# Development of a ROS2-based Photoacoustic-Robotic Visual Servoing System

Taylor R. Folk[a], Mardava R. Gubbi[b], and Muyinatu A. Lediju Bell[b,c,d,e]

[a]Department of Bioengineering, Harvard University, Boston, MA
[b]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD
[c]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD
[d]Department of Computer Science, Johns Hopkins University, Baltimore, MD
[e]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD

## ABSTRACT

A key component of many robotic and surgical procedures is the ability to effectively visualize and track surgical tool tips. In this paper, we introduce a real-time deep learning photoacoustic visual servoing system that uses ROS2 and Moveit2 to make and execute robot path planning decisions in order to track and maintain visualization of tool tips. This system also uses Detectron2 and 2D simulated photoacoustic channel data to train the deep neural network. The performance of this ROS2-based deep learning visual servoing system is compared to that of a deep learning-based visual servoing system that utilizes ROS as its software system, and Detectron and 2D simulated data to train its neural network. Experiments were conducted with a plastisol phantom. Needle tip tracking performance with the ROS2-based visual servoing system outperformed that of the ROS-based system by 23.53% in phantom tissue. These results demonstrate the benefits of upgrading the robot operating system to ROS2 for improved deep learning-based visual servoing and tracking of interventional tool tips.

## 1. INTRODUCTION

Tool tip tracking is an essential component of surgical and interventional procedures. The most commonly used tracking systems are ultrasound-based visual servoing systems, with more recent approaches incorporating deep learning to improve detection.[1,2] However, the effectiveness of ultrasound-based visual servoing is limited in acoustically challenging environments.[3] To overcome expected challenges with ultrasound imaging in these environments, photoacoustic imaging can be used. In contrast to ultrasound imaging, which relies on the transmission and reception of sound waves to form images, photoacoustic imaging uses light to produce acoustic signals that are detected by ultrasound detectors.[4,5] This photoacoustic approach combines the benefits of optical and acoustic imaging techniques. Photoacoustic imaging is particularly advantageous in acoustically challenging environments because it only requires one-way acoustic travel from the light source to the ultrasound receiver, unlike the round-trip acoustic travel necessary for conventional ultrasound imaging.

Previous work from our group demonstrated the successful use of photoacoustic image-based visual servoing systems for the continuous tracking of surgical tool tips.[6] In addition, we successfully implemented deep learning into the photoacoustic visual servoing system to better discriminate between sources and artifacts in images, as opposed to the conventional method of beamforming, which depends on mathematical models that do not account for multiple potential sources of photoacoustic image artifacts.[7] The use of deep learning to detect sources of interest in the raw sensor data before image formation is advantageous with respect to tool tip tracking. This detection was facilitated by using the Detectron[8] platform to train the neural network.

Our initial deep learning-based visual servoing system implementations utilized ROS as its software component. However, a newer and more advanced version of ROS called ROS2 has been developed and launched since these initial demonstrations. Improvements with ROS2 include increased message reliability, multi-thread

Send correspondence to: tfolk@college.harvard.edu, mardava.gubbi@jhu.edu, and mledijubell@jhu.edu

Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XXIII,
edited by Caroline Boudoux, James W. Tunnell, Proc. of SPIE Vol. 13306,
133060A · © 2025 SPIE · 1605-7422 · doi: 10.1117/12.3046395

execution, and real-time processing.[9] These improvements provide greater efficiency for real-time systems. One additional upgrade available is Detectron2[10] (as opposed to Detectron[8]), which offers greater speed, accuracy, and flexibility for object detection. This paper presents a deep learning-based photoacoustic-robotic visual servoing system that uses ROS2 for its software components and Detectron2 to train the deep neural network.

## 2. METHODS AND MATERIALS

### 2.1 Visual Servoing System

Fig. 1 shows a workflow diagram of the photoacoustic visual servoing system used in this work. First, a Phocus Mobile laser (Opotek, Carlsbad, CA, USA) interfaced to a 600 $\mu$m core diameter optical fiber was utilized to transmit pulsed laser light at a rate of 10 Hz and with a fixed wavelength of 750 nm. Each pulse of the laser triggered the acquisition of a frame of raw radiofrequency photoacoustic channel data with a Vantage 128 ultrasound scanner (Verasonics Inc., Kirkland, WA, USA) interfaced to a Verasonics P4-2v phased array ultrasound probe. Each channel data frame was provided to a deep learning-based point source localization system using the ROS2 framework. The detected coodinates correspond to the tip of the free end of the optical fiber. This fiber tip can be interfaced with the tip of a surgical or interventional tool (e.g., a catheter[11] or hollow core biopsy needle[7]), which was not implemented in this work for simplicity.

