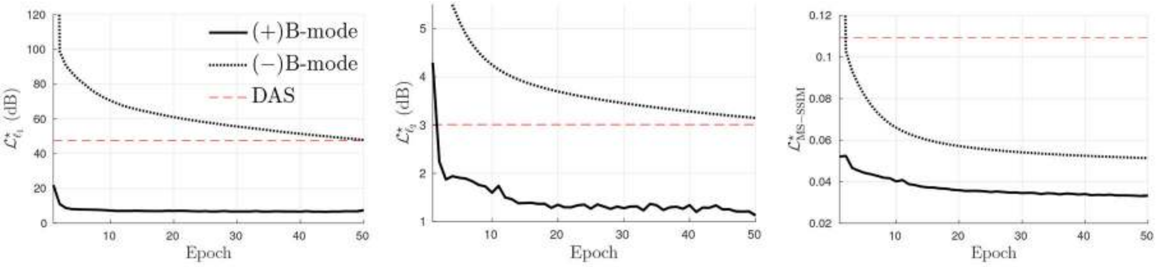
(d) $\mathcal{L}^{\star}_{\ell_1}$ loss vs. width

(e) $\mathcal{L}^{\star}_{\ell_2}$ loss vs. width

(f) $\mathcal{L}^{\star}_{\text{MS-SSIM}}$ loss vs. width

Fig. 3.

Validation losses versus network architecture. (*Top row*) Neural networks with 2, 4, 8, or 16 layers of convolution blocks and 32 filters per layer were used. (*Bottom row*) Neural networks with 16 layers of convolution blocks and 4, 8, 16, or 32 filters per layer were used. All networks were trained to minimize $\mathscr{L}^{\star}_{\text{Mix}}$. In general, it was observed that the deeper and wider networks trained more quickly and resulted in lower (a, d) $\mathscr{L}^{\star}_{\ell_1}$, (b, e) $\mathscr{L}^{\star}_{\ell_2}$, and (c, f) $\mathscr{L}^{\star}_{\text{MS-SSIM}}$ losses, as measured on the validation set.



(a) $\mathcal{L}^\star_{\ell_1}$ loss vs. concatenated B-mode (b) $\mathcal{L}^\star_{\ell_2}$ loss vs. concatenated B-mode (c) $\mathcal{L}^\star_{\text{MS-SSIM}}$ loss vs. concatenated B-mode

Fig. 4.

Validation losses with (+) and without (−) a concatenated B-mode (i.e., summed and envelope detected) image after the *M*-th convolution block. The networks were trained to minimize $\mathscr{L}^\star_{\text{Mix}}$. The network with B-mode concatenation (solid) significantly outperformed the network without B-mode (dotted) as measured by (a) $\mathscr{L}^\star_{\ell_1}$, (b) $\mathscr{L}^\star_{\ell_2}$, and (c) $\mathscr{L}^\star_{\text{MS-SSIM}}$. The network without B-mode outperformed DAS beamforming in only $\mathscr{L}^\star_{\text{MS-SSIM}}$.

