In Vivo Demonstration of Deep Learning-Based Photoacoustic Visual Servoing System

Mardava R. Gubbi, *Student Member, IEEE*, Aravindan Kolandaivelu, Nethra Venkatayogi, Jiaxin Zhang, *Student Member, IEEE*, Pankaj Warbal, Gareth C. Keene, Mawia Khairalseed, *Member, IEEE*, Jonathan Chrispin, and Muyinatu A. Lediju Bell, *Senior Member, IEEE*

***Abstract*— *Objective:* To develop the first known deep learning-based photoacoustic visual servoing system utilizing point source localization and hybrid position-force control to track catheter tips in three dimensions in real-time. *Methods:* We integrated either object detection or instance segmentation-based localization with hybrid position-force control to create our novel system. Cardiac catheter tips were then tracked across distances of 40 mm in a plastisol phantom and 25-64 mm in an *in vivo* swine in real-time in nine visual servoing trials total. *Results:* Object detection-based localization identified the cardiac catheter tip in 88.0-91.7% and 66.7-70.4% of phantom and *in vivo* channel data frames, respectively. Instance segmentation detection rates ranged 86.4-100.0% *in vivo*. These catheter tips were tracked with errors as low as 0.5 mm in phantom trials and 0.8 mm in the *in vivo* trials. The mean inference times were ≥145.3 ms and ≥516.3 ms with object detection-based and instance segmentation-based point source localization, respectively. Hybrid position-force control system enabled contact with the imaging surface during ≥99.43% of each visual servoing trial. *Conclusion:* Our novel deep learning-based photoacoustic visual servoing system was successfully demonstrated. Object detection-based localization operated with inference times that are more suitable for real-time implementations while instance segmentation had lower tracking errors. *Significance:* After implementing suggested optimization modifications, our novel system has the potential to track catheter tips, needle tips, and other surgical tool tips in real-time during surgical and interventional procedures.**

***Index Terms*— cardiac catheterization, deep learning, force control, imaging, instance segmentation, object detection, photoacoustics, visual servoing**

I. Introduction

MODERN interventional procedures such as cardiac catheterizations require real-time visual information to successfully guide surgical tool tips toward targets of interest inside the patient. Traditionally, this information is obtained using medical imaging modalities such as fluoroscopy [1]. However, fluoroscopy exposes patients and operators to ionizing radiation [2], increasing risks of cancer [3] and other adverse biological effects [4]. In addition, fluoroscopy does not provide depth information, limiting the ability to localize surgical tool tips in three dimensions with a single fluoroscopy image. Fluoroscopy machines are also large, expensive, and difficult to transport, limiting their ability to improve global access to quality healthcare. Ultrasound imaging overcomes these limitations with its low cost, portability, availability of depth information, and absence of ionizing radiation. However, ultrasound can fail to localize catheter tips in contact with tissue [5], [6]. In addition, ultrasound imaging fails in acoustically challenging environments characterized by significant acoustic clutter [7], sound scattering, and signal attenuation.

Photoacoustic imaging is an emerging imaging modality based on the photoacoustic effect, which enables acoustic propagation from an optical source (i.e., optical transmission to nearby optical absorbers causes local thermal expansion and generation of acoustic wave propagation). This modality offers utility when a region or target of interest has a higher optical absorption than the surrounding tissue (e.g., nerve visualization, blood vessel detection, surgical tool tracking, or cancer screening [8]). Potential applications of photoacoustic image guidance in monitoring treatment progression in minimally invasive interventional procedures include tool tracking in spinal surgeries [9], [10], photoacoustic-guided teleoperative robotic surgeries [11], [12], guidance of minimally invasive neurosurgeries [13]–[15], tumor boundary delineation [16], [17], large vessel tracking during liver procedures [18], and monitoring the proximity of tools to critical areas of interest during hysterectomies [19].

Photoacoustic imaging inherently benefits from reduced acoustic signal attenuation and increased tissue selectivity compared to ultrasound imaging [7], [20]. The reduction in acoustic signal attenuation is attributed to the one-way travel path of acoustic waves from optical sources (which may be located at surgical tool tips [6]) to acoustic receivers, as

This work was supported in part by NIH Trailblazer Award R21-EB025621, NIH R01 EB032358, NSF CAREER Award 1751522, and NSF SCH Award No. IIS-2014088.

Mardava R. Gubbi (e-mail: mgubbi1@jhu.edu), Jiaxin Zhang, Pankaj Warbal, Gareth C. Keene, and Mawia Khairalseed are with the Department of Electrical and Computer Engineering at the Johns Hopkins University, Baltimore, MD, USA.

Aravindan Kolandaivelu and Jonathan Chrispin are with the Division of Cardiology at the Johns Hopkins Medical Institutions, Baltimore, MD, USA.

Nethra Venkatayogi is with the Department of Computer Science at the Johns Hopkins University, Baltimore, MD, USA.

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

opposed to the two-way acoustic travel (from an ultrasound transducer to a tool tip and back) required with a single ultrasound transducer. While acoustic transmitters may be located at tool tips [21], this addition does not allow clear ultrasound signal distinction between the tool tip and surrounding tissue without advanced signal processing. Conversely, the optical attenuation encountered during photoacoustic imaging ensures that surgical tool tips, when coupled with optical fibers, are the brightest signals of interest in amplitude-based photoacoustic images.

Previous demonstrations with an *ex vivo* blood vessel revealed a photoacoustic imaging setup consisting of an optical fiber housed in a catheter tip, with an externally placed ultrasound transducer, can successfully visualize the catheter tip in the presence of surrounding tissue [5]. This ability of photoacoustic imaging was leveraged to design systems enabling the detection, localization, and autonomous tracking of surgical tool tips in phantom, *ex vivo*, and *in vivo* environments [5], [20], [22], [23]. These systems, collectively referred to as photoacoustic visual servoing systems, can perform amplitude-based segmentations of delay-and-sum beamformed images to detect and localize photoacoustic point sources (e.g., from needle and catheter tips), then provide the segmentations to robotic control algorithms to center ultrasound transducers above the identified point sources [5], [20], [22]. Despite the advantages of photoacoustic over ultrasound imaging, amplitude-based photoacoustic visual servoing systems have three limitations. First, these systems are sensitive to reflection artifacts (e.g., from nearby scattering structures such as bone), which limits consistent maintenance of tool tips in the field-of-view (FOV) of the transducer. Second, the lateral localization performance of these amplitude-based systems are limited by lateral resolution in beamformed images, which worsens with increasing target depth. Third, elevation displacement information is limited in individual beamformed images [24], limiting the ability of amplitude-based approaches to estimate three-dimensional surgical tool tip locations.

Deep learning-based approaches to photoacoustic-based surgical tool tip tracking have the potential to overcome the limitations of amplitude-based segmentation techniques. Previous work [25], [26] leveraged the dimensions of optical fibers integrated with surgical tool tips, which are typically smaller than the lateral and axial resolution of ultrasound transducers, enabling tool tips to be modeled as point sources. Using this model, Allman *et al.* [26] demonstrated a deep learning-based approach to distinguish point sources from reflection artifacts directly from raw photoacoustic channel data, highlighting robust lateral localization performance despite the poor lateral resolution with increasing target depth. This approach was then integrated with robotic control systems, resulting in deep learning-based photoacoustic visual servoing systems [27], [28] that were deployed to track needle tips in a plastisol phantom and *ex vivo* chicken breast tissue with 55.3-67.7% mean reductions in needle tip tracking errors relative to that of an amplitude-based image segmentation approach [27]. The images acquired in these environments contained a reflection artifact forming a larger bright region than the needle tip, causing misclassifications and failed detections with the amplitude-based segmentation approach. The deep learning approach was more robust to these misclassification errors compared to the amplitude-based approach, improving the failure rates by 60.6% [27]. While these visual servoing systems correctly identified and tracked needle tips within the ultrasound transducer FOV, localization of out-of-plane targets was not possible. As a result, additional search algorithms were required to find targets moving orthogonal to the imaging plane.

To detect and localize photoacoustic targets in three dimensions (3D) using beamformed images, Wang *et al.* [29] developed a 3D photoacoustic-based needle tip localization system by autonomously scanning the elevation dimension of the ultrasound transducer using a robotic arm. However, this system required 40 frames to generate each 3D photoacoustic image grid. With a 10 Hz laser pulse repetition frequency (PRF), this requirement resulted in each 3D image grid requiring 4 seconds to be generated (i.e., effective frame rate of 0.25 Hz).

To provide a computationally efficient, real-time alternative to 3D point source localization, our group leveraged the previously demonstrated point source model [26]–[28], [30] to counterintuitively introduce two deep learning-based photoacoustic point source localization systems offering 3D location estimates of catheter tips from a single two-dimensional frame of raw photoacoustic channel data [24]. The first system used an object detection-based approach, while the second system employed an instance segmentation-based approach with a theory-based gradient descent algorithm to improve localization performance. Both systems were demonstrated to detect and localize stationary catheter tips within and outside the imaging plane in phantom and *ex vivo* environments, with the instance segmentation-based system achieving mean elevation localization errors of 1.13 and 1.23 mm, respectively, from raw 2D photoacoustic channel data frames. However, these demonstrations were performed offline and did not include real-time robot-assisted tracking of the catheter tips. In addition, the improved performance of the instance segmentation-based system was achieved by increasing the required inference time due to the iterative gradient descent. Furthermore, prior visual servoing systems [5], [20], [27], [28] did not account for uneven surfaces, which caused the ultrasound transducer to lose contact with the skin, resulting in failed visual servoing attempts [5].

In this paper, we demonstrate deep learning-based 3D point source localization approaches to track surgical tool tips in real time during interventional procedures, with four novel contributions. First, we design an instance segmentation-based photoacoustic point source localization system with a time-optimized gradient descent algorithm, named WaveSegNet-1, to improve inference speed compared to our previous work [24], which employed a method that we call WaveSegNet-2 herein. Second, we integrate WaveSegNet-1 with our previous object detection-based point source localization system [24], named DetectionNet herein, adding hybrid position-force control and additional logic to form a novel real-time deep learning-based photoacoustic visual servoing system. Third, we compare the tracking performance of our previous

WaveSegNet-2-based point source localization system [24] with our real-time DetectionNet-based photoacoustic point source localization system (to demonstrate the advantages of instance segmentation over object detection in the context of photoacoustic-based target tracking) and with our real-time WaveSegNet-1-based system (to demonstrate the importance of selecting an appropriate gradient descent algorithm to accurately track catheter tips using instance segmentation on raw photoacoustic channel data frames). Finally, we provide results from the first known *in vivo* deployment of DetectionNet and WaveSegNet-1 in photoacoustic-based visual servoing, with force control to maintain required imaging contact at all times.

The remainder of this article is organized as follows. Section II presents the architecture of our deep learning-based photoacoustic visual servoing system and describes processes to demonstrate and assess the performance of our system during phantom and *in vivo* visual servoing trials. Section III reports the results of the presented methods. Section IV discusses the implications and future potential of our work. Section V presents a summary of our major findings.

