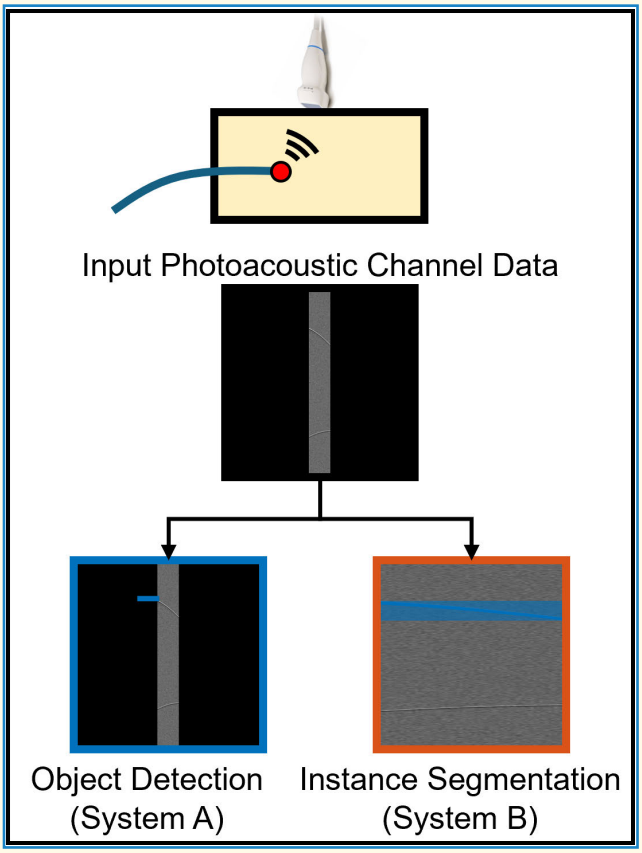
Deep Learning to Localize Photoacoustic Sources in Three Dimensions: Theory and Implementation

Mardava R. Gubbi, *Graduate Student Member, IEEE*, and Muyinatu A. Lediju Bell, *Senior Member, IEEE*

***Abstract*—Surgical tool tip localization and tracking are essential components of surgical and interventional procedures. The cross sections of tool tips can be considered as acoustic point sources to achieve these tasks with deep learning applied to photoacoustic channel data. However, source localization was previously limited to the lateral and axial dimensions of an ultrasound transducer. In this article, we developed a novel deep learning-based 3-D photoacoustic point source localization system using an object detection-based approach extended from our previous work. In addition, we derived theoretical relationships among point source locations, sound speeds, and waveform shapes in raw photoacoustic channel data frames. We then used this theory to develop a novel deep learning instance segmentation-based 3-D point source localization system. When tested with 4000 simulated, 993 phantom, and 1983 ex vivo channel data frames, the two systems achieved F1 scores as high as 99.82%, 93.05%, and 98.20%, respectively, and Euclidean localization errors (mean ± one standard deviation) as low as 1.46 ± 1.11 mm, 1.58 ± 1.30 mm, and 1.55 ± 0.86 mm, respectively. In addition, the instance segmentation-based system simultaneously estimated sound speeds with absolute errors (mean ± one standard deviation) of 19.22 ± 26.26 m/s in simulated data and standard deviations ranging 14.6–32.3 m/s in experimental data. These results demonstrate the potential of the proposed photoacoustic imaging-based methods to localize and track tool tips in three dimensions during surgical and interventional procedures.**

***Index Terms*—3-D localization, computer vision, deep learning, detection, imaging, phased arrays, photoacoustics, segmentation.**



Received 13 February 2025; accepted 15 April 2025. Date of publication 22 April 2025; date of current version 28 May 2025. This work was supported in part by the National Institutes of Health (NIH) Trailblazer Award R21-EB025621, in part by NIH Grant R01 EB032358, in part by NSF CAREER Award 1751522, and in part by NSF Smart and Connected Health (SCH) Award IIS-2014088. *(Corresponding author: Mardava R. Gubbi.)*

Mardava R. Gubbi is with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: mgubbi1@jhu.edu).

Muyinatu A. Lediju Bell is with the Department of Electrical and Computer Engineering, the Department of Computer Science, and the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: mledijubell@jhu.edu).

Digital Object Identifier 10.1109/TUFFC.2025.3562313

I. Introduction

SURGICAL tool tip localization and tracking are critical to the success of interventional procedures such as percutaneous liver biopsies [1] and cardiac catheter ablations [2]. In percutaneous liver biopsies, a needle is introduced through the skin and liver tissue to extract specimens for diagnostic evaluation of a variety of liver disorders [3], including nonalcoholic fatty liver disease which affects up to 100 million people in the USA [4]. In cardiac catheter ablations, performed on 18 000–45 000 people annually in the USA [5], a catheter is navigated from an insertion point in

1525-8955 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

Highlights

- **We developed two novel deep learning-based 3-D photoacoustic point source localization systems using object detection and instance segmentation paradigms with theory-based performance optimizations.**
- **The object detection approach successfully localized point sources in the elevation dimension, while the instance segmentation approach estimated both the 3-D point source location and sound speed.**
- **These systems can be deployed to track and visualize surgical tool tips during photoacoustic-guided interventional procedures implemented with or without robotic assistance.**

the thigh to the heart via the femoral vein. In the absence of sufficient navigation information, navigation errors could cause pain and intraperitoneal bleeding after percutaneous liver biopsies [6], [7] or perforation of heart tissue during cardiac catheterization procedures [5]. Therefore, these procedures are typically performed with the needle or catheter tip locations observed in real time to reduce the risk of patient injury and related complications [2], [7].

Traditional medical imaging modalities utilized to estimate the locations of surgical needle, catheter, and other tool tips during interventional procedures include computed tomography (CT) [8], [9], [10], magnetic resonance imaging (MRI) [11], [12], [13], and fluoroscopy [14], [15], [16]. However, these imaging systems are typically expensive, large, and difficult to transport, thereby limiting the ability of these modalities to improve global access to quality healthcare. In addition, CT and fluoroscopy expose both the patient and the interventionalist or surgeon to ionizing radiation [17], [18], resulting in potential biological effects [19] including radiodermatitis [20], [21], increased cancer risks [22], [23], [24], [25], and genetic defects [23], [25].

Ultrasound imaging is an alternative to CT, MRI, or fluoroscopy and overcomes the noted limitations of these techniques with its low cost, portability, and absence of ionizing radiation. With ultrasound, the transmission and reception of acoustic waves are employed to reconstruct human-interpretable images using beamforming algorithms, such as delay-and-sum (DAS) [26], [27]. Ultrasound imaging is commonly used to guide percutaneous liver biopsies [28] and cardiac procedures [29], yet fails in acoustically challenging environments characterized by significant acoustic clutter [30], sound scattering, and signal attenuation. Some examples of acoustically challenging environments include transcranial imaging [31], abdominal imaging [30], spinal imaging [32], or the imaging of obese patients [33].

Photoacoustic imaging is an emerging imaging modality with the potential to address the noted limitations of ultrasound imaging. Unlike ultrasound, which requires the transmission and reception of sound to make images, photoacoustic imaging transmits light to generate an acoustic response that can be received with the same ultrasound transducer. Therefore, photoacoustic imaging only requires one-way (as opposed to round-trip) acoustic travel from the transmission source to the ultrasound receiver, improving tool tip localization in acoustically challenging environments [34]. The received signals are typically processed using similar beamforming algorithms to ultrasound imaging (e.g, DAS) [35].

Su et al. [36] previously visualized needles in beamformed photoacoustic images overlaid on ultrasound images with potential applications to a variety of biopsy procedures. Lediju Bell and Shubert [37] developed a robotic system to autonomously identify, localize, and track a needle tip in real time in ex vivo tissue using amplitude information in DAS-beamformed photoacoustic images. This system was also utilized to track a catheter tip in real time during a cardiac catheterization procedure performed on an in vivo swine [38]. However, this amplitude-based system was sensitive to reflection artifacts from surrounding structures (e.g., bone), limiting the ability of the system to consistently maintain the catheter tip in the field-of-view (FOV) of the transducer for the duration of the procedure. Additional limitations include the loss of information between the input channel data and the output DAS image, which degrades target resolution with increasing depth, thus limiting the ability to track deep targets.

To overcome the limitations of amplitude-based systems, Allman et al. [39] modeled surgical tool tips as photoacoustic point sources. Using this model, Allman et al. [39] developed a deep learning-based approach to detect photoacoustic point sources directly in raw channel data frames and distinguish the identified sources from reflection artifacts. This approach formulated the point source localization problem as an object detection problem to identify waveforms corresponding to either sources or reflection artifacts followed by a two-class (i.e., source or artifact) classification problem to categorize the waveforms based on the underlying targets.

The use of raw channel data as inputs enabled the deep learning-based approach [39] to achieve consistent localization performance across a wide range of target depths. This deep learning-based system was integrated with a robotic control system to localize and track needle tips in real time in a plastisol phantom [40], [41] and ex vivo chicken breast [40], without the unnecessary computational overhead of reconstructing human-interpretable images using DAS beamforming. A similar deep learning-based approach was applied to photoacoustic channel data frames of catheter tips in ex vivo and in vivo swine hearts with potential applicability to cardiac catheterization procedures [42]. However, these deep learning-based systems are limited to estimating the lateral and axial displacements of targets and are unable to provide usable position information along the elevation dimension of the transducer. In addition, these systems assumed a fixed speed of sound in the surrounding medium.

Wang et al. [43] developed a 3-D photoacoustic-based needle tip localization system by autonomously scanning the

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

elevation dimension of the ultrasound transducer using a robotic arm. However, this system required 40 frames to generate each 3-D photoacoustic image grid. With a 10 Hz laser pulse repetition frequency, this requirement resulted in each 3-D image grid requiring 4 s to be generated (i.e., effective frame rate of 0.25 Hz). This low effective frame rate limits applicability to interventional procedures requiring real-time surgical tool tip location information.

In this article, we define three objectives to localize a photoacoustic point source in three spatial dimensions from a single channel data frame, enabling the design of two novel deep learning-based photoacoustic point source localization systems. First, we extend the two-class classification model of sources and artifacts proposed by Allman et al. [39] to a 22-class classification model with the elevation displacement information encoded in the class names. Second, we derive a theoretical framework using wave propagation time calculations to relate the 3-D point source location, the speed of sound in the underlying medium, and the shape of the waveform in the channel data frame acquired by the transducer. Third, we design a theory-based least squares optimization algorithm using our new theoretical framework to estimate the location of the point source and the surrounding medium sound speed. The first objective described above results in an object detection-based 3-D photoacoustic point source localization system, while a combination of the second and third objectives results in an instance segmentation-based 3-D point source localization and sound speed estimation system. We train, test, and evaluate the detection, segmentation, localization, and sound speed estimation performance of the two systems resulting from these three objectives.

The remainder of this article is organized as follows. Section II presents a theoretical framework relating the shape of the source waveform to the location of the corresponding point source, relative to the transducer and medium sound speed. Section III describes the simulation and data acquisition process, training and testing methods for our deep learning-based photoacoustic point source localization systems, and associated performance metrics. Section IV reports the associated results. Section V discusses the implications and future potential of our work. Finally, Section VI concludes this article with a summary of our key findings.

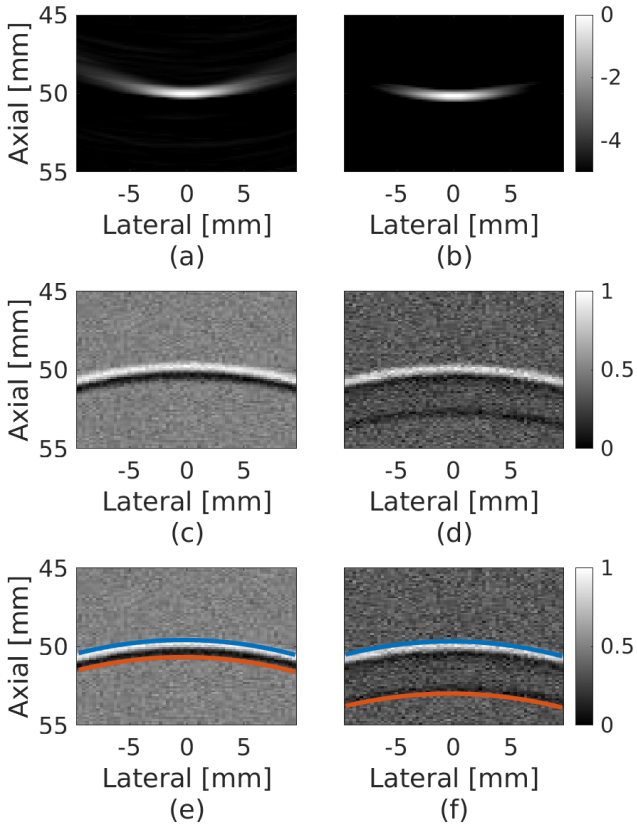


Fig. 1. Simulated photoacoustic images of point sources located (a) within and (b) outside the elevation width of an ultrasound transducer, with corresponding raw channel data frames (c) and (d) before and (e) and (f) after annotating the lower (blue) and upper (orange) bounds of the waveform limits.

II. Theory

A. Context and Overview

To contextualize our theoretical framework, Fig. 1 shows simulated DAS-beamformed images, followed by raw channel data frames from two photoacoustic point sources imaged by an ultrasound transducer. The first source [Fig. 1(a)] is centered in the lateral and elevation dimensions of the transducer. The second source [Fig. 1(b)] is centered in the lateral dimension and displaced outside the physical elevation limits of the transducer. The elevation displacement of the second source has reduced signal energy (e.g., due to attenuation), which translates to reduced brightness in the corresponding DAS-beamformed image (relative to the first source), yet the signal shape at the center of each source appears to be similar. This similarity is expected as a result of the information loss inherent to the DAS beamforming process.

The reduction in signal amplitudes associated with the elevation displacement is more evident in the corresponding channel data frames shown in Fig. 1(c) and (d). In addition, the increased elevation displacement of the second source (relative to the first source) results in a larger separation between the upper and lower bounds of the source waveform, as indicated by the annotations in Fig. 1(e) and (f). In particular, the "lower" bound is defined as having lower propagation time than the "upper" bound.

