sustainability

MDPI

Article

Research on Training Effectiveness of Professional Maintenance Personnel Based on Virtual Reality and Augmented Reality Technology

Xiao-Wei Liu [1], Cheng-Yu Li [1,2], Sina Dang [1], Wei Wang [1,*], Jue Qu [1,3], Tong Chen [1] and Qing-Li Wang [1]

[1] Academy of Air Defense and Antimissile, Air Force Engineering University, Xi'an 710051, China
[2] Graduate School, Air Force Engineering University, Xi'an 710051, China
[3] School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: lhww11@sina.com; Tel.: +86-131-8609-5680

Abstract: The maintenance training method based on Virtual Reality (VR) and Augmented Reality (AR) technology has the characteristics of safety, no space limitation, and good reusability. Compared with the traditional training method, it can reduce the training cost, shorten the training period, and improve training effectiveness. Therefore, more and more maintenance training use VR and AR to replace training based on actual equipment to improve training effectiveness. However, in the context of multi-level tasks, there is still no clear research conclusion on how to choose training methods, maximize the advantages of each training method, and achieve higher training effectiveness. In response to this problem, this study constructed three training platforms based on VR, AR, and actual equipment, designed three maintenance tasks at different levels, and created a comparative analysis of the training effects of 60 male trainees under the three tasks and three training platforms. The results show that for single-level maintenance tasks, the training effect of the traditional group was significantly better than that of the AR group and the VR group. For multi-level maintenance tasks, the training effect of AR group was significantly better than that of the VR group. With the increasing difficulty of maintenance tasks, the training efficiency of the AR group was more than 10% higher than that of the VR group and traditional group and the AR group had less cognitive load. The conclusions of this study can provide a theoretical basis for the selection of training methods and evaluation design and help to formulate training strategies, thereby shortening the training period of professional maintenance personnel.

Keywords: virtual reality; augmented reality; maintenance training; training effectiveness; cognitive load

Citation: Liu, X.-W.; Li, C.-Y.; Dang, S.; Wang, W.; Qu, J.; Chen, T.; Wang, Q.-L. Research on Training Effectiveness of Professional Maintenance Personnel Based on Virtual Reality and Augmented Reality Technology. *Sustainability* **2022**, *14*, 14351. https://doi.org/10.3390/su142114351

Academic Editors: Fisnik Dalipi and Arianit Kurti

Received: 6 October 2022
Accepted: 31 October 2022
Published: 2 November 2022

1. Introduction

With the iterative update of construction machinery and equipment, complex maintenance tasks are increasingly difficult to train, the time cost of training professional maintenance personnel is gradually rising, and traditional training based on actual equipment often causes economic losses and personal safety problems due to accidents [1]. This limits the training frequency of professionals, which in turn results in longer training cycles. With the iterative update of hardware equipment, VR and AR are more and more frequently used in the training process [2] and the platform used for professional training is gradually changing from actual to virtual platforms [3]. Virtual interaction based on VR and AR has brought a secure, economical, and reusable solution to these problems [4], which are widely used in military [5–7], aerospace [8,9], industrial maintenance [10–12], clinical [13–15], and fire protection [16].

The development of VR technology makes it possible to transform maintenance training from an actual equipment platform to a 3D virtual platform. The existing VR training platforms include VR training media for monoscopic screens and immersive VR training media and VR training media for monoscopic screens. It can provide a higher level

of real-world information than immersion [4], avoid the appearance of simulated diseases, and support long-term training. In Langley's research, the training effectiveness of VR based on monoscopic screens and actual equipment platforms was compared and analyzed. The results show that VR training platforms can effectively reduce errors in task execution and obtain better training effectiveness [17]. In Langley's training system, trainees interact with VR images in fixed projectors using Wii Mote controllers, which is different from monocular VR training based on mobile devices (such as smart phones and tablets). The latter is more in line with people's daily interaction habits and can reduce the time cost of learning to use interactive devices. For example, in [18], users can interact with virtual 3D models on smart phones directly by touch and learn the knowledge provided in VR space. From the previous research, it is meaningful to compare the training effectiveness between the latest AR Head-Mounted Display (HMD) and VR based on intelligent devices. However, Gavish's research points out that, compared with the real platform, VR training based on monoscopic screens needs longer training times and there is no significant difference in task accuracy. This difference is related to the type and difficulty of the task. Compared with VR training media, AR training requires shorter training times, higher accuracy, and higher training satisfaction [12].

The current AR training system gradually relies on commercial creative tools (such as Unity 3D, Unreal, and Amazon Sumerian Engine [19]) and expensive smart glasses equipment for development. AR HMD, a tool that frees hands, has become the main AR interactive medium. Microsoft released the HoloLens 2 in November 2019, which is the latest AR HMD. Compared with the old solutions such as AR projectors and AR handheld devices, the HoloLens 2 brings more functions and a more powerful platform. It has already spawned many applications for training in the field of Industrial Maintenance and Assembly (IMA) [13,20]. Many researchers build an AR training environment by binding external devices to HMD devices. AR HMD-based training is usually based on the 3D model or digital twin model of actual equipment. Ada has established a set of solutions about the future shipyard training and maintenance process based on AR digital twins. Trainees can perform training with full-scale models of actual equipment through HoloLens and receive step-by-step instruction [21]. Henderson and Feiner explored the benefits of AR display for mechanic maintenance training. Based on tracking HMD, they simplified task understanding in the form of text, labels, arrow and animation sequences, and enhanced interactive performance. After testing, the AR HMD improves the effectiveness of task execution [22,23]. Siyaev developed AR application for Boeing 737 aircraft maintenance training based on the Aircraft Maintenance Manual (AMM) and HoloLens 2 and also adopted the way of guidance and visual superposition. At the same time, trainees are provided with step-by-step instructions of a certain task, including videos, illustrations, and manual files, which improve the level of virtual interaction [24]. It is worth noting that the interactive devices of VR HMD or immersion AR are often accompanied by obvious simulation diseases due to their strong immersion. For this reason, Wiederhold suggests that the use time should be limited to 30 min, otherwise, the trainees may be dizzy or even vomit in different degrees [25]. However, in the test results of simulated diseases with 142 subjects, the symptoms of simulated diseases based on HoloLens can be neglected [26], which means that an AR training system based on HoloLens can support a longer training time.

Although VR and AR are both potential training methods, from the research of [4,27], these research experiments have drawn many contradictory conclusions. There is experimental evidence that training based on VR and AR is worse than traditional training based on actual equipment in terms of time, errors, and subjective experience. Therefore, there is always a debate among researchers about the effectiveness of AR/VR tools in training professionals and operators [28]. At present, it is not clear that AR technology is applicable to specific task types and crowd types [29] and researchers still cannot guarantee that AR/VR tools will show positive training effectiveness. Some studies aim at comparing the differences of training effectiveness through various evaluation indicators. For example,

Keighrey takes physiological indicators into account when evaluating the interactive quality and takes the perceived Quality of Experience (QoE) of users as the standard to measure the interactive experience. The results show that AR and tablets are better [30]. Werrlich compares the results of AR training and paper training in manual assembly tasks. The evaluation indicators include time, errors, System Usability Scale (SUS), User Experience Questionnaire (UEQ), and Cognitive Load Scale (NASA-TLX). The results show that AR is dominant in all but task completion time [31].

To sum up, although the training based on actual equipment has the most real and intuitive advantages, it also has the disadvantages of poor visualization effects and a lack of interactivity. The training based on VR platforms provides more interactivity and can provide more guidance information, but the disadvantage is also obvious, that is, it is out of touch with the real world. The training based on the AR platform emphasizes the connection with the real world and can ensure that the AR virtual model and the actual platform have the same three-dimensional positional relationship. The disadvantage is that it cannot provide a real operating experience and the visualization effect is not as clear as the VR platform. Therefore, under different levels of task difficulty, how to choose the appropriate training methods to cause the training effectiveness to be higher has become a problem to be studied. Aiming at different levels of task difficulty, this study selects VR, AR, and traditional training platforms for research.

According to the above research background, this paper aims to answer the following research questions:

1. What are the factors that measure the training effectiveness? How can we distinguish and quantify the influence of these factors on training effectiveness?
2. What kind of maintenance tasks are training based on VR, AR, and traditional methods suitable for and how to choose training media to achieve better training effectiveness?
3. For the same kind of maintenance tasks, how does the training platform affect the trainees' learning processes, resulting in better or worse training effectiveness?
4. Which method can help reduce the correction effect of individual differences on training effectiveness during training?

In order to answer these questions, this paper designs three types of maintenance tasks with different difficulties, and conducts comparative experiments based on the three types of training platforms. During the experiment, the homogeneity of the operating objects under the three types of training platforms is guaranteed. Sixty male trainees were equally divided into three groups. They received training on VR, AR, and traditional platforms, respectively, and then they were assessed on the actual equipment. The time and errors recorded in the assessment process would be taken as objective factors to measure the training effectiveness. At the same time, a NASA-TLX questionnaire was used to measure the cognitive load of the trainees as subjective factors to measure the training effectiveness.

In order to better quantify the training effect, the trainees' time, decision-making, and error data in the assessment process were quantified using a CPSI (Cognitive and Psychomotor Skill Index) model. Cognitive ability is the most important factor in individual differences. This study has tested each trainee's cognitive abilities through experiments and quantified them as a covariate in a one-way covariance analysis, which reduced the correction effect of individual differences on training effectiveness to a certain extent.

2. Experimental Platform

2.1. Setting up Training Environment

Three train platforms based on crane maintenance tasks were involved in this experiment, these include the traditional training platform (watching video learning and actual operation), VR training platform (3D model and animation tutorial), and AR training platform (holographic virtual model and no actual operation). Each type of training platform includes four parts: (1) main console: used for startup and fault transplantation; (2) crane operation console: it includes instrument, rocker switch, power supply, and relay for operating the actuator; (3) crane movement mechanism: execute the movement; and

(4) maintenance tools. The display modes of the experimental equipment under the three training platforms are shown in Figure 1.