II. Materials and Methods

A. Visual Servoing System Overview

Our photoacoustic imaging system consisted of an Opotek Phocus Mobile laser (Carlsbad, California, USA) with a PRF of 10 Hz (750 nm wavelength, mean energy of 2.0 mJ per pulse), connected to a 1 mm core-diameter optical fiber. The other end of the optical fiber was inserted into a 7F outer-diameter 60 cm long non-steerable catheter (Boston Scientific, Marlborough, Massachusetts, USA) to form a fiber-catheter pair with coincident tips [5]. To receive the photoacoustic signals originating from the catheter tip, a Verasonics (Kirkland, Washington, USA) P4-2v ultrasound transducer with 64 elements and a sampling frequency of 11.88 MHz was interfaced to a Verasonics Vantage 128 ultrasound scanner. This transducer was attached to the end effector of a UR5e robotic arm (Universal Robots, Odense, Denmark) via a custom 3D-printed adapter mounted on a Gamma NET-FT force sensor (ATI Industrial Automation, Apex, North Carolina, USA).

The software components of our deep learning-based photoacoustic visual servoing system are summarized in Fig. 1. Each laser pulse triggered the acquisition of a raw radiofrequency photoacoustic channel data frame of dimensions 64×926 pixels, corresponding to the number of transducer elements (64) and the imaging depth (926, based on 120 mm depth, 11.88 MHz sampling frequency, and an assumed sound speed of 1540 m/s). This channel data frame was then input to our real-time deep learning-based photoacoustic point source localization systems, which output the catheter tip position (Section II-B). The generated point source location estimates were the input to a multi-track linear Kalman filter (MTLKF), which determined multiple possible point source location candidates and output the most likely location of the point source (Section II-C).

To maintain contact with the imaging surface, force sensor readings were employed to estimate the contact force along the axial dimension of the transducer [31]. The point source location and contact force estimates were then input to a finite state machine (FSM), which generated robot motion plans to center the transducer above the catheter tip with the desired contact force (Section II-D). The motion plans were then executed by the robot and the cycle shown in Fig. 1 was repeated with the next laser pulse. The software components of the visual servoing system were implemented using the Robot Operating System [32].

B. Photoacoustic Point Source Localization

We developed three deep learning-based systems (i.e., DetectionNet, WaveSegNet-1, and WaveSegNet-2) to identify and localize photoacoustic point sources in raw photoacoustic channel data frames. Similar to previous work [24], these deep learning-based systems used algorithms belonging to the family of region-based convolutional neural networks (R-CNN) implemented in the Detectron2 platform [33]. These networks were pre-trained on the ImageNet dataset [34] and fine-tuned on custom datasets of point sources and reflection artifacts simulated using the k-Wave MATLAB toolbox [35]. Each dataset contained 16,000 channel data frames with network-specific image preprocessing and annotation strategies. Each network was fine-tuned with a batch size of four and a base learning rate of 0.001. The visual servoing system was designed to use either DetectionNet or WaveSegNet-1 in real time, while WaveSegNet-2 was used offline to provide a performance baseline. For each input channel data frame at iteration k, the selected deep learning-based point source localization system output $N(k)$ detections. Each detection consisted of a confidence score ranging 0 to 1 and an estimate of the source location in the transducer frame U, given by

$$ {}^{U}\vec{x}_i(k) = \left[{}^{U}\hat{x}_i(k), {}^{U}\hat{y}_i(k), {}^{U}\hat{z}_i(k)\right]^T, \quad (1) $$

where $k{\geq}0$, $0{\leq}i{<}N(k)$, and ${}^{U}\hat{x}_i(k)$, ${}^{U}\hat{y}_i(k)$, and ${}^{U}\hat{z}_i(k)$ are the lateral, elevation, and axial components, respectively, of source location estimate i at time instant k. Due to the elevation symmetry of the received waveforms, the point source localization systems were unable to distinguish between positive and negative source elevation displacements in the frame U. Therefore, the elevation displacement estimates were constrained to ${}^{U}\hat{y}_i(k) \geq 0$.

DetectionNet utilized an object detection-based approach to identify waveforms in the input channel data frames, categorize the waveforms by type (i.e., source or artifact) and elevation displacement rounded to the nearest millimeter (e.g., “Source-1.0”), and construct bounding boxes centered on the lateral and axial positions of the corresponding source or artifact. DetectionNet consisted of a Faster R-CNN network [36] with a ResNet-101 [37] feature extractor fine-tuned for 80 epochs on a simulated dataset of 16,000 bounding box annotated channel data frames [24]. To enable DetectionNet to generate bounding boxes outside the lateral dimensions of the transducer (i.e., ±9.6 mm), each input channel data frame was zero-padded to lateral and axial dimensions of 566 pixels and 926 pixels, respectively, corresponding to the phased array transducer beamformed data FOV dimensions of 169.7 mm and 120 mm, respectively (i.e., after scan conversion).

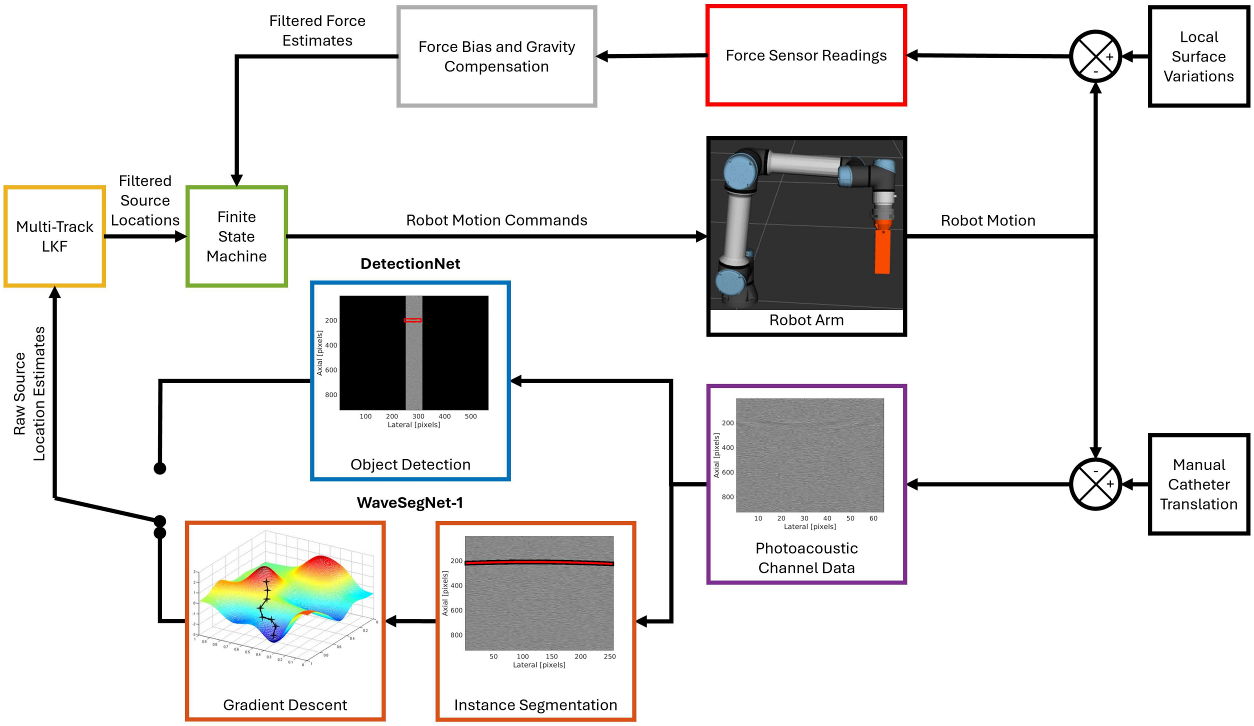

Fig. 1. Summary of deep learning-based photoacoustic visual servoing system. “DetectionNet” is an object detection-based photoacoustic point source localization system, while “WaveSegNet-1” utilizes instance segmentation followed by gradient descent to estimate point source locations. Both systems receive channel data frames as inputs and provide raw three-dimensional point source location estimates as outputs.

TABLE I

INITIAL, MINIMUM, AND MAXIMUM VALUES OF SOURCE AND MEDIUM PROPERTIES ESTIMATED USING GRADIENT DESCENT IN SECOND STAGE OF WAVESEGNET-1

Parameter	Initial	Minimum	Maximum
Lateral Position [mm]	-	-18.8	18.8
Elevation Position [mm]	0	0	10
Axial Position [mm]	-	20	100
Sound Speed [m/s]	1540	1440	1640
Wave Thickness [mm]	0.5	0.3	0.7

WaveSegNet-1 and WaveSegNet-2 utilized an instance segmentation-based approach followed by iterative gradient descent to segment waveforms in the input photoacoustic channel data frames, categorize the waveforms as corresponding to sources or artifacts, and estimate 3D source locations in the transducer frame U [24]. The instance segmentation component forming the first stage of WaveSegNet-1 and WaveSegNet-2 consisted of a Mask R-CNN network [38] with a ResNet-101 feature extractor [37] fine-tuned for 20 epochs on a simulated dataset of 16,000 segmentation mask annotated channel data frames [24]. To improve segmentation performance, the input channel data frames were laterally upsampled to 256 pixels before being input to the network [24]. The Mask R-CNN network output segmentations each consisting of the predicted object type (i.e., sources or artifacts), a confidence score ranging 0 to 1, and a segmentation mask corresponding to the waveform. For each waveform corresponding to the source class, the peak of the segmented waveform was used to obtain the initial estimates of the lateral and axial positions of the source. The initial estimates of the elevation position, sound speed, and wave thickness were set to the values reported in Table I. These initial estimates and the segmented waveform were provided to an iterative gradient descent algorithm, which ran for 128 iterations with the output of each iteration provided as an input to the next. WaveSegNet-1 used the Gauss Newton algorithm [39] which considered first-order gradient terms, while WaveSegNet-2 used Newton’s method [40], which considered first and second-order gradient terms. To compensate for the inaccuracies arising from neglecting second-order gradient terms in WaveSegNet-1, we saturated the outputs of each iteration to the maximum and minimum values provided in Table I. These values form the limits of the simulated parameters used to train the Mask R-CNN network within WaveSegNet-1 (such saturation was not required for WaveSegNet-2 due to the improved gradient descent design compared to WaveSegNet-1). The source location output by the final iteration of the gradient descent algorithm was retained as the estimate ${}^{U}\vec{x}_i(k)$ for the given source waveform. To optimize the achievable inference times of WaveSegNet-1 and WaveSegNet-2, we directly implemented the corresponding gradient descent algorithms in PyTorch [41] rather than using the automatic gradient computation facility provided with the PyTorch library.

