

A MIXED REALITY FRAMEWORK FOR EVALUATING AIRCRAFT PRE-FLIGHT WALK-AROUND TRAINING

Miguel A. Higinio[1], Michael J. Scott[1] & Wim J.C. Verhagen[1]

[1]Department of Aerospace Engineering, RMIT University, Melbourne, Australia

Abstract

Mixed Reality (MR) technologies are increasingly being explored to support aircraft inspection and maintenance-related training by embedding contextual task information, spatial guidance, and interactive feedback directly within the user's field of view. However, the evaluation of MR training systems remains fragmented, particularly in terms of how performance, workload, usability, and learning-related outcomes are selected and measured. This paper presents an MR training framework developed for aircraft pre-flight walk-around inspection, chosen as a procedure that sits at the intersection of maintenance awareness and operational readiness. The system was built in Unity using the Mixed Reality Toolkit 3 (MRTK3) and deployed on Microsoft HoloLens 2, with the Cessna 172S Skyhawk and Pilatus PC-9/A serving as aircraft testbeds. The framework delivers step-by-step spatial guidance, instructional panels, video support, manual aircraft alignment, and local logging of task execution data. Rather than selecting performance measures arbitrarily, a structured approach is proposed grounded in four explicit criteria: relevance, measurability, reliability, and comparability. For the current stage, the measures selected on that basis are task completion time, perceived workload using the NASA Task Load Index (NASA-TLX), usability through the System Usability Scale (SUS), and open-ended participant feedback. Initial indoor and outdoor trials were conducted to validate the MR workflow, alignment process, data logging pipeline, and post-task evaluation instruments. These trials are used to assess system feasibility and prepare for a delayed crossover comparison between MR-assisted and paper-based training. In the Pilatus PC-9/A aircraft, MR-guided sessions showed a lower mean task completion time than paper-based sessions (11:04 min vs. 17:18 min), lower perceived workload as measured by NASA-TLX (4.38 vs. 6.83), and higher usability as measured by SUS (83.13 vs. 47.50). Qualitative feedback further identified spatial alignment stability, outdoor visibility, and wayfinding as key design considerations for the next experimental stage. The broader evaluation of learning gain and knowledge retention will be addressed in the next experimental stage through pre/post knowledge assessments and delayed follow-up testing.

Keywords: aircraft inspection, mixed reality, performance measures, pre-flight walk-around, training

1. Introduction

Mixed Reality (MR) has gained traction in aircraft maintenance and training over the past decade, due to its ability to combine physical assets with digital holograms and information within the user's field of vision in real time. This ability opens a range of opportunities for maintenance personnel, pilots, and trainees to accomplish their tasks faster or more accurately, while enabling remote support and compliance checking. The foundational framework for understanding where MR fits within the reality-virtuality continuum proposed by Milgram and Kishino places MR in the region between fully real and fully virtual environments, where real and virtual objects coexist in a single display environment [1]. This taxonomy is represented in Figure 1. For aircraft inspection and training, that positioning is not merely theoretical: users may need to cross-reference physical components while simultaneously interpreting procedural instructions, inspection data, manuals, or task-specific information, all in real time.

https://doi.org/10.71945/icas2026_0464

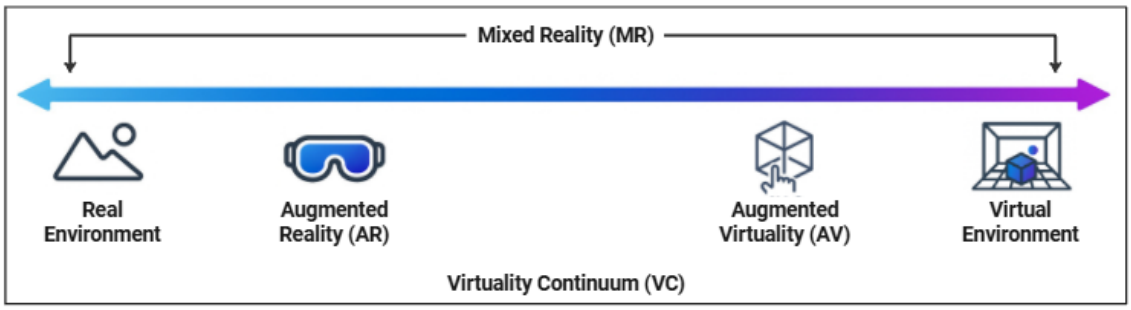


Figure 1 – Mixed Reality taxonomy based on Milgram & Kishino, 1994.

Aircraft inspection and maintenance-related procedures are commonly characterized by manual interaction with physical assets, high component variability, and the need to follow documented procedures for airworthiness, regulatory compliance, and safety purposes [2]. In routine aircraft operations, this also includes pre-flight walk-around checks, where users must inspect the physical aircraft while following a defined procedure. These characteristics create a need for training technologies that can present procedural information in context without removing the user's attention from the aircraft. Eschen et al. argue that the selection of a suitable immersive technology depends on the level of interaction between the worker, the physical object, and the digital model [3]. Their analysis draws a useful distinction: Virtual Reality (VR) may be better suited for virtual inspection tasks based primarily on digital models, while MR is more appropriate for process guidance and ground-test guidance, where the operator must continuously interact with the aircraft or physical equipment. This distinction guided the design of the present study, which uses MR not only to represent the aircraft environment, but to support users while performing routine aircraft inspection and pre-flight walk-around procedures on the physical aircraft.

This paper centers on MR-supported training for aircraft pre-flight walk-around inspection, treated as a procedure that sits at the intersection of maintenance awareness and operational readiness, and therefore representative of the kind of task where structured, repeatable training matters most. The Cessna 172S Skyhawk and Pilatus PC-9/A serve as aircraft testbeds, with the pre-flight inspection procedure forming the training scenario under evaluation. It is worth being explicit about what this work does and does not attempt: rather than claiming a fully longitudinal assessment of learning effectiveness, something that would require a study design well beyond this initial phase, the focus here is on presenting the MR training framework itself, establishing a method for selecting performance measures, and reporting first-round validation results drawn from task completion time, perceived workload, usability ratings, and participant feedback. Learning gain and knowledge retention are not ignored; they are deliberately deferred to a subsequent experimental stage, where pre/post knowledge assessments and delayed follow-up testing will allow those dimensions to be examined with the rigor they demand.

2. Related Work

2.1 Mixed Reality in Aircraft Inspection and Procedural Training

MR technologies have been explored across a wide range of aircraft lifecycle activities, such as design, assembly, inspection, maintenance support, training, and operational assistance. Tran et al. frame these developments within the broader ambition of smart aviation, where immersive interfaces are expected to support personnel operations, maintenance activities, product design, and training, with Artificial Intelligence (AI) and machine learning anticipated as future layers for decision-support [4]. From the perspective of this paper, the most relevant value of MR is its ability to place task information directly in the physical work environment, reducing dependence on paper manuals, static displays, or instructor demonstrations during inspection and procedural training tasks.