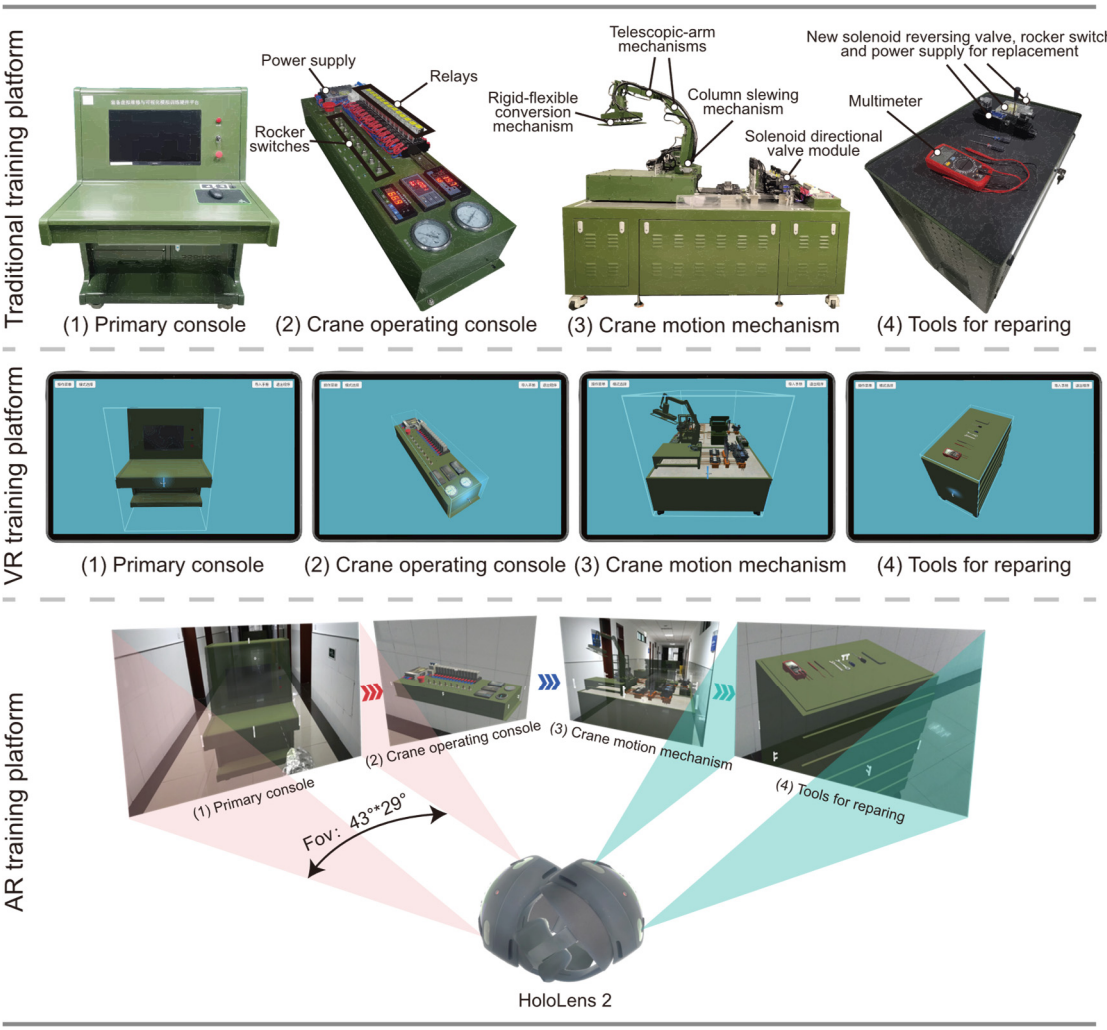


Figure 1. Comparison diagram of three kinds of experimental platforms showing the same equipment.

The traditional training platform can support fault transplantation and actual maintenance and the teaching and examination process is realized in the actual use process. For the VR training platform, all the devices and components are modeled and placed in the 3D virtual space based on a tablet computer. The specific model of the device is a Huawei Matepad 11. According to Daling's definition of this interaction mode [4], it is specifically named VR interaction based on monoscopic screens. The AR training platform was built in an empty room (Xi 'an, Shaanxi Province, China) based on a HoloLens 2 HMD device from the Microsoft Corporation (Redmond, WA, USA) and the virtual 3D models were placed in the real space. The size of the virtual models is the same as that of the actual equipment. Users can walk around in the space wearing the HoloLens 2. Based on the real-time registration and tracking principle of the device, these 3D models will remain relatively stationary with the actual space and users can browse the details of the models and operations by walking and rotating the visual angle.

2.2. Training Methods for Maintenance of Three Kinds of Platforms

The maintenance training methods of the three training platforms are different. The trainees who train based on the actual equipment all use the traditional tool (TT) and are named as the TT group. Gao summarized the traditional tools as: Toolbox talk, video

presentations, text-based handouts, practical training, and teaching seminars [1]. As shown in Figure 2, in this experiment, the trainees watched the video first and then performed the actual operation.

Figure 2. Demonstration video for traditional training, using the maintenance of electromagnetic directional valve as an example.

The maintenance training based on VR and AR training platforms is realized in the form of animation and an interactive dialog box. The VR maintenance training scenario is shown in Figure 3. The user can receive the step text description from the upper left corner in the interface, the object involved in the current step will also be highlighted, and the maintenance method will be conveyed in the form of an animation, which shows the animation of the disassembly and replacement of the electromagnetic directional valve. In the 3D space of VR, an indicating arrow will be provided, which can guide the user to adjust the angle of view to the correct direction when the user's angle of view is not on the relevant object of the current step.

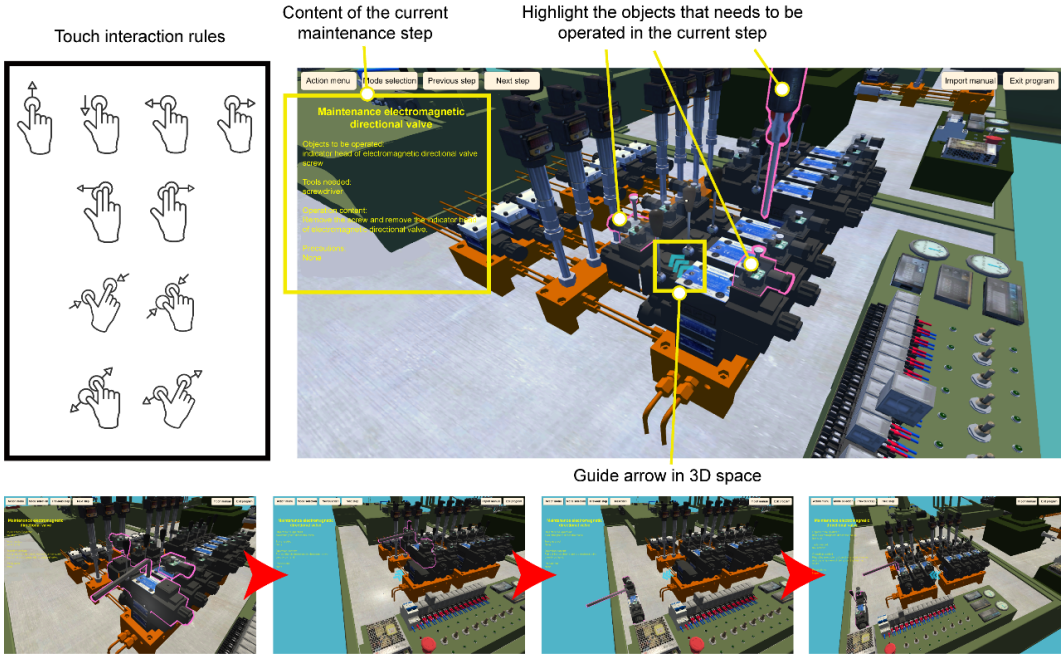


Figure 3. Training scene in VR, using maintenance of electromagnetic directional valve as an example.

In the AR training interface, visual guidance is a commonly used auxiliary strategy [32]. Highlighting, clues, 3D models, and animation enhancement methods are added to the AR training environment. This kind of real and dynamic 3D visual visualization can help trainees understand the space more easily [33], thus achieving better training effectiveness. The AR training ground scene in this study adopts the combination of highlighting, animation, and 3D models to enhance learning. As shown in Figure 4, trainees learn the current step by watching the animation demonstration and the objects to be operated in the space will be highlighted and prompted. Trainees can see the detailed instructions of the current

step in the floating window, such as the names of operating objects, operating methods, and precautions. These instructions will also be broadcast to trainees in the form of voice.

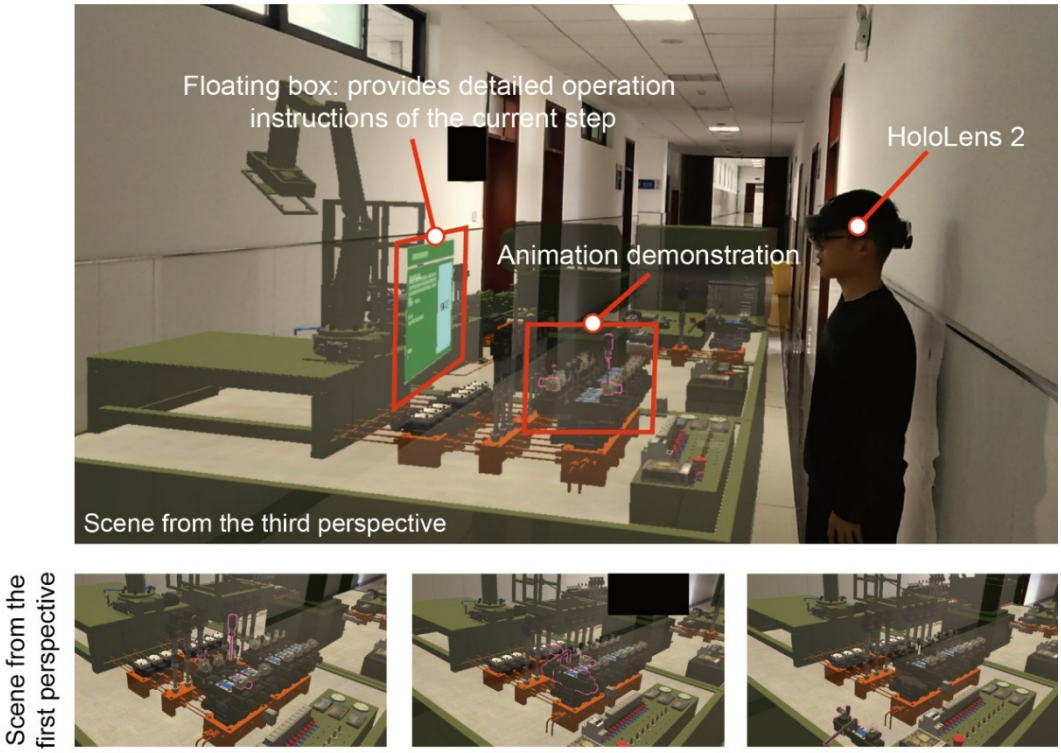


Figure 4. Training scene in AR, using maintenance of electromagnetic directional valve as an example.

