Pre-print, to be published in the International Journal of Computer Assisted Radiology and Surgery

Transformer-based Robotic Ultrasound 3D Tracking for Capsule Robot in GI Tract

Xiaoyun Liu[1], Changyan He[2], Mulan Wu[1], Ann Ping[1], Anna Zavodni[3], Naomi Matsuura[4,5], Eric Diller[1,5,6*]

[1]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada.
[2]Discipline of Medical Engineering, University of Newcastle, Newcastle, NSW, Australia.
[3]Department of Medicine, Division of Cardiology, University of Toronto, Toronto, ON, Canada.
[4]Department of Materials Science and Engineering, University of Toronto, Toronto, ON, Canada.
[5]Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada.
[6]University of Toronto Robotics Institute, University of Toronto, Toronto, ON, Canada.

*Corresponding author(s). E-mail(s): ediller@mie.utoronto.ca;

Abstract

Purpose: Ultrasound (US) imaging is a promising modality for real-time monitoring of robotic capsule endoscopes navigating through the gastrointestinal (GI) tract. It offers high temporal resolution and safety but is limited by a narrow field of view, low visibility in gas-filled regions, and challenges in detecting out-of-plane motions. This work addresses these issues by proposing a novel robotic ultrasound tracking system capable of long-distance 3D tracking and active re-localization when the capsule is lost due to motion or artifacts. **Methods**: We develop a hybrid deep learning-based tracking framework combining convolutional neural networks (CNNs) and a transformer backbone. The CNN component efficiently encodes spatial features, while the transformer captures long-range contextual dependencies in B-mode US images. This model is integrated with a robotic arm that adaptively scans and tracks the capsule. The system's performance is evaluated using ex-vivo colon phantoms under varying imaging conditions, with physical perturbations introduced to simulate realistic

clinical scenarios. **Results**: The proposed system achieved continuous 3D tracking over distances exceeding 90 cm, with a mean centroid localization error of 1.5 mm and over 90% detection accuracy. We demonstrated 3D tracking in a more complex workspace featuring two curved sections to simulate anatomical challenges. This suggests the strong resilience of the tracking system to motion-induced artifacts and geometric variability. The system maintained real-time tracking at 9–12 FPS and successfully re-localized the capsule within seconds after tracking loss, even under gas artefacts and acoustic shadowing. **Conclusion**: This study presents a hybrid CNN-transformer system for automatic, real-time 3D ultrasound tracking of capsule robots over long distances. The method reliably handles occlusions, view loss, and image artefacts, offering millimeter-level tracking accuracy. It significantly reduces clinical workload through autonomous detection and re-localization. Future work includes improving probe-tissue interaction handling and validating performance in live animal and human trials to assess physiological impacts.

Keywords: Capsule Robots, Robotic Ultrasound 3D Tracking System, Ultrasound Imaging, Deep Learning

1 Introduction

Minimally-invasive capsule-based technology is a promising alternative to standard endoscopic procedures for the diagnosis and treatment of diseases within the GI tract. Traditional endoscopic methods typically require anesthesia and poses the risk of bowel rupture, while allowing for very limited access to the deepest portions of the small intestine [1]. Capsule-based technology remains a promising means of effectively evaluating the small bowel, however, manipulating robot agents remotely to sample the intestinal-contents requires real-time and accurate tracking of the robot's location within the GI tract. Such capsule tracking remains an open research challenge. Extensive research has been conducted on magnetic localization technology [2]. However, the effective operating distance between the workspace and the sensor arrays is often constrained by the limited range of magnetic fields that diminish over distance. Additionally, magnetic tracking lacks the capability to localize the capsule position with respect to the anatomy of the GI tract, which is required for targeted sampling or drug delivery. US-based capsule tracking would overcome these barriers using a safe, non-invasive imaging method which can concurrently image the capsule and the surrounding GI tract anatomy in real-time.

Although US imaging is commonly used as a diagnostic tool by skilled sonographers in clinical practice, capsule tracking presents several unique challenges. US imaging is unable to image objects out of the scanning plane and has a limited field of view (FOV) within the scanning plane, which restricts the observable workspace to several millimeters [3]. The highly echogenic and heterogeneous tissue environment creates high-contrast imaging artefacts [4], which can obscure and create a visual resemblance to a pill-shaped capsule. Furthermore, intraluminal gas is highly echogenic and

produces different artefacts whether as bubbles or pockets of gas, which can cause substantial degradation of the ultrasonographic image [5].