Aircraft inspection has become one of the more mature areas of MR application. MR-assisted workflows have been applied to dent and deformation inspection in aircraft structures, demonstrating re-

ductions in inspection time under controlled evaluation conditions [5]. Similar work using HoloLens 2 has proposed spatially aligned workflows to help users search for, localize, and document aircraft damage [6]. What makes these contributions useful as reference points is not simply the performance gains they report, but rather what they reveal about the nature of the task itself: physical inspection requires the simultaneous interpretation of defect location, aircraft surface geometry, and inspection instructions. This creates a spatial and cognitive demand that MR is particularly well positioned to support. In this context, Gulrez et al. developed a physical load metric for assessing MR in aircraft inspection tasks, combining MR-supported inspection with workload-oriented evaluation [7]. Their work is especially relevant to the present study because it shows that MR inspection systems should not be evaluated only through task completion time, but also through how they affect the user's physical and cognitive response during the task.

Beyond structural inspection, technologies within the MR continuum have also been explored for procedural assistance and training. Satish and Kumar developed a tool-selection application for aircraft maintenance, demonstrating how virtual annotations and step-by-step animations can guide users through maintenance-related tasks without relying solely on paper instructions [8]. Lallai et al. proposed a task-based MR guidance environment for aircraft cockpit procedures, combining task models with holographic cues to help students interact with the correct cockpit elements in the correct sequence while allowing instructors to monitor progress in real time [9].

VR training studies provide additional evidence for the value of immersive procedural training, although the comparison with MR is not straightforward. In VR, the user operates inside a simulated environment; in MR-based walk-around inspection, the user works with the physical aircraft. That distinction matters for the present study because the training task depends on continuous interaction with real aircraft components. Dela Peña reported improvements in engagement, knowledge retention, task execution accuracy, and perceived competency in a VR aircraft maintenance training context [10]. Nguyen et al. developed a high-fidelity VR simulation for Airbus A320 nose landing gear wheel removal, built from full-scale 3D models and structured around aircraft maintenance documentation [11]. Both studies point to immersive environments as viable platforms for procedural learning, while also reinforcing why MR, rather than VR, is the more appropriate modality when the training task requires interaction with the physical aircraft.

Remote support studies offer a different angle. Vorraber et al. compared HoloLens-assisted and mobile-assisted remote maintenance in an industrial setting and found that visual feedback improved situation descriptions during remote tasks, although completion time varied by support modality [12]. Jing et al. showed that gaze- and hand-based MR collaboration interfaces can reduce workload and improve co-presence and mutual understanding during physical tasks [13]. These studies are not about walk-around inspection training specifically, but they are relevant: MR shapes how communication, workload, and shared situational awareness play out in technical tasks, and those effects will matter as the present framework evolves toward instructor-monitored or remotely supported inspection training.

At a broader systems level, recent work points to where MR maintenance platforms are heading. Liu et al. developed an MR-continuum maintenance platform for wind turbines that integrates real-time operational data, remote expert support, and interactive maintenance instructions into a single connected workflow [14]. Fuchs et al. applied model-based systems engineering, 3D scanning, and virtual reconfiguration to aircraft cabin variant evaluation, demonstrating how digital infrastructure can support traceability and rapid configuration assessment [15]. These examples reframe MR not as a standalone visual interface but as a node in a larger digital ecosystem—one that connects task data, system models, inspection outcomes, and decision-support tools. That is the long-term direction of the framework proposed here, even though the current paper focuses specifically on MR-supported pre-flight walk-around inspection training and its initial evaluation.

2.2 Performance Measures for Evaluating MR Training Impact

Evaluating MR training remains methodologically challenging because the field has not converged on a consistent evaluation methodology. Daling and Schlittmeier documented this directly: immersive training studies differ substantially in hardware, task type, training design, and evaluation approach, which makes comparison across studies difficult [16]. Their review points to objective measures, such as task completion time and accuracy, and subjective measures, such as usability and perceived workload, as a minimum basis for comparability. However, these measures capture different aspects of training performance and should not be interpreted in isolation. In particular, task completion during a guided session does not necessarily indicate that the user has learned the procedure or retained the knowledge required to perform it later without assistance.

In aircraft maintenance specifically, Loizeau et al. identified a more fundamental constraint: field evaluation of MR tools is limited by the absence of structured methods for integration and assessment in real industrial settings [17]. Their framework organizes evaluation around effectiveness, efficiency, and satisfaction, a structure that maps naturally onto measures such as task completion time, error-related indicators, SUS, and NASA-TLX. For the present study, this structure supports the use of multiple measures rather than relying on task time alone: a participant who completes a walk-around procedure faster has not necessarily performed it more correctly or learned it more effectively. An incorrect inspection sequence, a missed cue, unsustainable cognitive load, or poor retention after a delay would not appear in a task-time log. There is also a subtler risk: when participants know that task duration is being recorded, speed can become the implicit performance criterion, encouraging faster completion at the expense of inspection quality. Task completion time is useful, but only when interpreted alongside workload, usability, procedural correctness, and learning outcomes.

The case for including learning-related measures comes from studies that treat immersive technologies as training systems rather than only as assistance tools. Pre- and post-test designs have been used to compare MR-based instruction with traditional teaching materials in aircraft-related training contexts [18], and VR-based procedural training has been evaluated against traditional instruction and learning-by-teaching approaches to examine how delivery format affects knowledge gain [19]. These studies show that task performance during training, immediate learning gain, and delayed knowledge retention are distinct evaluation outcomes. A system may support efficient task completion during a guided session while producing limited retention once the guidance is removed. This distinction is particularly important for pre-flight walk-around training, where the long-term objective is not only to complete the guided task, but to develop procedural understanding that can be transferred to future aircraft inspection procedures.

This paper therefore treats MR training effectiveness as a multidimensional construct. In the current stage, the evaluation focuses on task completion time, perceived workload, usability, and open-ended participant feedback. Learning gain, retention, sequence accuracy, and inspection accuracy remain important outcomes, but they are planned for the next experimental stage once the assessment instruments and scoring protocols specific to the walk-around inspection task have been finalized.

2.3 Research Gap and Contribution

The literature shows that MR can support aircraft inspection, maintenance assistance, procedural training, and remote collaboration. However, three gaps remain directly relevant to the present study.