2.3. Analysis of the Advantages and Disadvantages of Three Types of Platform Training

As a traditional training method, watching demonstration videos has achieved good results in both teaching and training. The advantage is that the trainees can access the operating equipment and learn and train related maintenance subjects more intuitively. The disadvantage is that too long videos may cause trainees to miss or forget steps when operating the equipment, resulting in repeated viewing of videos, thus reducing training effectiveness.

The VR training scene in this study is based on the smart tablet, which is more interesting than the demonstration video, and supports trainees to browse freely in 3D space. The combination of highlighting, arrows, and words can bring more information to trainees and the touch interaction based on tablets is more in line with people's daily habits. The disadvantage is that the gap between the virtual maintenance environment and actual maintenance environment is too large, which may lead to a poor learning transfer effect.

Compared with the VR training scene, the AR training scene better simulates the actual maintenance training environment. Trainees can learn interactively with 3D virtual models in a real environment and a combination of highlighting, arrows, and floating windows is used to enhance learning. The disadvantage lies in the limitation of viewing angles and potential discomfort with the HoloLens. Compared with VR training, the interaction mode in AR training is brand new, which may cause interaction limitation.

To sum up, the performance of training effectiveness of the three kinds of training methods in the same task is unpredictable and whether the relative applicability of training methods changes due to the change of task complexity needs to be verified based on experimental methods. Therefore, this study carried out a comparative experiment, aiming at choosing the best training method based on the task characteristics.

3. Experiment

3.1. Experimental Process

This experiment recruited 77 postgraduate students majoring in telecommunications, aged 22–31 years (M = 24.43, SD = 1.720), all male, with normal or corrected visual acuity, no color blindness, color weakness, and right-handedness. No female trainees were involved in this experiment for objective reasons of the academy. The experiment scheme is shown in Figure 5. Figure 5a shows the screening process of the trainees. First, through the technical affinity test, the obtained test scores can be used to screen out possible individuals with a high technical affinity through the data analysis method of systematic clustering. For the remaining individuals, a cognitive ability test is needed, and the scores of cognitive ability test will be used as covariates in one-way covariance analysis in order to reduce the influence of individual cognitive differences on the experimental results. Figure 5b shows the situation of personnel grouping and the schematic diagram of the training mode of each group.

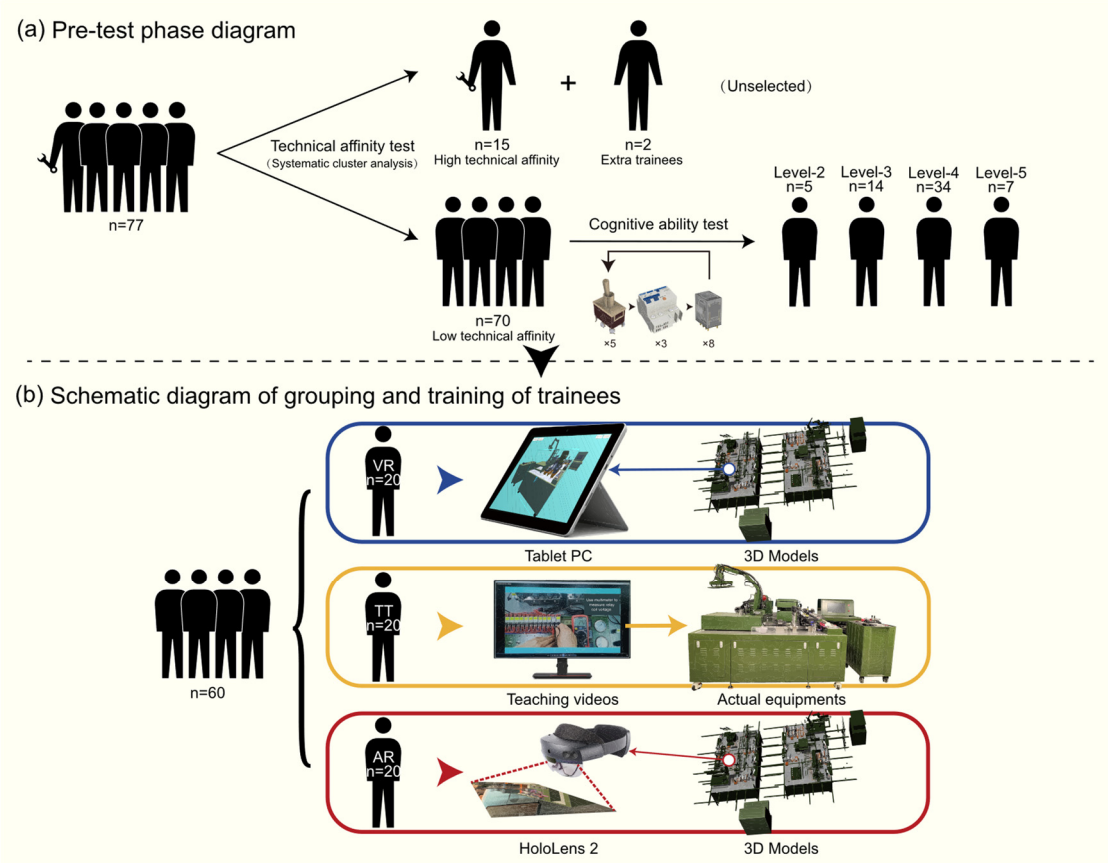


Figure 5. Schematic diagram of the experimental process.

First, a pre-test experiment was conducted to select individuals without AR experience and measure the cognitive ability of the experimental individuals. The number of operation steps of each trainee was measured by the cognitive ability experiment, which was used to measure the result of the combined effect of human factors. Then, the trainees entered the training preparation stage. The experimental steps and durations are shown in Figure 6. The trainees were equally divided into AR, TT, and VR groups and provided with a 10 min training lecture so that they could adapt to their respective training systems as much as possible and prepare for the training equipment to be used. After the preparation, the trainees completed the training and assessment tasks based on their respective training platforms. Each completed task assessment completed a NASA-TLX scale. The trainees

could rest for 10 min before proceeding to the next task. After completing all tasks, the trainees in VR and AR groups were subject to opinion polls.

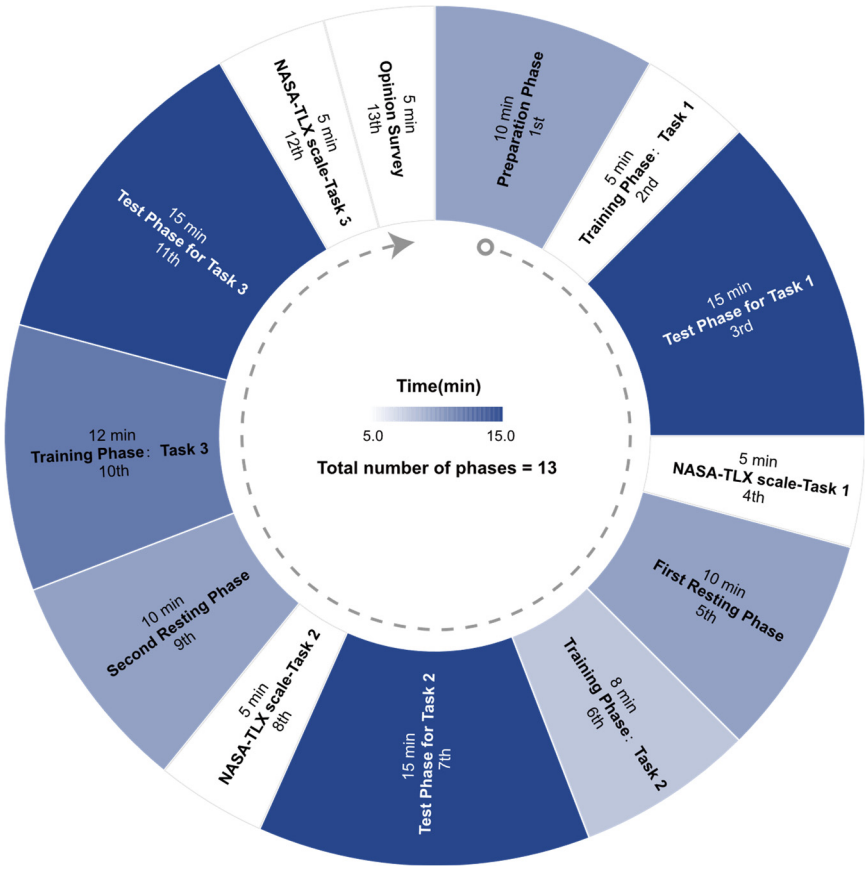


Figure 6. Steps and duration of the experiment.

3.2. Pre-Test Experiment

The purpose of the pre-test experiment is to screen the experimental individuals and reduce the individual differences between groups and within groups. The experimental plan is divided into two parts:

1. Screening out individuals with high technical affinity through questionnaires;
2. An experiment to test individual's cognitive ability aims to quantify the result of the comprehensive effect of individual factors.

