A Deep Learning-Based Model That Reduces Speed of Sound Aberrations for Improved *In Vivo* Photoacoustic Imaging

Seungwan Jeon, Wonseok Choi, Byullee Park, and Chulhong Kim, *Senior Member, IEEE*

***Abstract*— Photoacoustic imaging (PAI) has attracted great attention as a medical imaging method. Typically, photoacoustic (PA) images are reconstructed via beamforming, but many factors still hinder the beamforming techniques in reconstructing optimal images in terms of image resolution, imaging depth, or processing speed. Here, we demonstrate a novel deep learning PAI that uses multiple speed of sound (SoS) inputs. With this novel method, we achieved SoS aberration mitigation, streak artifact removal, and temporal resolution improvement all at once in structural and functional *in vivo* PA images of healthy human limbs and melanoma patients. The presented method produces high-contrast PA images *in vivo* with reduced distortion, even in adverse conditions where the medium is heterogeneous and/or the data sampling is sparse. Thus, we believe that this new method can achieve high image quality with fast data acquisition and can contribute to the advance of clinical PAI.**

***Index Terms*— Neural networks, image denoising, image enhancement, photoacoustic imaging.**

Manuscript received October 5, 2020; revised June 20, 2021 and August 20, 2021; accepted September 30, 2021. Date of publication October 19, 2021; date of current version October 28, 2021. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government Ministry of Science and ICT (MSIT) under Grant NRF-2019R1A2C2006269; in part by the Basic Science Research Program through the NRF funded by the Ministry of Education under Grant 2020R1A6A1A03047902; in part by the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health and Welfare, Republic of Korea, under Grant HI15C1817; in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Pohang University of Science and Technology) under Grant 2019-0-01906; in part by the Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea Government, Ministry of Trade, Industry and Energy (MOTIE); in part by the National Research and Development Program through the NRF funded by the MSIT under Grant 2020M3H2A1078045; and in part by the Brain Korea 21 FOUR (BK21 FOUR) Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiangqian Wu. *(Corresponding author: Chulhong Kim.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the Pohang University of Science and Technology.

Seungwan Jeon was with the Medical Device Innovation Center, Departments of Electrical Engineering, Convergence IT Engineering, Mechanical Engineering, and Interdisciplinary Bioscience and Bioengineering, Graduate School of Artificial Intelligence, Pohang University of Science and Technology, Pohang 37673, Republic of Korea. He is now with Samsung Electronics, Suwon 16677, Republic of Korea.

Wonseok Choi, Byullee Park, and Chulhong Kim are with the Medical Device Innovation Center, Departments of Electrical Engineering, Convergence IT Engineering, Mechanical Engineering, and Interdisciplinary Bioscience and Bioengineering, Graduate School of Artificial Intelligence, Pohang University of Science and Technology, Pohang 37673, Republic of Korea (e-mail: chulhong@postech.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2021.3120053, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3120053

I. Introduction

PHOTOACOUSTIC imaging (PAI) has become a trending medical imaging technique. This imaging modality detects ultrasound (US) signals generated through transient thermal expansion after optical absorbers are illuminated by pulsed light. Because two types of hemoglobin in living subjects mainly absorb visible and near-infrared light, high-resolution structural (e.g., total hemoglobin) and functional (e.g., hemoglobin oxygen saturation and blood flow) vascular images can be photoacoustically formed [1]–[3]. These high-resolution photoacoustic (PA) blood-vessel images provide significant information about such diseases as cancers, ischemic diseases, ocular neovascularization, and peripheral artery diseases [4], [5]. Potentially, PAI's diagnostic coverage and sensitivity can be further expanded and enhanced with various external contrast agents, further increasing its clinical importance [6]–[8].

In PAI, acoustic beamforming techniques are widely used to locate the initial PA pressure sources in a medium. In principle, the beamforming technique reconstructs PA images by synthesizing highly correlated PA signals. To achieve high-contrast, high-resolution, and artifact-reduced PA images, many beamforming algorithms have been explored, such as delay-and-sum (DAS), delay-multiply-and-sum (DMAS) [9], back-projection [10], [11], Fourier beamforming [12], time-reversal beamforming [13], model-based beamforming [14], and filter-aided beamforming [15], [16]. However, there are still limits to optimal PA image reconstruction. One serious issue concerns the assumed speed of sound (SoS) in biological tissues (Fig. 1a). To maximize the constructive interference at the main lobe point in the beamforming process, it is necessary to accurately estimate the time delay of each acoustic signal so that the signals are synthesized in phase. In conventional US imaging, Jaegar *et al.* introduced SoS imaging and aberration correction methods with multiple US transmission angles under some assumptions (e.g., no refraction/diffraction or hypoechoic region) [17], [18]. However, its estimated SoS distribution map was still blurred and its application for PA imaging was not been demonstrated. Alternatively, a model-based iterative reconstruction method was adopted to solve the SoS problem of PA computed tomography (PACT) images [19]–[21]. Recently, Poudel used a joint reconstruction method to simultaneously estimate initial pressure and SoS maps in PACT images [22], [23]. However, these tomographic approaches and the slow reconstruction times were clinically inappropriate. Because of these prob-

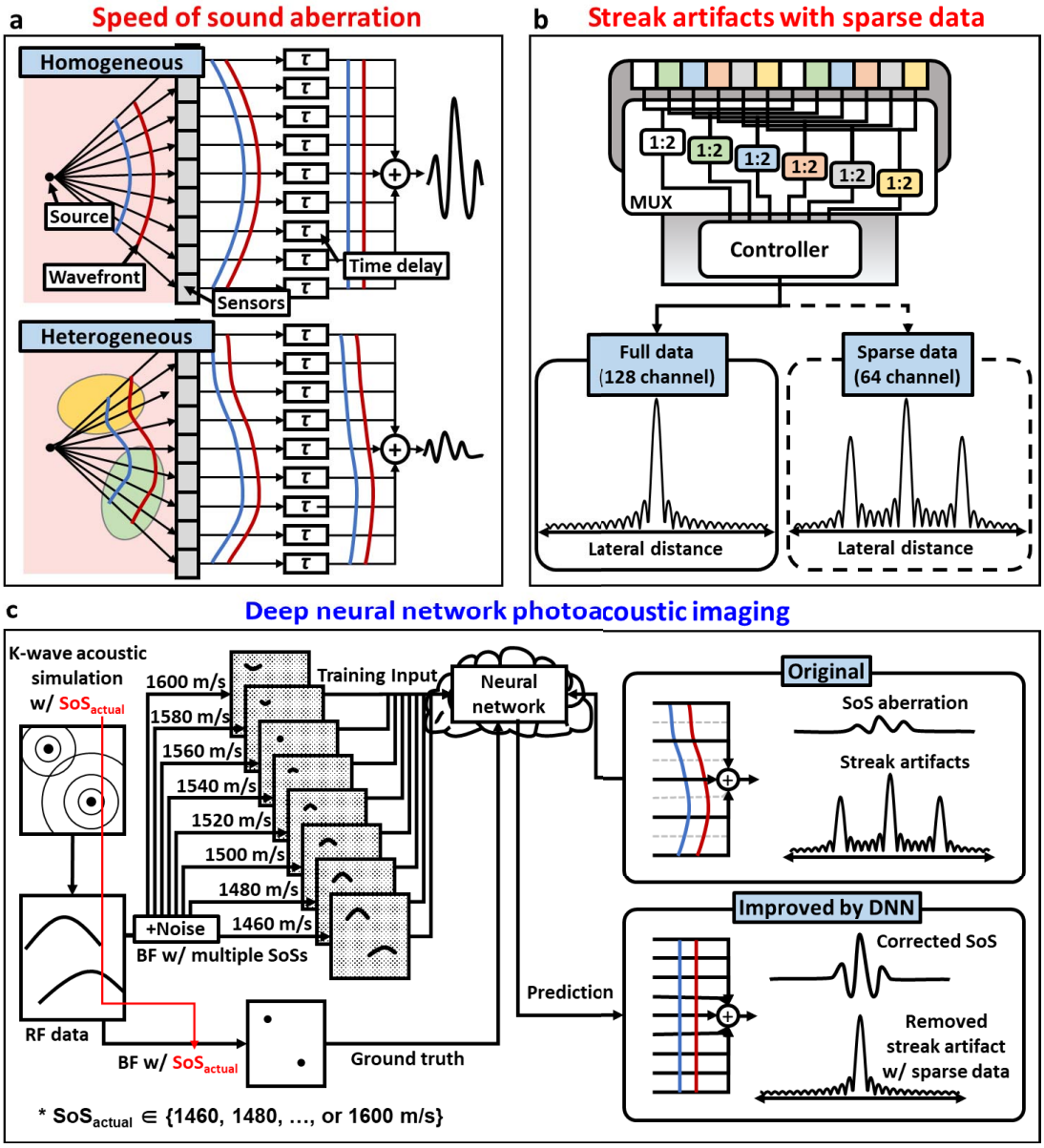


Fig. 1. (a) Speed of sound (SoS) aberration in photoacoustic imaging (PAI). (b) Streak artifacts with sparse data. (c) Proposed deep neural network (DNN) PAI. As inputs, eight beamformed photoacoustic (PA) images are used with eight different SoSs, increasing from 1460 to 1600 m/s in 20 m/s increments. As a ground truth, one PA image is reconstructed with the actual SoS. Based on the training dataset, the DNN is trained to correct the SoS aberration and streak artifacts in the PA images. MUX, multiplexer; RF, radiofrequency; and BF, beamforming.