The visible dependence of the waveform shape on the elevation source displacement motivates the development of a theoretical framework relating the 3-D location of the photoacoustic point source to the shape of the corresponding waveform in raw photoacoustic channel data. Our associated theoretical framework must meet two criteria: 1) characterize the shape of the waveform in the channel data frame as a function of the properties of the photoacoustic point source and medium and 2) estimate the 3-D location of the photoacoustic point source and sound speed in the surrounding medium using the segmented shape of this waveform.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

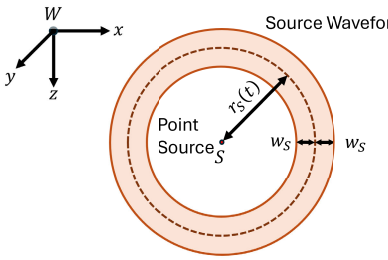


Fig. 2. Photoacoustic point source located at point S in a reference frame W with pressure wave modeled as a hollow sphere of thickness $2 \times w_S$ propagating outward from the source with time-varying radius $r_S(t)$.

To characterize the shape of the waveform as a function of the properties of the corresponding photoacoustic point source, Section II-B defines a point source in a homogeneous medium and models the pressure wave propagating from the point source. Section II-C models an ultrasound transducer as an array of transducer elements. Section II-D derives the shape of the waveform corresponding to the point source from the propagation time calculations for each individual transducer element. Section II-E models a photoacoustic point source localization system taking a raw channel data frame as input and providing estimates of the point source location with surrounding medium sound speed as outputs.

B. Photoacoustic Point Source

To characterize the shape of the waveform corresponding to a photoacoustic point source in a channel data frame acquired by an ultrasound transducer, we first define a photoacoustic point source in a homogeneous medium with speed of sound c. Let this point source be located at the point S, as shown in Fig. 2. The point S is represented by $\vec{x}_S$ in the reference frame W, given by

$$\vec{x}_S = \left[x_S, y_S, z_S\right]^T \tag{1}$$

where $z_S \geq 0$.

We define a pressure wave propagating spherically outward from point S with speed c, as shown in Fig. 2. We model this pressure wave as a hollow sphere of thickness $2 \times w_S$ and radius $r_S(t)$ given by

$$r_S(t) = ct \tag{2}$$

where t is the duration of time for which the wave has been propagating outward from the point source. To model the instantaneous amplitude $p_D(l,t)$ of the pressure wave at a distance l from the point source, we define a linear approximation of the N-shaped waveform presented by Diebold et al. [44] as follows:

$$p_D(l,t) = \begin{cases} \dfrac{p_0}{w_S}[l - r_S(t)], & \text{if } |l - r_s(t)| \leq w_S \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where p_0 is the peak amplitude of the pressure wave at distance l from the point S. Using (3) we define the instantaneous pressure amplitude at a point X as follows:

$$p_P\left(\vec{x}_X, t; \vec{x}_S\right) = p_D\left(\left\|\vec{x}_S - \vec{x}_X\right\|, t\right) \tag{4}$$

where $\vec{x}_X$ is the location of the point X in the reference frame W given by

$$\vec{x}_X = \left[x_X, y_X, z_X\right]^T. \tag{5}$$

From (3) and (5), we observe that the pressure $p_P(\vec{x}_X, t; \vec{x}_S)$ is nonzero in the time interval given by

$$\left\|\vec{x}_S - \vec{x}_X\right\| - w_S \leq r_S(t) \leq \left\|\vec{x}_S - \vec{x}_X\right\| + w_S. \tag{6}$$

C. Modeling a Transducer Array

Consider a transducer array consisting of N_T elements of height h_T, width w_T, and pitch p_T. For commercially available 1-D array transducers, the element width w_T is typically an order of magnitude smaller than the height h_T. Therefore, for ease of computation we model each transducer element as a line segment of height h_T. Let this transducer array be held stationary in contact with a homogeneous medium with speed of sound c. We refer to an individual element in this transducer array by the index n, where $0 \leq n < N_T$. Assuming that the center of the transducer array coincides with the origin of the reference frame W, we define the center $T(n)$ of transducer element n in the frame W with location $\vec{x}_T(n)$ given by

$$\vec{x}_T(n) = [x_T(n), 0, 0]^T \tag{7}$$

where

$$x_T(n) = \left(n - \frac{N_T - 1}{2}\right) p_T. \tag{8}$$

We also define points $N(n)$ and $F(n)$ on the surface of transducer element n which are nearest to and farthest from the point S, respectively. The coordinates of the point $N(n)$ are given by

$$\vec{x}_N(n) = \left[x_T(n), y_N, 0\right]^T \tag{9}$$

where y_N is given by

$$y_N = \begin{cases} h_T/2, & \text{if } \ y_S \geq h_T/2 \\ y_S, & \text{if } \ |y_S| < h_T/2 \\ -h_T/2, & \text{otherwise.} \end{cases} \tag{10}$$

The coordinates of the point $F(n)$ are given by

$$\vec{x}_F(n) = \left[x_T(n), y_F, 0\right]^T \tag{11}$$

where y_F is given by

$$y_F = \begin{cases} -h_T/2, & \text{if } y_S \geq 0 \\ h_T/2, & \text{otherwise.} \end{cases} \tag{12}$$

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

D. Shape of Waveform Corresponding to Photoacoustic Point Source in Channel Data Frame

Let the transducer acquire N_S samples of received pressure amplitude corresponding to a desired imaging depth D_I at a sampling rate of f_S. The value of N_S is given by

$$N_S = D_I\left(\frac{f_S}{c}\right). \tag{13}$$

These samples are organized to form a raw photoacoustic channel data frame of $N_T \times N_S$ pixels. We define a point Q located at

$$\vec{p}_Q = (m_Q, n_Q) \tag{14}$$

where m_Q and n_Q are the row and column indices, respectively, of point Q in the raw channel data frame. The received signal at point Q corresponds to the pressure amplitude received by the transducer element n_Q at time instant t_Q, defined as follows:

$$t_Q = \frac{m_Q}{f_S}. \tag{15}$$

Let the region $\Psi(\vec{x}_S, c)$ correspond to the source waveform in the channel data frame acquired by the transducer array. To characterize $\Psi(\vec{x}_S, c)$, we define the sets $L(\vec{x}_S, c)$ and $U(\vec{x}_S, c)$ forming the lower and upper bounds, respectively, of $\Psi(\vec{x}_S, c)$, as illustrated in Fig. 1(e) and (f).

To compute the relationship between m_Q and n_Q for a point Q in the set $L(\vec{x}_S, c)$, we must determine the instant at which the pressure wave from the point source first reaches the transducer element n_Q. Based on (2), (6), and (15), the locus of points in $L(\vec{x}_S, c)$ is given by

$$\left\|\vec{x}_S - \vec{x}_N(n_Q)\right\| - w_S = \frac{cm_Q}{f_S}. \tag{16}$$

Substituting (1) and (8)–(10) into (16), the locus of points in $L(\vec{x}_S, c)$ is given by

$$\left(\frac{cm_Q}{f_S} + w_S\right)^2 - \left[\left(n_Q - \frac{N_T - 1}{2}\right)p_T - x_S\right]^2 = z_S^2 \tag{17}$$

for point source locations S within the elevation limits of the transducer (i.e., $|y_S| \le h_T/2$), and

$$\left(\frac{cm_Q}{f_S} + w_S\right)^2 - \left[\left(n_Q - \frac{N_T - 1}{2}\right)p_T - x_S\right]^2 = \left(|y_S| - \frac{h_T}{2}\right)^2 + z_S^2 \tag{18}$$

for point source locations S outside the elevation limits of the transducer (i.e., $|y_S| > h_T/2$). Equations (17) and (18) can be represented as hyperbolic curves of the form

$$\left(\frac{m_Q - z_L(\vec{x}_S, c)}{b_L(\vec{x}_S, c)}\right)^2 - \left(\frac{n_Q - x_L(\vec{x}_S, c)}{a_L(\vec{x}_S, c)}\right)^2 = 1 \tag{19}$$

where z_L, x_L, b_L, and a_L, are defined as follows:

$$z_L(\vec{x}_S, c) = -\frac{w_S f_S}{c} \tag{20}$$

$$x_L(\vec{x}_S, c) = \left(\frac{N_T - 1}{2}\right) + \frac{x_S}{p_T} \tag{21}$$

$$b_L(\vec{x}_S, c) = \begin{cases} f_S z_S/c, & \text{if } |y_S| \le h_T/2 \\ \dfrac{f_S\sqrt{\left(|y_S| - \frac{h_T}{2}\right)^2 + z_S^2}}{c}, & \text{otherwise} \end{cases} \tag{22}$$

and

$$a_L(\vec{x}_S, c) = \begin{cases} z_S/p_T, & \text{if } |y_S| \le h_T/2 \\ \dfrac{\sqrt{\left(|y_S| - \frac{h_T}{2}\right)^2 + z_S^2}}{p_T}, & \text{otherwise.} \end{cases} \tag{23}$$

Using the parameters above, we obtain the following expression for the set $L(\vec{x}_S, c)$, which defines the lower bound of the waveform (i.e., the wavefront)

$$L(\vec{x}_S, c) = \left\{ \vec{p}_Q : \begin{array}{c} \left(\dfrac{m_Q - z_L(\vec{x}_S, c)}{b_L(\vec{x}_S, c)}\right)^2 \\ -\left(\dfrac{n_Q - x_L(\vec{x}_S, c)}{a_L(\vec{x}_S, c)}\right)^2 = 1 \end{array} \right\}. \tag{24}$$

To compute the relationship between m_Q and n_Q for a point Q in the set $U(\vec{x}_S, c)$, we must determine the instant at which the inner surface of the hollow spherical pressure wave from the point source passes the transducer element n_Q. Based on (2), (6), and (15), the locus of points in $U(\vec{x}_S, c)$ is given by

$$\left\|\vec{x}_S - \vec{x}_F(n_Q)\right\| + w_S = \frac{cm_Q}{f_S}. \tag{25}$$

Substituting (1), (8), (11), and (12) into (25), the locus of points in $U(\vec{x}_S, c)$, for point source locations S within and outside the elevation limits of the transducer, is given by

$$\left(\frac{cm_Q}{f_S} - w_S\right)^2 - \left[\left(n_Q - \frac{N_T - 1}{2}\right)p_T - x_S\right]^2 = \left(|y_S| + \frac{h_T}{2}\right)^2 + z_S^2. \tag{26}$$

Equation (26) can be represented as a hyperbolic curve of the form

$$\left(\frac{m_Q - z_U(\vec{x}_S, c)}{b_U(\vec{x}_S, c)}\right)^2 - \left(\frac{n_Q - x_U(\vec{x}_S, c)}{a_U(\vec{x}_S, c)}\right)^2 = 1 \tag{27}$$

where z_U, x_U, b_U, and a_U, are defined as follows:

$$z_U(\vec{x}_S, c) = \frac{w_S f_S}{c} \tag{28}$$

$$x_U(\vec{x}_S, c) = \left(\frac{N_T - 1}{2}\right) + \frac{x_S}{p_T} \tag{29}$$

$$b_U(\vec{x}_S, c) = \frac{f_S\sqrt{\left(|y_S| + \frac{h_T}{2}\right)^2 + z_S^2}}{c} \tag{30}$$

and

$$a_U(\vec{x}_S, c) = \frac{\sqrt{\left(|y_S| + \frac{h_T}{2}\right)^2 + z_S^2}}{p_T}. \tag{31}$$

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

Using the parameters derived above, we obtain the following expression for the set $U(\vec{x}_S, c)$, which defines the upper bound of the waveform:

$$U(\vec{x}_S, c) = \left\{ \vec{p}_Q : \left(\frac{m_Q - z_U(\vec{x}_S, c)}{b_U(\vec{x}_S, c)} \right)^2 - \left(\frac{n_Q - x_U(\vec{x}_S, c)}{a_U(\vec{x}_S, c)} \right)^2 = 1 \right\}. \quad (32)$$

E. Estimating Point Source Location and Speed of Sound From Waveform Shape

To estimate the location of a point source from the shape of the corresponding waveform, we model an instance segmentation-based point source localization system. This model utilizes the waveform shape $\Psi(\vec{x}_S, c)$ derived in Section II-D, the point source location, $\vec{x}_S$, and surrounding sound speed, c, as the ground-truth information. The corresponding system consists of two stages to estimate the location of the point source and sound speed. First, an instance segmentation algorithm takes the channel data frame containing the waveform $\Psi(\vec{x}_S, c)$ as input and outputs an estimate $\hat{\Psi}$ of the ground-truth waveform $\Psi(\vec{x}_S, c)$. The locations of the peaks of the hyperbolic curves forming the upper and lower bounds of $\hat{\Psi}$ are used to generate an initial estimate $\hat{x}_S$ of the lateral and axial position of the source, assuming a fixed sound speed of 1540 m/s and zero elevation displacement. Second, an iterative least squares optimization algorithm is implemented to improve the performance of the instance segmentation-based point source localization system.

Taking the segmented region $\hat{\Psi}$ and initial point source location estimate $\hat{x}_S$ from the first stage as inputs, the iterative least squares optimization algorithm first simplifies the input $\hat{\Psi}$ by extracting estimates $\hat{L}$ and $\hat{U}$ of the sets $L(\vec{x}_S, c)$ and $U(\vec{x}_S, c)$, respectively. With the ground-truth information unavailable to the point source localization system, the quality of the source location estimate $\hat{x}_S$ for an assumed speed of sound $\hat{c}$ may be represented by the residual function $r_L(\vec{p}_Q, \hat{x}_S, \hat{c})$ for any point Q in the set $\hat{L}$

$$r_L(\vec{p}_Q, \hat{x}_S, \hat{c}) = m_Q - z_L(\hat{x}_S, \hat{c}) - \left[b_L(\hat{x}_S, \hat{c}) \sqrt{1 + \left(\frac{n_Q - x_L(\hat{x}_S, \hat{c})}{a_L(\hat{x}_S, \hat{c})} \right)^2} \right] \quad (33)$$

and the residual function $r_U(\vec{p}_Q, \hat{x}_S, \hat{c})$ for any point Q in the set $\hat{U}$

$$r_U(\vec{p}_Q, \hat{x}_S, \hat{c}) = m_Q - z_U(\hat{x}_S, \hat{c}) - \left[b_U(\hat{x}_S, \hat{c}) \sqrt{1 + \left(\frac{n_Q - x_U(\hat{x}_S, \hat{c})}{a_U(\hat{x}_S, \hat{c})} \right)^2} \right]. \quad (34)$$

TABLE I
RANGES AND INCREMENT SIZES OF PARAMETERS USED TO GENERATE SIMULATED DATASETS

Parameters	Min	Max	Increment
Speed of Sound [m/s]	1440	1640	6
Axial Position [mm]	20	100	0.2
Lateral Position [mm]	-18.2	18.2	0.1
Elevation Position [mm]	0.0	10.0	0.1
Channel SNR [dB]	-5	2	continuous
Object Intensity (multiplier)	0.75	1.1	continuous

Beginning with the initial estimates $\hat{x}_S$ and $\hat{c}$, each iteration of the optimization algorithm minimizes the objective function

$$J(\hat{x}_S, \hat{c}, \hat{L}, \hat{U}) = \sum_{\vec{p}_Q \in \hat{L}} r_L^2(\vec{p}_Q, \hat{x}_S, \hat{c}) + \sum_{\vec{p}_Q \in \hat{U}} r_U^2(\vec{p}_Q, \hat{x}_S, \hat{c}). \quad (35)$$

Once optimized, the final outputs $\hat{x}_S$ and $\hat{c}$ are obtained.

