Amplitude-Aware Deep Learning-Based Tool Tip Localization in Raw Photoacoustic Channel Data

Nethra Venkatayogi*, Muyinatu A. Lediju Bell*†‡
*Department of Computer Science, Johns Hopkins University, Baltimore, MD
†Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD
‡Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

***Abstract*—Photoacoustic imaging is a promising modality for real-time surgical guidance to visualize critical structures during minimally invasive procedures. Deep learning-based visual servoing systems utilizing raw photoacoustic channel data have successfully tracked catheter tips in cardiac interventions. However, existing deep learning methods prioritize the proximal waveform and risk decreased generalizability due to low-amplitude artifacts. Therefore, we present an approach that leverages amplitude-aware training to improve surgical tool tip localization using raw photoacoustic channel data. A Faster R-CNN network was trained on 20,000 k-Wave-simulated frames and tested with experimental data. The system achieved a mean absolute tracking error of 0.42 mm and 0.91 mm in the axial and lateral image dimensions, respectively, with a frame rate of 10.9 Hz, which is compatible with our 10 Hz laser pulse repetition frequency. The proposed system promises to provide real-time guidance to track surgical tool tips while minimizing erroneous detections from low-amplitude signals and artifacts in multiple locations.**

***Index Terms*—Deep learning, Photoacoustic imaging, Surgical tool tracking system, Surgical guidance**

I. Introduction

Photoacoustic imaging is a promising modality for real-time surgical guidance to visualize critical anatomical structures during minimally invasive procedures [1], [2]. Photoacoustic imaging, based on the photoacoustic effect, utilizes pulsed laser light to excite optically absorbing chromophores, resulting in thermal expansion and the generation of an acoustic wave received by a standard ultrasound transducer. Potential applications of photoacoustic image guidance in minimally invasive interventional procedures include tool tracking in spinal surgeries [3], [4], photoacoustic-guided teleoperative robotic surgeries [5], [6], guidance of minimally invasive neurosurgeries [7]–[9], tumor boundary delineation [10], [11], large vessel tracking during liver procedures [12], and monitoring of the proximity of tools to critical areas of interest during hysterectomies [13].

Photoacoustic imaging has demonstrated particular success in cardiac interventions, where deep learning approaches utilizing raw photoacoustic channel data enable visual servoing systems to track catheter tips in real-time [14]–[18]. These systems detect catheter tips through a deep learning point source localization model trained on simulated images of 0.1 mm-diameter point sources. The translation of these visual servoing and localization systems from tracking smaller (e.g., catheter tips) to larger (e.g., cautery devices, drill bits, da Vinci surgical instruments, ureters, uterine arteries) imaging targets presents unique challenges due to varying target geometries and complex acoustic environments [5], [6], [13], [19].

Existing photoacoustic-based visual servoing systems face two critical limitations that restrict their broader surgical applicability. First, current methods prioritize the most proximal waveform detected in photoacoustic channel data images, operating under the assumption that reflection artifacts will appear distal to the true source location [18], [20]. This assumption fails in complex surgical environments as shape distortion arising from elevational displacements and out-of-plane absorbers can lead to false positive detections of the source. Second, low-amplitude artifacts surrounding photoacoustic sources can decrease system generalizability across different surgical scenarios and between simulation and experimental conditions.

We hypothesize that incorporating amplitude-aware training into existing photoacoustic-based visual servoing systems will improve the identification and tracking of surgical tool tips in real time. Low-amplitude, proximal artifacts selected as the source are expected to be reduced with this approach. This reduction is expected to improve the correct identification of source waveforms.

In this paper, we present the development and validation of an amplitude-aware photoacoustic tracking system with the creation of a model trained on a simulated dataset of sources and reflection artifacts differentiated by amplitude. In addition to amplitude-aware training, the signal amplitude at the predicted waveform peak location was utilized to compute a weighted sum of model confidence score and signal amplitude to confirm the tool prediction and choose among multiple possible detections. We then present experimental validation using robot-controlled tool motions to evaluate tracking performance, compared to our previously existing point source localization model.