First, most existing work addresses isolated maintenance or inspection tasks, while fewer studies treat pre-flight walk-around inspection as the primary focus, despite its character as a repeated, safety-relevant general visual inspection (GVI) task with a defined structure. The walk-around occupies a specific position in aircraft operations: it requires the user to identify visible aircraft conditions, follow a prescribed sequence, and form a judgment about aircraft state before operation. In operational contexts, that check also relates to aircraft availability and turnaround time (TAT), where inspection quality and procedural consistency are as important as speed [20, 21].

Second, evaluation approaches across the field remain inconsistent. Differences in hardware, task type, study design, and reported measures make it difficult to compare findings across studies or determine which measures are most appropriate for a specific aircraft procedure.

Third, while a number of systems have demonstrated technical feasibility, fewer studies explicitly connect MR workflow design, performance measure selection, and initial validation within a single evaluation framework. As a result, the relationship between system design decisions and evaluation choices often remains implicit.

This paper addresses those gaps through three contributions. First, it presents an MR training framework for aircraft pre-flight walk-around inspection using HoloLens 2, with the Cessna 172S Skyhawk and Pilatus PC-9/A as aircraft testbeds. Second, it proposes a structured performance measure selection approach organized around four criteria: relevance, measurability, reliability, and comparability. This makes the selection process transparent and repeatable. Third, it reports initial indoor and outdoor validation trials used to assess the MR workflow, alignment process, data logging pipeline, and post-task evaluation instruments. Together, these contributions establish the foundation for the next experimental stage, in which MR-assisted and paper-based training will be compared using a delayed crossover design.

3. Materials and Methods

This section describes the methodological basis of the MR pre-flight walk-around training study. The section is organized to first explain the technical implementation of the MR framework and the operational training task, before introducing the performance measure selection process and the planned experimental design. This order is deliberate: the relevance and interpretation of each performance measure depend on the system workflow, the aircraft testbeds, the inspection task structure, and the type of data that the MR system can reliably collect. The section therefore begins with the system architecture, aircraft testbeds, user interaction workflow, and data logging process, before describing the selected measures and the delayed crossover design.

3.1 Mixed Reality Training Framework

3.1.1 System Architecture

The MR training system is developed in Unity and deployed on Microsoft HoloLens 2. The implementation uses the MRTK3 for hand tracking, gaze-based interaction, and spatial user interfaces, together with the Mixed Reality OpenXR Plugin for HoloLens 2 deployment. The current system does not rely on image markers or QR-code tracking and uses manual alignment supported by spatial anchoring, which trades some automation for the flexibility needed to operate across varied indoor and outdoor training conditions.

The architecture is organized into four layers: device, edge, cloud, and application, as shown in Figure 2.

1. **Device layer:** This layer includes the physical aircraft testbeds, the Cessna 172S Skyhawk and the Pilatus PC-9/A, together with the HoloLens 2 headset. These define the real-world environment in which the walk-around inspection task takes place.

2. **Edge layer:** This layer contains the current development and runtime stack. Unity handles the MR interface, interaction logic, manual alignment workflow, step-by-step guidance, and data logging. Blender is used for preparing and adjusting the 3D aircraft models, and GitHub manages version control. All primary computation and interaction processing run locally on the HoloLens 2 during trials, which allows the system to operate without a continuous network connection.

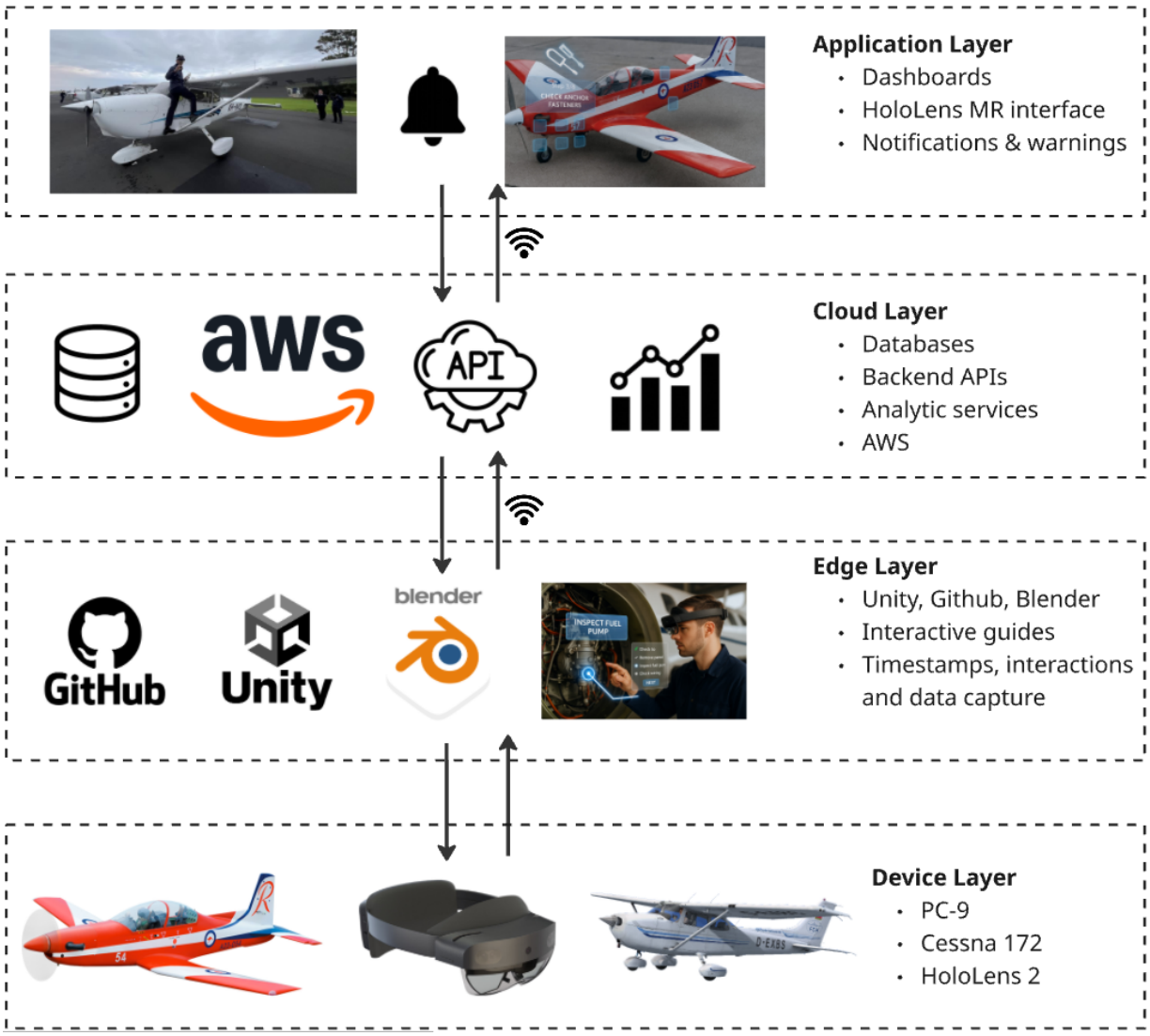


Figure 2 – Layered architecture of the proposed MR training and maintenance framework.