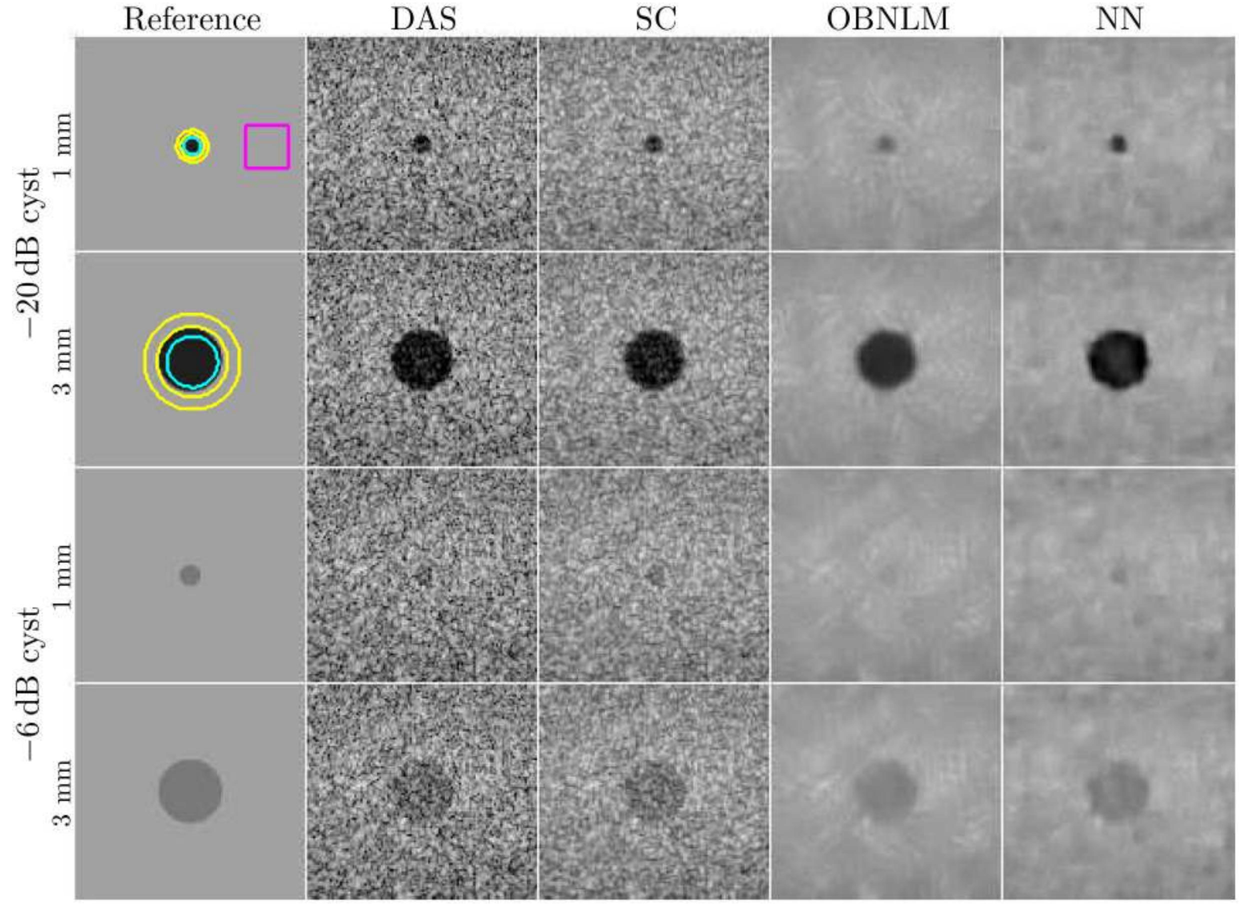
