Detection and Localization of Ultrasound Scatterers Using Convolutional Neural Networks

Youn, Jihwan; Ommen, Martin Lind; Stuart, Matthias Bo; Thomsen, Erik Vilain; Larsen, Niels Bent; Jensen, Jørgen Arendt

Published in:
IEEE Transactions on Medical Imaging

Link to article, DOI:
10.1109/TMI.2020.3006445

Publication date:
2020

Document Version
Peer reviewed version

Detection and Localization of Ultrasound Scatterers Using Convolutional Neural Networks

Jihwan Youn, Martin Lind Ommen, Matthias Bo Stuart, Erik Vilain Thomsen, Niels Bent Larsen, Jørgen Arendt Jensen, *Fellow, IEEE*

***Abstract*—Delay-and-sum (DAS) beamforming is unable to identify individual scatterers when their density is so high that their point spread functions overlap each other. This paper proposes a convolutional neural network (CNN)-based method to detect and localize high-density scatterers, some of which are closer than the resolution limit of DAS beamforming. A CNN was designed to take radio frequency channel data and return non-overlapping Gaussian confidence maps. The scatterer positions were estimated from the confidence maps by identifying local maxima. On simulated test sets, the CNN method with three plane waves achieved a precision of 1.00 and a recall of 0.91. Localization uncertainties after excluding outliers were $\pm 46\,\mu m$ (outlier ratio: 4%) laterally and $\pm 26\,\mu m$ (outlier ratio: 1%) axially. To evaluate the proposed method on measured data, two phantoms containing cavities were 3-D printed and imaged. For phantom study, training data were modified according to the physical properties of the phantoms and a new CNN was trained. On an uniformly spaced scatterer phantom, a precision of 0.98 and a recall of 1.00 were achieved with the localization uncertainties of $\pm 101\,\mu m$ (outlier ratio: 1%) laterally and $\pm 37\,\mu m$ (outlier ratio: 1%) axially. On a randomly spaced scatterer phantom, a precision of 0.59 and a recall of 0.63 were achieved. The localization uncertainties were $\pm 132\,\mu m$ (outlier ratio: 0%) laterally and $\pm 44\,\mu m$ with a bias of $22\,\mu m$ (outlier ratio: 0%) axially. This method can potentially be extended to detect highly concentrated microbubbles in order to shorten data acquisition times of super-resolution ultrasound imaging.**

***Index Terms*—high-density scatterers, convolutional neural network, super-resolution ultrasound imaging, ultrasound localization microscopy**

I. Introduction

DELAY-AND-SUM (DAS) beamforming [1] is simple and effective for B-mode image generation, but the spatial resolution is limited by wave diffraction. The resolution of conventional ultrasound imaging depends on wavelength, f-number, and excitation pulse bandwidth. Recently, ultrasound localization microscopy (ULM) and the resulting super-resolution ultrasound imaging (SRUS) were devised to overcome the diffraction limit [2]–[6]. The microvasculature, composed of vessels that are separated by less than a half-wavelength, was mapped by deploying microbubbles (MBs) as contrast agents. SRUS can be achieved by detecting and tracking the centroids of individual MBs over time.

ULM-based SRUS, however, requires long data acquisition times since the MB detection still relies on conventional ultrasound images. The ultrasound images are generally DAS beamformed and diffraction-limited as a consequence. Therefore, the MB concentration should be low to avoid overlapping point spread functions (PSFs) for accurate and reliable MB detection and localization. This constrains the number of detectable MBs in a frame, and it leads to long data acquisition times for mapping the entire target structure.

A novel method is proposed in this paper to detect and localize high-density scatterers by using convolutional neural networks (CNNs). Deep learning has had a profound impact on processing complex data and making associated decisions. By training deep neural networks with a large number of examples, impressive improvements were achieved in various challenging problems such as image classification [7]–[10], object detection [11], [12], semantic segmentation [13]–[15], and single-image super-resolution [16], [17]. It would be nearly impossible to attain such improvements using traditional logic programming or model-based approaches. The same principles can be applicable to ultrasound signals. It is hypothesized that a data-driven CNN-based method can identify scatterers laying closer than the resolution limit of DAS beamforming directly from radio frequency (RF) channel data.

In optics, where localization microscopy was firstly proposed [18]–[20], several studies have been conducted to incorporate deep learning in super-resolution localization microscopy [21]–[23]. These studies used CNNs to localize fluorescent molecules and showed that deep learning-based methods can drastically reduce data acquisition times and data processing times while achieving state-of-the-art performance.

Similar attempts also exist in SRUS. Van Sloun *et al.* [24] proposed Deep-ULM that outputs high-resolution images where the pixel values correspond to scattering intensities, given image patches of contrast-enhanced ultrasound (CEUS) acquisitions. This is similar to our approach in the sense that it handles high-density scatterer detection using CNNs, but Deep-ULM takes beamformed signals as input, whereas the proposed method only uses RF channel data without beamforming. Allman *et al.* [25] tried to locate and classify sources and artifacts from pre-beamformed photoacoustic channel data using Faster R-CNN [26] with VGG16 [27]. However, only up to 10 sources were considered, and classification for artifact removal is not necessary for scatterer detection.

Deep learning techniques have been applied to achieve better ultrasound image quality. A fully connected neural network beamformer improved image contrast by suppressing off-axis scattering [28]. Hyun *et al.* [29] proposed a CNN beamformer that reduces speckle and eventually enhances contrast while

This work was supported in part by the Fondation Idella.

The authors are with the Department of Health Technology, Technical University of Denmark, 2800 Lyngby, Denmark (email: jihyoun@dtu.dk).

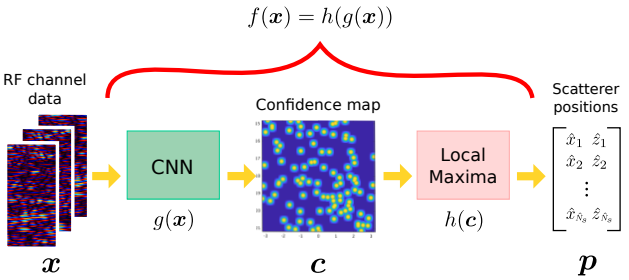


Fig. 1. Overview of the proposed scatterer detection and localization method.

preserving resolution. Generative Adversarial Network (GAN) [30], an architecture that generates output following the same distribution as training data, were applied to improve image quality without sacrificing frame rate. Multi-focus line-by-line images were synthesized from single-focus line-by-line images [31] and image quality comparable to using thirty one plane waves was achieved using three plane waves [32].

In this work, CNNs were trained to learn a mapping from RF channel data to confidence maps, and scatterer positions were then estimated from the confidence maps by identifying local maxima. The RF channel data were directly fed to the CNNs without beamforming to avoid the information loss caused by overlapping PSFs. The potential of the CNN-based method using RF channel data has been shown in [33]. Previously, however, the training was performed at a fixed scatterer density and its performance was not fully investigated. In this paper, two CNNs were trained and evaluated using simulated RF channel data generated using one plane wave or three plane waves. The training sets were generated using four different scatterer densities and the test sets were generated using ten different scatterer densities. The evaluation was performed with respect to three criteria, which are detection, localization, and resolution. Additionally, two phantoms with water-filled cavities were 3-D printed and imaged to examine the feasibility of the CNN method on measured data. Lastly, a comparison of the proposed method to Deep-ULM is discussed.

II. METHODS

Consider RF channel data $\boldsymbol{x} \in \mathbb{R}^{N_a \times N_l \times N_t}$ induced by scatterers $\boldsymbol{p} \in \mathbb{R}^{N_s \times 2}$, where N_a is the number of samples along the axial direction, N_l is the number of active elements of a transducer in reception, N_t is the number of transmissions, N_s is the number of scatterers, and 2 is the number of spatial dimensions (in the lateral and axial positions). The nonlinear mapping $f\colon \mathbb{R}^{N_a \times N_l \times N_t} \to \mathbb{R}^{N_s \times 2}$ needs to be found to estimate scatterer positions from the RF channel data, which satisfies

$$\boldsymbol{p} = f\left(\boldsymbol{x}\right). \tag{1}$$

Here N_s varies depending on the given RF channel data $\boldsymbol{x}$, so the mapping f needs to adjust N_s adaptively, but this is not straightforward. Therefore, the mapping f is decomposed into two functions g and h to handle the varying N_s. The mapping $g\colon \mathbb{R}^{N_a \times N_l \times N_t} \to \mathbb{R}^{N_h \times N_w}$ forms a confidence map $\boldsymbol{c} \in \mathbb{R}^{N_h \times N_w}$, where N_h and N_w are the number of samples in the axial and lateral directions, respectively. The confidence map $\boldsymbol{c}$ represents a region of interest (ROI) where the pixel values indicate confidences of scatterer presence in each pixel. The mapping $h\colon \mathbb{R}^{N_h \times N_w} \to \mathbb{R}^{N_s \times 2}$ detects and locates scatterers from the confidence map. The mapping in (1) can be rewritten using g and h as follows:

$$\begin{aligned}\boldsymbol{p} &= f\left(\boldsymbol{x}\right) \\ &= h\left(g\left(\boldsymbol{x}\right)\right) = h\left(\boldsymbol{c}\right),\end{aligned} \tag{2}$$

where

$$\boldsymbol{c} = g\left(\boldsymbol{x}\right). \tag{3}$$

The overview of the proposed method is illustrated in Fig. 1. The mapping g was modeled by a fully CNN and the mapping h corresponded to local maxima identification with thresholding. The RF channel data simulation and confidence map generation are explained in Section II-A and II-B, respectively. The architecture of the proposed CNN is introduced in Section II-C. Scatterer detection from the confidence maps is explained in II-D and the phantom fabrication is described in Section II-E. A baseline method for comparison is introduced in Section II-F.