lems, it is still common to apply a constant SoS for PA image reconstruction, which causes a SoS aberration artifact. If the SoS value could be locally and dynamically applied, SoS aberration in heterogeneous media could be mitigated, improving the overall image quality. In terms of data density, it is ideal to receive all signals in the entire detection area simultaneously and densely (Fig. 1b), but this requires high data throughput, which makes the imaging system complex and expensive. For that reason, as an alternative, many entry-level US imaging systems and some imaging devices that detect a large area use a multiplexer (MUX), at the expense of the temporal resolution. Under-sampling the signal overcomes the slow data acquisition rate of the imaging systems with the MUX, but can result in strong streak artifacts due to grating lobes. Successfully suppressing these artifacts from the sparse data would yield not only improved image quality, but also accelerated image acquisition and reduced system complexity.

Recently, deep learning techniques are being studied intensively to improve image quality beyond hardware limitations [24], [25]. Similarly, many studies are being conducted to explore the usefulness of deep learning techniques in medical images [26]–[28], including PA images. PA image reconstruction is hindered by such deleterious factors as the limited-view effect, streak artifacts [29], reflection artifacts [30], and erroneous SoS selection [31], and thus many researchers have

tried to overcome these problems by using deep neural networks (DNNs). In general, they trained the DNNs by pairing the observed data with the intact data and compensated for distortions inside the raw radiofrequency (RF) data [32], [33] or the reconstructed image [30], [34]. However, they have demonstrated image enhancement only on *in silico* or simple experimental phantoms. To test the DNNs on more realistic samples, several studies used simulated PA data based on *in vivo* human X-ray computed tomography (CT) images [8], [34]. However, no *in vivo* tests were present. Recently, using the PACT modality, Davoudi *et al*. developed a DNN that reduces streak artifacts from under-sampled data of *in vivo* PA images from animals [35], and Shan *et al.* developed a DNN that reconstructs both the acoustic initial pressure and speed [36]. Unfortunately, the use of a clinically uncommon transducer poses an obstacle to the wide-spread clinical use of this deep learning-based PA tomography approach. Hauptmann *et al.* also demonstrated a DNN to reduce noises in *in vivo* under-sampled PA images of a human palm by training the DNN with multiple CT images [37]. However, both studies only corrected the streak artifacts caused by under-sampled data on structural images, but did not correct for SoS aberration caused by acoustic heterogeneity. Despite the strong potential of deep learning techniques, little research has been reported on both structural and functional human imaging. A deep learning technique for clinical multi-parametric PAI, with fewer errors and affordable system complexity, remains to be developed.

Here, we demonstrate a deep learning technique that mitigates the artifacts caused by both SoS aberration and sparse radiofrequency (RF) data in structural and functional *in vivo* PA images of human soft tissue (Fig. 1c). We employed a clinically viable PA/US imaging system [38], [39] with a widely available linear US array transducer. As the training datasets, we prepared PA images from virtual phantoms in homogeneous media. Then, to train the DNNs, we first used the input PA images beamformed with eight different SoS values, and then with the ground truth PA image beamformed with the actual SoS. Finally, we corrected the *in silico* phantom and *in vivo* human limb PA images. In this way, we can realize the following benefits: 1) SoS aberration, streak artifacts and noise can be mitigated simultaneously in PA images; 2) suppressing streak artifacts due to sparse data can improve temporal resolution and reduce system complexity; and 3) high practicality can be expected using a commercial linear US array transducer. In the *in silico* phantom study, we quantitatively assessed the correction performance by measuring the structural similarity indices (SSIMs) and signal-to-noise ratios (SNRs). We found that the corrected PA images had, on average, up to about 0.24 higher SSIM than the pre-corrected image. Further, the trained DNNs successfully suppressed both SoS aberration and streak artifacts in both *in vivo* structural and functional PA images of healthy human limb and melanoma patients. It is difficult to objectively evaluate the correction performance in the *in vivo* images because they do not have ground-truth PA images. To deal with this problem, we tried a qualitative method to evaluate the correction performance as objectively as possible, taking advantage of the fact the side lobes become smallest at the optimal SoS. We also introduced a metric to quantify the difference between two functional PA images reconstructed from 128-/64-ch RF data to confirm that the proposed deep learning processing had little effect on the sO2 values, and to show that this DNN method is compatible with PA functional imaging. Our results indicate that the proposed deep learning-based method improves the overall PA image quality by mitigating the problems caused by heterogeneous SoS values and data sparsity. In addition, this method has been implemented with the linear US array transducer most commonly used in medical US imaging, and consequently it is expected to have great potential in many clinical applications.

II. Methods

A. Training Dataset Preparation

To prepare training datasets, we randomly generated initial pressure maps (Supplementary Fig. 1) and obtained their RF data, $r \in \mathbb{R}^{1801\times128}$, using the K-wave simulation toolbox [40]. The simulation was conducted in a homogeneous medium with an acoustic absorption coefficient of 0.5 dB/(MHz·cm)$^{-1}$; a maximum depth of 5.12 cm; a SoS, $c_{act} \in \{1460\text{m/s}, 1480\text{m/s}, \cdots, 1600\text{m/s}\}$; and a sampling rate of 40 MHz. Note that the lateral RF data size was determined by the element number (i.e., 128) while the temporal size was determined by the depth, SoS, and sampling rate (i.e., 1801 pixels $>$ 5.12 cm / 1500 m/s $\times$ 40 MHz). Gaussian random noise was added to the input RF data, $r + \mathcal{N}$, using the K-wave toolbox's addNoise function to randomly set the SNR from 1 to 5 dB. B-mode PA images were reconstructed via the Fourier beamforming method. Note that we laterally interpolated the beamformed image by three times by adding zero-padding before the inverse Fourier transform, $f_c(r+\mathcal{N}) \in \mathbb{R}^{1801\times384}$. Then we repeated this process to generate a group input data with eight different SoS values, $g(r+\mathcal{N}) \in \{f_{1460m/s}(r+\mathcal{N}), f_{1480m/s}(r+\mathcal{N}), \cdots, f_{1600m/s}(r+\mathcal{N})\} \in \mathbb{R}^{1801\times384\times8}$. In the same way, we prepared 64-ch inputs, $g(\ddot{r}+\mathcal{N})$, by replacing half of the 128-ch RF data with zeros considering the MUX configuration of our imaging system. Meanwhile, the ground truth images were beamformed using the noise-free RF data and the actual SoS as $f_{c_{act}}(r)$. Thus, during the DNN training, the input dataset was randomly selected between the $g(r+\mathcal{N})$ and $g(\ddot{r}+\mathcal{N})$ whereas $f_{c_{act}}(r)$ is always used as the ground truth.

B. Deep Neural Networks and Training

U-net is one of the most famous biomedical image segmentation DNNs [41] and is still widely adopted to solve many medical image segmentation and reconstruction problems. Like a classical autoencoder [42], this architecture consists of an encoder and decoder, but they are linked through concatenation layers at each depth. Segnet [43] has an encoder similar to that of U-net, but its decoder consists of unpooling and transposed convolution layers. As adopted by DeconvNet [44], this decoder is known to represent detailed patterns. In this study, we utilized both Segnet and U-net, and compared the output images. Complementarily to two existing DNNs, we created

