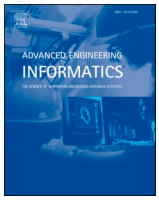
Advanced Engineering Informatics 64 (2025) 103081

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Full length article

Integrative human and object aware online progress observation for human-centric augmented reality assembly

Tienong Zhang, Yuqing Cui, Wei Fang *

School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing, PR China

ARTICLE INFO

Keywords:
Augmented reality
Assembly progress observation
Human-object interaction
Human-centric assembly

ABSTRACT

Augmented reality (AR) can provide step-by-step intuitive guidance for workers on the shop floor, enabling time-saving and error-avoid assembly actions. Nevertheless, existing AR-guided assembly methods have primarily paid attention to information on assembly objects and usually ignore the human factor in the assembly process. Further, there are a series of details regarding the AR system design that are frequently neglected, including systematic usability, human intervention, and AR perspective. To alleviate these limitations, this paper proposes a real-time two-branch approach that integrates human action-based human factor evaluation and object-based assembly progress observation. In the online human factor evaluation, a skeleton-based model is applied to predict the operator's assembly action, providing a quantitative analysis and optimized indicator for the ongoing AR assembly. In the assembly progress observation, the object-based model is deployed to recognize the assembly part, and the AR assembly status is checked automatically based on the prior sequential assembly knowledge without human intervention. Thus, a holistic human-object integrated framework is established for the human-centric AR assembly process inspection, as well as the quantitative analysis and optimized indicator output from the framework are actively feedback in the first-person AR perspective, where the operators can perceive the assembly stage and whether their working posture is appropriate or not intuitively. Finally, extensive experiments are carried out on the human-object integrated performance in the smart AR assembly, and results illustrate that the proposed method can monitor the online assembly observation from a holistic perspective, alleviate the cognitive load, and achieve superior performance for the AR assembly tasks.

1. Introduction

In the modern manufacturing industry, the growing number of product variants, complex structures, and shorter life cycles of products made assembly operations increasingly complicated, and it has been stated that the assembly process costs about 50 % of the developing time while nearly taking up about 20 % of the total manufacturing time [1]. Although assembly automation technology has made great achievements, some assembly operations for complex products, such as aerospace products, are still mainly accomplished by manual operations [2], and the cost and performance of manual assembly significantly affect final product quality. In current manual assembly tasks, operators often rely on paper-based instructions to activate the step-by-step operating tutorials, and the interpretations of the abstract 2D text or symbols are a time-consuming and boring task that imposes a great mental burden on shop floor operators [3]. Therefore, the ability to sense, monitor, and support assembly workers with intuitive and visual instructions against actual scenarios has become imperative. With the development of Industry 4.0, immersive technologies (e.g., augmented reality (AR) and virtual reality (VR)) and artificial intelligence (AI) have been applied by researchers in manual assembly assistance [4–6]. VR technology plays a vital role in training and educating operators to grasp complex product assembly procedures more rapidly [7], through simulating advanced 3D human–computer interaction, this technology allows the operator to be completely immersed in a synthetic scenario. However, there still exist limitations that hinder the development of VR in practical manual assembly. Specifically, in a VR environment, the physical connection between the operator and the real workbench will be removed, making the operator's realistic experience lost, and since it is almost impossible to perfectly and accurately simulate the actual work scenario, the VR experience can't be fully convincing.

In contrast, AR as a novel human–machine interface, can lead to an immersive knowledge delivery environment by superimposing virtual guidance on real workbenches, and it can be introduced to see what

* Corresponding author.
E-mail address: fangwei@bupt.edu.cn (W. Fang).

https://doi.org/10.1016/j.aei.2024.103081
Received 21 July 2024; Received in revised form 4 November 2024; Accepted 22 December 2024
Available online 29 December 2024

cannot be normally perceived in actual scenarios. For this reason, compared to traditional and theoretical model-based virtual assembly methods, our proposed AR-based assembly method is more suitable to assist operators in executing real-time assembly tasks on the shop floor, achieving improved performance in assembly quality and efficiency. By utilizing AR-based assistance, the accessibility and usability of working instructions in complex shop floor environments are improved [8]. It is also demonstrated that AR can ensure the worker's attention on the ongoing task rather than switching between the workpiece and the paper-based instruction back and forth constantly, allowing for easier reference and reducing the risk of error, as well as leading to lower workers' mental load and time-saving performance [9]. Additionally, AI as an emerging and revolutionary technology, provides a promising solution to increase the adaptability and versatility of AR systems. Park et al. [13] proposed a smart and user-centric AR assistance, which combines deep learning-based object detection and instance segmentation for more effective visual guidance and less cognitive load, the operator can identify and understand physical objects on the shop floor. Liu et al. [14] proposed InstruMentAR, a system that automatically generates AR tutorials by recording user demonstrations, while voice recognition and background capture are employed to automate the text and images as AR content. Instead of object-based recognition in AR assembly, more and more attention has been paid to the comfort and ergonomics of operators [15], and skeleton-based human action recognition approaches in assembly tasks are often used as they tend to generalize better to different people and environments [16].

Despite the advantages of AR and AI technologies, most current research on AR-assisted manual assembly primarily focuses on the technical community in AR view, registration performance [11], natural interaction [12], and object recognition in AR view [3]. There has been limited attention given to the analysis of human posture and human factors in the assembly process, where the human factor may be one of the key factors currently preventing AR from laboratory to field applications. An ideal AR-assisted assembly system should not only minimize operational errors, but also alleviate the potential cognitive workload on the operator. Therefore, the main research question of this paper is how to achieve a holistic progress observation by integrating the human factor and shop floor awareness simultaneously, which is more following the human-centric AR assembly in complex manufacturing environments. What is more, several concerns are not fully addressed in the prevailing works related to the design of AR-based assembly guidance systems. First concern: It has been found that the increase in performance by AR also depends on the complexity and nature of the task, as well as the experience level of the operator [10]. Generally, training workers without previous AR experience to fully familiarize the AR assembly procedures requires substantial resources, thus the AR-assisted assembly system should be easy-to-use for novice operators. Second concern: Currently, the step-by-step visual tutorials in most AR assembly systems are triggered by human intervention, so the operators must associate the corresponding instruction against the ongoing assembly status. However, it is time-consuming and has to be repeated for every new set of instructions, ignoring the confirmation of the actual assembly results that may be errors on the shop floor, and the observation of the assembly process mainly relies on the visual inspection of workers. Third concern: In the existing AR works, the key indicators are feedback from the third-person perspective. However, such a perspective fails to feedback the indicators intuitively and directly, requiring operators to alternate transferring vision during the manual assembly, which may not only distract their attention and cause increased mental load, but also result in hand-to-eye coordination efforts.

Therefore, to establish a better collaborative understanding between humans and the ongoing AR assembly process, we proposed an integrative human and object aware online progress observation for human-centric assembly task, and the novelty of this work is focused on establishing a holistic framework and active feedback for progress observation in human-in-the-loop AR assembly action, where the operators can perceive the current assembly status and whether their working posture is appropriate or not during the assembly progress from the first-person AR view. The main contributions of our paper are as follows:

(1) Based on the assembly part recognition and the skeleton model of the human action from an RGBD camera, an integrative context-aware assembly progress observation and validation for ongoing AR assembly procedure are achieved.

(2) A real-time ergonomics evaluation derived from human action recognition is established, providing a quantitative analysis and optimized indicator for the ongoing AR assembly in the first-person perspective, as well as checking the AR assembly procedures without human intervention.

(3) A holistic framework and system for progress observation in human-centric AR assembly is established by the integration of skeleton-based ergonomics awareness and assembly part recognition.

The rest of the paper is arranged as follows: Section 2 introduces the related works. Section 3 presents the systematic description of the proposed method, followed by the integrative human-object-aware AR assembly in Section 4. Extensive experiments are carried out in Section 5. Finally, the discussion and future works are introduced in Section 6.

2. Related works

2.1. Intelligent AR assembly

Rapidly emerging intelligent assembly technology is gradually replacing the original "semi-automatic" assembly technology, aiming to further enhance production efficiency and assembly quality. AR as a novel immersive technology has been deployed in assembly systems by a growing number of researchers to overcome a wide range of challenges throughout the assembly phase, where AR technology provides intuitive visual guidance in a realistic working environment to support the operator in executing the assembly task in real-time, enabling improved efficiency and quality compared to the traditional virtual assembly. It is also stated by Fang et al. [17] that, the majority of current industrial AR deployments are focused on assembly tasks, the reason may be that the assembly action involves lots of adjacent procedures that look similar, and the worker should exert great efforts to ensure the correctness of the step-by-step operations from abstract 2D text or symbols, while the AR-based visual instruction allow operators to perform the ongoing tasks intuitively.

Cardoso et al. [18] proposed a hand-held mobile AR system to satisfy the requirements of the structural assemblies in aeronautical industries, and also a field evaluation is conducted. The results indicated that AR offers a faster solution for highly complex assemblies despite several limitations regarding positioning tolerance. Eswaran et al. [19] proposed an AR-guided autonomous system to validate the input assembly sequences, followed by the optimization and generation of virtual content for AR instruction based on VR simulation, which can assist novice operators in performing complex assembly tasks with less cognitive load. Bahubalendruni et al. [20] presented an assembly sequence validation framework, taking textual assembly plans as input, and the generated virtual assembly task plans are utilized for the geometrical feasibility testing, finally the validated textual instructions will be automatically produced and visualized in the AR platform, leading to a huge time-saving in human confirmation. Simoes et al. [21] claimed that AR technology can enable an immersive working environment for the creation of intuitive assembly actions, supporting workers in the learning process of assembly tasks one at a time. Nevertheless, some researchers also state contradictory conclusions, Drouot et al. [22] show that AR does not provide obvious benefits in terms of assembly performance effectiveness in some specific tasks. Additionally, it has also been found that the increase in performance through AR depends on the complexity of the assembly task, and AR technology is particularly useful when operators are unfamiliar with the task at hand [23]. The

reason is mainly due to human-related factors, such as the interaction manner in AR assembly is not accepted as the paper-based one, as well as the easy-to-deployment needs to be further improved on the shop floor.

Additionally, instead of being technology-focused in the past, Industry 5.0 pays more attention to human-centricity toward the practical deployment of available technologies in industrial activities [24,25]. For the improvement of adaptability of AR assembly against varying tasks and shop floor operators, smart AR assembly systems that contain AI-based semantic information can improve the intelligence level of existing tasks, alleviating the operators' mental load further and enabling a human-centric AR assembly application. Lai et al. [26] introduced a worker-centered assembly system consisting of AR instructions with the support of a deep learning network for tool detection, where the integrated AR is designed to provide on-site visual instructions including various visual renderings with a fine-tuned Region-based Convolutional Neural Network (RCNN), and results showed that the proposed customizable smart AR system can help reduce the time and errors of the given assembly tasks by 33.2 % and 32.4 %, respectively. To solve the problem of occlusion handling for AR assistance assembly systems, Li et al. [27] proposed a monocular image-based real-time occlusion handling method for virtual parts and physical products in AR assembly tasks, and the depth relationship between real assembly scenes, assembly objects, and virtual augmented contents is determined for an intuitive AR view. Fu et al. [3] proposed a context-aware AR Assembly system to perform worker-centered manual assembly action, where context understanding of the current assembly status is achieved by the scene recognition algorithm, and thus the augmented guidance for manual operation can be activated accordingly. Zhao et al. [28] proposed a redundant object detection method based on computer vision and AR glasses, where the Feature Pyramid Networks-CenterNet is combined with multi-scale feature fusion to improve the detection accuracy of small-scale redundant targets, providing a new reference for the quality of the civil aircraft assembly process. Fang et al. [29] proposed a multi-modal context-aware on-site assembly stage recognition for AR assembly, which combined the sim-real point cloud-based semantic understanding and 2D image-based recognition for assembly parts, resulting in a closed-loop AR assembly system with assembly result confirmation automatically.

The studies aforementioned are mainly concerned with the technique and performance of AR assembly, nevertheless, it is obvious, that human-related factors often fall by the wayside, even if they are crucial factors that impact user acceptance and performance during working conditions [4,5]. Additionally, technological changes such as the use of AR are on the one hand positively changing the way of working but on the other hand, they are introducing new risks, potentially leading to not only normal but also post-normal accidents [30], and risk assessment of AR-equipped manufacturing systems is also meaningful. Based on the fact that assembly tasks for complex products mostly involve a huge collaboration between humans and machines, a basic necessity for AR assembly is the acceptance of AR technology by the workers in their daily work [31]. To alleviate this phenomenon, Generosi et al. [32] proposed a novel platform for a human-centric factory, based on objective, automatic, and real-time MSDs (musculoskeletal disorders) risk assessment, it can provide more effective health and safety awareness systems are needed for the ongoing assembly operations. However, this platform ignores the interaction information between operators and the workpiece. In summary, although aforementioned deterministic studies have made great improvements in object-based recognition for AR assembly, the consideration of human-factor-related shop floor AR deployment is still in its infancy, and the integrative human action and scene awareness during the AR assembly process deserves strong attention to share a holistic understanding of the ongoing progress, providing the workers with the information they are involved in for a human-centric AR assembly operation. Motivated by these limitations, we presented a human-object integrated method for AR-guided assembly, which simultaneously considers information about the contacted objects and human factors using human action recognition and object assembly progress observation, respectively.

2.2. Assembly process observation

Assembly usually consists of a series of consecutive operations, and workers are required to confirm their accomplished results step-by-step by intermittent passive observation, as well as to activate the next work instruction for AR view by human intervention, where skill and cognitive load are usually of must for shop floor operators. Based on the prior assembly procedural information, Chang et al. [33] presented a proof-of-concept interactive AR disassembly sequence planning method, where the text instructions and animated 3D models appeared accordingly based on a taxonomy link that has been established in advance. Yin et al. [34] proposed an assembly behavior awareness method based on gesture recognition, and the assembly process observation is also included with key-feature matching in the AR assembly task. Taking into consideration the temporal and spatial restrictions of assembly procedures, Hong et al. [35] proposed a marker-less assembly stage recognition method based on the 3D corner feature between an assembling product and its corresponding digital model, thus the assembly stage can be observed according to the similarity between these two parts. Nevertheless, these methods are mainly dependent on the structural and texture features of components on the workbench, and it is difficult to distinguish the status accurately based on these visual features from two adjacent assembly stages.