TABLE I
RF CHANNEL DATA SIMULATION PARAMETERS

Category	Parameter	Value
Transducer	Transmission frequency	5.2 MHz
	Pitch	0.20 mm
	Element width	0.18 mm
	Element height	6 mm
	Number of elements	192
Imaging	Number of TX elements	32
	Number of RX elements (N_l)	64
	Steered angles	$-15°$, $0°$, $15°$
Environment	Speed of sound (c)	1480 m/s
	Field II sampling frequency	120 MHz
	RF data sampling frequency	29.6 MHz
Scatterer	Number of scatterers (N_s)	$20 \cdot i$, $\forall i \in \{1, 2, \ldots, 10\}$
	Lateral position range	$(-3.2, 3.2)$ mm
	Axial position range	$(14.8, 21.2)$ mm

A. RF Channel Data Simulation

Field II pro [34]–[36] was used to simulate RF channel data to generate data sets for training, validation, and evaluation. The parameters for the simulation are listed in Table I. The transducer was modeled after a commercial 5.2 MHz 192-element linear array transducer, and a measured impulse response [37] was applied to make the simulated RF channel data as close to measured data as possible [38].

For each frame, a certain number of point scatterers were placed randomly within a region of $6.4\,\text{mm} \times 6.4\,\text{mm}$ where the center of the region was 18 mm away from the transducer, and three steered plane waves were transmitted using 32 elements. All the simulated scatterers had the same scattering intensity. Motion and flow were not considered, therefore, the scatterers used in each frame were static in the three plane wave transmissions and the scatterer positions were independent between frames. The aperture was shifted for each steered angle to insonify only the ROI, as shown in Fig. 2. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

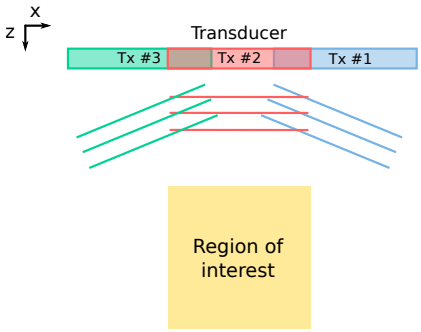


Fig. 2. Illustration of the imaging scheme. Scatterers were placed in the region of interest, and three steered plane waves were transmitted for each frame. The aperture was shifted to insonify only the region of interest.

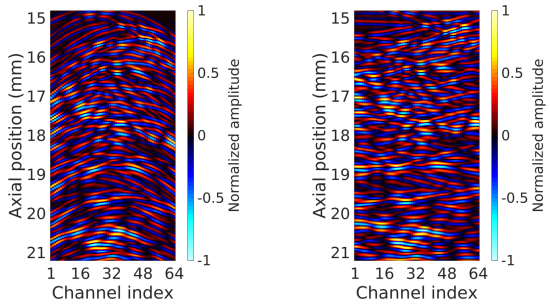


Fig. 3. An example of simulated RF channel data with one plane wave without steering. (a) is simulated raw RF channel data and (b) is delayed RF channel data. Note that the delay here is different from the delay for beamforming.

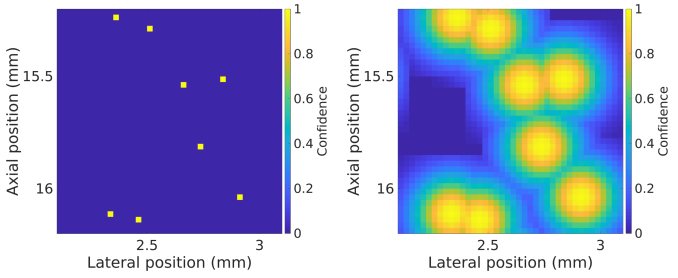


Fig. 4. An example of cropped confidence maps. (a) is a binary confidence map and (b) is a non-overlapping Gaussian confidence map created from (a).

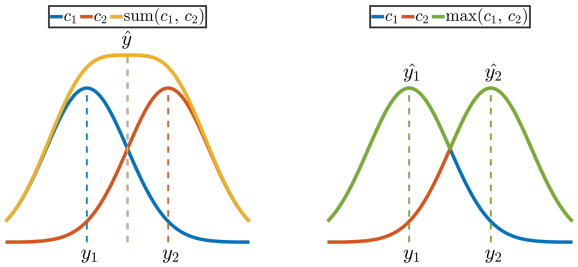


Fig. 5. A comparison of 1-D Gaussian confidence maps created by (a) summation and (b) maximum operation. There are two scatterers y_1 and y_2, and c_1 and c_2 are their confidence maps, respectively. The yellow line in (a) is the sum of c_1 and c_2. The green line in (b) is the maximum of c_1 and c_2. In (a), one scatterer $\hat{y}$ is found at a wrong position, whereas in (b), two scatterers $\hat{y}_1$ and $\hat{y}_2$ can be recovered at correct positions in the confidence map.

elements used in transmission were the 105th to the 136th ($-15°$), the 81st to the 112nd ($0°$), and the 57th to the 88th ($15°$) elements. Backscattered waves were received with 64 elements in the center of the transducer.

The simulated RF channel data were not beamformed but delayed based on the time-of-flight calculated as

$$\tau_i(x,z) = \left(\sqrt{(x-x_i)^2+z^2}+z\right)/c. \tag{4}$$

Here τ_i is the time-of-flight of the i-th transmission, (x,z) is the data point, x_i is the center of the i-th transmission aperture, and c is the speed of sound. This preprocessing helped the CNN solve the problem by making wavefronts appear more like straight lines, instead of parabolas, as shown in Fig. 3, so it is different from the delay for beamforming.

The input and output of the proposed CNN were required to have the same number of samples along the axial direction. Therefore, the delayed RF channel data were re-sampled to match the same number of samples as confidence maps along the axial direction ($N_a = N_h$). Essentially, the sampling frequency of the RF channel data was determined by the pixel size of the confidence maps, and N_a was determined by the sampling frequency and the ROI. After preprocessing, the size of RF channel data $\boldsymbol{x}$ for one frame was $256 \times 64 \times 3$ before being fed to the CNNs.

B. Non-overlapping Gaussian Confidence Map

Initially, binary confidence maps were created, where pixel values indicated presence (1) or absence (0) of a scatterer in the corresponding location, as shown in Fig. 4a. However, CNNs were not able to be trained using such confidence maps because most of their pixel values were zero. The sparse confidence maps provided small gradients during optimization and made the CNNs prone to converging to the wrong optimal solutions, returning only zero confidence maps regardless of input.

A non-overlapping Gaussian confidence map (Fig. 4b) was proposed to solve the imbalance problem of the binary confidence maps. Applying 2-D Gaussian filtering to sparse labels can improve training stability and guide CNNs to correct solutions [21], [24], [39]. But simply applying 2-D Gaussian filtering is problematic because the scatterer positions cannot be recovered in the confidence maps when the scatterers are closely spaced, as shown in Fig. 5a. To keep peaks at scatterer positions in the confidence maps, the Gaussian filter was applied one by one at each scatterer position in the binary confidence maps. Notably, when the Gaussian filter values induced by different scatterers were overlapped, the maximum values were taken instead of summation. By doing so, clearly separated peaks can be obtained at the true scatterer positions, as shown in Fig. 5b.

The parameters for non-overlapping Gaussian confidence maps are listed in Table II. The 2-D Gaussian filter is defined

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

TABLE II
CONFIDENCE MAP PARAMETERS

Parameter	Value
Pixel size	$25\,\mu\text{m}$
Confidence map size ($N_h \times N_w$)	256×256
Gaussian filter size	21 pixels
Gaussian filter standard deviation	5 pixels

by

$$G(u, v; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}}, \tag{5}$$

where u and v are the pixel distances from the scatterer position in the lateral and axial directions, respectively, and σ is the standard deviation. The filter size was fixed to $4\sigma+1$ and the standard deviation was chosen by cross-validation among 3, 5, and 7 pixels. The scatterer positions were quantized according to pixel size since the confidence maps are on discrete grids. Here the pixel size was set to $25\,\mu\text{m}$ ($\approx \lambda/10$); the lateral and axial localization uncertainties are $\pm 12.5\,\mu\text{m}$ in an ideal situation. The confidence map size was 256×256 ($N_w = N_h = 256$) given the pixel size and the area of the ROI.