3. **Cloud layer:** This layer is planned rather than fully implemented at this stage. Future iterations will incorporate backend APIs, databases, cloud storage, and analytics services for storing trial logs, maintenance records, inspection outcomes, and task performance data. This extension is intended to support a broader maintenance, repair, and overhaul (MRO) workflow in which inspection results feed into maintenance task generation and decision-support tools.

4. **Application layer:** This layer covers the user-facing components, including the aircraft selection panel, task selection panel, manual alignment controls, operating mode selection, instructional panels, and step navigation controls. These components define the interaction workflow used by participants during the MR-guided walk-around inspection.

The current paper focuses on the HoloLens-based training workflow and its initial validation for pre-flight walk-around inspection.

3.1.2 Aircraft Testbeds and Training Scenario

The framework was developed using two aircraft testbeds: the Cessna 172S Skyhawk and the Pilatus PC-9/A. These aircraft were selected because they were accessible through RMIT University, including support from the RMIT Aviation Academy, and provided physical aircraft platforms suitable for developing and validating the MR pre-flight walk-around workflow across different training and campus environments. The aircraft are treated as testbeds, while the training scenario is the pre-flight walk-around inspection itself.

The pre-flight walk-around requires the user to move around the physical aircraft, inspect visible components, follow a prescribed sequence, and confirm that key aircraft areas are in an acceptable condition before operation. Although the procedure is commonly associated with pilot preparation, it also reflects maintenance-related inspection awareness because it involves identifying visible defects, leaks, missing covers, surface damage, unsecured panels, and other observable conditions that may require further attention.

The path followed by the user during the Cessna 172S Skyhawk walk-around is shown in Figure 3. The numbered locations indicate the main inspection zones used to structure the MR guidance sequence:

1. **Cabin:** initial cockpit and cabin checks, including parking brake, controls, switches, annunciators, fuel quantity indications, and general cabin safety items.
2. **Empennage:** inspection of the baggage door, tail tiedown, rudder gust lock, control surfaces, elevator trim tab, and antenna condition.
3. **Right wing trailing edge:** inspection of the flap and aileron for security, condition, and freedom of movement.
4. **Right wing:** inspection of the leading edge, wing tiedown, main wheel tyre, and wing strut for damage, security, and general condition.
5. **Nose:** inspection of the engine cooling air inlets, alternator belt, propeller and spinner, exhaust system, air filter, nosewheel strut and tyre, and static source opening.
6. **Left wing leading edge:** inspection of the fuel tank vent opening, stall warning opening, and leading edge for blockage, operation, and visible damage.
7. **Left wing:** inspection of the wing tiedown, main wheel tyre, and wing strut for security, inflation, wear, damage, and general condition.
8. **Left wing trailing edge:** inspection of the aileron and flap for freedom of movement, security, and condition.

Consumable checks, including fuel quantity, fuel caps, sump drain valves, and engine oil level, are embedded within the relevant wing and nose inspection zones rather than shown as a separate numbered location in Figure 3.

These numbered zones were used to organize the MR instructional steps and to ensure that participants followed a consistent inspection route during the study. This structure also supports the collection of task completion time and step progression data, since each MR instruction can be associated with a specific aircraft zone. The purpose of the MR system is not to replace formal aircraft documentation or instructor supervision, but to provide an interactive training environment in which users can practice the inspection sequence while remaining physically oriented to the aircraft.

3.1.3 User Interaction and Guidance Workflow

The user interaction workflow was designed to guide participants from initial briefing to task completion with minimal setup complexity. The workflow is structured into six stages.

1. **Participant briefing:** The participant receives a short introduction to the study, the HoloLens 2 headset, and the MR interaction controls.
2. **Aircraft selection:** The user selects the aircraft testbed, either the Cessna 172S Skyhawk or the Pilatus PC-9/A, which loads the corresponding 3D model, inspection task data, and spatial guidance elements.

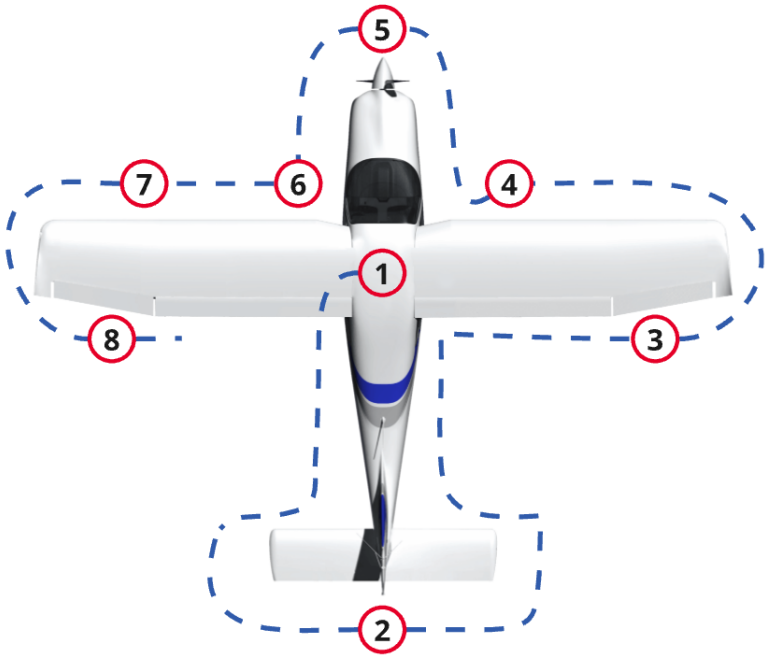


Figure 3 – Cessna 172S Skyhawk pre-flight walk-around path.

3. **Task selection:** The user selects the pre-flight walk-around inspection task, which defines the step sequence, inspection zones, and associated visual content for the guided procedure.
4. **Manual aircraft alignment:** Before the guided inspection begins, the user aligns the virtual aircraft model with the physical aircraft using manual controls for position, rotation, and scale. Once the alignment is accepted, the instructional content is spatially anchored relative to the aligned model for the remainder of the session.
5. **MR-guided walk-around inspection:** The participant follows the MR-guided inspection sequence, with each step presenting an instruction panel, spatial reference, and supporting visual material where relevant. Forward and backward navigation is available throughout, allowing participants to move along the inspection route while staying oriented to the physical aircraft.
6. **Data logging and post-task questionnaires:** During the task, the system records step progression and timing data. After the task, participants complete post-task questionnaires, including NASA-TLX and SUS.