III. METHODS

A. Channel Data Acquisition

1) Photoacoustic Point Source Simulations: Channel data were simulated using k-Wave [45], building on methods previously developed for linear [39] and phased array transducers [42]. In particular, each simulation consisted of a point source of radius 100 μm in a 3-D grid containing a homogeneous medium. The top face of the simulation grid was populated with sensing elements to record the local pressure distribution at each time instant during the simulation. The resolution of the simulation grid was 100 μm and its lateral, elevation, and axial dimensions were 38.4, 25, and 120 mm, respectively. These parameters were selected to simulate a Verasonics P4-2v phased array ultrasound transducer (Kirkland, WA, USA) with 64 elements and a sampling rate of 11.88 MHz. The pitch, element height, and aperture length of the transducer were 300 μm, 14 mm [46], and 19.2 mm, respectively. To reduce the GPU memory and processing times required for 3-D simulations, we modeled the transducer as a continuous aperture [47] rather than the more accurate discrete aperture model [39], [40], [42], [48].

A unique k-Wave simulation was conducted for each possible combination of sound speeds and source axial positions selected from the parameters in Table I (i.e., 34 sound speeds and 401 source axial positions yielded 13 634 unique simulations). The lateral and elevation positions of the point source were kept constant across simulations at 9.6 and 7 mm, respectively, from the corresponding lateral and elevation edges of the simulation grid. These fixed positions combined with the simulation grid dimensions selected above minimized the number of simulations required for all possible source positions summarized in Table I.

For each simulation, the initial pressure distribution of the point source was modeled as a sphere of radius 100 μm and smoothed using a Blackman filter [49]. To satisfy the Courant–Friedrichs–Lewy condition [50], the simulations were conducted with a time step of 17.5 ns, corresponding to a sampling rate of 57.14 MHz. Each simulation was conducted

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

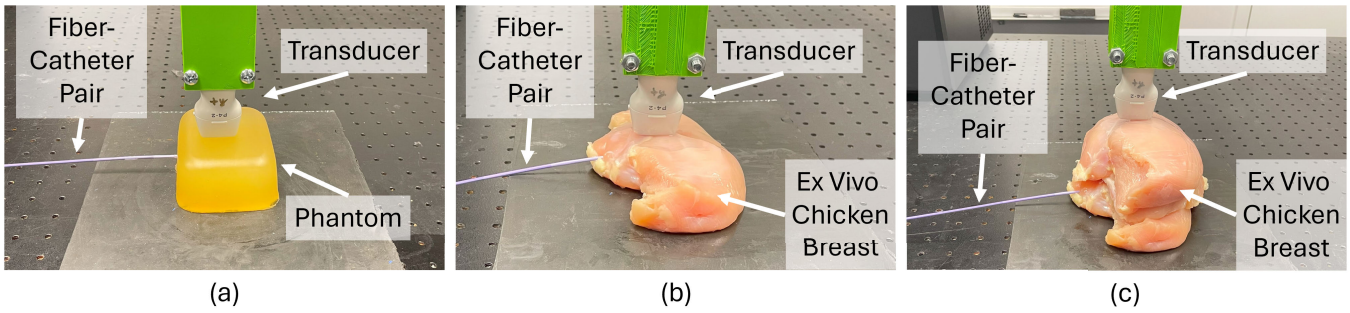


Fig. 3. Experimental setups to acquire phantom data and ex vivo photoacoustic channel data frames with a fiber–catheter pair depth of (a) 25 mm, (b) 23 mm, and (c) 46 mm relative to the imaging surface of the transducer. The transducer was attached to a robot arm and translated in the elevation dimension to acquire photoacoustic channel data frames of the tip of the fiber–catheter pair at varying elevation displacements from the transducer center.

for 4452 time steps, corresponding to an imaging depth of 12 cm at a sound speed of 1540 m/s. These simulations were performed on four NVIDIA (Santa Clara, CA, USA) Quadro RTX 8000 GPUs.

To train and test 3-D deep learning-based photoacoustic point source localization systems (described in Section III-D), each raw channel data frame consisted of one point source and at most one reflection artifact with parameters randomly sampled from the range reported in Table I, resulting in 20 000 raw photoacoustic channel data frames. Care was taken to ensure a uniform distribution of sound speeds across the raw channel data frames. For each frame, the simulation output corresponding to the selected sound speed and point source axial position was cropped to a 19.2 mm (lateral) × 14 mm (elevation) × 120 mm (axial) region of interest (ROI) centered on the selected source lateral and elevation positions. This cropped matrix was then integrated across the lateral and elevation dimensions to match the transducer pitch and element height. The integrated matrix was then axially resampled to the transducer sampling frequency to form a true photoacoustic source signal of dimensions 64 × 926 pixels.

Reflection artifacts were generated as described by Allman et al. [39] (i.e., a true photoacoustic source signal was shifted deeper into the image by the Euclidean distance between the source and reflector locations). Multiple simulation outputs were superimposed to form sources and reflection artifacts with a fixed radius of 500 μm. The superimposed source and reflection artifact corresponding to each raw channel data frame were multiplied by scalar object intensity multipliers (randomly sampled from the range in Table I) and the scaled frames were added. The resulting matrix was bandpass filtered to allow ±54.7% of the center frequency of the transducer. This bandwidth corresponded to the −20 dB threshold of the transducer specified by the manufacturer. Gaussian noise was then added to the filtered matrix using the `addNoise` function in the *k*-Wave toolbox [45] to form a raw channel data frame.

2) Experimental Photoacoustic Data: To acquire experimental phantom data, a 600 μm-core diameter optical fiber was inserted into a 7F outer diameter cardiac catheter (Boston Scientific, Marlborough, MA, USA), forming a fiber–catheter pair with the fiber and catheter tips coincident. This fiber–catheter pair was inserted into a plastisol phantom, as shown in Fig. 3(a). The other end of the optical fiber was interfaced with a Phocus Mobile laser (Opotek, Carlsbad, CA, USA) operating at a wavelength of 750 nm, a laser energy of 1.4 mJ, and a pulse repetition frequency of 10 Hz. A Verasonics Vantage 128 scanner connected to a P4-2v transducer was employed for imaging. The P4-2v transducer was mounted on an UR5e robot (Universal Robots, Odense, Denmark) via a custom 3-D-printed adapter. This entire system was designed to save raw channel data and the corresponding synchronized position of the center of the transducer with respect to the fixed robot base.

The catheter was initially aligned with the lateral dimension of the transducer, as shown in Fig. 3(a). To center the catheter tip in the lateral dimension of the image, the robot translated the transducer along its lateral dimension, until the peak of the hyperbolic waveform corresponding to the catheter tip was centered in the channel data. To center the catheter tip in the elevation dimension of the transducer, the robot first rotated the transducer by 90°, then translated the transducer along its lateral dimension until the associated photoacoustic signal was laterally centered in the image, followed by another 90° rotation to return to the original alignment between the catheter and lateral transducer dimension, with the catheter tip now centered in both the lateral and elevation dimensions of the transducer. The transducer was then translated 4 mm along the elevation dimension in steps of 1 mm. At each step, multiple channel data frames were acquired and stored with the corresponding robot pose information, resulting in a total of 993 raw channel data frames of the catheter tip at a measured depth of 25 mm in the phantom.

To acquire experimental ex vivo data, the fiber–catheter pair described above was inserted into ex vivo chicken breast, as shown in Fig. 3(b), at a measured depth of 23 mm. The procedure described above was employed to center the transducer above the catheter tip in the lateral and elevation transducer dimensions. A total of 992 channel data frames were acquired with the transducer translated 4 mm along the elevation dimension in steps of 1 mm. This experimental acquisition was then repeated with additional chicken breast tissue and the catheter tip placed at an increased depth of 46 mm, as shown in Fig. 3(c), and 991 channel data frames were acquired.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

B. Sound Speed Estimation

To estimate experimental sound speeds, we utilized the maximum lag one coherence (mLOC) introduced by Zhang et al. [51]. First, we validated the mLOC-based sound speed estimation technique on a subset of our simulated dataset consisting of 95 frames, with known ground truths. Each frame in this subset consisted of one source and at most one reflection artifact, with the lateral and elevation positions of the source constrained within <1 mm from the transducer center (based on previous results demonstrating greater coherence of photoacoustic sources and better estimation performance expectations near an array center [52]). The source axial positions and sound speeds were not constrained, ranging 20.6–98.4 mm and 1440–1638 m/s, respectively. Time delays corresponding to sound speeds ranging 1440–1640 m/s with an increment of 5 m/s were applied to each channel data frame in this subset, creating a total of 41 separate delayed channel data frames. The correlation of delayed channel data received by equally spaced elements (i.e., spatial lags) was calculated using the normalized spatial coherence function [53]

$$\hat{R}(m) = \frac{1}{N_T - m} \sum_{i=1}^{N_T - m} \frac{\sum_{n=n_1}^{n_2} s_i(n) s_{i+m}(n)}{\sqrt{\sum_{n=n_1}^{n_2} s_i^2(n) \sum_{n=n_1}^{n_2} s_{i+m}^2(n)}} \quad (36)$$

where m is the spatial lag (fixed at one), $s_i(n)$ is the time-delayed zero-mean photoacoustic signal received from the ith element at the nth sampling instant, and $n_2 - n_1$ is the axial correlation kernel size (fixed to approximately one acoustic wavelength). These calculations output a prescan converted matrix of coherence values with an angular width of $\pi/2$ radians and height 12 cm corresponding to the sector-shaped FOV of the phased array transducer. To fully enclose the target in each generated coherence function matrix, a rectangular ROI of angular width $\pi/3$ radians and height 10 mm was centered on the point (θ_S, z_B), where θ_S is the azimuthal component of the known target location in the prescan converted matrix. The axial position z_B is given by

$$z_B = \frac{z_S c_{\text{bf}}}{c_{\text{ch}}} \quad (37)$$

where c_{ch} is the sound speed in the channel data frame, c_{bf} is the beamformed sound speed, and z_S is the axial component of the known target location in the channel data frame. A representative sample of these ROIs is shown in Fig. 4(a). These rectangular ROIs corresponded to annular sectors in the scan converted images as shown in Fig. 4(b). The maximum value within the ROI was reported as the mLOC per channel data frame per sound speed. The sound speed corresponding to the maximum mLOC per channel data frame was validated against the known ground truth from simulated data.

The mLOC-based sound speed estimation process summarized above was repeated for a subset of the experimental datasets described in Section III-A2, after adding Gaussian noise corresponding to a channel SNR of −5 dB to each channel data frame using the `addNoise` function in the k-Wave toolbox (to improve the performance of the mLOC-based sound speed estimation technique). The selected raw channel data frames contained sources with elevation displacements

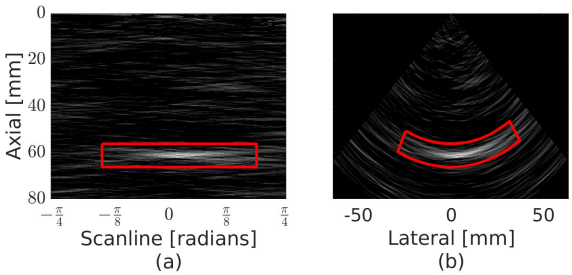


Fig. 4. Identical regions of interest in (a) pre- and (b) postscan converted short-lag spatial coherence images ($m = 1$) of a simulated photoacoustic point source.

<1 mm from the center of the transducer (based on previous results [52]), resulting in a total of 199 and 376 frames from the phantom and ex vivo datasets, respectively.

C. Image Annotation Process for Training and Testing

1) Object Detection-Based Annotated Images: As with previous implementations of object detection-based point source localization systems for phased array transducers [42], [54], [55], the FOV of the phased array transducer in a beamformed and scan-converted image extends laterally beyond the width of the raw channel data frame. Therefore, each raw channel data frame in the simulated, phantom, and ex vivo datasets was zero-padded to match this FOV to form a zero-padded channel data frame of dimensions 566 × 926 pixels.

To annotate the zero-padded channel data frames, we defined two super-classes for sources and artifacts. Each super-class was then divided into 11 distinct classes based on the corresponding elevation displacements rounded to and labeled based on the nearest millimeter (e.g., class 1 consisted of sources with elevation displacements constrained by -0.5 mm $\leq y_S <$ 0.5 mm and was named "Source-0.0"). For each zero-padded channel data frame, bounding boxes of dimensions 64 × 25 pixels were centered on the positions of the sources and artifacts within the frame. These bounding boxes were allowed to exist in the zero-padded regions if required, as shown in Fig. 5(a). A fully annotated image consisted of the zero-padded channel data frame, the coordinates of the bounding boxes, the elevation-encoded classes (e.g., "Source-0.0" "Artifact-10.0," etc.), the source and artifact position information, and the speed of sound corresponding to the frame. The ground-truth sound speed was known from simulated data, and the mLOC-based sound speed estimates (Section III-B) were used as ground-truth annotations in experimental data. The totality of fully annotated simulation images was randomly split into 16 000 training and 4000 test images. The fully annotated experimental images were only used during testing (and were not incorporated into the training set).

2) Instance Segmentation-Based Annotation: The 20 000 simulated, 993 phantom, and 1983 ex vivo channel data frames (described in Section III-A) were each laterally upsampled by a factor of 4 to form a resized channel data frame of dimensions 256 × 926 pixels. This lateral upsampling factor was selected to improve network performance. Unlike the object

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

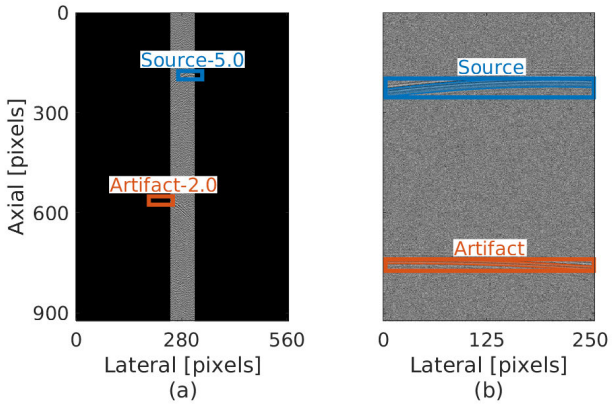


Fig. 5. Examples of (a) bounding box and (b) segmentation mask annotated images of simulated channel data frames containing a photoacoustic point source and reflection artifact.

detection-based point source localization systems previously presented by our group [40], [41], [42], [54], [55], an instance segmentation-based approach to point source localization does not require input images to match the scan-converted FOV of the phased array transducer. Therefore, despite the existence of point source locations outside the lateral limits of the transducer aperture, zero-padding was not applied to the resized channel data frames, as shown in Fig. 5(b).

