(017)

Stereo-Vision Localization of Millimeter-Scale Capsule Targets in a Clinically Inspired Maze: A Repeatability Study

I. Frajtag [1], L. Masjosthusmann [2], S. Misra [2] and F. Šuligoj [1]
[1] University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lučića 5, 10000 Zagreb, Croatia
[2] University of Twente, Faculty of Engineering Technology, Department of Biomechanical Engineering,
Surgical Robotics Laboratory, Drienerlolaan 5, 7522 NB Enschede, Netherlands
E-mail: ines.frajtag@fsb.unizg.hr

Summary: Reliable localization of magnetically actuated microrobots and capsule robotic systems is essential for safe navigation in complex anatomical pathways, yet remains challenging in tortuous and occlusion-prone environments. This study presents a stereo vision framework that combines YOLO-based detection in two camera views with triangulation to reconstruct metric 3D target center positions. When YOLO-OBB is used, the same framework also provides an orientation estimate that can be translated into a headin cue. The system is evaluated in a maze phantom designed to mimic bowel- and vessel-like pathways, using five targets of different sizes and geometries. On a test set, YOLO-OBB improves detection reliability over standard YOLO (precision/recall 95.6/94.5 % vs. 86.1/85.9 %), maintains comparable center error (4.31 vs. 4.22 px mean), and reducing CPU inference time by half. Repeatability analysis across ten maze points showed lower total 3D error for cylindrical than symmetric targets (E_{total} = 0.91/2.33 vs. 2.40/11.1), indicating higher robustness for navigation-relevant localization in medically relevant settings.

Keywords: Vision system, Microrobots, Detection, Error, Repeatability.

1. Introduction

Over the past decade, robotics has transitioned from specialized systems to a growing presence in routine clinical workflows. Robotic assistants are now used across many surgical disciplines [1-3] to support demanding procedures and improve precision. In parallel, the drive toward minimally invasive diagnosis and therapy has accelerated the development of microrobots and capsule robotic systems that can operate inside hard-to-reach anatomical spaces, including pathways such as the gastrointestinal (GI) tract and vasculature [4]. Among these technologies, magnetically actuated devices are particularly attractive because external magnetic fields can penetrate biological tissue and interact with magnetic components inside the device to generate force and torque locally. This allows the actuation source to remain outside the body, enabling compact and fully internalized device designs. As a result, magnetically actuated microrobots and capsule robots are increasingly explored for applications such as targeted drug delivery, localized therapy, and *in situ* diagnostics [5]. In this context, magnetically actuated microrobots and capsule robotic systems are especially relevant, but their practical deployment strongly depends on reliable localization and orientation tracking to support safe navigation in complex, occlusion prone anatomy. These requirements become especially strict in confined anatomies where visibility is intermittent and trajectories are highly curved. Recent surveys and experimental studies [6, 7] describe multiple localization modalities [8], such as magnetic sensing and magnetic localization, vision-based approaches such as visual odometry/SLAM or learning-based pose estimation from image sequences, as well as hybrid sensor-fusion strategies. Each modality entails practical constraints related to system complexity, susceptibility to artifacts, and achievable spatial and temporal resolution. Hybrid and vision-based solutions are appealing because they can provide high spatial detail using relatively low-cost imaging hardware. Yet robust performance remains difficult under clinically realistic conditions. Tortuous geometry, occlusions, reflections and low-texture scenes can all degrade tracking - especially for millimeter-scale targets.

In this work, we develop and evaluate a stereo-vision localization framework designed for a clinically inspired scenario: a maze phantom that emulates the vision challenges of gastrointestinal tract and vasculature anatomy. A central component of the framework are YOLO detectors trained on a dedicated dataset of capsule targets spanning multiple sizes and geometries. The detectors are used to detect the target in each camera view and to extract its 2D center; with YOLO-OBB (Oriented Bounding Box), it also provides an in-plane orientation, which we use to infere a heading cue. Given the synchronized left/right detections (cameras on Fig. 1a), we reconstruct the 3D target center via stereo triangulation and, when applicable, track position and heading. Our work focuses on repeatability and interpretable total error of capsule robotic systems of different sizes and shapes because absolute accuracy is limited by manual target placement and the absence of an external high-precision ground-truth tracking system. We decompose the error into variability induced by manual repositioning across repeated trials, within-trial measurement noise while the target remains stationary, and a detection/geometry related error. This separation enables identification of the dominant error source and provides an interpretable total error for navigation tasks, where stable relative localization is often more critical than absolute accuracy.

