SN Computer Science (2026) 7:437
https://doi.org/10.1007/s42979-026-05071-9

ORIGINAL RESEARCH

A Low-Cost Benchmarking Platform for Tracking Systems to Enable Augmented Surgery Tools in Clinical Practice

Federico Salerno[1] · Alessandro Contenti[1] · Luca Ulrich[1] · Sandro Moos[1] · Attilio Fiandrotti[2] · Enrico Vezzetti[1]

Received: 1 August 2024 / Accepted: 5 May 2026

Abstract

Augmented Reality (AR) is increasingly being adopted in surgical procedures. Identifying accurate and reliable tracking systems can enhance the effectiveness of AR assisted surgery. The study presents a benchmarking platform to evaluate the performance of different optical surgical tools tracking systems for AR applications, addressing the need for standardized comparison of tracking systems' accuracy, repeatability, and reliability. A custom-built, cost-effective benchmarking platform was developed, and different quantitative evaluation metrics were employed to assess performances of Marker-Based (MB) tracking systems. Two types of measurements were performed: static, where metrics such as Target Registration Error (TRE) and tip stability as a measure of jitter were estimated, and dynamic, involving a measure of the deviation between the tracked tooltip trajectory and its GT. Three AR MB-tracking methods were tested: method 1, mono-RGB sensor tracking a planar image target; method 2, RGB-depth sensor tracking passive spherical markers; and method 3, mono-IR sensor tracking IR active markers. Significant performance differences were observed. Method 3 achieved the lowest TRE value of 3.578 ± 2.836 mm. Method 1 exhibited the best tip stability with a jitter value of 1.435 ± 0.824 mm. Method 2 obtained the minimum trajectory distance in dynamic tests. The benchmarking platform demonstrated its effectiveness in evaluating and comparing different AR tracking methods. This study enables the optimization of tracking system selection for use in the operating room, and facilitating the integration of AR as a supportive technology in surgical practice.

Keywords Augmented surgery · Augmented reality · Pose estimation · Tooltip tracking · Marker-based · RealSense

✉ Enrico Vezzetti
enrico.vezzetti@polito.it

Federico Salerno
federico.salerno@polito.it

Alessandro Contenti
alessandro.contenti@polito.it

Luca Ulrich
luca.ulrich@polito.it

Sandro Moos
sandro.moos@polito.it

Attilio Fiandrotti
attilio.fiandrotti@unito.it

1 Department of Production and Management Engineering, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy

2 Department of Computer Science, Università di Torino, Corso Svizzera, 185, 10149 Turin, Italy

Introduction

Augmented Surgery refers to the adoption of Augmented Reality (AR) technology during surgical procedures and employs real-time imaging techniques to enhance surgical precision and safety, allowing for minimally invasive procedures and improved outcomes [1]. In this context, it is often required to precisely track instruments relative to a Reference Frame (RF) for accurate surgical navigation [2]. Instrument tracking in AR refers to the process of identifying and monitoring the position and orientation of a specific tool relative to an arbitrary RF. In the surgical field, the instrument tracking procedure is adopted to provide real-time guidance to the surgeon. Tracking feedback commonly consists of the superimposition of a 3D hologram of the tool on the real one. AR applications can be either monitor-based or Head-Mounted Display (HMD)-based [3–5]. In monitor-based systems, the augmented content is displayed on a screen.

Published online: 20 May 2026

In HMD-based systems, the augmented content is projected directly into the user's field of view through the headset. Monitor-based systems allow users to avoid moving their heads, offering a perfectly integrated real-virtual feedback; however, they lack the immersive experience provided by HMDs [6]. On the other hand, HMD-based systems offer a more immersive experience by directly projecting the augmented content into the user's field of view, at the cost of wearing discomfort.

In literature, two main types of tracking can be identified: inside-out and outside-in [7]. Inside-out tracking uses an AR HMD's cameras to track instruments, and shows results directly within the headset, making the system self-contained. Outside-in tracking relies on external cameras, displaying results on a monitor or within an HMD, providing visualization flexibility.

Acquisition cameras' setup and detection techniques are the key aspects of AR tracking procedures. The camera setup includes the type of camera used, namely RGB, grayscale, or infrared, and the technique for depth estimation [8], i.e., the use of stereoscopy, structured light, or Time-of-Flight sensors [9, 10]. Marker-Based (MB) detection technique employs specific fiducial markers [11] to identify and track instruments. More precisely, there are pattern-based markers [12, 13], e.g. QR codes, ArUco markers [14], and Vuforia image targets, and localization markers, either active, using IR or visible light LEDs, or passive, such as retro-reflective or colored spheres[15–17].

Surgical navigation instruments have markers and a tip to digitize touched points. The rigid connection between the tip and marker allows pose estimation to determine the tip's position. Tooltip instrument tracking is used in various applications, such as AR point-based registration [18], meaning the overlay of any virtual model on a determined anatomical district through the selection of at least three points on the physical part, and the visualization of the augmented operating field through an AR device. Moreover, it is also applied for guiding osteotomies [6, 19–21] namely for surgical guides anchoring [22], inserting pedicle screws [23, 24], and identifying specific anatomical points.

Nevertheless, tracking errors in AR can significantly impact application effectiveness, compromising functionality in various scenarios, particularly for applications involving point-based registration. In inside-out applications, the absolute tracking error along the optical axis might pass unnoticed [25]. This error can eventually propagate to estimated positions during point picking tasks, thereby hindering accurate positioning of virtual scene elements used to perform surgical treatments, as depicted in Fig. 1a. In outside-in tracking, there are two possible scenarios (Fig. 1b): in HMD-based applications [26], tracking errors affect all tracked objects in potentially different ways, resulting in unreliable recorded point coordinates [27]; in monitor-based applications, due to the fixed position of both the tracking and display systems, depth perception becomes more difficult the closer one is to the optical axis [28], making it difficult to perceive, and thus acknowledge, any tracking errors in the collected points.

Given a specific application context, arises the need to determine which tracking method ensures the best performance in terms of accuracy, measurement repeatability, and procedure reliability. Thus, there is a need to quantitatively and consistently compare different tracking methods using a dedicated evaluation tool.

Despite numerous studies in augmented surgery, tracking error is usually not measured in an absolute but a relative manner, meaning that a certain number of points are collected with the tracked instrument and then aligned with the

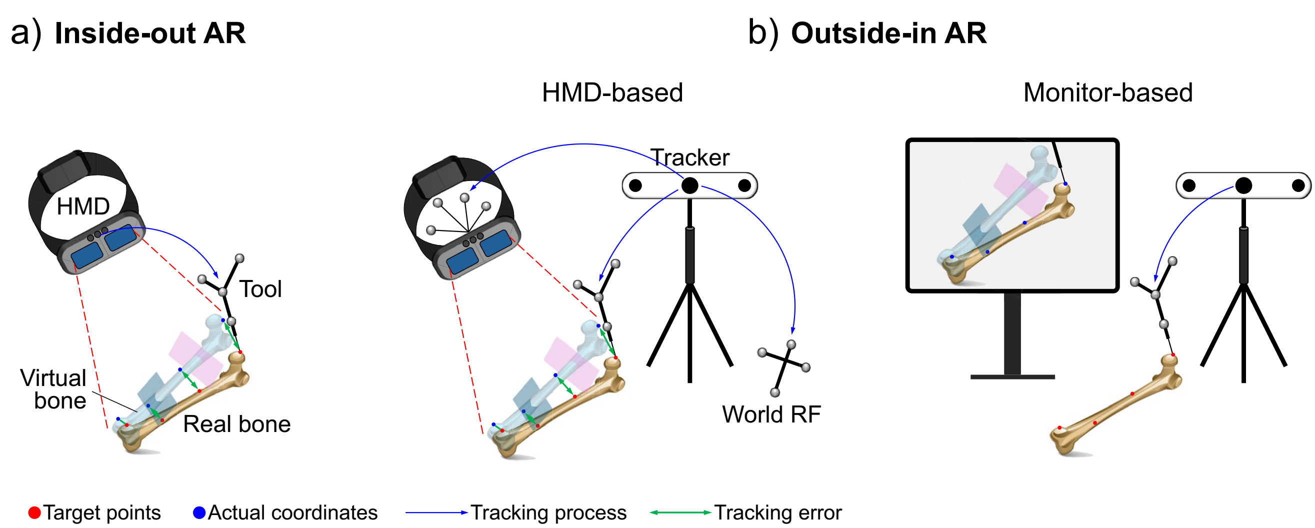


Fig. 1 How tracking error impacts in augmented content visualization in point picking-based AR surgical procedures. **a** Inside-out AR application; **b** Outside-in AR applications

theoretical pattern of points through a transformation that minimizes the alignment error, nullifying or compensating for any absolute tracking errors. Moreover, a comparison of different tracking methods using a unique and repeatable comparison procedure seems to be lacking. On the other hand, although several works compare marker detection, there is a lack of application to tracking tools where the tip needs to be tracked.

In this framework, the present work introduces a benchmarking platform designed for reproducible and modular evaluation of MB-AR tooltip tracking. The platform combines a standardized positioning structure, ground-truth estimation procedures, and a set of quantitative metrics, enabling consistent comparisons across different tracking pipelines.

The objective of the present work is to provide a testing platform that allows the evaluation of the absolute error of MB-AR tooltip tracking and the benchmarking of different tooltip tracking methods using quantitative evaluation metrics [29, 30], both in static and dynamic conditions. The proposed framework is intended to assess whether the achieved tracking accuracy lies within tolerance ranges that are considered acceptable in the targeted surgical context. Moreover, the platform is designed to be cost-effective, providing a reproducible methodology without the need for expensive or highly specialized equipment.

The article is organized as follows. In Sect. Related Works, a comprehensive overview of related works is provided. Section Materials and Methods provides a description of the testing platform structure, followed by the performance assessment methodology of some tested MB-AR tracking methods. In Sect. Results, both accuracy metrics results and a benchmarking of the tracking methods are presented. In Sect. Discussion, a discussion of the results is provided, while in Sect. Conclusions, the conclusions are drawn.

Related Works

Several works in the literature propose tracking methods, evaluating their performance through quantitative metrics. For clarity, a tracking method includes the video source employed, i.e., the type of camera used, the pose estimation algorithm to evaluate both the tool position and orientation, and the marker type. In 2015, Cesar et al. [31] evaluated the detection behavior of pattern-based markers in underwater environments, using metrics like the minimum dimension of the detected marker and the maximum permissible rotation with respect to the camera. Three years later, Ho et al. [32] employed the Kalman filter to enhance the pose estimation of a vibrating ArUco marker. In the same period, Romero et al. [33] proposed a “Fractal Marker” to enlarge the detection range in terms of distance from the camera and robustness to occlusions. Similarly, Sarmadi et al. [34] introduced a tracking method based on multi-camera pose estimation of prismatic objects equipped with planar ArUco markers. In 2020, Yu et al. [35] introduced TopoTag, a pattern-based fiducial marker designed to improve detection accuracy, vertex jitter, and pose jitter under occlusions. Concurrently, Popescu et al. [36] presented a multi-camera setup to enhance the pose estimation accuracy of ArUco markers. Brand et al. [37] explored the tracking accuracy of Microsoft HoloLens 1 using cubic markers with ArToolkit markers, comparing mono vision and stereo vision performances. In 2023, Zhang et al. [38] developed a novel color correction algorithm to increase detection capabilities for colored pattern-based markers. In the same year, Jurado et al. [39] investigated model-based tracking by combining edge detection, keypoints, and fiducial markers using a smartphone camera. Kulikov et al. [40] proposed an enhancement for traditional ArUco markers by adding a white frame and additional black rectangles to increase detection robustness.

Focusing on surgical applications, Ruppin et al. [41], in 2008, compared the positioning accuracy of surgical splints on cadaver mandibles under three different guidance systems: two optical tracking systems, RoboDent LapDoc Accedo and Artma Virtual Patient, and stereolithographic positioners. Quantitative evaluations involved comparing pre and postoperative CT scans of 20 cadaver mandibles to measure deviations in position and rotation of the implants. Statistical analysis showed no significant differences in application errors among the three systems, demonstrating comparable accuracy for both optical tracking and splint methods.

Two years later, a comparative study was proposed by Elfring et al. [42]. Authors compared the accuracy of three optical tracking hardware commonly adopted in Computer Assisted Surgery (CAS): the NDI Polaris P4, the NDI Polaris Spectra, both in passive and active mode, and the Stryker Navigation System II camera. Static accuracy tests involved tracking single markers over three hours to assess warm-up behavior. Volumetric accuracy was assessed using a Coordinate Measurement Machine across 900 grid positions to determine 3D positional accuracy. Length Measurement Error was calculated following VDI/VDE 2634 guidelines [43] using simulated clinical tool configurations. Tool-based evaluations extended these methods to multi-marker setups to estimate clinical accuracy, including Target Registration Error (TRE) calculations. The Stryker and Polaris Spectra active systems had the highest measured accuracy.

Effective detection and tracking of markers are beneficial for the application of patient-specific templates to the patient’s anatomical regions [44, 45]. An accuracy assessment of optical and electromagnetic tracking systems, commonly adopted in cranial surgery, was proposed by Kral et al. [46], in 2013. Six anatomical specimens were dissected to measure the TRE in and around the skull. Specimens were registered using passive optical markers and electromagnetic tracking. Static accuracy tests involved measuring the TRE using a navigated pointer and implanted titanium screws, repeated eight times for each tracking method. Deviations in the anterior–posterior, left–right, and head-to-foot directions were recorded. Optical tracking showed best and submillimeter accuracy, making it more suitable for high-resolution surgical procedures.