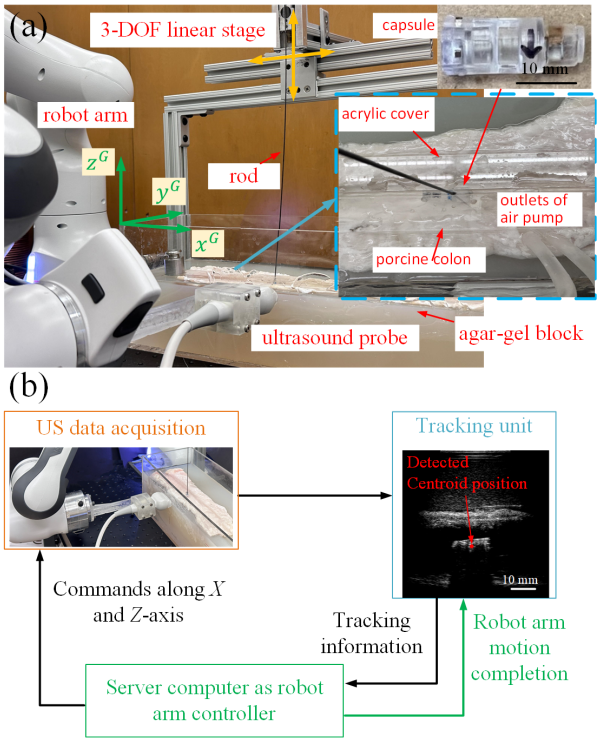

Fig. 1 The automatic robotic US system overview. (a) The proposed robotic US system for 3D tracking of capsule robots, where the inset is the defined channel-like workspace with the embedded porcine colon and simulated intraluminal gas. The yellow arrows represent the motions of the translational stage. The capsule agent is shown at the top corner. The image plane is X^GY^G-plane, the robot arm moves along X^G and Z^G-axis. (b) Robot arm control scheme with the US tracking feedback in a closed loop, which allows for 3D scanning of the workspace and real-time US feedback using a B-mode scanner.

One solution which allows for direct 3D visualization of target agents is the use of a 3D US probe which generates a volume visualization directly. Stationary 3D probe-based tracking of biopsy needles [6, 7] used deep learning-based semantic segmentation methods to detect needle position. However, in these works, the probe was fixed during scanning, which still renders a limited imaging area and is not feasible for tracking moving capsule robots over the entire GI tract. Thus, for tracking robotic agents over the longer GI tract, a portable, 2D probe is superior to 3D probes due to its higher speed and lower data processing requirements. For real-time tracking, a robot arm can be used to automatically move the US probe in response to motion of the robot agent. Although state-of-the-art robotic US tracking systems [8–10] achieved over 400 mm long-distance 3D tracking of microrobots, these methods were validated in less physiologically representative environments (i.e., in-vitro silicon and gelatin phantoms)

Table 1 System-level comparison with SOTA untethered microrobot ultrasound tracking works

Reference	Imaging mode	Tracking method	Imaging environment	Dimension/ Tracking distance (mm)	Tracking accuracy (mm)
Oliveria *et. al* [13]	2D B-mode imaging	Learning based	Non-tissue (water) environment	2D tracking/ 40-60	1.59
Yang *et. al* [8]	2D B-mode imaging	Non-learning based	In-vitro silicon phantom of human artery	3D tracking/ 400-500	Not provided
Pane *et. al* [11]	Ultrasound RF data	Non-learning based	In-vitro silicon phantom of a medium artery	2D tracking/ 80	0.368
Lu *et. al* [10]	2D B-mode and Doppler imaging	Non-learning based	Non-tissue (water) environment	2.5D tracking/ 170	8.12
Du *et. al* [9]	2D B-mode imaging	Non-learning based	In-vitro gelatin phantom	3D tracking/ Over 2000	Not provided
This work	2D B-mode imaging	Learning based	Ex-vivo tissue environment with air	3D tracking/ Over 900	1.46

instead of in more complex, heterogenous, tissue environments. One notable approach [11] utilized the radio-frequency (RF) data and US-acoustic phase analysis (APA) detection technique in a closed-loop visual-servoing system and demonstrated high tracking accuracy of a microrobot in sub-millimeter in a tissue-mimicking phantom. However, this method can only track 2D position of the microrobot without considering the out-of-plane motion and was not able to provide an automatic detection of the capsule in the search mode.