Similar to previous work from our group,[7,12,13] our point source localization system consisted of a Faster R-CNN network[14] with a ResNet-101[15] feature extractor, as shown in Fig. 2. This network was provided with a photoacoustic channel data frame acquired by the ultrasound probe as an input. The output was an estimate, ${}^{U}\hat{p}(n)$, of the fiber tip location in the coordinate frame $U$ along with a confidence score, $d\,(n)$, ranging zero to one indicating the likelihood that the output corresponded to the physical fiber tip. Because this point source localization system did not provide elevation displacement estimates, the y-dimension of ${}^{U}\hat{p}(n)$ was set to zero. For robustness, the fiber tip position estimates were compared across five consecutive frames. If the fiber tip was visible in each frame with a confidence score $d\,(n) > 0.7$ and the estimated position of the fiber tip did not change by more than 1 cm across 5 frames (i.e., corresponding to a maximum speed of 2 cm/s), then the location estimate was labeled as valid.

The coordinates extracted from the point source localization system were transformed to the robot coordinate frame, then provided to a MoveIt2-based (PickNik Robotics, Boulder, CO, USA) motion planning algorithm to center the probe above the fiber tip using a UR5e robot (Universal Robots, Odense, Denmark). The probe was attached to the end effector of the robot using a custom 3D-printed adapter. Moveit2 created a motion plan

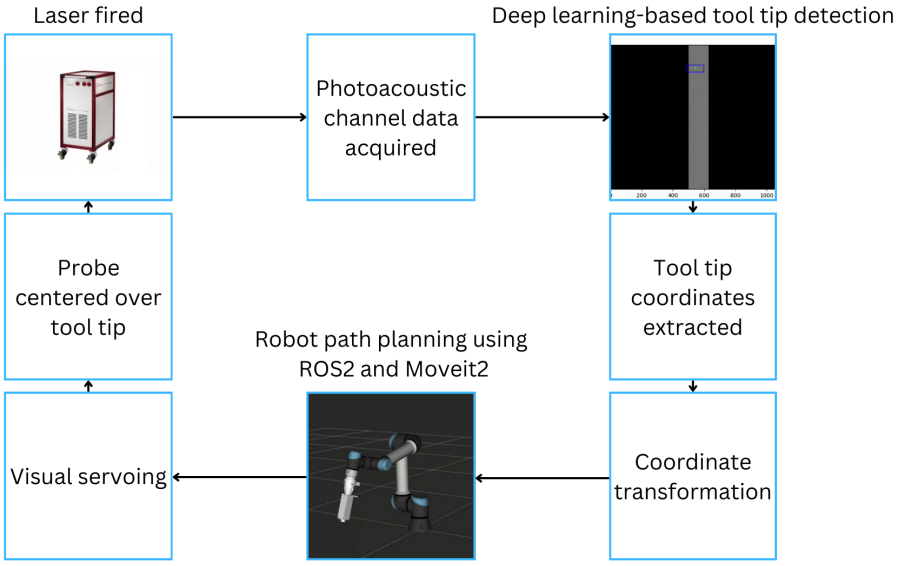



Figure 1. Block diagram illustrating the updated photoacoustic visual serving system.

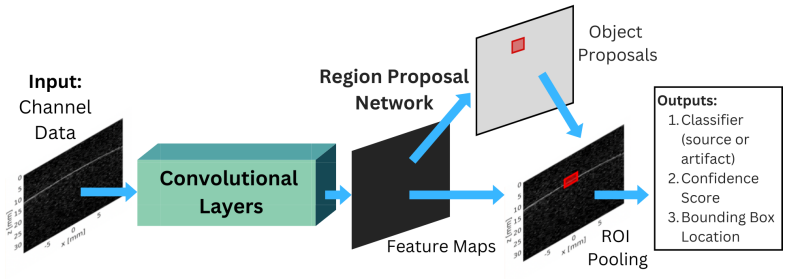



Figure 2. Architecture diagram of Faster R-CNN network with ResNet-101 feature extractor used in ROS2-based photoacoustic visual servoing system.

based on the current location of the probe relative to the robot base frame, and the desired location of the probe based on ${}^{U}\hat{p}(n)$. The successfully generated motion plan was then executed with the robot to center the ultrasound probe above the fiber tip. The cycle then repeated with the next pulse of the laser.