2. Related Research

Reliable pose estimation of the device is widely recognized as key enabler for magnetically actuated microrobots and capsule robotics systems, as they directly determine whether closed-loop control and re-navigation can be performed safely in confined, deformable anatomy. Recent surveys emphasize that no single sensing modality is universally reliable. Magnetic and vision-based approaches all face practical limitations when moving from controlled scenarios to clinically realistic conditions, including calibration burden, sensitivity to disturbances, and reduced robustness under occlusions and specular reflections.

A large body of work therefore uses camera-based perception as a first step, because it can provide fine spatial detail with relatively lightweight hardware. However, maintaining stable detection and tracking is challenging when the target is small and the scene appearance varies strongly with viewpoint. Beyond standard camera views, several studies focus on pose/orientation estimation under microscope or otherwise constrained optical setups. Choudhary *et al.* [9] propose a deep learning-based method for three-dimensional microrobot orientation estimation and tracking, enabling richer pose feedback than only object center localization. Related work also addresses 3D pose estimation for visually challenging targets, like transparent or low-contrast microrobots, by inferring depth and orientation from microscope images, where classical handcrafted visual cues are often unreliable [10]. While these approaches advance pose observability, they are validated in highly controlled imaging settings tied to a specific optical system, and experiments are not always structured to separate repeated repositioning, multi-location effects, and within-trial measurement noise. In addition to optical imaging, clinically relevant modalities such as ultrasound have also been investigated. Liu *et al.* [11] address capsule robot pose and mechanism-state detection in an anatomically representative environment using ultrasound imaging and deep learning, demonstrating that clinically relevant scenarios introduce failure modes that must be explicitly handled. Finally, datasets such as USMicroMagSet [12] highlight the importance of standardized evaluation for small device tracking under difficult imaging conditions, reinforcing that performance comparisons should be tied to well-defined metrics and protocols.

Against this background, machine learning has increasingly been adopted to enhance robustness in the visual perception stage. Deep learning object detectors such as YOLO [13] have become common in medical vision tasks due to their favorable balance between accuracy and speed, as well as the ease of deployment. In the microrobotics context, Li *et al.* [14] demonstrate real-time microrobot detection and tracking using deep-learning models, at the same time illustrating that YOLO detectors can remain effective even when the target is small, and the scene is visually challenging. Many vision-only demonstrations primarily report detection/tracking success, while less frequently quantifying how detection variability propagates into navigation-relevant 3D stability across repeated trials and multiple spatial locations. Hybrid and sensor-fusion pipelines are often advocated to mitigate modality-specific weaknesses and retain metric consistency under realistic disturbances [15].

In contrast to prior works, our study combines a calibrated stereo setup with YOLO-based detectors to introduce an evaluation protocol that explicitly quantifies repeatability by decomposing localization error into detection error, manual repositioning variability and within-trial measurement noise. This design directly supports navigation interpretation: not only whether the detector works, but how stable and predictable the resulting 3D localization is across repeated trials and multiple maze locations.

3. Materials and Methods

The goal of this work is to assess the detection of robustness and tracking repeatability in a clinically inspired experimental scenario. We are using a designed setup, shown in Fig. 1a, composed of a maze phantom and a calibrated stereo vision system with the dimensions shown in Fig. 1b.

The maze phantom is fabricated from plexiglass and rigidly mounted on a magnetized base plate, providing a stable and reproducible substrate during repeated trials. To reduce the impact of unwanted optical effects, the maze is lined with a white background layer, which improves contras and helps reduce the influence of reflections. The stereo rig remains unchanged throughout calibration and data collection sessions. Camera intrinsic, stereo extrinsic, and the world reference frame are obtained using an OpenCV ChArUco board pipeline, where ChArUco detections are used for per-camera intrinsic calibration, stereo calibration between C0 and C1, and a fixed board pose to estimate the rigid transform from C0 to the world coordinate system W, Fig. 1a. To support learning-based perception in this clinically inspired maze, we made a stereo dataset tailored to the geometry and materials of the experimental setup. The dataset is collected using both camera views and is designed to cover full experimental matrix: five capsule robots represented by magnets of different sizes and geometry (Fig. 2d).