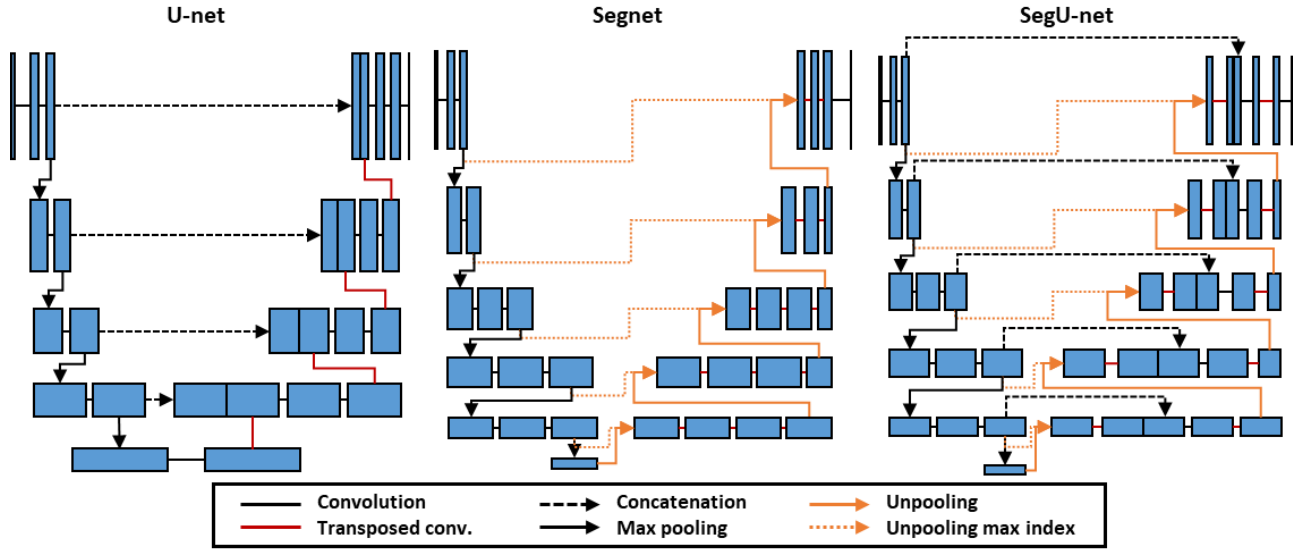


Fig. 2. Structures of U-net, Segnet, and SegU-net. The specification of each model are provided in Supplementary Fig. 2.

a hybrid DNN model, named SegU-net (Fig. 2). SegU-net has almost the same structure as Segnet but additionally connects the encoder and decoder through concatenation layers like U-net (Supplementary Fig. 2). The calculated numbers of parameters of U-net, Segnet, and SegU-net were about 32M, 29M, and 35M, respectively. In all the DNNs, batch normalization (BN) and a rectified linear unit (ReLu) were placed after every convolution and transposed convolution layer, except for the last 1×1 convolution layer. At the final regression layer, we used a loss function as follows:

$$L = \begin{cases} \frac{1}{2} \left\| f_{c_{act}}(r) - P\left(g\left(r+N\right)\right)\right\|_2^2 \ or \\ \frac{1}{2} \left\| f_{c_{act}}(r) - P\left(g\left(\ddot{r}+N\right)\right)\right\|_2^2 \end{cases} \quad (1)$$

where $P{:}\mathbb{R}^{N_x \times N_y \times N_z} \rightarrow \mathbb{R}^{N_x \times N_y}$ is the prediction function of the DNN, and (N_x, N_y, N_z) are the corresponding image sizes. In this case, we could confirm the effectiveness of the group input with multiple SoSs. As a comparison, we trained additional DNNs, $\bar{P} : \mathbb{R}^{N_x \times N_y} \rightarrow \mathbb{R}^{N_x \times N_y}$ that corrected the beamformed images only using a single SoS of 1540m/s as follows:

$$\bar{L} = \begin{cases} \frac{1}{2} \left\| f_{c_{act}}(r) - \bar{P}\left(f_{1540\mathrm{m/s}}\left(r+N\right)\right)\right\|_2^2 \ or \\ \frac{1}{2} \left\| f_{c_{act}}(r) - \bar{P}\left(f_{1540\mathrm{m/s}}\left(\ddot{r}+N\right)\right)\right\|_2^2 \end{cases} \quad (2)$$

To augment the training dataset, the DNNs were trained with randomly cropped patches. We extracted four patches from each dataset pair, and randomly flipped the cropped patches in the horizontal direction or scaled their amplitudes for data augmentation. The batch size was 32. The trainable weights were initialized with the He initialization method [45] and were updated with the ADAM optimizer [46]. The initial learning rate was set to 0.01. As the training strategy, if the validation loss did not decrease during one epoch, the training stopped and restarted with the decayed learning rate of 0.1. This process was repeated three times. In this study, the trainings were conducted for 9, 8, and 7 epochs in Segnet, U-net, and SegU-net, respectively. All processes were performed in the MATLAB 2019b environment using a desktop computer with an Intel Core i7-4790 processor, 24 GB RAM, and an Nvidia GeForce RTX 2070.

C. In Vivo PA Imaging Experiments

For this experiment, we used a programmable clinical US imaging system (ECUBE 12R, Alpinion Medical Systems, Republic of Korea) with a tunable pulsed laser (Phocus Mobile, OPOTEK Inc., USA) [38], [39]. The imaging system was equipped with a linear array transducer with 128 elements and a center frequency of 8.5 MHz. We attached the laser output port next to the transducer and adjusted the pulse energy to 10 mJ/cm^2. For the structural PA imaging, we scanned a healthy human's forearm or foot immersed in a water tank using a motorized scanner. When imaging the forearm, we set the scanner motor speed to 2 mm/s and used an 850 nm wavelength. For the foot imaging, the motor speed was set to 0.625 mm/s, and four wavelengths (700, 756, 796, and 866 nm) were used to calculate the hemoglobin sO_2. We also recruited a patient with a lentiginous melanoma on his heel to test melanoma PA unmixing. The melanoma region was imaged through a 3D handheld PA scanner [47], which is an imaging probe integrated with a linear motorized scanner and a water chamber, with a scanning speed of 0.5 mm/s and the same four wavelengths as for the foot imaging. All imaging procedures followed a protocol approved by the Institutional Review Boards of the POSTECH and Seoul St. Mary's Hospital (KC17DESI0201).

D. Hemoglobin Oxygen Saturation (sO_2) Calculation

Depending on the optical wavelength, λ, oxy- and deoxy-hemoglobin have different absorption coefficient curves. Thus,

the PA pressure amplitudes, p, generated at each wavelength are typically assumed to be proportional to the inner product between the concentrations (C_{HbO_2} and C_{Hb}) and the absorption coefficients ($\mu_{HbO_2}(\lambda)$ and $\mu_{Hb}(\lambda)$) of the two types of hemoglobin as follows:

$$(p_1 \cdots p_n) \propto \left(C_{HbO_2} C_{Hb}\right) \cdot M,$$

where

$$M = \begin{pmatrix} \mu_{HbO_2}(\lambda_1) \cdots \mu_{HbO_2}(\lambda_n) \\ \mu_{Hb}(\lambda_1) \cdots \mu_{Hb}(\lambda_n) \end{pmatrix}. \tag{3}$$

However, according to the optical wavelength, the optical fluence is differently attenuated along the depth, varying the PA pressure amplitudes. Hence, we compensated for the spectrally varying optical fluence before spectral unmixing as follow s [48]. First, we located the skin from the corresponding US B-mode images. Second, we made an optical compensation map by averaging the pixels under the skin layer by layer, excluding bright objects (i.e., blood vessels). Third, each PA B-mode image was normalized with the compensation map. Then we unmixed each hemoglobin concentration and calculated their ratio, called sO_2, by using the Moore-Penrose pseudo-inverse matrix [49] of M as follows:

$$\left(C_{HbO_2} C_{Hb}\right) = (p_1 \cdots p_n) M^T \left(M M^T\right)^{-1}, \tag{4}$$

when the MAP images were processed, the signals from the skin layers were removed.