## 2.2 Simulated Datasets to Train and Validate Photoacoustic Point Source Localization

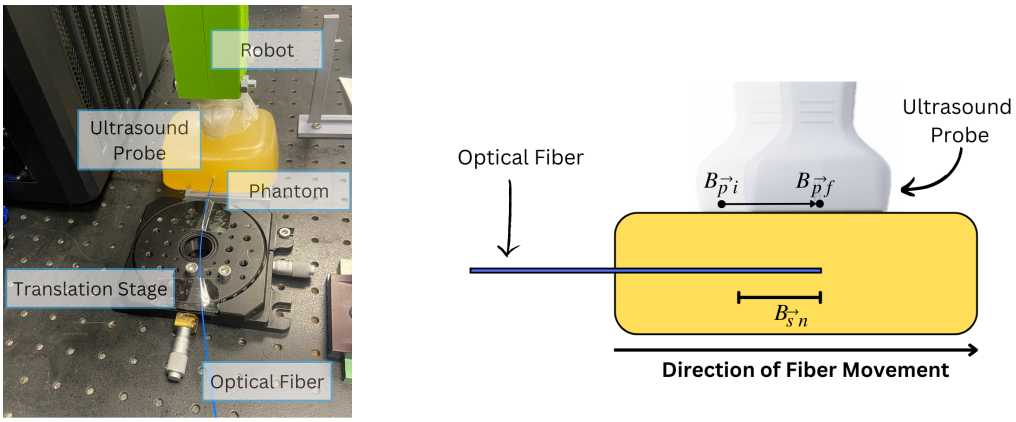
To train and validate our deep learning-based photoacoustic point source localization system, we simulated 20,000 photoacoustic channel data frames using the k-Wave[16] toolbox in MATLAB. We simulated a single source of diameter 100 $\mu$m and up to one reflection artifact in each photoacoustic channel data frame using the parameters reported in Table 1. These simulations were performed in a two-dimensional simulation grid with lateral and axial dimensions of 97 mm and 122 mm, respectively. Reflection artifacts were created using the method previously presented by our group[13,17] (i.e., waveforms originating from photoacoustic sources were axially downshifted by the euclidean distance between an actual source and the source representing the artifact).The source and reflection artifact corresponding to each raw channel data frame were multiplied by object intensity multipliers randomly sampled from Table 1 and added together. Gaussian noise was then added to the resulting matrix using the `addNoise` function in the k-Wave toolbox[16] to form a raw photoacoustic channel data frame. As with previous implementations of phased array transducer-based point source localization systems,[13,18,19] each raw channel data frame was zero-padded to match the FOV of a scan converted photoacoustic image to form a zero-padded channel data frame of dimensions $565 \times 926$ pixels. These zero-padded channel data frames were annotated using the method presented by Gubbi *et al.*[13] with class information (i.e., "source" or "artifact") and bounding boxes of dimensions $32 \times 16$ pixels centered on the positions of sources and artifacts to form annotated images. The totality of annotated images were separated into training (80%) and validation (20%) datasets.

## 2.3 Training and Validation Procedures

The Faster R-CNN network forming our point source localization system was initialized with pre-trained weights from the ImageNet dataset[20] and fine-tuned for 20 epochs with a batch size of 4 and an initial learning rate of $1 \times 10^{-3}$ on two NVidia (Santa Clara, California) Titan X (Pascal) GPUs. This fine-tuning process was performed using the training dataset described in Section 2.2 and the Detectron2 software package.[10] 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. The fine-tuned network was then validated offline on the simulated validation dataset using the process described by Gubbi *et al.*[13]

Table 1. 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 |
| Channel SNR [dB] | -5 | 2 | random |
| Object Intensity (multiplier) | 0.75 | 1.1 | random |




(a) (b)

Figure 3. (a) Photograph of experimental setup and (b) schematic diagram of fiber tip tracking experiment, showing direction of fiber movement with respect to ultrasound probe.