Retro-reflective MB-AR tracking is also of great interest in neurosurgery. In 2020, Kunz et al. [47] performed an *in vitro* study, developing a tracking method based on the HoloLens 1 HMD. Authors evaluated the accuracy of IR MB tracking using a custom phantom head on a robotic Stewart platform serving as a Ground Truth (GT). The HoloLens was aligned at a 60 cm distance, recording 10 movements along each x, y, and z axis, plus 10 combined movements. The real movements were compared to algorithm results based on vector length in space.

Martin-Gomez et al. [48] evaluated the tracking accuracy of retro-reflective markers using the built-in sensors of Microsoft HoloLens 2. The proposed framework tracks multiple tools in 6 Degrees of Freedom (DoF) without additional hardware. Passive retro-reflective markers were detected using the ToF depth sensor. Tracking accuracy was assessed by comparing the system's performance with NDI Polaris Spectra and visible spectrum markers (ArUco and ChArUco). Experiments involved translating and rotating markers along x and z axes and measuring deviations.

Li et al. [49] presented a framework to enhance surgical guidance during neurosurgical procedures through the use of HoloLens 2, employing ToF sensors in combination with retro-reflective markers and point clouds for accurate head tracking without rigidly attaching tracking targets to the patient's skull.

In 2024, Costa et al. [50] used the HoloLens 2 to compare four different tracking algorithms based on three pattern markers: ARToolkit, ArUco, and QR Code tracked both with MRTK2 Extension Service and Vuforia library. They employed retro-reflective spherical markers to create a reference frame for both the tracked marker and the HoloLens 2. For each algorithm, markers were located and tracked in 25 different positions. The GT was created by the Optitrack cameras used to track retro-reflective markers.

Brunzini et al. [51] evaluated the accuracy of a Mixed Reality application for maxillofacial osteotomies and repositioning, using a 3D-printed skull model with patient-specific cubic markers featuring QR codes. The application on Microsoft HoloLens 2 was assessed by making participants draw both cutting lines and drilling points, with accuracy measured using a 3D structured light scanner and analyzed with CloudCompare software.

Cubic pattern-based markers were adopted by Moreta-Martinez et al. (2021) [52] in an AR mobile application based on the Vuforia SDK to perform real-time tracking of patient-specific templates. Placement and tracking errors were separately investigated by comparing the recorded positions with those obtained from virtual models. For tracking error, users placed a tracked pointer on virtual spheres displayed on a smartphone screen and recorded the positions with an Optical Tracking System (OTS). The Euclidean distance between recorded and true positions determined the tracking error. For placement error, the surgical guide was attached to a phantom, and the positions of conical holes were recorded with a tracked pointer. The Euclidean distance between these coordinates and the virtual model's coordinates determined the placement error.

A "Γ"-shape marker was adopted by Akulauskas et al. (2023) [53] in an AR dental surgery application based on the Vuforia SDK to perform real-time tracking of dental models. Calibration and virtual model alignment errors were separately investigated by comparing the recorded positions with those obtained from virtual models. For calibration error, users placed a tracked pointer on predefined points on the dental model and recorded the positions with the HoloLens 2. The Euclidean distance between recorded and true positions determined the calibration error. For virtual model alignment error, a dental model with an attached "Γ"-shape marker was recorded from various positions and orientations. The Euclidean distance between these coordinates and the virtual model's coordinates determined the alignment error.

Materials and Methods

The present section details the constructed benchmarking platform in Subsection Mechanics and electronics, the quantitative analyses performed and the choice of GT are detailed in Subsection Evaluation Metrics, and presents the tested tracking methods in Subsection Tested Tracking Methods.

Mechanics and electronics

The testing platform consists of a modular aluminum frame supporting a robotic arm and a camera: both tool and camera supports integrate passive fiducial markers that define their reference frames. Details of the structure are shown in Fig. 2.

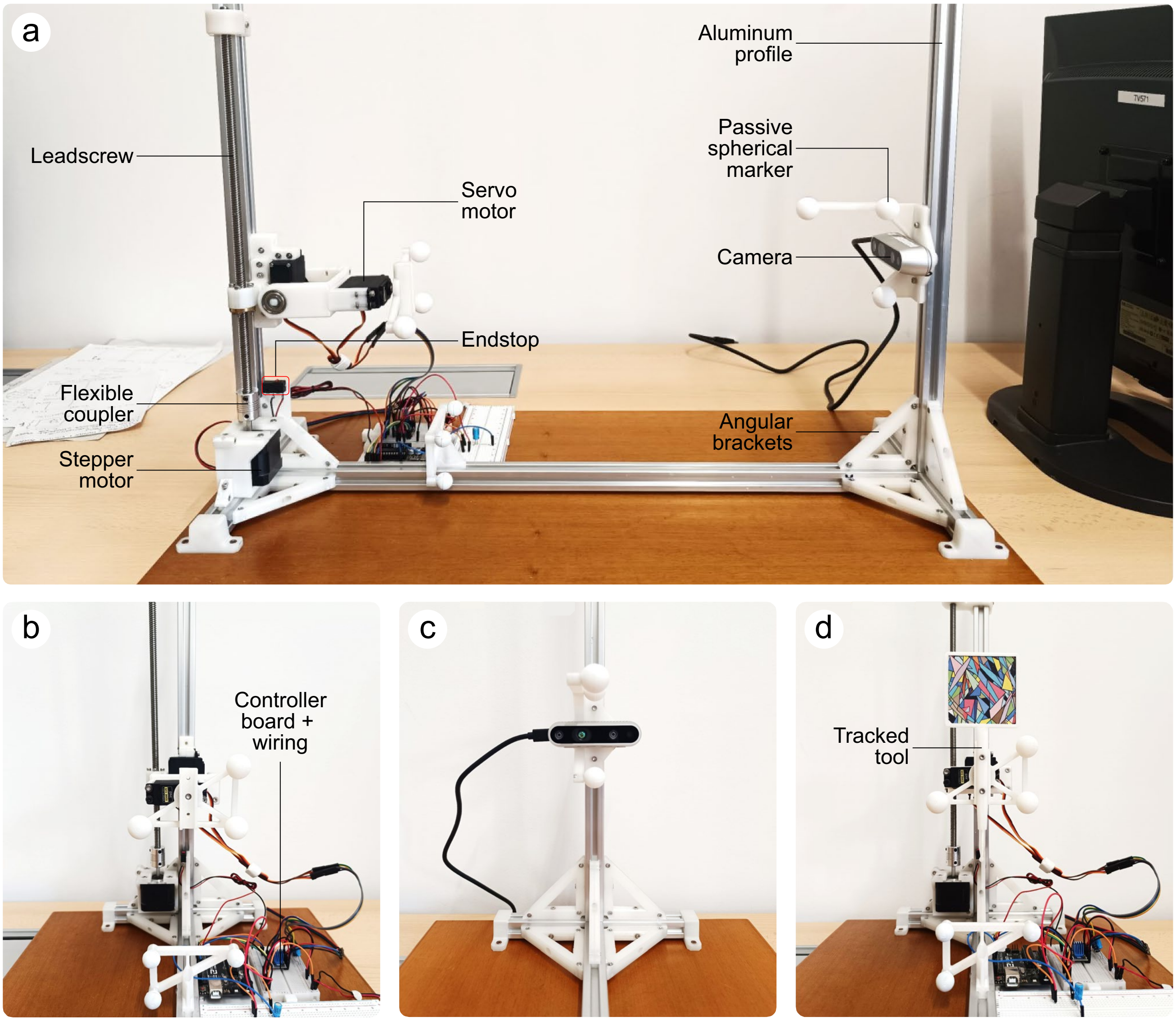


Fig. 2 Testing platform for tracking systems benchmarking: **a** lateral view and components details; **b** robotic arm frontal view; **c** camera support frontal view; d) tool holding frontal view

The robotic arm features three degrees of freedom, comprising one prismatic axis and two revolute joints, enabling controlled variations of both the distance and the orientation of the tracked tool relative to the camera. This configuration allows the exploration of non-planar and non-frontal relative poses, in which the tool undergoes simultaneous translational and rotational changes. The adopted design allows the investigation of a representative and controlled subset of motion configurations, conceived as a working approximation of naturalistic tool movements, while preserving the repeatability required for reliable ground-truth definition and quantitative benchmarking.

The arm is actuated through standard stepper and servo motors, driven by a microcontroller-based module interfaced with a PC command module. Implementation details, including component specifications, CAD models, and control electronics, are reported in the Supplementary Material.

Performance Assessment Methodology

The present subsection encompasses the theoretical formulations of evaluation metrics, contained in Paragraph Evaluation Metrics, the experimental procedure involving the collected measures (Paragraph Adopted Measures), and the GT definition detailed in Paragraph Ground Truth.

Evaluation Metrics

To ensure a uniform benchmarking of the different tracking methods, six evaluation metrics were considered:

Fiducial Localization Error (FLE), Fiducial Registration Error (FRE), Target Registration Error (TRE), Target Jitter Magnitude (TJM), Target Orientation Error (TOE), and Trajectories Hausdorff Distance (THD).

FLE, FRE, and TRE were originally introduced by Maurer et al. in 1997 [54] and are widely adopted to characterize the performance of marker-based tooltip tracking systems. FLE quantifies the localization error of individual fiducials, FRE measures the residual alignment error after pose estimation, and TRE evaluates the localization error of independent target points such as the tooltip.

To extend the evaluation, TJM was defined to assess tracking stability by measuring tooltip oscillations during stationary conditions, while TOE captures discrepancies in the orientation of the tracked instrument. Finally, THD compares the trajectories of the tooltip with respect to the GT, providing a measure of both precision and completeness.

The detailed mathematical formulations, derivations, and schematizations of these metrics are reported in the Supplementary Material.

Adopted Measures

The measurements are divided into two categories: static and dynamic. For static measurements, the goal is to assess the tracking performance of the specific method while the tool remains stationary. In dynamic measurements, the objective is to evaluate the tracking capability as the tool moves. For both measurements, the camera is kept fixed in the same position. Static measurements involve tracking the tool in five different configurations, representing various positions and orientations. These configurations are precisely controlled by a robotic arm to ensure repeatability. The five configurations are chosen to provide diverse views of the tracked tool and its markers, covering a significant portion of the robot arm's workspace, as shown in Fig. 3a.

For every configuration of each tracking method 300 frames were captured to ensure a consistent number of replications across experiments, For each frame in every configuration, FLE and FRE are calculated when the tracking method allows, while TRE and TOE are always calculated.

Dynamic tests involve the tool's tip following a predetermined trajectory, illustrated in Fig. 3b. Each tracking

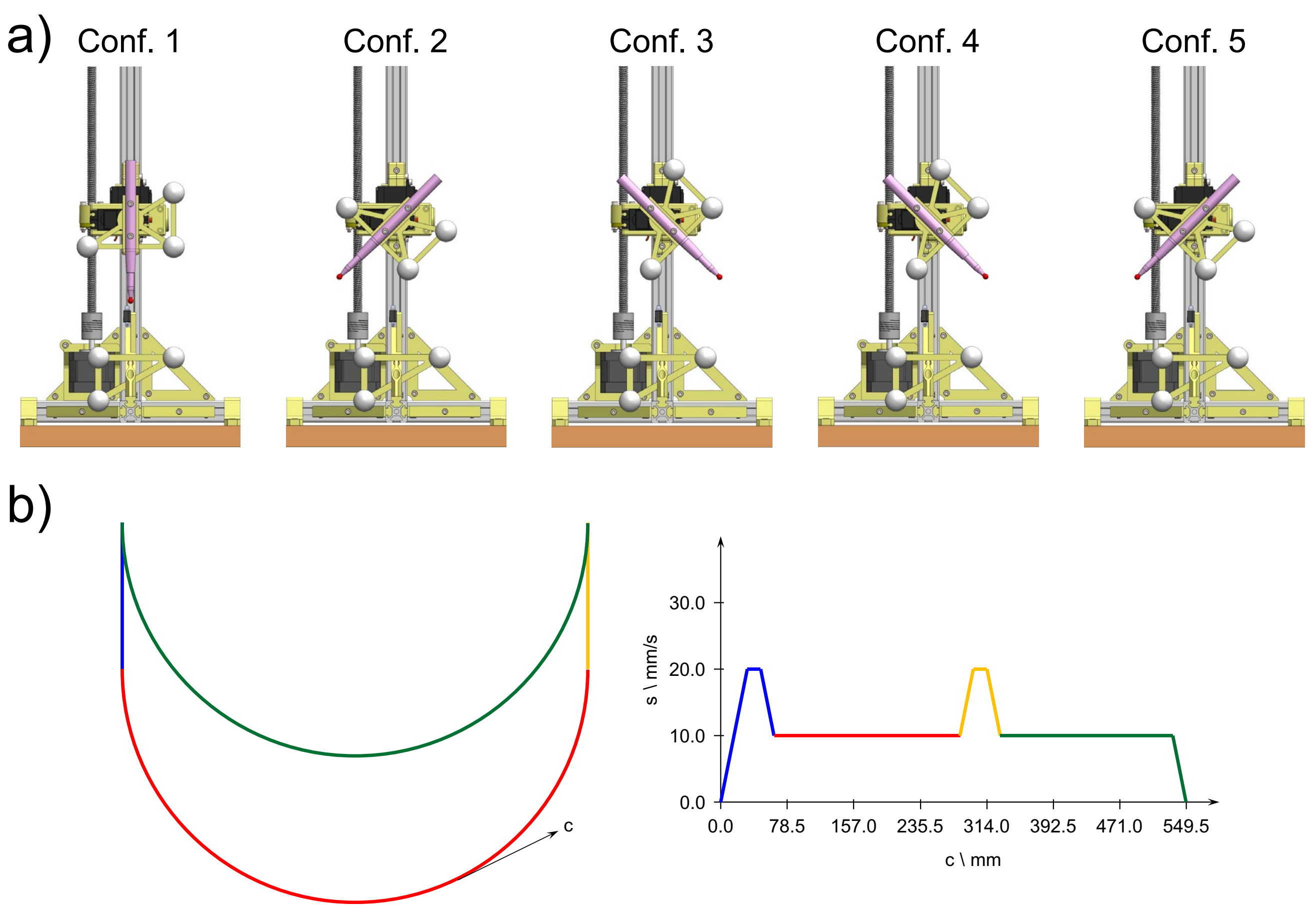


Fig. 3 **a** The five configurations, each corresponding to a specific position and orientation of the tracked tool, shown in purple. **b** On the left, the trajectory followed by the tooltip during the dynamic test. On the right, the tangential speed profile, where c is the curvilinear coordinate tangent to the trajectory at each point

method uses the same tangential velocity profile, controlled by the actuators on the tooltip. A video of the tool is recorded during the trajectory execution. Since the test duration is constant, the total number of frames collected, hence the number of sampled points, will vary depending on the tracking method. For each tested method, a THD couple of values is calculated (see Paragraph Evaluation Metrics).