To account for the possibility of negative elevation displacements in frame U, the $N(k)$ point source location estimates obtained in Eq. (1) were reflected about the imaging plane of the transducer to obtain $N(k)$ additional estimates given by

$${}^{U}\vec{x}_{i+N(k)}(k) = \left[{}^{U}\hat{x}_i(k), -{}^{U}\hat{y}_i(k), {}^{U}\hat{z}_i(k)\right]^T, \tag{2}$$

where $0 \leq i < N(k)$. The totality of $2N(k)$ point source location estimates were then transformed from the transducer frame U to the robot base frame B as

$$\begin{bmatrix} {}^{B}\vec{x}_i\left(k\right) \\ 1 \end{bmatrix} = {}^{B}T_U\left(k\right) \begin{bmatrix} {}^{U}\vec{x}_i\left(k\right) \\ 1 \end{bmatrix}, \tag{3}$$

where $0 \leq i < 2N(k)$ and ${}^{B}T_U\left(k\right)$ is the homogeneous transform from the transducer frame U to the robot base frame B at time instant k.

C. Multi-Track Linear Kalman Filter

Filtering (i.e., with MTLKF [42], [43]) was implemented to identify the correct point source location from the $2N(k)$ estimates obtained in Section II-B. This MTLKF consisted of $M(k)$ mutually independent linear Kalman filters [44] or tracks. Each track maintained the position and velocity of a source candidate given by

$${}^{B}\vec{s}_i\left(k|k\right) = \begin{bmatrix} {}^{B}\vec{p}_i\left(k|k\right) \\ {}^{B}\vec{v}_i\left(k|k\right) \end{bmatrix}, \tag{4}$$

where $0 \leq i < M(k)$, ${}^{B}\vec{s}_i\left(k|k\right)$ is the state of track i at time instant k, and ${}^{B}\vec{p}_i\left(k|k\right)$ and ${}^{B}\vec{v}_i\left(k|k\right)$ are the updated source position and velocity estimates, respectively, of source candidate i at time instant k in frame B. Each track also maintained the state covariance matrix $P_i\left(k|k\right)$. At the start of time instant $k{+}1$, each track first predicted the updated state ${}^{B}\vec{s}_i\left(k+1|k\right)$, given by

$${}^{B}\vec{s}_i\left(k+1|k\right) = A\, {}^{B}\vec{s}_i\left(k|k\right), \tag{5}$$

where A is the transition matrix, given by

$$A = \begin{bmatrix} I_3 & \Delta t I_3 \\ 0_3 & I_3 \end{bmatrix}. \tag{6}$$

Here, I_3 is an identity matrix of size three, 0_3 is a 3×3 matrix of zeroes, and Δt is the duration of time between time instants k and $k+1$. Each track also estimated the updated state covariance matrix $P_i\left(k+1|k\right)$, the measurement prediction $\vec{z}_i\left(k+1|k\right)$, and the measurement prediction covariance $S_i\left(k+1\right)$ given by

$$P_i\left(k+1|k\right) = \left(A\left[P_i\left(k|k\right)\right]A^T\right) + Q, \tag{7}$$

$$\vec{z}_i\left(k+1|k\right) = H\, {}^{B}\vec{s}_i\left(k+1|k\right), \tag{8}$$

and

$$S_i\left(k+1\right) = \left(H\left[P_i\left(k+1|k\right)\right]H^T\right) + R, \tag{9}$$

respectively, where Q is the state transition noise covariance, R is the measurement noise covariance, and H is the observation matrix given by

$$H = \begin{bmatrix} I_3 & 0_3 \end{bmatrix}. \tag{10}$$

Each track, i, was then associated with point source location estimate, j, obtained in Section II-B satisfying the conditions

$$j = \arg \min_{\substack{0 \leq l < 2N(k+1), \\ d_{il}(k+1) < 11.4}} d_{il}\left(k+1\right), \tag{11}$$

where $d_{il}\left(k+1\right)$ is the measurement prediction distance given by

$$d_{il}\left(k+1\right) = \left[\nu_{il}\left(k+1|k\right)\right]^T \left[S_i\left(k+1\right)\right]^{-1} \nu_{il}\left(k+1|k\right), \tag{12}$$

and

$$\nu_{il}\left(k+1|k\right) = {}^{B}\vec{x}_l\left(k+1\right) - \vec{z}_i\left(k+1|k\right). \tag{13}$$

The threshold of 11.4 corresponded to a 99% likelihood that the source location estimate ${}^{U}\vec{x}_l\left(k+1\right)$ could be obtained from a point source located at ${}^{B}\vec{p}_i\left(k+1|k\right)$. Multiple tracks associated with the same measurement were merged. Each track associated with a measurement was then updated to obtain

$$\begin{aligned} {}^{B}\vec{s}_i\left(k+1|k+1\right) = \\ {}^{B}\vec{s}_i\left(k+1|k\right) + W_i\left(k+1\right)\nu_{ij}\left(k+1|k\right), \end{aligned} \tag{14}$$

and

$$\begin{aligned} P_i\left(k+1|k+1\right) = P_i\left(k+1|k\right) \\ - W_i\left(k+1\right)S_i\left(k+1\right)\left[W_i\left(k+1\right)\right]^T, \end{aligned} \tag{15}$$

where

$$W_i\left(k+1\right) = P_i\left(k+1|k\right)H^T\left[S_i\left(k+1\right)\right]^{-1}. \tag{16}$$

Tracks not associated with a measurement for three consecutive time instants were deleted. The remaining unassociated tracks were then updated as

$${}^{B}\vec{s}_i\left(k+1|k+1\right) = {}^{B}\vec{s}_i\left(k+1|k\right), \tag{17}$$

and

$$P_i\left(k+1|k+1\right) = P_i\left(k+1|k\right). \tag{18}$$

Finally, each source location estimate without an associated track was used to generate a new track.

If the MTLKF contained at least one track at the end of time instant $k+1$, then the MTLKF output ${}^{B}\vec{x}_v\left(k+1\right)$ at time instant $k+1$ was computed as

$$\begin{bmatrix} {}^{U}\vec{x}_v\left(k+1\right) \\ 1 \end{bmatrix} = {}^{U}T_B\left(k+1\right) \begin{bmatrix} {}^{B}\vec{x}_m\left(k+1|k+1\right) \\ 1 \end{bmatrix}, \tag{19}$$

where

$${}^{U}\vec{x}_v\left(k+1\right) = \left[{}^{U}\hat{x}_v\left(k+1\right), {}^{U}\hat{y}_v\left(k+1\right), {}^{U}\hat{z}_v\left(k+1\right)\right]^T, \tag{20}$$

${}^{U}T_B\left(k+1\right)$ is the homogeneous transform from frame B to frame U at time instant $k{+}1$, and m is the index of the longest continuously running track.

D. Finite State Machine for Robotic Control

Table II describes the six states forming the FSM used in our visual servoing system. The FSM prioritized maintaining contact with the imaging surface, measured by the axial component ${}^{U}F_z$ of the estimated force in frame U. If the value of ${}^{U}F_z$ reduced below 0.5 N, the FSM entered the No Contact state and the transducer was translated vertically downward toward the imaging surface. In the remaining states listed in Table II, the robot translated the transducer along the axial

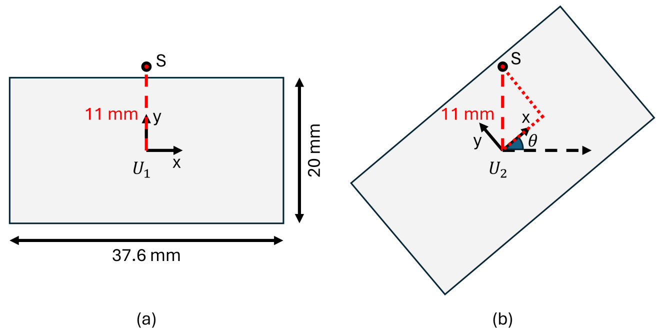

Fig. 2. Rectangular ranges of lateral and elevation positions of point sources detectable using a transducer (a) centered on frame $\boldsymbol{U_1}$ and (b) rotated by an angle $\boldsymbol{\theta}$ about the axial dimension to frame $\boldsymbol{U_2}$. This rotation brings a point source at a fixed location $\boldsymbol{S}$ from (a) outside to (b) inside the region of detectable positions.

dimension to maintain the contact force within a deadband ranging 0.5 N to 2.0 N. This deadband was implemented to minimize the vertical transducer motion required to meet our goal of maintaining physical contact between the transducer and the imaging surface. With the desired contact force maintained, if the MTLKF did not output a valid point source location estimate ${}^{U}\vec{x}_v\,(k)$ at time instant k, then the FSM entered the Search state. DetectionNet, WaveSegNet-1, and WaveSegNet-2 were able to detect point sources with lateral and elevation positions ranging -18.8 mm to 18.8 mm and -10 mm to 10 mm, respectively, as shown by the gray rectangle in Fig. 2(a). The lateral and elevation displacements of the point source from the center of the transducer were expected to be small during the visual servoing process. The Search state leveraged this expectation, rotating the transducer about the axial dimension (as it was translated to maintain contact with the surface) to bring the point source within the region of detectable positions, as shown in Fig. 2(b).

In the event of a valid source location estimate from the MTLKF, we relied on two separate strategies to center the point source in the lateral and elevation dimensions. The Center Lateral state translated the transducer along the lateral dimension to maintain the point source within 1 mm of the transducer axis. This strategy relied on the low lateral localization errors demonstrated with multiple deep learning-based photoacoustic point source localization systems across phantom, *ex vivo*, and *in vivo* environments [24], [26], [30]. To resolve the elevation symmetry about the transducer imaging plane, the Center Elevation state rotated the transducer about the axial dimension to reduce the elevation displacement of the point source (relative to the elevation center of the transducer), at the cost of increased lateral displacement (relative to the lateral center of the transducer). The resulting lateral displacement was corrected by transitioning back to the Center Lateral state until the lateral displacement was less than 1 mm. This strategy enabled elevation symmetry compensation and minimized the elevation localization errors observed in our previous work [24]. These rotations of the transducer brought the point source within 1 mm of the imaging plane with minimal deviation from the original trajectory. The transducer was not translated in the lateral and elevation dimensions during the Centering state (i.e., there was no motion if the target was not completely at the lateral or elevation center of the image), prioritizing consistent visualization over accurate centering.

E. Visual Servoing Applied to Plastisol Phantom

To characterize the detection, tracking, and contact performance of our visual servoing system, the fiber-catheter pair was inserted into an 83 mm-radius hemispherical phantom at a depth of approximately 30 mm as shown in Fig. 3(a). Two checkpoints separated by a distance of 40 mm were selected along the trajectory of the fiber-catheter pair within the phantom. These checkpoints were marked on the catheter at the insertion point into the phantom. With the fiber-catheter pair positioned at the first checkpoint, the transducer was placed in contact with the phantom. To center the transducer above the catheter tip in the lateral and elevation dimensions, the lateral dimension of the transducer was aligned with the catheter, as shown in Fig. 3(a). The robot translated the transducer along its lateral dimension until the peak of the photoacoustic 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 degrees about its axial dimension, then translated the transducer along its lateral dimension until the corresponding waveform was laterally centered in the signal, followed by another rotation by 90 degrees about the axial dimension of the transducer to return to the original alignment between the imaging plane and catheter, with the catheter tip now centered in both the lateral and elevation dimensions of the transducer.