All the recorded images are annotated in Roboflow using polygon-based labels to capture the true target extent for elongated and rotated objects. At each location, mentioned as P1-P10 in Fig. 1c, the magnet was manually repositioned, and the procedure was repeated 10 times to capture across-trial variability; within each repetition, 100 frames were recorded to characterize within-trial measurement fluctuations while the target remained stationary. Two deep learning detectors were evaluated in this work: YOLOv12n and YOLOv12n-OBB. We use standard YOLO model as a strong, widely adopted baseline for reliable and efficient 2D center detection and we use OBB to test whether explicitly modeling target orientation (via oriented bounding box) improves robustness and provides an additional heading cue- thereby quantifying the benefit of orientation-aware detection for navigation in curved, occlusion-prone scenes. For cylindrical targets, the 3D heading vector is obtained by triangulating the endpoints of the major axis estimated by OBB in the two camera views, and heading error is computed as the angular difference between the estimated and reference 3D direction. The standard YOLO and the OBB model were trained using the same configuration, with identical training parameters and a confidence threshold of 0.5 for inference. All experiments are conducted on a research workstation running Ubuntu 22.04, equipped with a 64-bit Intel Core i5 CPU operating at 2.80 GHz. In our system, the OBB orientation is used together with the object's 2D center to provide an orientation cue from both camera views. Quantitative 3D heading estimation is not the primary focus of the present study.