To alleviate the difficulties of distinguishing current assembly statuses, Marino et al. [36] proposed an AR tool to observe operators in the detection of production and assembly errors, where workers can detect the presence of design discrepancies in the final physically assembled product and report them by adding 3D annotations directly on virtual models. Based on the consecutive state before AR manual tasks, Stanescu et al. [37] enhanced a state-of-the-art object detector algorithm to observe the current assembly state by conditioning on the previous one, improving the performance for state-aware AR assembly configuration detection. Zhao et al. [28] proposed an object detection-based assembly process observation method with AR Glass, where the Feature Pyramid Networks-CenterNet is integrated into the multi-scale feature fusion for small-scale redundant targets, providing a new observation for the civil aircraft assembly process. Li et al. [6] proposed a novel encoding and decoding convolutional neural network (CNN) to realize accurate AR registration under a marker-less assembly environment, and the system can observe the AR assembly process and present AR work instruction guidance accordingly. We can find that the studies on assembly process awareness mainly rely on the detection of the workpiece configuration from 2D or 3D scene observation [38], nevertheless, poor image data, small object parts, and heavy occlusions during operator interaction often make it difficult to detect the ongoing assembly status at run-time.

In addition to the detection of the varying assembly parts, some research has begun to focus on assembly process observation by perceiving the user's physical interactions with the workpiece. Based on a neural network module to understand a user's hand movement in real-time, Kaimel et al. [39] made use of hand gesture classification to estimate the progression to the next instruction, thus the process observation in the AR assembly tutorial by dynamic hand gesture recognition is achieved. It also points out that if similar gestures are needed at a certain point in different assembly steps, the proposed method may not be able to distinguish them accurately. To enable a more reliable assembly process confirmation, Chidambaram et al. [40] proposed ProcessAR, it can locate and identify different tools/objects within the workspace when the author interacts against them by bare-hand gesture, thus the in-situ procedural AR instructions are generated automatically. By incorporating object and action detectors with computer vision, Kastner et al. [41] presented an AR-based human assistance system to assist workers in complex manual tasks, resulting a more intuitive and flexible workflows and a significant reduction in time to task completion. Based

on the repeatability and tool dependence of the assembly action, Chen et al. [42] employed deep learning methods to recognize repetitive assembly actions and estimate their operating times, it can monitor workers' operations and prevent assembly quality problems caused by the lack of key operational steps. Zhang et al. [43] proposed a human-object integrated approach for context-aware assembly intention recognition in human-robot collaborative assembly, which integrates the recognition of assembly actions and assembly parts to improve the accuracy of the operator's intention recognition. Motivated by these existing researches, in order to bridge the gap of integrating online ergonomic status evaluation with real-time progress observation in AR-assisted assembly, this study will try to provide an easy-to-deployment and human-factor aware assembly monitoring from a holistic perspective regarding operators and assembly parts, as well as confirm the ongoing AR procedural tasks friendly. Furtherly, our AR-guided system is designed by considering systematic usability, human invention, and AR perspective, which are easily neglected details, trying to provide a user-friendly solution.

3. Systematic descriptions of the proposed method

The systematic framework of the proposed human-object integrated progress observation system for human-centric AR assembly is shown in Fig. 1, which primarily incorporates an RGBD camera, AR device, and assembly target. Before performing the cognitive AR assembly tasks, some source files are prepared in advance, such as the assembly procedural files and 3D virtual components. Firstly, the RGBD camera serves as the primary context-aware sensor, collecting the on-site assembly process data (RGB and depth frames) that covers both the operator and the object, which is also the data source for the human-object integrated progress observation framework. Following this, on the one hand, the operator's assembly action is recognized and represented by a skeleton-based model, followed by the operator's specific joint prediction, human action-based RULA score qualification, and finally the current ergonomic hazard is determined to enable the online human factor evaluation. On the other hand, an improved assembly part recognition method is applied in parallel to monitor the ongoing assembly status, and validate the assembly stages based on the previous sequential assembly knowledge, resulting in a more robust operating progress observation.

Thus, an integrative human action and object recognition for real-time progress observation in human-centric AR assembly is achieved, which leverages the spatial and temporal information of assembly manipulations and human interactions with them. Finally, the quantitative analysis and optimized indicator will be actively transmitted to the AR device in real-time based on the cloud server and provided to the operator in the first-person perspective, where the operators can realize the current assembly stage and whether their working posture is appropriate or not. It is noted that the intuitive AR instruction is rendered through a hand-held device rather than the fatigue-prone HMD AR glass, as well as the AR device is adjusted in a suitable position to ensure that the operator only requires a slight up-and-down motion of the neck for a full view of the assembly components and AR guidance. Instead of traditional AR-only assembly, the proposed method establishes an integrative real-time monitoring and management of operator ergonomic risks and the assembly component status for successive AR assembly processes, alleviating the operators' physical and mental load during the manual assembly until the accomplishment of the scheduled tasks.

4. Integrative human action and object recognition for AR assembly

The integration of human action and object recognition not only augments the precision and efficiency of AR assembly tasks but also fosters a more human-centric approach by considering both the actions of operators and the status of assembly components. The following sections will detail the methods and technologies utilized to achieve this integration, with a focus on human action awareness, ergonomic status perception in procedural AR assembly, object-based assembly process inspection, and integrative human action-aware and object-aware progress observation.

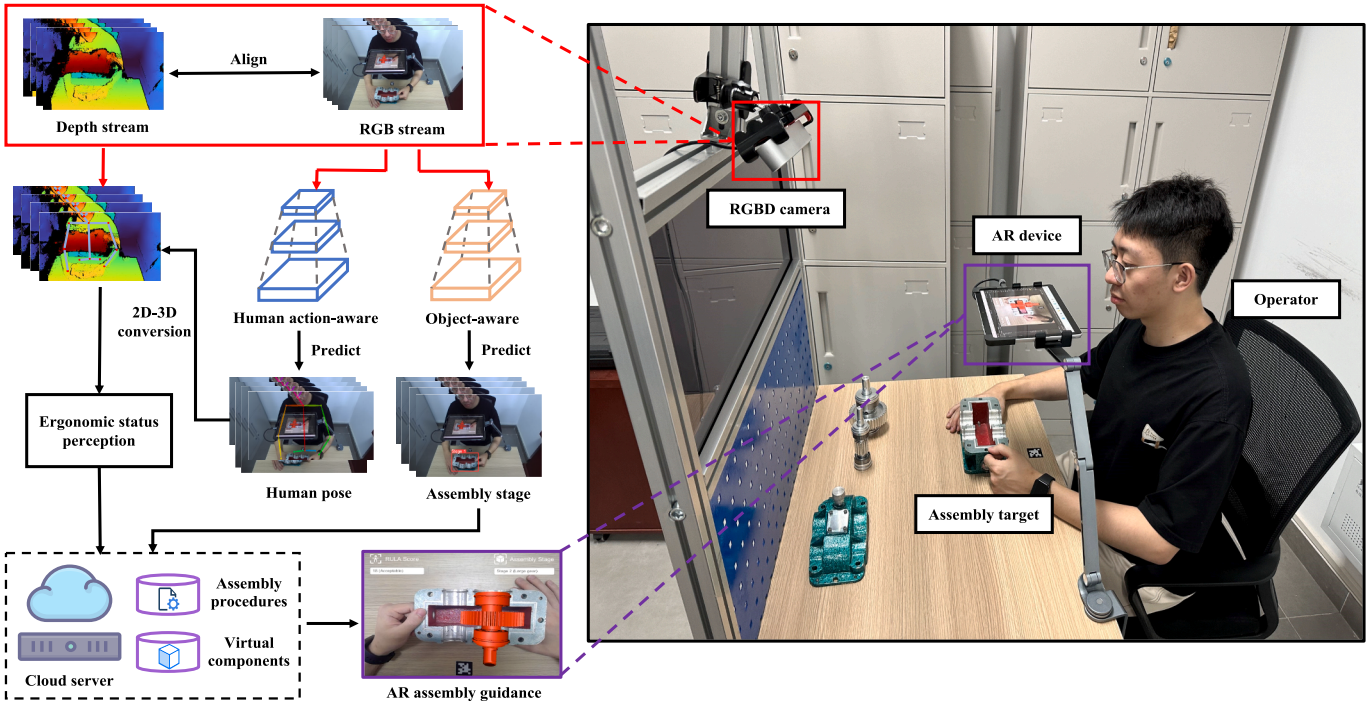


Fig. 1. The flowchart of the human-object integrated progress observation for human-centric AR assembly.

4.1. Human action-driven ergonomic-aware procedural AR assembly

The incorporation of human action recognition into AR assembly workplaces can enhance the AR system to be more sensitive to human-related factors in the operation process, leading to intelligent AR assembly with human-in-the-loop. In recent years, more and more human action recognition algorithms have been developed and packaged into libraries to provide ease-of-use. Therefore, the performance of these libraries is important when trying to integrate them into real-world applications for human-centric AR assembly. To choose the model that is better suitable in our case, three state-of-the-art skeleton-based human action recognition libraries are analyzed for their advantages and weaknesses, including OpenPose [44], BlazePose [45], and DeepLabCut [46]. According to the comprehensive survey, we found OpenPose to be more accurate and user-friendly for pose estimation compared to other models, which is recommended by many researchers to be deployed in future applications [47,48]. However, OpenPose has the shortcoming that processing speed is slower than BlazePose [49]. Considering that OpenPose not only satisfies the real-time requirements of our AR worksite (Over 20 FPS required, OpenPose can achieve 30 FPS) but also achieves superior performance in key points detection, where key points play a crucial role in the subsequent online human factors assessment, thus OpenPose is chosen as the foundational structure for the operator assembly action recognition network.

The architecture of the human action-aware network is illustrated in Fig. 2, which consists of a two-branch multi-stage CNN pipeline. The part confidence maps indicating the probability of key points' locations are predicted by the first branch, while the part affinity fields representing the spatial relationships between key points are predicted by the second branch. Moreover, both branches are refined in multiple stages, with the accuracy of the prediction improved in each stage based on the previous stage. It is indicated that the capacity of the human action-aware network to process complex poses and overlapping individuals can be enhanced by this multi-stage design, and making it suitable for precise and rapid human pose estimation. It is notable that although the original human motion features generated by the network are the positions of 18 body joints, only 12 of the operator's upper limb joint features are concerned, and the remaining 6 lower limb joint features are excluded from the data as they are not necessarily useful for human factors assessment, resulting in a more accurate prediction of the operator's ergonomic status as well as improved efficiency.

During assembly, RGB streams collected from the RGBD camera are processed by the human action-aware network to recognize and track the body key joints (2D human action representation), which are then converted into 3D representation with synchronized depth streams from the RGBD camera to enable the precise capture of operator postures and actions in 3D space. Following this, the joint angles are deduced by computing vectors from the differences in 3D joint coordinates (3D human action representation) and then using the dot product and the magnitudes of these vectors, as shown in equations (1) and (2), respectively.

$$\vec{v} = p_2 - p_1 = (x_2 - x_1, y_2 - y_1, z_2 - z_1) \tag{1}$$

Where $p_1 = (x_1, y_1, z_1)$ and $p_2 = (x_2, y_2, z_2)$ are the 3D coordinates of two connected joints.

$$\theta = \cos^{-1}\left(\frac{\vec{v}_1 \cdot \vec{v}_2}{\left|\vec{v}_1\right|\left|\vec{v}_2\right|}\right) \tag{2}$$

Where $\vec{v}_1 = p_2 - p_1$ and $\vec{v}_2 = p_2 - p_3$ are the 3D vectors of two connected joints. The flowchart for calculation of specific joint angles is depicted in Fig. 3.

Furtherly, since ergonomic-aware plays a crucial role in preventing work-related MSDs (musculoskeletal disorders) and ensuring the long-term well-being of operators, the RULA [50] (Rapid Upper Limb Assessment) criterion is adopted as a theoretical guideline for the perception of the operator's ergonomic status. The RULA scores are calculated in real-time during the manual assembly based on the deduced joint angles to quantify the risk of MSDs from each human action, which keeps the operator's assembly actions continuously supervised, ensuring that manual assembly is performed with a low physical workload and ergonomically acceptable action. It is noted that the standard RULA criterion is revised slightly to meet the requirements of the AR assembly scenario, and the details of the revision are as follows:

Firstly, a total of the operator's five specific joints are considered in the standard RULA criterion, but in this study, the wrist joint is excluded. The reason is that a large number of operators' assembly actions are recorded and further analyzed, there is an observation that the operator's wrist joint is almost constantly in a fixed and safe posture during the predefined assembly task. In addition, we also find that the majority of operators are used to leaning their trunk when they feel fatigued following a long-term manual assembly. On the contrary, the remaining three joints have a relatively limited range of action, thus the trunk joint is given a higher weight than the other three joints in the quantification of the overall RULA scores. In our case, the two revisions can be effective in enhancing the sensitivity of the ergonomic-aware method to the high-workload assembly actions, and then determine whether the operators' postures require to be corrected timely according to their ergonomic risk status.