II. Methods

A. Simulated datasets for training and validation

We simulated photoacoustic channel data images using k-Wave to create a dataset for model training and validation (similar to the method presented in [14], [16]). The ranges and increment sizes of the simulation variables are listed in Table I. Each simulation consisted of a 0.1 mm-diameter point source in a two-dimensional simulation grid consisting of a homogeneous medium, with lateral and axial dimensions of

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)	-74.3	74.3	0.1
Number of Sources	1	1	-
Number of Ref. Artifacts	0	1	Random
Channel SNR (dB)	-5	2	Random
Source and Ref. Artifact Diameter (mm)	0.9	2.1	Random
Source Amplitude	0.7	1.1	Random
Ref. Artifact Amplitude (95% of instances)	0.01	0.5	Random
Ref. Artifact Amplitude (5% of instances)	0.77	1.05	Random

97 mm and 122 mm, respectively. To generate sources with diameters that were more representative of surgical tool tips as opposed to catheter tips, we superposed stored simulated outputs at axially varied positions to create sources with diameters of 0.9-2.1 mm.

Reflection artifacts were created using the method previously presented in [21] (i.e., waveforms originating from photoacoustic sources were axially downshifted by the Euclidean distance between an actual source and the source representing the artifact). We simulated a discrete ultrasound probe model with a sampling frequency of 11.88 MHz, an aperture of 64 elements, an element width of 0.25 mm, and an inter-element spacing of 0.05 mm. These parameters were selected to match the specifications of the Verasonics P4-2v probe [22]. To incorporate amplitude-awareness into training, source and artifact amplitude scaling factors ranged from 0.7-1.1 and 0.01-0.5, respectively. A subset (i.e., 5%) of reflection artifacts were scaled to be brighter than the source in the image and had an amplitude range of 0.77-1.05. Gaussian noise was added to the resulting raw photoacoustic channel data frame using the addNoise function in the k-Wave toolbox [23].

As with previous implementations of phased array transducer-based point source localization systems [14], [15], [17], [24], [25], each raw channel data frame was zero-padded to match the field-of-view of a scan-converted photoacoustic image to form a zero-padded channel data frame of dimensions 565×926 pixels. The image was then laterally upsampled by a factor of 2 to form a final channel data frame of dimensions 1130x926 pixels. These zero-padded channel data frames were annotated using the method presented by Gubbi et al. [15] with class information (i.e., "source" or "artifact") and bounding boxes of dimensions 32×16 pixels centered on the positions of sources and artifacts to form annotated images. The total dataset included 20,000 photoacoustic channel data frames from which annotated images were randomly separated into training (80%) and validation (20%) datasets.

B. Network Architecture and Training Procedure

The Detectron-2 platform [26] was utilized for training and validation. A Faster R-CNN network [27] with a ResNet-101 [28] feature extractor was initialized with pre-trained weights from the ImageNet dataset [29] then fine-tuned for 80000 iterations with a batch size of 4 and a base learning rate of 1×10^{-3} on an NVidia (Santa Clara, California) Titan X (Pascal) GPU. The network was trained to detect and classify each waveform in the input photoacoustic channel data frame as a source or reflection artifact and position a bounding box around the peak of the detected waveform. If the peak was not visible in the photoacoustic channel data when the lateral location of the source or artifact resided in the zero-padded region, the network was required to classify the waveform and extrapolate the position of its peak using the visible portion of the waveform present in the input channel data frame. For each input image, the network outputs consisted of the identified class (i.e., source or artifact), the object location (i.e., bounding box pixel coordinates), and a confidence score (0-1), as summarized in Fig. 1.

After training, the network performed inference on input images at an average rate of 0.092 s per image, translating to an achievable frame rate of 10.9 Hz for real-time photoacoustic tool localization. Similar to [14], for robustness, the estimated tool tip were compared across five consecutive frames. If the tool tip was visible in each frame with a confidence score $>$ 0.7, and the estimated position of the tool tip did not change by more than 1 cm across 5 frames, then the location estimate was labeled as valid. To implement amplitude-aware localization when there were multiple tool tip detections, a weighted sum based on 0.4 times the confidence score and 0.6 times the signal amplitude at the bounding box location was used to select the tool tip position. If signal amplitude was 0 (i.e.,