Using the super-classes for sources and artifacts (Section III-C1), excluding the subdivision based on elevation displacement information, we annotated the resized channel data frames. Our theoretical framework (Section II) was then used to generate ground-truth segmentation masks of source and artifact waveforms, based on the ground-truth locations of sources and artifacts as well as the ground-truth sound speeds. The ground-truth sound speeds were known for simulated data and based on mLOC (Section III-B) for experimental data. The number of elements N_T and pitch p_T of the transducer modeled in Section III-A1 were multiplied and divided, respectively, by the lateral upsampling factor of 4 to match the properties of the resized channel data frames. The pressure wave thickness (illustrated in Fig. 2) was selected as $w_S = 500\ \mu$m.

For each photoacoustic source located at $\vec{x}_S$, our theoretical framework provided the coefficients $z_L(\vec{x}_S, c)$, $x_L(\vec{x}_S, c)$, $b_L(\vec{x}_S, c)$, $a_L(\vec{x}_S, c)$, $z_U(\vec{x}_S, c)$, $x_U(\vec{x}_S, c)$, $b_U(\vec{x}_S, c)$, and $a_U(\vec{x}_S, c)$ of the hyperbolic curves forming the upper and lower bounds of the corresponding waveform in the channel data frame with sound speed c, as described in (20)–(23) and (28)–(31). For each reflector located at $\vec{x}_R$, the parameters $z_L(\vec{x}_R, c; \vec{x}_S)$ and $z_U(\vec{x}_R, c; \vec{x}_S)$ of the hyperbolic curves forming the lower and upper bounds, respectively, of the corresponding reflection artifact waveform were given by

$$z_L(\vec{x}_R, c; \vec{x}_S) = -\frac{w_S f_S}{c} + \frac{\|\vec{x}_R - \vec{x}_S\| f_S}{c} \tag{38}$$

and

$$z_U(\vec{x}_R, c; \vec{x}_S) = \frac{w_S f_S}{c} + \frac{\|\vec{x}_R - \vec{x}_S\| f_S}{c} \tag{39}$$

respectively. The term $(\|\vec{x}_R - \vec{x}_S\| f_S/c)$ in (38) and (39) corresponds to the axial downshift of the reflection artifact waveform by the Euclidean distance between the source and reflector positions (i.e., $\vec{x}_S$ and $\vec{x}_R$, respectively), as presented in the method by Allman et al. [39] (noted in Section III-A1). The remaining parameters $x_L(\vec{x}_R, c)$, $b_L(\vec{x}_R, c)$, $a_L(\vec{x}_R, c)$, $x_U(\vec{x}_R, c)$, $b_U(\vec{x}_R, c)$, and $a_U(\vec{x}_R, c)$ were computed from (21) to (23) and (29) to (31), after replacing $\vec{x}_S$ with $\vec{x}_R$. These parameters were used to construct a segmentation mask for each source and reflection artifact in each image.

For each segmentation mask, a rectangular bounding box was defined with coordinates selected to minimize the area fully enclosing the segmentation mask. Unlike the object detection-based bounding boxes in Section III-C1, these bounding boxes were located within the transducer aperture limits even for sources and artifacts located outside the transducer aperture. Fully annotated segmentation mask images consisted of the source and/or artifact bounding boxes, the associated resized photoacoustic channel data frame combined with the segmentation masks, super-classes (i.e., source or artifact), ground-truth positions, and the speed of sound corresponding to the frame. The totality of fully annotated simulated segmentation images were randomly split into training and test datasets of 16 000 and 4000 images, respectively, while the fully annotated phantom and ex vivo segmentation images were only used for testing (i.e., they were not incorporated during training).

D. System Architectures and Training Procedures

Fig. 6 overviews the two deep learning-based 3-D photoacoustic point source localization systems we developed (i.e., Systems A and B). System A [Fig. 6(a)] implemented an object detection-based approach to 3-D point source localization, similarly to our previously published deep learning-based systems for two dimensions [39], [40], [41], [42], [47], [48], [54], [55], [56]. System A also consisted of a Faster R-CNN network [57] with a ResNet-101 [58] feature extractor. This network was initialized with pretrained weights from the ImageNet dataset [59], then fine-tuned on the simulated object detection training dataset described in Section III-C1 with a batch size of 4 and a base learning rate of 0.001 for 80 epochs using the Detectron2 platform [60]. This fine-tuning was performed using four NVIDIA Quadro RTX 8000 GPUs. The object detection network was trained to simultaneously detect each waveform present in an input zero-padded channel data frame, categorize the waveform into one of the 22 elevation-encoded classes (e.g., "Source-0.0," "Artifact-10.0," etc.), and locate the peak of the detected waveform within the imaging plane. The network outputs for each input image were formatted as a list of object detections consisting of the identified elevation-encoded class, the lateral and axial object location (i.e., bounding box pixel coordinates), and a confidence score between zero and one.

System B [Fig. 6(b)] implemented the two-stage point source localization system described in Section II-E. The first stage (i.e., the instance segmentation algorithm) consisted of a Mask R-CNN network [61] with a ResNet-101 feature extractor. This network was initialized with pretrained weights from

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

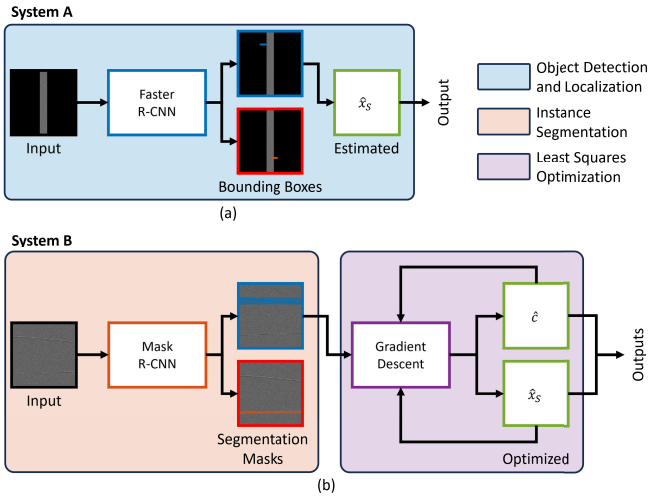


Fig. 6. Block diagrams illustrating (a) System A, a single-stage object detection-based photoacoustic point source localization system extended from previous work [39] on 2-D source localization to achieve 3-D source localization, and (b) System B, a novel instance segmentation-based simultaneous 3-D photoacoustic point source localization and sound speed estimation system.

the ImageNet dataset and fine-tuned on the simulated instance segmentation training dataset described in Section III-C2. This fine-tuning was performed with a batch size of 4 and a base learning rate of 0.001 for 20 epochs using the Detectron2 platform on the GPUs listed above. The network was trained to simultaneously detect acoustic waveforms in the input channel data frame (which was resized as described in Section III-C2), classify the detection as a source or artifact, construct a bounding box around the visible portion of the detected waveform, and generate a segmentation mask limited by the bounding box. The network outputs for each input image were formatted as a list of instance segmentations consisting of the identified super-class (i.e., source or artifact), the bounding box coordinates, the associated segmentation mask $\hat{\Psi}$, and a confidence score between zero and one.

The second stage of System B (i.e., simultaneous source location and sound speed estimation) implemented a least squares optimization algorithm performing gradient descent on the objective function $J(\hat{x}_S, \hat{c}, \hat{L}, \hat{U})$ defined in (35). For each segmentation mask $\hat{\Psi}$ corresponding to an identified source, an initial estimate of the source location was performed assuming an elevation displacement of zero and a sound speed of 1540 m/s. The peaks of the hyperbolic curves $\hat{L}$ and $\hat{U}$ forming the boundaries of the segmented region $\hat{\Psi}$ were identified and averaged to obtain the initial estimates of the lateral and axial displacements of the source from the center of the transducer. Gradient descent was performed using Newton's optimization method [62] for 128 iterations with the output of each iteration provided as an input to the next iteration. In each iteration, the first- and second-order derivatives of the objective function $J(\hat{x}_S, \hat{c}, \hat{L}, \hat{U})$ were computed using the in-built autogradient functionality of the PyTorch library [63]. A scaling factor of 0.1 was applied to each computed increment of the estimated source location, sound speed, and pressure wave thickness w_S. The estimated source elevation position was constrained to be non-negative after each iteration to account for the elevation symmetry of the segmented waveform. The final iteration of the gradient descent algorithm yielded the output source location and sound speed estimates of System B.

TABLE II
CONFIDENCE SCORE THRESHOLDS FOR SOURCES AND ARTIFACTS DETECTED BY SYSTEMS A AND B

Point Source Localization System	Source	Artifact
System A	0.577	0.586
System B	0.947	0.740

E. System Performance Metrics

1) Detection and Segmentation Performance: To quantify the detection performance of the object detection and instance segmentation networks on the simulated test datasets (Section III-C), we used the super-classes (i.e., source or artifact) and bounding box coordinates of network detections and ground truths to classify network detections as true positives, misclassifications, or false positives using the three criteria defined in our previous publication [42], i.e., 1) the confidence score of the given detection was above the super-class-specific confidence score threshold; 2) a ground truth of the same super-class was present in the associated annotations; and 3) the intersect-over-union (IOU) between the bounding boxes of the detection and ground-truth annotations exceeded 0.5. The confidence score threshold was varied from zero to one to generate the receiver operating characteristic (ROC) curves for each super-class and each network in the corresponding simulated test dataset. The area under the curve (AUC) corresponding to each ROC curve determines the quality of the network detections [64], [65]. From the ROC curves, the method presented by Allman et al. [39] determined the optimal confidence score thresholds for each super-class and each network, which is reported in Table II. Network detections above these confidence score thresholds were then retained to compute the recall, precision, and F1 scores [66] as well as the misclassification and missed detection rates [39], [42] for the source and artifact super-classes of simulated data and for the source class of experimental data.

For simulated data, the source detection rates were reported as functions of ground-truth lateral, elevation, and axial positions in the ranges -22.5 to 22.5 mm, -0.625 to 10.625 mm, and 15 to 105 mm, respectively. Each range was divided into nine groups with 5, 1.25, and 10 mm range, respectively (e.g., a group of lateral errors associated with ground-truth lateral positions ≥ -22.5 and < -17.5 mm). For experimental data, the source detection rates were reported as functions of the ground-truth elevation positions in the range -0.5 to 4.5 mm, separated into five groups, each containing a 1 mm range. There was insufficient variation to additionally report results as functions of lateral and axial positions.

To quantify the instance segmentation performance of the instance segmentation network on the simulated test dataset, the IOU between the segmentation masks of true positive detections and corresponding ground-truth annotations was

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

measured. The segmentation IOU of sources was reported separately for ground-truth source lateral, elevation, and axial positions in the groups defined above for the simulated test datasets. In addition, the segmentation IOU of sources was reported separately for ground-truth sound speeds in the range 1427.5–1652.5 m/s, separated into nine groups with a 25 m/s range per group. For each group of simulated data, box-whisker plots displaying the median (horizontal line), the interquartile range (box height), and the range (whisker height) excluding outliers (defined as deviations from the median by >1.5 times the interquartile range, displayed as dots) were employed to represent these characterizations. For experimental data, only the aggregated minimum, median, and maximum segmentation IOU values were reported (given the minimal variation in the lateral, elevation, and axial positions).

2) Point Source Localization Performance: To quantify point source localization performance on the simulated and experimental test datasets, the source location estimates output by each system were first extracted. For System A, the true positive object detection network detections corresponding to the source super-class were retained, and the lateral and axial components of each source location estimate were considered as the center of the corresponding bounding box annotation, assuming a fixed sound speed of 1540 m/s. The elevation source component was obtained from the corresponding class name (e.g., "Source-6.0" corresponded to an elevation displacement of 6 mm). For System B, the estimated point source location was obtained from the second stage output. These estimates were then compared with the corresponding ground-truth source locations, with the elevation components of both the ground truth and estimated source locations constrained to be non-negative to account for the elevation symmetry of the received waveform. The absolute lateral, absolute elevation, absolute axial, and Euclidean distance errors were measured for each retained network detection. To characterize the dependence of the source localization performance on the ground-truth source location and speed of sound in the simulated test datasets, the source location estimates were grouped based on the corresponding ground-truth lateral position, elevation position, axial position, and sound speed groupings described in Section III-E1. For each group of simulated data, box-whisker plots (displaying the details summarized in Section III-E1) were employed to represent these characterizations, in addition to reporting the aggregated mean $\pm$ standard deviation of the absolute lateral, absolute elevation, absolute axial, and Euclidean distance errors for simulated and experimental data.

3) Sound Speed Estimation Performance: To quantify the sound speed estimation performance of System A, System B, and the mLOC-based method described in Section III-B, we first collated the images in the simulated test datasets containing true positive outputs from these systems. We then subtracted either the assumed sound speed of 1540 m/s, the sound speed estimates output by the second stage of System B, or the mLOC-based sound speed estimates from the corresponding ground-truth sound speeds to obtain sound speed estimation errors. These errors were disaggregated into the previously defined groups of ground-truth source positions and sound speeds in simulated data (Section III-E1), in addition to reporting the mean $\pm$ one standard deviation of absolute sound speed estimation outputs of System B and mLOC applied to experimental data.

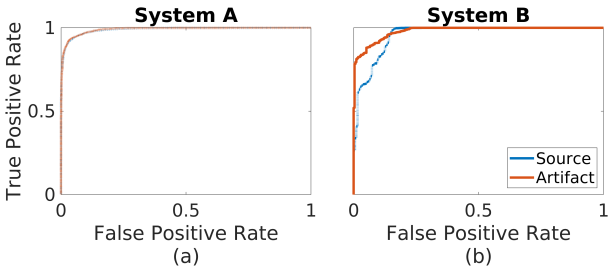


Fig. 7. ROC curves for sources and artifacts achieved with (a) System A and (b) System B applied to simulated data.