The Diagram of the revised RULA criterion with the score from 1 to 3 for each specific joint is presented in Fig. 4, and the located upper arm joint is taken as an example, its RULA score is determined by the angle of shoulder flexion or extension, with adjustments made for shoulder elevation and/or abduction. Shoulder flexion refers to the forward movement of the upper arm in the sagittal plane, while shoulder extension refers to the backward movement in the same plane. Adjustments to the score can be made based on shoulder elevation and

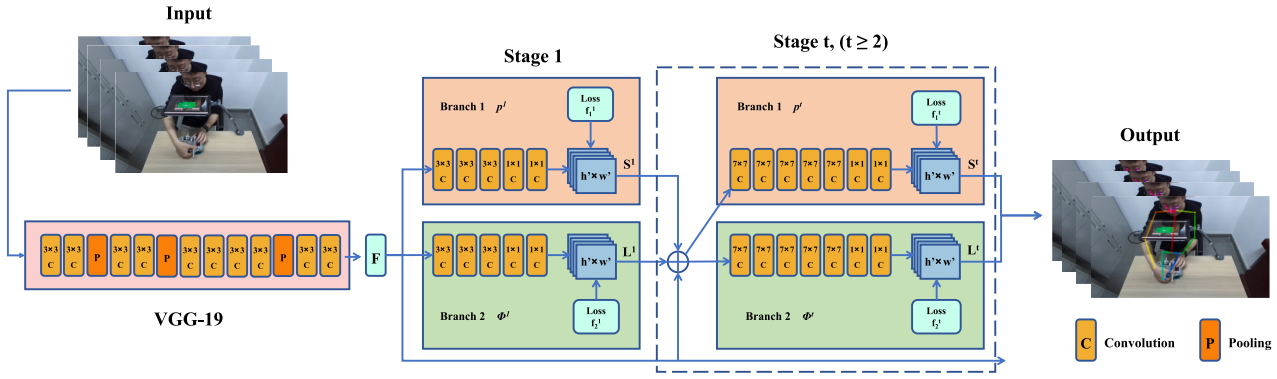


Fig. 2. The architecture diagram of the human action-aware network.

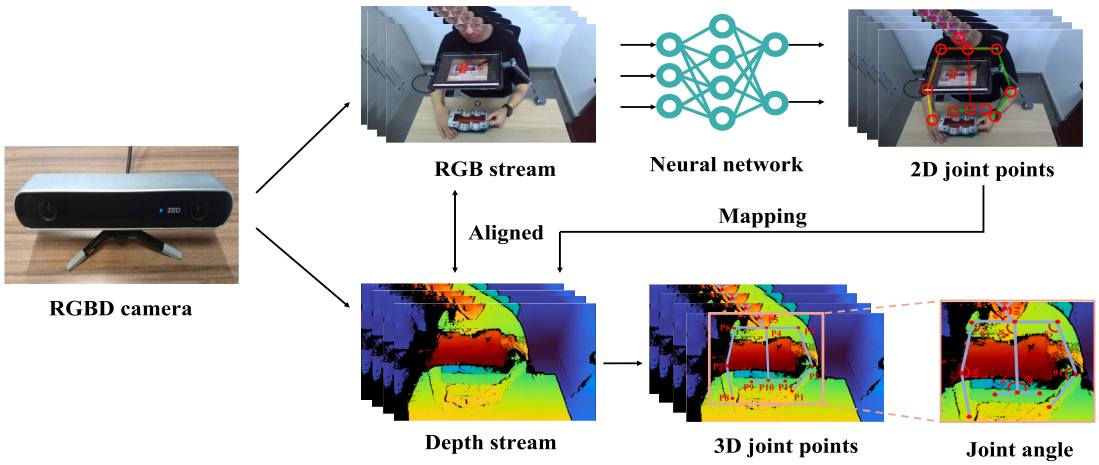


Fig. 3. Flowchart of calculation of specific joint angles.

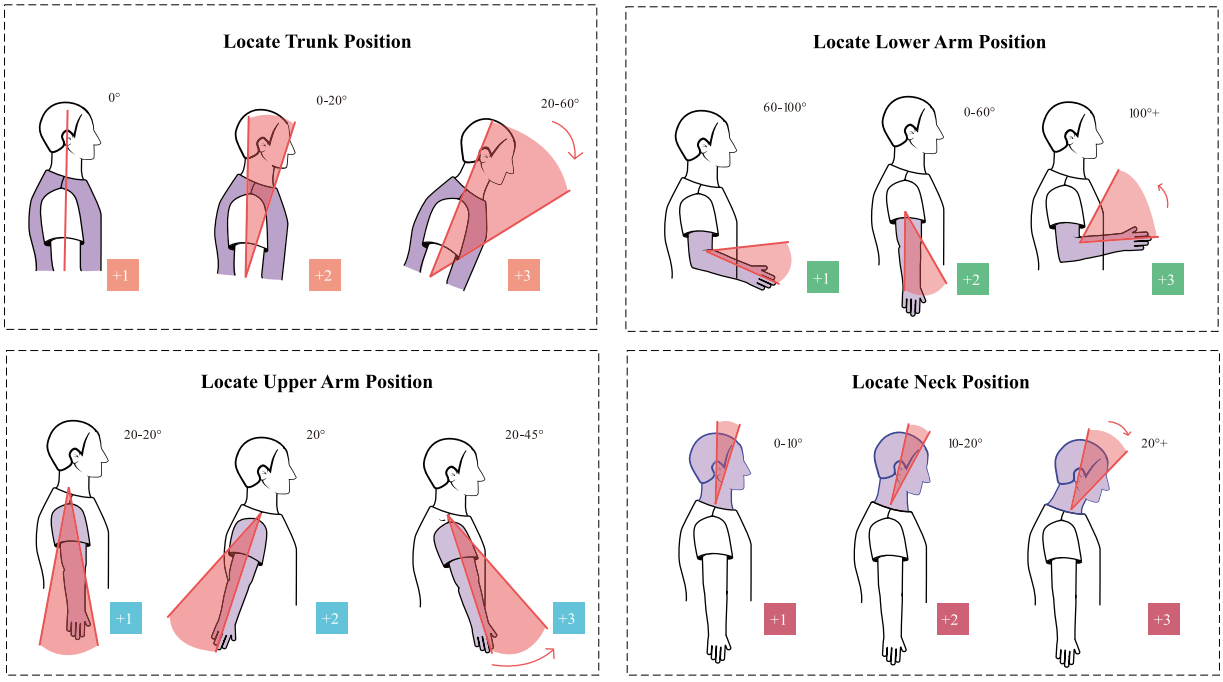


Fig. 4. The revised RULA criterion for the manual AR assembly action.

abduction. Additionally, a scoring adjustment of −1 can be applied if the worker's shoulder is supported or if the worker is leaning forward in a manner that allows gravity to assist the shoulder position.

However, in real applications, we encountered the challenge of abrupt changes in ergonomic perception data. For instance, if an operator's assembly action is extremely drastic and rapid, the ergonomic risk warning may be activated erroneously simultaneously, and then quickly return to normal. To overcome this challenge, the Savitzky-Golay filter [51] is applied to smooth the 3D human action representation. The Savitzky-Golay digital filter is particularly effective in preserving the higher moments of the data, such as peak shapes and widths, while eliminating noise and abrupt changes. By fitting a polynomial to a moving window of data points, the filter ensures that the essential trends and features of the data are maintained. It is observed that the utilization of the Savitzky-Golay filter can effectively solve the issue of abrupt changes in 3D human action data, thus enhancing the reliability of ergonomic risk status perception during operator assembly actions. The optimization process of data is presented in the equation (3).

$$(x_i', y_i', z_i') = \left(\sum_{j=-n}^{n} c_j x_{i+j}, \sum_{j=-n}^{n} c_j y_{i+j}, \sum_{j=-n}^{n} c_j z_{i+j}\right) \tag{3}$$

Where (x_i', y_i', z_i') represents the smoothed values of the respective coordinate components within the filtering window; c_j denotes the convolution coefficients; (x_i, y_i, z_i) illustrates the original data.

In the AR assembly workplace, RGB streams are processed by the human action-aware network in 30 FPS, and the 2D representation of the operator action output from the network is converted to the 3D representation based on aligned depth streams, which is followed by a series of pre-processing such as filter smoothing and specific joint angle deduction, finally quantify the ergonomic status according to the revised RULA, and then provided to the operator. It should be noted that the ergonomic risk warnings are displayed in the graphical user interface (GUI) of the AR assembly so that the operator can both perform manual assembly based on immersive AR instructions and rectify their assembly actions timely with text-based warnings.

By integrating ergonomic awareness into procedural AR assembly,

the real-time management of the operators' ergonomic risk status during assembly actions is achieved, resulting in the safeguarding of operators' health and well-being by pinpointing actions that may lead to muscle fatigue or discomfort. Early identification of ergonomic risks allows operators to implement targeted interventions, thus fostering a safer and healthier work environment. Furthermore, operator posture can also be tracked and analyzed over time by the ergonomic-aware system to detect potential musculoskeletal issues, followed by the assembly procedures that will be refined based on this longitudinal observation.

4.2. Object-based progress inspection for AR assembly

During the actual AR assembly execution, in addition to the operator's status, the assembly component's status is also the focus of our attention, which plays a crucial role in determining whether the assembly task is performed correctly or not. Most current AR-assisted assembly focuses mainly on virtual guidance superimposed on the real assembly site and assumes that the operator follows the visual instructions step-by-step to complete all the procedures correctly. Nevertheless, due to the lack of confirmation of the assembly process, it is difficult or even impossible to avoid operating errors on the shop floor for AR-only assembly. To this end, one of the aims of this study is to establish object-based assembly progress inspection to validate the assembly status of components in real-time until all the procedures are accomplished by the operator.

Fortunately, RGB video streams collected by the RGBD camera installed in the AR assembly scenario not only contain comprehensive information about the operator's status but also the assembly components, thus the RGB streams can also be utilized as the input source to the object-aware network for recognizing the current assembly stage. Further, regarding the choice of the object-based model applied in our case, a similar criterion is followed as the selection of human action recognition libraries, which pay more attention to accuracy while meeting the basic real-time requirements. To this end, several well-established models are compared in terms of accuracy and process speed, including the YOLO family [52], Faster RCNN [53], and SSD [54]. It is evident that the YOLO has the best performance with a well-balanced between object detection accuracy and computational efficiency [55]. In contrast, the accuracy of Faster-RCNN is relatively high, but hard to process images in real-time, and the SSD is the fastest, but its lightweight network architecture results in inferior accuracy. In the YOLO family, the YOLOv5 and YOLOv8 are the most popular versions due to their stability and usability, as well as the YOLOv8 is proven to have higher mAP in object detection compared to the YOLOv5 [56], thus the YOLOv8 is chosen as the basic framework for object-based inspection against the manual assembly process, enabling an object-aware assembly sequence stage check.

The architecture of the object-aware network is shown in Fig. 5, the modified CSPDarknet53[57] functions as the backbone network (Fig. 5 (a)), and the input features are down-sampled five times to obtain five different scale features. The CBS module (Fig. 5(d)) performs a convolution operation on the input information, followed by batch normalization, and finally activates the information stream using SiLU to obtain the output result. The C2f module (Fig. 5(e)) adopts a gradient shunt connection to enrich the information flow of the feature extraction network while maintaining a lightweight. Through the sequential application of multiple CBS and C2f modules, the object-aware backbone network progressively extracts multi-level features to enhance the model's understanding of the semantic information of the input RGB streams, thus improving both the performance and efficiency of

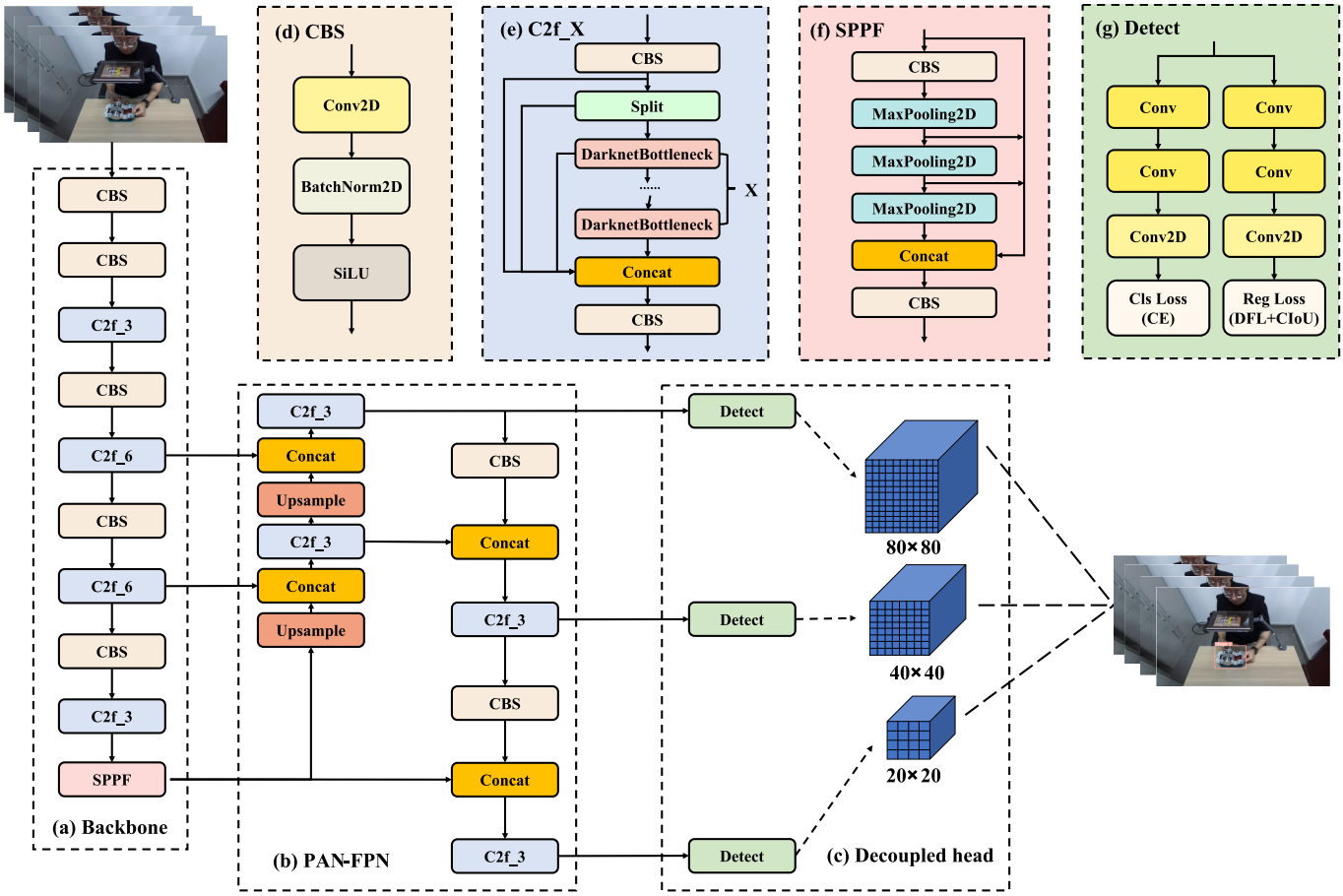


Fig. 5. Object-aware assembly progress inspection network architecture.

assembly progress inspection. In addition, the SPPF (spatial pyramid pooling fast) module (Fig. 5(f)) is utilized at the end of the backbone network to pool the input feature maps to a fixed-size map for adaptive size output.