Keeping the workflow consistent across participants was a deliberate design decision. It standardizes the user experience and makes step-level logging meaningful. The system records entry and exit events for each guided step, which allows later analysis of where participants slow down, backtrack, or spend more time during the inspection.

Figure 4 shows examples of the MR guidance interface used during the pre-flight walk-around tasks. The interface presents the selected aircraft task, step-by-step inspection instructions, spatially positioned guidance elements, and supporting visual content. These interface elements were designed to keep the participant oriented to both the physical aircraft and the current inspection step.

3.1.4 Data Logging

The MR system logs task progression locally on the HoloLens 2 during each guided walk-around session. A session identifier is generated at the start of each task, and the timer begins when the guided inspection sequence is initiated. The system records step entry and exit events, forward and backward navigation actions, time spent on each step, and total task completion time.

The logging structure serves two purposes. During the current stage, it provides a means of validating that the task sequence, navigation controls, alignment workflow, and timing pipeline behave



Figure 4 – Examples of the MR guidance interface for the Pilatus PC-9/A and Cessna 172S Skyhawk pre-flight walk-around tasks.

reliably across indoor and outdoor conditions. This validation is a necessary step before treating collected data as experimental. In the formal stage, the same structure will support comparison of task completion time and step-level progression between MR-assisted and paper-based training conditions.

Beyond timing, the logs also enable behavioral analysis of the inspection process. Step-level data can reveal where participants slow down, backtrack, or spend more time during the walk-around, patterns that are not visible from total task completion time alone. This granularity will support future work on sequence accuracy, error identification, and learning-related outcomes once task-specific scoring criteria have been defined for each aircraft testbed.

3.2 Performance Measure Selection

Performance measure selection formed a methodological step in the study design, rather than a secondary reporting decision. MR training effectiveness cannot be captured through a single metric, and informal measure selection is one of the primary sources of inconsistency across immersive training studies. Prior work commonly combines objective indicators such as task completion time and accuracy with subjective ones such as perceived workload and usability [16]. In aircraft maintenance field evaluation, Loizeau et al. argue that immersive tools should be assessed through effectiveness, efficiency, and satisfaction dimensions [17]. Aircraft inspection studies reinforce the need to evaluate not only task outcomes, but also the user's physical and cognitive response to the technology [7, 22].

3.2.1 Initial Pool of Measures

An initial pool of candidate measures was compiled from prior MR training studies, aircraft maintenance research, and human factors evaluation methods. The pool included task completion time, error rate, sequence accuracy, inspection accuracy, learning gain, knowledge retention, perceived workload, usability, interaction performance, physical workload, system performance, authoring effort, perceived usefulness, and safety perception.

The purpose of building a broad pool was not to deploy every measure in this stage, as that would increase participant burden and require scoring protocols that are not yet validated for the pre-flight walk-around inspection task. Instead, the pool provided a structured basis for selecting measures that were appropriate to the current system maturity, task context, and validation objectives.

3.2.2 Selection Criteria

Each candidate measure was evaluated against four criteria: relevance, measurability, reliability, and comparability. These criteria were introduced to ensure that measure selection was driven by fitness for purpose, not convenience. Their definitions are provided in Table 1.
Measures that aligned strongly with the study objectives, could be collected consistently at the current validation stage, and supported comparison between training conditions were prioritized for the core evaluation set. Measures that were conceptually important but required additional scoring protocols,

Table 1 – Criteria used for selecting performance measures.

Criterion	Definition
Relevance	Assesses whether the measure directly reflects the learning or performance outcome targeted in the aircraft maintenance task.
Measurability	Assesses whether the measure can be captured using the current MR system, structured observation, system logs, or validated questionnaires.
Reliability	Assesses whether the measure is sufficiently stable, repeatable, or supported by prior literature and standardized instruments.
Comparability	Assesses whether the measure supports meaningful comparison between MR-assisted and traditional training conditions.

instructor validation, or delayed follow-up assessment were retained for the next experimental stage, deferred rather than discarded.

3.2.3 Selected Measures for the Study

For the initial validation stage, the selected measures are task completion time, perceived workload, usability, and open-ended participant feedback. Task completion time is recorded automatically through the MR system logs. Perceived workload is assessed using the NASA-TLX and usability through the SUS, both administered after the task via Qualtrics. Open-ended feedback captures qualitative observations about user experience, interaction friction, and perceived training value that structured scales tend not to surface.

These four measures were selected because they can be collected consistently within the current system without requiring task-specific scoring criteria that have not yet been validated. They also match the immediate purpose of the study: confirming that the MR workflow functions as intended, that the guidance sequence is navigable, and that the data capture pipeline is reliable across indoor and outdoor conditions.

Learning gain, knowledge retention, sequence accuracy, error rate, and inspection accuracy remain important outcomes; they are planned for the formal experimental stage, not treated as secondary. Learning gain will require pre- and post-task knowledge assessments, while knowledge retention will require delayed follow-up testing. Sequence accuracy, error rate, and inspection accuracy will require task-specific scoring criteria developed with input from qualified aviation or maintenance instructors, so that what counts as correct reflects operational standards rather than procedural completion alone.

3.3 Experimental Design

The formal experimental stage is planned as a delayed crossover study comparing MR-guided training with traditional paper-based instruction. Each participant completes the same pre-flight walk-around inspection task under both conditions, a within-participant structure that reduces the influence of individual differences in prior experience, confidence, and task familiarity on the comparison between training methods.

Participants will be assigned to one of two condition orders. Group 1 completes the MR-guided condition first and the paper-based condition approximately one week later; Group 2 follows the reverse order. The counterbalancing is intended to distribute order effects across the sample rather than allow them to accumulate in one direction. The crossover structure is shown in Figure 5.

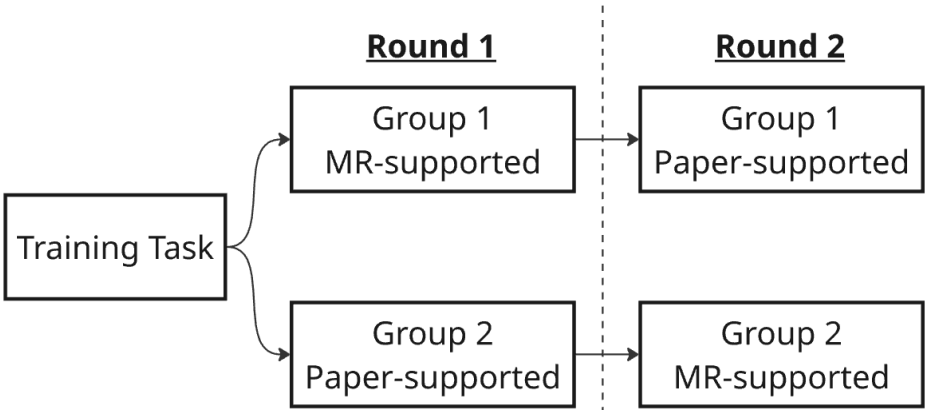


Figure 5 – Planned crossover design comparing paper-based and MR-guided pre-flight walk-around training conditions.