IV. Results

A. Detection Performance With Systems A and B

Fig. 7 shows ROC curves for simulated source and artifact detections output by System A and System B in the object detection and instance segmentation test datasets, respectively. These ROC curves reveal that the object detection network forming System A was more robust to false positive errors than the instance segmentation network forming the first stage of System B. This observation is consistent with the increased AUC values (first row of Table III) achieved by System A compared to System B. After confidence score-based filtering (detailed in Section III-E), both Systems A and B achieved similarly high precision scores (second row of Table III) ranging 99.73%–99.97% across sources and artifacts. Overall, the quality of detections was high for Systems A and B for both sources and artifacts with AUC values ranging 0.958–0.990. Table III also reports the recall, $F1$ scores, misclassification rates, and missed detection rates of the object detection network (System A) and the instance segmentation network (System B) in the object detection and instance segmentation test datasets, respectively, for both sources and artifacts. These results highlight the robustness of our two photoacoustic point source detection approaches (i.e., object detection in System A and instance segmentation in System B) to false positives and false negatives associated with detecting photoacoustic point sources and artifacts.

Fig. 8 shows the source detection rates obtained with the simulated test datasets. The source detection rate of System A depended on the ground-truth lateral, axial, and elevation displacements of the source from the center of the transducer, with detection rates ranging 73.61%–92.34%, 84.67%–91.89%, and 78.99%–93.87%, respectively, as shown in Fig. 8(a), (c), and (e), respectively. In comparison, System B achieved a more consistent source detection performance across the lateral, axial, and elevation positions, with detection rates ranging 98.61%–100%, 98.80%–100%, and 99.30%–99.81%, respectively, as shown in Fig. 8(b), (d), and (f), respectively. These results demonstrate the improved ability of the instance segmentation approach utilized in System B to detect photoacoustic point sources (relative to the object detection approach utilized in System A).

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

TABLE III
DETECTION PERFORMANCE ACHIEVED WITH OBJECT DETECTION (SYSTEM A) AND INSTANCE SEGMENTATION (SYSTEM B) NETWORKS IN SIMULATED, PHANTOM, AND EX VIVO TEST DATASETS

Performance Metric	Simulated Sources		Simulated Artifacts		Phantom Sources		*Ex Vivo* Sources	
	System A	System B	System A	System B	System A	System B	System A	System B
Area Under Curve	0.989	0.958	0.990	0.980	-	-	-	-
Precision [%]	99.73	99.97	99.74	99.90	100.00	73.32	100.00	98.39
Recall [%]	88.90	99.67	90.07	98.87	87.01	92.45	96.47	90.17
F1 Score [%]	94.00	99.82	94.66	99.38	93.05	81.78	98.20	94.10
Misclassifications [%]	0.25	0.05	0.20	0.30	0.00	0.00	0.00	0.00
Missed Detections [%]	10.85	0.27	9.73	0.84	12.99	7.55	3.53	9.83

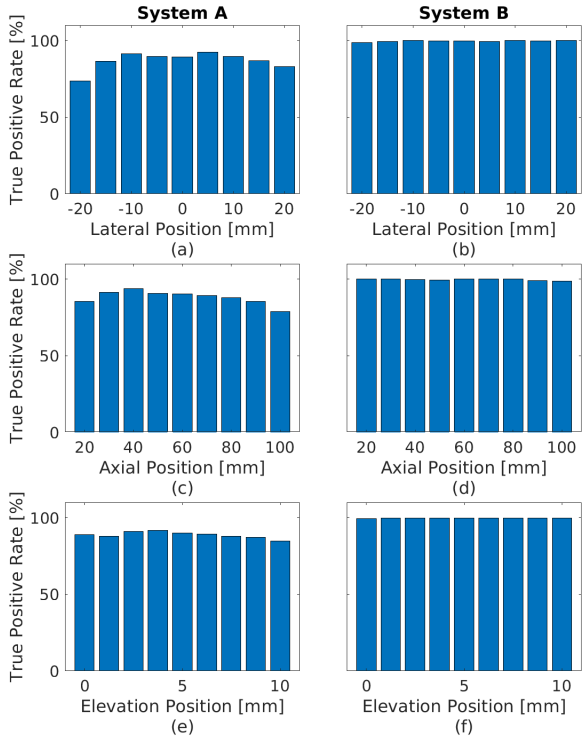


Fig. 8. Summary of detection performance quantified by source detection rates (i.e., recall values) based on bounding boxes associated with System A and System B as functions of ground-truth lateral, axial, and elevation positions in the simulated test datasets.

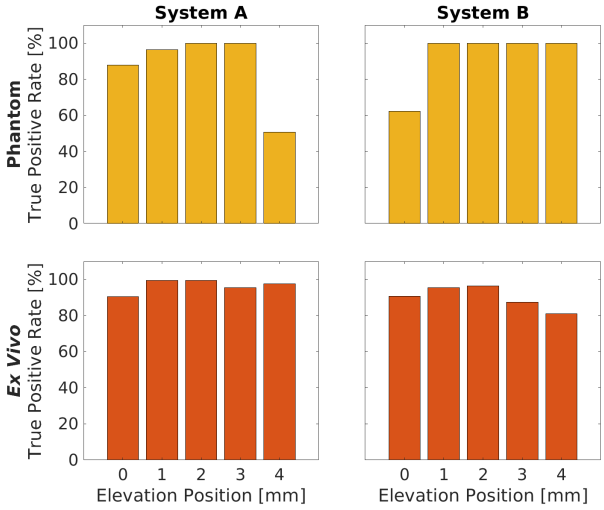


Fig. 9. Source detection rates based on source bounding boxes as functions of the ground-truth elevation positions in the phantom and ex vivo datasets.

Fig. 9 shows the source detection rates of Systems A and B in the phantom and ex vivo datasets as functions of the ground-truth source elevation position relative to the center of the transducer. The ground-truth axial and lateral positions were constant within each dataset, and the ground-truth elevation positions are based on known robot translations, as noted in Section III-A2. The elevation source detection rates ranged 87.94%–100%, with the exception of the phantom results at elevation positions 4 and 0 mm with Systems A and B, respectively (which were 50.75% and 62.31%, respectively). The corresponding recall values, $F1$ scores, misclassification rates, and missed detection rates are reported in the phantom and ex vivo columns of Table III. The generally similar performance (i.e., within <10%) across simulated and ex vivo results demonstrates the ability of our simulation-trained networks to detect photoacoustic targets in real data.

B. Segmentation Performance With System B

Fig. 10 shows the IOU between segmentation masks corresponding to the ground-truth waveforms and outputs of the instance segmentation network (which forms the first stage of System B), as functions of the ground-truth lateral [Fig. 10(a)], elevation [Fig. 10(b)], and axial [Fig. 10(c)] positions and as a function of sound speed [Fig. 10(d)], evaluated for sources in the simulated instance segmentation test dataset. In Fig. 10(a), the median segmentation performance was highest (i.e., 0.977) for sources laterally centered in the transducer FOV and decreased to 0.949 with lateral displacement from the center. In Fig. 10(b), the median segmentation IOU increased from 0.956 to 0.972 as the source elevation position increased from 0 to 10 mm. The segmentation performance remained consistently high across the simulated ranges of source axial positions and sound speeds in Fig. 10(c) and (d), respectively. These results highlight the ability of System B to accurately segment waveforms corresponding to 3-D photoacoustic point sources across a wide range of simulated source positions and sound speeds.

Table IV reports the minimum, median, and maximum IOU between ground-truth segmentation masks and outputs of the instance segmentation network. These values were reduced for the phantom and ex vivo datasets relative to the corresponding

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

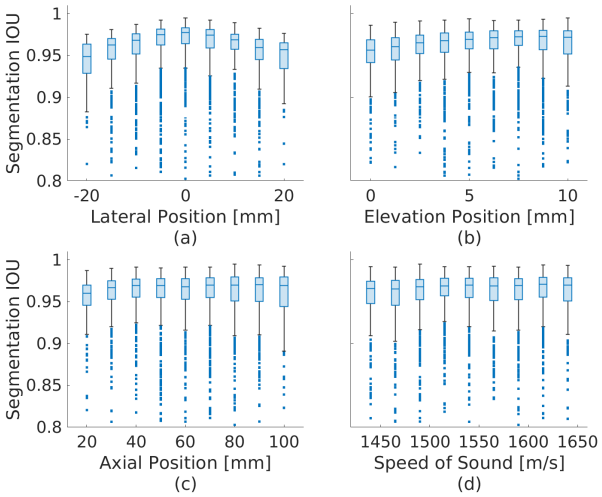


Fig. 10. Summary of segmentation performance based on the IOU between ground-truth segmentation masks and true positive segmentation masks output by the instance segmentation network (i.e., first stage of System B) as functions of the ground-truth (a) lateral position, (b) elevation position, (c) axial position, and (d) sound speed, evaluated for sources in the simulated instance segmentation test dataset.

TABLE IV
MINIMUM, MEDIAN, AND MAXIMUM IOU VALUES BETWEEN GROUND-TRUTH SEGMENTATION MASKS AND TRUE POSITIVE SEGMENTATION MASKS OUTPUT BY INSTANCE SEGMENTATION NETWORK (I.E., FIRST STAGE OF SYSTEM B) IN SIMULATED, PHANTOM, AND EX VIVO DATASETS

	Minimum	Median	Maximum
Simulated	0.608	0.968	0.995
Phantom	0.500	0.626	0.831
Ex Vivo	0.501	0.703	0.905

values in the simulated dataset (indicating reduced segmentation performance). In particular, the median IOU values obtained with the experimental datasets deviated by 27%–35% from that obtained with simulated data.

C. Point Source Localization Performance

The simulation columns of Table V report the mean ± one standard deviation of absolute lateral, elevation, and axial errors, as well as Euclidean distance errors, between the ground-truth locations of point sources defined as true positives and the corresponding source location estimates output by Systems A and B. While System A achieved reduced lateral and elevation errors compared to System B, both systems achieved mean absolute errors <1 mm in these dimensions. However, the Euclidean distance errors are most representative of overall localization performance. System B achieved smaller axial and Euclidean distance errors compared to System A. The improvements in axial and Euclidean localization performance are likely due to the simultaneous estimation of sound speeds by System B, as opposed to the assumed fixed sound speed of 1540 m/s with System A. These results demonstrate the ability of System A (an extension of our previously demonstrated object detection-based approach [39], [42]) and System B (our novel instance segmentation and optimization-based approach) to localize photoacoustic point targets in three dimensions in simulated data across a wide range of source locations and sound speeds.

Fig. 11 shows localization errors for true positives in the simulated test datasets, as functions of the ground-truth source lateral, elevation, and axial positions. Overall, Systems A and B achieved comparable lateral [Fig. 11(a)–(c)] and elevation [Fig. 11(d)–(f)] localization errors across the range of simulated source positions. In Fig. 11(a), System B generally achieved smaller lateral errors than System A for ground-truth source lateral positions ranging -5 to 5 mm. However, unlike System A, both the median and interquartile range of the lateral errors of System B increased with lateral displacement from the center of the transducer. Elevation errors with System B also increased as the elevation displacement increased from 0 to 2.5 mm [Fig. 11(e)]. System B achieved consistently smaller axial errors than System A across the ranges of source lateral, elevation, and axial positions investigated [Fig. 11(g)–(i), respectively]. As shown in Fig. 11(i), axial errors were reduced with System B as axial displacements increased, while the axial errors of System A increased with axial displacements. Overall, these results demonstrate that the localization performance of Systems A and B largely depends on the source position along the corresponding dimension.

Fig. 12 shows localization errors for true positives in the simulated test datasets, as functions of the ground-truth sound speed. System B achieved smaller median lateral errors and larger median elevation errors than System A across the range of simulated sound speeds (Fig. 12(a) and (b), respectively). In addition, the median lateral and elevation errors of Systems A and B remained consistent across the simulated range of sound speeds. In Fig. 12(c), System B (which simultaneously estimated source locations and sound speeds) had smaller axial errors than System A (which assumed a fixed sound speed of 1540 m/s). The axial errors of System A increased as the ground-truth sound speed deviated from 1540 m/s; this relationship was not observed with System B. These results demonstrate the advantage of simultaneously estimating source locations and sound speeds to improve point source localization performance.

Fig. 13 shows example photoacoustic channel data frames in the phantom and ex vivo datasets overlaid with results generated by Systems A and B, along with corresponding ground-truth and source location estimates overlaid on DAS-beamformed images. These DAS images were reconstructed with sound speeds output by System B (i.e., 1485 and 1570 m/s in the phantom and ex vivo data, respectively). The bounding boxes generated by System A are axially centered on the waveform peak, but corresponding source location estimates are axially shifted relative to the ground truth, due to the 1540 m/s sound speed assumed by System A. The bounding box in Fig. 13(a) is laterally displaced from the center of the image, resulting in a 0.9 mm lateral localization error. The distal boundary of the segmentation mask in Fig. 13(c) is distorted from the expected hyperbolic shape, resulting in a 1.9 mm elevation localization error.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

TABLE V
MEAN ± STANDARD DEVIATION OF DISTANCE ERRORS OF SYSTEMS A AND B IN SIMULATED, PHANTOM, AND EX VIVO TEST DATASETS

	Simulation		**Phantom**		***Ex Vivo***	
	System A	System B	System A	System B	System A	System B
Absolute Lateral Error (mm)	0.60 ± 0.36	0.91 ± 1.13	0.72 ± 0.20	0.55 ± 0.40	0.19 ± 0.13	0.20 ± 0.14
Absolute Elevation Error (mm)	0.59 ± 0.93	0.83 ± 0.71	1.83 ± 1.29	1.13 ± 0.86	2.89 ± 1.71	1.23 ± 1.04
Absolute Axial Error (mm)	2.01 ± 1.45	0.20 ± 0.21	0.32 ± 0.09	0.58 ± 1.16	0.51 ± 0.11	0.68 ± 0.18
Euclidean Distance Error (mm)	2.39 ± 1.46	1.46 ± 1.11	2.13 ± 1.07	1.58 ± 1.30	2.99 ± 1.64	1.55 ± 0.86