C. Convolutional Neural Network Architecture

The proposed CNN has an encoder-decoder structure with pooling and unpooling, similar to U-Net [13] but without skip connections. The encoder-decoder structure was adopted to transform the input in the channel data domain to the confidence map in the ultrasound image domain. In the encoding path, information is extracted from the RF channel data, and in the decoding path, the confidence maps are reconstructed based on the extracted information.

The overview of the CNN architecture and its components are shown in Fig. 6. It mainly consists of four *down-blocks*, one *conv-block*, and four *up-blocks*. In the *down-blocks*, the feature map size is decreased by strided convolution to reduce the amount of parameters, and in the *up-blocks*, the feature map size is increased to the confidence map size by pixel shuffle [40]. An 11×1 convolution layer prior to the encoding path extracts per-channel features, and two convolution layers after the decoding path refine the feature maps and return the confidence maps.

The pre-activation residual units [9] (Fig. 6a) were used instead of common convolution and rectified linear unit (ReLU) layers to improve the network performance. Batch normalization (BN) in the residual units helped ease the optimization, limited covariate shift, and had the effect of regularization [41]. Dropout [42] was additionally attached after the shortcut for further regularization. Leaky ReLU [43] and Sigmoid were chosen as non-linear activation. CoordConv [44] was added to transfer spatial information over convolution layers.

The same CNN architecture was used for both one and three plane wave data. For three plane waves, the preprocessed RF channel data from each transmission in a frame were stacked along the third dimension before applied to a CNN.

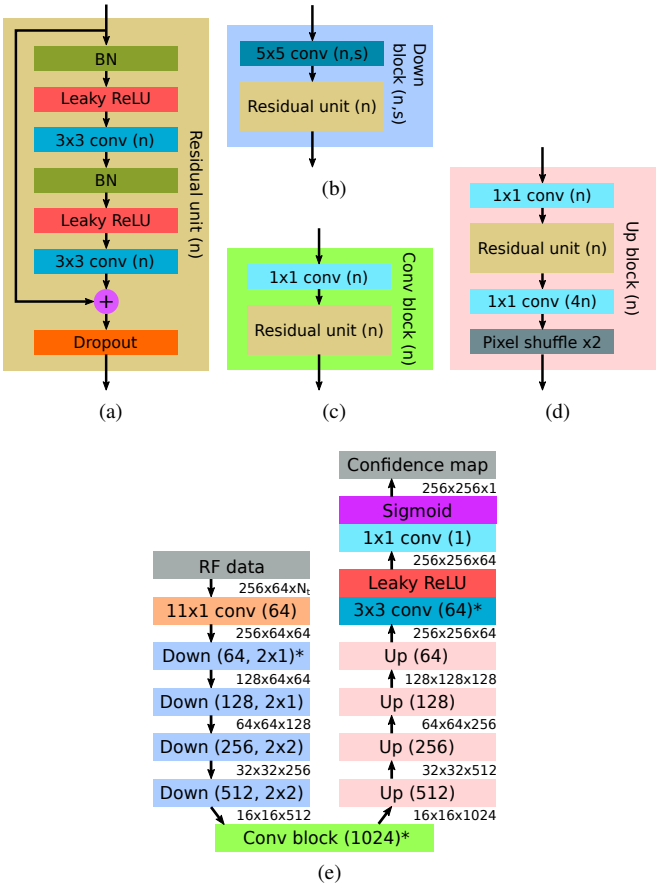


Fig. 6. The proposed CNN architecture and its components: (a) residual unit, (b) down-block, (c) conv-block, (d) up-block, and (e) the network overview. The n and s in the parenthesis are the number of kernels and stride. In (e), the sets of three numbers are the feature map size between two blocks, and the asterisk indicates that CoordConv was applied at the first convolution in the block.

D. Scatterer Detection from Confidence Maps

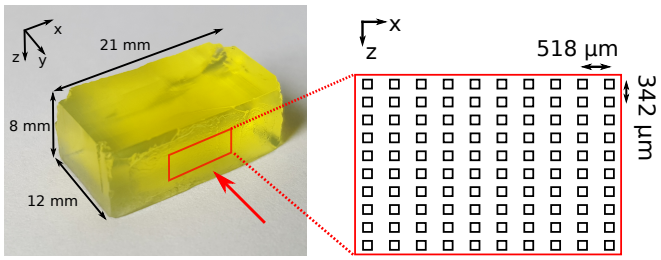
The scatterer positions can be found by locating the pixels whose confidences are one in the true confidence map $\boldsymbol{c}$. However, the estimated confidence map $\hat{\boldsymbol{c}} = g(\boldsymbol{x})$ acquired from a trained CNN is an approximation of $\boldsymbol{c}$. It is not guaranteed that the confidences are one where scatterers are located in $\hat{\boldsymbol{c}}$. Therefore, the algorithm relied on the fact that pixels containing scatterers are local peaks. The scatterer positions were recovered by finding the local maxima whose confidence is larger than a certain decision value. The chosen decision value was 0.9 in this work.

E. Phantom Fabrication

Two PEGDA $700\,\text{g/mol}$ hydrogel phantoms were 3-D printed [45], [46] to assess the CNN method on measured data. The phantoms contained water-filled cavities which acted as scatterers. The volume of each cavity was $45\,\mu\text{m} \times 1000\,\mu\text{m} \times 45\,\mu\text{m}$. The cavities were designed to be elongated in the elevation direction to increase the intensity of received signals.

In the first phantom, 100 cavities were placed on a 10×10 grid with a spacing of $518\,\mu\text{m}$ in the lateral direction and $342\,\mu\text{m}$ in the axial direction, as illustrated in Fig. 7. This grid scatterer phantom had the spacing larger than the resolution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging



(a) (b)

Fig. 7. Fabricated 3-D phantom with uniformly spaced cavities: (a) photograph of the phantom and (b) 100 cavities placed on a 10×10 grid.

limit of DAS to show that the CNN method works on measured data. The second phantom, on the other hand, had 100 cavities randomly distributed with a minimum spacing of $190\,\mu\mathrm{m}$ to demonstrate that the CNN method can resolve targets closer than the conventional resolution limit. The minimum spacing between cavities were constrained due to the cavity size and the 3-D printer voxel size.

F. Baseline Method

Local peak detection on the beamformed images was chosen as a baseline method for comparison. RF channel data were DAS beamformed in the region of interest with the same pixel size as the confidence map, and, for three plane wave transmissions, beamformed images in a frame were coherently compounded [47]. The baseline method detected and located scatterers in the envelope detected and log-compressed B-mode images with a dynamic range of $40\,\mathrm{dB}$. The B-mode images were smoothed to avoid more than one pixel corresponding to a peak, and scatterer positions were estimated by finding local maxima.

Deconvolution using an estimated PSF is one of the commonly used techniques for microbubble localization [5]. However, it was not considered in this work since its performance has been found to be sensitive to parameters when the PSFs were highly overlapped, and the spatially varying PSF of ultrasound imaging resulted in imprecise scatterer localization.

III. Experiments

A. Training Details

CNNs, which correspond to the mapping g in (2), were trained to return the corresponding confidence map $\boldsymbol{c}_i$ given RF channel data $\boldsymbol{x}_i$ by minimizing the mean squared error (MSE), given by

$$\mathcal{L}_{\mathrm{MSE}}\left(\boldsymbol{x}_i, \boldsymbol{c}_i; g\right) = \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{c}_i - g\left(\boldsymbol{x}_i\right) \right\|_F^2, \tag{6}$$

where N is the number of samples and $\|\cdot\|_F$ is the Frobenius norm.

One data set consisted of frames simulated at the same scatterer density, and four training sets and four validation sets were generated at the scatterer densities of $0.49\,\mathrm{mm}^{-2}$, $0.98\,\mathrm{mm}^{-2}$, $2.44\,\mathrm{mm}^{-2}$, and $4.88\,\mathrm{mm}^{-2}$, i.e., the numbers of scatterers were 20, 40, 100, and 200 in one frame, respectively. Each training set and validation set had $10\,240$ and 1280 frames, respectively.

The kernel weights were initialized by orthogonal initialization [48] and optimized with ADAM [49] by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. Firstly, the training was performed using only the training set at the scatterer density of $2.44\,\mathrm{mm}^{-2}$. The initial learning rate was 10^{-4} and it was halved every 100 epochs. After 600 epochs, the learning rate was set to 10^{-5} and the training continued using all the training sets while the learning rate was halved every 50 epochs. The mini-batch size was 32, and each batch was composed of frames from all four training sets after 600 epochs. The CNN was implemented in Python using Tensorflow [50], and were trained on a server equipped with a NVIDIA TESLA V100 16 GB PCIe graphics card. The total number of training epochs was 800, and the training took approximately 40 hours.

During training, the RF channel data and confidence maps were flipped along the lateral direction at random with a probability of 0.5 to augment the training sets. White Gaussian noise was added to the RF channel data for generalization along with BN and dropout. The signal-to-noise ratio after noise addition was $6\,\mathrm{dB}$, and the dropout rate was 0.3. The RF channel data and confidence maps were then normalized to be in the range $[-1, 1]$ and $[0, 1]$, respectively. Validation was performed every epoch to monitor the training, and also for cross-validation to choose hyper-parameters.