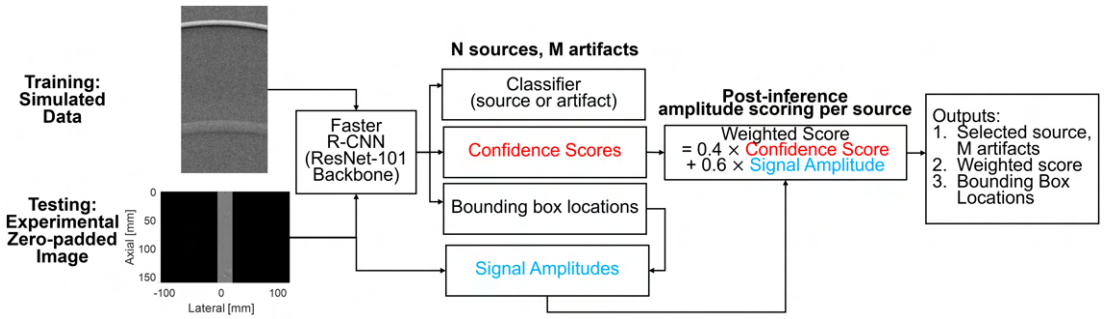

Fig. 1. Summary of model architecture and workflow.

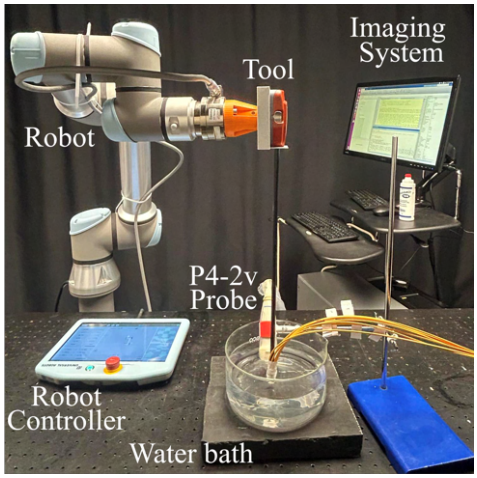

Fig. 2. Experimental setup for model testing.

the detected peak was in the zero-padded region), only the confidence score was used to determine the tool tip position.

C. Experimental data for testing

To test our network, our experimental set up consisted of a Vantage 128 ultrasound scanner (Verasonics Inc., Kirkland, WA, USA), a Verasonics P4-2v phased array ultrasound probe, a Phocus Mobile laser (Opotek, Carlsbad CA, USA) emitting 5 ns pulses at a rate of 10 Hz with wavelength of 750 nm, and a 1-to-7 fiber splitter [30], as shown in Fig. 2. A custom designed, 3-D printed fiber holder (described in [6]) secured the seven output fibers to surround the da Vinci curved scissor tool. The tool tip was extended from the fiber holder tip by 5 mm to illuminate the tool tip. The tool was affixed to the end effector of a UR5e robot arm (Universal Robots, Denmark), using a 3D-printed holder. A water bath in a glass beaker was used to perform controlled tool motions. The lateral, elevation, and axial dimensions of the P4-2v ultrasound probe were aligned with the x-, y-, and z-dimensions of the robot by moving the tool tip (attached to the robot end effector) to fixed positions and localizing the tool tip in photoacoustic images. The transducer remained fixed throughout experiments.

The robot was commanded to perform seven back-and-forth tool translations, across the lateral image dimension, spanning 38 mm per direction. Each motion was performed at 9 mm/s velocity. This velocity was determined based on maximum expected velocities of 17.26 mm/s (axial) and 10.68 mm/s (lateral), determined from tool motions performed by a board-certified gynecological surgeon during an open procedure performed on a human cadaver [19]. Robot positions were published at a rate of 500 Hz to the Robot Operating System (ROS) topic /tf, and corresponding photoacoustic channel data were published to the same topic at the data acquisition rate of 10 Hz. The closest robot position to each channel data acquisition were correlated, resulting in corresponding robot and photoacoustic-based tool displacements. The transformation between the robot base and the tool tip was determined by constructing directed acyclic graphs from the /tf messages.