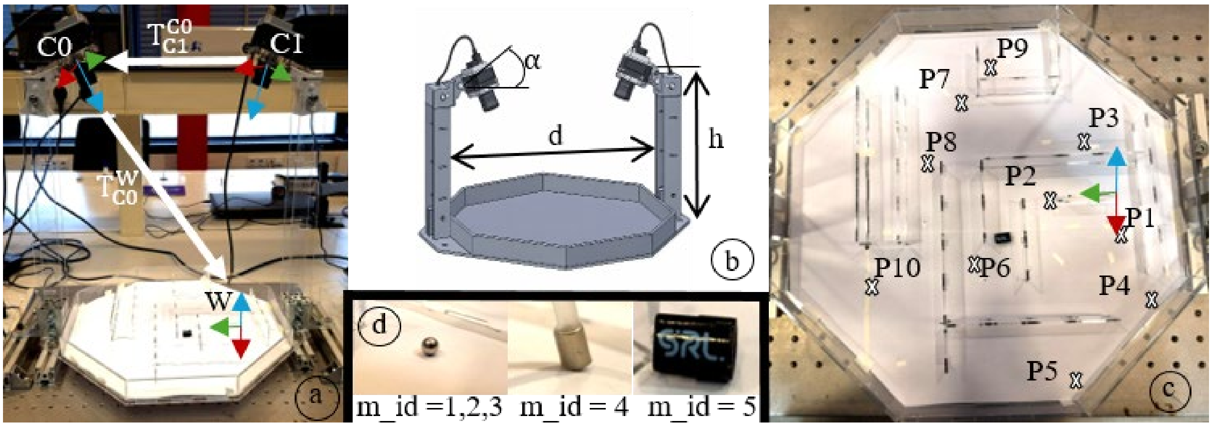

Fig. 1. a) Clinically inspired maze phantom setup with a calibrated stereo vision rig. Two fixed cameras define the camera coordinate frames C0 and C1. The stereo calibration provides rigid transformation between cameras, $\boldsymbol{T}_{C1}^{C0}$, while an extrinsic calibration defines the transformation from camera 0 to the world coordinate system, $\boldsymbol{T}_{C0}^{W}$. The world coordinate system – W is used as the reference for reporting metric 3D positions and repeatability results. The orientation of coordinate axes and the direction of transformations used in the reconstruction process is shown with arrow, b) Monochrome Basler cameras, model avA1000-100gm, are used. Each camera is operating at resolution 1024x1024 with 6 mm focal length lenses. Cameras are mounted at a fixed height of h = 420 mm and tilted by 28° to satisfy the required field of view and working distance. The maze area observed by the stereo pair corresponds to an approximately circular field of view with a diameter d = 350 mm, c) Top view of the maze phantom with the 10 predefine spatial locations (P1 – P10) used for repeatability testing. The marked points span straight segments and curved regions, capturing visibility changes, occlusions, and reflections induced by the maze geometry. These positions define systematic evaluation grid for comparing target-dependent and location-dependent effects, d) Representative targets used in the study. Capsule /microrobots surrogate were realized using five magnets of different size and shapes. Circular magnets are m_id = 1, 2, 3, and their dimensions are: m_id = 1 has diameter 2 mm, m_id = 2 has diameter 3 mm, and m_id = 3 has 5 mm. Two magnets are cylindrical sizes, m_id = 4 (Ø3.5 x 6 mm) and m_id = 5 (Ø9 x 15 mm). This set allows evaluation of detection and localization repeatability across varying target appearance, scale and aspect ratio.

4. Results

We assessed detection performance and positioning error in a bowel-/vessel-like maze phantom that mimics clinically relevant visual constraints, using multiple target locations that would be difficult to probe in vivo. Our goal was to determine which factor has the greatest impact on positioning accuracy and which error sources dominate in repeatability.

Detection performance was evaluated on a test image set using a standard YOLOv12n detector and a YOLOv12n-OBB detector. The annotated dataset comprises 10 948 frames and was split 80/10/10 into train/validation/test, yielding a test set of 1094 images used for the quantitative results in Table 1. We report IoU (Intersection-over-Union) based overlap statistics to quantify spatial agreement between predicted regions and the reference annotations. IoU is defined as the ratio between the intersection area and the union area of the prediction and the ground truth. To summarize the IoU distribution, we report IoU_{mean}, IoU_{p95} and the proportion of detections with $IoU \geq 0.75$, a commonly used criterion for tight localization. On the test set, IoU statistics indicate that the standard

YOLO achieves slightly tighter spatial alignment with the ground-truth regions than OBB, as reflected by higher average IoU, a higher p95 values of IoU and a larger fraction of detections exceeding the IoU≥0.75 threshold.

In contrast, OBB shows a modest drop in IoU overlap, which is expected because the overlap is more sensitive to small orientation deviations. Even minor angle errors can reduce IoU while detection remains visually correct. Because our downstream objective is stereo triangulation, we prioritize detection properties that directly affect 3D reconstruction: image-level detection rate, stability of the estimated 2D center, and for OBB orientation error, since it provides an angle estimate. Overall, OBB yields more reliable image-level detections and a more consistent precision-recall behavior, which is reflected in the mAP metrics reported in Table 1**.** It also triggers fewer false positives in difficult scenes with partial occlusions and specular reflections, compared with the standard YOLO model. For stereo triangulation, two properties matter most: consistent detection in both camera views and stable 2D center localization. In our experiment, both detection models achieve comparable center accuracy in the image plane. To better characterize center stability beyond mean/p95, we report the full distribution of center errors using an empirical cumulative distribution function (CDF), Fig. 2b, which reveals how frequently small pixel-level errors occur.

Table 1. Detection and 2D localization performance of the standard YOLO model and the YOLO-OBB evaluated on the test set.

Metric	YOLOv12n	YOLO-OBBn
IoU_{mean}	0.72	0.70
IoU_{p95}	0.96	0.93
IoU≥0.75	0.51	0.49
Test images	1094	
Detection rate	98.61 %	99.11 %
Precision	86.11 %	95.6 %
Recall	85.93 %	94.5 %
mAP50	85.74 %	95.25 %
mAP50-95	48.61 %	66.53 %
Center error mean/p95 (px)	4.22/11.68	4.31/11.99
Angle error mean/p95 (°)	-	4.63/11.1
Inference time mean (ms)	427.33 (~3 FPS)	203.90 (~5 FPS)