III. Results

A. In Silico Phantom Study in Homogeneous Media

Fig. 3 illustrates the DNN test in both homogeneous and heterogeneous media. The dataset preparation processes and symbols are detailed in the Methods section and Supplementary Table I, respectively. Briefly, we prepared 340 virtual phantoms (270 for training, 30 for validation, and 40 for testing) with homogeneous media and generated PA RF datasets, r, via an acoustic propagation simulation tool (K-wave toolbox) [40]. We set one SoS (one of the eight SoSs varying from 1460 to 1600 m/s with an increment of 20 m/s) for one virtual phantom and produced one beamformed PA image with the same SoS as the ground truth, $f_{c_{act}}(r)$. Note that r is 128-ch RF data and c_{act} is the actual SoS of the corresponding r. As inputs, we added noise to the RF data, $r + \mathcal{N}$, and generated a group of eight PA images beamformed with eight different SoSs, $g(r + \mathcal{N})$. We also generated additional 340 datasets using the sparse RF data (i.e., 64-ch inputs, $g(\ddot{\ddot{r}} + N)$) where $\ddot{\ddot{r}}$ represents the undersampled RF data. With the 600 datasets (270 training datasets and 30 validation datasets reconstructed with the 128- and 64-ch RF data each), we trained three DNNs (e.g., Segnet, U-net, and SegU-net). Particularly, the DNN model, SegU-net, was newly created by combining Segnet and U-net (detailed in the Methods section). We compared the DNN-corrected PA images, $P(g(r + \mathcal{N}))$, with $f_{c_{act}}(r)$. Fig. 3a and Supplementary Fig. 3 show representative results with a c_{act} of 1560 m/s, and so $f_{c_{act}}(r + \mathcal{N})$ was placed in the fifth segment of g (r + N) (highlighted with the red boundary in Fig. 3a). Note that this sequence information was not provided in the DNN prediction process. The results show noticeable noise and streak artifacts in $f_{c_{act}}(r + \mathcal{N})$, whereas these artifacts are dramatically suppressed in $P(g(r + \mathcal{N}))$. We also tested the DNN-based correction with $\ddot{\ddot{r}}$ (bottom of Fig. 3a). The corrected PA image with 64-ch data by DNN, $P(g(\ddot{\ddot{r}} + \mathcal{N}))$, is qualitatively identical to the DNN-corrected PA image with 128-ch data, $P(g(r + \mathcal{N}))$, while the beamformed PA image with sparse data, $f_{c_{act}}(\ddot{\ddot{r}} + \mathcal{N})$, exhibits considerably strong streak artifacts. Interestingly, the side lobes and streak artifacts are even more suppressed in the DNN-corrected images compared with the ground truth. We quantitatively evaluated the *in silico* phantom study results by measuring the SNRs. In Supplementary Fig. 3a, we observe that the SNRs of both $P(g(r + \mathcal{N}))$ and $P(g(\ddot{\ddot{r}} + \mathcal{N}))$ are at least about 20 dB higher than those of $f_{c_{act}}(r + \mathcal{N})$ and $f_{c_{act}}(\ddot{\ddot{r}} + \mathcal{N})$, and they are also higher than that of $f_{c_{act}}(r)$. We also quantified the SSIMs of the PA images before and after applying the DNNs compared to the ground-truth PA image, $f_{c_{act}}(r)$ (Supplementary Figs. 1b and 1c). In this particular case with an actual SoS of 1560 m/s, the input PA image with an SoS of 1560 m/s, $f_{1560m/s}(r + \mathcal{N})$ has the highest SSIM (i.e., 0.822) compared with the ground truth, $f_{1560m/s}(r)$. However, all DNN-improved PA images had more than 0.95 of SSIMs regardless of the number of RF channels and DNN models. Fig. 3b shows the statistical SSIMs using the 40 test datasets. On average, the SSIM between $f_{c_{act}}(r + \mathcal{N})$ and $f_{c_{act}}(r)$ is 0.797 ± 0.086, and that with the sparse data is 0.723 ± 0.122. However, the SSIMs between $P(g(r + \mathcal{N}))$ and $f_{c_{act}}(r)$ for all three DNN models are higher than 0.94, and U-net and SegU-net provide better SSIMs than Segnet. The maximum SSIM improvement is about 21.3 %. These improved SSIMs are still valid with sparse data, but even more significantly, they show a maximum SSIM improvement of 32.9 %.

B. In Silico Phantom Study in a Heterogeneous Medium and Comparison With Existing Methods

We also tested our DNN-correction method in a heterogeneous *in silico* phantom consisting of three layers with SoSs of 1480 m/s, 1450 m/s, and 1575 m/s. The DNN-corrected PA images are compared with PA images processed with the conventional beamformer and two SoS aberration correction methods, multi-stencil fast marching (MSFM) and automatic SoS selection (Figs. 2c and 2d). The conventional beamformer is the Fourier beamformer with a single SoS (i.e., 1540 m/s) [50] but the medium is acoustically heterogeneous. The MSFM method can accurately synthesize the PA signals in a heterogeneous medium because this method estimates the acoustic time of flight (ToF) using the eikonal equation and utilizes the ToF as the exact time delay in the beamforming process (Supplementary Method 1) [51]. However, the MSFM is not practical because of its prior requirement of the actual SoS map to solve the Eikonal equation. In our case, we used the PA image processed with the MSFM as the reference because the actual SoS map was priorly incorporated for beamforming. The automatic SoS selection method processes the PA image

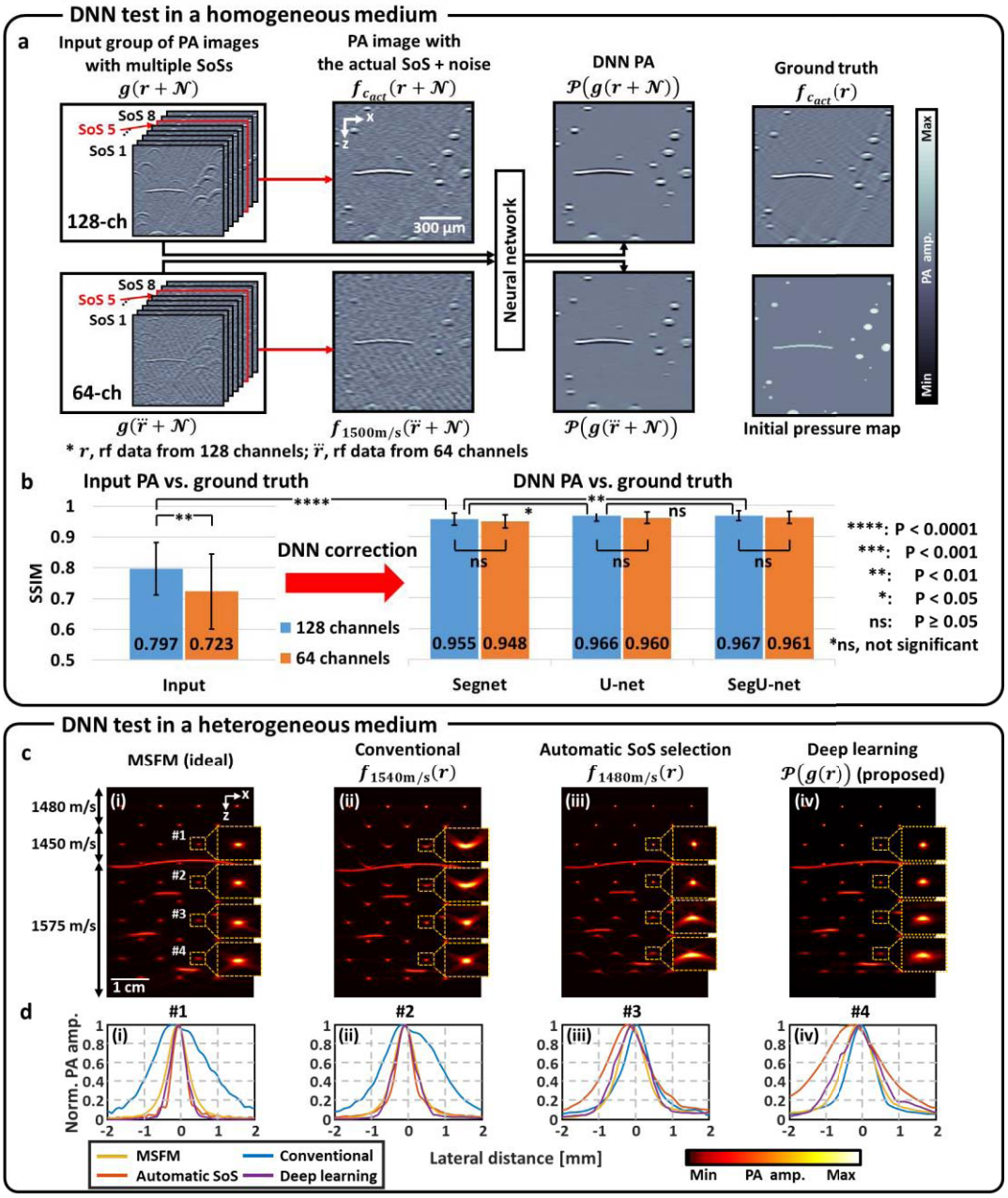


Fig. 3. (a) Deep neural network (DNN) training and testing in a *In silico* homogeneous medium. $g\,(r+\mathcal{N})$ is a group of photoacoustic (PA) images with multiple assigned speeds of sound (SoSs varying from 1460 to 1600 m/s, in increments of 20 m/s) and used as an input. $f_{c_{act}}\,(r+\mathcal{N})$ is a PA image with the actual SoS and noise values (i.e., 1500 m/s in this case). $P\,(g\,(r+\mathcal{N}))$ is a PA image corrected by DNN. $f_{c_{act}}\,(r)$ is the ground truth PA image. r stands for the data from 128 channels; $\ddot{r}$ represents 64 channels. Full-size images and the SNR measurement results are provided in Supplementary Fig. 3. (b) Statistical comparison of SSIMs of the input vs. ground truth PA images (left) and the corrected PA images by DNN vs. ground truth (right). n = 40. The asterisks represent the statistical significance by two-sample t-test. (c) DNN test in a heterogeneous medium consisting of three layers with three different SoSs of 1480 m/s, 1450 m/s, and 1575 m/s. (i) A beamformed PA image corrected for aberration with the multi-stencil fast marching (MSFM) method. (ii) A conventionally beamformed PA image with an SoS of 1540 m/s. (iii) A beamformed PA image corrected with the automatic SoS selection method. The SoS was estimated to be 1480 m/s from the Brenner focus function. (iv) A PA image corrected by SegU-net. (d), Comparison of normalized PA lateral profiles acquired from regions #1 – #4 in (c).