Therefore, robust long-distance 3D tracking in tissue environments where loss-of-view of the capsule robot is common, remains an open challenge. To fill the research gap, we propose an automatic robotic ultrasound tracking system, as shown in Figure 1. Our contributions include: (1) Our proposed tracking approach is the first to provide fully automatic detection, long-distance accurate 3D tracking, and search of capsule robots using 2D B-mode imaging in a physiologically representative tissue environment; (2) leveraging the approach of large language models (LLM) and proposing a hybrid CNN and transformer-combined deep learning method for automatic detection (capsule detected or lost) and localizing the centroid position of capsule robots in *ex-vivo* porcine GI tract with intraluminal gas; (3) enhancing the US-guided 3D tracking performance using conventional B-mode imaging in a clinically representative imaging environment to reduce the gap between the tracking approach validated in the lab environment and real clinical applications as summarized in Table 1.

2 Results

2.1 System Overview and Experimental Setup

The proposed robotic ultrasound tracking system and the control scheme for mobile 2D US-enabled 3D tracking of capsule robots are shown in Figure 1. The linear array transducer (L15-7H40-A5) is attached to the 7-DOF robot arm (Franka Emika) via a

fixture for stably fastening the probe to the robot arm end-effector and is positioned to image the capsule from the side of the tank. Two computers work in parallel, in which the client computer is connected to the US system (Telemed ArtUs EXT-1H) which accesses raw B-mode frames in real time with a frequency of 40 Hz and runs the tracking algorithm to provide in-plane position and state estimations (capsule detected or lost) of the capsule robot. The tracked information is sent to the server computer as feedback for actuating the robot arm to perform a 2-DOF translational motion along the x and z-axis of the global coordinate frame G via the bilateral User Datagram Protocol (UDP). The US system operates in the standard B-mode imaging with an imaging depth and frequency of 70 mm and 7.5 MHz, respectively, which are appropriate settings in clinical abdominal US imaging [12]. The rod used to connect the capsule is a low-stiffness Nitinol rod with a length of 50 cm and a diameter of 0.5 mm. One end of the rod is attached to the linear stage that generates the 3D motions, while the other end is attached to the capsule. We assume the out-of-plane orientation is negligible.

The anatomically-representative imaging setup (Figure 1(a)) was constructed by placing a porcine colon tissue sample inside of an agar gel block, which was placed in a 50 $cm(L) \times 15$ $cm(W) \times 15$ $cm(H)$ acrylic tank. The channel with a length $a = 40$ cm, width $b = 30$ mm, and depth of 50 mm was designed to ensure that the capsule moves within the valid imaging depth range of 60-80 mm with sufficient space for out-of-plane motions. To replicate clinical GI ultrasonography, we used an anechoic non-absorbable polyethylene glycol (PEG) solution mixed with digestive enzymes powder (Webber Naturals) to fill the channel-like workspace. Two outlets of the air pump were placed inside the workspace to produce the simulated intraluminal gas continuously. The acrylic cover has a length of 45 cm and a 1.5 mm wide slot to provide space for the rod to move, which was used to facilitate trapping large air pockets. The capsule with a diameter of 8 mm and length of 21 mm consists of three permanent magnets encapsulated in a plastic sample chamber.

2.2 *Ex-vivo* Long-distance 3D Tracking of the Capsule Robot with Air

We demonstrated the 3D tracking in a section of the porcine colon that exhibits alternating echogenicity in B-mode images due to the layered structure of the colon wall and the presence of air-fluid interfaces.

First, we demonstrated the in-plane long-distance tracking where the capsule was maneuvered to move along the channel without any out-of-plane motion. The total tracking distance was around 90 cm with 1339 tracking frames. The proposed system detected and localized the capsule in all frames with a 0.31 ± 0.25 mm which corresponds to 1.5% of the capsule's body length. The tracked positions with the US frames at three different locations in the workspace and the centroid tracking error distribution are shown in Figure 2(a) and (b), respectively.

Second, we conducted *ex-vivo* 3D tracking to replicate the capsule operations in real clinical procedures. In the first trial, the capsule traveled for a round-trip in the workspace with air and several random motions at different locations, which caused the view loss. The tracking system detected it as lost and switched to the searching mode.