Since IoU is often used as a localization quality proxy, we further analyze how IoU relates to the 2D center error for OBB detections, Fig. 2c. It shows that higher overlap generally corresponds to smaller center errors, but with increased variability at lower IoU. This suggests that the primary performance gap is driven by detection robustness (whether the target is detected at all), rather than by systematic differences in the predicted center once a detection is present. A key benefit of OBB is the additional orientation signal. This enables estimation of target heading and extends localization beyond 3D position toward direction estimation, which is directly relevant for navigation in narrow and tortuous paths where heading informs steering and collision avoidance. It is important to note that orientation is not equally meaningful for all target geometries. For circular or symmetric targets, the heading can be ambiguous because multiple orientations look identical in the image. In those cases, even a correct detection can yield an unstable OBB angle, since small appearance changes can rotate the fitted box without reflecting a true physical rotation. This is why orientation estimates are inherently more reliable for elongated targets, which provide stronger directional visual cues. In our navigation setting, this implies that pose estimation should be interpreted as a robust 3D position estimate for all targets, and a reliable heading estimate primarily for non-symmetric targets.

Runtime was measured on a CPU configuration. Even under these constraints, OBB runs approximately two times faster than the standard model in our setup. Although the current system is not yet real-time, these measurements provide a clear CPU baseline and motivate GPU deployment in future work.

For navigating a microrobot or a capsule the stability of the 3D pose estimate across repetitions is critical. The system must return to a consistent location and orientation to support motion planning, heading correction, and revisiting regions of interest. to quantify this, we designed an experiment with 5 magnets (representing different capsule/microrobot geometries) and 10 spatially defined points within the maze. Each point was repeated 10 times per magnet. Within each repetition, we recorded 100 frames while the object remains stationary. This design lets us separately quantify variability across repetitions and variability within a repetition under stationary conditions. Table 2 reports the 3D localization errors grouped by magnet type. As expressed in Equation (1), the total error is decomposed into components associated with repositioning variability, within-trial noise, and repeatability, allowing a more interpretable analysis of localization performance.

$$E_{total} = \sqrt{E_{place}^2 + E_{noise}^2 + E_{det}^2} \qquad (1)$$

All components are reported as mean/p95. The mean summarizes average behavior, and the p95 value indicates a conservative bound below which most measurements fall, and it is useful for safety-aware navigation. The results show that total error is primarily driven by manual repositioning. Within-trial noise is smaller and reflects the intrinsic stability of the measurement process when the target does not move. This indicates that the largest improvements in repeatability will come from standardizing the placement procedure and reducing manual variability, for example through fixtures, guided placement, or automated positioning. At the same time, the measurement process itself behaves relatively consistently during stationary measurements, which is encouraging for close loop use once placement variability is controlled. Finally, the spatial dependence of the worst-case 3D localization error is visualized as a heatmap of E_{total} at p95 across all points and magnet types, Fig. 2a, highlighting specific maze locations that act as error hotspots.

We also observe a clear dependence on target geometry, Fig. 2a. Circular/symmetric capsules are more sensitive to imaging conditions such as bends, partial occlusions, and specular reflections. This typically appears as higher across repetition variability, because small changes in viewpoint and visibility produce larger changes in the detected outline and thus in triangulated depth. Cylindrical targets provide a stronger, more stable visual signature across viewpoints. The largest cylindrical magnet (m_id = 5) achieves the best repeatability and the most stable total error, consistent with its higher visual salience and more reliable detections in both camera views. The smaller cylindrical target (m_id = 4) preserves the same geometric advantage but is more sensitive to detection instability and center jitter under occlusions and reflections, because its projected image area is reduced and the influence of background clutter becomes relatively larger.

Table 2. Decomposition of 3D localization error in the maze phantom for each magnet type, reported as mean/p95 (mm) across all evaluated points and repetitions using the final selected model, YOLOv12n-OBB. The p95 statistic provides a conservative upper-tail summary that is particularly relevant to navigation robustness. We decompose errors into interpretable components. E_{place} captures variability induced by manual repositioning between repetitions. E_{noise} captures within-trial measurement noise while the target is stationary. E_{repeat} summarizes repeatability at the same point (how constantly the system returns to the same 3D location over repeated placements). E_{total} represents the total error.