with an optimal SoS by maximizing the PA image sharpness (Supplementary Method 2) [52]. In this study, the image sharpness was calculated through the Brenner gradient focus function and the optimal SoS was estimated to be 1480 m/s. Note that all beamformed PA images were enveloped by the quadrature demodulation. We selected four regions at different depths (highlighted in Fig. 3c(i)) and compared their line profiles laterally and axially (Fig. 3d and Supplementary Fig. 4, respectively). These results show the significant suppression of side lobes with our DNN based on the multiple-SOS inputs. We extracted the line profiles by projecting the maximum PA amplitude in each direction within the regions to represent the 2D information of the targets. As shown in Figs. 2c(i, iv) and 2d, the MSFM- and DNN-PA images and

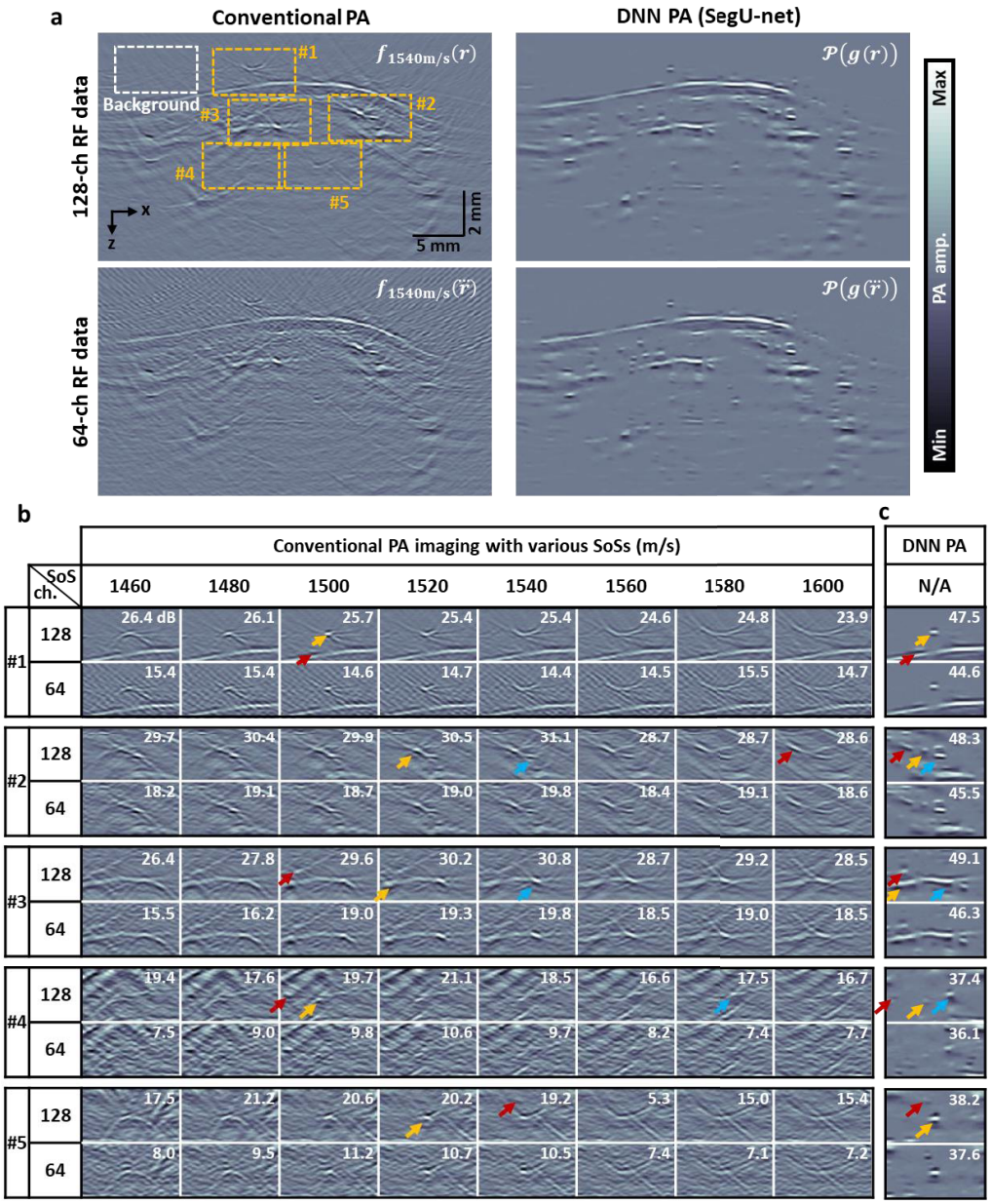


Fig. 4. (a) *In vivo* human forearm photoacoustic (PA) images reconstructed via conventional beamforming (left) and deep neural network (DNN; SegU-net, right) using 128-ch (top) and 64-ch (bottom) RF data. (b and c) Zoomed conventional (b) and DNN-corrected (c) PA images of the highlighted regions in (a). We conventionally beamformed the PA images with eight different SoSs in (b). The regions marked with colored arrows are qualitatively compared. The signal-to-noise ratios (upper right corner) were measured by dividing the maximum amplitude in each region-of-interest by the standard deviation of the area highlighted with the white dashed box in (a). SoS, speed of sound.

their associated line profiles are highly correlated in all four regions. However, significant SoS aberrations are visible in the #1 and #2 areas of the conventionally beamformed PA image (Fig. 3c(ii)), and they are also obvious in the #3 and #4 areas of the PA image processed with the automatic SoS selection method (Fig. 3c(iii)). These results are strongly related to the SoS distribution in the medium from each target to the sensors. The PA waves generated from the targets #1 and #2 mainly travel through the 1st and 2nd layers (e.g.,1480 and 1450 m/s, respectively), and thus the targets are well reconstructed with the automatic SoS selection method (e.g., estimated to be 1480 m/s). Yet, the conventional beamformer does not work properly for targets #1 and #2. In contrast, targets #3 and #4 are better reconstructed with the conventional beamformer because the targets are in the 3rd layer where the medium (1575 m/s) and preset (1540 m/s) SoSs are close. We also compared the processing time of each beamformer. It takes 0.10s, 3.31s, 9.21s, and 1.32s for the conventional Fourier, automatic SoS selection, MSFM, and proposed deep learning beamformers, respectively, to make a single B-mode image.

C. In Vivo DNN Photoacoustic Imaging of Human Forearms

Fig. 4a compares structural PA images of a human forearm obtained before and after DNN correction with both