Inspired by PANet[58], a PAN-FPN (Fig. 5(b)) is applied to the output of the backbone network. The FPN (Feature Pyramid Network) conveys the high-level semantic information to the low-level feature maps from the top-down and enhances the semantic information by feature fusion, but some object localization information will be lost. To overcome this challenge, the PAN is applied to transfer the low-level texture features to the high-level feature maps in a down-top form to realize path enhancement. The PAN-FPN constructs a bidirectional network structure, which realizes the complementarity of shallow positional information and deep semantic information through feature fusion, resulting in feature diversity and completeness. Finally, adaptive detection of assembly parts with different statuses is performed by a decoupled head structure (Fig. 5(c)), which consists of two separate branches (Fig. 5(g)) for assembly status classification and predicted bounding box regression of assembly components, thus different loss functions are used for these two types of tasks. For the classification task of assembly status, CEL (Cross Entropy Loss) is used, as given in Equation (4).

$$L_{cls} = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{4}$$

Where M represents the amount of assembly status categories; y_{ic} denotes the sign function (0 or 1), which takes 1 if the true category of the status sample i is equal to c, otherwise it takes 0; p_{ic} indicates the predicted probability that the observed assembly status i belongs to the category c.

For the regression task of assembly components, DFL[59] and CIoU [60] are employed, as depicted in Equation (5).

$$L_{reg} = L_{DFL} + L_{CIoU} \tag{5}$$

The meaning of DFL is to optimize the probability of the two positions which is the closest one to the label, one left and one right, in the form of cross-entropy, so that the network can focus on the distribution of adjacent area of the target position faster. The formula of DFL is shown in Equation (6).

$$L_{DFL} = -((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})) \tag{6}$$

Where DFL changes the single value of coordinate regression to output $n+1$ values, each value represents the probability of the corresponding regression distance, and the integral is calculated to obtain the final regression distance. DFL can make the network focus on the label y faster nearby values, thus increasing their probability.

CIoU is an improved bounding box similarity metric that provides a more efficient and accurate evaluation of bounding box localization by considering several factors, including IoU, centroid distance, and aspect ratio differences. Compared to IoU and DIoU, CIoU reflects the differences between predicted and real boxes more comprehensively, particularly when the bounding box has low overlap. The formula of CIoU is illustrated in Equation (7).

$$L_{CIoU} = 1 - IoU + \frac{p^2(b, b_{gt})}{(c_w)^2 + (c_h)^2} + \frac{4}{\pi^2}(\tan^{-1}\frac{w_{gt}}{h_{gt}} - \tan^{-1}\frac{w}{h}) \tag{7}$$

Where IoU represents the intersection ratio of the prediction box and the real box; $p(b, b_{gt})$ denotes the Euclidean distance between the centroids of the real box and the prediction box; h and w indicate the height and width of the prediction box; h_{gt} and w_{gt} reflect the height and width of the real box (ground truth); c_h and c_w refer the height and the width of the minimum enclosing box formed by the prediction box and the real box.

Further, the Task-Aligned Assigner[61] is also utilized to dynamically assign samples, which improves the detection accuracy and robustness of the object-based assembly progress inspection network. The formula of the Task-Aligned Assigner is presented in Equation (8).

$$t = s^{\alpha} \times u^{\beta} \tag{8}$$

Where s represents the prediction score corresponding to the labeled category; u denotes the IoU of the prediction and ground truth boxes. The multiplication of s and u provides a metric for the alignment between the classification task and regression task.

In the reducer assembly scenario, since the operator's assembly actions and the component's assembly status are simultaneously monitored by an RGBD camera, the object detection task has a high proportion of small-sized assembly parts, as well as partial occlusion is unavoidable during manual assembly. It is noted that the multi-scale feature fusion, bidirectional PAN-FPN structure, and a reasonable loss function applied to the object-aware network can improve the model's performance in selecting the object's edge and texture features across various scales. On the one hand, the detailed shape and contour information of the object from these features can be used to localize the object precisely in case of partial occlusion. On the other hand, the multi-scale features can assist the model in detecting the small-sized assembly components such as small gears and reducer handles, which are essential parts that play a decisive role in the observation of the assembly process.

In contrast to AR-only assembly, which simply superimposes visual guidance directly on the shop floor without validation of the assembly process, the integration of the object-aware network into the AR system enables real-time detection of component status throughout the assembly process. If the operator performs an erroneous assembly procedure, in addition to text-based warnings displayed on the GUI, corresponding 3D-based visual instructions will also be activated based on the results of the assembly progress inspection to rectify the operating errors timely, resulting in a flexible and efficient AR assembly workflow.

4.3. Human-object integrated progress observation in AR assembly

In real assembly workplaces, the two factors of human (operator) and object (component) should be fully considered for human-centric operations. The former decides whether the long-term mental and physical effects caused by the assembly procedures on the operator are positive or not, while the latter determines whether the final assembly targets are completed, a human-object integrated assembly progress observation method for more human-centric AR assembly is expected. To this end, this study aims to propose a generalized and ease-to-deploy framework that integrates ergonomic status perception and assembly process inspection to achieve the supervision and management regarding the status of the operator and component during AR assembly in the first-person perspective, alleviating cognitive workload while leading to an error-corrected intelligent AR assembly. In principle, the proposed framework works with all kinds of skeleton-based human action recognition models and object-based detection models (In our case, the OpenPose and YOLOv8 are deployed, respectively). The workflow of this framework mainly consists of three stages: Multi-modal data collection, Integrative human-object awareness, and Human-centric AR assembly assistance.

Multi-modal data collection: The RGBD camera is applied to monitor both the operator action and assembly sequence, where the RGB and depth video streams with the same resolution of 1920 × 1080 pixels are available as data sources. In addition, the multi-modal data collected by the RGBD sensor are further temporally and spatially precisely aligned based on the timestamps and camera parameters, respectively, thus preparing for the subsequent 2D and 3D human action coordinate conversion.

Integrative human-object awareness: The two-branch Integrative human-object awareness framework consists of the human action-driven ergonomic-aware branch and the object-aware branch. In the

ergonomic-aware branch, the aligned multi-modal data from the RGBD camera are used as the input source, and the human action is recognized by an OpenPose-based network to obtain the 2D joint coordinates, followed by the 2D-3D transformations and Savitzky-Golay smoothing, resulting in the corresponding specific joint angles, which are finally utilized by a revised RULA to quantify the ergonomic risk status of operator's assembly actions. For the object-aware branch, given the same RGB streams, the part stage status can be deduced by a Yolo-based network, which can understand and confirm the current assembly sequence. It is noted that the computational time is considered during the experimentation, in order to enable integrative human-object awareness fast enough, the discussed two branches are computed in multi-threaded parallel, resulting in a minimal overall inference time and meeting the real-time requirement for practical applications. Besides, due to the difference in computational speed between the two branches, in order to maintain the alignment of ergonomic status perception with assembly process inspection, "Thread Event" is applied in each thread to monitor the computational process of the branches.

Human-centric AR assembly assistance: In most workshops, in addition to the operators who are directly associated with the product assembly process, safety supervisors are also an essential part of the workflow, assisting operators in executing safe and regular procedures and playing a crucial role in preventing possible operating safety hazards. To achieve human-centric assistance in the operation and supervision during AR assembly, different software systems are designed and developed to satisfy the respective requirements of the two professions. The GUI of the PyQt5-based software system for supervisors is illustrated in Fig. 6, which allows the supervisor to remotely monitor the operator's status and assembly progress in real-time, ensuring that all assembly procedures are performed safely and successfully. Following this, the operator can identify unreasonable procedures with possible negative physical effects in their AR view, which is unique to our knowledge. Furtherly, the human-object aligned results are also transmitted by the cloud server to the Unity-based software system for the operator, which not only enables the operator to carry out the manual assembly according to the automatically activated AR guidance, but also rectifies their high-workload actions or operating errors based on the text-based instructions.

5. Experiments

To evaluate the proposed human-object integrated AR assembly system, extensive experiments are performed in the reducer assembly scenario. Firstly, the performance of the ergonomic-aware method is evaluated based on the self-established dataset. Then, the accuracy of the object-aware network is measured by recognizing typical reducer assembly processes under four different interference conditions. Finally, a series of user studies on four dimensions: task performance, learning effect, cognitive workload, and system usability are carried out with many participants.

5.1. Evaluations on human action-driven ergonomic-aware in manual assembly

To achieve robust human action recognition during AR assembly, the OpenPose is chosen as the basic framework for the ergonomic status perception of the operator, and the pre-training parameters of the OpenPose are utilized to perform the pose estimation directly. Furthermore, since there is no available assembly ergonomics dataset, a dataset including four typical RULA ergonomics assessment levels is established in the present study. Based on the four risk levels defined in the established ergonomics standard [50], following the growing trend of risk level represented by the ground scores, which are linearly mapped to the percent scores (Score range from 0 to 100) in our case. On the one hand, a fine-grained score range provides a more precise reflection of the changes in the operator's ergonomic status during the manual assembly, resulting in improved sensitivity to the human factor in the AR system. On the other hand, the intuitive percent scores make it easier to combine the RULA criteria with other evaluation tools if needed, thus enhancing the flexibility for additional extensions. In addition, practical applications also prove the validity of such score range, the four ergonomic risk scales with the corresponding score ranges are defined as: Acceptable (RULA score 0–24): negligible risk, no action required; Change needed (RULA score 25–49): low risk, change may be needed; Change soon (RULA score 50–74): medium risk, further investigation; Change immediately (RULA score 75–100): very high risk, implement change now. It is worth noting that during the data collection, the high-precision human joint angle measuring device is used to limit the specific range of each operator's joint motion, resulting in the precise ergonomic risk groundtruth.

The creation of the assembly ergonomic dataset consists of two stages. In the action recording stage, the assembly actions of 12 operators are recorded in the actual assembly scenario. Each operator is required to perform four levels of ergonomic risk manual assembly

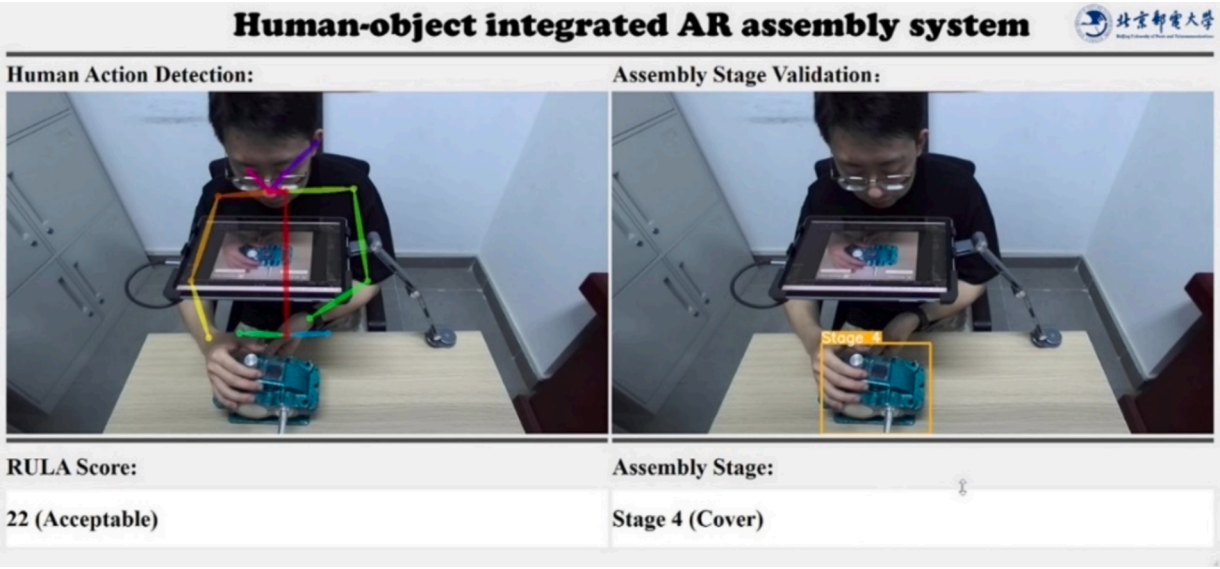


Fig. 6. The software system designed for supervisors.

actions each two times. In addition, two actions from the same level of ergonomic risk are requested to be noticeably distinct. For instance, if the operator leans the body towards one side for the first action, then, he will lean the body towards the reverse direction during the second action. Therefore, 96 video clips (aligned RGB and depth image streams) are collected. In the video frames extraction stage, all videos are classified according to the level of ergonomic risk. Following this, the most representative frame of the manual assembly motion will be extracted from each video. Accordingly, 96 video frames (aligned RGB and depth images) are obtained. Further, these assembly action data with ergonomic risk ratings are utilized in the perception performance evaluation of the proposed ergonomic-aware method. Some typical test results during the experiment are shown in Fig. 7, the images from left to right indicate that the operator performs manual assembly with increasingly high-workload actions. It is observed that different ergonomic risk levels of the operator can be effectively recognized by the ergonomic-aware method, resulting in a human-centric assembly workflow.