The one-week interval between sessions provides a practical delay between conditions. In practice, a scheduling window of five to ten days will be used to accommodate participant availability, aircraft access, and weather dependencies associated with outdoor testing. The exact number of days between sessions will be recorded for each participant and included in the dataset, as this interval may affect the interpretation of learning and retention outcomes.

Each session follows the same general structure: a short briefing, the inspection task under the assigned training method, and post-task questionnaires. For the current stage, the primary post-task measures are perceived workload, usability, and open-ended feedback. In the MR condition, task progression and completion time are captured automatically through the HoloLens 2 logging system. In the paper-based condition, equivalent timing and observational data will be collected manually or through external observation procedures. This difference in data collection method will be considered when interpreting task completion time.

The crossover design also provides a natural structure for incorporating learning-related measures in the next stage. Learning gain will require pre- and post-task knowledge assessments, while knowledge retention will require a delayed follow-up. Sequence accuracy, error rate, and inspection accuracy will be added once scoring criteria have been developed with input from qualified aviation or maintenance instructors, so that what is being measured reflects operational standards, not just procedural completion.

4. Initial Results

The initial validation stage was used to verify the MR workflow, data logging pipeline, questionnaire delivery, and participant interaction process before conducting the full delayed crossover experiment. The results reported in this section are descriptive only. Inferential statistical testing was not conducted because the sample is small and unbalanced, with only three participants having completed both the paper-based and MR-guided conditions at this stage.

Eight MR-guided task logs were collected across the two aircraft testbeds: four for the Cessna 172S Skyhawk and four for the Pilatus PC-9/A. Of the Cessna 172S Skyhawk MR sessions, three were conducted indoors and one outdoors, while the Pilatus PC-9/A MR sessions were conducted outdoors. Questionnaire data were available for seven MR-guided sessions and three paper-based sessions. The paper-based sessions were conducted only for the Pilatus PC-9/A task, which means that direct method comparison is currently limited to the three participants who completed both Pilatus PC-9/A conditions. Table 2 summarizes the available initial validation sessions by method and aircraft testbed.

Table 2 – Overview of initial validation sessions by method and aircraft testbed.

Method	Aircraft testbed	Environment	n
MR-guided	Cessna 172S Skyhawk	Indoor	3
MR-guided	Cessna 172S Skyhawk	Outdoor	1
MR-guided	Pilatus PC-9/A	Outdoor	4
Paper-based	Pilatus PC-9/A	Outdoor	3

4.1 Task Completion Time

Task completion time was analysed descriptively by aircraft testbed and method. Because the Cessna 172S Skyhawk and Pilatus PC-9/A procedures differ in task length, step count, and inspection content, times were not aggregated across aircraft. The Cessna 172S Skyhawk MR sessions served primarily to verify logging and deployment feasibility for that aircraft, while method comparison was restricted to the Pilatus PC-9/A subset, where both training conditions were available.

The MR system successfully recorded task completion time for all MR-guided sessions. Descriptive completion times by aircraft testbed and method are summarized in Table 3. For the Pilatus PC-9/A subset, where both methods were available, paper-based sessions had a longer mean task time than MR-guided sessions.

Table 3 – Descriptive task completion times by aircraft testbed and method. Times are reported in min:s and rounded to the nearest second. SD denotes standard deviation.

Condition	n	Mean time	SD	Range
MR-guided, Cessna 172S Skyhawk	4	18:41	2:12	15:29–20:09
MR-guided, Pilatus PC-9/A	4	11:04	1:41	9:47–13:24
Paper-based, Pilatus PC-9/A	3	17:18	1:49	15:22–19:00

The SD values indicate the variability in completion time within each condition. The larger SD for the Cessna 172S Skyhawk MR sessions reflects greater variation across the indoor and outdoor validation runs, while the Pilatus PC-9/A MR and paper-based sessions showed broadly comparable variability.

For the Pilatus PC-9/A task, three participants completed both the paper-based and MR-guided conditions. In this paired subset, MR-guided task times were consistently shorter than paper-based times. The mean task time decreased from 17:18 min in the paper-based condition to 11:27 min in the MR-guided condition, corresponding to an average reduction of 5:51 min, or approximately 33.8% relative to the paper-based mean task time. This pattern is reported as an observation rather than an inferential finding, given the small paired subset and the difference between manual timing in the paper-based condition and automatic logging in the MR-guided condition.

4.2 Questionnaire Results

Post-task questionnaires were used to capture perceived workload and usability after each session. Workload was assessed using the NASA-TLX, and usability was assessed using the SUS. NASA-TLX scores were calculated using the six workload dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Scores were processed so that higher values consistently represented higher perceived workload before calculating the overall average, as shown in Eq. 1.

$$NASA\text{-}TLX = \frac{\text{Mental} + \text{Physical} + \text{Temporal} + \text{Performance} + \text{Effort} + \text{Frustration}}{6} \tag{1}$$

Questionnaire results were analysed separately by aircraft testbed. This was necessary because the Cessna 172S Skyhawk and Pilatus PC-9/A datasets served different purposes in the current study stage. The Cessna 172S Skyhawk sessions were used to evaluate MR deployment and interaction under indoor and outdoor conditions, whereas direct method comparison was only available for the Pilatus PC-9/A subset.

For the Cessna 172S Skyhawk, questionnaire data were available for three MR-guided sessions, including two indoor sessions and one outdoor session. These results are summarized in Table 4. The indoor sessions produced a lower mean NASA-TLX score and a higher mean SUS score than the outdoor session. This suggests that, under more controlled conditions, participants were able to interact with the MR guidance with less perceived workload and higher perceived usability. The outdoor session produced the highest workload score and the lowest usability score within the Cessna 172S Skyhawk subset. This result was associated with an earlier alignment stability issue, where the virtual aircraft model could shift after manual positioning, rather than with the inspection task itself.

Table 4 – Descriptive questionnaire results for the Cessna 172S Skyhawk MR-guided sessions.