Ground Truth

Operational Ground Truth Definition

Except for FRE and TJM, all the metrics introduced in Sect. Evaluation Metrics require a operational GT data. The benchmarking platform relies on five reference frames (RF): the scanner $\{S\}$, the world $\{W\}$, the tool $\{T\}$, the camera $\{C\}$, and the lens $\{L\}$, as illustrated in Fig. 4.

For FLE, TRE, and TOE, the true positions and orientations of the tool relative to the lens RF are required, so the key transformation is T_T^L, which expresses the pose of the tool with respect to the lens. This transformation is obtained by relating the triplets of spherical markers attached to the tool and the camera through an auxiliary world RF, making it possible to decompose T_T^L into two main parts:

$$\mathrm{T}_T^L = \mathrm{T}_W^L \cdot \mathrm{T}_T^W. \tag{1}$$

Here, T_W^L is computed once since the world RF is rigidly fixed to the support structure, while T_T^W is obtained for each configuration from the scans of the tool–world pair. In the case of THD, the GT trajectory corresponds to the programmed tool path, which is aligned with the tracked one using ICP.

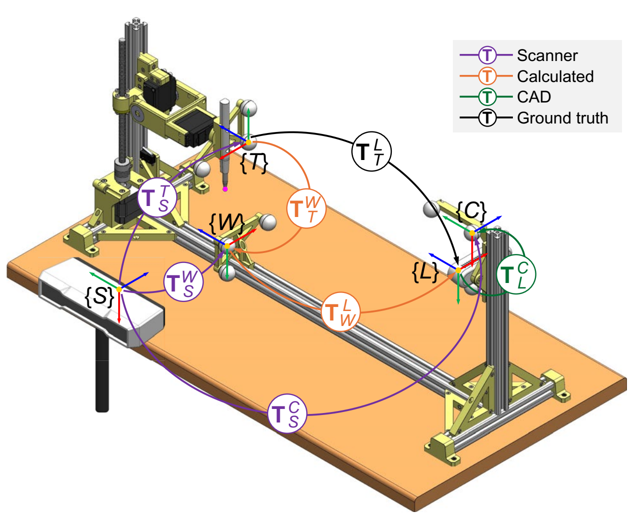


Fig. 4 Reference frames and transformations involved in the GT definition. Violet transformations are directly retrieved from 3D scans, orange are calculated from the previous ones, green is retrieved from CAD, and black corresponds to the GT transformation

To simplify the setup and avoid unnecessary complexity, the following assumptions are made: (i) differences between CAD geometry and the actual tools are neglected; (ii) the geometric data of the camera, including lens position, are assumed accurate; (iii) the tool tip position is derived from the CAD model starting from the Tool RF; (iv) the lens position is derived from the CAD model starting from the Camera RF; (v) for THD calculation, the difference between the programmed and executed trajectory is neglected.

Although the adopted procedure provides a reliable reference, each step involved in the GT definition introduces a potential source of error. The main contributions can be summarized as follows: (i) the dimensional accuracy of the 3D printed custom parts is within ±0.2 mm on the XY plane and ±0.1 mm along the Z axis, according to the Prusa MK3s manufacturer specifications [55]; (ii) the 3D scanner employed (EinScan SP V2, Shining 3D®, Hangzhou, Zhejiang, CN) has a declared single-shot accuracy of $\leq 0.05\mathrm{mm}$ [56]; (iii) the Gaussian fitting procedure used to extract marker centers has been reported to introduce sub-voxel errors, typically below $0.01\mathrm{mm}$ given the scanner resolution [57]; (iv) the ICP alignment procedure used for trajectory matching is associated with residual errors on the order of $0.02-0.05\mathrm{mm}$, depending on the overlap and distribution of sampled points [58]; (v) the simplifying assumption that the programmed trajectory perfectly matches the executed one neglects possible deviations of the motion stage. For linear axes driven by stepper motors, typical positioning errors are reported in the range of 0.01–0.05mm [59].

These factors show that the GT, while precise, is still affected by small errors from printing tolerances, scanner accuracy, fitting, and alignment steps. Further details on the definition of the reference frames, the derivation of all intermediate transformations, and the geometric description of the camera lenses are provided in the Supplementary Material.

Metrological Qualification

In order to quantify the accuracy of the operational GT acquisition chain described in the previous subsection, an independent metrological qualification was performed using a high-precision industrial optical measurement system. Specifically, an optical 3D coordinate measuring system (CMS) [60], namely the ATOS ScanBox 4105 (Zeiss® Industrial Quality Solutions, Oberkochen, Germany), was employed for the qualification procedure. All acquisitions were performed in the metrology laboratory of the Department of Management and Production Engineering (DIGEP), Politecnico di Torino, Italy. The measurements were processed using the dedicated metrology software GOM Inspect Pro for inspection and geometric analysis.

All tool configurations corresponding to those shown in Fig. 3a were replicated. For each configuration, all spherical features were acquired using both the metrology-grade system and the operational GT scanner described previously.

For the purpose of this qualification, three additional Grade 5 [61] ceramic calibration spheres with a diameter of 30 mm [62] were introduced, as depicted in Fig. 5 (left). The spheres were selected for their high form accuracy and dimensional stability. Their centers, extracted from the scanned point clouds, were used to define a local calibration reference frame dedicated to the metrological assessment. All comparisons between the two acquisition systems were performed within this frame, independently of any global alignment procedure. This approach minimizes potential registration biases and ensures a consistent spatial reference across all measurements.

For each configuration shown in Fig. 5, both systems acquired four triads corresponding to: (i) the local calibration reference frame; (ii) the tool reference frame $\{T\}$; (iii) the world reference frame $\{W\}$; (iv) the camera reference frame $\{C\}$.

Sphere centers were extracted from both datasets using the same Gaussian fitting procedure described previously. All coordinates associated with the $\{T\}$, $\{W\}$, and $\{C\}$ triplets were then expressed in the local calibration reference frame defined by the three calibration spheres.

For each spherical feature k, the coordinate discrepancy between the metrology-grade system acquisition and the operational acquisition was computed along the three Cartesian components of the local calibration reference frame as:

$$\Delta \mathrm{p}_k = \begin{bmatrix} \Delta x_k \\ \Delta y_k \\ \Delta z_k \end{bmatrix} = \mathrm{p}_k^{\mathrm{met}} - \mathrm{p}_k^{\mathrm{op}}, \tag{2}$$

where $\mathrm{p}_k^{\mathrm{met}}$ and $\mathrm{p}_k^{\mathrm{op}}$ denote the coordinates of the k-th sphere center measured by the metrology-grade system and by the operational scanner, respectively.

To improve the robustness of the assessment, the acquisitions performed with the operational scanner were repeated ten times for each of the five tested configurations.

Let $\Delta q_{k,i,j}$ denote the discrepancy along component $q \in \{x, y, z\}$ for the k-th sphere, configuration i, and repetition j.

The resulting dataset enables the estimation of repeatability and the separation of systematic and random contributions to the GT uncertainty. The systematic component, i.e. the bias, and repeatability were estimated as:

$$\overline{\Delta q} = \frac{1}{N} \sum_{m=1}^{N} \Delta q_m, \qquad s(\Delta q) = \sqrt{\frac{1}{N-1} \sum_{m=1}^{N} \left(\Delta q_m - \overline{\Delta q}\right)^2}, \tag{3}$$

where N is the total number of discrepancy samples considered for the selected component (aggregated over spheres, configurations, and repetitions). The bias term $\overline{\Delta q}$ was not corrected but retained as part of the measurement performance assessment.

The combined standard uncertainty associated with the discrepancy component was then evaluated according to the GUM framework [63], under the assumption of statistical independence between the metrology-grade and operational contributions, as:

$$u_{\Delta q} = \sqrt{\left(u_{\Delta q}^{\mathrm{op}}\right)^2 + \left(u_{\Delta q}^{\mathrm{met}}\right)^2}, \tag{4}$$

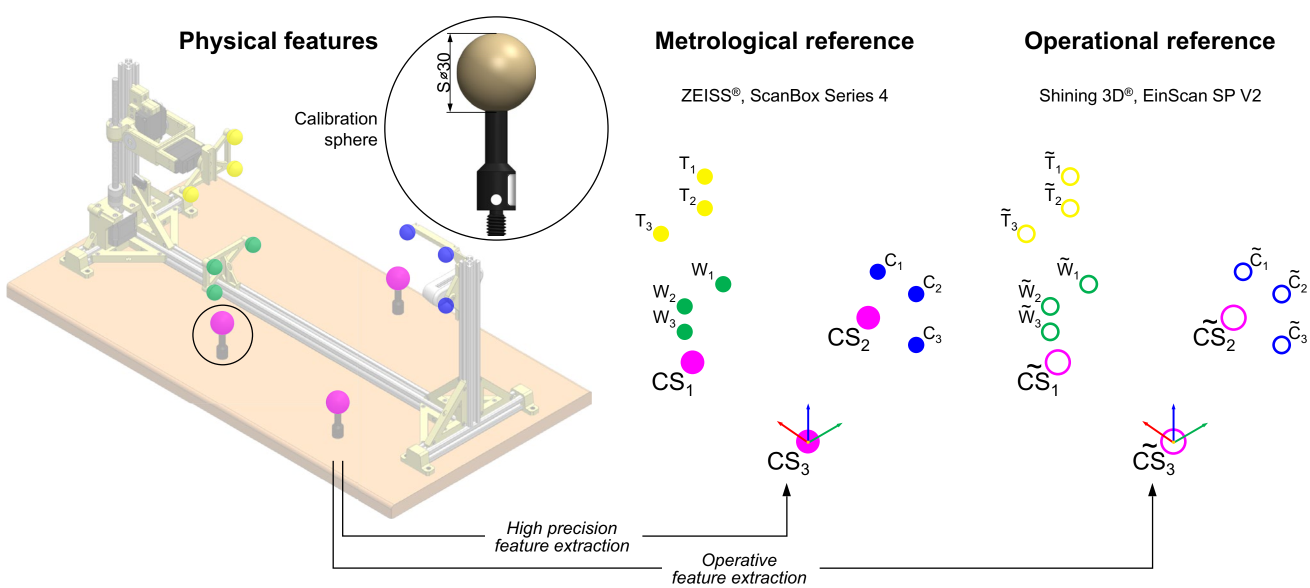


Fig. 5 Metrological qualification process: spherical features are independently acquired by the metrology-grade system (ATOS ScanBox 4105) and the operational GT scanner. Three calibration spheres define a local reference frame used for the comparison of the measured features

where $u^{\mathrm{op}}_{\Delta q}$ was estimated from the repeated acquisitions as $u^{\mathrm{op}}_{\Delta q} = s(\Delta q)$, while $u^{\mathrm{met}}_{\Delta q}$ represents the uncertainty contribution of the metrology-grade system, treated as a type-B component derived from its declared performance (or from repeated reference acquisitions when available).

Finally, the expanded uncertainty at a conventional 95% coverage level was computed as:

$$U_{\Delta q} = k\, u_{\Delta q}, \qquad k \approx 2, \tag{5}$$

where k is the coverage factor. When required, k can be determined from the Student's t distribution as $k = t_{0.975,\nu}$, with ν degrees of freedom estimated through the Welch–Satterthwaite equation [63].

For the sake of completeness, an aggregated 3D expanded uncertainty was computed by combining the component-wise combined standard uncertainties as:

$$U_{3D} = k\sqrt{u^2_{\Delta x} + u^2_{\Delta y} + u^2_{\Delta z}}. \tag{6}$$

Tested Tracking Methods

In the present study, the Intel® the RealSense™ D435i camera [64] was adopted for its multiple sensors, which facilitate the development of various integrated tracking methods. For the sake of completeness, some details of the camera are provided in Fig. 6. RGB image sensor OmniVision OV2740 achieves an image resolution of 1920×1080 pixels, with a $16:9$ aspect ratio, and is capable of outputting 10-bit RAW RGB data. Specifically we employed the YUY2 stream format, at a frame rate of 30FPS. Optically, the camera features an aperture of *f*/2.0 and a focal length of $1.88\mathrm{mm}$. The focus mechanism is fixed, and fields of view are $69.4°$ horizontally, $42.5°$ vertically, and $77°$ diagonally, with distortion specified to be $\leq 1.5\%$.