The judgment results of ergonomic risk levels from 12 operators are counted, and the statistical results are illustrated in Fig. 8, we can find that in a total number of 96 assembly action samples, 88 action samples can be correctly identified so that the OA (Overall Accuracy) is 91.7 %. Besides, the CA (Class Accuracy) of four ergonomic risk levels are counted as: Acceptable: 95.8 %; Change needed: 91.7 %; Change soon: 87.5 %; Change immediately: 91.7 %. Experimental results indicate that the proposed ergonomic-aware method can accurately detect the operators' current ergonomic status based on the manual assembly actions, thus warning the operators to properly adjust their working gestures in time to alleviate their physical workload, especially when performing assembly tasks with long working hours.

5.2. Evaluations on object-aware assembly process inspection

The present study chooses the YOLOv8 as the backbone of the object-aware network and trains it with 600 RGB images collected in the reducer assembly scenario. During the training process of the object-aware network model, the epoch is set as 200, the ratio of the training data set to the validation data set is 6:4, and the number of batches is selected as 24 according to the GPU (RTX4090) performance. A stochastic gradient descent (SGD) optimizer that optimizes network training by setting its momentum factor to 0.937 and weight decay to 5e-4. In addition, the initial and final learning rates of the weight model training are set to 1e-2. The discussed parameters are controlled to preserve stability. Finally, the object-aware network is trained to achieve the assembly process inspection of the reducer based on RGB streams. The training process and result are depicted in Fig. 9, where the descending process of the losses of training and validating are illustrated in Fig. 9(a), and the anchor window of the training model is shown in Fig. 9(b). Furthermore, the F1 confidence curve of different assembly stages is depicted in Fig. 9(c), which plays an important role in selecting a reasonable confidence threshold for object detection.

To evaluate the inspection performance of the object-aware network on the reducer assembly process, a total of 100 RGB images covering four typical assembly stages are collected during the actual reducer manual assembly as the testing data set. It is noted that these original test data are obtained in a relatively ideal experimental environment, but in real industrial scenarios, where the RGB streams are affected by different factors, such as the background noise, motion blurring, and light intensity, which affects the detection performance of the object-aware network. Therefore, four types of interferences of different

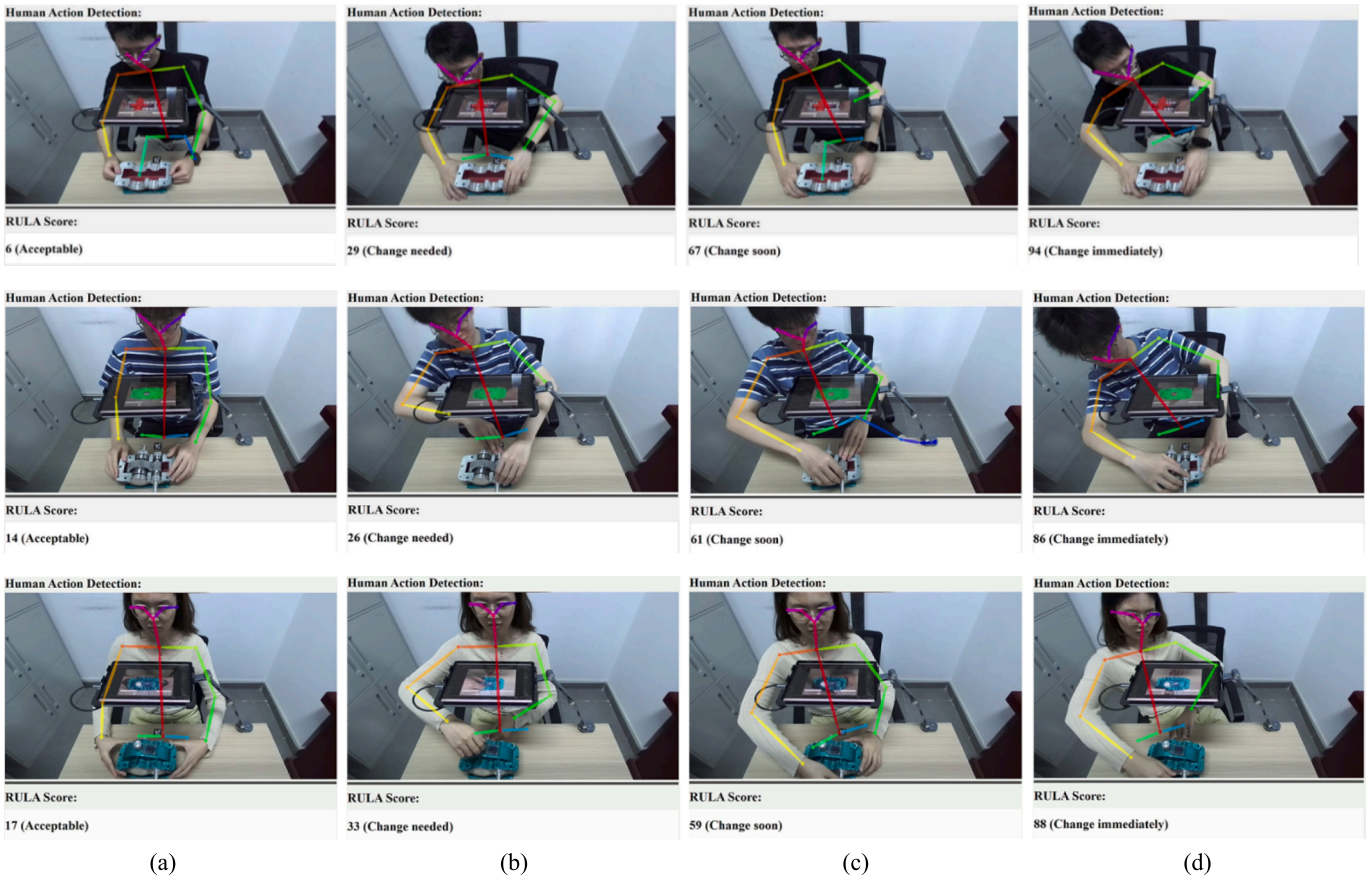


Fig. 7. The typical snapshots of performance evaluation for the ergonomic-aware method: (a) Test results of the ergonomic-aware method for negligible risk assembly actions, (b) Test results of the ergonomic-aware method for low risk assembly actions, (c) Test results of the ergonomic-aware method for medium risk assembly actions, (d) Test results of the ergonomic-aware method for high risk assembly actions.

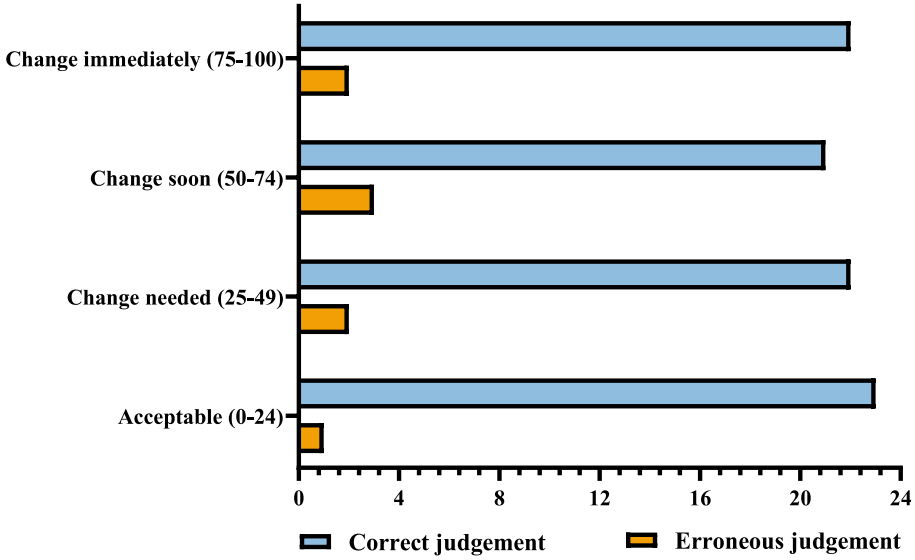


Fig. 8. The judgement results of different ergonomic risk levels.

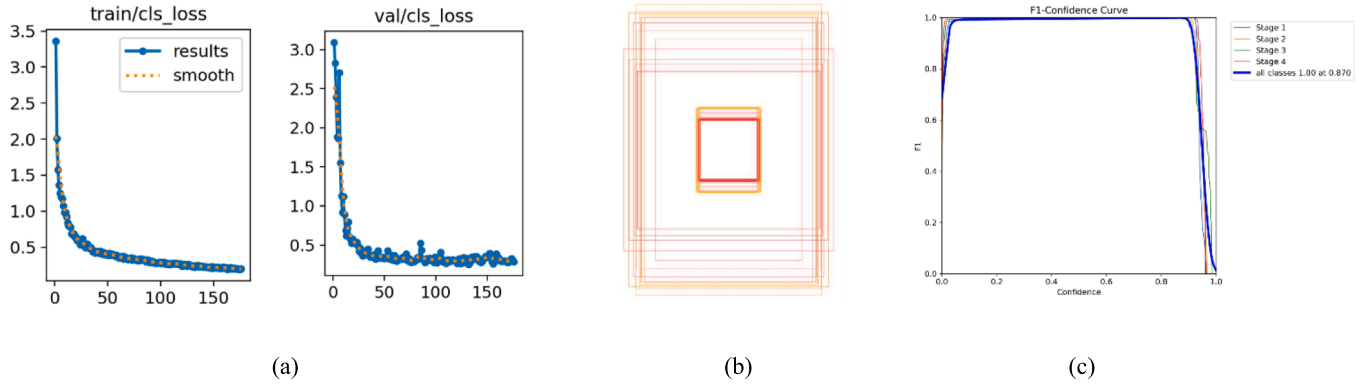


Fig. 9. The object-aware training process for assembly process inspection. (a) Decaying process of the training and validating process. (b) The anchor window of the training model. (c) The F1 confidence curve of different assembly stage.

intensities including salt and pepper noise, Gaussian noise, Gaussian blur, and brightness change are applied to corrupt the original test set to simulate the actual detection environment. Moreover, the inspection performance of the object-aware network for the reducer assembly process under different types and intensities of interferences is shown in Fig. 10.

The statistical results of the detection accuracy of the object-aware network under different types and intensities of interference are illustrated in Fig. 11. We can find that the object-aware network performs well on the original test set (no noise interference), achieving the assembly process inspection accuracy of 95 %. As the interference intensity in the initial RGB image increases, the detection accuracy decreases continuously in the four interference situations, which indicates that the applied simulated interferences have a certain impact on the detection performance of the object-aware network. Further analyzing the accuracy curves, it is observed that the detection accuracy of the object-aware network remains stable even under medium-intensity interferences, and the reducer assembly stage is recognized precisely regardless of interference conditions (accuracy exceeding 90 %). Experimental results also indicate that the object-aware network can enable reliable assembly process inspection to assist workers in avoiding operating errors in most general assembly scenarios and challenging industrial environments.

However, there is a sharp decrease in the detection accuracy of the object-aware network as soon as the interference intensity transcends the robustness threshold (breakdown point, corresponding to 10 % salt and pepper noise, 1 Gaussian noise standard deviation, (31, 31) Gaussian blur kernel size, and 0.3 brightness factor, respectively, in our case). Furthermore, only under extremely poor interference conditions (collapse point, corresponding to 30 % salt and pepper noise, 4 Gaussian noise standard deviation, (51, 51) Gaussian blur kernel size, and 0.2 brightness factor, respectively), the object-aware network will result in a large amount of false and missed inspections for the reducer assembly process (accuracy below 60 %), depicted in Fig. 11 (e), (j), (o), and (t).