Environment	n	NASA-TLX	SUS
Indoor	2	4.42	85.00
Outdoor	1	15.00	40.00

For the Pilatus PC-9/A, questionnaire data were available for four MR-guided sessions and three paper-based sessions. These results are summarized in Table 5. This subset provides the most relevant early method comparison because both MR-guided and paper-based sessions were available for the same aircraft testbed. Within this subset, the MR-guided condition produced a lower mean NASA-TLX score than the paper-based condition, 4.38 compared with 6.83, suggesting lower perceived workload during MR-guided task execution. The MR-guided condition also produced a higher mean SUS score, 83.13 compared with 47.50 for the paper-based condition, indicating stronger perceived usability of the MR guidance workflow.

Table 5 – Descriptive questionnaire results for the Pilatus PC-9/A sessions by method.

Method	n	NASA-TLX	SUS
MR-guided	4	4.38	83.13
Paper-based	3	6.83	47.50

The questionnaire results suggest that the MR-guided workflow was generally perceived as usable and manageable, particularly in the Pilatus PC-9/A subset. The results also point to the importance of stable spatial alignment. After the alignment workflow was updated so that the accepted manual position remained fixed through spatial anchoring, the Pilatus PC-9/A outdoor sessions showed more consistent usability and workload outcomes. These values are treated as preliminary indicators rather than inferential evidence, as the sample remains too small and unbalanced to support conclusions about workload reduction or usability improvement. These effects will be evaluated in the formal experimental stage.

4.3 Qualitative Feedback

Open-ended responses were reviewed to identify usability and deployment issues not fully captured by the quantitative measures. Although the number of responses was limited, the comments provided useful context for interpreting the task time, NASA-TLX, and SUS results.

Three issues surfaced across sessions. Spatial alignment stability emerged as the most consequential. The lower usability and higher workload scores observed in the outdoor Cessna 172S Skyhawk session were associated with an earlier version of the alignment workflow, in which the virtual aircraft

model could drift after manual positioning. That issue directly informed the subsequent update in which accepted alignments are locked using spatial anchoring for the duration of the guided session.

Outdoor visibility was a separate and persistent concern. One participant in the Pilatus PC-9/A outdoor sessions reported that sunlight affected display visibility and interaction comfort, indicating that even with stable spatial anchoring, HoloLens 2 use in direct sunlight remains a practical constraint that system design must account for.

Navigation clarity also drew comment. One participant suggested adding a persistent next-step indicator to reduce the visual search effort involved in locating the next inspection point. Another noted that attending to the virtual instructions could reduce awareness of the surrounding aircraft structure. This reflects a tension between interface engagement and environmental awareness that future interface revisions will need to address explicitly.

The indoor Cessna 172S Skyhawk sessions were particularly useful for evaluating the interaction workflow in a controlled setting, where participants could focus on the virtual guidance without the additional constraints of working around the physical aircraft. The Pilatus PC-9/A outdoor sessions, conducted after spatial anchoring was introduced, provided a more realistic validation context, as participants interacted with the MR guidance directly around the aircraft with more stable hologram placement. Together, these sessions confirm that the system is functional and generally usable, while identifying alignment stability, outdoor visibility, and wayfinding as the priority design considerations for the formal experimental stage.

5. Discussion

The initial validation results confirm that the proposed MR framework is technically feasible for supporting aircraft pre-flight walk-around training across both testbeds. The system guided users through the inspection workflow, recorded task progression and completion time, and supported post-task evaluation through NASA-TLX and SUS questionnaires. For the validation stage, this was the intended outcome: to confirm that the workflow and evaluation pipeline functioned as intended before committing to a larger delayed crossover experiment.

The Pilatus PC-9/A subset provides the most informative early comparison, as it is the only aircraft for which both training conditions were available. MR-guided sessions showed shorter task completion times, lower perceived workload, and higher SUS scores than the paper-based sessions. That pattern is encouraging, but it does not constitute evidence that MR improves task performance or learning effectiveness. The sample is small, the dataset is unbalanced, and the two conditions used different timing methods: automatic logging for MR and manual timing for paper-based sessions. Each of those factors introduces variability that cannot be separated from any genuine method effect at this stage. The results are therefore best interpreted as a signal worth investigating in a controlled experimental stage, rather than as a basis for performance claims.

The Cessna 172S Skyhawk sessions served a different purpose: validating the MR workflow across indoor and outdoor environments. The indoor sessions produced high SUS scores and low workload ratings, suggesting that participants could engage with the virtual guidance effectively in a controlled setting. The outdoor session told a different story. The high workload and low usability score in that trial were tied to an earlier alignment workflow in which the virtual aircraft model could drift after manual positioning, a problem that was visible in the data before it was fully understood as a design issue. That observation directly motivated the spatial anchoring update applied in the subsequent Pilatus PC-9/A sessions.

The qualitative feedback adds something the quantitative results cannot: it shows that technical feasibility and usability are not the same thing in outdoor aircraft environments. Stable alignment, adequate display visibility under direct sunlight, and clear wayfinding around the aircraft are not refinements; they are prerequisites for the system to function as a training tool rather than an interaction exercise. A participant who is managing hologram drift or searching for the next inspection point is

not learning the walk-around procedure; they are managing the interface. Resolving that distinction is what the next experimental stage is designed to address, through a formal delayed crossover comparison with a larger and more balanced dataset, extended to cover task efficiency, workload, usability, sequence accuracy, learning gain, and retention.

6. Conclusions

This paper presented an MR training framework for aircraft pre-flight walk-around inspection using HoloLens 2, Unity, MRTK3, and OpenXR, implemented across two aircraft testbeds: the Cessna 172S Skyhawk and the Pilatus PC-9/A. The framework supports step-by-step spatial guidance, manual aircraft alignment, instructional content delivery, and local task logging. A structured performance measure selection approach was also proposed, organized around relevance, measurability, reliability, and comparability.

Initial validation confirmed that the system can support guided walk-around inspection, record task timing data, and collect post-task workload and usability measures. The PC-9/A subset showed favourable preliminary trends for MR-guided instruction across task time, perceived workload, and usability. The Cessna 172S Skyhawk sessions were more instructive in a different sense: they exposed the consequences of unstable spatial alignment under outdoor conditions and established the design requirements that the spatial anchoring update was built around. Neither result should be read as evidence of learning effectiveness, as that is not what a validation stage is designed to show. What it does establish is that the framework is functional, the evaluation pipeline is operational, and the key design variables for the formal experiment are now better understood.

The next step is the delayed crossover experiment with a larger and more balanced participant sample, comparing MR-guided and paper-based training across task completion time, NASA-TLX, SUS, sequence accuracy, task errors, learning gain, and retention. Parallel development work will focus on outdoor display visibility, wayfinding clarity, and spatial alignment robustness, the three factors the validation stage identified as most relevant to deployment and evaluation at aircraft scale.