Once the transducer was centered above the catheter tip, the visual servoing system was engaged, and the fiber-catheter pair was manually translated to the second checkpoint. Once the

TABLE II
NAME, ENTRY CONDITIONS, AND TRANSDUCER MOTION ASSOCIATED WITH EACH STATE IN THE FINITE STATE MACHINE

State Name	Contact Force ${}^{U}F_z$	Valid Position	Lateral Position $\lvert{}^{U}\hat{x}_v\,(k)\rvert$	Elevation Position $\lvert{}^{U}\hat{y}_v\,(k)\rvert$	Transducer Motion
Initialize	-	-	-	-	Remain stationary
No Contact	<0.5 N	-	-	-	Move vertically downward
Search	≥ 0.5 N	No	-	-	Rotate about and translate along axial dimension
Center Lateral	≥ 0.5 N	Yes	≥ 1 mm	-	Translate along lateral and axial dimensions
Center Elevation	≥ 0.5 N	Yes	<1 mm	≥ 1 mm	Rotate about and translate along axial dimension
Centered	≥ 0.5 N	Yes	<1 mm	<1 mm	Translate along axial dimension

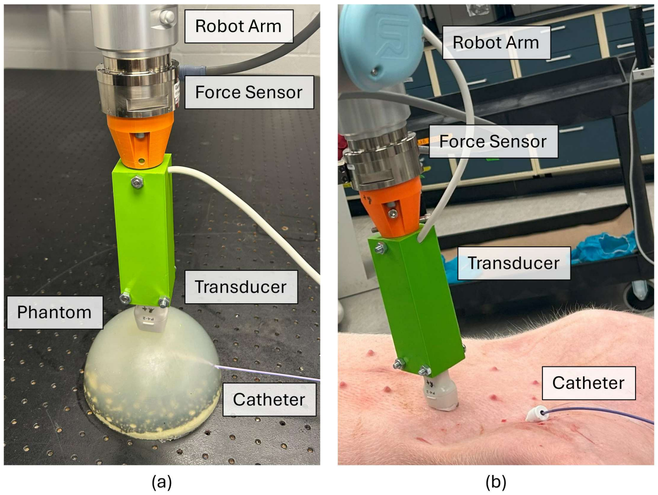

Fig. 3. Photographs of (a) phantom and (b) *in vivo* experimental setups to characterize system performance. See Supplementary Video 1 for motion of transducer relative to catheter *in vivo*.

transducer was autonomously centered above the catheter tip at the second checkpoint, the visual servoing system was disengaged. We recorded raw photoacoustic channel data frames, point source localization system outputs, MTLKF outputs, processed force readings, and robot kinematic information at each instant of time during the visual servoing trial. We conducted a total of five visual servoing trials in the phantom with DetectionNet as the selected real-time point source localization system, as noted in Table III. As WaveSegNet-1 was unable to detect the catheter tip in the phantom across multiple trials, we are unable to report phantom results with this approach.

F. *In Vivo* Demonstration

To demonstrate the viability of our visual servoing system in an *in vivo* setting, we performed a catheterization procedure on an adult female Yorkshire swine weighing 32.2 kg. After the swine was fully anesthetized with isoflurane, a 9F vascular sheath was placed in the right femoral vein. The fiber-catheter pair was inserted into this sheath and advanced into the IVC. Two checkpoints were selected within the IVC using a General Electric (Boston, Massachusetts, USA) OEC 9800 C-arm fluoroscopy system. These two checkpoints were marked in two places: (1) on the skin of the swine and (2) at the insertion point of catheter in the vascular sheath. The catheter tip was manually translated to the first checkpoint within the IVC (using the mark on the catheter). The transducer was placed in contact with the abdominal surface and centered on the first skin checkpoint, with the imaging plane aligned with the intended trajectory of the catheter tip. Real-time ultrasound imaging was used to confirm that the transducer was centered in the elevation dimension above the catheter tip, with robotic translations performed to maximize the amplitude of the catheter tip signal in the ultrasound images.

The visual servoing system was engaged and the catheter was manually translated to the second checkpoint in an approximately linear path at depths ranging approximately 63 mm to 95 mm from the skin surface. The robot was allowed to autonomously move the transducer to follow the motion of the catheter tip. Once the transducer was centered above the catheter tip at the second checkpoint, the visual servoing system was disengaged. We recorded raw photoacoustic channel data frames, real-time point source localization system outputs, MTLKF outputs, processed force readings, and robot kinematic information at each instant of time during the visual servoing trial. We conducted a total of four visual servoing trials with the catheter tip manually translated within the IVC with either DetectionNet or WaveSegNet-1 as the selected real-time point source localization system. The total travel distance (i.e., distance between the catheter checkpoints) per trial per point source localization system are listed in Table III. This study was approved by the Johns Hopkins University Animal Care and Use Committee.

TABLE III
REAL-TIME POINT SOURCE LOCALIZATION SYSTEMS AND CORRESPONDING GROUND TRUTH DISTANCES TRAVELED BY CATHETER TIP DURING EACH PHANTOM OR *in vivo* VISUAL SERVOING TRIAL

Trial Number	Real-Time System	Distance [mm]
Phantom Trial 1	DetectionNet	40
Phantom Trial 2	DetectionNet	40
Phantom Trial 3	DetectionNet	40
Phantom Trial 4	DetectionNet	40
Phantom Trial 5	DetectionNet	40
In Vivo Trial 1	DetectionNet	25
In Vivo Trial 2	DetectionNet	64
In Vivo Trial 3	WaveSegNet-1	38
In Vivo Trial 4	WaveSegNet-1	38

G. Performance Characterization and Comparison with WaveSegNet-2

To compare the performance of the real-time point source localization systems DetectionNet and WaveSegNet-1 (as noted in Table III) with the more computationally expensive WaveSegNet-2, each channel data frame acquired during the real-time visual servoing trials (Sections II-E and II-F) was processed offline using WaveSegNet-2. The outputs of WaveSegNet-2 were synchronized with the real-time photoacoustic point source localization system outputs and robot kinematic information obtained during each visual servoing trial. Detection, localization, tracking, and contact performance were characterized.

To characterize detection performance, each detection or segmentation output was defined as a true positive if: (1) the confidence score of the detection was ≥ 0.5, and (2) the axial position of the detection was within the ranges 25-35 mm and 63-95 mm in the phantom and *in vivo* data, respectively. The confidence score threshold of 0.5 was chosen to minimize the rate of missed detections, assuming that the corresponding increase in false positives will be filtered by the MTLKF. The axial position range for the phantom data accommodated the hemispherical shape of the phantom, while the axial position range for the *in vivo* data is based on the depth information in Section II-F. We did not require additional filtering of true positive detections in the lateral and elevation dimensions because the Faster R-CNN network forming DetectionNet and the Mask R-CNN network forming the first stage of WaveSegNet-1 and WaveSegNet-2 contained layers acting on potential regions of interest (ROI) named ROI-Pool and ROI-Align, respectively, designed to merge candidate detections or segmentations with sufficient overlap [36], [38]. These layers ensured that a given object (i.e., photoacoustic waveform) corresponded to at most one network output. In addition, both networks were fine-tuned with bounding box annotations spanning the width of the channel data in each image [24]. Therefore, source candidates corresponding to the same axial position but with different lateral or elevation position components were combined prior to being output by the networks. Detections which did not satisfy one or both of the criteria above were defined as false positives. The axial component of each initial position estimate was used to evaluate the second criterion to ensure that the detection performance of WaveSegNet-1 and WaveSegNet-2 was assessed independently of gradient descent errors. Based on these true and false positive definitions, the precision, recall, and F1 scores [45] were reported per system per visual servoing trial, using the acquired channel data frames.

To characterize localization performance, the elevation and axial location estimates corresponding to true positive detections in the transducer frame U were assessed. The corresponding lateral position estimates in the frame U were ignored for this assessment, because the motion of the catheter tip was primarily along the lateral transducer dimension. Therefore, variations in lateral position estimates were more reflective of tracking performance (see next paragraph), rather than localization performance. The catheter tip trajectories in the phantom and *in vivo* trials corresponding to the lowest F1 scores achieved by DetectionNet and WaveSegNet-2 were reconstructed using location estimates from the robot base frame B. For ease of plotting, these trajectories were translated to a frame B' parallel to the original frame B. The origin of B' in each visual servoing trial coincided with the transducer center at the first checkpoint of the trial.

To characterize tracking performance, the distance traveled by the robot and the ground truth distance traveled by the catheter were compared per visual servoing trial. The absolute difference between these distances is the catheter tip tracking error. In addition, the number of channel data frames with at least one MTLKF track with a valid point source location candidate were counted, ignoring frames which were not processed by the real-time point source localization systems.

To validate our tracking performance characterizations, we confirmed catheter tip positions with fluoroscopic images before and after each visual servoing trial using the fluoroscopy system noted in Section II-F. In addition, the fluoroscopic image acquired after *In Vivo* Trial 2, corresponding to the largest travel distance of the catheter (Table III), was compared with the corresponding catheter tip location estimates of WaveSegNet-2. To perform this comparison, a subset of the catheter tip estimates from WaveSegNet-2 (i.e., 10% of the total) and the catheter appearance in the fluoroscopy image were used to estimate the transformation between the 3D robot base frame, B, and the 2D fluoroscopy frame, F, using Horn's quaternion-based method [46]. The rotational component of the estimated transform was limited to the axis of B most aligned with the axial dimension of the transducer, considering the single x-ray projection of the anterior-posterior view, resulting in F primarily aligning with the lateral-elevation transducer plane. This transform was then applied to the full set of WaveSegNet-2 outputs for *In Vivo* Trial 2. The root mean square error (RMSE), median error, and range of errors between each output of WaveSegNet-2 transformed to frame F and the corresponding closest point along the catheter in the fluoroscopy image were reported as quantitative performance metrics.

To characterize contact performance, the component of the measured force along the axial dimension of the transducer was determined. In addition, the contact time duration was measured per visual servoing trial (indicated by non-negative contact forces along the axial dimension of the transducer).