3.2.1. Technical Affinity Test

A technical affinity questionnaire was designed for the training experiments of this system, as shown in Table A1 in Appendix A. The questionnaire adopted the Likert 5-level scoring structure and the content validity of the five items was excellent (I-CVI = 1.00) for the experts in the same group. The specific content is as follows:

- Number of AR/VR interactions;
- Number of times to learn about the mechanical maintenance system;
- Number of times to use the HMD device;
- Number of times you experience AR/VR games;
- Knowledge of AR/VR technology.

The reliability analysis of the data of 77 questionnaires shows that the Cronbach's alpha coefficient of the technical affinity questionnaire is 0.904, which is greater than 0.7 and indicates that there is a high internal consistency among the survey items. The construct validity was verified using exploratory factor analysis. The results of the KMO test and

Bartlett sphericity test showed that the data were suitable for factor analysis (KMO = 0.759, $p < 0.001$). A common factor F1 whose eigenvalue is greater than one is extracted by orthogonal rotation using principal component analysis, the results are shown in Table 1. It can be seen that the cumulative contribution rate of the common factors is 72.972%, the Combined Reliability (*CR*) is greater than 0.7, and the Average Variance Extraction (AVE) is greater than 0.5, which indicates that the scale has a good convergent validity.

Table 1. Construct validity check of technical affinity questionnaire.

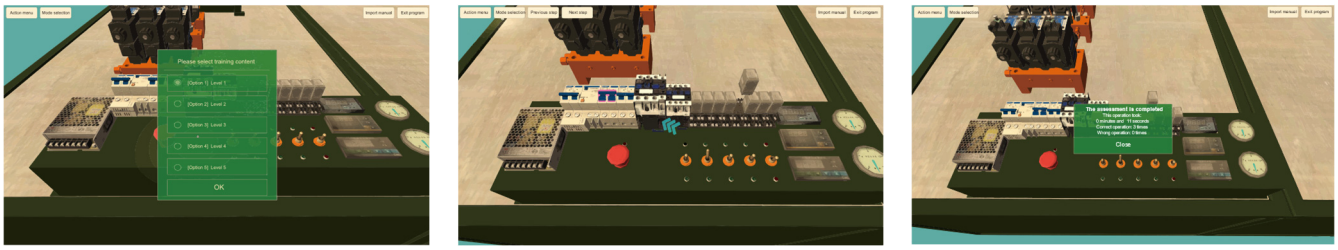
Factors	Factor Loading	Cumulative Contribution Rate	Cronbach's α	*CR*	AVE
F1	Q1:0.913 Q2:0.712 Q3:0.946 Q4:0.780 Q5:0.897	72.972%	0.904	0.930	0.730

Due to the small number of samples (n < 200), the average Euclidean distance of intervals is measured by the average inter-group method in the system clustering to realize the two-classification of data. According to the results of systematic clustering, the subjects with high score significance (n = 15) were removed from the experiment. To ensure that each group had the same number of subjects, the individual with the highest score (n = 2) was also removed.

3.2.2. Measure Individual Cognitive Ability

After screening, 60 trainees were involved in the experiment. The cognitive load theory (CLT) assumes that an individual's working memory is limited [34,35], which can explain the learning mechanism in a virtual environment [36]. The experiment measured the maximum number of maintenance steps that each trainee could remember in a short time, aiming to measure the results of the comprehensive effect of individual factors (such as interest, motivation, responsibility, self-efficacy [37], and working memory capacity [38]).

- The cognitive ability measurement experiment is based on a crane maintenance task. The operational element Q in the experimental task includes three basic elements: Q_{r1}~Q_{r8} (relays), Q_{cb1}~Q_{cb3} (circuit breakers), and Q_{rs1}~Q_{rs5} (rocker switches).
- As shown in Figure 7, the trainee needs to first select the task level to perform, which will start at level 1. Then, watch the operation demonstration, each step has a time limit of 5s, and then complete the assessment task in the assessment mode; the assessment data will be recorded by the system. There are two attempts for each person before moving to the next level without errors.
- The operation elements of each level are randomly arranged and combined by three basic elements of Q_r, Q_{cb}, and Q_{rs}, the number of operation elements in level 1 is $X_1 = 3$, the number of operation elements in level n is $X_n = X_{n-1} + 1$, $n \geq 2$.
- The final grades were all within the range of Level 2 to Level 5, and the number of steps Q completed by the trainee was recorded. It can be seen that Q is a continuous variable. It was set as a covariate in the one-way covariance analysis to reduce the influence caused by individual cognitive ability differences.



(**a**) (**b**) (**c**)

Figure 7. Cognitive Ability Test. (**a**) Interface for selecting mission level. (**b**) Prompt the trainee to the object that needs to be operated in the current step, which will last for 5 s. (**c**) Assessment results, including time, number of correct and incorrect.

3.3. Training and Assessment Tasks

As shown in Figure 8, three tasks with different difficulties are set in training and assessment, which involve different maintenance steps, such as detection, replacement, repair, and calibration. The figure shows the number of steps and descriptions for each process. The difficulty of the task is defined by the number of steps and the number of failures.

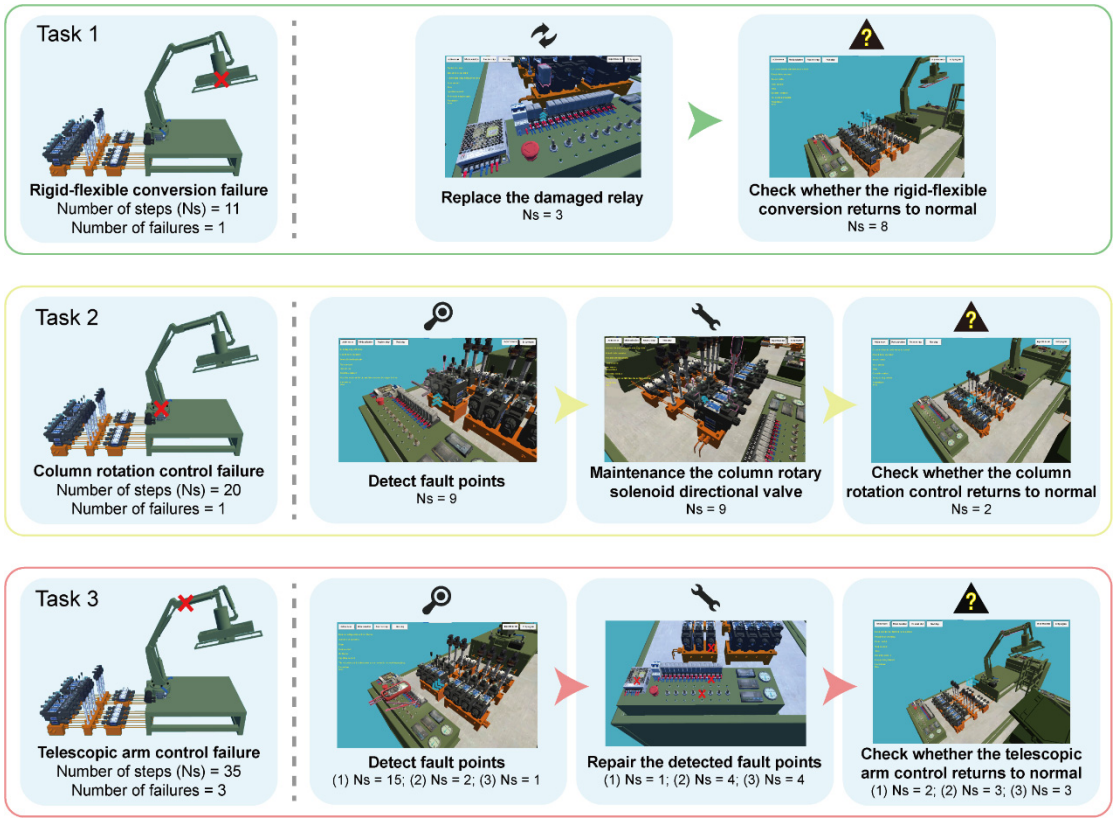


Figure 8. Three types of maintenance tasks of different difficulty.

In the first task, the task characteristic is set to be single-threaded, which means that there is no decision-making link in the task. This mission setting simulates the situation that maintenance is performed by maintenance personnel when the point of failure is known. The failure point selected "damage of rigid–flexible conversion relay". During the task, the trainee needs to replace the rigid–flexible conversion relay and check whether the rigid–flexible conversion is back to normal.

In the second task, the steps of measurement and diagnosis are added. This task setting simulates the situation that the maintenance personnel perform detection, diagnosis, and emergency repair based on the phenomenon when the system fails. The selected fault point is "column rotation control failure". The trainee only needs to detect a minimum fault cut set (column rotation–electromagnetic reversing valve), where the column rotation control fails. After inspection, check whether the column rotation returns to normal.

In the third task, multiple points of failure are set. Different from the first two types of tasks, this task tests the trainee's self-judgment and logical thinking ability, while the first two tasks are procedural tasks in nature, mainly examining the trainee's ability based on working memory. The selected fault point is "telescopic arm control failure" and trainees need to detect all possible failure points (hydraulic lock, rocker switch, power supply, and wiring) for the telescopic arm. To avoid learning effects, there are no duplicate failure points in all tasks.

3.4. Post-Test

The NASA-TLX scale was selected to measure cognitive load in the post-test, which will be performed in time after each task is completed. In addition, in order to better improve the system and increase the usability, the trainees in the VR group and AR group will also accept the system's opinion survey at the end of the experiment.

3.5. Quantization Method Based on Improved CPSI Model

As one of the objective factors, errors in experiments need to be quantified and weighted [4]. After referring to the error analysis procedure proposed by Stanton [39] and Renson's error classification system [40], the following six basic error types were summarized in this experiment:

- Touch by mistake;
- Wrong location;
- Missing a step;
- Operation error;
- Decision error;
- Repair task not completed.

The CPSI model developed by Nalin's team quantifies the time factor, accuracy factor, decision-making factor, and relationship factor based on task performance [41], but it is not suitable for the simultaneous evaluation of multiple tasks, nor is it compatible with the situation that there are multiple sub factor variables under the same factor. Therefore, this part was supplemented, the coefficients were readjusted, and the scores were normalized.