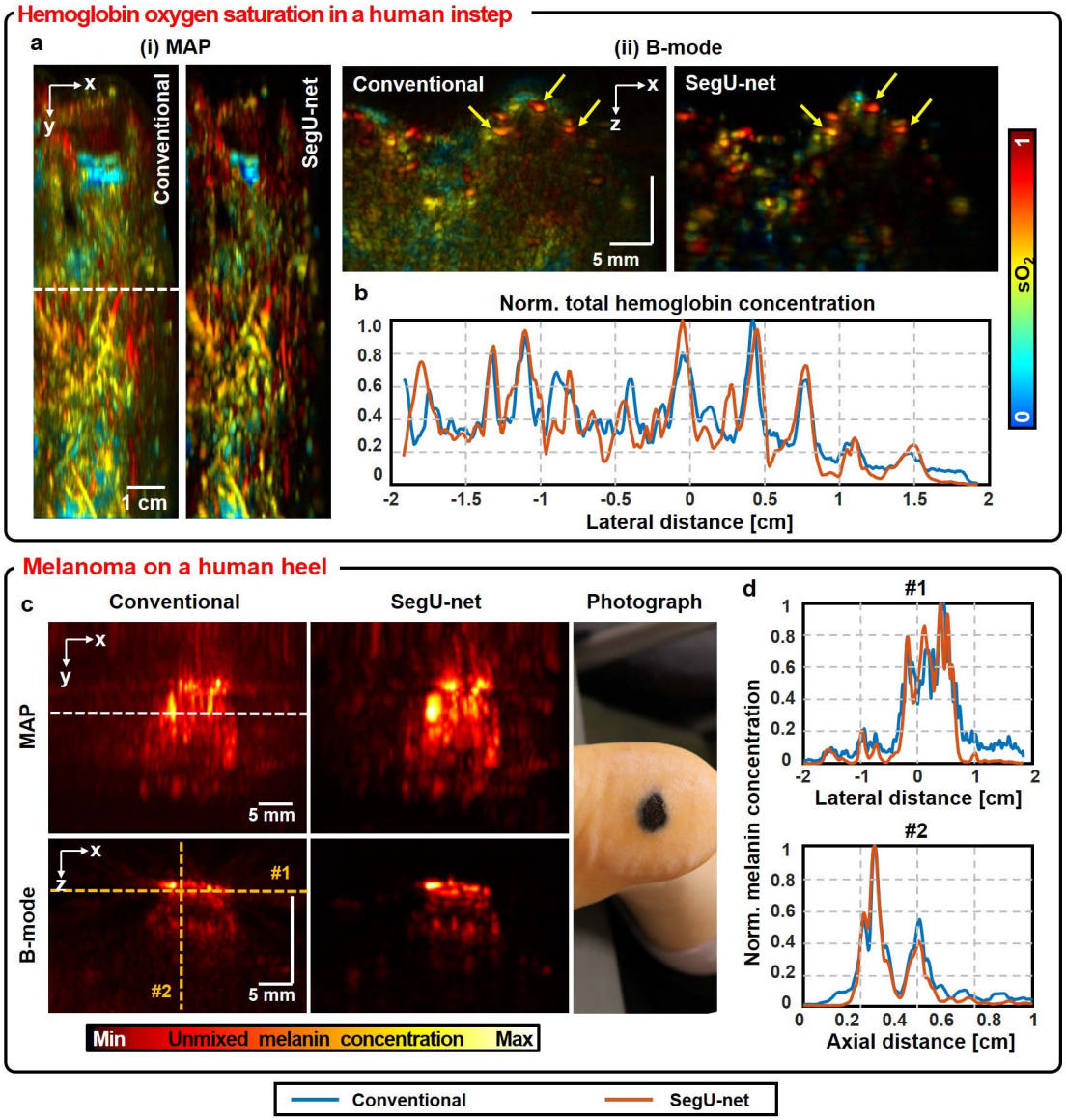


Fig. 5. (a) *In vivo* photoacoustic (PA) hemoglobin oxygen saturation (sO_2) images of human feet. (i) Maximum amplitude projection (MAP) and (ii) B-mode images reconstructed by conventional beamforming and deep neural network (DNN; SegU-net). The B-mode images were obtained along the white dashed line. (b) Lateral line profiles acquired by projecting the maximum hemoglobin concentration of the B-mode PA images in (a). (c) MAP (top) and B-mode (bottom) PA images of the spectrally unmixed melanoma with conventional beamforming (left) and SegU-net (middle). The B-mode PA images were extracted along the white dashed line. The photograph shows the melanoma on the patient's heel. (d) Lateral (left) and axial (right) line profiles acquired from highlighted regions #1 and #2, enclosed by the yellow dashed lines in (c).

128- and 64-ch RF data. Note that all beamformed PA images are before demodulation. In the conventional PA images, we observe noticeable SoS aberration and streak artifacts near main signal lobes, and the streak artifacts become even worse with the 64-ch RF data. In contrast, the SoS aberration and streak artifacts are remarkably reduced in the DNN-corrected PA images regardless of the data channel numbers. To analyze the network performance in detail, five regions were selected, indicated by the dashed boxes in Fig. 4a and shown in zoomed views in Figs. 3b and 3c. The enlarged DNN-PA images of the five regions (Fig. 4c) are compared with the zoomed conventional PA images with eight different SoSs (Fig. 4b). Owing to the absence of the ground-truth PA images *in vivo*, we assumed that one of the PA signals among the eight conventional PA images was close to the optimal PA signal, and was locally used as a reference for the PA signal in the corresponding DNN-PA image. We qualitatively selected the optimal main signal lobes in each region where the streak artifacts are minimal, indicted with colored arrows in Fig. 4b, and paired those with the main signal lobes (the same colored arrows) in Fig. 4c. Additionally, we measured the SNRs of each enlarged area by dividing their maximum amplitudes by the standard deviations in the background regions highlighted with the white dashed box in Fig. 4a. The measured SNRs of the 128- and 64-ch DNN-PA images were at least about 16 dB and 26 dB, respectively, higher than those of the conventional PA images. It is clear that the DNNs can exclusively extract the main signal lobes by purposely suppressing the surrounding SoS aberration and streak artifacts in both the 128- and 64-ch RF data. Moreover, all DNNs' noise cancellation and

main lobe corrections are quite notable in the lateral line profiles (Supplementary Fig. 5). All these improvements by the DNNs are also readily visible in the maximum amplitude projection (MAP) images (Supplementary Fig. 6). One thing to emphasize is that we used the multiple PA images with various SoSs as inputs for the DNNs. It is confirmed that DNNs with a single-SoS input do not properly alleviate the SoS aberration and streak artifacts (Supplementary Fig. 7). Among the three DNNs, the Segnet-corrected PA image contains somewhat more noise than the others. Still, U-net and SegU-net might not effectively suppress reverberation artifacts (red arrows in Supplementary Fig. 7a). Choosing the best DNN model remains controversial here due to the absence of the ground-truth image.

D. In Vivo Multi-Spectral Functional DNN Photoacoustic Imaging of Human Feet

Multi-spectral functional imaging using the unique optical absorbance of different intrinsic chromophores (e.g., oxy- and deoxy hemoglobins and melanin) is an important feature that differentiates PAI technology from other medical imaging technologies. This feature allows us to detect oxygen saturation abnormalities around diseased areas (e.g., cancers and brain disorders) and to capture specific molecular activities in PA images without exogenous contrast agents. To demonstrate that our proposed method can also be used for this function, we applied our method to multi-wavelength *in vivo* PA images obtained from the feet of healthy human and melanoma patients. For the healthy volunteers, the DNNs were used to correct the 3D PA hemoglobin oxygen saturation (sO_2) images (Figs. 4a and 4b). For the melanoma patients, we corrected the spectrally unmixed PA melanoma images (Figs. 4c and 4d). In both studies, we demodulated the DNN-corrected PA images for each wavelength and then applied a spectral unmixing technique (Figs. 4 and Supplementary Fig. 8). For sO_2 imaging, a human instep near the big toe was imaged at 700, 756, 796, and 866 nm. The conventional PA MAP sO_2 image exhibits blurry vessels and a noisy background, whereas the DNN (SegU-net) PA sO_2 shows much improved contrast. This improvement is also observed in the B-mode images (Fig. 5a). In the conventional PA B-mode sO_2 image, some vessels are distorted due to SoS aberration (yellow arrows in Fig. 5a), and strong speckle patterns are visible. In the SegU-net PA image, however, the SoS aberration and speckle patterns are significantly suppressed. Background suppression and signal enhancement are notable in the total hemoglobin concentration line profiles as well (Fig. 5b). We also compared the PA sO_2 images with the channel numbers and DNN models (Supplementary Fig. 8). The conventional PA images with the 64-ch RF input have a blurry background and strong speckle patterns compared to the images with the 128-ch RF input. In contrast, the DNN PA images do not show a notable difference between the 128- and 64-ch RF datasets. To quantitatively compare the difference between the 128- and 64-ch PA images, the overall and pixel-wise sO_2 image differences, D and d, were calculated (Supplementary Method 3 and Supplementary Fig. 9). All Ds of the DNN PA sO_2 images are shorter than that of the conventional PA image. Further, both U-net and SegU-net perform similarly, and are better than Segnet. In addition, we calculated the hemoglobin sO_2 difference of the main blood vessels before and after DNN correction to quantify the effect of DNN on sO_2 change (Supplementary Method 4 and Supplementary Table II). In the result, we observe little difference between the 128- and 64-ch DNN PA images. Among the DNN, SegU-net changed the hemoglobin sO_2 the least with the 128-ch RF data. With 64-ch RF data, U-net and SegU-net showed similar hemoglobin sO_2 changes (6.24 % vs. 6.29 %, respectively). For melanoma imaging, the heel region of a melanoma patient was imaged with the same four wavelengths [53] (Figs. 4c and 4d). Like the sO_2 images, the DNNs improve the PA image qualities in all the MAP, B-mode, and line profiles. This result is evidence that the DNN-PA imaging method could be a vital tool for delineating melanoma depth, a critical factor in staging melanoma patients.

IV. Discussion

SoS aberration and streak artifacts are major causes of image degradation in PA images. The autofocusing technique of searching for an optimal SoS value can mitigate the SoS aberration, but it is still not ideal [52]. The MSFM method can correct the SoS aberration in a heterogeneous medium, but it is impractical because the prior SoS map is inaccessible [51]. To suppress streak artifacts, a sufficient density of data is required in the beamforming process, but using an imaging system with a high number of channels increases the system complexity and cost. To solve these key problems, we propose a DNN-based PAI method that can correct SoS aberration and streak artifacts without the prior SoS map or excessive hardware cost. In this study, using *in silico* phantoms and *in vivo* healthy volunteers and melanoma patients, we demonstrate the effectiveness and utility of our DNN PA imaging method. Crucially, we train the DNNs with multiple-SoS inputs, which significantly improves the resulting image quality. Our experimental input datasets are beamformed with various SoSs. Consequently, the main signal lobes in the images are fixed, while the side and grating signal lobes are varied. As a basis for comparison, a ground truth image is beamformed with the actual SoS. Thus, this input and ground truth pairing can train the DNNs to retain the main signal lobes exclusively and properly reconstruct them with optimal shapes in the heterogeneous medium. These exclusive enhancements in the main lobes and the suppression in the side and grating lobes further reduce the streak artifacts from sparse data. Furthermore, random background noises are also minimized because they are variable. We could confirm that the DNN-corrected image shows weaker streak artifacts and a clearer background than the ground truth image in both *in vivo* and *in silico* studies. We also observe that the DNN outputs with multiple-SoS inputs provide better correction performance than those with a single-SoS input (Supplementary Fig. 7). The calculated difference in sO_2 values within the main blood vessels before and after the DNN correction is less than 9 % (Supplementary Table II). This result implies that the amplitude linearity of each main lobe can be relatively well maintained for spectral unmixing during the DNN process.