### 2.4 Fiber Tip Tracking Experiment

An experiment was conducted to estimate the tip-tracking error of the ROS2-based photoacoustic visual servoing system, with the experimental setup shown in Fig. 3. At the beginning of the experiment, the optical fiber was inserted into the plastisol phantom and the translation stage was set to 0 mm. The probe was placed on the phantom with the lateral dimension of the probe approximately aligned with the length of the fiber (to visualize as much of the intended trajectory of the optical fiber tip as possible). The probe was then centered over the fiber tip. The translation stage advanced the fiber tip in the lateral dimension of the probe, in 2 mm increments, with a total travel distance of 10 mm. The visual servoing system was employed to center the probe over the fiber tip with each fiber tip displacement increment. Three trials were conducted for each lateral displacement, resulting in 15 total trials. The tip tracking error $e$ was calculated using the equation:

$$e = \left\| {}^{B}\hat{p}_f - {}^{B}\hat{p}_i - {}^{B}\hat{s}_n \right\|, \tag{1}$$

where ${}^{B}\hat{p}_i$ and ${}^{B}\hat{p}_f$ are the initial and final robot end effector positions, respectively, , and ${}^{B}\hat{s}_n$ is the measured displacement of the fiber tip. The mean and standard deviation of the errors for each lateral displacement were calculated. These results were compared to measurements obtained when this experiment was previously performed with our ROS-based visual serving system.[7]

## 3. RESULTS

### 3.1 Validation with Simulated Dataset

Table 2 reports the detection performance (i.e., precision, recall, F1 scores, misclassification rates, and missed detection rates) of the point source localization system when applied to the simulated dataset. The precision, recall, and F1 scores indicate similarly high performance for both sources and artifacts. The system missed 3.36% more artifacts than sources (i.e., 10.94% vs. 7.58%). In addition, the system was robust against misclassification errors for both sources and artifacts. The system also achieved lateral, axial, and Euclidean localization errors of $0.65 \pm 0.74$, $0.24 \pm 0.22$, and $0.72 \pm 0.75$, respectively, for detected sources.

### 3.2 Experimental Tracking Performance

Fig. 4 shows the mean and standard deviation of tip-tracking errors for the previous ROS1-based system and the current ROS2-based system. The former system produced tracking errors ranging 0.59-1.03 mm with a mean error of 0.82 mm across the lateral displacements tested, based on results acquired during a previously conducted experiment.[7] The updated system produced tracking errors ranging 0.35-0.98 mm with a mean error of 0.65 mm across the lateral displacements.

Table 2. Detection performance for sources and artifacts achieved by point source localization system on simulated validation dataset

| Performance Metric | Sources | Artifacts |
|---|---|---|
| Precision [%] | 95.65 | 98.60 |
| Recall [%] | 92.22 | 88.06 |
| F1 score [%] | 93.91 | 93.04 |
| Misclassifications [%] | 0.20 | 0.06 |
| Missed detections [%] | 7.58 | 10.94 |

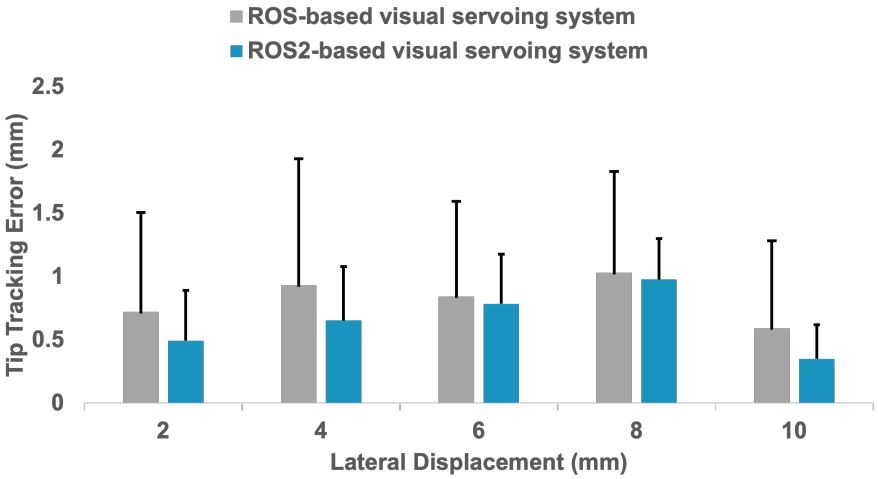



Figure 4. Mean needle tip tracking errors as functions of the lateral shift for the previous[7] and updated visual servoing systems. The black error bars represent the standard deviation of each set of measured errors.

## 4. DISCUSSION

The results presented herein evaluate the benefits of using ROS2 and Detectron2 to create a deep learning photoacoustic visual servoing system. The main advantage seems to be lower tip tracking errors across a range of displacements, when compared to the previous system.[7] In particular, the reduced mean fiber tip tracking errors with the ROS2-based system (i.e., 0.82 mm) can be compared to the tool tip tracking errors obtained with the ROS-based system (i.e., 0.98 mm).[7] The overall mean improvement with the ROS2-based visual servoing system translates to a 23.53% reduction in fiber tip tracking errors in the phantom tissue.