The resulting photoacoustic-based (p_i) and robot-based (r_i) tool displacements were used to determine the motion tracking error. In particular, the mean absolute error (MAE) was determined, per lateral (x) or axial (z) dimension, as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \tag{1}$$

where n is the number of corresponding robot and photoacoustic-based tool displacements per programmed robot trajectory. The standard deviation of the absolute error was also calculated along the x and z dimensions. If no tool tip was detected based on the above conditions, tool positions were excluded from comparison with robot displacements. Frames without a detected tool tip based on the above conditions were excluded from comparison with robot displacements. For comparison, results from a point-source localization model trained on 0.1 mm-diameter point sources without amplitude-awareness, referred to as *Faster R-CNN Model*, are presented, while the proposed network model is referred to as the *Amplitude Aware Faster R-CNN Model*.

III. Results and Discussion

Fig. 3 compares *Faster R-CNN* and *Amplitude Aware Faster R-CNN* detections on experimental channel data images and delay-and-sum beamformed images. *Faster R-CNN* detects a low-amplitude proximal artifact as a source, while *Amplitude Aware Faster R-CNN* correctly localizes the source at the ∼40 mm expected depth. In addition, the *Amplitude Aware Faster R-CNN* correctly identifies the tool despite multiple signals in the beamformed images, overcoming a key limitation of non-deep learning-based, amplitude-based tool tip tracking methods using beamformed images [18].

Across 1824 photoacoustic images, the *Amplitude Aware Faster R-CNN* achieved MAEs of 0.42 mm and 0.91 mm in the axial and lateral dimensions, respectively. The associated rejection rate (i.e., percentage of excluded frames without a detected tool tip based on the conditions reported in Section II-C) was 1.32% (i.e., 24 out of 1824 frames). The *Faster R-CNN* achieved MAEs of 4.42 mm and 3.03 mm in the axial

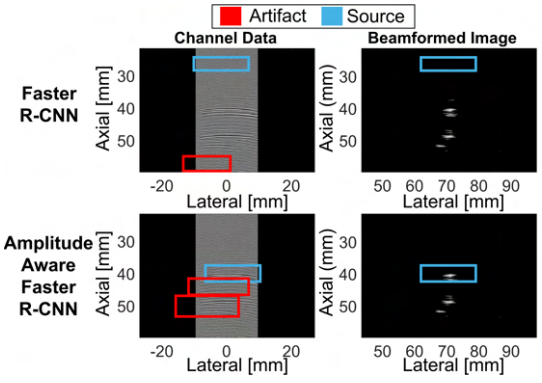

Fig. 3. Visualization of *Faster R-CNN* and *Amplitude Aware Faster R-CNN* outputs on Channel Data and Delay-and-Sum Beamformed Images. In beamformed images, artifact predictions are omitted to highlight source localization.

and lateral dimensions, respectively, with a rejection rate of 90.84% (i.e., 1657 out of 1824 frames) when validity checks were applied. Removing the motion validity constraint reduced the rejection rate to 2.41% (i.e., 44 out of 1824 frames). However, the MAEs increased to 1.78 mm and 1.80 mm in the axial and lateral dimensions, respectively, when these constraints were removed, resulting in an increase by factors of 4.24 and 1.98, respectively, relative to that achieved with the *Amplitude Aware Faster R-CNN*.

IV. Conclusion

This work is the first to utilize amplitude-aware training to detect surgical tool tips in photoacoustic channel data. By incorporating amplitude information, the model achieves MAEs of 0.42 and 0.91 mm in the axial and lateral dimensions, which are 4.24x and 1.98x lower, respectively, than those achieved with *Faster R-CNN*. The enhanced tool tip localization with the *Amplitude Aware Faster R-CNN* indicates strong potential to improve existing photoacoustic-based visual servoing systems by minimizing erroneous detections caused by low-amplitude signals and artifacts. Future work will extend this approach to detect multiple sources simultaneously (e.g., both the surgical tool tip and ureter to assist with preventing accidental ureteral injuries during hysterectomies). Model fine-tuning may also be performed on experimental data and model validation could be further verified with human cadaver studies.

Acknowledgment

This work is supported by NIH R01 EB032358. We thank Mardava Gubbi and Manik Kakkar for their insights.

References

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

[2] 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.

[3] E. A. González, A. Jain, and M. A. L. Bell, "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.

[4] 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.