The accuracy score (*Acc*) is calculated using Formula (1)–(2) and E_i represents the error set in the task i, including e_1 (false touch), e_2 (sequence or position error), e_3 (missing step), and e_4 (dangerous operation). Acc_i represents the score of the task i, act_i represents the total number of steps in the task i, and X represents the weight matrix, as shown in Table A2 in Appendix A.

$$Acc_i = (act_i - XE_i)/act_i \quad (1)$$

$$E_i = [e_{1i}, e_{2i}, e_{3i}, e_{4i}]^{\mathrm{T}} \quad (2)$$

The decision score (*Dec*) measures the trainee's decision-making level and the errors generated in the decision-making process, normalized by Formula (3), where Dec_i represents the score of the task i and the apt_i value is shown in Table A1 in Appendix A.

$$Dec_i = \lg apt_i \quad (3)$$

The completion score (*Com*) measures the completion degree of the trainee for the maintenance task, which is normalized by Formula (4), where Com_i represents the score of the task i, R_i represents the completed steps in task i, and R_t represents the total steps number.

$$Com_i = R_i/R_t \quad (4)$$

Due to the separate analysis of the time T for the trainee to complete the task, the quantitative method of time score in the CPSI model is not adopted. The final *CPSI* is determined by Formula (5), where n represents the number of *CPSI* sub-elements, $P_i = (r_j)_{1\times n}$ $(j = 1, 2, \ldots, n)$, $r_j = 1/n$.

$$CPSI_i = P_i\,[Acc_i, Dec_i, Com_i]^{\mathrm{T}} \quad (5)$$

To sum up, the objective performance the OP_i of the task i in the experiment can be determined by Formula (6).

$$OP_i = [T, CPSI_i] \quad (6)$$

3.6. Data Analysis

The data analysis work in this experiment was all based on IBM's statistical products SPSS Statistics 25 and Amos 26 (IBM, Armonk, NY, USA).

3.6.1. One-Way Analysis of Covariance for Time Parameters

To verify the difference in time performance of each training group, the training method was set as an independent variable that affected the task time, the number of steps Q was set as a covariate, and the time differences (T_1, T_2, T_3) from the start to the end of the three types of tasks were set as a dependent variable, with the significance level set to 0.05.

The detailed results for the time parameter T are presented in Appendix B, Table A3. T_{ij} represents the execution time of the jth group under the ith task. The Shapiro–Wilk test results showed that variables T_1, T_2, and T_3 were in normal distribution ($p > 0.05$). The Levin's statistical results and inter-subjective effect test results show that T_1, T_2, and T_3 variables satisfy the hypothesis of homogeneity of variance ($p < 0.05$) and homogeneity of slope ($p > 0.05$).

The results of statistical analysis of the effects of training methods on T_1, T_2, and T_3 are presented in Table 2.

Table 2. Test whether the difference in training method constitutes a significant difference in T_1, T_2, T_3 parameters between groups.

Items	ss	Partial η^2	df	MS	*F*	Sig.
T_1	878.525	0.583	2	439.262	39.140	<0.001 ***
T_2	2060.539	0.262	2	1030.270	9.096	<0.001 ***
T_3	3651.474	0.117	2	1825.737	3.702	0.031 *

* When the significant level is 0.05 (two-tailed), the difference is significant. *** At a significant level of 0.001 (two-tailed), the difference is significant.

It can be seen that for the variables T_1 ($F = 39.140$, partial $\eta^2 = 0.583$, $p < 0.001$), T_2 ($F = 9.096$, partial $\eta^2 = 0.262$, $p < 0.001$), and T_3 ($F = 3.702$, partial $\eta^2 = 0.117$, $p < 0.05$), the differences in training methods caused significant differences between groups. Multiple comparisons were adjusted using the Bonferroni method and the results of the pairwise comparisons are presented in Table 3, with the data presented in the table indicating:

Table 3. About T_1, T_2, T_3 descriptive statistics and post hoc pairwise comparisons.

Groups		Descriptive Statistics			Pairwise Comparison	
		n	Average	*SD*	Comparison Group	Sig. [a]
T_1	VR	20	35.689	3.196	TT	<0.001 ***
	TT	20	28.816	2.918	AR	<0.001 ***
	AR	20	37.789	3.877	VR	0.123
T_2	VR	20	170.669	12.148	TT	<0.001 ***
	TT	20	155.570	9.820	AR	0.033 *
	AR	20	163.833	9.972	VR	0.228
T_3	VR	20	172.439	20.347	TT	1.000
	TT	20	175.486	27.491	AR	0.034 *
	AR	20	156.697	20.862	VR	0.168

[a] Multiple comparison adjustment: Bonfiglioni method. * When the significant level is 0.05 (two-tailed), the difference is significant. *** At a significant level of 0.001 (two-tailed), the difference is significant.

For the variable of T_1, the task completion time T_1 of trainees in the VR group (M = 35.689, SD = 3.196) and AR group (M = 37.789, SD = 3.877) was significantly higher than that of the TT group (M = 28.816, SD = 2.918), but there was no significant difference between the VR group and AR group.

For the T_2 variable, the task completion time T_2 was significantly higher in the VR (M = 170.669, SD = 12.148) and AR groups (M = 163.833, SD = 9.972) than in the TT group (M = 155.570, SD = 9.820) and there was no significant difference between the VR and AR groups.

For the T_3 variable, the task completion time T_3 of the trainees in the TT group (M = 172.439, SD = 20.347) was significantly higher than that of the trainees in the AR group (M = 156.697, SD = 20.862) and there was no significant difference between the other groups.

The time parameter is one of the factors in the objective effectiveness model. The results of descriptive statistics and inter-group significance analysis for the time parameters are presented in Figure 9. It can be concluded that the time parameter T is significantly affected by the training method in the three types of maintenance tasks with different difficulties ($p < 0.05$). With the gradual increase in task difficulty, the trainees in VR and AR groups gradually eliminated the significant difference from the trainees in the TT group. Particularly in Task 3, the task time of the trainees in the AR group was significantly less than that of the other two groups, which meant that the trainees in the VR and AR groups performed better and better in time compared with the trainees in the TT group. In addition, there is no significant difference in task time between the VR group and AR group.

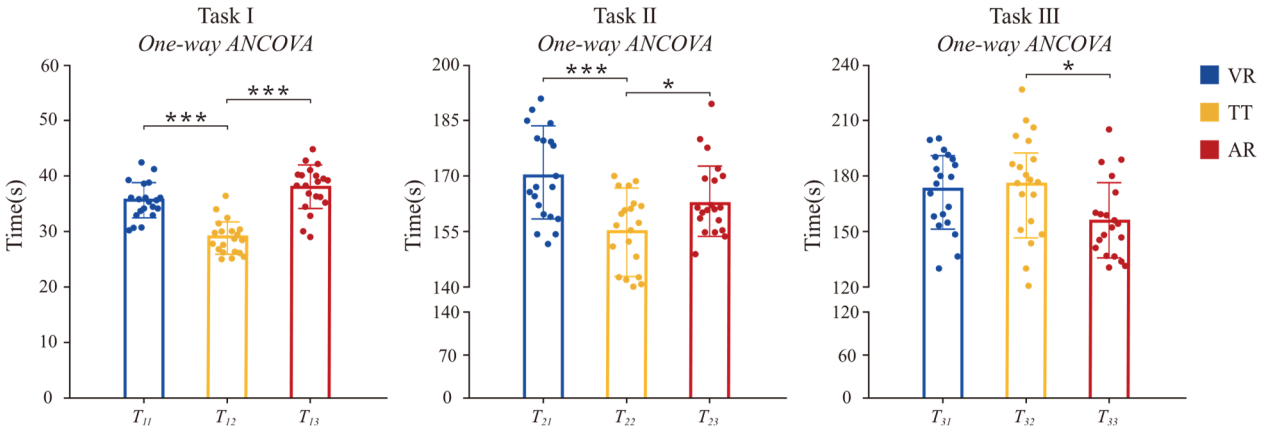


Figure 9. Results of between-group significance analysis of the time parameter in objective performance. * means p value ≤ 0.05, *** means p value ≤ 0.001, if not indicated, means p value > 0.05.

3.6.2. Nonparametric Analysis of CPSI Score

In this experiment, CPSI score is another factor of objective performance OP, which builds a mathematical model through three indicators of error, completion, and decision-making. The error metrics are classified and weighted. The higher the CPSI score, the better. All data were normalized in the mathematical model; detailed statistical results about the CPSI scores can be seen in Table A4 in Appendix B. $CPSI_{ij}$ represents the CPSI score of the jth group under the ith task.

According to the S–W test, $CPSI_{1j}$, $CPSI_{2j}$, and $CPSI_{3j}$ did not meet the normality test ($p < 0.05$), which was further tested using the Kruskal–Wallis test nonparametric method. The results are shown in Table 4. It can be seen that for Task 1, different training methods did not constitute a group difference in the CPSI score ($p > 0.05$), while Task 2 and Task 3 constituted a significant difference between the groups ($p < 0.05$).

Table 4. Kruskal–Wallis test results for CPSI scores in three tasks.

	Task I			**Task II**			**Task III**		
	H(K)	**df**	**Sig.**	**H(K)**	**df**	**Sig.**	**H(K)**	**df**	**Sig.**
CPSI	1.809	2	0.405	7.498	2	0.024 *	8.687	2	0.013 *

* When the significant level is 0.05 (two-tailed), the difference is significant.

Based on the post-event Mann–Whitney U test shown in Table 5, the VR group (M = 0.921, SD = 0.054) had significantly lower CPSI scores than the TT group (M = 0.963, SD = 0.024) on the completion of Task 2. For Task 3, the CPSI scores of the AR group (M = 0.930, SD = 0.109) were significantly higher than those of the TT group (M = 0.808, SD = 0.168) and the mean CPSI scores were 10.45% and 15.10% higher than those of the VR and TT groups, respectively. There were no significant differences among the other groups.

Table 5. Results of post hoc pairwise comparisons of CPSI scores using the Mann–Whitney U test.