Future possible improvements to the proposed deep learning visual servoing system include increasing the number of degrees of freedom of the robot end effector motion and using 3D simulated data to train the neural network (rather than 2D data). Regarding the degrees of freedom of the end effector, the nominal motion of the robot end effector is limited to one dimension in our visual servoing system, and a second dimension is used to search for and find the tool tip when it is not in the imaging plane of the probe. Although these two degrees of freedom are sufficient to achieve visual servoing, future work will explore the extent to which additional degrees of freedom are necessary to achieve more complicated path planning outcomes. Regarding the use of 3D simulated data,given the greater similarity to experimental photoacoustic data, a 3D dataset is anticipated to both provide more accurate position estimates and enable localization of out-of-plane tool tips.

## 5. CONCLUSION

This work is the first to integrate of ROS2, Detectron2, and 2D simulated photoacoustic channel data with a deep learning-based photoacoustic-robotic visual servoing system. The ROS2-based system is more accurate (e.g., 0.35-0.98 mm needle tracking errors) when compared to the former ROS-based system (which produced tracking errors of 0.59-1.03 mm).[7] This improvement enables more effective, real-time, tip tracking of needles, catheters, and other surgical tools that are similarly essential to surgical and interventional procedures.

## ACKNOWLEDGMENTS

This work was completed in partnership with the Computational Sensing and Medical Robots Research Experience for Undergraduates (NSF EEC-1852155). Financial support was provided by NSF SCH Award IIS-2014088, NSF CAREER Award ECCS-1751522, NIH Trailblazer Award R21 EB025621, and NSF Alan T. Waterman Award IIS-2431810.

## REFERENCES

[1] Pourtaherian, A., Ghazvinian Zanjani, F., Zinger, S., Mihajlovic, N., Ng, G. C., Korsten, H. H., and de With, P. H., "Robust and semantic needle detection in 3d ultrasound using orthogonal-plane convolutional neural networks," *International journal of computer assisted radiology and surgery* **13**, 1321–1333 (2018).

[2] Groves, L. A., VanBerlo, B., Peters, T. M., and Chen, E. C., "Deep learning approach for automatic out-of-plane needle localisation for semi-automatic ultrasound probe calibration," *Healthcare technology letters* **6**(6), 204–209 (2019).

[3] Lediju, M. A., Pihl, M. J., Dahl, J. J., and Trahey, G. E., "Quantitative assessment of the magnitude, impact and spatial extent of ultrasonic clutter," *Ultrasonic Imaging* **30**(3), 151–168 (2008).

[4] Xu, M. and Wang, L. V., "Photoacoustic imaging in biomedicine," *Review of scientific instruments* **77**(4) (2006).

[5] Bell, M. A. L., "Photoacoustic imaging for surgical guidance: principles, applications, and outlook," *Journal of Applied Physics* **128**(6), 060904 (2020).

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

[7] Gubbi, M. R. and Bell, M. A. L., "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).

[8] Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., and He, K., "Detectron." https://github.com/facebookresearch/detectron (2018).

[9] Thomas, D., "Changes between ros 1 and ros 2," *ROS2 Design* (2017).

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

[11] Graham, M., Assis, F., Allman, D., Wiacek, A., González, E., Gubbi, M., Dong, J., Hou, H., Beck, S., Chrispin, J., and Bell, M. A. L., "In vivo demonstration of photoacoustic image guidance and robotic visual servoing for cardiac catheter-based interventions," *IEEE Transactions on Medical Imaging* **39**(4), 1015–1029 (2020).

[12] Allman, D., Reiter, A., and Bell, M. A. L., "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Transactions on Medical Imaging* **37**(6), 1464–1477 (2018).

[13] Gubbi, M. R., Assis, F., Chrispin, J., and Bell, M. A. L., "Deep learning in vivo catheter tip locations for photoacoustic-guided cardiac interventions," *Journal of Biomedical Optics* **29**(S1), S11505 (2023).

[14] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems* **28** (2015).

[15] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[16] Treeby, B. E. and Cox, B. T., "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of Biomedical Optics* **15**(2), 021314 (2010).

[17] Allman, D., Assis, F., Chrispin, J., and Bell, M. A. L., "Deep neural networks to remove photoacoustic reflection artifacts in ex vivo and in vivo tissue," in [*Proceedings of the IEEE International Ultrasonics Symposium (IUS)*], 1–4, IEEE (2018).

[18] Allman, D., Assis, F., Chrispin, J., and Bell, M. A. L., "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)*], 1–3, IEEE (2019).

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

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