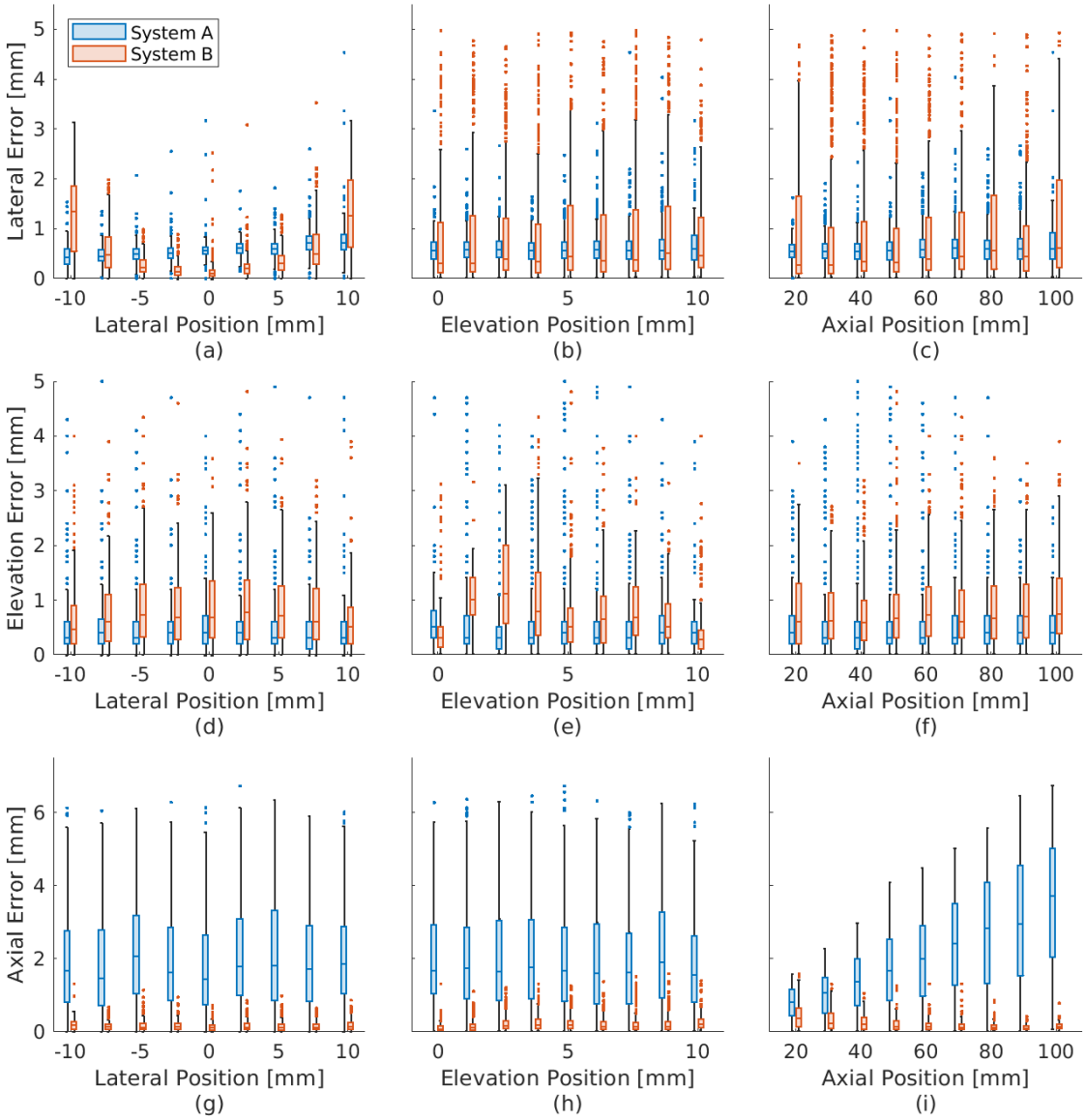


Fig. 11. Summary of localization performance based on absolute (a), (b), and (c) lateral, (d), (e), and (f) elevation, and (g), (h), and (i) axial position errors of true positive sources output by Systems A and B in the simulated test datasets as functions of ground-truth source (a), (d), and (g) lateral, (b), (e), and (h) elevation, and (c), (f), and (i) axial positions.

To summarize the overall localization performance of experimental data, the phantom and ex vivo columns of Table V report the mean ± one standard deviation of the absolute lateral, elevation, and axial distance errors, as well as the Euclidean distance errors, achieved with Systems A and B. The errors of System B were generally lower than or similar to those of System A, with the mean Euclidean distances being most representative of the overall localization errors. These results demonstrate the advantage of the instance segmentation and optimization-based approach utilized in System B, relative to the object detection-based approach utilized in System A, when tasked with accurately

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

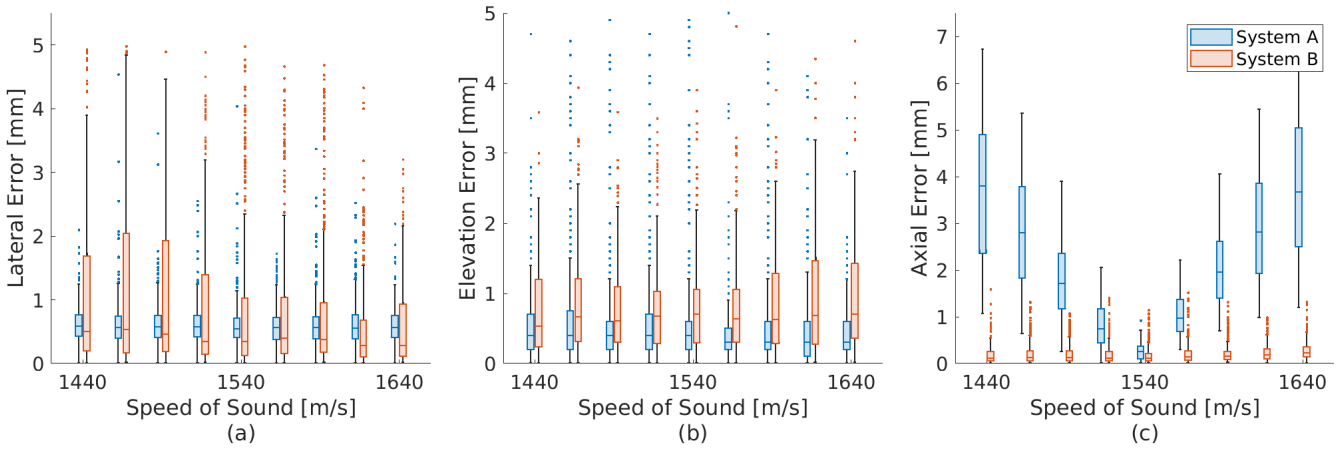


Fig. 12. Summary of localization performance based on absolute (a) lateral, (b) elevation, and (c) axial errors of true positive sources output by Systems A and B in the simulated test datasets as functions of the ground-truth sound speed.

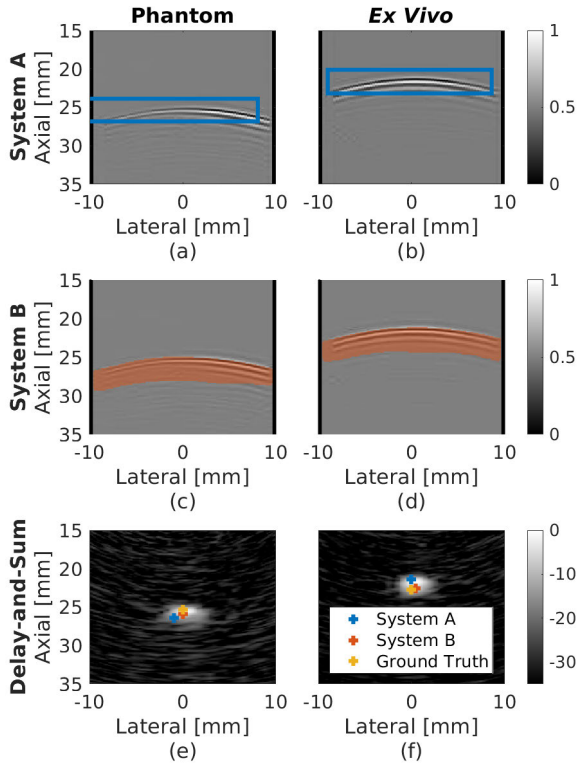


Fig. 13. Example photoacoustic channel data frames of the catheter tip in (a) phantom and (b) ex vivo tissue overlaid with source detections output by System A. Corresponding source segmentations output by System B for the same (c) phantom and (d) ex vivo channel data. Corresponding DAS beamformed images overlaid with estimates output by Systems A and B and compared to the ground-truth source locations for the same (e) phantom and (f) ex vivo data.

localizing photoacoustic point sources in experimental data.

D. Sound Speed Estimation Performance

Fig. 14 shows the mean and standard deviation of the absolute sound speed estimation errors of Systems A and B as functions of the ground-truth source positions and medium sound speeds. System B (which estimated the sound speed) consistently achieved lower sound speed estimation errors compared to System A (which assumed a fixed sound speed of 1540 m/s) across the simulated ranges of lateral, elevation, and axial source positions (i.e., Fig. 14(a)–(c), respectively). The sound speed errors of System A increased as the ground-truth sound speed deviated from the value of 1540 m/s assumed by System A, while System B achieved more consistent sound speed estimation errors across the simulated range of sound speeds [Fig. 14(d)]. Overall, the mean ± one standard deviation of absolute sound speed estimation errors achieved with System B was 19.22 ± 26.26 m/s. These results highlight the ability of System B to estimate the sound speed across wide ranges of source positions and sound speeds.

TABLE VI
COMPARISON OF MEAN ± ONE STANDARD DEVIATION OF SOUND SPEED ESTIMATES OBTAINED WITH SYSTEM B AND MLOC APPLIED TO PHANTOM AND EX VIVO DATASETS

Dataset	Target Depth	Sound Speed [m/s]	
		System B	mLOC
Phantom	25 mm	1505.8 ± 29.3	1524.7 ± 53.3
Ex Vivo	23 mm	1580.2 ± 14.6	1579.3 ± 44.9
	46 mm	1565.2 ± 32.3	1563.6 ± 52.3

Table VI reports the mean ± one standard deviation of the sound speed estimates from System B and mLOC, when applied to experimental photoacoustic sources with no lateral or elevation displacement from the transducer center. To set baseline expectations for these results, Fig. 14(c) and (d) report mLOC errors obtained with a subset of the simulated data containing lateral and elevation displacements <1 mm from the transducer center. In Fig. 14(c), the mLOC-based method achieved reduced sound speed errors at shallower target depths, while the opposite trend was obtained with System B. In Fig. 14(d), larger mLOC errors were achieved as the true sound speed deviated from 1540 m/s, while System B was relatively unaffected by the magnitude of the true sound

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

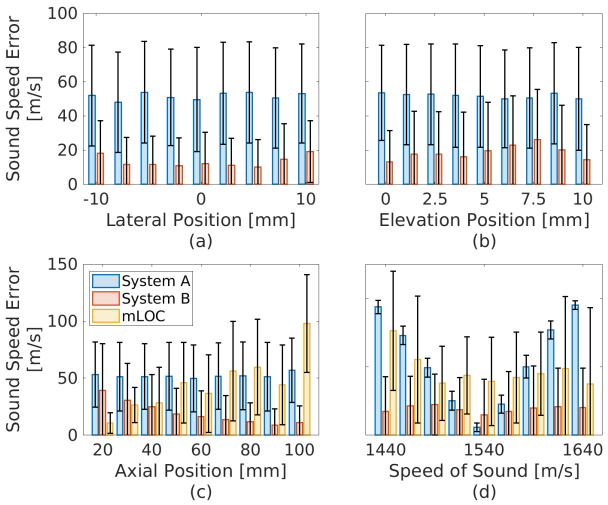


Fig. 14. Absolute sound speed estimation errors achieved with System A (assuming a fixed sound speed of 1540 m/s), System B (after least squares optimization of sound speeds), and mLOC applied to the simulated test datasets, reported as functions of the ground-truth source (a) lateral, (b) elevation, and (c) axial positions and (d) medium sound speeds. The errors obtained with mLOC set expectations for experimental results (reported in Table VI). Error bars show ± one standard deviation.

speed. Despite the observed differences, similar sound speed values (i.e., within one standard deviation) were achieved with both approaches applied to experimental data, as reported in Table VI, which further demonstrates the successful translation of System B to experimental data.

V. Discussion

A. Overview of Contributions

This article presents two novel approaches (i.e., Systems A and B) to train and implement deep learning-based techniques that detect and localize a photoacoustic source in three spatial dimensions, based on the input of a single 2-D channel data frame. First, we extended the two-class classification model proposed by Allman et al. [39] to a 22-class classification model to encode elevation displacement information, resulting in System A [Fig. 6(a)]. Second, we introduced a novel theoretical framework (Section II) relating 3-D point source locations to the corresponding waveform shapes in photoacoustic channel data. Third, we utilized our novel theoretical framework to relate photoacoustic point source positions to channel data waveform shapes, then we employed theory-based least squares optimization to simultaneously estimate source locations and sound speeds, resulting in System B [Fig. 6(b)]. The success of Systems A and B offers a significant improvement over previously published 2-D point source localization systems [39], [40], [42], [43], [47], [48], [54], [55], [56], [67], [68], [69], [70], [71]. This success was characterized with respect to detection (Figs. 7–9 and Table III), segmentation (Fig. 10 and Table IV), localization (Figs. 11 and 12, and Table V), and sound speed estimation (Fig. 14 and Table VI) performance in simulated and experimental data.

The 22-class classification model utilized in System A, enabled differentiation among 11 elevation displacements of point sources and reflection artifacts. This approach is an upgrade from 2-D to 3-D photoacoustic point source localization, utilizing an object detection-based approach to detect, classify, and localize sources in 3-D. As reported in Table III, the $F1$ scores of System A obtained from simulated data (i.e., 94.00% and 94.66% for sources and artifacts, respectively) are comparable to those of our previously published 2-D system obtained with simulated data (98.3% and 90.6% for sources and artifacts, respectively [42]), despite the increased complexity of the 22-class classification problem.

System B is the first known deep learning-based system utilizing the instance segmentation paradigm to detect and localize photoacoustic point sources in three dimensions. The ≥90.17% recall performance of System B on simulated and experimental data (Table III) validates the relationship between the point source location $\vec{x}_S$ and the corresponding waveform shape $\Psi(\vec{x}_S, c)$ derived in Section II-D. Allman et al. [39] previously hypothesized that the object detection-based point source localization systems were learning this underlying relationship to accurately detect and distinguish between sources and artifacts. We introduced our theoretical framework to explicitly characterize this relationship and thereby improve the performance of photoacoustic point source localization systems. This framework enabled an improved formulation of the point source localization problem as an instance segmentation problem compared to previous object detection-based formulations [39], [42].