For both simulation and phantom experiment, two CNNs were trained and compared: one CNN acting on the data from one plane wave ($0°$) and the other CNN acting on the data from three plane waves ($-15°$, $0°$, $15°$).

B. Simulation Experiment

The CNNs were evaluated on simulated test sets firstly. One test set consisted of 3840 frames simulated at the same scatterer density, and ten test sets were created at scatterer densities from $0.49\,\mathrm{mm}^{-2}$ to $4.88\,\mathrm{mm}^{-2}$ by varying the number of scatterers from 20 to 200 with intervals of 20. The parameters in Table I were used again, apart from the number of scatterers. The evaluation was performed on the frames simulated at various scatterer densities to evaluate how the performance changes over different scatterer densities and how well the CNNs were generalized in terms of scatterer density.

C. 3-D Printed Phantom Experiment

1) RF Channel Data Acquisition: The 3-D printed phantoms were scanned using the $5.2\,\mathrm{MHz}$ 192-element linear array transducer which has the same parameters as in Table I. The raw RF channel data were acquired by the synthetic aperture real-time ultrasound system (SARUS) experimental ultrasound scanner [51]. The same imaging scheme and processing as in the simulation were applied.

The experimental setup is shown in Fig. 8. The transducer was fixed, and a water tank containing the phantom was placed on a motion stage. The phantom was aligned with the transducer by the motion stage, capable of translating

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

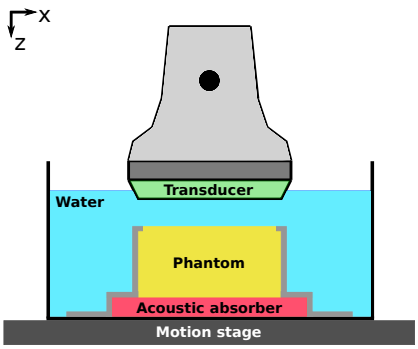


Fig. 8. Illustration of the experimental setup for phantom measurement.

in the x- and y-axis, and rotating around the z-axis. During measurement, the motion stage was translated along the x-axis in steps of $50\,\mu\text{m}$ between frames, and 33 frames were acquired for each phantom experiment.

2) Training Set Modification: The training sets were modified and new CNNs were trained from scratch for the phantom experiment. Transfer learning was also considered but it did not show as good performance as training from scratch. In the simulation, it was assumed that scatterers were infinitesimally small points. However, the cavities in the phantoms were squares, as shown in Fig. 7b, if the elevation direction is ignored. Scattering, therefore, happens twice at each cavity: once when a wave goes into the cavity and the other when the wave comes out of the cavity. Additionally, the first scattering experiences a phase reversal because the acoustic impedance of the phantoms is higher than that of water.

RF channel data for training were accordingly re-simulated by modeling each scatterer using two points separated by the cavity size axially and with a phase reversal. To remain consistent, the same scatterer positions of the original training set were used.

3) Depth Correction: The speed of sound in the phantoms is higher than in water. The axial positions of the estimated scatterers were corrected to compensate for the different speed of sound in the phantoms by

$$\hat{z}^* = (\hat{z} - d_{\text{pht}}) \cdot \frac{c_{\text{water}}}{c_{\text{pht}}} + d_{\text{pht}}, \tag{7}$$

where $\hat{z}$ and $\hat{z}^*$ are the axial position before and after correction, c_{water} and c_{pht} are the speed of sound in water and in the phantoms, respectively, and d_{pht} is the distance from the transducer to the surface of the phantoms.

D. Evaluation Metrics

Three evaluation criteria were considered to assess the CNNs: detection, localization, and resolution. The positive and negative detections were determined by pairing estimated scatterers with true scatterers based on their pair-wise distances, as stated in Algorithm 1. Namely, to be a positive detection, an estimated scatterer should be exclusively matched with a true scatterer within a certain localization precision. This localization precision can be translated to the target resolution of ULM without tracking. It was set to be half of the full width at half maximum (FWHM) in this work. Specifically, an

Algorithm 1 Algorithm for determining positive or negative detections

Input: $\boldsymbol{p} \in \mathbb{R}^{N_s \times 2}$ and $\hat{\boldsymbol{p}} \in \mathbb{R}^{\hat{N}_s \times 2}$, where $\boldsymbol{p}$ is true scatterer positions and $\hat{\boldsymbol{p}}$ is estimated scatterer posions

Output: Positive or negative detection $\boldsymbol{a} \in \mathbb{R}^{\hat{N}_s \times 1}$

$\boldsymbol{a} \leftarrow \boldsymbol{0} \in \mathbb{R}^{\hat{N}_s \times 1}$
$D \leftarrow \left\{ (d_{ij}) \in \mathbb{R}^{N_s \times \hat{N}_s} \,\middle|\, d_{ij} = \|p_i - \hat{p}_j\|_2 \right\}$
for $j = 1$ to $\hat{N}_s$ **do**
  $\hat{i} \leftarrow \arg\min D_{*,j}$
  if $j = \arg\min D_{\hat{i},*}$ **and** $\frac{(p_{\hat{i}1} - \hat{p}_{j1})^2}{(\text{FWHM}_x/2)^2} + \frac{(p_{\hat{i}2} - \hat{p}_{j2})^2}{(\text{FWHM}_z/2)^2} < 1$ **then**
    $a_j \leftarrow 1$
  else
    $a_j \leftarrow 0$
  end if
end for

ellipse whose major axis and minor axis were half of FWMH_x and half of FWMH_z, respectively, was used as the desired localization precision, where FWMH_x is the lateral FWHM and FWMH_z is the axial FWHM. This bi-directional matching process was extended from the left-right consistency check [52], [53] for stereo matching in computer vision. It conforms to the uniqueness constraint; one true scatterer can be paired with at most one estimated scatterer.

Detection capability was assessed by quantifying wrong detections and missed detections using precision, recall, and F_1 score, which are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{9}$$

and

$$\text{F}_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{10}$$

where TP is the number of true positives (correct detections), FP is the number of false positives (wrong detections), and FN is the number of false negatives (missed detections).

Localization uncertainties were measured by calculating the lateral and axial position errors. Only positive detections were considered for the localization assessment.

Spatial resolution, meaning the ability to separate two points that are close together, was investigated statistically. For two isolated true scatterers, it was checked whether they were detected. A pair of scatterers was set to *resolved* if both scatterers were detected. It was set to *non-resolved* if only one of them was detected. And it was not considered if none of them were detected, as this would be a detection problem. The resolved rates were calculated in $20\,\mu\text{m} \times 20\,\mu\text{m}$ bins by

$$\text{Resolved rate} = \frac{N_{\text{res}}}{N_{\text{res}} + N_{\text{non-res}}}, \tag{11}$$

where N_{res} is the number of resolved pairs and $N_{\text{non-res}}$ is the number of non-resolved pairs in a bin.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

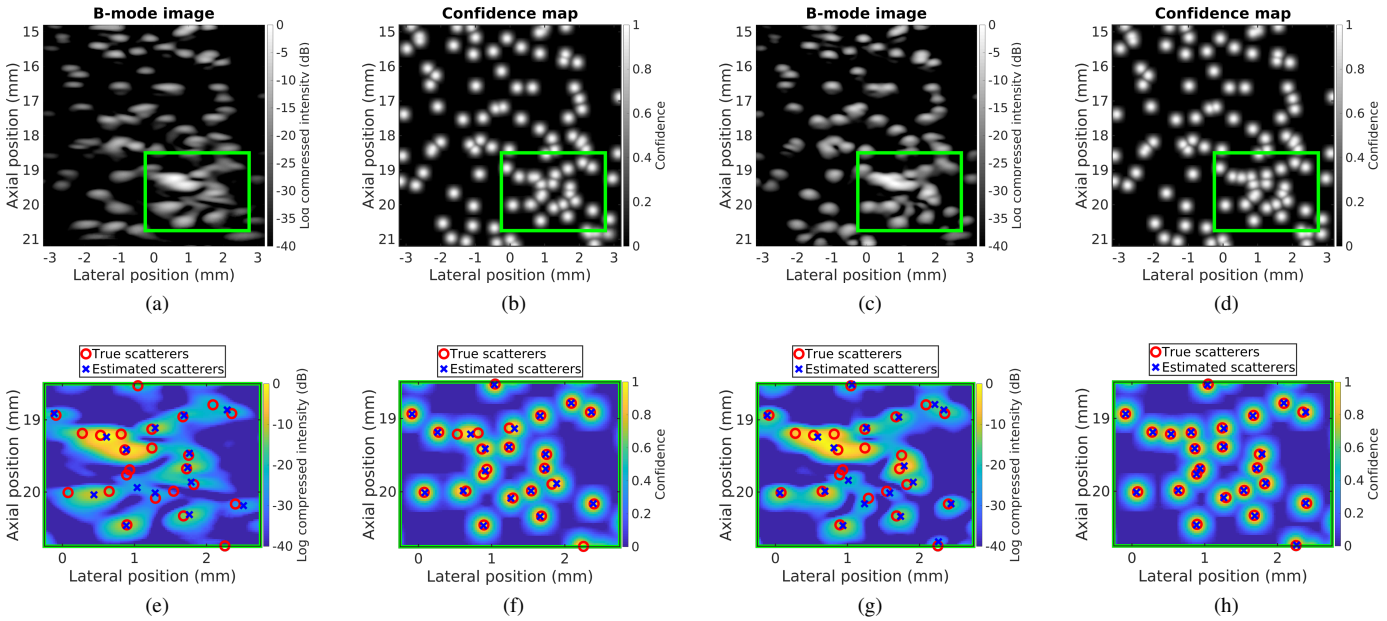


Fig. 9. A comparison of scatterer detection between baseline method and CNN method on a simulated test frame. (a) and (c) are DAS beamformed B-mode images with one and three plane waves, respectively. (b) and (d) are estimated confidence maps by CNNs with one and three plane waves, respectively. (e) - (h) show true scatterers and estimated scatterers from their corresponding results above in the same column in the green box region.