The depth module integrates two identical stereo infrared sensors with a baseline of $50\mathrm{mm}$ and an IR projector. Each OmniVision OV9282 IR sensor captures images at a resolution of 1280×800 pixels, with an $8:5$ aspect ratio, outputting in 10-bit RAW format. These imagers feature fixed focus, an aperture of *f*/2.0, and a focal length of

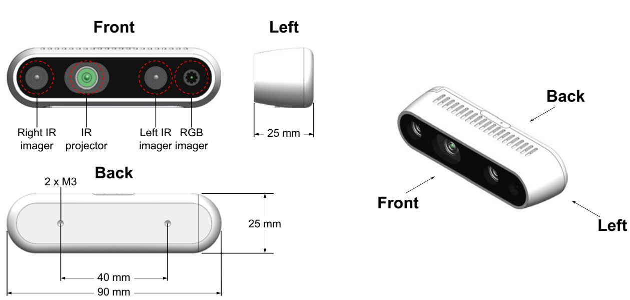


Fig. 6 Intel® the RealSense™ D435i camera details

$1.93\mathrm{mm}$. The fields of view is $91.2°$ horizontally, $65.5°$ vertically, and $100.6°$ diagonally, maintaining distortion at $\leq 1.5\%$. The infrared projector utilizes a static pattern and a vertical-cavity surface-emitting laser (VCSEL), with optics for illumination, and enhances depth perception and object detection in diverse applications.

Three AR tracking methods were benchmarked using the testing platform as shown in Fig. 7.

The first one employs the RGB video source of the camera to track a pattern-based marker. Additionally, tool tracking was developed using C# within the Unity engine, integrated with the Software Development Kit (SDK) Vuforia Engine [65] for augmented and virtual reality, which utilizes proprietary computer vision techniques, along with the graphical Engine for real-time recognition and tracking of planar and three-dimensional image-based markers. An image target was generated using the image generating software Fooocus [66]. Since all tracking logic is implemented within the Vuforia library, the only accessible information is the position and orientation of the tracked object relative to the camera, thus the pose estimation technique remains hidden and proprietary to Vuforia.

The Unity project comprises a virtual scene containing the following main objects: (i) an ARCamera object, representing the virtual camera used by Vuforia to track objects, replicating the behavior of a real camera; (ii) an ImageTargetBehaviour object representing the Image Target to be tracked, containing: (a) a Mesh object, representing the 3D model of the instrument equipped with the Image Target; (b) a Tip object, a 5 mm radius sphere placed at the instrument's tip.

The virtually tracked targets thus replicate the position and orientation of their real-world counterparts, along with all child objects, such as the 3D model of the instrument and the Tip sphere. Additionally, by observing the virtual scene through the ARCamera during the application's execution, it is possible to observe real-time tracking of virtual objects overlaid on the real objects in the camera's video feed.

The second method employed a combination of RGB and depth video sources. The tracked tool was equipped with four spherical passive markers, painted with fluorescent yellow paint to enhance contrast and detectability from the surrounding environment. The RGB video stream was converted in real-time to Hue Saturation Value (HSV) to manually threshold the H, S, and V values. Thresholding was employed to selectively isolate the spherical bodies of the tools and create a binary mask to exclude everything else. Through a blob detection algorithm, the 2D regions in the binarized mask frame corresponding to a value of 1 were identified, representing the potential tool markers. By applying the minimum enclosing circle algorithm and filtering out extreme values, the centers and radii of the minimum

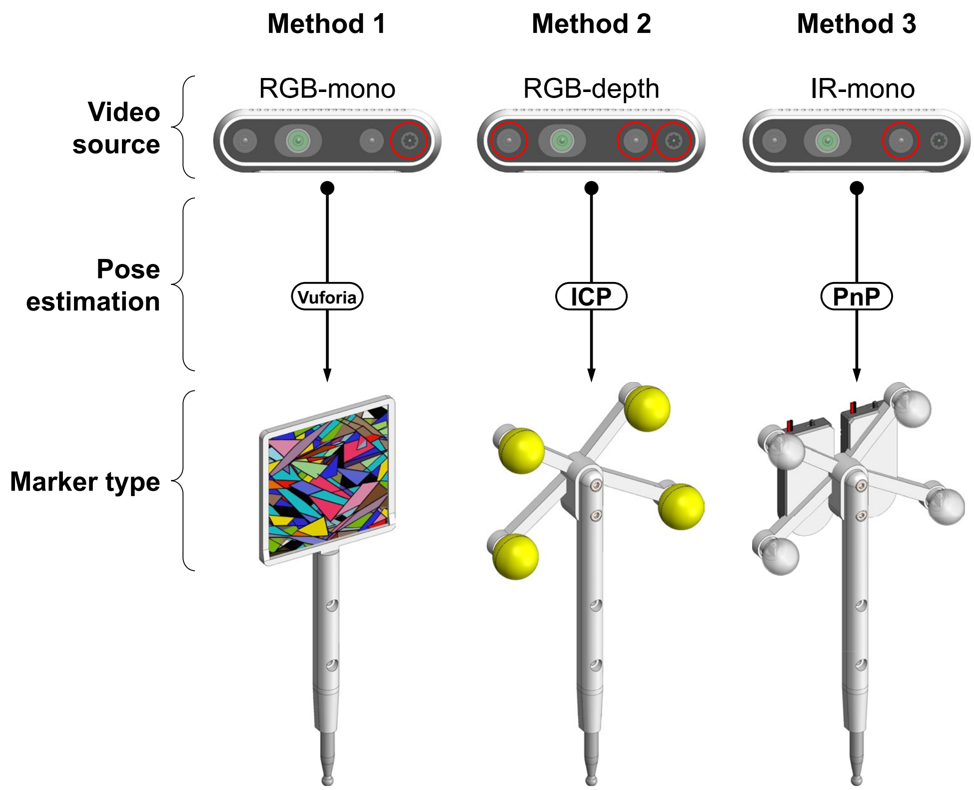


Fig. 7 Tested AR tracking systems methods: **a** RGB-mono video source, Vuforia pose estimation and image target marker; **b** RGB-depth video source, Iterative Closest Point (ICP) pose estimation, and passive spherical marker; **c** IR-mono video source, Perspective-n-Point pose estimation, and active IR spherical marker

enclosing circles of the blobs were obtained in the *uv* space of the image. Aligning the depth video with the RGB video allowed us to find the depth value corresponding to the centers of the detected circles. Thus, with the *u*, *v*, and *z* coordinates, i.e., the depth value, of each center, the complete 3D position of each spherical marker was determined relative to the IR left lens reference system. Knowing the coordinates of the marker centers and the tooltip from the CAD model, the ICP algorithm was used to align the theoretical center pattern in each frame to the tracked one, resulting in the transformation matrix T_T^L of the estimated tool pose.

The third method employed IR mono video source. The tracked tool was equipped with four spherical active IR markers, consisting of an IR LED enclosed in a spherical transparent shell. The IR stream grayscale frames were thresholded by manually adjusting the brightness value to selectively isolate the spherical bodies of the tools and create a binary mask. Also in this case, a blob detection algorithm was applied to find the areas of the potential markers, followed by the tracking of the minimum enclosing circles. Using the pose estimation function `solvePnP` from OpenCV, it was possible to compute the transformation matrix T_T^L of the tool relative to the IR left lens RF, utilizing the *uv* coordinates of the detected markers, their corresponding 3D coordinates relative to the Tool RF, and the camera calibration parameters.

The second and third methods were implemented in Python, utilizing the OpenCV library. For the first method,

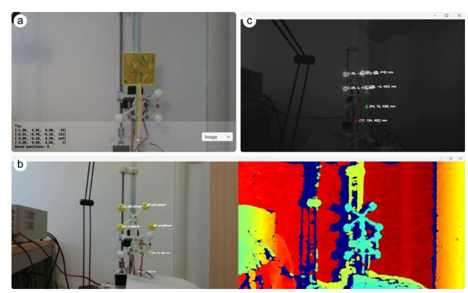


Fig. 8 Real-time tracking feedback visualization. **a** Method 1: Unity application interface with real-time updates of the T_T^L transformation and tool hologram augmentation. **b** Method 2: On the left, the augmented content is shown as visualization feedback on the RGB video, and on the right, the corresponding depth stream. **c** Method 3: tracking feedback is provided on the IR video stream

the developed application provided in real-time the transformation matrix T_T^L and the hologram of the tool, overlaid in transparency onto the real tool (Fig. 8a). For the other two methods, the visualization feedback was available on a monitor in real-time, including smaller circles to indicate the centers of the detected markers, points on each marker that are part of the aligned pattern, and a circle on the tip, as depicted in Fig. 8b, c.

Results

Ground-Truth Metrological Qualification

Following the qualification procedure described in Paragraph 3.2.3, the accuracy of the operational GT acquisition chain was assessed against the metrology-grade reference system. Discrepancy samples were computed in the local calibration reference frame and aggregated over all spherical features, configurations, and repetitions to estimate bias, mean absolute deviation, and repeatability, and to derive the associated standard and expanded uncertainty bounds of the operational GT.

Table 1 reports the bias *m*, the mean absolute discrepancy m_{abs}, the repeatability *s*, the combined standard uncertainty $u_{\Delta q}$, and the expanded uncertainty $U_{\Delta q}$ (95%, $k=2$) for each Cartesian component. An aggregated 3D expanded uncertainty U_{3D} (95%, $k=2$), computed as in Eq. (6), is also reported as a compact summary metric.

The bias terms *m* remain limited in magnitude compared to the dispersion component *s*, indicating the absence of relevant systematic offsets within the explored workspace. The mean absolute discrepancies m_{abs} are consistent with the observed repeatability levels, confirming that random variability represents the dominant contribution to the uncertainty budget. The expanded uncertainty bounds $U_{\Delta q}$ define the intrinsic confidence interval of the operational GT acquisition chain. Accordingly, tracking performance differences approaching values below approximately 0.7–0.8 mm should be interpreted in light of U_{3D}.

Table 1 Metrological qualification of the operational GT acquisition chain, aggregated over spheres, configurations, and repetitions

Component	*m* / mm	m_{abs} / mm	*s* / mm	$u_{\Delta q}$ / mm	$U_{\Delta q}$ / mm
Δx	0.011	0.170	0.225	0.225	0.450
Δy	−0.077	0.190	0.240	0.240	0.480
Δz	0.0023	0.155	0.195	0.195	0.391
Aggregated U_{3D} (95%, $k=2$) / mm					0.766

m: bias, m_{abs}: mean absolute discrepancy, s: repeatability, $u_{\Delta q}$: combined standard uncertainty, $U_{\Delta q}$: expanded uncertainty (95%, $k=2$)

Static Test

General statistics of FLE and FRE are reported in Table 2. Bar plots of FLE and FRE are depicted in Fig. 9. The first tracking method, based on the Vuforia Engine, does not support the FLE and FRE metrics, as it does not provide marker position information. For each configuration, the average values of the 300 collected frames are reported along with the overall mean. A FLE value of 2.334 ± 0.621 mm and 1.544 ± 0.571 mm was estimated for method 2 and 3, respectively, while the FRE was equal to 2.030 ± 1.532 mm and 0.210 ± 0.074 mm, in order.