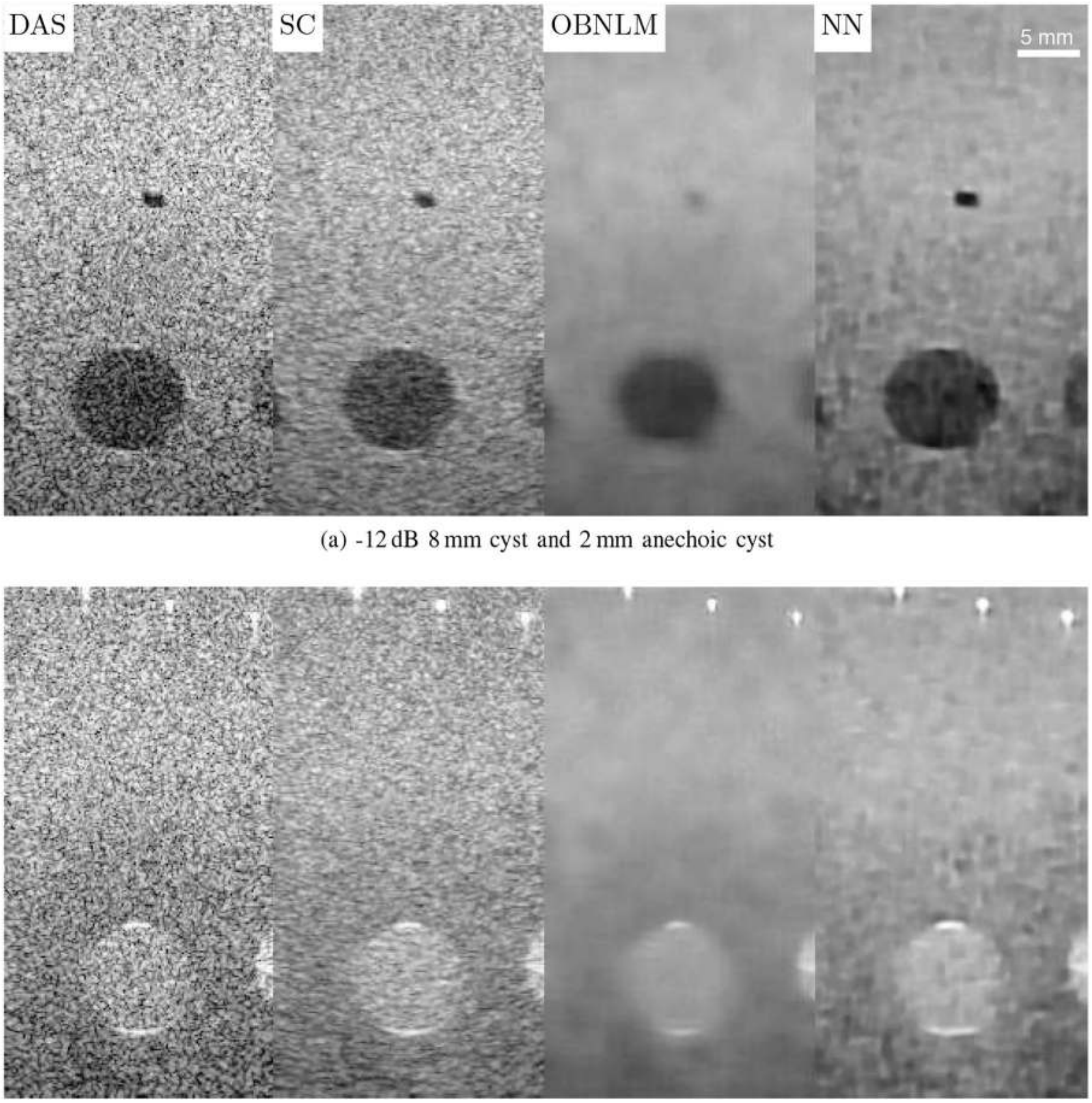