TABLE III
PRECISION, RECALL, AND F_1 SCORE COMPARISON ON THE SIMULATED TEST SETS

Method	One plane wave			Three plane waves		
	Precision	Recall	F_1	Precision	Recall	F_1
Baseline	0.83	0.51	0.63	0.93	0.62	0.75
CNN	0.99	0.83	0.90	1.00	0.91	0.96

IV. RESULTS

The CNN method results on the simulated data and the measured data of the 3-D printed phantoms presented in this Section. Quantitative evaluation comparing one plane wave and three plane waves was performed as specified in Section III-D. The results of the baseline method on the same test data are also presented for comparison.

A. Simulation Experiment

The qualitative comparison between the baseline and CNN methods is shown in Fig 9. The proposed CNN method successfully detected and localized high-density scatterers when the baseline method failed due to overlapping PSFs.

The detection results on the simulated test sets are shown in Table III. The CNN method achieved the better precision, recall, and F_1 score for both one and three plane transmissions. Also, when the higher number of transmissions was involved, the detection performance was improved for both methods. The detection capabilities over different scatterer densities were investigated, as shown in Fig. 10. The recalls dropped as the scatterer density increased while the precisions were relatively kept high. In addition, the recalls of the baseline method decreased more drastically as the scatterer density increased, which led to the lower F_1 scores.

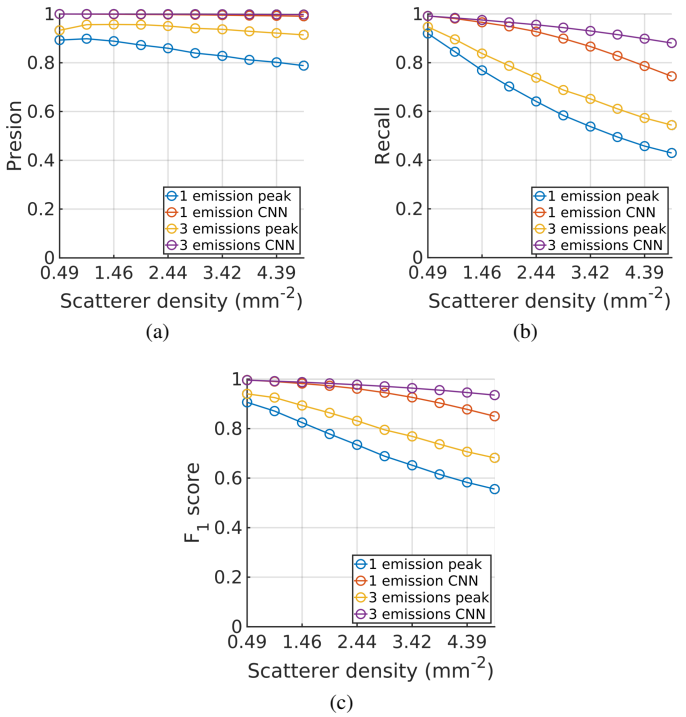


Fig. 10. Detection capabilities of the baseline and CNN methods over different scatterer densities on the simulated test sets with one and three plane waves: (a) precision, (b) recall, and (c) F_1 score.

The comparison of localization uncertainties between the baseline and CNN methods on the simulated test sets are presented in Fig. 11, using box-and-whisker plots along with violin plots. The bottom and top edges of the blue boxes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

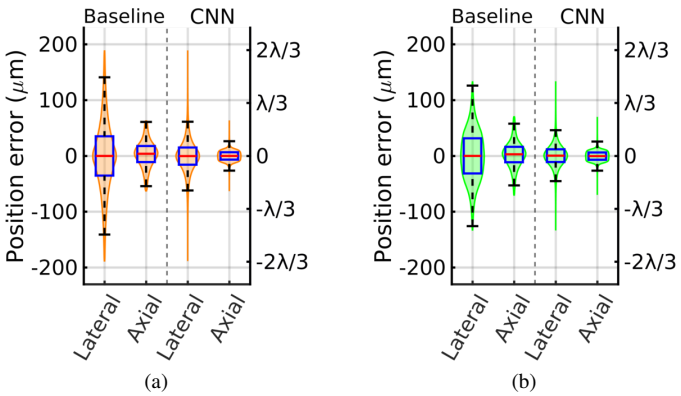


Fig. 11. Localization uncertainties of baseline and CNN methods on the simulated test sets. (a) and (b) are the results with one plane wave and three plane waves, respectively.

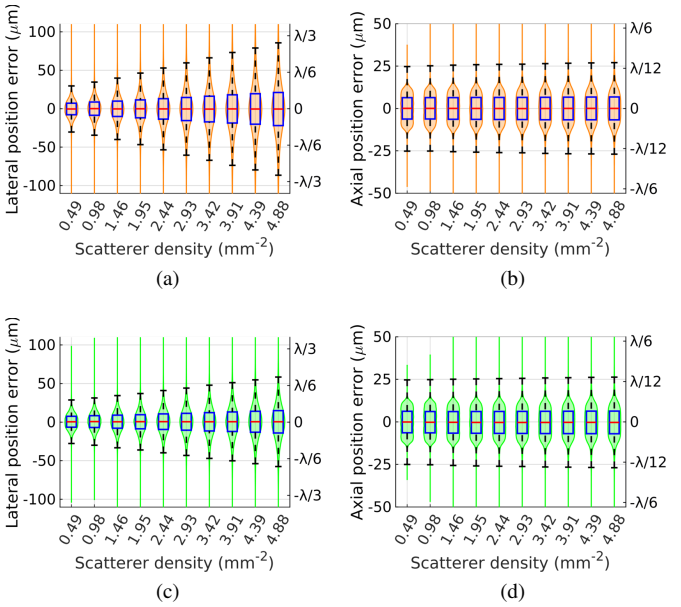


Fig. 12. Localization uncertainties of the CNN method on the simulated test sets at different scatterer densities: the lateral position errors with (a) one plane wave and (b) three plane waves, and the axial position errors with (c) one plane wave and (d) three plane waves.

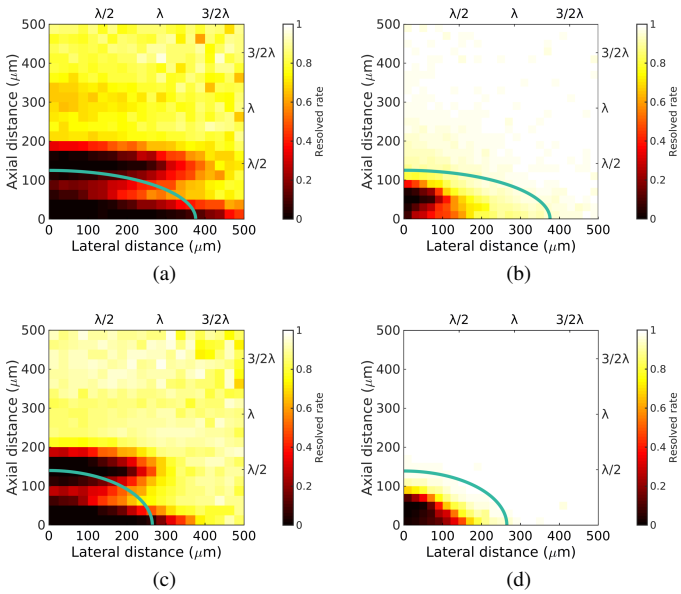


Fig. 13. Resolved rate of (a), (c) baseline methods and (b), (d) CNN methods on the simulated test sets where (a) and (b) are with one plane wave and (c) and (d) are with three plane waves. The green lines represent the theoretical resolution limit of DAS beamforming.

indicate the 25 th (q_1) and 75 th percentiles (q_3), and the center red lines indicate the medians. The whiskers, vertically extended lines from the boxes, indicate the range of values except outliers, which are greater than $q_3 + 1.5 \times (q_3 - q_1)$ or less than $q_1 - 1.5 \times (q_3 - q_1)$. The violin plots were overlaid as shaded area to demonstrate the error distribution directly. For both methods, the lateral position error was higher than the axial position error, and the CNN method achieved clearly better localization than the baseline method. For the most part the medians were very close to zero, indicating that the scatterer position estimation was unbiased in both directions. The localization was also improved when more plane waves were transmitted. Localization uncertainties of the CNN method at different scatterer densities are shown in Fig. 12. Neither the scatterer density nor the number of transmissions had much impact on the axial position errors. The lateral position errors, on the other hand, gradually increased as the scatterer density increased.