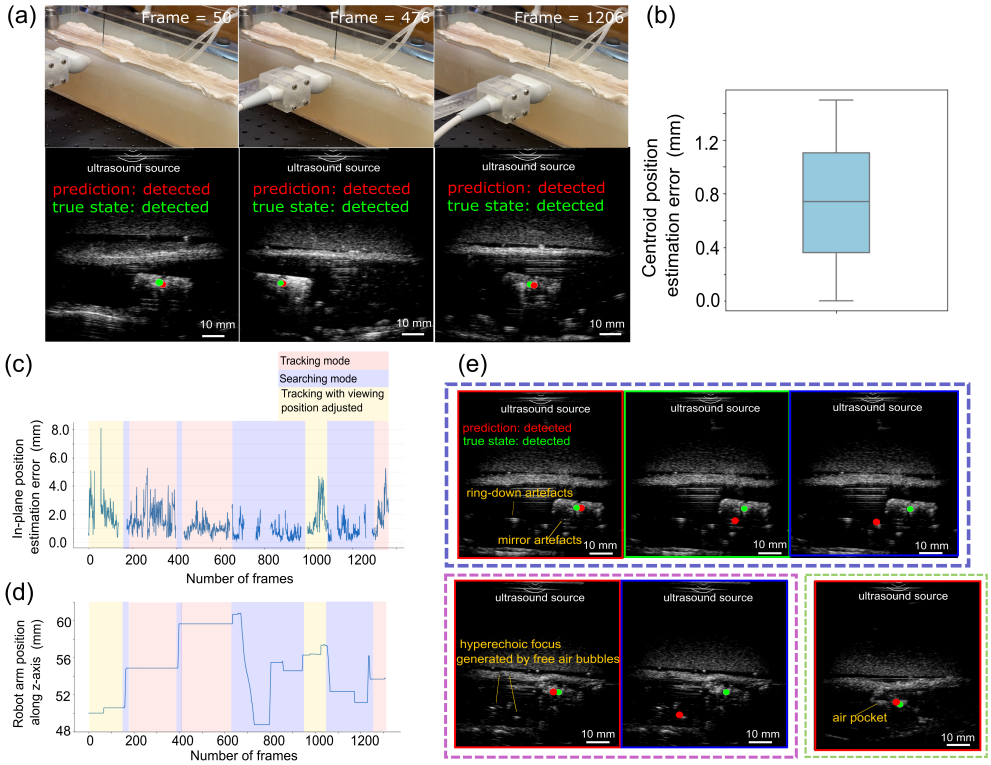

Fig. 2 *Ex-vivo* real-time 3D tracking of the capsule robot in porcine colons. (a) shows the tracked positions with US frames at three different locations during the long-distance in-plane tracking with the centroid position error distribution in (b). (c) and (d) plot the in-plane tracking error and the robot arm position perpendicular to the image plane in the first trial of the 3D tracking experiments, respectively. The slopes in the figure represent the search process, during which the capsule becomes misaligned with the image plane, resulting in view-loss. The horizontal segments depict the tracking process, where the image plane aligns with the capsule, corresponding to the capsule's out-of-plane positions along the z-axis. (e) demonstrates the comparison of the retrained ResVit model (denoted by red frames) with non-retrained ResVit (green) and the ResNet with channel attention (blue) for tracking the capsule in *ex-vivo* porcine colon in three sample test images (denoted by the boxes with dotted lines in different colors) with different air-related artefacts (purple), the capsule in close proximity to the tissue boundary (pink) and is obscured due to clusters of air bubbles, and is occluded by the large air pocket (green). The retrained ResVit mode can provide accurate tracking in all the scenarios while the other two models either provide biased localization or fail to detect the capsule.

Once the capsule was detected and the tracking was resumed, the system switched back to the tracking mode automatically. The mean centroid estimation error of in-plane tracking is 1.46±0.86 mm, which corresponds to 6.9% of the capsule body length. The robot arm position along z-axis with working mode switches and the in-plane tracking error in each frame are plotted in Figure 2(c) and Figure 2(d), respectively.

Last, we assessed the 3D tracking of the capsule with an initial search without *a-priori* knowledge of the capsule state. To test the robustness and searching capability of the proposed tracking method, the capsule moved randomly perpendicular to the

image plane at the air region, where the pump speed was increased to generate a higher concentration of air bubbles to simulate the conditions of air-intensive regions in the GI tract. The system demonstrated prolonged and reliable detection and tracking, as shown in Figure 2(e).