	Task II				**Task III**			
Items	**Mean Rank**	**U**	**Z**	**Sig(2-Tailed)**	**Mean Rank**	**U**	**Z**	**Sig(2-Tailed)**
VR vs. TT	15.88 25.13	107.5	−2.528	0.011 *	22.53 18.48	159.5	−1.097	0.273
TT vs. AR	22.48 18.52	160.5	−1.087	0.277	15.00 26.00	90.0	−2.987	0.003 *
VR vs. AR	17.00 24.00	130.0	−1.912	0.056	17.40 23.60	138.0	−1.686	0.092

* When the significant level is 0.05 (two-tailed), the difference is significant.

The results of descriptive statistics and inter-group significance analysis of CPSI scores are plotted in Figure 10. From the results, it can be concluded that different training methods caused the significant difference on this factor in Task 2 and Task 3 ($p < 0.05$), but this was not reflected in Task 1.

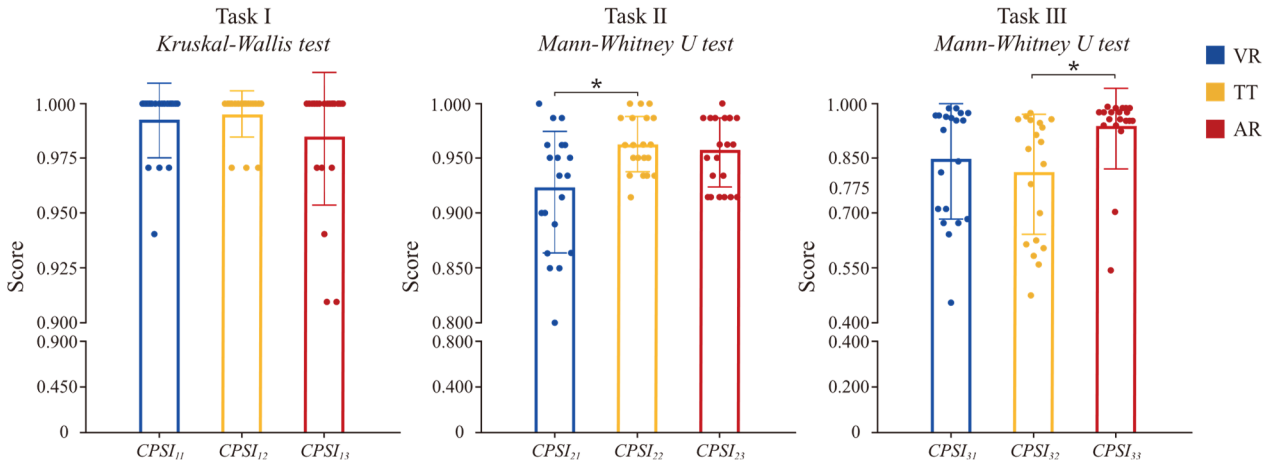


Figure 10. Results of between-group significance analysis of CPSI scores in objective performance. * means p value ≤ 0.05, if not indicated, means p value > 0.05.

According to the data results of two indicators of objective performance, for tasks with different levels of difficulty, differences in training methods resulted in differences in time parameters and CPSI scores, which helped to select the most appropriate training method according to the difficulty of completing the task faster and minimizing errors in the process.

3.6.3. One-Way Analysis of Covariance on Cognitive Load Scores

In this study, the NASA-TLX scale was used to measure the variable of the cognitive load. This measurement process would be conducted in time after the completion of each task to finally obtain the cognitive load score of each task. The lower the cognitive load score, the better. In addition, in order to better improve the system and increase the availability, the trainees in the VR group and AR group at the end of the experiment provided their opinions about the system.

Table A5 in Appendix B presents a hypothesis test for the cognitive load scores for the three tasks, with N_{ij} representing the cognitive load score for the jth group under the ith task. After testing, it was found that the N_{ij} all met the assumption of normality, homogeneity of variance, and homogeneity of slope ($p > 0.05$).

Table 6 gives the statistical analysis results of the influence of the training method on N_1, N_2, and N_3. It can be concluded that there is a significant difference between the groups in Task 3 ($p < 0.05$), but there is no significant difference between the groups in Task 1 and Task 2 under the influence of training methods.

Table 6. The significance test of the influence of training method on cognitive load index.

Items	ss	Partial η^2	df	MS	*F*	Sig.
N_1	1303.642	0.043	2	651.821	1.256	0.293
N_2	6365.045	0.099	2	3182.522	3.071	0.054
N_3	6111.054	0.129	2	3055.527	4.145	0.021 *

* When the significant level is 0.05 (two-tailed), the difference is significant.

According to the pairwise comparison results in Table 7, there was no significant difference in cognitive load scores between the groups in Task 1 and Task 2. In Task 3, the AR group (M = 189.80, SD = 23.539) had significantly lower cognitive load scores than the VR group (M = 212.85, SD = 30.505), but there were no significant differences among the other groups.

Table 7. Post hoc pairwise comparison of cognitive load scores in Task 1 to Task 3 among the three groups.

	Task I		Task II		Task III	
Items	**SE**	**Sig.** [b]	**SE**	**Sig.** [b]	**SE**	**Sig.** [b]
VR vs. TT	7.233	0.999	10.222	0.983	8.588	0.999
TT vs. AR	7.206	0.364	10.183	0.051	8.621	0.134
VR vs. AR	7.251	0.999	10.247	0.447	8.642	0.022 *

[b] The least significant difference method. * When the significant level is 0.05 (two-tailed), the difference is significant.

The results of descriptive statistics and inter-group significance analysis on cognitive load scores are plotted in Figure 11. From the results, it can be concluded that with the increase in task difficulty, the TT group with the lowest cognitive load gradually lost its advantage, while the AR group gradually showed its advantage. Specifically, in Task 3, the mean cognitive load was reduced by 10.83% and 8.20% in the AR group compared with the VR combined with the TT group.

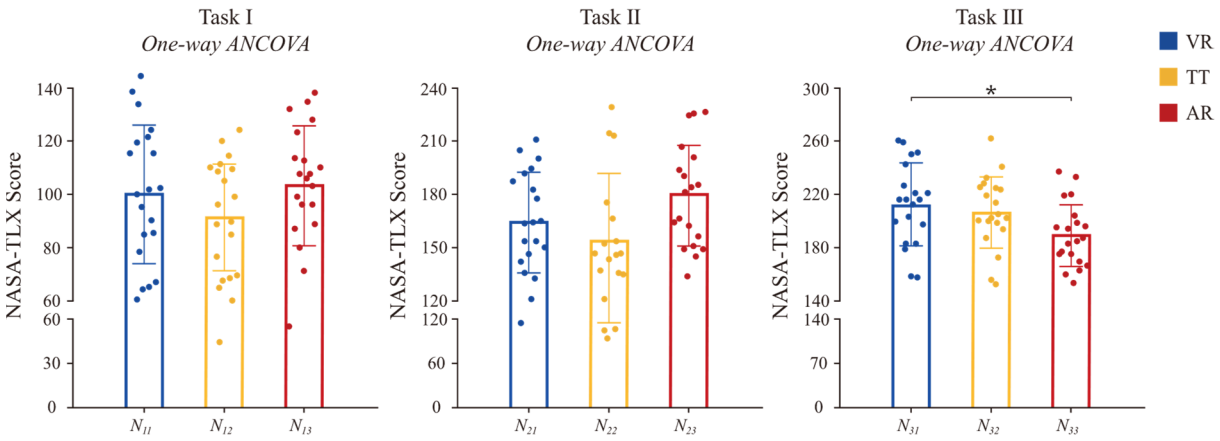


Figure 11. Results of between-group significance analysis of cognitive load scores. * means p value $\leq$ 0.05, if not indicated, means p value > 0.05.

4. Results and Discussion

In this study, the training effectiveness of VR, AR, and actual equipment training platforms for different crane maintenance tasks was compared and analyzed. Based on CPSI and CLT, the subjective and objective data of 60 trainees were measured, including time, errors, and cognitive load. From the experimental results, it provides a strong support for the research problems mentioned in the previous article.

4.1. General Conclusions from the Analysis of Experimental Results

From the experimental results of objective performance, for simple maintenance tasks, the TT group trained directly through the actual platform has a greater time advantage, especially compared with the task time of AR training. Being new to the training system or not adapting to it is one of the reasons, other reasons may be that AR complicates simple maintenance tasks. Redundant action prompts and information display in AR can cause trainees unable to find the key information, resulting in a lower effectiveness and CPSI score. Some trainees indicated that they were not well adapted to the interactive environment of AR at first and the fixed FOV limited the visual field to a certain extent, which led to the unnatural process of AR interaction. In particular, some trainees with glasses indicated that wearing this overlaid caused them to be uncomfortable. However, most of the trainees confirmed this training method and believed that AR can mobilize their interests, provide more information than learning video streams, and can interact with virtual devices. As the difficulty of the task gradually increased, individuals who did not complete the experimental task gradually appeared in the VR group, which was

reflected in the low CPSI score. The method of learning in the 3D environment using their fingers on the tablet was limited to a certain extent. Some participants indicated that this training work gradually became more difficult as the task steps gradually increased. The trainees in the TT group and AR group interacted with objects in the real environment and the operation process was much simpler than that of the tablet. In the experiment, this advantage was also reflected in the task time and CPSI scores. With the increase in the number of task steps and thread branches, the AR group participants achieved a significant advantage in CPSI scores, which was attributed to the multi-channel interaction, such as voice prompts and thread labeling. Compared with the video streaming of the TT group, AR was able to present more information and words to help the trainees to remember and understand the maintenance steps in the task, resulting in fewer errors and incomplete steps. This conclusion reached a consensus with the research work of Amanda [42], which indicated that AR training was especially suitable for understanding and understanding the causal relationship in complex tasks and helping the trainees to form psychomotor skills (such as logical thinking and step planning).

From the experimental results of the cognitive load, the objective performance showed a certain correlation with the cognitive load index. In Task 1 and Task 2, the higher cognitive load level of the VR group and AR group trainees caused their task time to be longer than the TT group trainees, although there was no significant difference. In Task 3, the significantly lower cognitive load level of trainees in AR group resulted in fewer errors and task time compared with the other two groups. This demonstrates the correlation among the cognitive load level, task time, and CPSI score.