The 2-D histograms in Fig. 13 show the resolved rates of two isolated scatterers measured in $20\,\mu\text{m} \times 20\,\mu\text{m}$ bins. The green lines represent the theoretical resolution limit of DAS beamformed images, assuming that the $6\,\text{dB}$ contour of a PSF is an ellipse. The FWHM was measured on a simulated PSF in the center of the ROI. For one plane wave, the FWHM was $376\,\mu\text{m}$ ($1.32\,\lambda$) laterally and $125\,\mu\text{m}$ ($0.44\,\lambda$) axially. For three plane waves, the FWHM was $265\,\mu\text{m}$ ($0.93\,\lambda$) laterally and $140\,\mu\text{m}$ ($0.49\,\lambda$) axially. The resolution results show that the CNN method can resolve scatterers closer than the DAS limit. The mean resolved rates in the area under the green line for the baseline and CNN methods were 0.16 and 0.68 with one plane wave, and 0.17 and 0.67 with three plane waves, respectively.

B. 3-D Printed Phantom Experiment

For the phantom study, CNNs were applied to measured data without evaluation on simulated test data. The qualitative results of the baseline and CNN methods on the grid and random scatterer phantoms are presented in Fig. 14 and their quantitative comparison is shown in Table IV and Fig. 15.

With one plane wave, side lobe level was high, and side lobes were added up when the scatterers were placed in a grid. Therefore, the DAS beamforming was unable to identify individual scatterers of the grid phantom properly, as shown in Fig. 14a. The CNN method also achieved poor detection with one plane wave on the grid phantom, as shown in Fig. 14b. The CNN was not generalized sufficiently to handle regularly placed scatterers as the training frames were generated by placing scatterers randomly. Most of the scatterers in the first and the last columns were correctly detected, but the other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

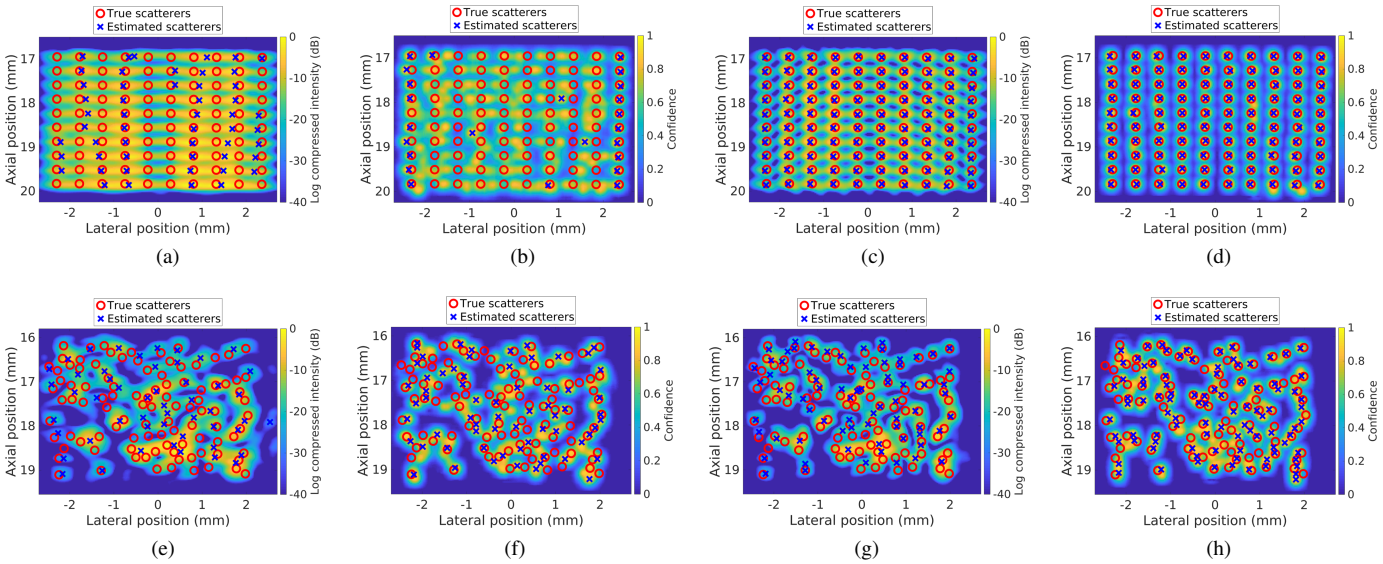


Fig. 14. A comparison of scatterer detection between baseline method and CNN method on phantom measured frames. (a) - (d) are results of the grid phantom and (e) - (h) are results of the random phantom. B-mode images with (a), (e) one plane wave and (c), (g) three plane waves and confidence maps with (b), (f) one plane wave and (d), (h) three plane waves are shown with true scatterers and estimated scatterers.

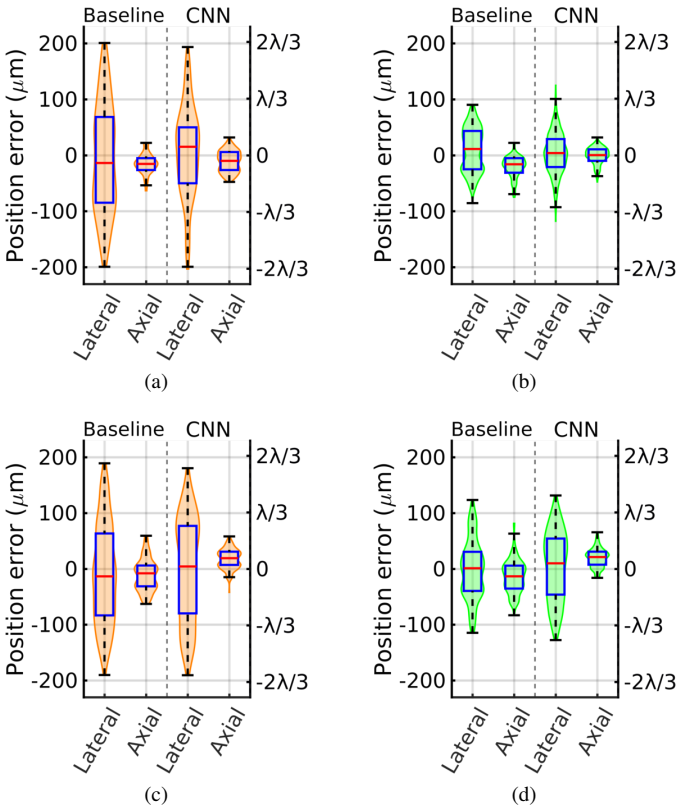


Fig. 15. Localization uncertainties of baseline and CNN methods on phantom measured data: (a) and (b) are results on the grid scatterer phantom with one and three plane waves, respectively. (c) and (d) are results on the random scatterer phantom with one and three plane waves, respectively.

TABLE IV
PRECISION, RECALL, AND F_1 SCORE COMPARISON ON THE PHANTOM TEST SETS

Phantom	Method	One plane wave			Three plane waves		
		Precision	Recall	F_1	Precision	Recall	F_1
Grid	Baseline	0.82	0.41	0.54	1.00	1.00	1.00
	CNN	0.89	0.22	0.35	0.98	1.00	0.98
Random	Baseline	0.47	0.23	0.31	0.49	0.32	0.39
	CNN	0.53	0.37	0.44	0.59	0.63	0.61

scatterers were missed. Thus, the precision was higher than the baseline but the recall was lower. On the contrary, with three plane waves, the baseline method found all the scatterers without any false detection. The CNN method also achieved comparable detection results with three plane waves, showing that more transmissions for a frame helped generalization of the CNN. For localization, the CNN method showed slightly smaller uncertainties except the axial localization with one plane wave.

On the random scatterer phantom, the CNN method achieved better detection for both one and three plane waves. For localization, the CNN method showed smaller axial uncertainties but little higher lateral uncertainties. With three plane waves, the detection and localization were improved but, in general, it was more challenging to identify scatterers for both methods on the random scatterer phantom.

V. DISCUSSION

A CNN-based scatterer detection and localization method is presented. Instead of end-to-end training, the CNNs were trained to learn the mapping from RF channel data to non-overlapping Gaussian confidence maps, and scatterers were detected and localized from the confidence maps by looking

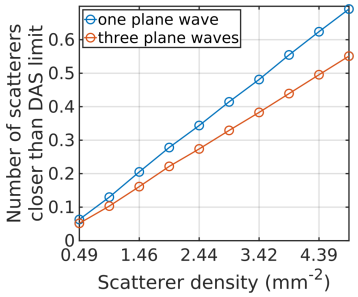
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3006445, IEEE Transactions on Medical Imaging

for local maxima. This two-step framework made it possible to handle varying numbers of scatterers (N_s). By obtaining non-overlapping Gaussian confidence maps from RF channel data without beamforming, it was able to identify high concentrations of scatterers which cannot be separated by conventional ultrasound imaging due to the overlapping PSFs. This method also has an advantage of fast processing by exploiting GPU computation. The proposed CNN implicitly included beamforming since it is a mapping from the channel domain to the ultrasound image domain, which is a bottleneck of current ultrasound imaging. For the CNNs, processing time for a frame was $16\,\mathrm{ms}$ on average in a PC equipped with a NVIDIA Titan V graphics card.