m_id	E_{place} [mm]	E_{noise} [mm]	E_{repeat} [mm]	E_{total} [mm]
1	1.21/8.34	0.28/1.37	1.36/8.38	1.64/8.52
2	1.80/8.34	0.28/0.99	1.85/8.38	2.06/8.52
3	1.33/6.80	0.49/2.61	1.70/8.45	2.40/11.1
4	1.54/7.44	0.18/0.91	1.58/7.59	1.77/7.78
5	0.45/1.11	0.48/2.02	0.72/2.28	0.91/2.33

5. Conclusions

In this paper, we present an evaluate a stereo-vision-based experimental setup for capsule robotic systems in a bowel-/vessel-inspired maze phantom, intended to support the testing and validation of localization and future control strategies, with emphasis on repeatability rather than absolute accuracy. The proposed pipeline combines YOLO-based 2D detections in both camera views with stereo triangulation to reconstruct the 3D target center, while YOLO-OBB additionally provides an estimate of target heading. On a test set, YOLO-OBB demonstrates superior detection reliability compared with a standard YOLO model, achieving higher precision/recall and overall detection quality, while maintaining comparable 2D center localization accuracy. These properties translate into a higher likelihood of obtaining valid, synchronized detections in both views, which is critical for robust stereo triangulation and consistent 3D reconstruction.

Beyond detector performance, a key contribution of this work is an interpretable repeatability analysis that decomposes 3D localization variability into variability induced by manual repositioning and within-trial measurement noise while the target remains stationary. This decomposition provides a practical total error that aligns with navigation requirements, where stable relative localization is often more important than global absolute accuracy. These results indicate that within-trial noise is low across most maze regions, whereas the upper-tail error is primarily driven by manual repositioning variability and localized failure modes. Heatmap and worst-case analyses identify a few spatial hotspots that dominate the error tail under visually challenging conditions. Finally, the experiment highlights a clear geometry dependence: cylindrical targets exhibit more robust behavior, while symmetric provide limited directional cues, making orientation estimates inherently less informative and more sensitive under adverse imaging conditions. Overall, the framework supports stable relative localization (and heading where meaningful) across most maze segments, establishing a solid basis for closed-loop navigation of magnetically actuated devices in tortuous anatomy.

Future work will focus on reducing dominant error sources and increasing real-time capability. We will introduce automated repositioning to isolate sensing error from setup variability. Second, we will improve robustness in the identified hotspot regions by collecting additional data. Hard-negative mining and failure-case reweighting will be explored to further suppress false positives in reflective scenes. The full process will be running on GPU and optimize end-to-end latency, with the goal of achieving update rates suitable for closed-loop control. Finally, we will extend evaluation to more diverse lighting, materials, and phantom configuration, and investigate multimodal fusion (magnetic sensing an/or inertial cues) to increase robustness under clinically realistic variability.

Acknowledgements

This research was funded by the project INSPIRATION non-INvaSive PatIent RegistrATIon for rObotic Neurosurgery (grant no. NPOO.C3.2.R2-I1.06.0153), financed by the European Union through the National Recovery and Resilience Plan (NPOO).

This work was supported by the European Commission under the Horizon Europe program under Grant no. 101070066 (RĚGO).