The initial search involved 100–200 frames during which the capsule was not within the field of view of the ultrasound probe. The neural network model demonstrated robustness against imaging artifacts and noise, achieving 100% detection accuracy once the capsule appeared. Similar to the relocalization scenario, the capsule was occasionally lost due to out-of-plane motion. Relocalization was performed automatically by the robot and also achieved 100% detection accuracy. These results highlight the robustness and reliability of the proposed robotic tracking system.

Besides these results, we demonstrated the 3D tracking experiment in the workspace with a more complex profile, in which the channel-like workspace was designed with two curved sections, each having a diameter of 3.5 cm and lengths of 8 cm and 16 cm, respectively to simulate the curves of colons. The imaging depth was adjusted from 70mm to 80mm, as the curve made the depth deeper. The proposed system was able to track the capsule accurately and was robust to the varying imaging conditions. Furthermore, the model was robust and detected the capsule as lost when encountering the lack of acoustic windows. The searching continued until acoustic windows came back. The complete tracking trials can be found in the supporting videos.

We also conducted both the deep learning-based model-level (Supplementary Materials) and system-level comparison (Table 1) of the proposed tracking system with state-of-the-art works on automatic ultrasound tracking of untethered microrobots. The results in Supplementary Materials suggest that best performance is attained by the fine-tuned ResVit model (adopted in this work) by a large margin over the two other baseline models: ResNet-50 with channel attention and the non-fine-tuned ResVit model..

3 Automatic Robotic Ultrasound 3D Tracking Using Mobile 2D Ultrasound

3.1 Transformer-based Deep Learning Method for In-Plane Tracking of Capsule Robot

Convolutional neural networks (CNN) have been successfully used in detection and tracking of needles, catheters [7, 13] and microrobots in US images [14]. CNN has natural inductive bias and translation invariance for learning local features in images, while having difficulty capturing contextual information [15]. The attention mechanism built in the transformer [16] provides a global receptive field and enables the model to capture long-range relationships more efficiently. Transformers are also more robust to common corruptions and perturbations, such as noise, occlusions, and contrast variation [17]. Therefore, we integrate CNN with transformer and propose a hybrid network that has the capability for handling with feature misalignment and occlusion issues.

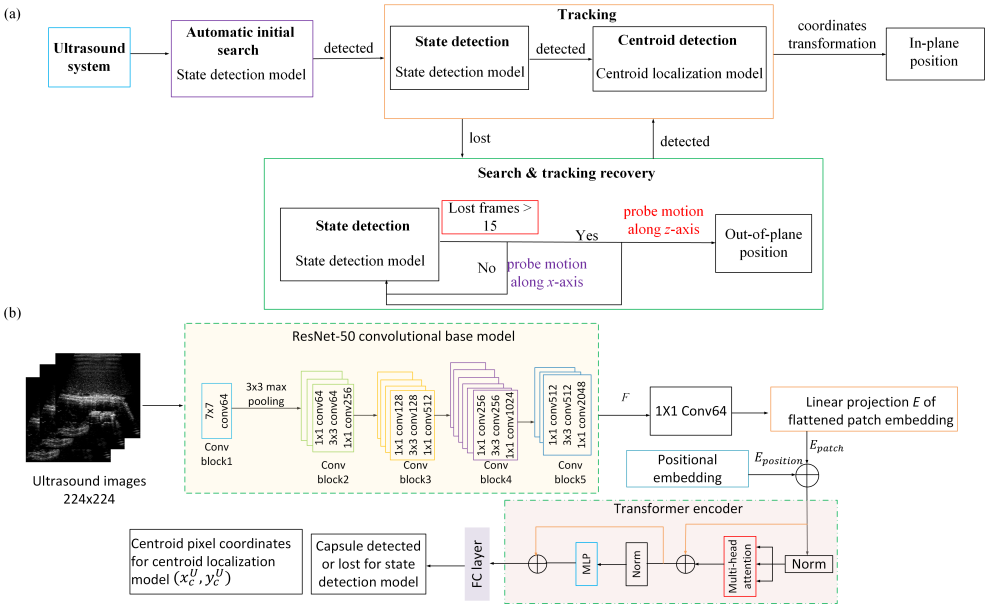

Fig. 3 (a) The workflow of the proposed transformer-based robotic ultrasound 3D tracking of the capsule. (b) The hybrid ResNet and Vision Transformer (ResViT) deep neural networks model for state detection and centroid localization of the capsule robot.