Overall, AR has the effectiveness potential for maintenance training and can even bring better task performance in complex maintenance tasks, which benefits from its unique display and interaction mode and is beneficial to the trainees' knowledge acquisition, causing them more interest [43] and providing a better interaction experience [44]. However, it should be emphasized that not all tasks are suitable for AR implementation. More importantly, it is necessary to ensure the simplicity of learning new tasks, being error-free, and reducing the cognitive load [45]. The inherent advantages of AR training lie in the fact that it does not need the support of actual equipment, has a good economy, has a low degree of risk, is not subject to spatial restrictions, and has a strong reusability. These factors are sufficient to prove its value in the field of teaching and training. One of the purposes of this study is to verify the system effectiveness through comparative experiments, ensure that AR tools can enable trainees to complete the training migration from virtual to actual, and sum up the methods and strategies under different maintenance tasks. AR is a potential training tool that still needs further exploration and improvement, but evaluation is a complex process with many factors and instability.

4.2. Limitations

Defects in the display device remained the single most important factor affecting training performance, with the HoloLens 2 device having a limited FOV and being unable to deliver the immersive experience provided by the HTC Vive. Actually, immersion and presence are very important for the user experience, but the benefit of less presence is less simulation effect, which can better support the trainees in completing the training task for a long time.

With regard to the selection of trainees, no female trainees participated in this experiment due to the objective restrictions of the academy. Seventy-seven male trainees were recruited through questionnaires and sixty trainees finally took part in formal training. This experiment was implemented as a bonus item of the course, which can arouse the enthusiasm of trainees to some extent.

The evaluation model still needs further testing and more comparative and post-test experiments are needed to measure the difference between the telepresence and immersion under different training methods. System availability is also worth testing. However, due to the real-time nature of questionnaire effectiveness, it is too dependent on the trainees'

memory ability for the experimental process, and individuals are easily affected by the occurrence time at the end of the experiment. Therefore, filling in multiple scales in the post-test items might result in larger individual differences in the results. We accounted for this in the experiment and we will also measure the other subjective feelings of the trainees step by step in the next experiment.

4.3. Future Work

Based on Daling's definition of interactive devices [4], this experiment can be summarized as a comparison of the training effectiveness of actual equipment, AR HMD, and VR based on monoscopic screens to prove whether AR and VR based on monoscopic screens can achieve the same or even better training effectiveness as the actual equipment. Based on the opinion of Catal et al. [46], the training method of a virtual–real fusion that combines actual objects and AR may have better training effectiveness. This is part of the consideration of designing experiments and conducting comparative evaluations in the next step. However, in order to obtain the advantages of AR in multi-information, cue display and multi-channel prompting, economy, reusability, and space convenience are sacrificed. This means that a virtual–real fusion may not be an economical solution. In addition, we are committed to design more effective spatial cue tips to improve the visual stimulation to the trainees, so as to achieve better AR interaction effects, reduce the complexity and redundancy of information, and reduce the cognitive load of the trainees when performing complex training tasks.

The objective performance in this experiment refers to the predictive validity under the influence of working memory. For the research on long-term memory, according to Uttal's suggestion [47], training should have a lasting effect, and it is meaningful to delay the evaluation of training effectiveness. However, it is obvious that such experiments are not easy to carry out. We will consider designing another group of experiments to measure the durability of training in our future work, which will be a long process.

5. Conclusions

This study aims to provide a reference for choosing the best training mode among VR, AR, and traditional training under multi-level tasks. Based on three types of crane maintenance tasks with different levels of difficulty, an experiment involving 60 trainees was conducted on three training platforms of VR, AR, and actual equipment in this study. The experimental results show the differences of training effectiveness of the three training platforms under the influence of the task difficulty.

The conclusions are as follows. For single-level maintenance tasks, the traditional training based on actual equipment is the most suitable. Compared with AR, VR has more training effectiveness advantages, which is reflected in less task time, errors, and cognitive load. The reason is that VR based on monoscopic screens is more in line with the habit of daily interaction and the redundant information in AR indicates that it reduces the ability of trainees to capture key information. For multi-level maintenance tasks, AR and traditional training based on actual equipment have the training effectiveness advantage over VR, which is reflected in less errors and cognitive load. The reason is that VR training based on monoscopic screens is limited by information display, but AR can provide more useful tips and guidance to help trainees understand complex maintenance logic. In addition, there is still a lack of experimental cases considering individual differences in the evaluation of training platforms. Customizing the training scheme for each trainee based on individual characteristics and task characteristics is the next research direction. At present, VR and AR display devices are still limited in view angle, definition, wearing, and interaction experience. With the development of hardware technology and the optimization of the interaction mode, training based on VR and AR technology is expected to achieve better performance.

Author Contributions: Conceptualization, X.-W.L. and W.W.; methodology, X.-W.L.; software, C.-Y.L. and T.C.; validation, W.W., J.Q., and Q.-L.W.; formal analysis, X.-W.L.; investigation, Q.-L.W.; data analysis, Q.-L.W. and C.-Y.L.; writing—original draft preparation, C.-Y.L. and S.D.; writing—review and editing, W.W., S.D., and J.Q.; visualization, S.D. and Q.-L.W.; supervision, W.W. and J.Q.; project administration, J.Q.; funding acquisition, J.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 51405505.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks to Xi'an E-Fly Aviation Electric Technology Co., Ltd. (Xi'an, China) for hardware support.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Parameter Table in CPSI Model

Table A1. AR and VR technology affinity questionnaire.

	Options and Scores				
Items	**One Point**	**Two Points**	**Three Points**	**Four Points**	**Five Points**
Q1: How many times have you interacted with AR or VR?	0	1	2	3	>4
Q2: How many times have you studied mechanical maintenance systems?	0	1	2	3	>4
Q3: How many times have you experienced the head-mounted display device?	0	1	2	3	>4
Q4: How many times have you experienced AR or VR games?	0	1	2	3	>4
Q5: How well do you know AR or VR?	Not at all	Little	Not sure	Know something	Know it quite well

Table A2. Parameter setting in CPSI model.

Items	**Task I**	**Task II**	**Task III**
X	[1–3, 5]	[1–3, 5]	[1–3, 5]
apt	5	8	10

Appendix B. The Experimental Data Sheet That Needs to Be Explained

Table A3. Hypothesis test about task time T.

	Descriptive Analysis		**S-W Test**		**Levene Test**		**Slope Homogeneity Test**			
Items	**Average**	***SD***	**df**	**Sig.**	***F***	**Sig.**	**Ss(III)**	**Ms**	***F***	**Sig.**
T_{11}	35.689	3.196	20	0.678						
T_{12}	28.817	2.918	20	0.174	0.604	0.550	21.714	10.857	0.966	0.387
T_{13}	37.789	3.877	20	0.644						

Table A3. *Cont.*

	Descriptive Analysis		S-W Test		Levene Test		Slope Homogeneity Test			
Items	**Average**	***SD***	**df**	**Sig.**	***F***	**Sig.**	**Ss(III)**	**Ms**	***F***	**Sig.**
T_{21}	170.669	12.148	20	0.094						
T_{22}	155.570	9.820	20	0.158	1.760	0.181	54.677	27.338	0.257	0.774
T_{23}	163.833	9.972	20	0.078						
T_{31}	172.439	20.347	20	0.218						
T_{32}	175.486	27.491	20	0.989	0.819	0.446	701.595	350.797	0.704	0.499
T_{33}	156.697	20.862	20	0.082						

Table A4. Descriptive statistics of the CPSI scores of the three groups in Task 1 to 3.

	Task I		Task II		Task III	
Groups	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
VR	0.992	0.017	0.921	0.054	0.842	0.157
TT	0.995	0.011	0.963	0.024	0.808	0.168
AR	0.983	0.030	0.953	0.030	0.930	0.109

Table A5. Hypothesis test about cognitive load index *N*.

	Descriptive Analysis		S-W Test		Levene Test		Slope Homogeneity Test			
Items	**Average**	***SD***	**df**	**Sig.**	***F***	**Sig.**	**Ss(III)**	**Ms**	***F***	**Sig.**
N_{11}	100.50	25.673	20	0.417						
N_{12}	92.75	19.950	20	0.309	0.980	0.382	619.685	309.843	0.588	0.559
N_{13}	103.85	22.885	20	0.517						
N_{21}	164.80	28.382	20	0.664						
N_{22}	154.95	37.884	20	0.416	0.727	0.488	2.883	1.441	0.001	0.999
N_{23}	180.10	28.591	20	0.165						
N_{31}	212.85	30.505	20	0.506						
N_{32}	206.75	26.929	20	0.646	0.424	0.656	3242.488	1621.244	2.302	0.110
N_{33}	189.80	23.539	20	0.482						

References

1. Gao, Y.; Gonzalez, V.A.; Yiu, T.W. The effectiveness of traditional tools and computer-aided technologies for health and safety training in the construction sector: A systematic review. *Comput. Educ.* **2019**, *138*, 101–115. [CrossRef]
2. Daponte, P.; De Vito, L.; Picariello, F.; Riccio, M.J. State of the art and future developments of the Augmented Reality for measurement applications. *Measurement* **2014**, *57*, 53–70. [CrossRef]
3. Qin, Z.; Tai, Y.; Xia, C.; Peng, J.; Huang, X.; Chen, Z.; Li, Q.; Shi, J. Towards Virtual VATS, Face, and Construct Evaluation for Peg Transfer Training of Box, VR, AR, and MR Trainer. *J. Heal. Eng.* **2019**, *2019*, 6813719. [CrossRef] [PubMed]
4. Daling, L.M.; Schlittmeier, S.J. 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. *Hum. Factors* **2022**, 1–38. [CrossRef]
5. Mao, C.C.; Chen, C.H. Augmented Reality of 3D Content Application in Common Operational Picture Training System for Army. *Int. J. Hum. Comput. Interact.* **2021**, *37*, 1899–1915. [CrossRef]
6. Champney, R.; Lackey, S.J.; Stanney, K.; Quinn, S. Augmented Reality Training of Military Tasks: Reactions from Subject Matter Experts. In *Virtual, Augmented and Mixed Reality*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 251–262.
7. Abich, J.; Eudy, M.; Murphy, J.; Garneau, C.; Raby, Y.; Amburn, C. Use of the Augmented REality Sandtable (ARES) to Enhance Army CBRN Training. In Proceedings of the 20th International Conference on Human-Computer Interaction (HCI International), Las Vegas, NV, USA, 15–20 July 2018; pp. 223–230.
8. Schaffernak, H.; Moesl, B.; Vorraber, W.; Braunstingl, R.; Herrele, T.; Koglbauer, I. Design and Evaluation of an Augmented Reality Application for Landing Training. In *Human Interaction, Emerging Technologies and Future Applications IV—Proceedings of the 4th International Conference on Human Interaction and Emerging Technologies, Strasbourg, France, 28–30 April 2021*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 107–114.