As mentioned above, our proposed DNNs work well with sparse data, which helps us to further improve the temporal resolution and reduce the system complexity. Our US transducer array has 128 elements, but the US imaging system has only 64 data acquisition channels. To acquire full 128-channel RF PA data, two laser shots are required, and because the laser repetition rate is 10 Hz, it takes 0.2 seconds to acquire one full-channel PA image. Using the DNNs method, however, only one laser shot is needed to reconstruct one full-channel PA image, potentially improving the temporal resolution by a factor of 2. Moreover, 64 data acquisition channels now sufficient to recover the full signal width, potentially reducing the system complexity and cost. These two potential improvements can be critical for wide clinical and commercial translation.

Identifying the optimal DNN model remains to be done. Initially, we explored two encoder-decoder segmentation DNNs: Segnet and U-net. We anticipated that Segnet could better reconstruct high-frequency information than U-net, because Segnet's decoder consists of unpooling and transposed convolution layers [44]. As expected, we could see the more detailed structures in the Segnet-PA images than in the U-net images. However, background noises were also relatively more distinct in the Segnet images. The high-frequency information was rather suppressed in the U-net images, relatively dominating the low-frequency information. Next, we implemented a new DNN model, called SegU-net, by adding U-net's skip layer to Segnet's basic structure. The SegU-net images were relatively similar to the U-net images rather than the Segnet images, which could indicate that the skip layer in U-net and SegU-net enhances the low-frequency information. When we compared the 128- and 64-ch DNN PA sO_2 images (Supplementary Fig. 8), the differences between the 128- and 64-ch images were minimal with U-net. However, with SegU-net, the hemoglobin sO_2 in the main vessels were maintained slightly better than with U-net (6.21 vs. 5.03 %, Supplementary Table II). We believe at this time that it is difficult to choose the optimal DNN model without ground-truth images for *in vivo* situations. Both the DNN model and the SoS range of the input data need to be optimized to further improve our method. Based on the SoS in living soft tissues, we set the SoS range of the input data at 20 m/s intervals (Supplementary Table III). Thus, an input consisting of the eight SoS values within the range could be sufficient. Iincreasing the size of the multi-SoS input data by setting the SoS range wider and finer might provide better output image quality, but it would negatively affect the processing speed and the amount of memory required.

In summary, we introduced a new training strategy using the multiple SoS inputs, an alternative method to evaluate the correction performance on non-reference *in vivo* PA images, and a metric to confirm the usability of the proposed method in functional imaging. Our proposed method has the following advantages: it can improve the PA image qualities by simultaneously reducing SoS aberration and streaking artifacts; its use of multiple SoS inputs effectively maximizes the image enhancement performance; and it allows low-channel systems to obtain high-quality, high-speed PA images without additional hardware costs. However, spatially-varying impulse responses and optical fluence changes were not considered in this study, but certainly these factors will be considered in the future.

V. Conclusion

Our developed DNN-PA imaging methods could correct SoS aberrations and suppress streak artifacts from sparse data in *in vivo* images from healthy volunteers and melanoma patients. After training, the DNNs could exclusively encode the main signal lobes from multiple beamformed images with various SoSs and could encode the main signal lobes with optimal shapes. Using an *in silico* phantom study, we demonstrated the superior practical performance of the DNNs over existing SoS correction methods. Next, using *in vivo* structural and functional PA imaging studies, we confirmed that the trained DNNs with multiple-SoS inputs could also successfully mitigate SoS aberrations and streak artifacts caused by medium heterogeneity and sparse data, respectively. We believe that this DNN-based method can potentially contribute to the fast clinical translation and practical commercial deployment of PAI.

Acknowledgment

The author Seungwan Jeon thanks Changyeop Lee for providing the motorized imaging scanner. The author Byullee Park thanks Chul Hwan Bang for providing human melanoma photoacoustic data.

References

[1] L. Li *et al.*, "Single-impulse panoramic photoacoustic computed tomography of small-animal whole-body dynamics at high spatiotemporal resolution," *Nature Biomed. Eng.*, vol. 1, no. 5, pp. 1–11, May 2017.

[2] J. Kim, J. Y. Kim, S. Jeon, J. W. Baik, S. H. Cho, and C. Kim, "Super-resolution localization photoacoustic microscopy using intrinsic red blood cells as contrast absorbers," *Light: Sci. Appl.*, vol. 8, no. 1, pp. 1–11, Dec. 2019.

[3] L. V. Wang and J. Yao, "A practical guide to photoacoustic tomography in the life sciences," *Nature Methods*, vol. 13, no. 8, p. 627, Aug. 2016.

[4] A. Taruttis and V. Ntziachristos, "Advances in real-time multispectral optoacoustic imaging and its applications," *Nature Photon.*, vol. 9, no. 4, p. 219, 2015.

[5] K. Haedicke *et al.*, "High-resolution optoacoustic imaging of tissue responses to vascular-targeted therapies," *Nature Biomed. Eng.*, vol. 4, no. 3, pp. 286–297, Mar. 2020.

[6] J. Qi *et al.*, "Light-driven transformable optical agent with adaptive functions for boosting cancer surgery outcomes," *Nature Commun.*, vol. 9, no. 1, pp. 1–12, Dec. 2018.

[7] L. Li *et al.*, "Small near-infrared photochromic protein for photoacoustic multi-contrast imaging and detection of protein interactions *in vivo*," *Nature Commun.*, vol. 9, no. 1, pp. 1–14, Dec. 2018.

[8] H. Shan, G. Wang, and Y. Yang, "Accelerated correction of reflection artifacts by deep neural networks in photo-acoustic tomography," *Appl. Sci.*, vol. 9, no. 13, p. 2615, Jun. 2019.

[9] G. Matrone, A. S. Savoia, G. Caliano, and G. Magenes, "The delay multiply and sum beamforming algorithm in ultrasound B-mode medical imaging," *IEEE Trans. Med. Imag.*, vol. 34, no. 4, pp. 940–949, Apr. 2015.

[10] M. Xu and L. V. Wang, "Universal back-projection algorithm for photoacoustic computed tomography," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, no. 1, 2005, Art. no. 016706.

[11] L. Li, L. Zhu, Y. Shen, and L. V. Wang, "Multiview Hilbert transformation in full-ring transducer array-based photoacoustic computed tomography," *J. Biomed. Opt.*, vol. 22, no. 7, Jul. 2017, Art. no. 076017.

[12] K. P. Köstli, M. Frenz, H. Bebie, and H. P. Weber, "Temporal backward projection of optoacoustic pressure transients using Fourier transform methods," *Phys. Med. Biol.*, vol. 46, no. 7, p. 1863, 2001.

[13] B. E. Treeby, E. Z. Zhang, and B. T. Cox, "Photoacoustic tomography in absorbing acoustic media using time reversal," *Inverse Problems*, vol. 26, no. 11, 2010, Art. no. 115003.