3.1.1 Network Architecture

In this work, a hybrid ResNet [18] and Vision Transformer [19] (ResViT)-based deep learning method is proposed to detect the capsule robot state (detected or lost) and track the capsule in-plane position via centroid detection. The ResViT model (Figure 3(b)) consists of a pre-trained ResNet-50 backbone for extracting low-level features, a transformer encoder to facilitate long-range spatial dependencies across the entire image locations, and an output head to predict either the centroid pixel coordinates (x_c^U, y_c^U) in the US image or the capsule state as detected or lost.

Given an input image $I \in R^{H_I \times W_I \times C_I}$, the CNN feature extraction component outputs a 2D spatial feature map $F \in R^{H_F \times W_F \times C_F}$. A 1×1 convolutional layer is applied to transform the CNN feature map to $F \in R^{H_F \times W_F \times D_F}$, in which $D_F = 64$ is the constant latent vector size used in the transformer encoder through all its layers. Then the transformed feature map is flattened into a sequence of patch embeddings E_{patch} with a trainable linear projection $[F_{p1}E; F_{p2}E; \ldots; F_{pN}E]$ where F_{pi} denotes the i^{th} image patch in the feature map F and $E \in R^{(H_F \cdot W_F) \times D_F}$ is the linear projection layer. To retain each position of the patch embeddings in the feature map, the 1D learnable position embedding is applied to obtain position embeddings $E_{position} \in R^{(H_F \cdot W_F) \times D_F}$, which are added to the patch embeddings $[F_{p1}E; F_{p2}E; \ldots; F_{pN}E] + E_{position}$.The resulting embedding sequence is the input to the transformer encoder with $N = 4$ encoder layers, in which each encoder layer consists of a multi-head self-attention sub-layer, a fully-connected (FC) sub-layer with

a GELU activation function, a residual connection around each of the two sub-layers followed by layer normalization. The self-attention mechanism allows each position in the encoder layer to attend all other locations in the input sequence by computing the attention scores $Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V$ is used to determine contributions of different image locations for making the prediction, where Q, K, V are the query, key, and value matrices obtained by mapping the input sequence through a linear layer with weights W_q, W_k, and W_v. Multi-block of self-attention modules are concatenated for paying attention to different regions of the image. We used a multi-head of self-attention layer with 4 hidden blocks. A feed-forward layer that predicts the centroid position coordinates is attached to the transformer encoder with an output $(H_F \cdot W_F) \times D_F$ in the centroid estimation model, while a classification output head with 2 classes is attached to the encoder in the state detection model.

3.1.2 Model Training and Fine-tuning

We used a custom *ex-vivo* dataset [20] that was generated in *ex-vivo* porcine stomachs with a different capsule robot [21] in our earlier work for training the ResVit model. Details of data collection and generation can be found in [20]. During training, we split the two training datasets into train, validation, and test set with a ratio of 7:2:1. Two models were trained separately for the capsule state classification and centroid position estimation. To further enhance the tracking accuracy and robustness of the deep neural network models with the existence of unseen artefact patterns in the original training data, the trained model was fine-tuned using a small-scale *ex-vivo* porcine colon dataset with a size of 5200 B-mode images that includes air-related artefacts and mirror artefacts caused by the bouncing of acoustic waves within the hollow cavity of the new capsule.

The real-time US image frames are processed on a GPU (NVIDIA GeForce RTX 2070). The system's processing speed is evaluated by calculating the neural model's processing time per frame (input to prediction), which is equivalently 9-12 fps. The latency of the data transmission between the robot arm controller and the US system is negligible compared to the model processing time.

3.2 Robotic 3D Tracking Strategy

As shown in Figure 3, the tracking process starts with an automatic initial search without any *a-priori* knowledge of the capsule position and state. The capsule moves in the workspace with an average in-plane speed of 1 mm/s by manually maneuvering the 3-DOF stage along and perpendicular to the lateral direction of the agar tank. The system operates in two working states: a) tracking state, b) search and recovery tracking state. The state detection model first detects the capsule, if the capsule is detected, the in-plane centroid detection model then provides real-time centroid localization, which is transformed to the workspace coordinate frame G and used for actuating the robot arm with the controller showed in Eq. (1) to align the image plane with the capsule for real-time tracking. The search and recovery tracking state allows for the mobility of the US probe to scan the workspace orthogonal to the image plane and thus adjusting the image plane to coincide with the capsule. Once the tracking

is recovered and the capsule is detected for consecutive n frames, the working state automatically switches back to the tracking state. The initial search at the start of the tracking is performed manually. An operator holds the robotic US probe to scan the workspace until the model outputs a prediction on the capsule's position. The average time in this procedure is around 40 seconds. Re-localization if the capsule is lost during tracking is performed automatically by the robotic arm, with an average re-localization search time of 1.5 minutes.