Table 2 General statistics of FLE and FRE metrics

	FLE / mm				FRE / mm			
	Method 2		Method 3		Method 2		Method 3	
C	m	s	m	s	m	s	m	s
1	2.050	0.204	1.186	0.020	1.315	0.132	0.155	0.009
2	2.027	0.187	1.161	0.014	1.092	0.156	0.291	0.013
3	1.977	0.218	1.208	0.015	1.812	0.132	0.212	0.007
4	2.949	0.252	2.656	0.035	3.283	2.857	0.107	0.008
5	2.587	0.877	1.510	0.017	2.825	0.618	0.287	0.036
	2.334	0.621	1.544	0.571	2.030	1.532	0.210	0.074

C configuration, *m* mean, *s* standard deviation

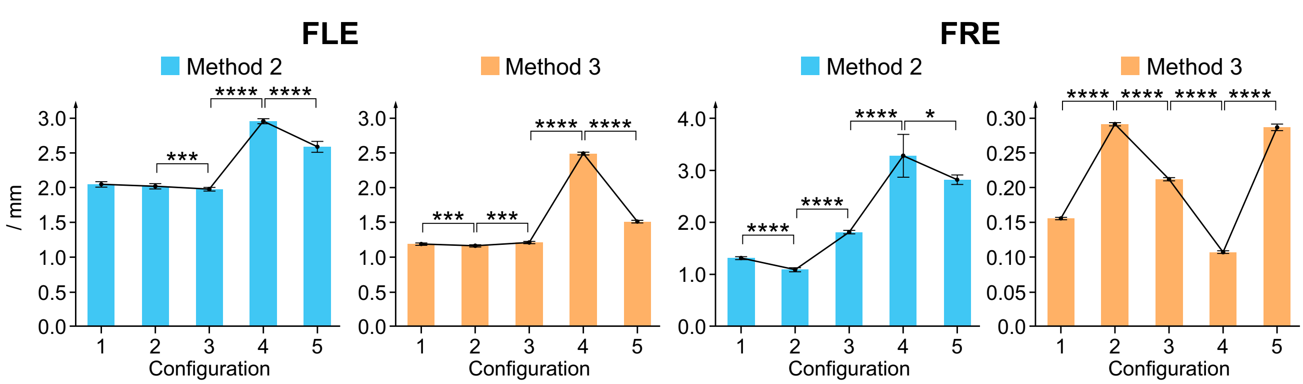


Fig. 9 FLE and FRE bar charts: each bar represents the mean and its 95% confidence interval; the smaller the *p*-value is than 0.05, the greater the number of asterisks

Table 3 General statistics of TRE metrics

		TRE_x / mm		TRE_y / mm		TRE_z / mm		$\|TRE\|$ / mm	
C	M	m	s	m	s	m	s	m	s
1	1	−2.341	0.015	−11.391	0.008	−0.562	0.145	11.682	0.254
2		−0.793	0.025	−13.291	0.022	−0.273	0.150	13.247	0.468
3		−3.277	0.097	−8.638	0.067	5.718	0.301	11.033	0.975
4		−2.084	0.193	−17.382	0.357	−4.637	0.177	18.117	0.348
5		−4.487	0.050	−0.946	0.046	−16.467	0.114	17.084	0.100
		−2.596	1.239	−10.330	5.491	−3.244	7.396	14.207	2.908
1	2	−3.115	0.272	−1.425	0.227	0.051	1.859	3.861	0.634
2		−1.657	0.338	−2.313	0.328	−0.550	2.120	3.521	0.836
3		−3.471	0.283	−0.038	0.190	−1.034	2.160	4.151	0.805
4		−2.138	0.663	−4.055	0.466	−1.022	2.043	5.109	0.877
5		−3.917	0.864	0.631	0.552	4.160	1.854	5.972	1.363
		1.470	0.073	1.152	0.026	−8.355	0.362	4.488	1.291
1	3	1.470	0.073	1.152	0.026	−8.355	0.362	8.561	0.362
2		0.842	0.040	0.504	0.078	0.810	0.339	1.317	0.077
3		0.810	2.057	−2.308	1.393	0.625	4.183	3.813	1.153
4		0.794	0.043	1.370	0.160	0.217	0.307	1.623	0.205
5		0.272	0.326	1.056	0.190	1.289	0.413	1.719	0.456
		1.247	0.907	0.570	0.956	−1.893	3.694	3.578	2.836

C configuration, *M* method, *m* mean, *s* standard deviation

Table 4 General statistics of TOE metrics

		TOE_x / °		TOE_y / °		TOE_z / °		$\|TOE\|$ / °	
C	M	m	s	m	s	m	s	m	s
1	1	7.324	0.192	3.681	0.116	8.158	0.181	11.566	0.258
2		2.637	0.159	1.794	0.182	3.016	0.217	4.391	0.311
3		7.975	0.055	10.473	0.569	13.032	0.436	18.527	0.617
4		0.861	0.065	0.823	0.090	1.038	0.085	1.582	0.117
5		0.975	0.080	1.250	0.052	1.215	0.133	1.999	0.149
		3.959	3.093	3.602	3.578	5.295	4.653	7.616	6.534
1	2	1.342	0.305	1.218	0.324	1.190	0.549	2.210	0.559
2		1.652	0.618	1.399	0.555	2.180	0.509	3.149	0.686
3		1.080	0.322	1.236	0.512	1.274	0.574	2.127	0.697
4		1.664	0.506	1.784	0.556	2.184	0.524	3.325	0.708
5		1.626	0.240	1.327	0.264	1.417	0.376	2.556	0.383
		1.472	0.485	1.390	0.504	1.657	0.676	2.677	0.788
1	3	2.430	0.022	4.023	0.054	3.684	0.059	5.972	0.075
2		4.233	0.060	4.876	0.053	6.060	0.059	8.856	0.082
3		8.433	0.139	2.430	0.070	8.121	0.161	11.957	0.215
4		3.618	0.071	10.250	0.172	10.811	0.140	15.331	0.203
5		1.672	0.023	7.555	0.061	7.368	0.061	10.684	0.087
		4.077	2.356	5.827	2.768	7.209	2.352	10.560	3.125

C configuration, *M* method, *m* mean, *s* standard deviation

Regarding the TRE and TOE, metric general statistics are collected in Table 3 and 4 respectively, while TRE bar charts are grouped in Fig. 10. Method 3 outperformed the other two, resulting in a euclidean value, $\|TRE\|$, of 3.578 ± 2.836 mm, compared to 14.207 ± 2.908 mm and 4.448 ± 1.291 mm for method 1 and 2, respectively. More specifically, the largest component of TRE in method 1 is TRE_y, equal to -10.330 ± 5.491 mm, while in methods 2 and 3, it is TRE_z, equal to -8.355 ± 0.362 mm and -1.893 ± 3.694 mm, respectively.

With respect to TOE, method 2 had the lowest $\|TOE\|$ value of $2.677 \pm 0.788°$, whereas, for method 1 and 3, TOE was equal to $7.616 \pm 6.534°$ and $10.560 \pm 3.125°$, respectively.

Minitab software was employed to conduct the statistical analysis. Non parametric Kruskal–Wallis statistic was performed to find the statistical significance of configuration

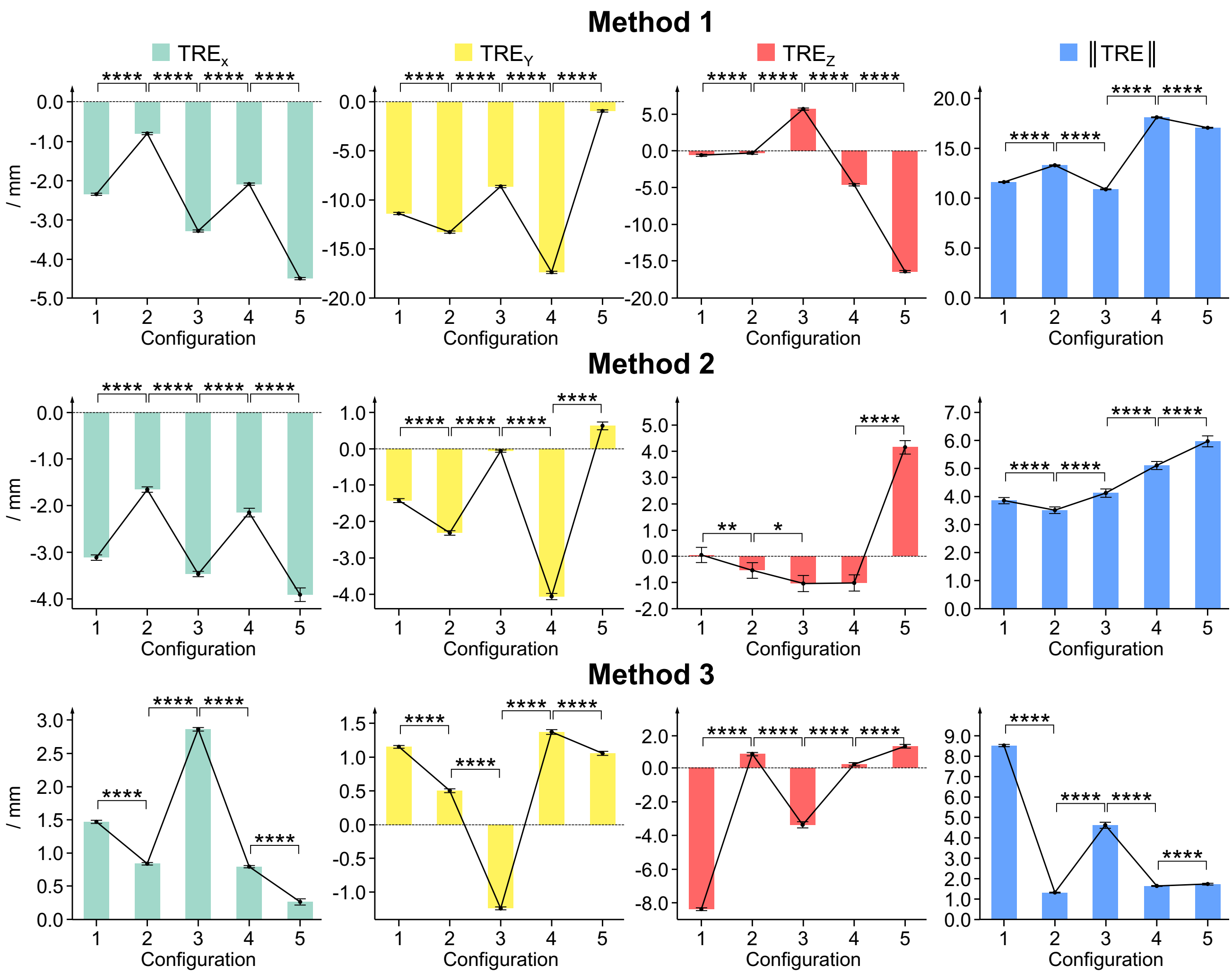


Fig. 10 TRE bar charts: each bar represents the mean and its 95% confidence interval; the smaller the *p*-value is than 0.05, the greater the number of asterisks

Table 5 General statistics of TJM metrics

	TJM_x / mm		TJM_y / mm		TJM_z / mm		$\|\mathrm{TJM}\|$ / mm	
M	m	s	m	s	m	s	m	s
1	0.456	0.435	0.599	0.871	1.062	0.435	1.435	0.824
2	2.896	1.593	2.111	0.925	12.015	0.863	12.651	0.762
3	0.779	0.725	0.690	0.390	3.202	2.432	3.473	2.385

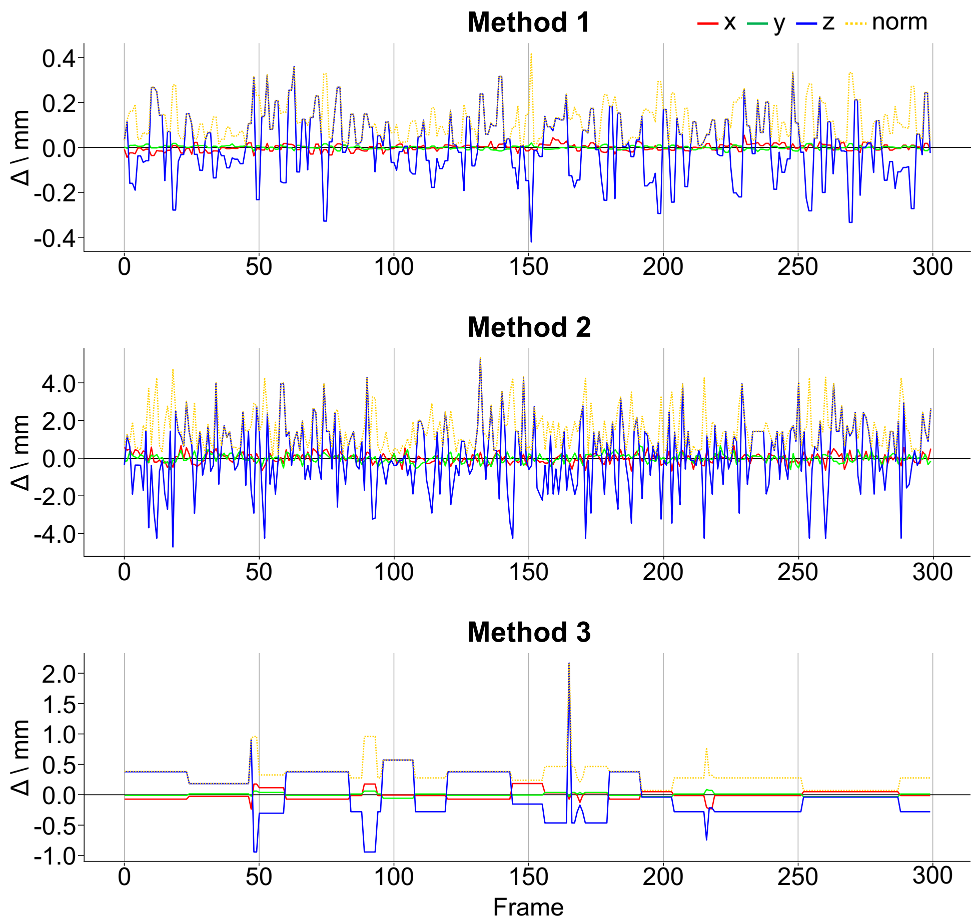
M method, *m* mean, *s* standard deviation

change in FLE, FRE, TRE, and TOE variability: H_0 assumes that there are no differences between the means of the individual configurations. In all cases, the calculated *p*-value is significantly less than the conventional value of 0.05, thus rejecting H_0 and demonstrating the significance of the configuration. Furthermore, a 2-sample *t*-test was performed between each configuration: H_0 assumes that the means of each configuration are equal.