It was essential to use non-overlapping Gaussian confidence maps to make training work. Binary confidence maps were initially used to train CNNs with advanced loss functions such as weighted cross entropy [13], jaccard loss [54], or focal loss [55], as well as simple loss functions such as MSE or mean absolute error, but all of them failed. The binary confidence maps were too sparse to be handled by simply manipulating the loss function. However, non-overlapping Gaussian confidence maps relaxed the sparsity of the binary confidence maps while being able to recover scatterer positions by taking the maximum of overlapping Gaussians. Therefore, the larger gradients were provided during training and the CNNs were able to be guided to the correct solutions stably.

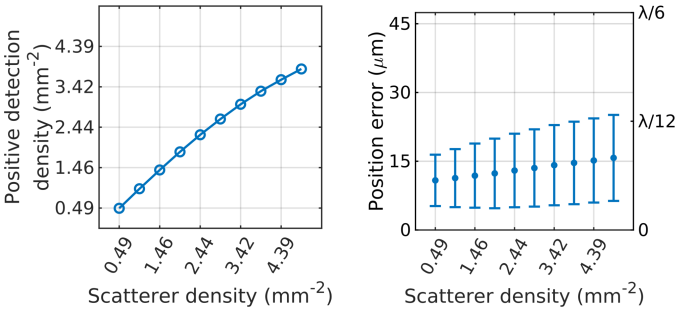
The training was firstly performed in the training set at the scatterer density of $2.44\,\mathrm{mm}^{-2}$, and was further performed on the whole training sets later. Interestingly, the CNNs trained at the scatterer density of $2.44\,\mathrm{mm}^{-2}$ were already well generalized at the scatterer densities higher than $2.44\,\mathrm{mm}^{-2}$. On the other hand, the CNNs achieved poor precision and localization at the lower scatterer densities as two Gaussian peaks appeared laterally near a true scatterer position in the confidence maps. Therefore, the training sets had more frames at the lower scatterer densities. It was also investigated to train CNNs using the whole training sets from the beginning of the training but the proposed way was more efficient; CNNs converged to the solutions with fewer iterations.

The delayed RF signal induced by a scatterer lies across all the channels and at several depths depending on the lateral location of the scatterer. Hence, large receptive fields were required for a CNN, so four *down* and four *up blocks* were used. We tried to incorporate skip connections into the proposed CNN by, if necessary, applying upsampling to the feature maps in the contracting path to match the size of their corresponding feature maps in the expanding path. For image segmentation, the skip connections play an important role to recover lost spatial information during downsampling [13], [56]. The resulting reconstructed images have more fine details and, as a result, provide better localized semantic segmentation. However, the skip connections hindered successful training for the task in this paper and the CNNs learned zero confidence maps. We presume that the feature maps extracted from RF channel data in the contracting path are not directly related to the reconstruction of confidence maps, unlike image segmentation. Instead, CoordConv [44] was applied to cope with the spatial information loss. The



(a)

Fig. 16. The average numbers of scatterers closer than the theoretical resolution limit of DAS beamforming given a scatterer at different scatterer densities in the simulated test sets.



(a) (b)

Fig. 17. Recall and localization precision re-calculated to compare CNN method to Deep-ULM: (a) Positive detection density and (b) median of Euclidean position errors with one standard deviation bars at different scatterer densities.

CNNs with CoordConv localized non-overlapping Gaussians more precisely and achieved the better recall and localization precision on the validation sets.

On the simulated test sets, the proposed method outperformed the baseline method. The performance drop was much more severe for the baseline method at high scatterer densities, where the more scatterers were placed with in the resolution limit. Fig.16 shows the average numbers of scatterers within the FWHM (the $6\,\mathrm{dB}$ ellipse contour) given a scatterer in the simulated test sets.

Deep-ULM is another CNN-based method which localizes high-density targets from beamformed images that contain overlapping PSFs. To compare the proposed method with Deep-ULM, the recall and localization errors were re-calculated following the method which van Sloun *et al.* used to generate the results in the supplementary Fig. 1 in [24]. The threshold value for determining positive detection was $\lambda/7$ and Euclidean distances between the true and estimated scatterers were calculated. The evaluation results depend on the threshold value. As it increases, recall improves while localization precision degrades. The threshold $\lambda/7$ was chosen following [24] for a fair comparison. The results are presented in Fig. 17. Both methods showed good performance at high densities but the proposed method achieved slightly better recall and localization precision. Deep-ULM recovered roughly $1.80\,\mathrm{mm}^{-2}$, while the proposed method recovered $2.26\,\mathrm{mm}^{-2}$ at the density of $2.44\,\mathrm{mm}^{-2}$, and Deep-ULM recovered

roughly $2.10\,\mathrm{mm}^{-2}$ at the density of $3.53\,\mathrm{mm}^{-2}$ when the proposed method recovered $3.00\,\mathrm{mm}^{-2}$ at the density of $3.42\,\mathrm{mm}^{-2}$. The median of Euclidean errors of Deep-ULM was approximately $\lambda/12$ but the proposed method achieved smaller errors than that. It is difficult to conclude that the proposed method outperforms Deep-ULM since the evaluation was not performed on the same test data. This, however, shows the potential of the methods directly employing RF channel data.

To assess the proposed method for real world applications, two 3-D printed phantoms were imaged. One of the benefits of using the 3-D printed phantoms is that true scatterer positions and the dimensions of the phantom and scatterers (cavities) are known. It was important to modify the scatterers in the training sets to match the cavity dimensions. The CNNs trained for the simulation experiment failed on the measured data, showing too many false positive detections axially. However, the CNNs trained with the modified training sets successfully identified scatterers to some extent except some scatterers on the grid phantom when one plane wave was transmitted as seen in Fig. 14b. It is notable that this was achieved only with the simulated training data, since it is extremely difficult to obtain sufficient training data with ground truth for these kinds of experiments.

The phantom experiments show that the CNN method is transferable to measured data by modeling scatterers properly in the training data simulation. The baseline method performed slightly better for the most trivial case, namely the grid scatterer phantom with three plane waves, but the CNN method performed better on the random scatterer phantom. Even so, the CNN method on the random scatterer phantom presented a relatively large number of false positives compared to the simulation results. This could be because of the discrepancy between the training (simulated) data and the test (phantom) data. There are factors not considered in the simulation such as attenuation, different scattering intensities of the cavities, and different speed of sound in the phantom medium. Moreover, a further degradation of the performance is expected on *in vivo* data since the discrepancy between the training data and the *in vivo* data would become larger due to scatterer response variations, refraction, reverberation artifacts, etc. A more versatile simulation using various parameters to cover possible *in vivo* variations of RF channel data and a more generalized CNN model could increase the CNN method performance on the measured phantom data and overcome the potential limits in *in vivo* scenarios.

The proposed method gives 2-D images using a 1-D transducer. This limits the view of target structure along the elevation direction. The 3-D printed phantoms are essentially 2-D phantoms which have elongated cavities and the dimension along the elevation direction was not captured in the results. This limitation can be solved by using 2-D transducers such as fully addressed transducers or row-column addressed transducers.

Several problems are expected to occur if the CNN method is applied to MB detection for SRUS. MBs are not static but move with different velocities depending on the vessel size. This should be considered during training data generation. Also, it is important to model MBs properly in simulations since their sizes and other physical properties vary. It was necessary to remodel scatterers following the real physical structure for the phantom experiment. This is expected to be an important factor when applying the CNN method on measured MB signals.

Background scattering from tissue was not dealt with here since this work focused on a proof-of-concept of CNNs' ability to detect and localize high concentrations of scatteres from RF channel data. For *in-vivo* scenarios, the tissue signals may hinder the CNN method, so a way of rejecting them without hurting the performance of CNNs needs to be investigated. For example, clutter filtering based on singular value decomposition (SVD) or contrast-enhanced ultrasound (CEUS) imaging such as pulse inversion [57] or amplitude modulation [58] can be applied. However, the drawbacks of such methods are that it is difficult to find an optimal singular value for SVD to separate MB signals, and the CEUS imaging limits the frame-rate. In addition, both methods have a chance to distort the signals from the MBs, which would make the detected MB signals different from the data used for training. Alternatively, another neural network such as CORONA [59] can be deployed, which is a Robust PCA-based unfolded neural network that performs clutter filtering. By incorporating CORONA with the proposed CNN method, clutter filtering and MB localization can be learned simultaneously.