The robot arm is controlled to perform searching within the range of $[-30, +40]$ mm along the z-axis of the workspace. The capsule's centroid position (x_c^U, y_c^U) in the US image pixel coordinate frame U is transformed to the global coordinate frame G and used to calculate the robot velocity:

$$\begin{cases} \dot{x}_c^G = \begin{cases} s\alpha(x_c^U - x_m^U), & \text{if } \dot{z}_c^G = 0 \\ 0, & \text{if } \dot{z}_c^G > 0 \end{cases} \\ \dot{y}_c^G = 0 \\ \dot{z}_c^G = \begin{cases} sgn(\delta t)v, & \text{if OutPlaneCnt}> n \\ 0, & \text{otherwise} \end{cases} \end{cases} \tag{1}$$

where s is the scale factor from the US image pixel to mm, α is a control parameter for controlling the in-plane motion of the robot arm. x_m^U is the X coordinate of the US image's geometrical centroid point. v is a preset value for controlling the robot upward/downward. $sgn(\delta t) = \begin{cases} 1, \delta t < t_c \\ -1, \delta t \geq t_c \end{cases}$, δt is the robot's motion accumulated time at each up/down searching loop. t_c is a preset value for changing the robot's moving direction. We assume that the side wall of the tank is parallel to the XZ plane of the global frame, such that the robot is motionless in the Y direction.

4 Discussion and Conclusion

This paper proposes a hybrid CNN and transformer-based automatic robotic ultrasound system for long-distance US-guided 3D tracking of capsule robots in tissue environments using 2D B-mode imaging. Experimental results manifest that the proposed method can automatically detect the existence of the capsule and reliably track the capsule for over 90 cm while addressing view-loss of the capsule and recovering tracking of the lost capsule in 3D space. Our work focuses on robotic capsule applications for microbial sampling within the GI tract, which do not necessarily demand very high localization accuracy for the robotic agents. However, from a control perspective, it is desirable for the robot to be controlled with motion accuracy on the order of several millimeters. The proposed system provides a fully automatic detection, tracking, and search of capsule robots, which largely reduces the workload of physicians. Furthermore, the attention mechanism of the transformer allows for capturing long-range dependencies across the image and localizing the occluded and obscured capsule. The fine-tuned ResVit model achieves high detection and localization accuracies on both unseen test data and new imaging scenarios with varying imaging parameters through

a small-scale fine-tuning. In our future work, we will improve the scanning mechanism to accommodate external pressures exerted by the probe when it is pushed against the tissues during imaging. We will also validate our system in *in-vivo* environment of both live animals and human that can account for acoustic properties of interior fluids and tissue contractions.

References

[1] Otuya, D.O., Verma, Y., Farrokhi, H., Higgins, L., Rosenberg, M., Damman, C., Tearney, G.J.: Non-endoscopic biopsy techniques: a review. Expert Review of Gastroenterology & Hepatology **12**(2), 109–117 (2018) https://doi.org/10.1080/17474124.2018.1412828

[2] Hu, C., Meng, M.Q., Mandal, M.: Efficient magnetic localization and orientation technique for capsule endoscopy. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp. 628–633. IEEE, Edmonton (2005). https://doi.org/10.1109/IROS.2005.1545490

[3] Wang, Q., Yang, L., Yu, J., Chiu, P.W.Y., Zheng, Y.P., Zhang, L.: Real-Time Magnetic Navigation of a Rotating Colloidal Microswarm under Ultrasound Guidance. IEEE Transactions on Biomedical Engineering **67**(12), 3403–3412 (2020) https://doi.org/10.1109/TBME.2020.2987045

[4] Kimmey, M.B., Martin, R.W., Haggitt, R.C., Wang, K.Y., Franklin, D.W., Silverstein, F.E.: Histologic correlates of gastrointestinal ultrasound images. Gastroenterology **96**(2 PART 1), 433–441 (1989) https://doi.org/10.1016/0016-5085(89)91568-0

[5] Nasir, A.I.: Artifact in the Image of Ultrasound. Australian Journal of Basic and Applied Sciences **12**(12), 131–143 (2018) https://doi.org/10.22587/ajbas.2018.12.12.21