Stability and Dynamic Test

Estimated TJM values are reported in Table 5. It should be noted that TJM has only one value for each configuration, as it accounts for the variability across all frames. Therefore, the values reported are the mean of all configurations and the corresponding standard deviation. Method 1 exhibited the greatest stability with a $\|\mathrm{TJM}\|$ of 1.435 ± 0.824 mm followed by method 3 with 3.473 ± 2.385 mm, both significantly lower than method 2, with an estimated $\|\mathrm{TJM}\|$

Fig. 11 Point jitter on first configuration



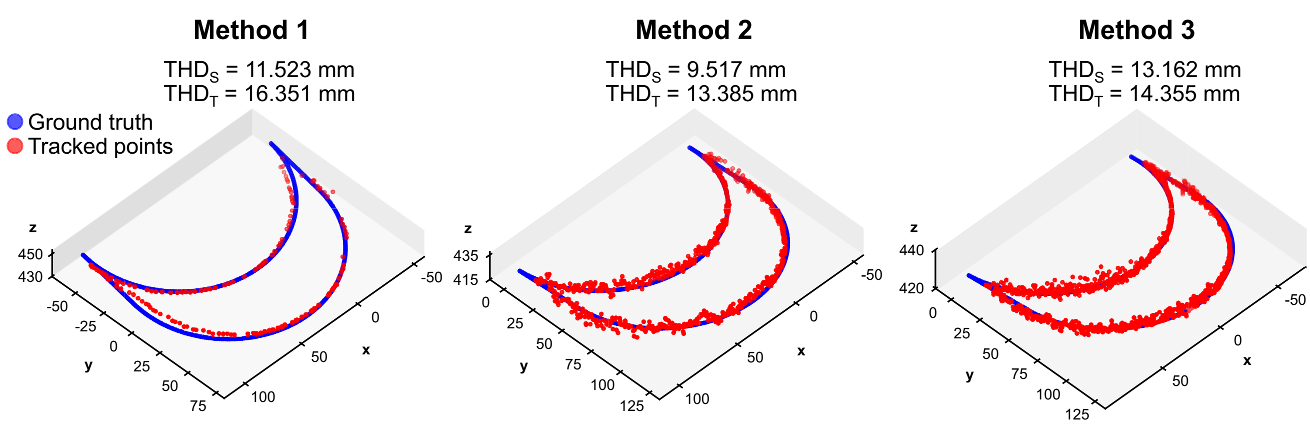


Fig. 12 Dynamic test tracked trajectories

of an order of magnitude greater, i.e., 12.651 ± 0.726 mm. In Fig. 11, details of tip coordinates fluctuations for each method are reported for the first configuration. For each method, greater fluctuation values for the z component are observed. The greater stability along recorded frames of the first method can be appreciated, having an order of magnitude lower fluctuation value along the z axis compared to the second one.

Details of dynamic test tracked trajectories are reported in Fig. 12: only mean values are reported as they were obtained from the average of all frames. Method 1 had the highest THD_S value at 11.523 mm and the highest THD_H value at 16.351 mm, while method 2 had the lowest THD_S value at 9.157 mm, and method 3 had the lowest THD_H value at 14.355 mm.

Discussion

For augmented surgery to be effective, it is essential to continuously track anatomical districts or surgical tools to align the virtual content seamlessly with the real-world environment [67]. Moreover, significant advancements have been made, encompassing the proposing of novel tracking systems [68, 69] and the evaluation in different working conditions [70]. Moreover, accurate and precise pose estimation of tools with tips is required [48] to identify specific points of the patient or perform point-based registration of virtual content [71, 72] on the anatomical area of interest [71]. Several studies focused on individual aspects of AR systems, such as marker detection accuracy [73] or pose estimation algorithms, without providing a standardized method for evaluating overall performance in surgical contexts. This study addresses this gap by presenting a benchmarking platform designed to evaluate the performance of different optical tracking methods suitable for AR applications.

The focus of this study is the definition of platform specifications and procedures that standardize evaluation across tracking pipelines. Three tracking methods were evaluated as case studies to showcase the benchmarking procedure applied to AR tool tracking using the proposed platform.

A cost-effective benchmarking platform was developed and proposed in this study, encompassing both a semi-automatic tool positioning structure and a performance assessment methodology. The structure allows for reliable and repeatable tool positioning, and its controlled movement along a path, combined with scene acquisition enabled by the camera attached to the structure. Moreover, the evaluation methodology leverages a series of quantitative metrics and testing procedures for benchmarking tracking methods. More in detail, for the metrics that require it, a procedure for accurately estimating the GT has been defined, employing a 3D scanner with an acquisition accuracy below $0.1\ \mathrm{mm}$.

Within a benchmarking framework, the reliability of the GT reference directly affects the interpretability of the measured tracking performance. For this reason, the operational GT acquisition chain was independently qualified against a metrology-grade optical CMS, as detailed in Paragraph 3.2.3. The qualification procedure assessed the coordinate consistency within a local calibration frame, thus evaluating the effective pose measurement capability of the operational pipeline rather than relying solely on nominal scanner specifications.

The resulting expanded uncertainty bounds at 95% coverage ($k = 2$) were $0.450\,\mathrm{mm}$, $0.480\,\mathrm{mm}$, and $0.391\,\mathrm{mm}$ along the three Cartesian components, leading to an aggregated value of $U_{3D} = 0.766\,\mathrm{mm}$. Incorporating both random variability and systematic contributions according to the GUM framework, the expanded uncertainty defines the intrinsic confidence interval of the operational GT within the considered workspace. Accordingly, the adopted GT can be regarded as suitable for comparative benchmarking in the millimetric regime. The aggregated expanded uncertainty U_{3D} provides a conservative confidence bound for the interpretation of the results: performance differences substantially exceeding this value can be considered clearly distinguishable, while smaller differences remain measurable but should be interpreted taking into account the inherent measurement uncertainty.

Method 3 demonstrated better accuracy in terms of TRE with a euclidean value of $3.578 \pm 2.836\ \mathrm{mm}$. However, method 1 exhibits better precision in terms of tip stability, achieving a TJM euclidean value of $1.062 \pm 0.435\ \mathrm{mm}$; nonetheless, method 3 also achieves a contained value of $3.473 \pm 2.385\ \mathrm{mm}$, both significantly outperforming method 2, which is characterized by higher instability ($12.651 \pm 0.762\ \mathrm{mm}$). In contrast, method 2 seems to have slightly better behavior during dynamic tests, where THD_S and THD_T are equal to $9.517\ \mathrm{mm}$ and $13.385\ \mathrm{mm}$, respectively. Overall, method 3 achieves a better balance between accuracy and precision in static tests. Meanwhile, method 2 demonstrates better performance in dynamic tests.

From a clinical perspective, these results should be interpreted with respect to accuracy levels commonly considered acceptable in surgical navigation. Several studies report that positional errors on the order of a few millimeters and angular errors of a few degrees, typically up to approximately $\pm 5\,\mathrm{mm}$ and $\pm 5^\circ$, are regarded as clinically acceptable, with reference to application domains including orthopedic tumor resection [74], AR-guided orthopedic implant positioning [75], dental implant surgery [76], and pediatric oncologic surgical navigation [77].

Accordingly, the error values obtained in this study fall within the tolerance ranges reported in the literature, indicating compatibility with clinically acceptable accuracy requirements.

The superior performance of method 3, which utilizes active IR markers [78, 79] and the PnP algorithm, indicates its potential for precise surgical applications. This can be attributed to the use of active markers, which mitigate issues such as lighting variability, common in other tracking methods. Moreover, due to the wider field of view of the IR sensor, the same displacement along the optical axis, captured with the RGB camera, is affected by greater distortion due to lens aberration compared to the IR camera. This reflects in the ICP algorithm effectiveness, which depends on the values of the reference point cloud coordinates. On the other hand, the richness of corners and detectable features of the image target itself could have positively impacted the higher tip stability of the first method, while the planar geometry is penalizing for tool placements where the image target is not facing the RGB sensor [80]. This effect can be appreciated in Fig. 10, where the TRE_z increases the 4 and 5 configurations, where the image target is tilted relative to the camera. Conversely, the significant errors observed in method 1 during dynamic scenarios suggest that image-based tracking with Vuforia may not be suitable for real-time surgical guidance, where accurate and stable tracking is critical. Significantly higher values of THD_T compared to THD_S indicate a scarcity of points collected in the tracked trajectory or a lack of continuity, meaning accumulations of points in a specific area and absence in others as mathematically defined by the metric. Equal THD_S values indicate that the compared curves have the same shape and spatial point distribution but are rotated and/or shifted relative to each other.

In literature, several studies compare different tracking methods through specific test setups and evaluation methodologies. In 2022, Marino et al. [81] developed a benchmarking framework to evaluate various mobile-based AR tracking methods. Specifically, a planar pattern-based marker is tracked by different mobile devices using the following experimental setup: the tracking device is mounted on a tripod and the tracked marker is placed in front of it. Both are equipped with retro-reflective markers, allowing for GT estimation by an OptiTrack motion capture system. Meanwhile, the mobile device is moved along a predefined path and its trajectory is recorded using both the tracked marker and the OptiTrack system. The proposed and evaluated metrics encompass Absolute Trajectory Error ATE and Relative Pose Error, which respectively evaluate the overall consistency and the relative shift of recorded trajectories. The reported results showed average ATE values of $0.121\,\mathrm{m}$ for the iPad Pro and $0.162\,\mathrm{m}$ for the Xiaomi Mi 11 Lite.

A year later, Jurado et al. [80] conducted a comparative analysis of planar pattern-based markers. The authors used an experimental setup with a camera mounted on a tripod and sequentially positioned in a grid of locations. The camera captures the marker of the evaluated tracking method, using a series of ArUco markers scattered around as GT, following the methodology detailed in [82]. The metrics used for the evaluation are the True Positive Rate, expressing the rate of correctly detected markers in various positions both free of and under occlusion, and the vertex jitter magnitude as the deviation from the mean position, analogous to the TJM metric defined in this work. In terms of vertex jitter, ArUco3 achieved errors below $0.08\,\mathrm{m}$ even at 6 m, while QR Codes exceeded $0.3\,\mathrm{m}$.

Focusing on tested methods quantitative results, Stenmark et al. (2022) [83], manufactured a multi pattern-based marker die fitted on a surgical DeBakey forceps tracked by an off-the-shelf camera. Authors measured an average translational error of $2.6\ \mathrm{mm}$,and rotational error was $1.1°$. Differently from our work, these accuracy estimates were made for the marker die centroid, and not the tip of the instrument.

In the same year Thabit et al. [84] performed an evaluation of the absolute translational error of planar pattern-based marker centroids studying the effects of marker size, marker-camera distance, and camera resolution, achieving best performances at $50\ \mathrm{cm}$ and $10\ \mathrm{cm}$ marker size with a mean translation error of $1.8 \pm 0.6\ \mathrm{mm}$.

An IR tracking system employing different tools equipped with active markers, i.e., IR LEDs, and a tip was developed by Lin et al. [85] in 2016. They compared the relative distance of tracked markers with real ones measured with a caliper, collecting a root mean square error of less than $0.3\ \mathrm{mm}$. The relative distance between the tracked tip and markers was $0.025\ \mathrm{mm}$, and the tracking error of adjacent corner points of an A4 checkerboard with $14 \times 14\ \mathrm{mm}$ squares was less than $0.2714\ \mathrm{mm}$.

Some limitations of this study are highlighted in the following. Firstly, the benchmarking platform was tested in a controlled environment, which may not fully replicate the complexities of actual surgical settings. On the one hand, maintaining constant environmental conditions is partially justified by operating practices, where lighting is controlled. Besides, a potential drawback of the custom-built setup is that results can be affected by different hardware components and structural assembly. However, consistent results are achievable when the same platform is used throughout the experiments.

In the tested configurations, although large variations in the tool pose were considered, it might be interesting to test a larger working volume Future studies should consider a broader range of hardware configurations and real-world testing scenarios to validate the platform's effectiveness comprehensively.

In addition to these setup-related limitations, it is also worth considering that AR system errors do not originate solely from the tracking component. Various influence factors [86] can be identified and grouped into several families, as summarized in Fig. 13. A first family relates to AR interaction dynamics and ergonomics, including the type of device (wearable or handheld), the feedback mechanisms, and the distance between the tracked tool and the tracking device. A second family concerns the tracking system setup, such as camera configuration, resolution, algorithmic approach, and marker arrangement [87]. A third family refers to digitalization procedures, influenced by tool orientation, operator handling, and object pose. Another family refers to the interaction object, whose geometry and accessibility can influence the measurement [88, 89]. Finally, environmental aspects such as illumination, line of sight, and background conditions can also affect the tracking process.

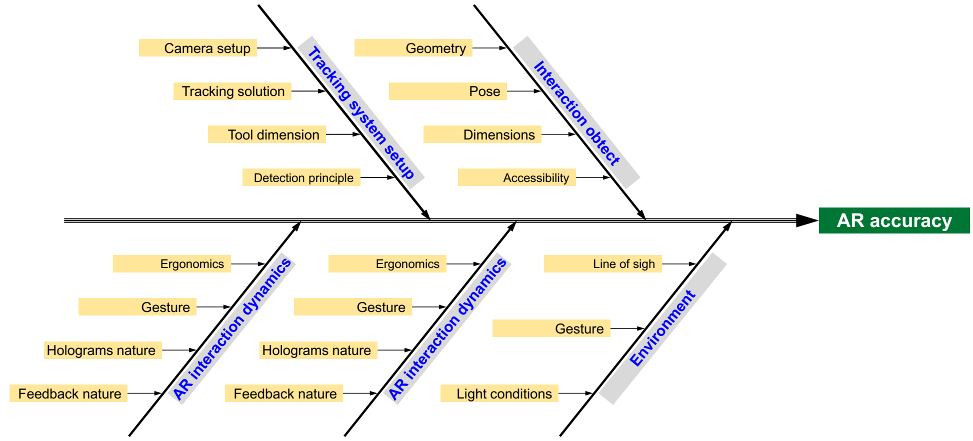


Fig. 13 Ishikawa diagram of influence factors in AR-based tracking

Infrared optical systems are often used in surgical contexts to mitigate the influence of visible light variability. This study focused specifically on tracking accuracy, whereas the other families highlight complementary error sources that can be addressed separately.