5.3. *Evaluation on human-object integrated progress observation in AR assembly*

In this section, comprehensive user surveys are carried out to further evaluate the performance of the proposed human-object integrated progress observation method in AR reducer assembly. Firstly, 30 participants are randomly divided into three groups according to different assembly methods, followed by task performance experiments performed by comparison of the efficiency in executing the predefined assembly task among the three groups. Then, in order to assess the user's learning effect, 10 participants out of the original groups are required to repeat the assembly task with paper-based work instruction compared to human-object integrated AR assistance, respectively. Following this, the

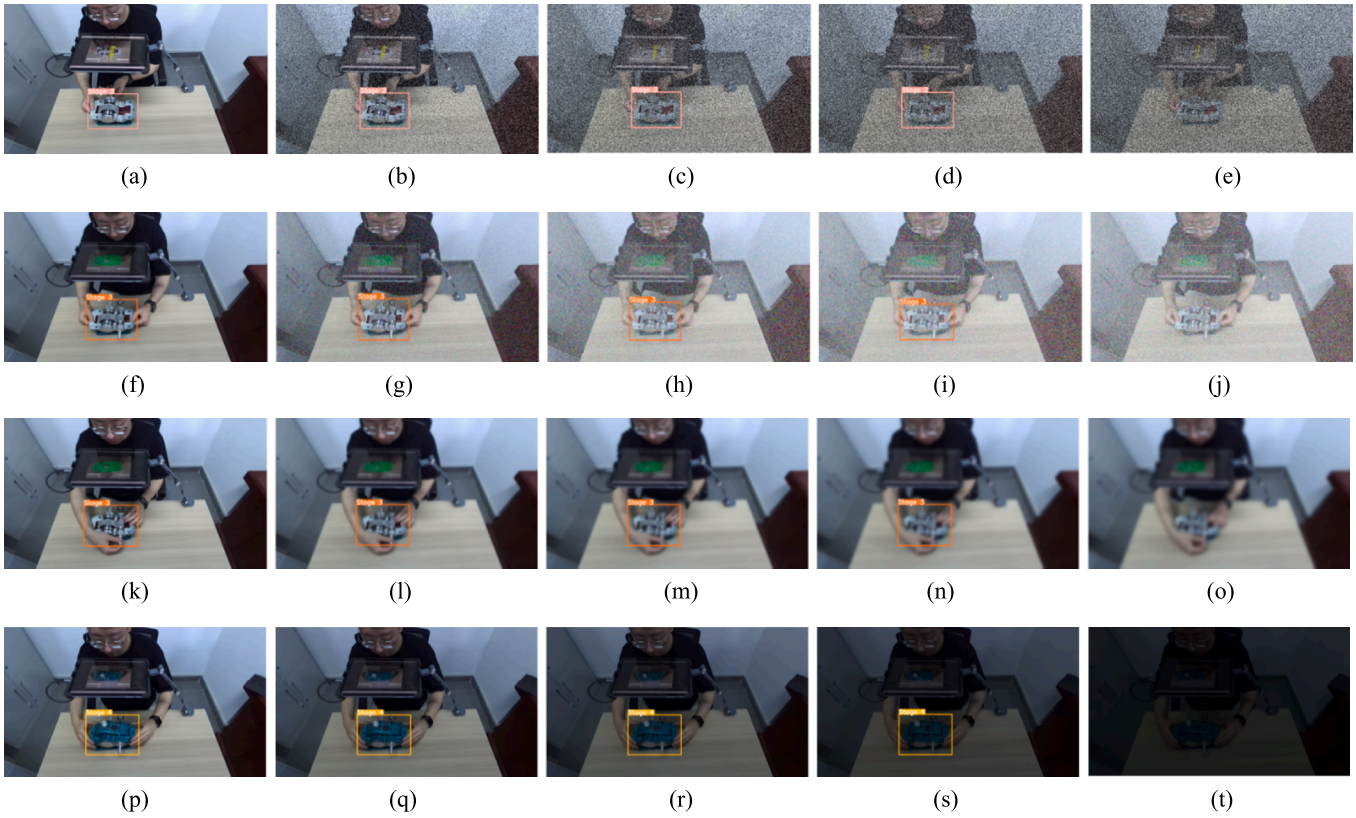


Fig. 10. Test results for detecting the reducer assembly stage under different types and intensities of interference conditions. (a)-(e) 0%, 10%, 20%, 25%, and 30% (failed) salt and pepper noise. (f)-(j) 0, 1, 2, 3, 4 (failed) Gaussian noise standard deviation. (k)-(o) (0,0), (21,21), (31,31), (41,41), (51,51) (failed) Gaussian blur kernel size. (p)-(t) 1.0, 0.8, 0.6, 0.4, 0.2 (failed) brightness factor.

NASA-TLX (NASA Task Load Index) is applied to measure the overall cognitive workload of participants during the use of the two aforementioned assembly methods. Finally, the system usability of the human-object integrated AR assembly is evaluated based on the SUS (System Usability Scale).

5.3.1. *Task performance*

By integrating the human-object integrated assembly progress observation into the AR assembly system, a human-centric AR assembly for a reducer is achieved. As shown in Fig. 12, taking the images in column (b) as an example (assembly stage 2), the first image represents the immersive AR assembly guidance that is automatically activated when step 1 is validated (images in column(a)). Then, the operator accomplishes the manual reducer assembly of stage 2 with the help of visual guidance, illustrated in the second image. It is noted that since the assembly progress observation and data transmission in real-time, the AR assistance for the next assembly stage is activated almost simultaneously as soon as the current assembly stage is confirmed, which is the reason that stage 3 guidance is present in the second image of stage 2. The third image indicates the skeleton tracking result of the human action-aware network, which is also utilized to drive the ergonomic-aware method for the prediction of the RULA score in the AR view, thus achieving the perception and supervision of the operators' ergonomic status throughout the whole manual assembly process in the first-person perspective. The last image represents the observation of the current assembly progress by the object-aware network, and when the current assembly stage 2 is validated as finished, the subsequent AR assembly guidance will be automatically activated (stage 3, displayed in the first image of column (c)). The columns from (a) to (d) correspond to assembly stages 1 to 4, and finally complete all the assembly procedures for a reducer.

To evaluate the proposed human-centric AR assembly system in the actual manual assembly task, the task performance evaluation experiment is conducted with 30 participants. Since the AR assembly system is expected to support novice operators, we intentionally invited participants without prior knowledge about reducer assembly. Each participant is required to complete a pre-study survey with questions about their demographic information before the study. Among the 30 participants, there are 21 males and 9 females, their mean (M) age is 24 (standard deviation (SD = 2), and they are science and engineering students. Then, these participants are randomly divided into three groups (10 in each group): Group A, Group B, and Group C. The detailed descriptions for the three groups are given as follows: Group A: participants accomplish the assembly task with the traditional paper-based assembly outline; Group B: participants complete the assembly task according to the AR-only assembly guidance; Group C: participants perform assembly tasks with the human-object integrated AR assembly, which automatically activates the subsequent AR assembly instruction based on the assembly progress observation. In contrast to the AR-only method, there is no need to perform additional human operations related to the stage switching.

Before carrying out the experiment, all participants are provided with a pre-training session with 20 min of introduction and operation instruction to familiarize them with the task procedures. During the experiment, the detailed completion time for the reducer assembly task of each participant is recorded and depicted in Fig. 13(a), and we can find that all participants can complete the assembly task correctly, which may benefit from limited assembly steps. However, the accomplished time varies with different assembly methods. The average completion time of Group A is 50.8 s, while 40.2 s with AR-only guidance for Group B, the reason may be that the operators can obtain a holistic understanding with intuitive AR guidance, which is helpful for a high-efficiency operation. In contrast to AR-only instruction, the proposed human-object integrated AR assembly can assist the operator in

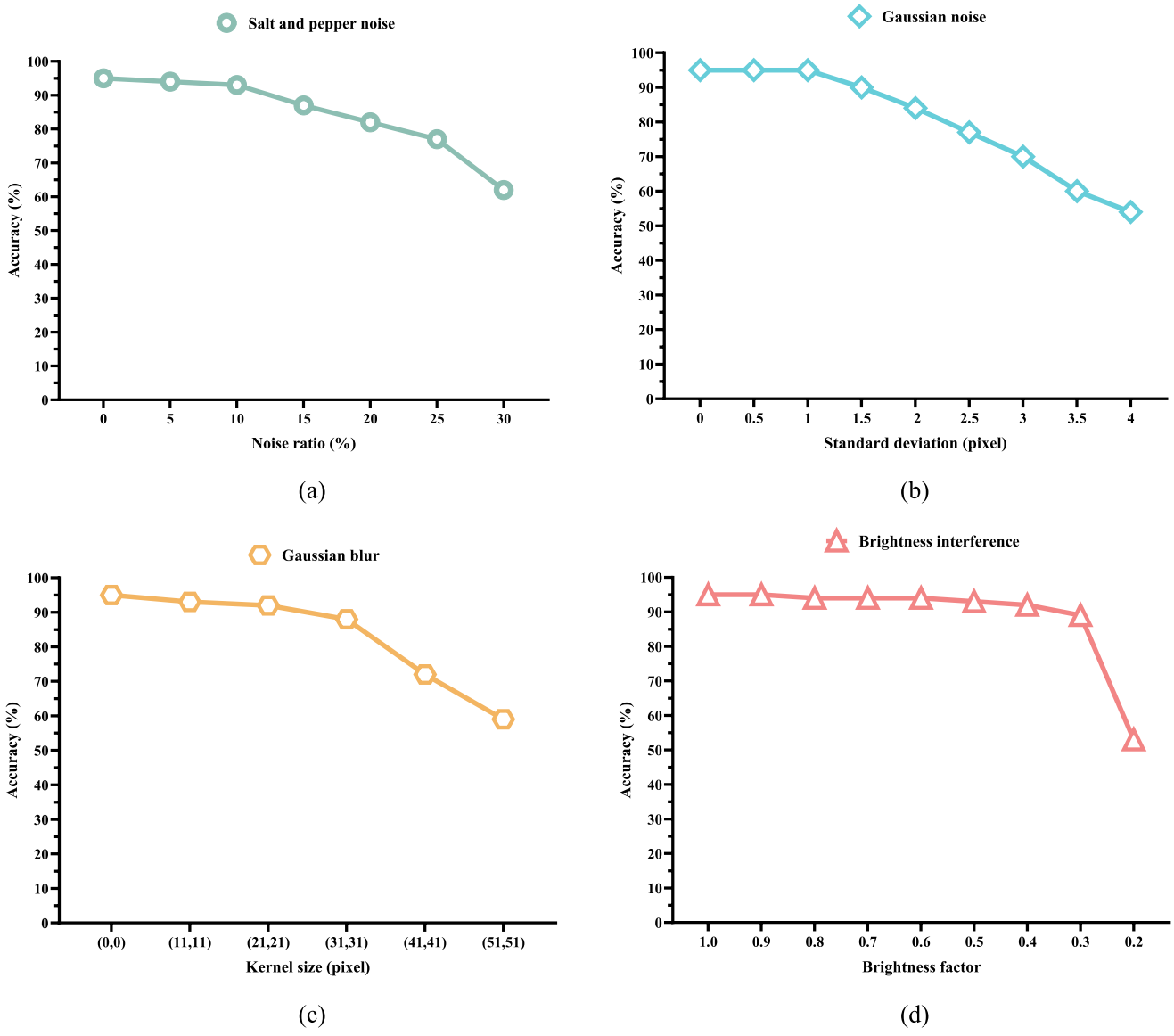


Fig. 11. Detection accuracy of the object-aware network under different types and intensities of interference situations. (a) Accuracy curve for various intensities of salt and pepper noise. (b) Accuracy curve for various intensities of Gaussian noise. (c) Accuracy curve for various intensities of Gaussian blur. (d) Accuracy curve for various intensities of brightness interference.

avoiding the extra time to switch and confirm the sequential assembly process by human intervention, thus achieving superior efficiency with an average accomplish time of 32.6 s (Group C), 19.4 % improvement compared to Group B. Moreover, the standard derivations for different groups have a similar tendency as the completion time, Group C (SD= 2.4 s) still shows superior performance compared to Group B (SD=3.3 s) and Group A (SD=5.8 s). Experimental results illustrate that not only the efficiency improvement is achieved with the human-object integrated AR assembly, but also a more convergent result is achieved.

Furtherly, to test for statistical significance between the three assembly methods, a one-way ANOVA (suitable for our independent measures design with the three methods) is used to compare mean completion times, since the applicability conditions normality and homoscedasticity are met. The established confidence level for all analyses is set as 95 % (α=0.05), and normality is confirmed using the Shapiro-Wilk test (p_A= 0.97, p_B= 0.82, p_B= 0.73), and homoscedasticity using the Levine test (p= 0.08). The results show that the average completion times are significantly different among the three groups (p= 8.78E-10, F= 49.77). To specify which pairs of groups have significantly different mean completion times, thus post-hoc pairwise comparisons are performed using the Tukey HSD test. Experimental results are shown in Fig. 13(b), it is observed that highly significant differences are found between each of the two groups, which provides strong evidence that our human-object integrated AR assembly system reduces completion time significantly in untrained operators when performing the given task for the first time, and results in more intelligent human-centric assistance for manual reducer assembly.

5.3.2. *Learning effect*

In addition to the evaluation regarding the task performance of operators without prior assembly experience using the proposed AR system, in order to assess the user's learning effect, 10 participants (labeled as P1, P2, …, and P10) are picked out of the original groups and requested to do the same assembly task 15 times with traditional paper-based work instruction and the human-object integrated AR assistance system. It is noted that to eliminate possible bias due to the experiment order and operator fatigue, each participant performed the two methods in alternating order. For instance, the participants labeled as P1, P3, …,