Acknowledgments

The authors thank RMIT University, the School of Engineering and the RMIT Aviation Academy for supporting this research. Special thanks are due to Leighton Powell and the Aviation Academy instructors and staff for facilitating aircraft access, supporting the trial sessions, and providing operational guidance during the pre-flight walk-around activities. The authors also thank the study participants for their time and feedback during the initial validation stage.

Contact

Corresponding author: Miguel Higinio (S4139028@student.rmit.edu.au)

Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] Fumio Kishino Paul Milgram. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information*, E77-D(12):1321–1329, December 1994.

[2] Civil Aviation Safety Authority. Aircraft logbook. `https://www.casa.gov.au/aircraft/airworthiness/continuing-airworthiness/aircraft-logbook`, 2026.

[3] Henrik Eschen, Tobias Kötter, Rebecca Rodeck, Martin Harnisch, and Thorsten Schüppstuhl. Augmented and virtual reality for inspection and maintenance processes in the aviation industry. *Procedia Manufacturing*, 19:156–163, 2018. Proceedings of the 6th International Conference in Through-life Engineering Services, University of Bremen, 7th and 8th November 2017.

[4] Trung Hieu Tran, Yirui Jiang, and Leon Williams. Applications of mixed reality for smart aviation industry: Opportunities and challenges. *Modern Development and Challenges in Virtual Reality*, 2022.

[5] Ann-Kathrin Koschlik, Michael J. Scott, Thore Keser, Fiete Rauscher, Hendrik Meyer, Wim Verhagen, Pier Marzocca, Florian Raddatz, and Gerko Wende. Mixed reality for aircraft structural dent and buckle non-destructive evaluation. In *34th Congress of the International Council of the Aeronautical Sciences, ICAS 2024*, September 2024.

[6] Thore Keser, Rahel Louise Schmied-Kowarzik, Ann-Kathrin Koschlik, Lina Kaschub, Michael J. Scott, Rebecca Rodeck, and Gerko Wende. Augmented reality workflow for aircraft dent & buckle inspection. In *34th Congress of the International Council of the Aeronautical Sciences, ICAS 2024*, September 2024.

[7] Tauseef Gulrez, Vlado Kekoc, Doug Williams, Tony Mills, Wim Verhagen, James Ong, and Pier Marzocca. A physical load metric development for assessment of mixed reality in aircraft inspection tasks. In *Proceedings of the 19th Australian International Aerospace Congress (AIAC 2021)*, pages 167–172, 2021.

[8] Nikhil Satish and C. R. S. Kumar. Artsam: Augmented reality app for tool selection in aircraft maintenance. In Neha Sharma, Amol Goje, Amlan Chakrabarti, and Alfred M. Bruckstein, editors, *Data Management, Analytics and Innovation*, pages 569–581, Singapore, 2023. Springer Nature Singapore.

[9] Giorgia Lallai, Giovanni Loi Zedda, Célia Martinie, Philippe Palanque, Mauro Pisano, and Lucio Davide Spano. Engineering task-based augmented reality guidance: Application to the training of aircraft flight procedures. *Interacting with Computers*, 33(1):17–39, 01 2021.

[10] Arthur Dela Peña. Virtual reality in aircraft maintenance training: Transforming student engagement and competency development. *Journal of Interdisciplinary Perspectives*, 3(3):360–371, Feb. 2025.

[11] Hoang The Nguyen, An Hoang Huynh, Thuan Van Luu, and Son The Nguyen. High-fidelity vr simulation for aircraft maintenance training. *Aerospace*, 13(5), 2026.

[12] Wolfgang Vorraber, Johannes Gasser, Helena Webb, Dietmar Neubacher, and Philipp Url. Assessing augmented reality in production: remote-assisted maintenance with hololens. *Procedia CIRP*, 88:139–144, 2020. 13th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 17-19 July 2019, Gulf of Naples, Italy.

[13] Allison Jing, Kunal Gupta, Jeremy McDade, Gun A. Lee, and Mark Billinghurst. Comparing gaze-supported modalities with empathic mixed reality interfaces in remote collaboration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 837–846, 2022.

[14] Jui-Hung Liu, Chen-Yi Tsai, Jia-Zhou Li, and Chun-Chieh Wang. Design and implementation of an augmented reality-based interactive and real-time wind turbine maintenance auxiliary platform system. *Sensors & Materials*, 35, 2023.

[15] Mara Fuchs, Florian Beckert, Fiete Rauscher, Carina Goetz, Jörn Biedermann, and Björn Nagel. Virtual reconfiguration and assessment of aircraft cabins using model-based systems engineering. In *33rd Congress of the International Council of the Aeronautical Sciences, ICAS 2022*, 2022.

[16] Lea M Daling and Sabine J Schlittmeier. Effects of augmented reality-, virtual reality-, and mixed reality–based training on objective performance measures and subjective evaluations in manual assembly tasks: a scoping review. *Human factors*, 66(2):589–626, 2024.

[17] Quentin Loizeau, Florence Danglade, Fakhreddine Ababsa, and Frédéric Merienne. Methodology for the field evaluation of the impact of augmented reality tools for maintenance workers in the aeronautic industry. *Frontiers in Virtual Reality*, Volume 1 - 2020, 2021.

[18] Yuk-Ming Tang, Kin Man Au, Henry CW Lau, George TS Ho, and Chun-Ho Wu. Evaluating the effectiveness of learning design with mixed reality (mr) in higher education. *Virtual Reality*, 24(4):797–807, 2020.

[19] Federico De Lorenzis, Filippo Gabriele Pratticò, Maurizio Repetto, Enrico Pons, and Fabrizio Lamberti. Immersive virtual reality for procedural training: Comparing traditional and learning by teaching approaches. *Computers in Industry*, 144:103785, 2023.

[20] Xueyan Oh, Leonard Loh, Shaohui Foong, Zhong Bao Andy Koh, Kow Leong Ng, Poh Kang Tan, Pei Lin Pearlin Toh, and U-Xuan Tan. Initialisation of autonomous aircraft visual inspection systems via cnn-based camera pose estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11047–11053, 2021.

[21] Marco Picchi Scardaoni, Fabio Magnacca, Andrea Massai, and Vittorio Cipolla. Aircraft turnaround time estimation in early design phases: Simulation tools development and application to the case of box-wing architecture. *Journal of Air Transport Management*, 96:102122, 2021.

[22] Tauseef Gulrez, Vlado Kekoc, Lei Hou, Wim Verhagen, Riyaz Bhoyro, Shaoze Wu, Pier Marzocca, and Tony Mills. *Measure Future User Experience of Mixed Reality System for Complex Aircraft Maintenance Task Using Machine Learning*. Engineers Australia, Melbourne, 2023.