The platform is intended as a benchmarking environment that can incorporate alternative sensors, detection strategies, and additional metrics under the same standardized procedures. This study opens several opportunities for further research. Future studies could expand the work envelope, as the platform supports such enhancements due to its flexible and modular design, which includes standardized components. Additionally, in identifying the most suitable tracking method for Augmented Surgery, it is suggested to benchmark different alternatives, such as targets using the Vuforia engine, and to explore three-dimensional targets, such as cubic and cylindrical shapes. Enhancing the integration of RGB and depth data in tracking systems, as demonstrated by the performance of method 2, presents a valuable area for further exploration. Furthermore, investigating hybrid tracking approaches that combine the strengths of different modalities could improve overall performance. Clinical trials in real-world settings are necessary to validate the findings and ensure their practical application in surgical environments.

The platform is intended as a benchmarking environment that can incorporate alternative sensors, detection strategies, and additional metrics under the same standardized procedures. This study therefore opens several opportunities for further research, enabled by the flexible and modular design of the framework. Future developments may include the extension of the work envelope, allowing the evaluation of tracking performance under a wider range of acquisition conditions. In the context of identifying the most suitable tracking methods for Augmented Surgery, different alternatives should be benchmarked, such as targets based on the Vuforia engine and three-dimensional geometries including cubic and cylindrical shapes. Enhancing the integration of RGB and depth information, as suggested by the performance of method 2, represents another promising direction, together with the investigation of hybrid tracking approaches that combine the strengths of multiple modalities to improve robustness and accuracy. Furthermore, future work will extend the proposed framework toward the analysis of naturalistic motions representative of clinical practice, including freehand tool movements characterized by variable orientations, distances, and motion velocities. While preserving a controlled acquisition setup, these tests will enable the evaluation of tracking stability and failure modes under realistic kinematic variability. In this perspective, the platform will serve as a quantitative pre-selection tool for identifying suitable tracking systems prior to their integration and validation in application-specific clinical workflows.

Conclusions

This study presents a novel benchmarking platform for AR tracking systems in surgery, providing a performance assessment throughout a set of quantitaive evaluation metrics. IR active spherical markers-based tracking method emerged as the most accurately balanced solution. Therefore, it is well-suited as a starting point for future research in the surgical context. These insights pave the way for future innovations and improvements, ultimately contributing to the advancement of AR in surgical practice. Future research should focus on real-world clinical trials, the integration of various tracking data, and the exploration of hybrid approaches to enhance overall performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42979-026-05071-9.

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement. This article was produced while attending the PhD programme in Gestione, Produzione e Design at Politecnico di Torino, Cycle XXXVIII, with the support of a scholarship co-financed by the Ministerial Decree no. 352 of 9th April 2022, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4C2, Investment 3.3, and by the Foundation ITS Biotecnologie e Nuove Scienze della Vita - Piemonte.

Data Availability Data sets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest All authors have no conflict of interest.

References

1. Azagury DE, Dua MM, Barrese JC, Henderson JM, Buchs NC, Ris F, et al. Image-guided surgery. Curr Probl Surg. 2015;52(12):476–520. https://doi.org/10.1067/j.cpsurg.2015.10.001.
2. Keereweer S, Kerrebijn JDF, Van Driel PBAA, Xie B, Kaijzel EL, Snoeks TJA, et al. Optical image-guided surgery-where do we stand? Mol Imag Biol. 2011;13(2):199–207. https://doi.org/10.1007/s11307-010-0373-2.
3. Carmigniani J, Furht B, Anisetti M, Ceravolo P, Damiani E, Ivkovic M. Augmented reality technologies, systems and applications. Multimedi Tools Appl. 2011;51(1):341–77. https://doi.org/10.1007/s11042-010-0660-6.
4. Salerno F, Ulrich L, Marullo G, Moos S, Vezzetti E. Advancing surgical cutting guide flexibility: a hybrid physical and augmented reality solution for cranio-maxillofacial surgery. J Craniomaxillofac Surg. 2026. (In press).
5. Kim SK, Kang S-J, Choi Y-J, Choi M-H, Hong M. Augmented-reality survey: from concept to application. KSII Trans Internet Inf Syst (TIIS). 2017;11(2):982–1004. https://doi.org/10.3837/tiis.2017.02.004.
6. Ulrich L, Salerno F, Moos S, Vezzetti E. How to exploit augmented reality (ar) technology in patient customized surgical tools: a focus on osteotomies. Multimed Tools Appl. 2024;83(1):783–806. https://doi.org/10.1007/s11042-023-18058-y.
7" GrobidStyleApplied="Changed. Gsaxner C, Li J, Pepe A, Schmalstieg D, Egger J. Inside-out instrument tracking for surgical navigation in augmented reality. In: Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology. 2021;1–11. https://doi.org/10.1145/3489849.3489863.
8. Maculotti G, Ulrich L, Olivetti EC, Genta G, Marcolin F, Vezzetti E, et al. A methodology for task-specific metrological characterization of low-cost 3d camera for face analysis. Measurement. 2022;200:111643. https://doi.org/10.1016/j.measurement.2022.111643.
9. Sanz PR, Mezcua BR, Peña J. Current Advancements in Stereo Vision. In: Zhang A, editor. IntechOpen, London; 2012:93–118. https://doi.org/10.5772/45904. (**Depth estimation-an introduction**).
10. Ulrich L, Vezzetti E, Moos S, Marcolin F. Analysis of rgb-d camera technologies for supporting different facial usage scenarios. Multimed Tools Appl. 2020;79(39):29375–98. https://doi.org/10.1007/s11042-020-09479-0.
11" GrobidStyleApplied="Changed. Fiala M. Artag, a fiducial marker system using digital techniques. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2. San Diego, CA, USA: IEEE; 2005:590–6. https://doi.org/10.1109/CVPR.2005.74.
12. Čepon G, Ocepek D, Kodrič M, Demšar M, Bregar T, Boltežar M. Impact-pose estimation using aruco markers in structural dynamics. Exp Tech. 2024;48(2):369–80. https://doi.org/10.1007/s40799-023-00646-0.
13" GrobidStyleApplied="Changed. Flohr D, Fischer J. A lightweight id-based extension for marker tracking systems. In: Eurographics Symposium on Virtual Environments (EGVE) Short Papers and Posters. The Eurographics Association, Goslar, Germany; 2007:59–64. https://doi.org/10.2312/PE/VE2007Short/059-064.
14. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ. Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recogn. 2014;47(6):2280–92. https://doi.org/10.1016/j.patcog.2014.01.005.
15. Van Liere R, Mulder JD. Optical tracking using projective invariant marker pattern properties. In: IEEE Virtual Reality. In: Proceedings, pp. 191–198. IEEE Computer Society, Los Angeles, CA, USA; 2003. https://doi.org/10.1109/VR.2003.1191138.
16. Odmins J, Slics K, Fenuks R, Linina E, Osmanis K, Osmanis I. Comparison of passive and active fiducials for optical tracking. Latv J Phys Tech Sci. 2022;59(5):46–57. https://doi.org/10.2478/lpts-2022-0040.
17. Wu H, Lin Q, Yang R, Zhou Y, Zheng L, Huang Y, et al. An accurate recognition of infrared retro-reflective markers in surgical navigation. J Med Syst. 2019;43(6):153. https://doi.org/10.1007/s10916-019-1257-x.
18. Ma L, Huang T, Wang J, Liao H. Visualization, registration and tracking techniques for augmented reality guided surgery: a review. Phys Med Biol. 2023;68(4):04–02. https://doi.org/10.1088/1361-6560/acaf23.
19. Geisbüsch A, Auer C, Dickhaus H, Niklasch M, Dreher T. Electromagnetic bone segment tracking to control femoral derotation osteotomy-a saw bone study. J Orthop Res. 2017;35(5):1106–12. https://doi.org/10.1002/jor.23348.
20. Pflugi S, Vasireddy R, Lerch T, Ecker TM, Tannast M, Boemke N, et al. Augmented marker tracking for peri-acetabular osteotomy surgery. Int J Comput Assist Radiol Surg. 2018;13(2):291–304. https://doi.org/10.1007/s11548-017-1690-6.
21. Han Z, Tian H, Vercauteren T, Liu D, Li C, Duan X. Collaborative human-robot surgery for mandibular angle split osteotomy: optical tracking based approach. Biomed Signal Process Control. 2024;93:106173. https://doi.org/10.1016/j.bspc.2024.106173.
22" GrobidStyleApplied="Changed. Salerno F, Moos S, Ulrich L, Novaresio A, Vezzetti E. A methodology for the dimensional and mechanical analysis of surgical guides. In: International Conference of the Italian Association of Design Methods and Tools for Industrial Engineering. Cham: Springer; 2023:184–93. https://doi.org/10.1007/978-3-031-52075-4_22.
23. Peh S, Chatterjea A, Pfarr J, Schäfer JP, Weuster M, Klüter T, et al. Accuracy of augmented reality surgical navigation for minimally invasive pedicle screw insertion in the thoracic and lumbar

spine with a new tracking device. The Spine J. 2020;20(4):629–37. https://doi.org/10.1016/j.spinee.2019.12.009.
24. Müller F, Roner S, Liebmann F, Spirig JM, Fürnstahl P, Farshad M. Augmented reality navigation for spinal pedicle screw instrumentation using intraoperative 3d imaging. The Spine J. 2020;20(4):621–8. https://doi.org/10.1016/j.spinee.2019.10.012.
25. Dho Y-S, Lee BC, Moon HC, Kim KM, Kang H, Lee EJ, et al. Validation of real-time inside-out tracking and depth realization technologies for augmented reality-based neuronavigation. Int J Comput Assist Radiol Surg. 2024;19(1):15–25. https://doi.org/10.1007/s11548-023-02993-0.
26. Marullo G, Innocente C, Ulrich L, Lo Faro A, Porcelli A, Ruggieri R, et al. Home-based mirror therapy in phantom limb pain treatment: the augmented humans framework. Multimed Tools Appl. 2025;84(28):34145–77. https://doi.org/10.1007/s11042-025-20628-1.
27. Andrews CM, Henry AB, Soriano IM, Southworth MK, Silva JR. Registration techniques for clinical applications of three-dimensional augmented reality devices. IEEE J Transl Eng Health Med. 2020;9:1–14. https://doi.org/10.1109/JTEHM.2020.3045642.
28. Bogdanova R, Boulanger P, Zheng B. Depth perception of surgeons in minimally invasive surgery. Surg Innov. 2016;23(5):515–24. https://doi.org/10.1177/1553350616639141.
29. Tsoy T, Safin R, Sultanov R, Saha SK, Magid E. Recommended criteria for qualitative comparison of fiducial markers performance. In: Proceedings of the 2022 International Siberian Conference on Control and Communications (SIBCON). Tomsk, Russia: IEEE; 2022:1–5. https://doi.org/10.1109/SIBCON56144.2022.10003018.
30. Salerno F, Contenti A, Ulrich L, Marullo G, Moos S, Vezzetti E. Mott modular optical tool tracking framework enabling efficient benchmarking. Sci Rep. 2025;15(1):37205. https://doi.org/10.1038/s41598-025-21044-z.
31" GrobidStyleApplied="Changed. Cesar S, Gaudig DB, Fritsche C, Reis M, Kirchner MA, F. An evaluation of artificial fiducial markers in underwater environments. In: OCEANS 2015 - Genova. Genova, Italy: IEEE; 2015:1–6. https://doi.org/10.1109/OCEANS-Genova.2015.7271491.
32. Kam HC, Yu YK, Wong KH. An improvement on aruco marker for pose tracking using kalman filter. In: 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Busan, South Korea: IEEE; 2018; 65–9. https://doi.org/10.1109/SNPD.2018.8441049.
33. Romero-Ramírez FJ, Muñoz-Salinas R, Medina-Carnicer R. Fractal markers: a new approach for long-range marker pose estimation under occlusion. IEEE Access. 2019;7:169908–19. https://doi.org/10.1109/ACCESS.2019.2951204.
34. Sarmadi H, Muñoz-Salinas R, Berbís MA, Medina-Carnicer RJ. Simultaneous multi-view camera pose estimation and object tracking with squared planar markers. IEEE Access. 2019;7:22927–40. https://doi.org/10.1109/ACCESS.2019.2896648.
35. Yu G, Hu Y, Dai J. Topotag: a robust and scalable topological fiducial marker system. IEEE Trans Vis Comput Graph. 2020;27(9):3769–80. https://doi.org/10.1109/TVCG.2020.2988466.
36. Popescu DC, Dumitrache I, Caramihai SI, Cernaianu MO. High precision positioning with multi-camera setups: adaptive kalman fusion algorithm for fiducial markers. Sensors. 2020;20(9):2746. https://doi.org/10.3390/s20092746.
37. Brand M, Wulff LA, Hamdani Y, Schüppstuhl T. Accuracy of marker tracking on an optical see-through head mounted display. In: Krüger J, Scholz-Reiter B, Wahl F, (eds) Annals of Scientific Society for Assembly, Handling and Industrial Robotics. Mechanisms and Machine Science, vol. 91. Cham: Springer; 2020:21–31. https://doi.org/10.1007/978-3-662-61755-7_3.
38. Zhang G, Song S, Panescu J, Shapiro N, Dannemiller KC, Qin R. A novel systems solution for accurate colorimetric measurement through smartphone-based augmented reality. PLoS ONE. 2023;18(6):0287099. https://doi.org/10.1371/journal.pone.0287099.
39. Jurado-Rodriguez D, Muñoz-Salinas R, Garrido-Jurado S, Romero-Ramirez FJ, Medina-Carnicer R. 3d model-based tracking combining edges, keypoints and fiducial markers. Virtual Real. 2023;27(4):3051–65. https://doi.org/10.1007/s10055-023-00853-5.
40. Kulikov V, Spirin E, Pikalov I, Saramud M. Methods for determining spatial position and production objects orientation based on the proposed fiducial markers for technical vision system. Int J Adv Manuf Technol. 2024;132(5):2549–62. https://doi.org/10.1007/s00170-024-13424-y.
41" GrobidStyleApplied="Changed. Ruppin J, Popovic A, Strauss M, Spüntrup E, Steiner A, Stoll C. Evaluation of the accuracy of three different computer-aided surgery systems in dental implantology: optical tracking vs. stereolithographic splint systems. Clin Oral Implant Res. 2008;19(7):709–16. https://doi.org/10.1111/j.1600-0501.2007.01430.x-i2.
42. Elfring R, Fuente M, Radermacher K. Assessment of optical localizer accuracy for computer aided surgery systems. Comput Aided Surg. 2010;15(1–3):1–12. https://doi.org/10.3109/10929081003647239.
43" GrobidStyleApplied="Changed. VDI/VDE: VDI/VDE 2634 Blatt 3 Optical 3D Measuring Systems—Multiple View Systems Based on Area Scanning. In: VDI Standard; 2016. https://www.dinmedia.de/en/technical-rule/vdi-vde-2634-blatt-3/109737809.
44. Hafez MA, Chelule KL, Seedhom BB, Sherman KP. Computer-assisted total knee arthroplasty using patient-specific templating. Clin Orthop Relat Res. 2006;444:184–92. https://doi.org/10.1097/01.blo.0000201148.06454.ef.
45. Matsukawa K, Kaito T, Abe Y. Accuracy of cortical bone trajectory screw placement using patient-specific template guide system. Neurosurg Rev. 2020;43(4):1135–42. https://doi.org/10.1007/s10143-019-01140-1.
46. Kral F, Puschban EJ, Riechelmann H, Freysinger W. Comparison of optical and electromagnetic tracking for navigated lateral skull base surgery. Int J Med Robot Comput Assist Surg. 2013;9(2):247–52. https://doi.org/10.1002/rcs.1502.
47. Kunz C, Maurer P, Kees F, Henrich P, Marzi C, Hlaváč M, et al. Infrared marker tracking with the hololens for neurosurgical interventions. Curr Dir Biomed Eng. 2020;6(1):20200027. https://doi.org/10.1515/cdbme-2020-0027.
48. Martin-Gomez A, Li H, Song T, Yang S, Wang G, Ding H, et al. Sttar: surgical tool tracking using off-the-shelf augmented reality head-mounted displays. IEEE Trans Vis Comput Graph. 2023. https://doi.org/10.1109/TVCG.2023.3238309.
49. Li H, Yan W, Liu D, Qian L, Yang Y, Liu Y, et al. Evd surgical guidance with retro-reflective tool tracking and spatial reconstruction using head-mounted augmented reality device. IEEE Trans Vis Comput Graph. 2025;31(9):6497–511. https://doi.org/10.1109/TVCG.2024.3518258.
50. Costa GM, Petry MR, Martins JG, Moreira APGM. Assessment of multiple fiducial marker trackers on hololens 2. IEEE Access. 2024;12:14211–26. https://doi.org/10.1109/ACCESS.2024.3356722.
51. Brunzini A, Mazzoli A, Pagnoni M, Mandolini M. An innovative mixed reality approach for maxillofacial osteotomies and repositioning. Virtual Real. 2023;27(4):3221–37. https://doi.org/10.1007/s10055-023-00867-z.
52. Moreta-Martinez R, Pose-Díez-de-la-Lastra A, Calvo-Haro JA, Mediavilla-Santos L, Pérez-Mañanes R, Pascau J. Combining augmented reality and 3d printing to improve surgical workflows in orthopedic oncology: smartphone application and clinical