9. Moesl, B.; Schaffernak, H.; Vorraber, W.; Holy, M.; Herrele, T.; Braunstingl, R.; Koglbauer, I.V. Towards a More Socially Sustainable Advanced Pilot Training by Integrating Wearable Augmented Reality Devices. *Sustainability* **2022**, *14*, 2220. [CrossRef]
10. Velosa, J.D.; Cobo, L.; Castillo, F.; Castillo, C. Methodological proposal for use of Virtual Reality VR and Augmented Reality AR in the formation of professional skills in industrial maintenance and industrial safety. In *Online Engineering & Internet of Things*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 987–1000.
11. Bosch, T.; Van Rhijn, G.; Krause, F.; Könemann, R.; Wilschut, E.S.; de Looze, M. Spatial augmented reality: A tool for operator guidance and training evaluated in five industrial case studies. In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June–3 July 2020; pp. 1–7.
12. Gavish, N.; Gutierrez, T.; Webel, S.; Rodriguez, J.; Peveri, M.; Bockholt, U.; Tecchia, F. Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interact. Learn. Environ.* **2015**, *23*, 778–798. [CrossRef]
13. Vergel, R.S.; Tena, P.M.; Yrurzum, S.C.; Cruz-Neira, C. A Comparative Evaluation of a Virtual Reality Table and a HoloLens-Based Augmented Reality System for Anatomy Training. *IEEE Trans. Hum. Mach. Syst.* **2020**, *50*, 337–348. [CrossRef]
14. Ferrer-Torregrosa, J.; Jiménez-Rodríguez, M.Á.; Torralba-Estelles, J.; Garzón-Farinós, F.; Pérez-Bermejo, M.; Fernández-Ehrling, N.J. Distance learning ects and flipped classroom in the anatomy learning: Comparative study of the use of augmented reality, video and notes. *BMC Med. Educ.* **2016**, *16*, 230. [CrossRef]
15. Koutitas, G.; Smith, S.; Lawrence, G.J. Performance evaluation of AR/VR training technologies for EMS first responders. *Virtual Real.* **2021**, *25*, 83–94. [CrossRef]
16. Papakostas, C.; Troussas, C.; Krouska, A.; Sgouropoulou, C. Measuring User Experience, Usability and Interactivity of a Personalized Mobile Augmented Reality Training System. *Sensors* **2021**, *21*, 3888. [CrossRef] [PubMed]
17. Langley, A.; Lawson, G.; Hermawati, S.; D'Cruz, M.; Apold, J.; Arlt, F.; Mura, K. Establishing the Usability of a Virtual Training System for Assembly Operations within the Automotive Industry. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2016**, *26*, 667–679. [CrossRef]
18. Chang, Y.S.; Hu, K.J.; Chiang, C.W.; Lugmayr, A. Applying Mobile Augmented Reality (AR) to Teach Interior Design Students in Layout Plans: Evaluation of Learning Effectiveness Based on the ARCS Model of Learning Motivation Theory. *Sensors* **2019**, *20*, 105. [CrossRef] [PubMed]
19. Gonzalez, A.V.; Koh, S.; Kapalo, K.; Sottilare, R.; Garrity, P.; Billinghurst, M.; LaViola, J. A comparison of desktop and augmented reality scenario based training authoring tools. In Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Beijing, China, 14–18 October 2019; pp. 339–350.
20. Moghaddam, M.; Wilson, N.C.; Modestino, A.S.; Jona, K.; Marsella, S.C. Exploring augmented reality for worker assistance versus training. *Adv. Eng. Inform.* **2021**, *50*, 101410. [CrossRef]
21. Vidal-Balea, A.; Blanco-Novoa, O.; Fraga-Lamas, P.; Vilar-Montesinos, M.; Fernández-Caramés, T.M. Collaborative Augmented Digital Twin: A Novel Open-Source Augmented Reality Solution for Training and Maintenance Processes in the Shipyard of the Future. *Eng. Proc.* **2021**, *7*, 10.
22. Henderson, S.J.; Feiner, S. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality, Orlando, FL, USA, 19–22 October 2009; pp. 135–144.
23. Henderson, S.; Feiner, S.J. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE Trans. Vis. Comput. Graph.* **2010**, *17*, 1355–1368. [CrossRef]
24. Siyaev, A.; Jo, G.S. Towards Aircraft Maintenance Metaverse Using Speech Interactions with Virtual Objects in Mixed Reality. *Sensors* **2021**, *21*, 2066. [CrossRef]
25. Wiederhold, B.K.; Bouchard, S. *Advances in Virtual Reality and Anxiety Disorders*; Springer: Berlin/Heidelberg, Germany, 2014.
26. Vovk, A.; Wild, F.; Guest, W.; Kuula, T. Simulator Sickness in Augmented Reality Training Using the Microsoft HoloLens. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–9.
27. Kaplan, A.D.; Cruit, J.; Endsley, M.; Beers, S.M.; Sawyer, B.D.; Hancock, P.A. The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis. *Hum. Factors* **2021**, *63*, 706–726. [CrossRef]
28. Borsci, S.; Lawson, G.; Broome, S. Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance. *Comput. Ind.* **2015**, *67*, 17–26. [CrossRef]
29. Daling, L.M.; Abdelrazeq, A.; Isenhardt, I. Abdelrazeq, A.; Isenhardt, I. A Comparison of Augmented and Virtual Reality Features in Industrial Trainings. In *Virtual, Augmented and Mixed Reality. Industrial and Everyday Life Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 47–65.
30. Keighrey, C.; Flynn, R.; Murray, S.; Murray, N. A Physiology-Based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-Based Applications. *IEEE Trans. Multimed.* **2021**, *23*, 333–341. [CrossRef]
31. Werrlich, S.; Daniel, A.; Ginger, A.; Nguyen, P.-A.; Notni, G. Comparing HMD-Based and Paper-Based Training. In Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 16–20 October 2018; pp. 134–142.
32. Young, K.-Y.; Cheng, S.-L.; Ko, C.-H.; Su, Y.-H.; Liu, Q.-F. A novel teaching and training system for industrial applications based on augmented reality. *J. Chin. Inst. Eng.* **2020**, *43*, 796–806. [CrossRef]

33. Münzer, S. Facilitating spatial perspective taking through animation: Evidence from an aptitude–treatment-interaction. *Learn. Individ. Differ.* **2012**, *22*, 505–510. [CrossRef]
34. Sweller, J. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* **1988**, *12*, 257–285. [CrossRef]
35. Sweller, J. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* **2010**, *22*, 123–138. [CrossRef]
36. Gasteiger, N.; van der Veer, S.N.; Wilson, P.; Dowding, D. How, for Whom, and in Which Contexts or Conditions Augmented and Virtual Reality Training Works in Upskilling Health Care Workers: Realist Synthesis. *JMIR Serious Games* **2022**, *10*, e31644. [CrossRef]
37. Tziner, A.; Fisher, M.; Senior, T.; Weisberg, J. Assessment, Effects of trainee characteristics on training effectiveness. *Int. J. Sel. Assess.* **2007**, *15*, 167–174. [CrossRef]
38. Baddeley, A. Working memory. *Science* **1992**, *255*, 556–559. [CrossRef] [PubMed]
39. Stanton, N.A. Hierarchical task analysis: Developments, applications, and extensions. *Appl. Erg.* **2006**, *37*, 55–79. [CrossRef]
40. Reason, J. *Human Error*; Cambridge University Press: Cambridge, UK, 1990.
41. Randeniya, N.; Ranjha, S.; Kulkarni, A.; Lu, G. Virtual reality based maintenance training effectiveness measures—A novel approach for rail industry. In Proceedings of the 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), Vancouver, BC, Canada, 12–14 June 2019; pp. 1605–1610.
42. Bond, A.; Neville, K.; Mercado, J.; Massey, L.; Wearne, A.; Ogreten, S. Evaluating Training Efficacy and Return on Investment for Augmented Reality: A Theoretical Framework. In *Advances in Human Factors in Training, Education, and Learning Sciences*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 226–236.
43. Aziz, F.A.; Alsaeed, A.S.; Sulaiman, S.; Mohd Ariffin, M.K.A.; Al-Hakim, M.F. Mixed Reality Improves Education and Training in Assembly Processes. *J. Eng. Technol. Sci.* **2020**, *52*, 598. [CrossRef]
44. Papakostas, C.; Troussas, C.; Krouska, A.; Sgouropoulou, C. User acceptance of augmented reality welding simulator in engineering training. *Educ. Inf. Technol.* **2021**, *27*, 791–817. [CrossRef]
45. Nee, A.Y.C.; Ong, S.K.; Chryssolouris, G.; Mourtzis, D. Augmented reality applications in design and manufacturing. *CIRP Ann.* **2012**, *61*, 657–679. [CrossRef]
46. Catal, C.; Akbulut, A.; Tunali, B.; Ulug, E.; Ozturk, E. Evaluation of augmented reality technology for the design of an evacuation training game. *Virtual Real.* **2019**, *24*, 359–368. [CrossRef]
47. Uttal, D.H.; Meadow, N.G.; Tipton, E.; Hand, L.L.; Alden, A.R.; Warren, C.; Newcombe, N.S. The malleability of spatial skills: A meta-analysis of training studies. *Psychol Bull.* **2013**, *139*, 352–402. [CrossRef] [PubMed]