Fig. 5.

Field II simulation images of cysts were reconstructed using DAS, spatial compounding, OBNLM, and the neural network. The top two rows show 1 mm and 3 mm −20 dB cysts, while the bottom two rows show the same for −6 dB cysts. The ground truth echogenicity is shown in the leftmost column, overlaid with the circular ROIs used for contrast measurements and the square ROI used for SNR measurement.



(a) -12 dB 8 mm cyst and 2 mm anechoic cyst

(b) +12 dB 8 mm cyst and point targets

Fig. 6.

Phantom images were reconstructed using DAS, spatial compounding, OBNLM, and neural network. Each image is normalized and displays 40 dB of dynamic range. (a) A 2 mm anechoic cyst and a 8 mm −12 dB cyst are pictured. (b) A 8 mm +12 dB cyst and several bright point targets are pictured.

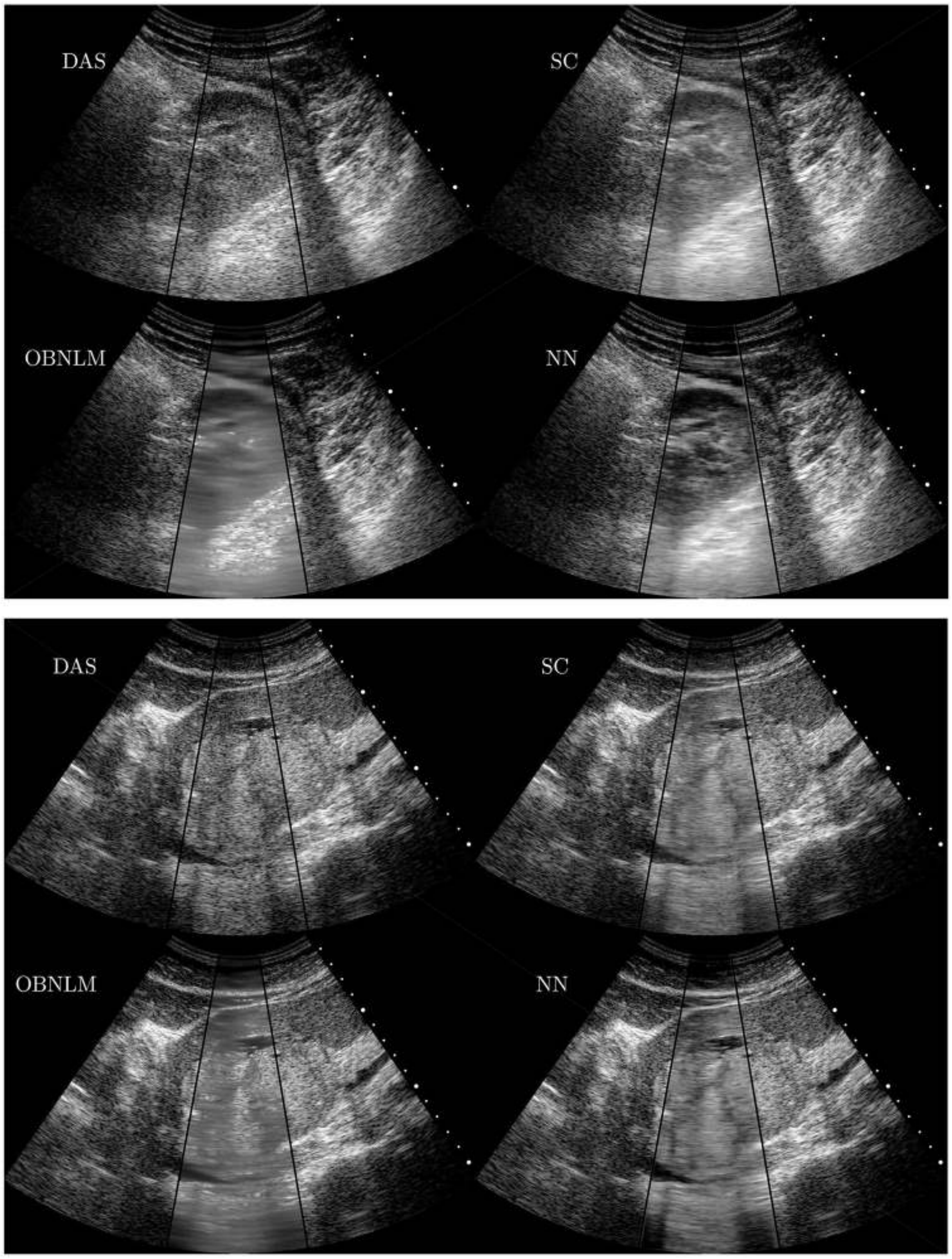


Fig. 7.

Harmonic B-mode images of a (top) kidney and (bottom) a complex focal lesion surrounded by an anechoic fluid are shown. The center sectors (bounded with black lines) were reconstructed using DAS, spatial compounding, OBNLM, and the neural network, and are overlaid on the full B-mode sector scan. The tick marks show 1 cm spacing. Each image displays 50 dB of dynamic range.