Lastly, further research on the optimal imaging scheme and scalability of CNN is required. Plane waves were used to support the hypothesis in a small region. In practice, however, a larger field of view is needed. Also, the more correlated data are available, the better estimation can be achieved. The CNNs with three plane waves achieved better performance than the CNN with one plane wave in all evaluation criteria, but this increases the required GPU memory. In addition, the imaging scheme would affect the capability of the CNN method and plane waves might not be the optimal choice. It is necessary to examine how other imaging schemes, such as focused or defocused waves affect the CNN method, or a new imaging scheme could be developed.

VI. Conclusion

The CNN-based scatterer detection and localization method is presented. CNNs were trained to return non-overlapping Gaussian confidence maps from simulated RF channel data, and the scatterer positions were estimated from the confidence maps. The simulation results show that the proposed method can identify high-density scatterers successfully even when some of them are closer than the resolution limit of conventional ultrasound imaging. It is also shown that the CNN method can be applied to real measured data by modeling scatterers following the true scatterer structure. The CNN method can potentially be extended to replace DAS beamforming for high concentration MB detection and thus reduce the long data acquisition times of SRUS using ULM.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V graphics card used for

this research.

REFERENCES

[1] F. L. Thurstone and O. T. von Ramm, "A new ultrasound imaging technique employing two-dimensional electronic beam steering," in *Acoustical Holography*, P. S. Green, Ed., vol. 5. New York: Plenum Press, 1974, pp. 249–259.

[2] O. Couture, B. Besson, G. Montaldo, M. Fink, and M. Tanter, "Microbubble ultrasound super-localization imaging (MUSLI)," in *Proc. IEEE Ultrason. Symp.*, 2011, pp. 1285–1287.

[3] O. M. Viessmann, R. J. Eckersley, K. Christensen-Jeffries, M. X. Tang, and C. Dunsby, "Acoustic super-resolution with ultrasound and microbubbles," *Phys. Med. Biol.*, vol. 58, pp. 6447–6458, 2013.

[4] M. A. O'Reilly and K. Hynynen, "A super-resolution ultrasound method for brain vascular mapping," *Med. Phys.*, vol. 40, no. 11, pp. 110 701–7, 2013.

[5] C. Errico, J. Pierre, S. Pezet, Y. Desailly, Z. Lenkei, O. Couture *et al.*, "Ultrafast ultrasound localization microscopy for deep super-resolution vascular imaging," *Nature*, vol. 527, pp. 499–502, November 2015.

[6] K. Christensen-Jeffries, R. J. Browning, M. Tang, C. Dunsby, and R. J. Eckersley, "In vivo acoustic super-resolution and super-resolved velocity mapping using microbubbles," *IEEE Trans. Med. Imag.*, vol. 34, no. 2, pp. 433–440, February 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012, pp. 1097–1105.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[9] ——, "Identity mappings in deep residual networks," in *Eur. Conf. Computer Vision*, 2016, pp. 630–645.

[10] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Computer Vision*, 2017, pp. 2980–2988.

[12] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," `arXiv:1804.02767v1 [cs.CV]`, 2018.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[15] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Eur. Conf. Computer Vision*, 2018, pp. 801–818.

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 105–114.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 1132–1140.

[18] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino *et al.*, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, no. 5793, pp. 1642–1645, 2006.

[19] S. T. Hess, T. P. K. Girirajan, and M. D. Mason, "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy," *Biophysical Journal*, vol. 91, no. 11, pp. 4258–4272, 2006.

[20] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature methods*, vol. 3, no. 10, pp. 793–795, 2006.

[21] E. Nehme, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Deep-STORM: super-resolution single-molecule microscopy by deep learning," *Optica*, vol. 5, no. 4, pp. 458–464, Apr 2018.

[22] W. Ouyang, A. Aristov, M. Lelek, X. Hao, and C. Zimmer, "Deep learning massively accelerates super-resolution localization microscopy," *Nature biotechnology*, 2018.

[23] N. Boyd, E. Jonas, H. Babcock, and B. Recht, "Deeploco: Fast 3d localization microscopy using neural networks," `bioRxiv 267096`, 2018.

[24] R. J. G. van Sloun, O. Solomon, M. Bruce, Z. Z. Khaing, H. Wijkstra, Y. C. Eldar *et al.*, "Super-resolution ultrasound localization microscopy through deep learning," `arXiv:1804.07661v2 [eess.SP]`, 2018.

[25] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1464–1477, 2018.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learning Representations*, 2015.

[28] A. C. Luchies and B. C. Byram, "Deep neural networks for ultrasound beamforming," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2010–2021, 2018.

[29] D. Hyun, L. L. Brickson, K. T. Looby, and J. J. Dahl, "Beamforming and speckle reduction using neural networks," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 66, no. 3, pp. 898–910, 2019.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014, pp. 2672–2680.

[31] S. Goudarzi, A. Asif, and H. Rivaz, "Multi-focus ultrasound imaging using generative adversarial networks," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2019, pp. 1118–1121.

[32] X. Zhang, J. Li, Q. He, H. Zhang, and J. Luo, "High-quality reconstruction of plane-wave imaging using generative adversarial network," in *Proc. IEEE Ultrason. Symp.*, 2018, pp. 1–4.

[33] J. Youn, M. L. Ommen, M. B. Stuart, E. V. Thomsen, N. B. Larsen, and J. A. Jensen, "Ultrasound multiple point target detection and localization using deep learning," in *Proc. IEEE Ultrason. Symp.*, 2019.

[34] J. A. Jensen and N. B. Svendsen, "Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 39, no. 2, pp. 262–267, 1992.

[35] J. A. Jensen, "Field: A program for simulating ultrasound systems," *Med. Biol. Eng. Comp.*, vol. 10th Nordic-Baltic Conference on Biomedical Imaging, Vol. 4, Supplement 1, Part 1, pp. 351–353, 1996.

[36] ——, "A multi-threaded version of Field II," in *Proc. IEEE Ultrason. Symp.* IEEE, 2014, pp. 2229–2232.

[37] B. G. Tomov, S. E. Diederichsen, E. V. Thomsen, and J. A. Jensen, "Characterization of medical ultrasound transducers," in *Proc. IEEE Ultrason. Symp.*, 2018, pp. 1–4.

[38] J. A. Jensen, "Safety assessment of advanced imaging sequences, II: Simulations," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 63, no. 1, pp. 120–127, 2016.

[39] A. Gomariz, W. Li, E. Ozkan, C. Tanner, and O. Goksel, "Siamese networks with location prior for landmark tracking in liver ultrasound sequences," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2019, pp. 1757–1760.

[40] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Machine Learning*, 2015, pp. 448–456.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[43] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[44] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev *et al.*, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Neural Information Processing Systems*, 2018, pp. 9605–9616.

[45] M. L. Ommen, M. Schou, R. Zhang, C. A. V. Hoyos, J. A. Jensen, N. B. Larsen *et al.*, "3D printed flow phantoms with fiducial markers for super-resolution ultrasound imaging," in *Proc. IEEE Ultrason. Symp.*, 2018, pp. 1–4.

[46] M. L. Ommen, M. Schou, C. Beers, J. A. Jensen, N. B. Larsen, and E. V. Thomsen, "3d printed calibration micro-phantoms for validation of super-resolution ultrasound imaging," in *Proc. IEEE Ultrason. Symp.*, 2019, pp. 1–4.

[47] G. Montaldo, M. Tanter, J. Bercoff, N. Benech, and M. Fink, "Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 56, no. 3, pp. 489–506, March 2009.

[48] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," arXiv:1312.6120v3 [cs.NE], 2013.
[49] D. Kingma and L. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 [cs.LG], 2015.
[50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2011, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/
[51] J. A. Jensen, H. Holten-Lund, R. T. Nilsson, M. Hansen, U. D. Larsen, R. P. Domsten *et al.*, "SARUS: A synthetic aperture real-time ultrasound system," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 60, no. 9, pp. 1838–1852, 2013.
[52] C. Chang, S. Chatterjee, and P. R. Kube, "On an analysis of static occlusion in stereo vision," in *IEEE Conf. Computer Vision and Pattern Recognition*, 1991, pp. 722–723.
[53] P. Fua, "A parallel stereo algorithm that produces dense depth maps and preserves image features," *Mach. Vis. Appl.*, vol. 6, no. 1, pp. 35–49, 1993.
[54] P. Jaccard, "The distribution of the flora in the alpine zone," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[55] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE Int. Conf. Computer Vision*, 2017, pp. 2999–3007.
[56] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," arXiv:1608.04117v2 [cs.CV], 2016.
[57] D. H. Simpson, C. T. Chin, and P. N. Burns, "Pulse inversion Doppler: a new method for detecting nonlinear echoes from microbubble contrast agents," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 46, no. 2, pp. 372–382, 1999.
[58] V. Mor-Avi, E. G. Caiani, K. A. Collins, C. E. Korcarz, J. E. Bednarz, and R. M. Lang, "Combined assessment of myocardial perfusion and regional left ventricular function by analysis of contrast-enhanced power modulation images," *Circulation*, vol. 104, no. 3, pp. 352–357, 2001.
[59] O. Solomon, R. Cohen, Y. Zhang, Y. Yang, Q. He, J. Luo *et al.*, "Deep unfolded robust PCA with application to clutter suppression in ultrasound," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1051–1063, 2020.