B. Promise of Instance Segmentation Approach

System B leveraged the instance segmentation-based formulation to achieve improved precision, recall, $F1$ scores, and missed detection rates in simulated data relative to System A (Table III). While System A achieved higher AUC values than System B for both sources and artifacts (indicating that System A is less likely to output false positive detections), System B achieved higher recall and lower missed detection rates compared to System A (indicating that System B is less likely to miss ground-truth sources). These competing differences necessitate a single metric (e.g., $F1$ score) combining metrics such as precision and recall to enable comparisons across systems.

With better $F1$ scores, System B emerges as providing improved detection performance relative to System A when applied to simulated data. The segmentation performance of System B remained consistently high across the simulated ranges of source positions and sound speeds [Fig. 10]. In addition, System B only misclassified two simulated sources as artifacts and only missed eleven sources, out of the 4000 total sources in the simulated dataset (Table III), further demonstrating its ability to learn the relationships among the source axial position, sound speed, and corresponding waveform shape.

The $F1$ scores ranging 81.78%–94.17% achieved by System B with the experimental datasets (Table III) demonstrate the ability of our deep learning-based point source localization system to correctly identify photoacoustic point sources

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

in experimental data. The similar mean Euclidean distance errors of 1.46, 1.58, and 1.55 mm achieved by System B in the simulated, phantom, and ex vivo datasets, respectively (Table V), validate our theoretical framework and instance segmentation-based approach to point source localization. The reduced mean Euclidean distance errors achieved by System B compared to System A in the simulated and experimental datasets (Table V) further demonstrate the advantages of our novel theoretical framework over the extension of our previously presented object detection-based approach [39], [42] to 3-D point source localization. Overall, these results offer a promising new direction for a theory-based instance segmentation approach to photoacoustic point source localization in three dimensions (which are also applicable to two dimensions).

C. Sound Speed and Localization Performance Insights

There are two insights related to the characterization of the localization errors of Systems A and B as functions of the ground-truth sound speed. First, the lateral [Fig. 12(a)] and elevation [Fig. 12(b)] errors of System A were independent of the sound speed, while the axial errors increased as sound speed deviated from 1540 m/s [Fig. 12(c)]. This increase is likely due to the axial position estimates of System A being derived from the axial positions of the corresponding bounding boxes using the fixed sound speed of 1540 m/s as a scaling factor. The second insight is that System B, which estimated sound speed in addition to point source locations, achieved consistent median lateral, elevation, and axial errors across the simulated range of sound speeds (Fig. 12(a)–(c), respectively). These two insights, combined with the overall localization errors of Systems A and B (Table V), as well as the localization errors corresponding to sound speeds within and outside the range 1527.5–1552.5 m/s (Fig. 12), demonstrate the importance of estimating sound speed to improve point source localization performance.

D. Impact of Image Size

Systems A and B both benefited from network performance improvement techniques (i.e., zero-padding and image resizing) that were previously presented [42]. System A was required to extrapolate source positions from partially visible waveforms (similar to 2-D photoacoustic point source localization systems for phased array transducers [40], [42], [54], [55]). Therefore, zero-padding was applied to channel data frames in the object detection dataset to accommodate the placement of bounding boxes outside the visible channel data region, but these zero-padded channel data frames were not required to be resized. While zero-padding was not applied to channel data frames for System B, channel data frames were laterally upsampled by a factor of four to enable the high detection performance of System B. These results indicate the existence of an optimal size of the channel data frames for a given selection of neural network, transducer, imaging depth, and other simulation parameters. If it is necessary to include axial resampling and zero-padding in the presented theoretical framework in future applications, we detail the required modifications in the Appendix.

E. Limitations

One potential limitation of our theoretical framework is the absence of effects related to signal amplitude (e.g., attenuation, sensor directivity, etc.) or waveform shapes (e.g., distortion and attenuation arising from heterogeneities in tissue [72], [73] or bone [37], [74]). However, our previous deep learning-based point source localization systems have performed well by applying histogram matching to experimental data using simulated data as a reference [42]. In addition, previous 2-D object detection deep learning-based systems successfully detected and localized optical fibers [39], needle tips [40], [41], and catheter tips [42] in phantom [39], [40], [41], ex vivo [40], [42], and in vivo environments [42], despite assumptions of a homogeneous medium with a uniform speed of sound. Methods that compensate for heterogeneity-induced waveform distortions [75], [76], [77], [78] could potentially be incorporated, if necessary.

Considering memory limitations, the increase in GPU memory requirements for the 3-D photoacoustic simulations (relative to the 2-D simulations performed in our previous publication [42]) necessitated the selection of the continuous model of the transducer (Section III-A1). This necessity conflicts with the previous recommendation by Allman et al. [48] to use a discrete model for improved network performance in experimental data, suggesting room for additional improvements with more memory. The associated increased memory requirement also resulted in a reduced range of simulated lateral positions in the object detection and instance segmentation datasets (Table I) compared to our previous work [42]. Although the lateral dimension was reduced due to GPU memory limitations, sources both within and outside the lateral aperture limits of the transducer were included in the simulation, which is necessary to enable deep learning-based point source localization systems to detect sources outside the lateral limits of the transducer [54], [55]. Therefore, it is promising that despite the GPU memory limitations, we developed two 3-D point source localization systems (i.e., Systems A and B) that can detect point sources outside the lateral limits of the transducer.

F. Potential Future Applications

The proposed object detection-based and instance segmentation-based systems for photoacoustic point source localization have four potential applications. First, these methods may be integrated with our previously presented deep learning-based photoacoustic visual servoing systems [40], [41] to autonomously track surgical tool tips such as needle tips and catheter tips during interventional procedures such as percutaneous liver biopsies and cardiac catheterizations, respectively. Due to the elevation symmetry of the received waveform and the non-negative elevation displacements output by Systems A and B (Section III), additional logic is required to track surgical tool tips with negative displacements relative to the transducer center. Second, the provided point source locations may be overlaid on ultrasound images that offer real-time visualization of the anatomical details surrounding tool tips [39]. Third, the sound speed estimates

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

from System B may be provided in real time to assess tissue properties, as well as to assist with real-time ultrasound and/or photoacoustic image formation. Finally, the associated techniques may be extended to other applications of computer vision [79] and deep learning in photoacoustics [68] and biomedical optics [80], [81], including the potential to disambiguate tool tips from nearby chromophores with novel multispectral approaches [82], [83], [84], [85], [86].

VI. Conclusion

This work is the first to present two deep learning-based approaches to detect and localize photoacoustic point sources with 3-D displacements relative to an ultrasound transducer in channel data, with potential applicability to numerous surgical and interventional procedures. We successfully trained an object detection-based approach to detect and localize photoacoustic point sources using an elevation-encoded classification model (System A). We also derived a theoretical framework relating the 3-D point source location and speed of sound to the shape of the waveform in the corresponding channel data frame, then trained an instance segmentation network to identify and segment waveforms corresponding to photoacoustic point sources in resized channel data frames, and estimate the corresponding point source locations from the segmented waveform shapes (System B). We characterized the detection, localization, and sound speed estimation performance of this network after theory-based optimizations, validating our theoretical framework. We then demonstrated the improvement in localization performance with simultaneous sound speed estimates, demonstrating the importance of accurate sound speed information to the task of point source localization. The two systems presented in this article have the potential to localize and track needle tips, catheter tips, and other surgical tool tips in numerous surgical and interventional procedures, with System B being the recommended approach going forward considering its overall performance.

Appendix
Incorporating Network Performance Improvement Techniques Into Theoretical Framework

Our theoretical framework may be adapted to accommodate the network performance improvement techniques of image resizing and zero-padding presented in our previous work [42]. In particular, as described in Section III-C2, the lateral resampling operation required corresponding modifications of the transducer parameters N_T and p_T. In addition, if the channel data frames were required to be resampled along the axial dimension by a factor z_{new}, the corresponding transducer sampling frequency f_S' could be obtained from the original sampling frequency f_S as

$$f_S' = f_S z_{\text{new}} \tag{40}$$

to obtain the corrected waveform shapes in the resampled frames. Finally, while the zero-padding operation was deemed inapplicable to the instance segmentation datasets in Section III-C2, performing this operation would require corrected parameters $x_L'(\vec{x}_S, c)$ and $x_U'(\vec{x}_S, c)$ given by

$$x_L'(\vec{x}_S, c) = x_L(\vec{x}_S, c) + (n_Z p_T) \tag{41}$$

and

$$x_U'(\vec{x}_S, c) = x_U(\vec{x}_S, c) + (n_Z p_T) \tag{42}$$

respectively, where n_Z is the number of columns added to the left side of the channel data frame during the zero-padding process.

References

[1] B. Al Knawy and M. Shiffman, "Percutaneous liver biopsy in clinical practice," *Liver Int.*, vol. 27, no. 9, pp. 1166–1173, Nov. 2007.

[2] H. Calkins et al., "2012 HRS/EHRA/ECAS expert consensus statement on catheter and surgical ablation of atrial fibrillation: Recommendations for patient selection, procedural techniques, patient management and follow-up, definitions, endpoints, and research trial design: A report of the heart rhythm society (HRS) task force on catheter and surgical ablation of atrial fibrillation," *Europace*, vol. 14, no. 4, pp. 528–606, 2012.

[3] M. Chan and V. J. Navarro, *Percutaneous Liver Biopsy*. Treasure Island, FL, USA: StatPearls Publishing, 2020.

[4] N. E. Rich et al., "Racial and ethnic disparities in nonalcoholic fatty liver disease prevalence, severity, and outcomes in the United States: A systematic review and meta-analysis," *Clin. Gastroenterol. Hepatol.*, vol. 16, no. 2, pp. 198–210, Feb. 2018.

[5] S. M. Hosseini et al., "Catheter ablation for cardiac arrhythmias: Utilization and in-hospital complications, 2000 to 2013," *JACC, Clin. Electrophysiol.*, vol. 3, no. 11, pp. 1240–1248, 2000.

[6] V. Filingeri, D. Sforza, and G. Tisone, "Complications and risk factors of a large series of percutaneous liver biopsies in patients with liver transplantation or liver disease," *Eur. Rev. for Med. Pharmacological Sci.*, vol. 19, no. 9, p. 1621, Jan. 2015.

[7] L. F. Whitmire et al., "Imaging guided percutaneous hepatic biopsy: Diagnostic accuracy and safety," *J. Clin. Gastroentero.*, vol. 7, no. 6, pp. 511–515, Dec. 1985.

[8] L. Thanos, A. Zormpala, G. Papaioannou, K. Malagari, E. Brountzos, and D. Kelekis, "Safety and efficacy of percutaneous CT-guided liver biopsy using an 18-gauge automated needle," *Eur. J. Internal Med.*, vol. 16, no. 8, pp. 571–574, Dec. 2005.

[9] A. Krishnaswamy, E. M. Tuzcu, and S. R. Kapadia, "Three-dimensional computed tomography in the cardiac catheterization laboratory," *Catheterization Cardiovascular Intervent.*, vol. 77, no. 6, pp. 860–865, May 2011.

[10] R. Gupta, C. J. Walsh, I. S. Wang, M. Kachelrieß, J. Kuntz, and S. Bartling, "CT-guided interventions: Current practice and future directions," *Intraoperative Imag. Image-Guided Therapy*, vol. 2013, pp. 173–191, Nov. 2013.

[11] M. Moche et al., "Navigated MRI-guided liver biopsies in a closed-bore scanner: Experience in 52 patients," *Eur. Radiol.*, vol. 26, no. 8, pp. 2462–2470, Aug. 2016.

[12] K. Pushparajah, A. Tzifa, and R. Razavi, "Cardiac MRI catheterization: A 10-year single institution experience and review," *Interventional Cardiol.*, vol. 6, no. 3, pp. 335–346, Jun. 2014.

[13] S. G. Hushek, A. J. Martin, M. Steckner, E. Bosak, J. Debbins, and W. Kucharzyk, "MR systems for MRI-guided interventions," *J. Magn. Reson. Imag.*, vol. 27, no. 2, pp. 253–266, Feb. 2008.

[14] K. Ahrar, "Fluoroscopy-guided biopsy," in *Percutaneous Image-Guided Biopsy*. Cham, Switzerland: Springer, 2014, pp. 65–72.

[15] L. Yatziv, M. Chartouni, S. Datta, and G. Sapiro, "Toward multiple catheters detection in fluoroscopic image guided interventions," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 770–781, Jul. 2012.

[16] K. P. Kim and D. L. Miller, "Minimising radiation exposure to physicians performing fluoroscopically guided cardiac catheterisation procedures: A review," *Radiat. Protection Dosimetry*, vol. 133, no. 4, pp. 227–233, Feb. 2009.

[17] B. D. Lindsay, J. O. Eichung, H. D. Ambos, and M. E. Cain, "Radiation exposure to patients and medical personnel during radiofrequency catheter ablation for supraventricular tachycardia," *Amer. J. Cardiol.*, vol. 70, no. 2, pp. 218–223, Jul. 1992.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

[18] C. M. Stahl, Q. C. Meisinger, M. P. Andre, T. B. Kinney, and I. G. Newton, "Radiation risk to the fluoroscopy operator and staff," *Amer. J. Roentgenol.*, vol. 207, no. 4, pp. 737–744, Oct. 2016.

[19] M. Mahesh, "Fluoroscopy: Patient radiation exposure issues," *RadioGraphics*, vol. 21, no. 4, pp. 1033–1045, Jul. 2001.

[20] L. S. Rosenthal et al., "Acute radiation dermatitis following radiofrequency catheter ablation of atrioventricular nodal reentrant tachycardia," *Pacing Clin. Electrophysiol.*, vol. 20, no. 7, pp. 1834–1839, Jul. 1997.

[21] G. T. Nahass, "Acute radiodermatitis after radiofrequency catheter ablation," *J. Amer. Acad. Dermatol.*, vol. 36, no. 5, pp. 881–884, May 1997.

[22] P. Kovoor, M. Ricciardello, L. Collins, J. B. Uther, and D. L. Ross, "Risk to patients from radiation associated with radiofrequency ablation for supraventricular tachycardia," *Circulation*, vol. 98, no. 15, pp. 1534–1540, Oct. 1998.

[23] K. Perisinakis, J. Damilakis, N. Theocharopoulos, E. Manios, P. Vardas, and N. Gourtsoyiannis, "Accurate assessment of patient effective radiation dose and associated detriment risk from radiofrequency catheter ablation procedures," *Circulation*, vol. 104, no. 1, pp. 58–62, Jul. 2001.

[24] L. S. Rosenthal et al., "Predictors of fluoroscopy time and estimated radiation exposure during radiofrequency catheter ablation procedures," *Amer. J. Cardiol.*, vol. 82, no. 4, pp. 451–458, Aug. 1998.