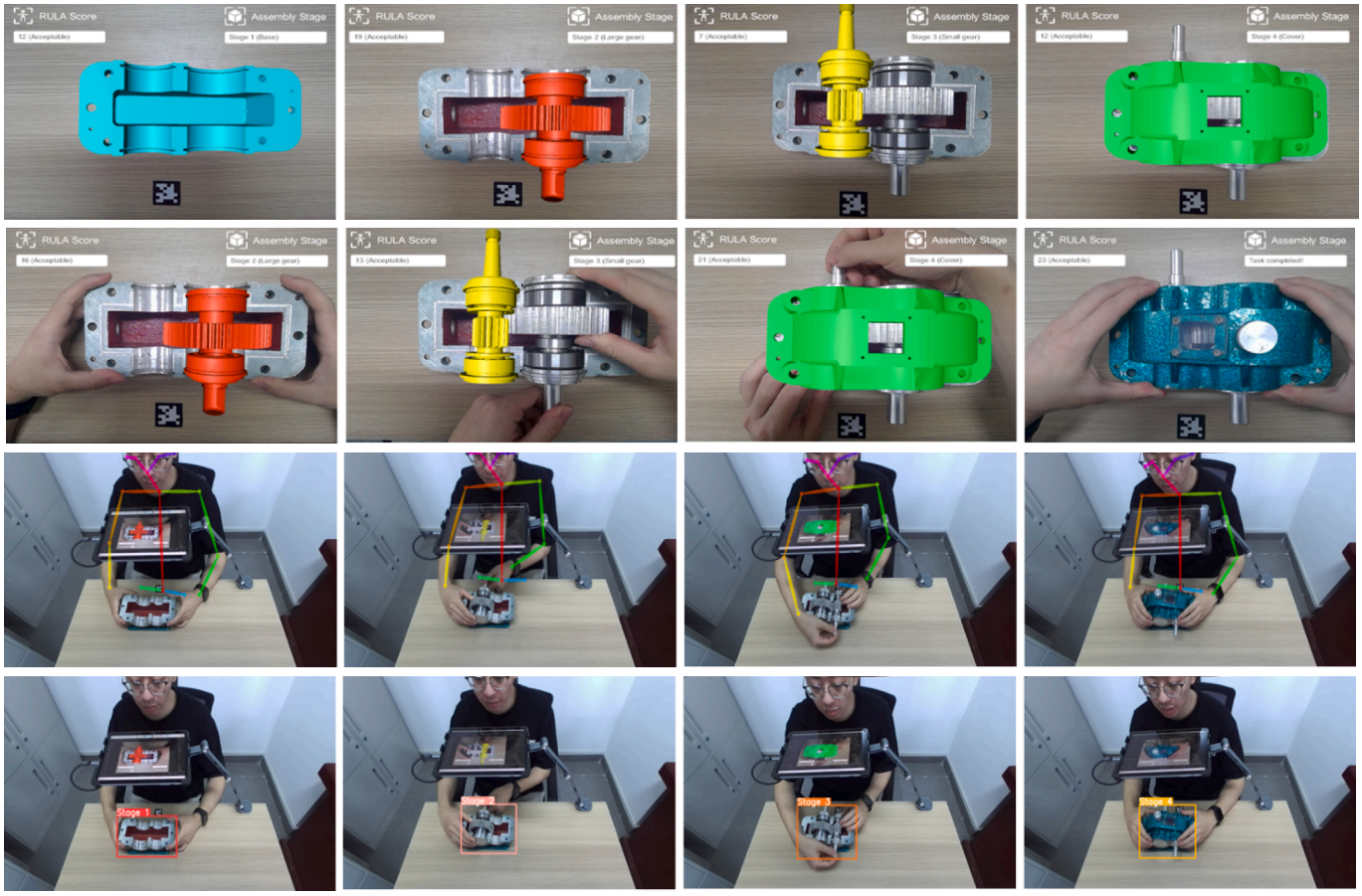


Fig. 12. Some typical assembly snapshots of the human-object integrated AR assembly process for a reducer. (a) Assembly stage 1, placing the reducer base. (b) Assembly stage 2, installing the large gear. (c) Assembly stage 3, installing the small gear. (d) Assembly stage 4, installing the reducer cover.

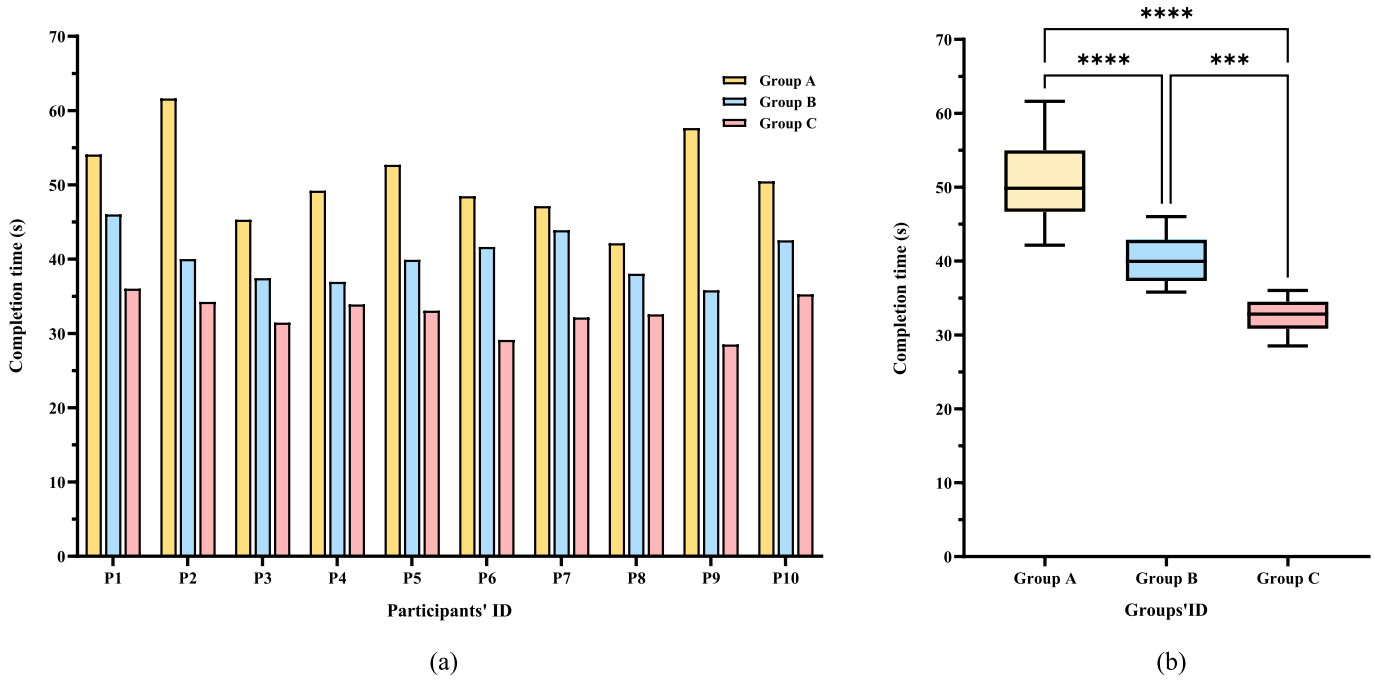


Fig. 13. Completion time statistics with different assembly methods. (a) Recorded time for each participant to complete the reducer assembly task. (b) Statistical significance of completion time among different groups.

and P9 assemble the reducer with the paper-based instruction and then reassemble them under the human-object integrated AR assistance. On the contrary, the participants labeled as P2, P4, …, and P10 assemble the reducer under the human-object integrated AR assistance and then reassemble them with the paper-based instruction. Besides, two experiments for the same participant are executed at different periods with a large interval to ensure that the participant has enough rest time to eliminate possible assembly fatigue from the first experiment when performing the second experiment.

Following this, participants' durations to complete the task each time are measured and recorded, the users learning curves as illustrated in Fig. 14. We can find that the average completion time converges at the 9th iteration while using method 1 (paper-based assembly instruction) and at the 5th iteration in method 2 (human-object integrated AR assistance), which indicates that the human-object integrated AR assembly system is more user-friendly and usable than the traditional assembly method (paper-based work instruction). Notably, without the AR assistance, participants achieve similar or even slightly faster task completion after convergence is reached, with the best result requiring only 22.65 s while the shortest completion time when using the assistance system is 23.36 s. The reason may be that when operators are fully familiar with the assembly procedures (external support is not necessary), the additional features of AR assistance may detract from the operators' attention. Nevertheless, this study also gives empirical evidence that the proposed human-object integrated AR assembly system can help novice operators faster learn and familiarize themselves with workflows, so it is still meaningful to change the traditional assembly mode.

5.3.3. Cognitive workload

To determine if the proposed AR assembly assistance can lead to a reduced cognitive workload on the operator, following the learning effect assessment experiment, each participant is required to answer the NASA-TLX questionnaire according to their experience with the paper-based assembly instruction compared to the human-object integrated AR assembly, respectively. It is noteworthy that since a single reducer assembly procedure may be limited and relatively simple, it's not representative of the fatigue-prone manual assembly task. To provide more evidence that the proposed human-object integrated AR-assisted system is effective in alleviating long-term cognitive workload, a specific task of high-workload and fatigue-prone is designed in the learning effect experiment, which consisted of many iterations for the assembly and disassembly task with intensive operational actions, as well as all participants lacking experience in long-term assembly. It is observed that participants suffer from fatigue and soreness of the upper limbs at approximately 20 iterations (For each participant, there are 30 iterations during the experiment, with a duration of around 35 min).

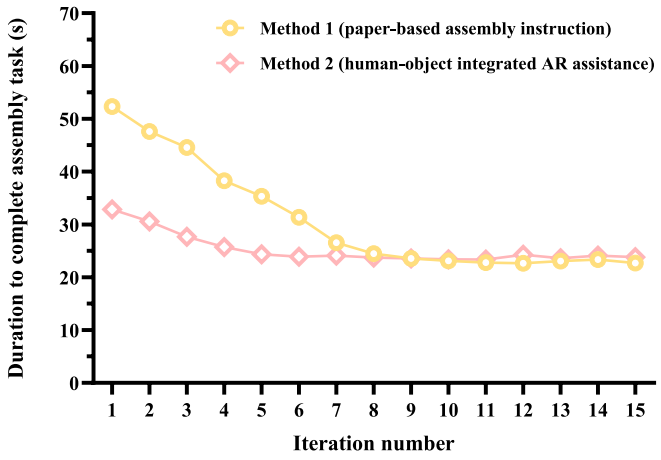


Fig. 14. Learning curves of paper-based instruction (yellow) and human-object integrated AR assistance (pink).

These quantitative NASA-TLX data are shown in Fig. 15(a), it is observed that generally the average cognitive workload is reduced by the use of a human-object integrated AR system (M= 22.83, SD= 4.68) compared with the paper-based assembly instruction (M= 33.83, SD= 2.43) for both operators. Moreover, since the datasets for the two methods don't follow the normal distribution, thus a Wilcoxon signed-rank test is applied to specify if there is an overall significantly different TLX score between the two methods. As depicted in Fig. 15(b), the results of the Wilcoxon signed-rank test (p= 0.002) also provide strong evidence that the human-object integrated AR system reduces workload significantly.

Furthermore, to ascertain the reason for this result, the six dimensions of the NASA-TLX include Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration are also tested for significant differences (seen in Fig. 15(b)). We can find that there is a highly significant difference between the two methods on three dimensions, Mental Demand, Physical Demand, and Effort, with a minor significant difference in Performance, as well as no significant difference found in Temporal Demand and Frustration. The reason may be that the current assembly stage is confirmed by the assembly process inspection in real-time, also the corresponding intuitive AR guidance will be automatically activated, thus alleviating the operator's extra effort to understand the text-based work instruction and perform the intervention. Meanwhile, the level of the operator's ergonomic risk is supervised by the ergonomic status perception depending on the human assembly action, thereby reducing the operator's physical workload. It is also evidence that the integration of the human action-aware driven ergonomic status perception with object-aware driven assembly progress observation can result in a more human-centric AR assembly.

5.3.4. Systematic usability

Lastly, the standard SUS is applied to evaluate the systematic usability of the human-object integrated AR assembly, 10 participants are also asked to answer the SUS statements according to their intuitive perception, and the questionnaire is shown to the operator after accomplishing the reducer assembly task with the human-object integrated AR assembly. The scores of the SUS from all participants are shown in Fig. 16, and the mean score of SUS (M=81.75, SD= 3.34) is above average 68, proving the great usability of the human-object integrated AR assembly [62].

Moreover, according to the post-experiment feedback from the participants, most of them show a positive opinion of the proposed AR assembly system. As well as there exists a minority of participants who point out that integrating comfortable human–computer interaction will lead to a more intelligent and holistic human-centric AR assembly. For instance, integrating audio interaction into the AR system can be more effective than 2D display in warning operators when their assembly actions are not ergonomically friendly, which indicates that more human factors should be fully considered in the future.

6. Conclusion and future works

The paper describes an integrative human action and object recognition-driven progress observation for online smart AR assembly, enabling the human-factor and status-aware human-centric intuitive guidance, which is unique to the best of our knowledge. According to the assembly part and human activity recognition from the RGBD camera, an integrative context-aware assembly progress observation and validation from the human-object AR assembly procedure is achieved, which can check the ongoing AR assembly procedures without human intervention. Then, the online and quantitative human factor indicator for the manual operation in the AR view is established, thus, the operators can perceive whether their working posture is appropriate or not during the assembly progress from the first-person perspective, and then adjust to a more appropriate working state as soon as possible, enabling

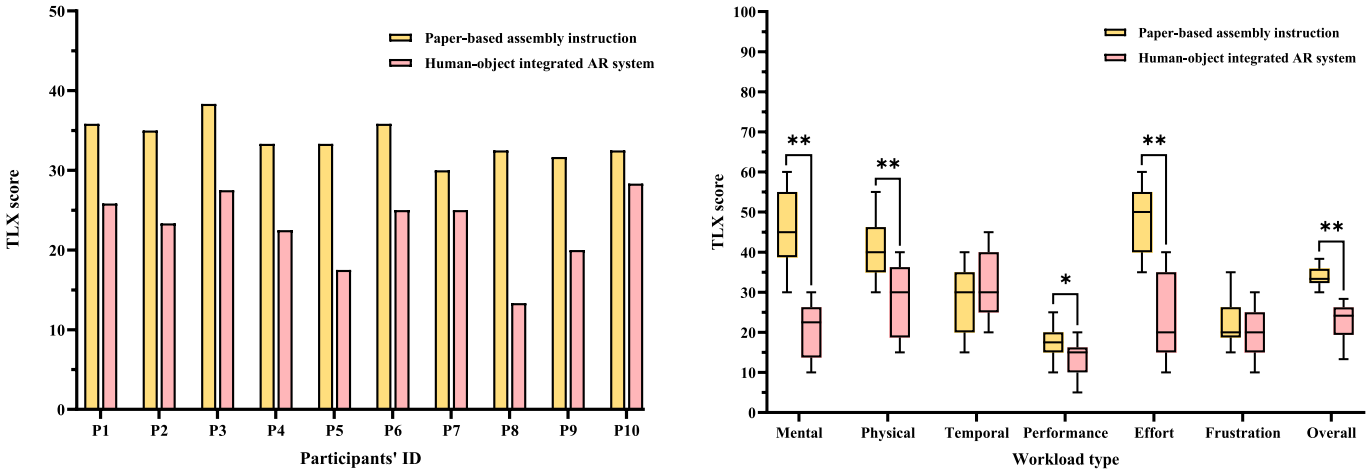


Fig. 15. Statistical results of TLX scores between paper-based assembly instruction and human-object integrated AR system. (a) Summary TLX score of participants. (b) Statistical significance of TLX score on the different dimensions.