[14] S. Bu *et al.*, "Model-based reconstruction integrated with fluence compensation for photoacoustic tomography," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1354–1363, May 2012.
[15] P. K. Yalavarthy, S. K. Kalva, M. Pramanik, and J. Prakash, "Non-local means improves total-variation constrained photoacoustic image reconstruction," *J. Biophoton.*, vol. 14, no. 1, 2021, Art. no. e202000191.
[16] D. Van de Sompel, L. S. Sasportas, J. V. Jokerst, and S. S. Gambhir, "Comparison of deconvolution filters for photoacoustic tomography," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, Art. no. e0152597.
[17] M. Jaeger, E. Robinson, H. G. Akarcay, and M. Frenz, "Full correction for spatially distributed speed-of-sound in echo ultrasound based on measuring aberration delays via transmit beam steering," *Phys. Med. Biol.*, vol. 60, no. 11, p. 4497, 2015.
[18] M. Jaeger, G. Held, S. Peeters, S. Preisser, M. Grünig, and M. Frenz, "Computed ultrasound tomography in echo mode for imaging speed of sound using pulse-echo sonography: Proof of principle," *Ultrasound Med. Biol.*, vol. 41, no. 1, pp. 235–250, Jan. 2015.
[19] C. Huang, K. Wang, L. Nie, L. V. Wang, and M. A. Anastasio, "Full-wave iterative image reconstruction in photoacoustic tomography with acoustically inhomogeneous media," *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1097–1110, Jun. 2013.
[20] M. Haltmeier and L. V. Nguyen, "Analysis of iterative methods in photoacoustic tomography with variable sound speed," *SIAM J. Imag. Sci.*, vol. 10, no. 2, pp. 751–781, 2017.
[21] S. Gutta, M. Bhatt, S. K. Kalva, M. Pramanik, and P. K. Yalavarthy, "Modeling errors compensation with total least squares for limited data photoacoustic tomography," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 1, pp. 1–14, Jan. 2019.
[22] J. Poudel and M. A. Anastasio, "Joint reconstruction of initial pressure distribution and spatial distribution of acoustic properties of elastic media with application to transcranial photoacoustic tomography," *Inverse Problems*, vol. 36, no. 12, Dec. 2020, Art. no. 124007.
[23] T. P. Matthews, J. Poudel, L. Li, L. V. Wang, and M. A. Anastasio, "Parameterized joint reconstruction of the initial pressure and sound speed distributions for photoacoustic computed tomography," *SIAM J. Imag. Sci.*, vol. 11, no. 2, pp. 1560–1588, 2018.
[24] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2503–2516, Aug. 2020.
[25] Z. Shao, L. Wang, Z. Wang, and J. Deng, "Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.
[26] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.
[27] A. C. Luchies and B. C. Byram, "Deep neural networks for ultrasound beamforming," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2010–2021, Sep. 2018.
[28] D. Hyun, L. L. Brickson, K. T. Looby, and J. J. Dahl, "Beamforming and speckle reduction using neural networks," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 5, pp. 898–910, May 2019.
[29] S. Antholzer, M. Haltmeier, and J. Schwab, "Deep learning for photoacoustic tomography from sparse data," *Inverse Problems Sci. Eng.*, vol. 27, no. 7, pp. 987–1005, 2018.
[30] 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, Jun. 2018.
[31] E. M. A. Anas, H. K. Zhang, C. Audigier, and E. M. Boctor, "Robust photoacoustic beamforming using dense convolutional neural networks," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11042, S. Aylward *et al.*, Eds. Springer-Verlag, 2018, pp. 3–11, doi: 10.1007/978-3-030-01045-4_1.
[32] S. Gutta, V. S. Kadimesetty, S. K. Kalva, M. Pramanik, S. Ganapathy, and P. K. Yalavarthy, "Deep neural network-based bandwidth enhancement of photoacoustic data," *J. Biomed. Opt.*, vol. 22, no. 11, 2017, Art. no. 116001.
[33] N. Awasthi, G. Jain, S. K. Kalva, M. Pramanik, and P. K. Yalavarthy, "Deep neural network-based sinogram super-resolution and bandwidth enhancement for limited-data photoacoustic tomography," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2660–2673, Dec. 2020.
[34] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 568–576, Feb. 2020.
[35] N. Davoudi, X. L. Deán-Ben, and D. Razansky, "Deep learning optoacoustic tomography with sparse data," *Nature Mach. Intell.*, vol. 1, no. 10, pp. 453–460, Oct. 2019.
[36] H. Shan, C. Wiedeman, G. Wang, and Y. Yang, "Simultaneous reconstruction of the initial pressure and sound speed in photoacoustic tomography using a deep-learning approach," *Proc. SPIE*, vol. 11105, Sep. 2019, Art. no. 1110504.
[37] A. Hauptmann *et al.*, "Model-based learning for accelerated, limited-view 3-d photoacoustic tomography," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1382–1393, Jun. 2018.
[38] J. Kim *et al.*, "Programmable real-time clinical photoacoustic and ultrasound imaging system," *Sci. Rep.*, vol. 6, p. 35137, Oct. 2016.
[39] J. Kim, E.-Y. Park, B. Park, W. Choi, K. J. Lee, and C. Kim, "Towards clinical photoacoustic and ultrasound imaging: Probe improvement and real-time graphical user interface," *Exp. Biol. Med.*, vol. 254, no. 4, 2020, Art. no. 1535370219889968.
[40] B. E. Treeby and B. T. Cox, "K-wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *J. Biomed. Opt.*, vol. 15, no. 2, 2010, Art. no. 021314.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer, 2015, doi: 10.1007/978-3-319-24574-4_28.
[42] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. Workshop Unsupervised Transf. Learn. (ICML)*, 2012, pp. 37–49.
[43] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[44] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[47] C. Lee, W. Choi, J. Kim, and C. Kim, "Three-dimensional clinical handheld photoacoustic/ultrasound scanner," *Photoacoustics*, vol. 18, Jun. 2020, Art. no. 100173.
[48] A. A. Oraevsky, B. Clingman, J. Zalev, A. T. Stavros, W. T. Yang, and J. R. Parikh, "Clinical optoacoustic imaging combined with ultrasound for coregistered functional and anatomical mapping of breast tumors," *Photoacoustics*, vol. 12, pp. 30–45, Dec. 2018.
[49] R. Penrose, "A generalized inverse for matrices," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3. Cambridge, U.K.: Cambridge Univ. Press, 1955, pp. 406–413.
[50] M. K. Feldman, S. Katyal, and M. S. Blackwood, "US artifacts," *Radio Graph.*, vol. 29, no. 4, pp. 1179–1189, Jul. 2009.
[51] X. Lin, M. Sun, Y. Liu, Z. Shen, Y. Shen, and N. Feng, "Variable speed of sound compensation in the linear-array photoacoustic tomography using a multi-stencils fast marching method," *Biomed. Signal Process. Control*, vol. 44, pp. 67–74, Jul. 2018.
[52] B. E. Treeby, T. K. Varslot, E. Z. Zhang, J. G. Laufer, and P. C. Beard, "Automatic sound speed selection in photoacoustic image reconstruction using an autofocus approach," *J. Biomed. Opt.*, vol. 16, no. 9, 2011, Art. no. 090501.
[53] R. A. Nicolaus, Melanins, Hermann, MO, USA, 1968.

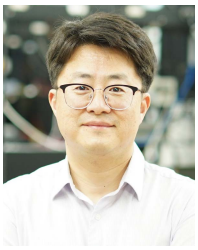
Seungwan Jeon was born in Bucheon, Republic of Korea, in 1989. He received the B.S. degree in biomedical engineering from Yonsei University, Wonju, Republic of Korea, in 2014, and the Ph.D. degree in creative IT engineering from POSTECH, Pohang, Republic of Korea, in 2020.

He is currently a Researcher with Samsung Electronics, Republic of Korea. He has authored/coauthored about 16 articles on international journals. His research interests include photoacoustic/ultrasound imaging techniques using image/signal processing, beamforming, and deep learning.

Wonseok Choi was born in Seoul, South Korea, in 1990. He received the bachelor's *(summa cum laude)* and Ph.D. degrees from the Department of Electrical Engineering, Pohang University of Science and Technology (POSTECH), in 2012 and 2020, respectively.

He is currently a Research Assistant Professor with the Department of Convergence IT Engineering, POSTECH, working on the clinical feasibility evaluation of photoacoustic imaging. He has authored/coauthored about 20 articles on international journals. His Ph.D. research focused on the development of dual-modal photoacoustic and ultrasound imaging technique for multi-structural imaging of human peripheral vasculature and the monitoring of high intensity focused ultrasound therapy. His current research interests include the ultrafast ultrasound imaging, tomographic image reconstruction techniques, and the convolutional neural network for photoacoustic/ultrasound image enhancement.

Byullee Park received the Ph.D. degree in creative IT engineering from the Pohang University of Science and Technology (POSTECH), Republic of Korea, in 2020. He has been undertaking a Postdoctoral Researcher training course at POSTECH since 2021. His research interests include development of novel biomedical imaging techniques, including sub-wavelength photoacoustic microscopy and clinical photoacoustic/ultrasound imaging.



Chulhong Kim (Senior Member, IEEE) received the B.S. degree in biomedical engineering from Kyungpook National University, Daegu, Republic of Korea, in 2004, and the Ph.D. degree in biomedical engineering from Washington University, St. Louis, in 2009.

He studied for his Ph.D. degree and post-doctoral training under the supervision of Dr. Lihong V. Wang, Dr. Younan Xia, and Dr. Samuel Achilefu. From 2010 to 2013, he was an Assistant Professor of biomedical engineering at the University at Buffalo, the State University, New York. He currently holds Mueunjae Chair Professorship and is a Professor of electrical engineering, creative it engineering, mechanical engineering, and interdisciplinary bioscience and bioengineering at POSTECH. He is also the Director of the Medical Device Innovation Center, POSTECH, and a Visiting Scholar of radiology at Stanford University. He has published 146 articles and coauthored six book chapters. His research interests include development of novel biomedical imaging techniques including photoacoustic tomography/microscopy, and ultrasound-modulated optical tomography.

Prof. Kim was a recipient of the 2020 Microscopy Today Innovation Award, the 2020–2021 IEEE EMBS Distinguished Lecturer, and the 2017 IEEE EMBS Academic Early Career Achievement Award. He has served as an Associate Editor for IEEE TRANSACTIONS ON MEDICAL IMAGING and a Topical Editor for IEEE ACCESS.