[25] H. Calkins, L. Niklason, J. Sousa, R. El-Atassi, J. Langberg, and F. Morady, "Radiation exposure during radiofrequency catheter ablation of accessory atrioventricular connections," *Circulation*, vol. 84, no. 6, pp. 2376–2382, Dec. 1991.

[26] V. Perrot, M. Polichetti, F. Varray, and D. Garcia, "So you think you can DAS? A viewpoint on delay-and-sum beamforming," *Ultrasonics*, vol. 111, Mar. 2021, Art. no. 106309.

[27] A. Rodriguez-Molares et al., "The ultrasound toolbox," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[28] Y. Chang et al., "Clinical application of ultrasonography-guided percutaneous liver biopsy and its safety over 18 years," *Clin. Mol. Hepatol.*, vol. 26, no. 3, pp. 318–327, Jul. 2020.

[29] M. Kanj, O. Wazni, and A. Natale, "Pulmonary vein antrum isolation," *Heart Rhythm*, vol. 4, no. 3, pp. S73–S79, Mar. 2007.

[30] M. A. Lediju, M. J. Pihl, J. J. Dahl, and G. E. Trahey, "Quantitative assessment of the magnitude, impact and spatial extent of ultrasonic clutter," *Ultrason. Imag.*, vol. 30, no. 3, pp. 151–168, Jul. 2008.

[31] G. T. Clement and K. Hynynen, "A non-invasive method for focusing ultrasound through the human skull," *Phys. Med. Biol.*, vol. 47, no. 8, pp. 1219–1236, Apr. 2002.

[32] E. A. Gonzalez, A. Jain, and M. A. L. Bell, "Combined ultrasound and photoacoustic image guidance of spinal pedicle cannulation demonstrated with intact ex vivo specimens," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 8, pp. 2479–2489, Aug. 2021.

[33] L. Gesualdo et al., "Percutaneous ultrasound-guided renal biopsy in supine antero-lateral position: A new approach for obese and non-obese patients," *Nephrol. Dialysis Transplantation*, vol. 23, no. 3, pp. 971–976, Oct. 2007.

[34] A. Wiacek and M. A. L. Bell, "Photoacoustic-guided surgery from head to toe," *Biomed. Opt. Exp.*, vol. 12, no. 4, pp. 2079–2117, 2021.

[35] M. Xu and L. V. Wang, "Photoacoustic imaging in biomedicine," *Rev. Sci. Instrum.*, vol. 77, no. 4, Apr. 2006, Art. no. 041101.

[36] J. Su, A. Karpiouk, B. Wang, and S. Emelianov, "Photoacoustic imaging of clinical metal needles in tissue," *J. Biomed. Opt.*, vol. 15, no. 2, 2010, Art. no. 021309.

[37] M. A. Lediju Bell and J. Shubert, "Photoacoustic-based visual servoing of a needle tip," *Sci. Rep.*, vol. 8, no. 1, p. 15519, Oct. 2018.

[38] M. Graham et al., "In vivo demonstration of photoacoustic image guidance and robotic visual servoing for cardiac catheter-based interventions," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1015–1029, Apr. 2020.

[39] 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.

[40] M. R. Gubbi and M. A. L. Bell, "Deep learning-based photoacoustic visual servoing: Using outputs from raw sensor data as inputs to a robot controller," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 14261–14267.

[41] T. R. Folk, M. R. Gubbi, and M. A. L. Bell, "Development of a ROS2-based photoacoustic-robotic visual servoing system," *Proc. SPIE*, vol. 13306, pp. 58–64, Mar. 2025.

[42] M. R. Gubbi, F. Assis, J. Chrispin, and M. A. L. Bell, "Deep learning in vivo catheter tip locations for photoacoustic-guided cardiac interventions," *J. Biomed. Opt.*, vol. 29, no. S1, p. 11505, Nov. 2023.

[43] H. Wang, S. Liu, T. Wang, C. Zhang, T. Feng, and C. Tian, "Three-dimensional interventional photoacoustic imaging for biopsy needle guidance with a linear array transducer," *J. Biophotonics*, vol. 12, no. 12, Dec. 2019, Art. no. 201900212.

[44] G. J. Diebold, T. Sun, and M. I. Khan, "Photoacoustic monopole radiation in one, two, and three dimensions," *Phys. Rev. Lett.*, vol. 67, no. 24, pp. 3384–3387, Dec. 1991.

[45] 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.

[46] A. Cigier, F. Varray, and D. Garcia, "SIMUS: An open-source simulator for medical ultrasound imaging. Part II: Comparison with four simulators," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106774.

[47] D. Allman, A. Reiter, and M. A. L. Bell, "A machine learning method to identify and remove reflection artifacts in photoacoustic channel data," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[48] D. Allman, M. A. L. Bell, and A. Reiter, "Exploring the effects of transducer models when training convolutional neural networks to eliminate reflection artifacts in experimental photoacoustic images," *Proc. SPIE*, vol. 10494, pp. 499–504, Feb. 2018.

[49] P. Podder, T. Z. Khan, M. H. Khan, and M. M. Rahman, "Comparative performance analysis of hamming, Hanning and blackman window," *Int. J. Comput. Appl.*, vol. 96, no. 18, pp. 1–7, Jun. 2014.

[50] C. A. De Moura and C. S. Kubrusly, "The courant-friedrichs-lewy (CFL) condition," *AMC*, vol. 10, no. 12, pp. 45–90, 2013.

[51] J. Zhang, K. Ding, and M. A. L. Bell, "Flexible array curvature and sound speed estimations with a maximum spatial lag-one coherence metric," *Proc. SPIE*, vol. 12379, pp. 309–314, Jan. 2024.

[52] J. Zhang, K. Ding, and M. A. L. Bell, "Impact of photoacoustic source location on flexible array curvature estimation with a maximum lag-one spatial coherence metric," in *Proc. IEEE Ultrason., Ferroelectr., Freq. Control Joint Symp. (UFFC-JS)*, Sep. 2024, pp. 1–4.

[53] M. A. Lediju, G. E. Trahey, B. C. Byram, and J. J. Dahl, "Short-lag spatial coherence of backscattered echoes: Imaging characteristics," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 58, no. 7, pp. 1377–1388, Jul. 2011.

[54] D. Allman, F. Assis, J. Chrispin, and M. A. Lediju Bell, "Deep learning to detect catheter tips in vivo during photoacoustic-guided catheter interventions: Invited presentation," in *Proc. 53rd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2019, pp. 1–3.

[55] D. Allman, M. A. L. Bell, J. Chrispin, and F. R. Assis, "A deep learning-based approach to identify in vivo catheter tips during photoacoustic-guided cardiac interventions," *Proc. SPIE*, vol. 10878, pp. 454–460, Feb. 2019.

[56] D. Allman, F. Assis, J. Chrispin, and M. A. L. Bell, "Deep neural networks to remove photoacoustic reflection artifacts in ex vivo and in vivo tissue," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2018, pp. 1–4.

[57] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[60] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. (2019). *Detectron2*. [Online]. Available: https://github.com/facebookresearch/detectron2

[61] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[62] R. Fletcher, *Practical Methods of Optimization*. Hoboken, NJ, USA: Wiley, 2000.

[63] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 1–11.

[64] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.

[65] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.

[66] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Santiago de Compostela, Spain: Springer, Mar. 2005, pp. 345–359.
[67] T. Tanaka, R. Imai, and H. Takeshima, "Split-based elevational localization of photoacoustic guidewire tip by 1D array probe using spatial impulse response," *Phys. Med. Biol.*, vol. 69, no. 6, Mar. 2024, Art. no. 065013.
[68] K. Johnstonbaugh et al., "A deep learning approach to photoacoustic wavefront localization in deep-tissue medium," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2649–2659, Dec. 2020.
[69] R. de Jong, H. Moradi, V. Vousten, R. Rohling, and S. E. Salcudean, "Multiple photoacoustic sources localization using deep learning," *Proc. SPIE*, vol. 10064, p. 30, Apr. 2024.
[70] V. Vousten, H. Moradi, Z. Wu, E. M. Boctor, and S. E. Salcudean, "Laser diode photoacoustic point source detection: Machine learning-based denoising and reconstruction," *Opt. Exp.*, vol. 31, no. 9, p. 13895, 2023.
[71] A. Yazdani, S. Agrawal, K. Johnstonbaugh, S.-R. Kothapalli, and V. Monga, "Simultaneous denoising and localization network for photoacoustic target localization," *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2367–2379, Sep. 2021.
[72] B. Liang et al., "Impacts of the murine skull on high-frequency transcranial photoacoustic brain imaging," *J. Biophotonics*, vol. 12, no. 7, Jul. 2019, Art. no. 201800466.
[73] S. Na and L. V. Wang, "Photoacoustic computed tomography for functional human brain imaging," *Biomed. Opt. Exp.*, vol. 12, no. 7, pp. 4056–4083, 2021.
[74] E. A. Gonzalez and M. A. L. Bell, "Photoacoustic imaging and characterization of bone in medicine: Overview, applications, and outlook," *Annu. Rev. Biomed. Eng.*, vol. 25, no. 1, pp. 207–232, Jun. 2023.
[75] Y. Shen et al., "Acoustic-feedback wavefront-adapted photoacoustic microscopy," *Optica*, vol. 11, no. 2, p. 214, 2024.
[76] S. Na, X. Yuan, L. Lin, J. Isla, D. Garrett, and L. V. Wang, "Transcranial photoacoustic computed tomography based on a layered back-projection method," *Photoacoustics*, vol. 20, Dec. 2020, Art. no. 100213.
[77] S. Jeon, W. Choi, B. Park, and C. Kim, "A deep learning-based model that reduces speed of sound aberrations for improved in vivo photoacoustic imaging," *IEEE Trans. Image Process.*, vol. 30, pp. 8773–8784, 2021.
[78] H. Jin, S. Liu, R. Zhang, S. Liu, and Y. Zheng, "Frequency domain based virtual detector for heterogeneous media in photoacoustic imaging," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 569–578, 2020.
[79] M. R. Gubbi, E. A. Gonzalez, and M. A. L. Bell, "Theoretical framework to predict generalized contrast-to-noise ratios of photoacoustic images with applications to computer vision," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 69, no. 6, pp. 2098–2114, Jun. 2022.
[80] L. Tian et al., "Deep learning in biomedical optics," *Lasers Surg. Med.*, vol. 53, no. 6, pp. 748–775, May 2021.
[81] G. Volpe et al., "Roadmap on deep learning for microscopy," 2023, *arXiv:2303.03793*.
[82] E. A. Gonzalez, C. A. Graham, and M. A. L. Bell, "Acoustic frequency-based approach for identification of photoacoustic surgical biomarkers," *Frontiers Photon.*, vol. 2, Oct. 2021, Art. no. 716656.
[83] E. A. Gonzalez and M. A. Lediju Bell, "Dual-wavelength photoacoustic atlas method to estimate fractional methylene blue and hemoglobin contents," *J. Biomed. Opt.*, vol. 27, no. 9, Sep. 2022, Art. no. 096002.
[84] A. Wiacek, K. C. Wang, H. Wu, and M. A. L. Bell, "Photoacoustic-guided laparoscopic and open hysterectomy procedures demonstrated with human cadavers," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3279–3292, Dec. 2021.
[85] J. Zhang, J. Arroyo, and M. A. L. Bell, "Multispectral photoacoustic imaging of breast cancer tissue," in *Proc. IEEE Ultrason., Ferroelectr., Freq. Control Joint Symp. (UFFC-JS)*, Sep. 2024, pp. 1–4.
[86] J. Zhang, J. Arroyo, and M. A. L. Bell, "Multispectral photoacoustic imaging of breast cancer tissue with histopathology validation," *Biomed. Opt. Exp.*, vol. 16, no. 3, pp. 995–1005, 2025.

Mardava R. Gubbi (Graduate Student Member, IEEE) received the B.Tech. degree in electrical engineering and the M.Tech. degree with a focus in wireless communications and signal processing from IIT Madras, Chennai, India, in 2015. He is currently pursuing the Ph.D. degree in ECE at Johns Hopkins University, Baltimore, MD, USA.

From 2015 to 2018, he was a System Engineer at Axiom Research Labs Private Ltd., Bengaluru, India, where he implemented real-time computer vision, guidance, navigation, and control algorithms for a lunar lander and rover. His research interests include photoacoustic imaging, robotics, and machine learning with a view to surgical applications.

Muyinatu A. Lediju Bell (Senior Member, IEEE) received the S.B. degree in mechanical engineering (minor in biomedical engineering) from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006, and the Ph.D. degree in biomedical engineering from Duke University, Durham, NC, USA, in 2012. From 2009 to 2010, she conducted research as a Whitaker International Fellow at the Institute of Cancer Research and Royal Marsden Hospital, Sutton, U.K. From 2012 to 2016, she was a Postdoctoral Fellow with the Engineering Research Center for Computer-Integrated Surgical Systems and Technology, Johns Hopkins University, Baltimore, MD, USA.

She is currently the John C. Malone Associate Professor of electrical and computer engineering with the Departments of Biomedical Engineering, Computer Science, and Oncology, Johns Hopkins University, where she founded and directs the Photoacoustic and Ultrasonic Systems Engineering Laboratory. Her research interests include ultrasound and photoacoustic imaging, coherence-based beamforming, deep learning for ultrasound and photoacoustic image formation, image-guided surgery, robotics, and medical device design.

Dr. Bell is a fellow of SPIE, Optica, and American Institute for Medical and Biological Engineering (AIMBE). She has received multiple significant awards and honors, including the NSF Alan T. Waterman Award in 2024, the IEEE Ultrasonics Early Career Investigator Award in 2022, the SPIE Early Career Achievement Award in 2021, the Alfred P. Sloan Research Fellowship in 2019, the NSF CAREER Award in 2018, the NIH Trailblazer Award in 2018, and the MIT Technology Review Innovator Under 35 Award in 2016. She is the Editor-in-Chief of the *Journal of Biomedical Optics*. She has served as an Associate Editor for IEEE TRANSACTIONS ON MEDICAL IMAGING since 2020, an Editorial Advisory Board Member for *GEN Biotechnology* since 2021, and a member of the Technical Program Committee for the IEEE International Ultrasonics Symposium since 2022. She has also served as an Associate Editor from 2018 to 2022 and an Associate Editor-in-Chief from 2019 to 2021 for IEEE TRANSACTIONS ON ULTRASONICS, FERROELECTRICS, AND FREQUENCY CONTROL.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on May 30,2025 at 11:40:58 UTC from IEEE Xplore. Restrictions apply.