[6] Pourtaherian, A., Ghazvinian Zanjani, F., Zinger, S., Mihajlovic, N., Ng, G.C., Korsten, H.H.M., With, P.H.N.: Robust and semantic needle detection in 3D ultrasound using orthogonal-plane convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery **13**(9), 1321–1333 (2018) https://doi.org/10.1007/s11548-018-1798-3

[7] Yang, H., Shan, C., Kolen, A.F., With, P.H.N.: Catheter localization in 3D ultrasound using voxel-of-interest-based ConvNets for cardiac intervention. International Journal of Computer Assisted Radiology and Surgery **14**(6), 1069–1077 (2019) https://doi.org/10.1007/s11548-019-01960-y

[8] Yang, L., Zhang, M., Yang, Z., Yang, H., Zhang, L.: Mobile Ultrasound Tracking and Magnetic Control for Long-Distance Endovascular Navigation of Untethered Miniature Robots against Pulsatile Flow. Advanced Intelligent Systems **4**(3), 2100144 (2022) https://doi.org/10.1002/aisy.202100144

[9] Du, X., Wang, Q., Jin, D., Chiu, P.W.Y., Pang, C.P., Chong, K.K.L., Zhang, L.: Real-Time Navigation of an Untethered Miniature Robot Using Mobile Ultrasound Imaging and Magnetic Actuation Systems. IEEE Robotics and Automation Letters **7**(3), 7668–7675 (2022) https://doi.org/10.1109/LRA.2022.3184445

[10] Lu, Y., Zhao, H., Becker, A.T., Leclerc, J.: Steering Rotating Magnetic Swimmers in 2.5 Dimensions Using only 2D Ultrasonography for Position Sensing. IEEE Robotics and Automation Letters **7**(2), 3162–3169 (2022) https://doi.org/10.1109/LRA.2022.3146560

[11] Pane, S., Faoro, G., Sinibaldi, E., Iacovacci, V., Menciassi, A.: Ultrasound Acoustic Phase Analysis Enables Robotic Visual-Servoing of Magnetic Microrobots. IEEE Transactions on Robotics **38**(3), 1571–1582 (2022) https://doi.org/10.1109/TRO.2022.3143072

[12] Steinsvik, E.K., Hatlebakk, J.G., Hausken, T., Nylund, K., Gilja, O.H.: Ultrasound imaging for assessing functions of the GI tract. Physiological Measurement **42**(2) (2021) https://doi.org/10.1088/1361-6579/abdad7

[13] Yang, H., Shan, C., Kolen, A.F., De With, P.H.N.: Efficient Medical Instrument Detection in 3D Volumetric Ultrasound Data. IEEE Transactions on Biomedical Engineering **68**(3), 1034–1043 (2021) https://doi.org/10.1109/TBME.2020.2999729

[14] De Oliveira, A.J.A., Batista, J., Misra, S., Venkiteswaran, V.K.: Ultrasound Tracking and Closed-Loop Control of a Magnetically-Actuated Biomimetic Soft Robot. IEEE International Conference on Intelligent Robots and Systems **2022-Octob**, 3422–3428 (2022) https://doi.org/10.1109/IROS47612.2022.9981635

[15] Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems (Nips), 4905–4913 (2016)

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **2017-Decem**(Nips), 5999–6009 (2017)

[17] Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding Robustness of Transformers for Image Classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10211–10221 (2021). https://doi.org/10.1109/ICCV48922.2021.01007

[18] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. CoRR **abs/1512.03385** (2015)

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby,

N.: An Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. In: ICLR 2021 - 9th International Conference on Learning Representations (2020). http://arxiv.org/abs/2010.11929

[20] Liu, X., Esser, D., Wagstaff, B., Zavodni, A., Matsuura, N., Kelly, J., Diller, E.: Capsule robot pose and mechanism state detection in ultrasound using attention-based hierarchical deep learning. Scientific Reports **12**(1) (2022) https://doi.org/10.1038/s41598-022-25572-w

[21] Shokrollahi, P., Lai, Y.P., Rash-Ahmadi, S., Stewart, V., Mohammadigheisar, M., Huber, L.A., Matsuura, N., Zavodni, A.E.H., Parkinson, J., Diller, E.: Blindly Controlled Magnetically Actuated Capsule for Noninvasive Sampling of the Gastrointestinal Microbiome. IEEE/ASME Transactions on Mechatronics **26**(5), 2616–2628 (2021) https://doi.org/10.1109/TMECH.2020.3043454