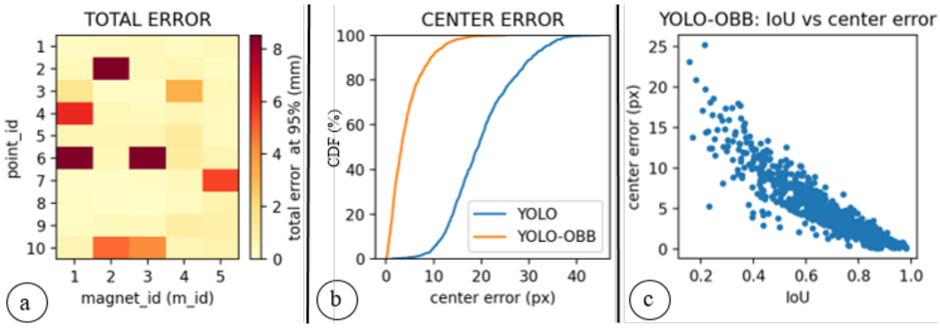

Fig. 2. a) Heatmap of the 95th percentile total 3D localization error (E_{total}), evaluated for each combination of maze position P1-P10 (y-axis) and magnet type m_id = 1-5 (x-axis). The most critical spots are concentrated at point 2 for magnet 2 and point 6 for magnets 1 and 3, where the total 3D error at p95 increases substantially relative to the remaining locations. This pattern indicates that worst-case performance is driven by a specific combination of maze geometry and magnet type and is further amplified by local visibility constraints and material-induced optical artifacts, b) Empirical cumulative distribution function (CDF) of the 2D center localization error in the image plane for both detection models, where the x-axis denotes center error of the target and the y-axis denotes the cumulative percentage of detections. Curves that rise faster and lie further left correspond to more frequent small pixel errors and thus higher center stability. This plot complements mean/p95 summaries by revealing the full distribution and the presence of long-error tails that can affect stereo triangulation, c) Scatter plot relating IoU (x-axis) to 2D center error (y-axis) for YOLO-OBB detection model. The negative trend indicates that higher overlap with the reference annotation generally coincides with smaller center errors, while lower IoU values are associated with larger variability in center estimates. This relationship helps interpret IoU as a proxy for localization quality: high IoU detections tend to provide stable centers, whereas low IoU detections include a wider range of center errors and are more likely to degrade 3D reconstruction.

References

[1]. D. Dlaka, M. Švaco, et al., Frameless stereotactic brain biopsy: A prospective study on robot-assisted brain biopsies performed on 32 patients by using the RONNA G4 system, *The International Journal of Medical Robotics and Computer Assisted Surgery*, Vol. 17, Issue 3, 2021, e2244.

[2]. A. Handa, et al., Role of robotic-assisted surgery in public health: Its advantages and challenges, *Cureus*, Vol. 16, Issue 6, 2024, e62331.

[3]. T. Haidegger, et al., Robot-assisted minimally invasive surgery-surgical robotics in the data age, *Proceedings of the IEEE*, Vol. 110, Issue 7, 2022, pp. 835-846.

[4]. Q. Cao, et al., Robotic wireless capsule endoscopy: Recent advances and upcoming technologies, *Nature Communications*, Vol. 15, 2024, 9355.

[5]. D. Zhang, et al., Advanced medical micro-robotics for early diagnosis and therapeutic interventions, *Frontiers in Robotics and AI*, Vol. 9, 2023, 1056508.

[6]. B. J. Nelson, et al., Magnetically actuated medical robots: An in vivo perspective, *Proceedings of the IEEE*, Vol. 110, Issue 7, 2022, pp. 1028-1037.

[7]. A. Schmidt, et al., Tracking and mapping in medical computer vision: A review, *Medical Image Analysis*, Vol. 94, 2024, 103134.

[8]. M. A. Ali, et al., Recent advancements in localization technologies for wireless capsule endoscopy: A technical review, *Sensors*, Vol. 25, Issue 1, 2025, 23.

[9]. S. Choudhary, et al., Three-dimensional optical microrobot orientation estimation and tracking using deep learning, *Robotica*, Vol. 43, Issue 2, 2025, pp. 616-637.

[10]. D. Zhang, et al., Micro-object pose estimation with sim-to-real transfer learning using small data, *Communications Physics*, Vol. 5, 2022, 80.

[11]. X. Liu, et al., Capsule robot pose and mechanism state detection in ultrasound using attention-based hierarchical deep learning, *Scientific Reports*, Vol. 12, 2022, 18671.

[12]. K. Botross, USMicroMagSet: Using deep learning analysis to benchmark the performance of microrobots in ultrasound images, *IEEE Robotics and Automation Letters*, Vol. 8, Issue 6, 2023, pp. 3254-3261.

[13]. A. Chandrashekhar, et al., An efficient YOLOv12-based framework for detecting extremely small-scale objects, *Scientific Reports*, Vol. 16, 2026, 2062.

[14]. H. Li, et al., Magnetic-controlled microrobot: Real-time detection and tracking through deep learning approaches, *Micromachines*, Vol. 15, Issue 2, 2024, 255.

[15]. X. Tang, Vision-based automated control of magnetic microrobots, *Micromachines*, Vol. 13, Issue 2, 2022, 269.