III. Results

A. Validation of Catheter Tip as a Point Source

Fig. 4 shows DAS-beamformed simulated (as described in Section II-B) and experimental photoacoustic images to validate the point source model of our visual servoing approach. In Fig. 4(a), the simulated 1 mm-diameter photoacoustic source was located at a depth of 76.2 mm. In Fig. 4(b), the tip of the fiber-catheter pair was inserted in the swine IVC at Checkpoint 1 of *In Vivo* Trial 4. The targets in both images are qualitatively similar, which supports the rationale

TABLE IV

Recall, precision, and F1 scores of real-time (i.e., DetectionNet or WaveSegNet-1, as indicated in Table III) and offline (i.e., WaveSegNet-2) axial point source detection performance achieved during phantom and in vivo visual servoing trials

Trial Number	Real Time			Offline		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Phantom Trial 1	90.5%	60.7%	72.7%	70.0%	69.7%	69.8%
Phantom Trial 2	94.2%	68.5%	79.3%	69.6%	87.9%	77.7%
Phantom Trial 3	89.1%	76.2%	82.1%	78.5%	88.0%	83.0%
Phantom Trial 4	91.7%	69.6%	79.2%	82.0%	83.7%	82.8%
Phantom Trial 5	88.0%	74.9%	80.9%	65.5%	90.0%	75.8%
In Vivo Trial 1	70.4%	96.2%	81.3%	100.0%	100.0%	100.0%
In Vivo Trial 2	66.7%	97.5%	79.2%	86.4%	98.8%	92.2%
In Vivo Trial 3	100.0%	100.0%	100.0%	100.0%	97.7%	98.8%
In Vivo Trial 4	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

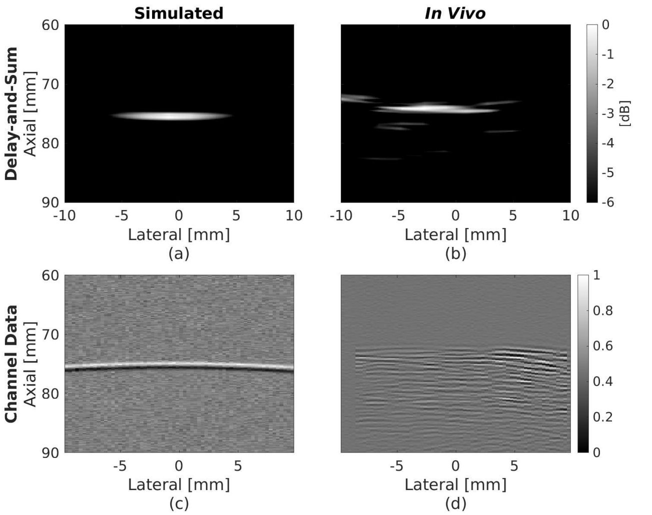

Fig. 4. Delay-and-sum beamformed photoacoustic images of (a) a simulated 1 mm-diameter point source and (b) the tip of the fiber-catheter pair in an *in vivo* swine inferior vena cava, with corresponding raw photoacoustic channel data frames in (c) and (d), respectively.

provided in Section I and in previous work [5], [20], [22] to model the catheter tip as a photoacoustic point source. A similar photoacoustic point source observation was previously demonstrated with a catheter-fiber pair inserted in the plastisol phantom (e.g., see Fig. 13 in [24]). The channel data corresponding to the example simulated and *in vivo* beamformed images herein are presented in Figs. 4(c) and 4(d), respectively, representing example inputs to DetectionNet, WaveSegNet-1, and WaveSegNet-2. See Supplementary Video 1 for example network outputs from channel data, overlaid on beamformed images, from *In Vivo* Trial 4.

B. Detection Performance, Localization Performance, and Efficiency of Point Source Localization Systems

Table IV reports the recall, precision, and F1 scores per network per phantom or *in vivo* visual servoing trial. As DetectionNet, WaveSegNet-1, and WaveSegNet-2 were each fine-tuned with bounding boxes spanning the width of the photoacoustic channel data and multiple network candidates at a given target depth were merged prior to our filtering process, this performance primarily represents performance in the axial dimension. DetectionNet achieved comparable F1 scores in the phantom and *in vivo* trials (ranging 72.7-82.1% and 79.2-81.3%, respectively). These F1 scores corresponded to high recall rates (i.e., ≥88.0%) in the phantom trials and high precision rates (i.e., ≥96.2%) in the *in vivo* trials. In the phantom visual servoing trials, WaveSegNet-2 achieved comparable F1 scores to DetectionNet (ranging 69.8-83.0%). However, WaveSegNet-1 and WaveSegNet-2 outperformed DetectionNet during the *in vivo* trials with F1 scores ranging 92.2-100.0%. These results demonstrate the dependence of the detection performance of our deep learning-based photoacoustic point source localization systems on the imaging environment used to acquire raw photoacoustic channel data.

Fig. 5 shows box-and-whisker plots of the elevation and axial components of point source location estimates from the phantom and *in vivo* visual servoing trials. In the elevation dimension (Figs. 5(a) and 5(b)), DetectionNet outputs median elevation displacement estimates ranging 6-7 mm and 7-10 mm of the catheter tip in the phantom and *in vivo* trials, respectively. WaveSegNet-1 estimated median elevation displacements of 10 mm during *In Vivo* Trials 3 and 4. In comparison, WaveSegNet-2 consistently output reduced elevation displacement estimates compared to DetectionNet and WaveSegNet-1 with median values ranging 0.0-0.1 mm in the phantom and *in vivo* environments. Elevation localization errors can be determined from the distance between the elevation position estimates in Figs. 5(a) and 5(b) and the transducer center at an elevation position of zero. Similarly, axial localization error can be determined from the distance between the axial position estimates in Figs. 5(c) and 5(d) and the axial depths reported in Sections II-E and II-F (i.e., 30 mm for the phantom and 63-95 mm for the *in vivo* trials).

In the axial dimension of the phantom trials (Fig. 5(c)),

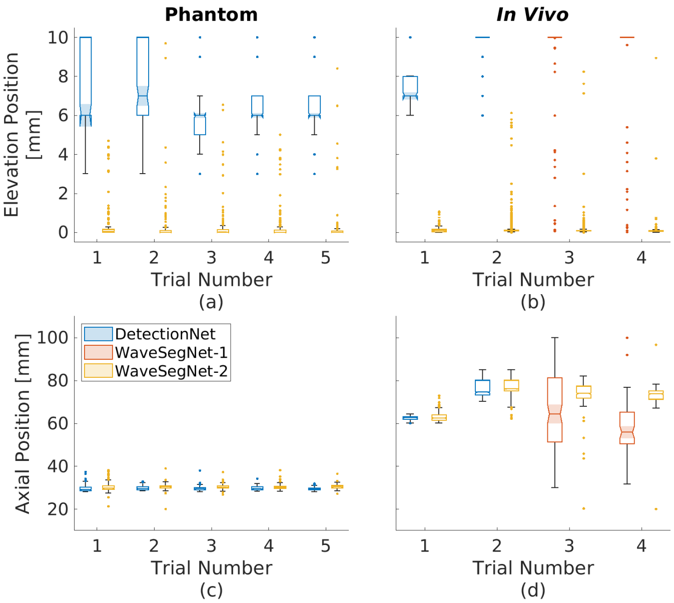

Fig. 5. (a,b) Elevation and (c,d) axial position estimates of the catheter tip output by photoacoustic point source localization systems from (a,c) phantom and (b,d) *in vivo* visual servoing trials. The notches, box heights, whiskers, and dots denote the medians, the interquartile ranges, 1.5 times the interquartile ranges, and outliers, respectively.

TABLE V
RANGES OF VERTICAL (I.E., AXIAL) TRANSDUCER MOTION TO MAINTAIN DESIRED CONTACT FORCE DURING VISUAL SERVOING TRIALS

Trial Number	Vertical Motion Range (mm)
Phantom Trial 1	2.25
Phantom Trial 2	1.70
Phantom Trial 3	1.20
Phantom Trial 4	1.26
Phantom Trial 5	1.23
In Vivo Trial 1	3.69
In Vivo Trial 2	13.19
In Vivo Trial 3	7.48
In Vivo Trial 4	5.53

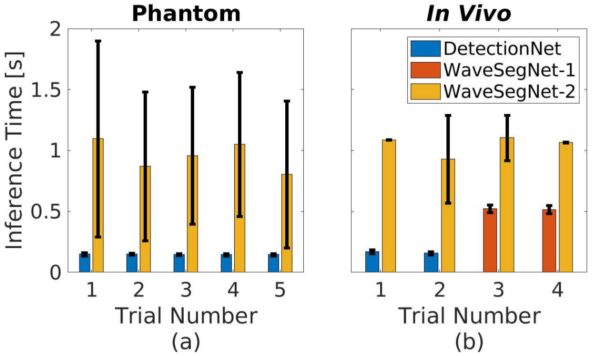

Fig. 6. Inference times achieved with (a) DetectionNet and WaveSegNet-2 in phantom visual servoing trials and (b) DetectionNet, WaveSegNet-1, and WaveSegNet-2 during *in vivo* trials. Error bars show ± one standard deviation.

DetectionNet and WaveSegNet-2 output similar axial position estimates of the catheter tip in the phantom, with median values ranging 29.0-29.6 mm and 29.7-30.4 mm, respectively, and interquartile ranges ranging 0.9-1.8 mm and 1.0-1.8 mm, respectively. In the axial dimension of *In Vivo* Trials 1 and 2 (Fig. 5(d)), DetectionNet and WaveSegNet-2 output comparable axial location estimates with median values ranging 62.7-74.6 mm and 62.5-76.1 mm, respectively (interquartile ranges ranging 1.2-6.8 mm and 2.4-5.0 mm, respectively). The larger interquartile ranges relative to the phantom trials were partially caused by the larger vertical motion of the transducer to maintain the desired contact force, which is reported in Table V. However, the even larger interquartile ranges of in axial position estimates obtained with WaveSegNet-1 during *In Vivo* Trials 3 and 4 (i.e., 30.0 mm and 14.8 mm, respectively) relative to those of WaveSegNet-2 (i.e., 6.5 mm and 3.8 mm, respectively, for the same *in vivo* trials) are not due to differences in transducer motion. Instead, these results demonstrate the improved axial localization accuracy that can be achieved by WaveSegNet-2 relative to WaveSegNet-1.

Fig. 6 shows the mean ± one standard deviation of inference times achieved by the photoacoustic point source localization systems in each of the phantom and *in vivo* visual servoing trials. DetectionNet consistently achieved the lowest mean inference times among the three systems across the phantom and *in vivo* environments, with mean inference times ranging 145.3 ms to 158.0 ms. WaveSegNet-1 achieved mean inference times of 522.7 ms and 516.3 ms in *In Vivo* Trials 3 and 4, respectively. WaveSegNet-2 performed inference slower than both DetectionNet and WaveSegNet-1 with mean inference times ranging 805.0 ms to 1103.8 ms across the phantom and *in vivo* visual servoing trials. In addition, DetectionNet and WaveSegNet-1 achieved comparably lower standard deviations of inference times (ranging 7.3 ms to 14.0 ms and 32.5 ms to 32.9 ms, respectively), relative to that of WaveSegNet-2 (i.e., 562.3 ms to 802.8 ms and 2.7 ms to 360.7 ms during the phantom and *in vivo* trials, respectively). Hence, there are increased computational costs associated with WaveSegNet-2.

C. Lateral Tracking Performance

Fig. 7 shows the reconstructed trajectories of the transducer and catheter tip as estimated by the real-time (i.e., DetectionNet and WaveSegNet-1) and offline (i.e., WaveSegNet-2) photoacoustic point source localization systems in the fixed frame B' defined in Section II-E during Phantom Trial 1, *In Vivo* Trial 2, and *In Vivo* Trial 3, which correspond to the lowest F1 scores achieved by DetectionNet and WaveSegNet-2 (based on Tables III and IV). In Fig. 7(a), the transducer moved in a curved path following the hemispherical surface of the phantom, while the catheter tip moved in a straight line inside the phantom. This linear motion of the catheter was captured by DetectionNet with a small number of outliers, as shown by the blue dots in Fig. 7(a). The catheter trajectory reconstructed using WaveSegNet-2 contained a larger number of outliers from the linear trajectory compared to DetectionNet, as shown by the yellow dots in Fig. 7(a).