[5] N. Gandhi, M. Allard, S. Kim, P. Kazanzides, and M. A. L. Bell, "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.

[6] M. Allard, J. Shubert, and M. A. L. Bell, "Feasibility of photoacoustic-guided teleoperated hysterectomies," *Journal of Medical Imaging*, vol. 5, no. 2, p. 021213, 2018.

[7] M. A. L. Bell, A. K. Ostrowski, K. Li, P. Kazanzides, and E. M. Boctor, "Localization of transcranial targets for photoacoustic-guided endonasal surgeries," *Photoacoustics*, vol. 3, no. 2, pp. 78–87, 2015.

[8] M. T. Graham, J. Huang, F. X. Creighton, and M. A. L. Bell, "Simulations and human cadaver head studies to identify optimal acoustic receiver locations for minimally invasive photoacoustic-guided neurosurgery," *Photoacoustics*, vol. 19, p. 100183, 2020.

[9] M. T. Graham, F. X. Creighton, and M. A. L. Bell, "Validation of eyelids as acoustic receiver locations for photoacoustic-guided neurosurgery," in *Proceedings of SPIE Photonics West*, vol. 11642, pp. 162–167, SPIE, 2021.

[10] E. Najafzadeh, H. Ghadiri, M. Alimohamadi, P. Farnia, M. Mehrmohammadi, and A. Ahmadian, "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.

[11] J. Zhang, J. Arroyo, and M. A. Lediju Bell, "Multispectral photoacoustic imaging of breast cancer tissue with histopathology validation," *Biomedical Optics Express*, vol. 16, no. 3, pp. 995–1005, 2025.

[12] K. M. Kempski, A. Wiacek, M. Graham, E. González, B. Goodson, D. Allman, J. Palmer, H. Hou, S. Beck, J. He, and M. A. L. Bell, "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.

[13] A. Wiacek, K. C. Wang, H. Wu, and M. A. L. Bell, "Photoacoustic-guided laparoscopic and open hysterectomy procedures demonstrated with human cadavers," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3279–3292, 2021.

[14] 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)*, pp. 14261–14267, IEEE, 2021.

[15] M. R. Gubbi, F. Assis, J. Chrispin, and M. A. L. Bell, "Deep learning in vivo catheter tip locations for photoacoustic-guided cardiac interventions," *Journal of Biomedical Optics*, vol. 29, no. S1, p. S11505, 2023.

[16] T. R. Folk, M. R. Gubbi, and M. A. L. Bell, "Development of a ROS2-based photoacoustic-robotic visual servoing system," in *Proceedings of SPIE Photonics West*, SPIE, 2025.

[17] 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.

[18] M. R. Gubbi, A. Kolandaivelu, N. Venkatayogi, J. Zhang, P. Warbal, G. C. Keene, M. Khairalseed, J. Chrispin, and M. A. L. Bell, "In vivo demonstration of deep learning-based photoacoustic visual servoing system," *IEEE Transactions on Biomedical Engineering*, 2025.

[19] N. Venkatayogi, K. C. Wang, and M. A. L. Bell, "Velocity-based filtering approach to photoacoustic-guided hysterectomy demonstrated with a human cadaver," in *Proceedings of SPIE Photonics West*, vol. 13319, pp. 110–115, SPIE, 2025.

[20] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1464–1477, 2018.

[21] 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 *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2018.

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

[23] 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, pp. 021314–021314, 2010.

[24] D. Allman, F. Assis, J. Chrispin, and M. A. L. Bell, "Deep learning to detect catheter tips in vivo during photoacoustic-guided catheter interventions: Invited presentation," in *Proceedings of the 53rd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–3, IEEE, 2019.

[25] D. Allman, F. Assis, J. Chrispin, and M. A. L. Bell, "A deep learning-based approach to identify in vivo catheter tips during photoacoustic-guided cardiac interventions," in *Proceedings of SPIE Photonics West*, vol. 10878, p. 108785E, SPIE, 2019.

[26] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2." https://github.com/facebookresearch/detectron2, 2019.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

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

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009.

[30] B. Eddins and M. A. L. Bell, "Design of a multifiber light delivery system for photoacoustic-guided surgery," *Journal of Biomedical Optics*, vol. 22, no. 4, p. 041011, 2017.