evaluation. Sensors. 2021;21(4):1370. https://doi.org/10.3390/s21041370.
53. Akulauskas M, Butkus K, Rutkūnas V, Blažauskas T, Jegelevičius D. Implementation of augmented reality in dental surgery using hololens 2: an in vitro study and accuracy assessment. Appl Sci. 2023;13(14):8315. https://doi.org/10.3390/app13148315.
54. Maurer CR, Fitzpatrick JM, Wang MY, Galloway RL, Maciunas RJ, Allen GS. Registration of head volume images using implantable fiducial markers. IEEE Trans Med Imaging. 1997;16(4):447–62. https://doi.org/10.1109/42.611354.
55. Research P. Original Prusa i3 MK3S Technical Specifications. Online; 2025. https://www.prusa3d.com/product/original-prusa-i3-mk3s-3d-printer-3/. Accessed 18 Sep 2025.
56. Shining 3D: EinScan-SP Specifications. Online 2021. https://www.einscan.com/einscan-sp/einscan-sp-specs/.
57. Kanai T, Suzuki H. Approximate fitting of spheres and cylinders to range data. Comput Aided Des. 2001;33(4):235–46. https://doi.org/10.1016/S0010-4485(00)00128-6.
58" GrobidStyleApplied="Changed. Rusinkiewicz S, Levoy M. Efficient variants of the icp algorithm. In: Proceedings of the Third International Conference on 3-D Digital Imaging and Modeling (3DIM). IEEE, Quebec City. 2001:145–52. https://doi.org/10.1109/IM.2001.924423.
59. Slocum AH. Precision machine design. Dearborn: Society of Manufacturing Engineers; 1992.
60. ISO 10360–13:2021 Geometrical Product Specifications (gps)—Acceptance and reverification tests for coordinate measuring systems (CMS)—part 13: Optical 3D CMS. Standard, International Organization for Standardization, Geneva, CH; 2021.
61" GrobidStyleApplied="Changed. ISO 3290–1:2014 Rolling bearings—Balls—Part 1: Steel balls. Standard, International Organization for Standardization, Geneva, CH; 2014.
62" GrobidStyleApplied="Changed. Q-Mark Manufacturing Inc.: CAL-DK30 Grade 5 Ceramic Calibration Sphere, 30.0 mm Diameter. Online; 2026. https://www.cmms.com/products/cal-dk30. Accessed 01 Mar 2026.
63. JCGM 100:2008 Evaluation of measurement data—Guide to the expression of Uncertainty in Measurement (GUM). In: Standard, International Organization for Standardization, Geneva, CH; 2008.
64" GrobidStyleApplied="Changed. Intel Corporation: Intel RealSense D400 Series Datasheet. Online. Document Number: 336051–002; 2017. https://www.realsenseai.com/wp-content/uploads/2020/06/Intel-RealSense-D400-Series-Datasheet-June-2020.pdf.
65. Vuforia: Vuforia Engine Developer Portal. Online. Disponibile su; 2024. https://developer.vuforia.com/.
66. lllyasviel: Fooocus: Tool for Image Generation. GitHub repository. Disponibile su; 2024. https://github.com/lllyasviel/Fooocus.
67" GrobidStyleApplied="Changed. Milgram P, Takemura H, Utsumi A, Kishino F. Augmented reality: a class of displays on the reality-virtuality continuum. In: Telemanipulator and Telepresence Technologies, vol. 2351, pp. 282–292. SPIE, Philadelphia, PA, USA; 1995. https://doi.org/10.1117/12.197321.
68. Rabbi I, Ullah S. A survey on augmented reality challenges and tracking. Acta graphica znanstveni časopis za tiskarstvo i grafičke komunikacije. 2013;24(1–2):29–46.
69. Jun J, Yue Q, Qing Z. An extended marker-based tracking system for augmented reality. In: Second International Conference on Modeling, Simulation and Visualization Methods (MSV'10). Las Vegas, NV, USA: IEEE; 2010. p. 94–7. https://doi.org/10.1109/WMSVM.2010.52.
70. Ribeiro LG, Suominen OJ, Durmush A, Peltonen S, Ruiz Morales E, Gotchev A. Retro-reflective-marker-aided target pose estimation in a safety-critical environment. Appl Sci. 2020;11(1):3. https://doi.org/10.3390/app11010003.
71. Shao L, Yang S, Fu T, Lin Y, Geng H, Ai D, et al. Augmented reality calibration using feature triangulation iteration-based registration for surgical navigation. Comput Biol Med. 2022;148:105826. https://doi.org/10.1007/s00464-009-0568-2.
72. Hong J, Hashizume M. An effective point-based registration tool for surgical navigation. Surg Endosc. 2010;24:944–8. https://doi.org/10.1007/s00464-009-0682-8.
73. Ehambram A, Hemme P, Wagner B. An approach to marker detection in ir-and rgb-images for an augmented reality marker. In: Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO), vol. 2. 2019. p. 190–7. https://doi.org/10.5220/0007900701900197.
74. Cartiaux O, Docquier P-L, Paul L, Francq BG, Cornu OH, Delloye C, et al. Surgical inaccuracy of tumor resection and reconstruction within the pelvis: an experimental study. Acta Orthop. 2008;79(5):695–702. https://doi.org/10.1080/17453670810016731.
75. Orozco-Martínez J, Fernández-Fernández T, Aguilera-Jiménez E, Iribar-Zabala A, Gregorio-Bermejo C, Pose-Díez-de-la-Lastra A, et al. Validation of augmented reality-guided implant positioning in cadaveric models: a step toward clinical application. J Orthop Res Ther. 2025;10:1385. https://doi.org/10.29011/2575-8241.001385
76. Mai H-N, Dam VV, Lee D-H. Accuracy of augmented reality-assisted navigation in dental implant surgery: systematic review and meta-analysis. J Med Internet Res. 2023;25:42040. https://doi.org/10.2196/42040.
77. Groot NT, Zee JM, Bökkering GMJ, Littooij AS, Hulsker CC, Scheltinga CEJ, et al. Introducing holographic surgical navigation in pediatric wilms' tumor patients: a feasibility study during total nephrectomy. Bioengineering. 2025;12(8):896. https://doi.org/10.3390/bioengineering12080896.
78. Kuo C-H, Huang F-C, YF-C. Development of active ir-based surgical marker tracking and positioning systems. In: IEEE International Conference on Systems, Man and Cybernetics (SMC 2005), vol. 3. Waikoloa, Hawaii, USA: IEEE; 2005. p. 2443–8. https://doi.org/10.1109/ICSMC.2005.1571515.
79. Urtans E, Nikitenko A. Active infrared markers for augmented and virtual reality. Markers. 2016;9:10.
80. Jurado-Rodriguez D, Muñoz-Salinas R, Garrido-Jurado S, Medina-Carnicer R. Planar fiducial markers: a comparative study. Virtual Real. 2023;27(3):1733–49. https://doi.org/10.1007/s10055-023-00772-5.
81. Marino E, Bruno F, Barbieri L, Lagudi A. Benchmarking built-in tracking systems for indoor ar applications on popular mobile devices. Sensors. 2022;22(14):5382. https://doi.org/10.3390/s22145382.
82. Muñoz-Salinas R, Marín-Jimenez MJ, Yeguas-Bolivar E, Medina-Carnicer R. Mapping and localization from planar markers. Pattern Recogn. 2018;73:158–71. https://doi.org/10.1016/j.patcog.2017.08.010.
83. Stenmark M, Omerbašić E, Magnusson M, Andersson V, Abrahamsson M, Tran P-K. Vision-based tracking of surgical motion during live open-heart surgery. J Surg Res. 2022;271:106–16. https://doi.org/10.1016/j.jss.2021.10.025.
84. Thabit A, Niessen WJ, Wolvius EB, Van Walsum T. Evaluation of marker tracking using mono and stereo vision in microsoft hololens for surgical navigation. In: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 12034, pp. 253–262. SPIE, San Diego, CA, USA 2022. https://doi.org/10.1117/12.2607262.
85. Lin Q, Cai K, Yang R, Chen H, Wang Z, Zhou J. Development and validation of a near-infrared optical system for tracking surgical instruments. J Med Syst. 2016;40:1–14. https://doi.org/10.1007/s10916-016-0462-0.

86. Salerno F, Ulrich L, Maculotti G, Moos S, Genta G, Vezzetti E, et al. A metrological approach for augmented reality tooltip tracking assessment. Comput Ind. 2026;175:104430. https://doi.org/10.1016/j.compind.2025.104430.
87. Maurer CR, Fitzpatrick JM, Wang MY, Galloway RL, Maciunas RJ, Allen GS. Registration of head volume images using implantable fiducial markers. IEEE Trans Med Imaging. 1997;16(4):447–62. https://doi.org/10.1109/42.611354.
88. West JB, Khadem R, Maurer CRJ. Nonlinear averaging method to optimize the accuracy of optical tracking. In: Fleming MV, Holmes DR (eds) Medical Imaging 2003: Visualization, Image-Guided Procedures, and Display, vol. 5029, pp. 311–318. SPIE, San Diego, CA, USA 2003. https://doi.org/10.1117/12.479697.
89. West JB, Maurer CR. Designing optically tracked instruments for image-guided surgery. IEEE Trans Med Imaging. 2004;23(5):533–45. https://doi.org/10.1109/TMI.2004.825614.