In Figs. 7(b) and (c), the height of the transducer increased with the insertion of the catheter tip, following the abdominal surface of the swine. However, the catheter tip continued moving along an approximately linear trajectory within the IVC. DetectionNet produced variations in the axial location estimates of the catheter tip, which affecting the reconstructed trajectory (Fig. 7(b)), while WaveSegNet-1 was unable to capture the linear motion of the catheter (Fig. 7(c)). In comparison, WaveSegNet-2 successfully reconstructed linear trajectories in both cases. These results demonstrate the ability of WaveSegNet-2 to accurately track the trajectory of a catheter tip during the *in vivo* visual servoing trials.

Table VI reports real-time tracking errors, based on the difference between the ground truth start and end positions in the x dimension, which are plotted in Fig. 7, and the tracked start and end positions for each trial. Comparably low tracking errors were achieved with DetectionNet during the phantom and *in vivo* trials (i.e., 0.5-3.6 mm and 2.2-3.2 mm, respectively). Although Fig. 7 shows large deviations in the axial dimension with WaveSegNet-1, and Fig. 5 shows large deviations in the axial and elevation dimensions with WaveSegNet-1, one benefit of WaveSegNet-1 is the reduced lateral tracking errors during the real-time *in vivo* trials in Table VI (i.e., 0.8-1.3 mm with *In Vivo* Trials 3-4), when compared to the tracking errors achieved with DetectionNet

TABLE VI
REAL-TIME LATERAL CATHETER TIP TRACKING ERRORS ACHIEVED WITH DETECTIONNET OR WAVESEGNET-1 DURING PHANTOM AND IN VIVO VISUAL SERVOING TRIALS, AS INDICATED IN TABLE III

Trial Number	Tracking Error (mm)
Phantom Trial 1	0.5
Phantom Trial 2	2.7
Phantom Trial 3	3.6
Phantom Trial 4	3.0
Phantom Trial 5	3.1
In Vivo Trial 1	2.2
In Vivo Trial 2	3.2
In Vivo Trial 3	0.8
In Vivo Trial 4	1.3

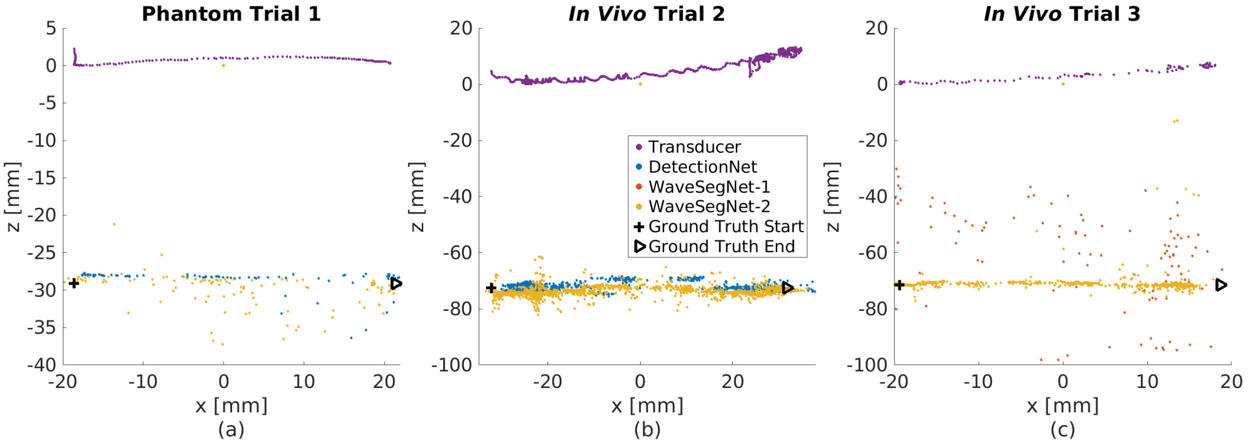

Fig. 7. Transducer and catheter tip position during (a) Phantom Trial 1, (b) *In Vivo* Trial 2, and (c) *In Vivo* Trial 3, as noted in Section II-G. The real-time photoacoustic point source localization systems DetectionNet and WaveSegNet-1 are compared with corresponding offline results obtained with WaveSegNet-2. The $+$ and $\triangleright$ symbols indicate the start and end, respectively, of ground truth travel in the x dimension, based on checkpoints marked on the inserted catheter (described in Section II-F).

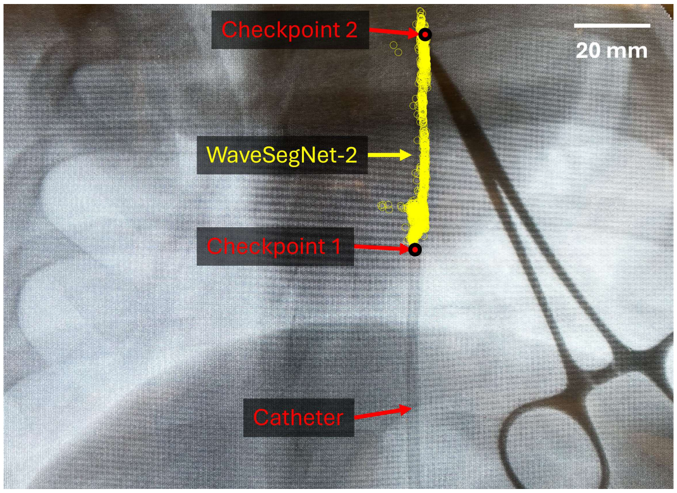

Fig. 8. Fluoroscopic image of catheter acquired after *In Vivo* Trial 2 within inferior vena cava with catheter tip location estimates from WaveSegNet-2 superimposed. Clamp forceps rested on the chest to externally mark the skin with the catheter tip location at Checkpoint 2.

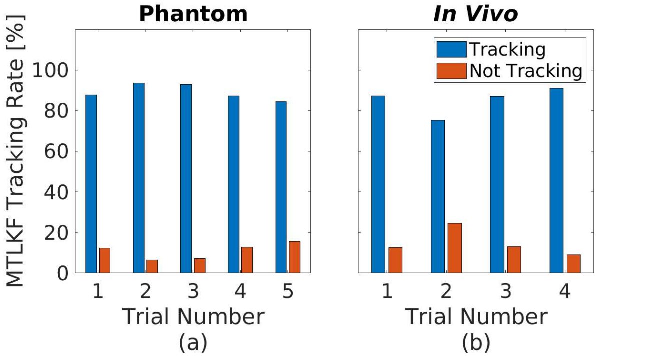

Fig. 9. Tracking success and failure rates of the MTLKF during (a) phantom and (b) *in vivo* visual servoing trials.

(i.e., 2.2-3.2 mm with *In Vivo* Trials 1-2).

Fig. 8 shows the catheter tip location estimates output by WaveSegNet-2 corresponding to *In Vivo* Trial 2, overlaid on the fluoroscopy image of the catheter acquired after *In Vivo* Trial 2, with starting and ending positions marked by Checkpoints 1 and 2, respectively. The RMSE between each output of WaveSegNet-2 in the fluoroscopy reference frame F and the corresponding closest point along the fluoroscopy-based catheter trajectory was 1.1 mm. The median error was 0.5 mm, and the error range was 0-7.0 mm, which are both larger than the 0.1 mm median and 0-6 mm error range between 0 mm elevation location and the datapoints for *In Vivo* Trial 2 in Fig. 5(b), indicating better accuracy with the methods used to obtain the results presented in Figs. 5 and 7.

Fig. 9 shows the percentage of time during which the output state of the MTLKF was tracking or not during each visual servoing trial, which ultimately measures tracking success or failure rates, respectively. In Fig. 9(a), the catheter tip was successfully tracked (i.e., MTLKF maintained at least one track associated with a point source location estimate from within the previous three time instants) for time durations ranging 84.5% to 93.6% of each total phantom visual servoing trial length. The corresponding tracking rates were ≥75.4% during the *in vivo* trials (Fig. 9(b)). These results demonstrate the ability of the MTLKF to utilize the outputs of the real-time deep learning-based point source localization systems to consistently identify the position of the catheter tip.

D. Contact Performance

Fig. 10 shows box-and-whisker plots of the contact force during the phantom and *in vivo* trials. In Fig. 10(a), our hybrid position-force control-based visual servoing system maintained contact 100.0% of the time during each phantom trial (median and interquartile ranges of contact forces ranging 1.38 N to 1.51 N and 0.27 N to 0.32 N, respectively). In Fig. 10(b), our system maintained contact with the abdomen of the swine (i.e., positive force readings) between 99.43% and 100.0% of the total time duration of each trial (median and interquartile ranges of contact forces ranging 1.29 N to

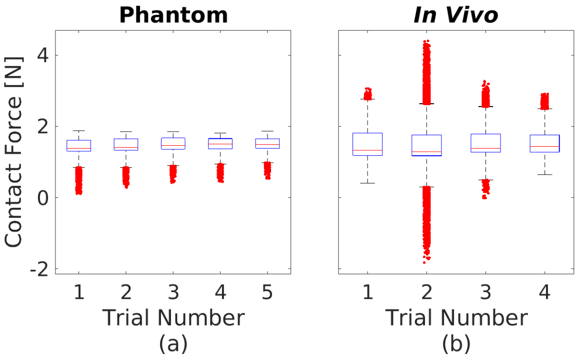

Fig. 10. Contact forces during (a) phantom and (b) *in vivo* visual servoing trials. The horizontal red lines, box heights, whiskers, and red dots denote the medians, the interquartile ranges, 1.5 times the interquartile ranges, and outliers, respectively.

1.43 N and 0.48 N to 0.63 N, respectively). The contact loss (which only occurred during *In Vivo* Trial 2) and the increased interquartile ranges of forces in the *in vivo* trials compared to the phantom trials were caused by respiratory motion, resulting in the robot moving the transducer to reacquire contact and the initiation of the search state of the FSM if catheter tip motion resided outside the transducer FOV (Table II). Our force control implementation successfully enabled the transducer to mostly remain in contact to provide necessary real-time photoacoustic data of the catheter tip.

IV. Discussion

This paper is the first to present a real-time deep learning-based photoacoustic visual servoing system utilizing both object detection and instance segmentation to estimate catheter tip positions in three spatial dimensions. This system was designed with two features not present in previous amplitude-based [5], [20] and deep learning-based [27], [28] photoacoustic visual servoing systems. First, our point source localization systems estimated the location of the catheter tip along the elevation dimension of the transducer. Second, the integration of force control into our visual servoing system improved the tracking of targets across uneven imaging surfaces, overcoming this stated limitation of our previous systems [5], [20], [27], [28]. These features improve the potential of our novel visual servoing system for clinical translation (e.g., cardiac catheterizations, other interventional procedures).