TABLE I

Range of Parameters for Random Patches

Attribute	Minimum Value	Maximum Value
Simulation Index	1	1024
Lateral Position	−1.9 mm (−10λ)	1.9 mm (10λ)
Axial Position	18.1 mm (94λ)	21.9 mm (114λ)
Lateral Size	0.4 mm (2λ)	6.2 mm (32λ)
Axial Size	0.4 mm (2λ)	6.2 mm (32λ)
Thermal Noise	−40 dB	+6 dB
Acoustical Noise	−40 dB	+6 dB

TABLE II

Network Training Hyperparameters

Hyperparameter	Value
Learning Rate	1×10^{-4}
Filter ℓ_1 Regularization	1×10^{-3}
Filter ℓ_2 Regularization	1×10^{-3}
Batch Size	128

TABLE III

Test Dataset 1: Cyst Simulation Reconstruction Metrics

Target	Metric	DAS	SC	OBNLM	NN[*]
All simulations	$\mathscr{L}^{\star}_{\ell_1}$ (dB)	28.09	12.49	4.17	4.88
All simulations	$\mathscr{L}^{\star}_{\ell_2}$ (dB)	2.84	1.13	0.20	0.18
All simulations	$\mathscr{L}^{\star}_{\text{MS-SSIM}}$	0.108	0.050	0.005	0.005
−20 dB 1 mm cyst	Contrast	−17 dB	−15 dB	−7 dB	−13 dB
−20 dB 3 mm cyst	Contrast	−20 dB	−19 dB	−18 dB	−20 dB
−6 dB 1 mm cyst	Contrast	−7 dB	−7 dB	−2 dB	−3 dB
−6 dB 3 mm cyst	Contrast	−6 dB	−6 dB	−5 dB	−5 dB
−20 dB 1 mm cyst	CNR	−1.8	−2.2	−3.6	−4.6
−20 dB 3 mm cyst	CNR	−1.6	−2.2	−6.5	−6.7
−6 dB 1 mm cyst	CNR	−1.0	−1.2	−2.4	−2.8
−6 dB 3 mm cyst	CNR	−0.8	−1.1	−3.1	−2.7
Background	SNR	1.87	2.50	11.16	11.00

[*] Trained to minimize $\mathscr{L}^{\star}_{\text{Mix}}$ for 30 epochs

TABLE IV

Test Dataset 2: Phantom Reconstruction Metrics

Target	Metric	DAS	SC	OBNLM	NN*
−12 dB cyst	Contrast	−10 dB	−10 dB	−9 dB	−11 dB
−6 dB cyst	Contrast	−5 dB	−5 dB	−5 dB	−5 dB
+6 dB cyst	Contrast	+2 dB	+2 dB	+2 dB	+2 dB
+12 dB cyst	Contrast	+5 dB	+5 dB	+5 dB	+6 dB
−12 dB cyst	CNR	−1.2	−1.6	−2.9	−2.3
−6 dB cyst	CNR	−0.8	−1.1	−2.2	−1.6
+6 dB cyst	CNR	+0.2	+0.5	+0.7	+0.8
+12 dB cyst	CNR	+0.7	+1.2	+2.9	+2.2
Background	SNR	1.90	2.57	11.60	7.73

*Trained to minimize $\mathscr{L}^{\star}_{\mathrm{Mix}}$ for 30 epochs

TABLE V

Test Dataset 3: In Vivo Reconstruction Metrics

Target	Metric	DAS	SC	OBNLM	NN*
Liver lesion	SNR	1.63	2.62	1.83	3.09
Surrounding fluid	Contrast	−20 dB	−18 dB	−19 dB	−22 dB
Surrounding fluid	CNR	−1.5	−2.3	−1.6	−2.8

*Trained to minimize $\mathscr{L}^{\star}_{\text{Mix}}$ for 15 epochs