It is encouraging that our novel visual servoing system achieved catheter tip tracking errors as low as 0.5 mm and 0.8 mm in the phantom and *in vivo* environments, respectively (Table VI). These errors are comparable to needle tip tracking errors ranging 0.6-1.0 mm with our previous deep learning-based visual servoing system [27], when tracking needle tips in plastisol phantoms and *ex vivo* tissue. Notably, the lowest lateral tracking errors were achieved with WaveSegNet-1, which is promising for the implementation of real-time, theory-based, instance-segmentation point source localization approaches. In addition, both real-time and offline instance segmentation-based point source localization systems (i.e., WaveSegNet-1 and WaveSegNet-2, respectively) successfully detected the catheter tip in 100.0% of the input channel data frames during a majority of the *in vivo* visual servoing trials (i.e., 100% recall in Table IV).

There are a few suspected tradeoffs among source location outputs, signal amplitudes, accuracy, and inference times. First, the unexpectedly large elevation outputs of DetectionNet (Figs. 5(a) and 5(b)) are likely due to differences in the signal amplitude between the simulated data used to train DetectionNet and the experimental data provided to the system during the visual servoing trials. DetectionNet likely misinterpreted lower signal amplitudes as source displacements relative to the elevation center of the transducer. This hypothesis indicates that the performance of DetectionNet may be improved by increasing the range of signal amplitudes in the simulated training set. While the location estimation process of WaveSegNet-1 depended on the shape of the segmentation masks rather than signal amplitudes, the large elevation (Fig. 5(b)) and axial (Fig. 5(d)) outputs were likely caused by inaccuracies from neglecting second order terms during gradient descent. WaveSegNet-2 provided source locations based on the same segmentation masks as WaveSegNet-1, albeit with more consistent location estimates, likely due to the inclusion of second order terms during gradient descent. However, this inclusion negatively impacted inference times (Fig. 6). WaveSegNet-1 and WaveSegNet-2 had the greatest inference times due to the associated iterative gradient descent algorithms being the most computationally expensive step.

The large inference time standard deviations achieved with WaveSegNet-2 (Fig. 6) are caused by the large number of false positives output by the Mask R-CNN algorithm implemented prior to gradient descent (corresponding to decreased precision values in Table IV). In comparison, WaveSegNet-1 had greater precision (i.e., less false positives), and the detections output by the Faster R-CNN network forming DetectionNet were immediately output to the MTLKF, both resulting in smaller variations in inference times compared to WaveSegNet-2. DetectionNet achieved mean inference times as low as 145.3 ms, which is slower than the 10 Hz PRF of the laser (i.e., 100 ms between pulses), but was demonstrated in real-time nonetheless, with faster inference times than either WaveSegNet-1 or WaveSegNet-2.

The MTLKF is a computationally efficient alternative to the consistency check presented in previous visual servoing systems from our group [5], [20], [27], which required valid outputs in five consecutive channel data frames and additional position-based calculations to identify a tracked target as valid. It is promising that this filter successfully resolved positive and negative elevation source displacements arising from the elevation symmetry of photoacoustic waveforms encountered by our visual servoing system, enabling the generally low tracking errors of 0.5-3.6 mm (Table VI). In addition, while our visual servoing system required only a single output from the MTLKF, this filter potentially enables the development of a visual servoing system that simultaneously tracks multiple targets [26].

Two additional design choices contributed to the low catheter tip tracking errors measured in the phantom and *in vivo* environments (Table VI), despite the large elevation source displacements output by DetectionNet and WaveSegNet-1 (Fig. 5). First, the FSM relied more on the lateral rather than elevation position estimates. This choice was derived from the more accurate lateral localization performance of object detection-based and instance segmentation-based systems compared to elevation performance observed in our previous phantom and *ex vivo* experiments [24]. Second, no robot motion was implemented if point sources were within 1 mm of the lateral or elevation transducer center (Table II), which reduced the transducer motion required to maintain the catheter tip near the center of the imaging plane, relative to previous visual servoing systems [5], [20], [27], [28].

As expected, force control along the axial dimension of the transducer enabled consistent contact with the imaging surface (Fig. 10). While the transducer experienced motion in the vertical direction following the shape of the hemispherical phantom (Fig. 7(a)), the trajectory of the transducer did not follow a circular arc matching the radius of the phantom. This discrepancy was caused by the fixed orientation of the transducer, which resulted in a different point on the transducer being in contact with the phantom at each time instant. During the *in vivo* trials, the transducer exhibited rapid oscillations in the vertical direction (Figs. 7(b) and 7(c)) as a consequence of the respiratory motion of the swine and the limited force control model (i.e., hybrid-position force control in the transducer axial dimension, ignoring effects of robot dynamics on measured force readings). An example of the oscillations is available in Supplementary Video 1. The force control model was sufficient to achieve smoother transducer motion during the phantom trials and did not appear to negatively impact the *in vivo* tracking results relative to the phantom results (Table VI), indicating that lateral tracking is robust to respiratory motion. This unwanted axial transducer motion due to respiratory and related effects during *in vivo* trials could potentially be addressed with alternative control strategies [47]. Sterility concerns with required transducer contact can be addressed with commercially available transducer covers that are readily available for ultrasound-guided patient procedures, customized cardiovascular incise drapes [48], or a miniaturized system that operates under existing sterile drapes [5].

Considering the multiple factors noted above (e.g., tradeoffs, inference speed variability, localization and tracking performance), DetectionNet offers the best real-time potential among the three deep learning-based systems considered herein (Fig. 6), whereas WaveSegNet-1 is the first implementation of a real-time instance segmentation-based point source localization system, offering the best real-time axial point source detection performance (Table IV) and real-time lateral tracking performance (Table VI). The poorer axial and elevation position localization performance of WaveSegNet-1 relative to WaveSegNet-2 (Fig. 5) indicates that second-order gradient terms are critical for accurate localization of point sources using our theory-based wave segmentation approach, which is an unexpected outcome that we did not anticipate when developing WaveSegNet-1. Despite current inference times with WaveSegNet-2 prohibiting real-time implementation (Fig. 6), the *in vivo* potential of this approach to outperform object detection-based methods (e.g., DetectionNet) has been successfully demonstrated (Figs. 5 and 7), with additional benefits likely achievable with future system optimizations (e.g., enhanced speed via alternative network architectures [49], gradient descent algorithms [50], and reduced image dimensions [30], [51], [52]).

Although successful *in vivo* performance is more important than unsuccessful phantom performance when determining clinical translatability and future system optimizations, WaveSegNet-1 failed to track the catheter tip in the phantom likely because of first-order gradient terms, which are suspected to have caused the large axial and elevation variations in Figs. 5(b), 5(d), and 7(c), as indicated above. A similar failure was not achieved with the *in vivo* trials, likely due to the MTLKF (yellow box in Fig. 1) receiving consistent catheter tip location estimates from WaveSegNet-1 (see real-time performance with *In Vivo* Trials 3 and 4 in Table IV). Considering the reduced offline detection performance of WaveSegNet-2 in the phantom relative to *in vivo* trials (Table IV) and the identical first stages of WaveSegNet-1 and WaveSegNet-2 (i.e., Mask R-CNN), the detection performance of WaveSegNet-1 in the phantom trials likely degraded the consistency of the catheter tip location estimates provided to the MTLKF, resulting in no phantom results with WaveSegNet-1.

The objective of an optimal photoacoustic visual servoing system is to simultaneously maximize surgical tool tip localization performance and achievable frame rates. Given the failure of WaveSegNet-1 to successfully perform in the phantom trials (Section II-E) or achieve our axial and elevation localization goals in the *in vivo* trials (Fig. 5), we must consider other possible approaches to leverage the improved localization performance of WaveSegNet-2 [24] without suffering from the associated low frame rates with existing computing hardware. For example, amplitude-based [20] or coherence-based [23] photoacoustic visual servoing approaches were previously demonstrated to operate on delay-and-sum or short-lag spatial coherence (SLSC) beamformed images in real time (i.e., with execution times $\leq$100 ms, corresponding to the 10 Hz laser PRF). These real-time approaches could potentially be combined with WaveSegNet-2 to receive periodic 3D source location estimates (e.g., at the 1-2 Hz frame rates demonstrated in Fig. 6) using techniques similar to sensor fusion algorithms [53] employed in automotive [54] and aerospace [55] applications. The periodic elevation information from WaveSegNet-2 would provide periodic robustness to reflection artifacts and lateral localization performance, while informing the system of rotations about the axial dimension required to periodically compensate for out-of-plane motion in lengthy surgical and interventional procedures.

These multiple possible photoacoustic visual servoing options could be deployed either as a standalone system to replace fluoroscopy or as an add-on to conventional imaging modalities (e.g., ultrasound, fluoroscopy). As a standalone system, photoacoustic visual servoing could be used to track catheter tips, then the photoacoustic imaging component can be used to assess ablated lesion boundaries [56]. As an add-on, photoacoustic visual servoing may be augmented with ultrasound to provide additional anatomical information or with fluoroscopy images that provide intermittent checks of catheter tip positions.

One limitation of our study is the absence of a continuous synchronized ground truth using an external system (e.g., fluoroscopy) for the visual servoing trials. While catheter tip locations were confirmed with fluoroscopy at the start and end of four *in vivo* trials, future work could potentially utilize fluoroscopic videos to characterize the instantaneous localization performance of our visual servoing system across each visual servoing trial. These videos would need to be acquired from multiple fixed reference frames to compensate for the lack of depth information in individual fluoroscopy images, which would significantly extend the time required to complete a single visual servoing trial. As an alternative, previous work by Graham *et al.* [5] used an electromagnetic tracking system to validate an amplitude-based photoacoustic visual servoing system. This tracking system could potentially be used with intermediate checkpoints separated by small distances (e.g., 5 mm) to provide additional points of comparison with our deep learning-based photoacoustic visual servoing system.

A second potential study limitation is that the mean laser energy of 2.0 mJ corresponds to a laser fluence of 254.6 mJ/cm^2 within the IVC. While this laser fluence value exceeded the 25.2 mJ/cm^2 laser safety limit defined for skin at a wavelength of 750 nm [57], no such safety limit has been published for internal tissue. Previous work by our group demonstrated the use of higher laser energy levels during *in vivo* swine studies without observable tissue damage in post-exposure histopathological and immunohistochemistry studies [5], [58], [59]. In addition, our group previously introduced a theoretical framework linking predictions of required laser energies to visual servoing performance through the generalized contrast-to-noise ratio (gCNR) [60]. When evaluating this theory alongside data acquired from a previous *in vivo* cardiac catheterization experiment [61], gCNR values $\geq$0.56 were achieved with laser energies $\geq$104.7 μJ (i.e., $\geq$13.3 mJ/cm^2 fluence), corresponding to $\geq$97.8% predicted segmentation success rates (reported as segmentation accuracy in [61]), which could be viewed as expected success rates of visual servoing with lower energies [60]. Although these values refer to achievements that are possible with delay-and-sum beamforming, similar achievements are anticipated to be possible for successful visual servoing within current safety standards, with appropriate modifications to deep learning approaches applied to photoacoustic channel data (e.g., histogram matching [30], SLSC beamforming [23], [62], [63] combined with pulsed laser diodes [64], acquisitions with optical wavelengths that allow higher energies within fluence safety limits [6], [23], [57]).

V. Conclusion

This work demonstrates a novel deep learning-based photoacoustic visual servoing system tracking a catheter tip during an *in vivo* catheterization procedure. We successfully integrated object detection-based and instance segmentation-based 3D point source localization systems (i.e., DetectionNet and WaveSegNet-1, respectively), with MTLKF and a hybrid position-force control system to ultimately track a catheter tip *in vivo*. We also characterized the ability of our visual servoing system to detect and localize the catheter tip in phantom and *in vivo* environments, using raw photoacoustic channel data frames as the input. Our system successfully followed the catheter tip while continuously maintaining contact with the imaging surface. In addition to real-time demonstrations, we validated the potential of an offline instance segmentation-based point source localization system using second order gradient terms (i.e., WaveSegNet-2) to improve catheter tip localization at the cost of increased inference times, with the potential for additional optimizations to increase implementation speeds. These contributions are promising to autonomously track and visualize catheter tips, needle tips, and other surgical or interventional tool tips in real time.

References

[1] L. Yatziv, *et al.*, “Toward multiple catheters detection in fluoroscopic image guided interventions,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 770–781, 2012.

[2] B. D. Lindsay, *et al.*, “Radiation exposure to patients and medical personnel during radiofrequency catheter ablation for supraventricular tachycardia,” *The American Journal of Cardiology*, vol. 70, no. 2, pp. 218–223, 1992.

[3] L. S. Rosenthal, *et al.*, “Predictors of fluoroscopy time and estimated radiation exposure during radiofrequency catheter ablation procedures,” *The American Journal of Cardiology*, vol. 82, no. 4, pp. 451–458, 1998.

[4] M. Mahesh, “Fluoroscopy: Patient radiation exposure issues,” *Radiographics*, vol. 21, no. 4, pp. 1033–1045, 2001.

[5] M. Graham, *et al.*, “In vivo demonstration of photoacoustic image guidance and robotic visual servoing for cardiac catheter-based interventions,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1015–1029, 2020.

[6] M. A. L. Bell, “Photoacoustic imaging for surgical guidance: principles, applications, and outlook,” *Journal of Applied Physics*, vol. 128, no. 6, p. 060904, 2020.

[7] M. A. Lediju, *et al.*, “Quantitative assessment of the magnitude, impact and spatial extent of ultrasonic clutter,” *Ultrasonic Imaging*, vol. 30, no. 3, pp. 151–168, 2008.

[8] A. Wiacek and M. A. L. Bell, “Photoacoustic-guided surgery from head to toe,” *Biomedical Optics Express*, vol. 12, no. 4, pp. 2079–2117, 2021.

[9] E. A. González, *et al.*, “Combined ultrasound and photoacoustic image guidance of spinal pedicle cannulation demonstrated with intact ex vivo specimens,” *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2479–2489, 2021.

[10] J. Shubert and M. A. L. Bell, “Photoacoustic imaging of a human vertebra: implications for guiding spinal fusion surgeries,” *Physics in Medicine & Biology*, vol. 63, no. 14, p. 144001, 2018.

[11] N. Gandhi, *et al.*, “Photoacoustic-based approach to surgical guidance performed with and without a da Vinci robot,” *Journal of Biomedical Optics*, vol. 22, no. 12, p. 121606, 2017.

[12] M. Allard, *et al.*, “Feasibility of photoacoustic-guided teleoperated hysterectomies,” *Journal of Medical Imaging*, vol. 5, no. 2, p. 021213, 2018.

[13] M. A. L. Bell, *et al.*, “Localization of transcranial targets for photoacoustic-guided endonasal surgeries,” *Photoacoustics*, vol. 3, no. 2, pp. 78–87, 2015.

[14] M. T. Graham, *et al.*, “Simulations and human cadaver head studies to identify optimal acoustic receiver locations for minimally invasive photoacoustic-guided neurosurgery,” *Photoacoustics*, vol. 19, p. 100183, 2020.

[15] ——, “Validation of eyelids as acoustic receiver locations for photoacoustic-guided neurosurgery,” in *Proceedings of SPIE Photonics West*, vol. 11642. SPIE, 2021, pp. 162–167.

[16] E. Najafzadeh, *et al.*, “Application of multi-wavelength technique for photoacoustic imaging to delineate tumor margins during maximum-safe resection of glioma: A preliminary simulation study,” *Journal of Clinical Neuroscience*, vol. 70, pp. 242–246, 2019.

[17] J. Zhang, *et al.*, “Multispectral photoacoustic imaging of breast cancer tissue with histopathology validation,” *Biomedical Optics Express*, vol. 16, no. 3, pp. 995–1005, 2025.

[18] K. M. Kempski, *et al.*, "In vivo photoacoustic imaging of major blood vessels in the pancreas and liver during surgery," *Journal of Biomedical Optics*, vol. 24, no. 12, p. 121905, 2019.

[19] A. Wiacek, *et al.*, "Photoacoustic-guided laparoscopic and open hysterectomy procedures demonstrated with human cadavers," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3279–3292, 2021.

[20] M. A. L. Bell and J. Shubert, "Photoacoustic-based visual servoing of a needle tip," *Scientific Reports*, vol. 8, p. 15519, 2018.

[21] D. Ostler-Mildner, *et al.*, "The sound of surgery-development of an acoustic trocar system enabling laparoscopic sound analysis," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2024.

[22] J. Shubert and M. A. L. Bell, "Photoacoustic based visual servoing of needle tips to improve biopsy on obese patients," in *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2017, pp. 1–4.

[23] E. A. González and M. A. L. Bell, "GPU implementation of photoacoustic short-lag spatial coherence imaging for improved image-guided interventions," *Journal of Biomedical Optics*, vol. 25, no. 7, p. 077002, 2020.

[24] M. R. Gubbi and M. A. L. Bell, "Deep learning to localize photoacoustic sources in three dimensions: Theory and implementation," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 72, no. 6, pp. 786–805, 2025.

[25] A. Reiter and M. A. L. Bell, "A machine learning approach to identifying point source locations in photoacoustic data," in *Proceedings of SPIE Photonics West*, vol. 10064. SPIE, 2017, pp. 504–509.

[26] D. Allman, *et al.*, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1464–1477, 2018.

[27] 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 *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 14 261–14 267.

[28] T. R. Folk, *et al.*, "Development of a ROS2-based photoacoustic-robotic visual servoing system," in *Proceedings of SPIE Photonics West*. SPIE, 2025.

[29] H. Wang, *et al.*, "Three-dimensional interventional photoacoustic imaging for biopsy needle guidance with a linear array transducer," *Journal of Biophotonics*, vol. 12, no. 12, p. e201900212, 2019.

[30] M. R. Gubbi, *et al.*, "Deep learning in vivo catheter tip locations for photoacoustic-guided cardiac interventions," *Journal of Biomedical Optics*, vol. 29, no. S1, p. S11505, 2023.

[31] Y. Yu, *et al.*, "Bias estimation and gravity compensation for wrist-mounted force/torque sensor," *IEEE Sensors Journal*, vol. 22, no. 18, pp. 17 625–17 634, 2021.

[32] M. Quigley, *et al.*, "ROS: An open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[33] Y. Wu, *et al.*, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[34] J. Deng, *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[35] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of Biomedical Optics*, vol. 15, no. 2, p. 021314, 2010.

[36] S. Ren, *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[37] K. He, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[38] ——, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[39] Y. Wang, "Gauss–Newton method," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 4, pp. 415–420, 2012.

[40] J. C. Meza, "Newton's method," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 1, pp. 75–78, 2011.

[41] A. Paszke, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[42] Y. Bar-Shalom, "Multitarget-multisensor tracking: Advanced applications," *Norwood*, 1990.

[43] Y. Bar-Shalom, *et al.*, *Estimation with applications to tracking and navigation: Theory, algorithms, and software*. John Wiley & Sons, 2004.

[44] G. Welch and G. Bishop, "An introduction to the kalman filter," 1995.

[45] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval*. Springer, 2005, pp. 345–359.

[46] B. K. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the optical society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[47] H. Arenbeck, *et al.*, "Control methods for robot-based predictive compensation of respiratory motion," *Biomedical Signal Processing and Control*, vol. 34, pp. 16–24, 2017.

[48] F. A. Czajka and R. A. Lockwood, "Medical drape," June 6 2017, US Patent 9,668,822.

[49] J. Huang, *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7310–7311.

[50] G. Grisetti, *et al.*, "Least squares optimization: From theory to practice," *Robotics*, vol. 9, no. 3, p. 51, 2020.

[51] M. A. Lediju, *et al.*, "Sources and characterization of clutter in cardiac b-mode images," in *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2009, pp. 1419–1422.

[52] M. A. L. Bell, *et al.*, "In vivo liver tracking with a high volume rate 4D ultrasound scanner and a 2D matrix array probe," *Physics in Medicine & Biology*, vol. 57, no. 5, p. 1359, 2012.

[53] J. Z. Sasiadek, "Sensor fusion," *Annual Reviews in Control*, vol. 26, no. 2, pp. 203–228, 2002.

[54] D. J. Yeong, *et al.*, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[55] S. Blackman and T. Broida, "Multiple sensor data association and fusion in aerospace applications," *Journal of Robotic Systems*, vol. 7, no. 3, pp. 445–485, 1990.

[56] M. T. Graham, *et al.*, "Photoacoustic image guidance and robotic visual servoing to mitigate fluoroscopy during cardiac catheter interventions," in *Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XVIII*, vol. 11229. SPIE, 2020, pp. 80–85.

[57] A. N. S. Institute, *American national standard for safe use of lasers*. Laser Institute of America, 2022.

[58] J. Huang, *et al.*, "Empirical assessment of laser safety for photoacoustic-guided liver surgeries," *Biomedical Optics Express*, vol. 12, no. 3, pp. 1205–1216, 2021.

[59] J. J. Arroyo, *et al.*, "Predictive model for laser-induced tissue necrosis with immunohistochemistry validation," *Biophotonics Discovery*, vol. 1, no. 2, pp. 025 003–025 003, 2024.

[60] M. R. Gubbi, *et al.*, "Theoretical framework to predict generalized contrast-to-noise ratios of photoacoustic images with applications to computer vision," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 6, pp. 2098–2114, 2022.

[61] E. A. González, *et al.*, "A beamformer-independent method to predict photoacoustic visual servoing system failure from a single image frame," in *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2021, pp. 1–4.

[62] A. Wiacek, *et al.*, "CohereNet: A deep learning architecture for ultrasound spatial correlation estimation and coherence-based beamforming," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2574–2583, 2020.

[63] J. Timaná, *et al.*, "Application of CohereNet to photoacoustic data for non-invasive, in vivo, subcutaneous imaging," in *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2023, pp. 1–4.

[64] M. A. L. Bell, *et al.*, "Improved contrast in laser-diode-based photoacoustic images with short-lag spatial coherence beamforming," in *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2014, pp. 37–40.