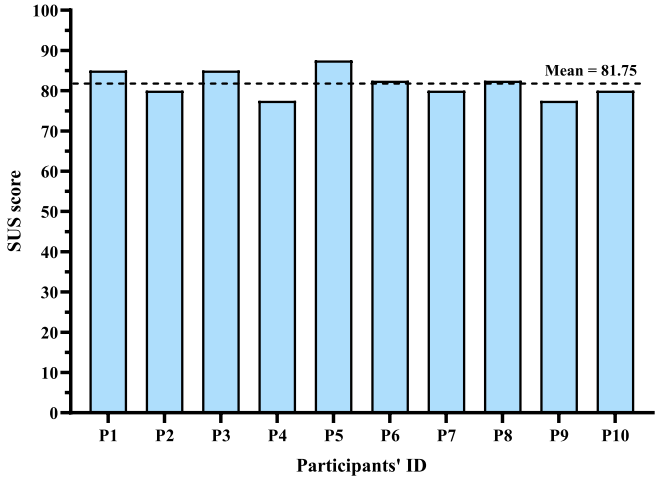


Fig. 16. The statistical SUS score of the integrative human-object aware smart AR assembly system.

a holistic framework and active feedback for the human-in-the-loop AR assembly operations. The combination of human action recognition with object awareness enables operators to perform manufacturing tasks rapidly and accurately with a low cognitive workload, and this integrative AI-driven assembly addresses how expert system AI can improve productivity and profitability in smart manufacturing.

Finally, extensive experiments are carried out to validate several hypotheses related to the human-object integrated performance in the smart AR assembly. First hypothesis: The human-object integrated AR assembly can effectively mitigate the operator's task completion time, leading to increased productivity (Task performance evaluation in section 5.3.1). Second hypothesis: The human-object integrated AR assembly can effectively alleviate the operator's cognitive workload during manual assembly (Cognitive workload evaluation in section 5.3.3). Third hypothesis: The human-object integrated AR assembly is friendly and easy-to-use for novice operators (Learning effect evaluation and systematic usability evaluation in sections 5.3.2 and 5.3.4, respectively). Experimental results demonstrate that the proposed method can result in the online assembly observation from a holistic perspective, alleviate the cognitive load, and achieve superior performance on the AR assembly tasks for novice operators, enabling a more human-centric manual assembly operation.

Although remarkable performance is achieved in the proposed integrated human-object AR application, it's still worth pointing out that the challenge with the existing practices for our method is occlusion and drastic actions may occur from time to time during manual assembly operations, which degrade the accuracy of the human activity recognition and cause abrupt changes in the online human-factor identification results. To alleviate the issue of abrupt changes, the Savitzky-Golay filter is utilized to smooth the 3D human action data in the paper. Instead of the filter-based method, the prior assembly knowledge graph is helpful in establishing continuous association among adjacent assembly operations, which can alleviate the inaccurate human factor indicator for ongoing AR assembly and deserve further attention in future.

CRediT authorship contribution statement

Tienong Zhang: Writing – original draft, Validation. **Yuqing Cui:** Writing – review & editing. **Wei Fang:** Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors gratefully acknowledge the support of the National Natural Science Foundation of China (52105505) and the Beijing Natural Science Foundation (3204050).

Data availability

Data will be made available on request.

References

[1] G. Michalos, S. Makris, N. Papakostas, et al., Automotive assembly technologies review: challenges and outlook for a flexible and adaptive approach, CIRP. J. Manuf. Sci. Tec. 2 (2) (2010) 81–91.
[2] H. Xiao, Y. Duan, Z. Zhang, et al., Detection and estimation of mental fatigue in manual assembly process of complex products, Assembly. Autom. 38 (2) (2018) 239–247.
[3] M. Fu, W. Fang, S. Gao, et al., Edge computing-driven scene-aware intelligent augmented reality assembly, Int. J. Adv. Manuf. Technol. 119 (2022) 7369–7381.
[4] M. Eswaran, M.V.A. Raju Bahubalendruni, Challenges and opportunities on AR/VR technologies for manufacturing systems in the context of industry 4.0: A state of the art review, J. Manuf. Syst. 65 (2022) 260–278.

[5] M. Eswaran, A.K. Gulivindala, A.K. Inkulu, et al., Augmented reality-based guidance in product assembly and maintenance/repair perspective: a state of the art review on challenges and opportunities, Expert. Syst. Appl. 213 (2023) 118983.
[6] W. Li, A. Xu, M. Wei, et al., Deep learning-based augmented reality work instruction assistance system for complex manual assembly, J. Manuf. Syst. 73 (2024) 307–319.
[7] L.P. Berg, J.M. Vance, Industry use of virtual reality in product design and manufacturing: a survey, Virtual. Real. 21 (2017) 1–17.
[8] Z. Wang, X. Bai, S. Zhang, et al., The role of user-centered AR instruction in improving novice spatial cognition in a high-precision procedural task, Adv. Eng. Inform. 47 (2021) 101250.
[9] M. Eswaran, M. Bahubalendruni, Augmented reality aided object mapping for worker assistance/training in an industrial assembly context: exploration of affordance with existing guidance techniques, Comput. Ind. Eng. 185 (2023) 109663.
[10] A.E. Uva, M. Gattullo, V.M. Manghisi, et al., Evaluating the effectiveness of spatial augmented reality in smart manufacturing: a solution for manual working stations, Int. J. Adv. Manuf. Technol. 94 (2018) 509–521.
[11] K. Wang, D. Liu, Z. Liu, et al., A fast object registration method for augmented reality assembly with simultaneous determination of multiple 2D-3D correspondences, Robot. Comput. Integr. Manuf. 63 (2020) 101890.
[12] K. Su, G. Du, H. Yuan, et al., A natural bare-hand interaction method with augmented reality for constraint-based virtual assembly, IEEE T. Instrum. Meas. 71 (2022) 1–14.
[13] K.B. Park, M. Kim, S. Choi, et al., Deep learning-based smart task assistance in wearable augmented reality, Robot. Comput. Integr. Manuf. 63 (2020) 101887.
[14] Z. Liu, Z. Zhu, E. Jiang, F. Huang, A.M. Villanueva, X. Qian, T. Wang, K. Ramani, Instrumentar: Auto-generation of augmented reality tutorials for operating digital instruments through recording embodied demonstration, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–17.
[15] D. Ariansyah, J. Erkoyuncu, I. Eimontaite, et al., A head mounted augmented reality design practice for maintenance assembly: Toward meeting perceptual and cognitive needs of AR users, Appl. Ergon. 98 (2022) 103597.
[16] D. Aganian, M. Kohler, S. Baake, et al., How Object Information Improves Skeleton-based Human Action Recognition in Assembly Tasks, in: 2023 International Joint Conference on Neural Networks, 2023, pp. 1–9.
[17] W. Fang, L. Chen, T. Zhang, et al., Head-mounted display augmented reality in manufacturing: A systematic review, Robot. Comput. Integr. Manuf. 83 (2023) 102567.
[18] L. Cardoso, F. Mariano, E. Zorzal, Mobile augmented reality to support fuselage assembly, Comput. Ind. Eng. 148 (2020) 106712.
[19] M. Eswaran, V.S.S. Vara Prasad, M. Hymavathi, et al., Augmented reality guided autonomous assembly system: A novel framework for assembly sequence input validations and creation of virtual content for AR instructions development, J. Manuf. Syst. 72 (2024) 104–121.
[20] M.V.A. Raju Bahubalendruni, P. Bhavasagar, Assembly sequence validation with feasibility testing for augmented reality assisted assembly visualization, Processes. 11 (7) (2023) 2094.
[21] B. Simoes, R. Amicis, I. Barandiaran, et al., Cross reality to enhance worker cognition in industrial assembly operations, Int. J. Adv. Manuf. Tech. 105 (2019) 3965–3978.
[22] M. Drouot, N.L. Bigot, E. Bricard, et al., Augmented reality on industrial assembly line: Impact on effectiveness and mental workload, Appl. Ergon. 103 (2022) 103793.
[23] E. Marino, L. Barbirei, F. Bruno, et al., Assessing user performance in augmented reality assembly guidance for industry 4.0 operators, Comput. Ind. 157–158 (2024) 104085.
[24] L. Wang, A futuristic perspective on human-centric assembly, J. Manuf. Syst. 62 (2022) 199–201.
[25] C. Zhang, Z. Wang, G. Zhou, et al., Towards new-generation human-centric smart manufacturing in Industry 5.0: A systematic review, Adv. Eng. Inform. 57 (2023) 102121.
[26] Z. Lai, W. Tao, M. Leu, et al., Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing, J. Manuf. Syst. 55 (2020) 69–81.
[27] W. Li, J. Wang, M. Liu, et al., Real-time occlusion handling for augmented reality assistance assembly systems with monocular images, J. Manuf. Syst. 62 (2022) 561–574.
[28] Q. Zhao, Y. Kong, S. Sheng, et al., Redundant object detection method for civil aircraft assembly based on machine vision and smart glasses, Meas. Sci. Technol. 33 (2022) 105011.
[29] W. Fang, T. Zhang, Z. Wang, et al., A multi-modal context-aware sequence stage validation for human-centric AR assembly, Comput. Ind. Eng. 194 (2024) 110355.
[30] S. Bahaei, B. Gallina, Assessing risk of AR and organizational changes factors in socio-technical robotic manufacturing, Robot. Comput. Integr. Manuf. 88 (2024) 102731.
[31] F. Schuster, B. Engelmann, U. Sponholz, et al., Human acceptance evaluation of AR-assisted assembly scenarios, J. Manuf. Syst. 61 (2021) 660–672.
[32] A. Generosi, T. Agostinelli, S. Ceccacci, et al., A novel platform to enable the future human-centered factory, Int. J. Adv. Manuf. Tech. 122 (2022) 4221–4233.
[33] M. Chang, A.Y.C. Nee, S.K. Ong, Interactive AR-assisted product disassembly sequence planning, Int. J. Prod. Res. 58 (2020) 4916–4931.
[34] X. Yin, X. Fan, W. Zhu, et al., Synchronous AR assembly assistance and monitoring system based on ego-centric vision, Assembly. Autom. 39 (1) (2019) 1–16.
[35] Y. Hong, J. Zhang, H. Fan, et al., A marker-less assembly stage recognition method based on corner feature, Adv. Eng. Inform. 56 (2023) 101950.
[36] E. Marino, L. Barbieri, B. Colacino, et al., An augmented reality inspection tool to support workers in Industry 4.0 environments, Comput. Ind. 127 (2021) 103412.
[37] A. Stanescu, P. Mohr, M. Kozinski, et al., State-aware configuration detection for augmented reality step-by-step tutorials, in: Proceedings of IEEE International Symposium on Mixed and Augmented Reality, 2023, pp. 157–166.
[38] Y. Ghasemi, H. Jeong, S. Choi, et al., Deep learning-based object detection in augmented reality: A systematic review, Comput. Ind. 139 (2022) 103661.
[39] T. Kaimel, A. Stanescu, P. Mohr, et al., Progress Observation in Augmented Reality Assembly Tutorials Using Dynamic Hand Gesture Recognition, in: Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, 2024, pp. 1–2.
[40] S. Chidambaram, H. Huang, F. He, X. Qian, A.M. Villanueva, T.S. Redick, W. Stuerzlinger, K. Ramani, Processar: An augmented reality-based tool to create in-situ procedural 2d/3d ar instructions, in: Proceedings of the 2021 ACM Designing Interactive Systems Conference, 2021, pp. 234–249.
[41] L. Kästner, L. Eversberg, M. Mursa, Integrative Object and Pose to Task Detection for an Augmented-Reality-based Human Assistance System using Neural Networks, in: 2020 IEEE Eighth International Conference on Communications and Electronics, 2021, pp. 332–337.
[42] C. Chen, T. Wang, D. Li, et al., Repetitive assembly action recognition based on object detection and pose estimation, J. Manuf. Syst. 55 (2020) 325–333.
[43] Y. Zhang, K. Ding, J. Hui, et al., Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly, Adv. Eng. Inform. 54 (2022) 101792.
[44] Z. Cao, T. Simon, S.E. Wei, et al., Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.
[45] V. Bazarevsky, BlazePose: On-device real-time body pose tracking, arXiv: 2006.10204, 2020.
[46] A. Mathis, P. Mamidanna, K.M. Cury, et al., DeepLabCut: markerless pose estimation of user-defined body parts with deep learning, Nat. Neurosci. 21 (9) (2018) 1281–1289.
[47] F. Zhang, P. Juneau, C. McGuirk, et al., Comparison of OpenPose and HyperPose Artificial Intelligence Models for Analysis of Hand-held Smartphone Videos, in: IEEE International Symposium on Medical Measurements and Applications, 2021, pp. 1–6.
[48] E.P. Washabaugh, T.A. Shanmugam, R. Ranganathan, et al., Comparing the accuracy of open-source pose estimation methods for measuring gait kinematics, Gait. Posture. 97 (2022) 188–195.
[49] S. Mroz, N. Baddour, C. McGuirk, et al., Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose, in: International Conference on Bio-Engineering for Smart Technologies, 2021, pp. 1–4.
[50] L. McAtamney, E.N. Corlett, RULA: a survey method for the investigation of work-related upper limb disorders, Appl. Ergon. 24 (2) (1993) 91–99.
[51] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (8) (1964) 1627–1639.
[52] J. Redmon, You Only Look Once: Unified, Real-time Object Detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[53] S. Ren, K. He, R. Girshick, et al., Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks, in, Adv. Neural Inf. Proces. Syst. (2015) 91–99.
[54] W. Liu, D. Anguelov, D. Erhan, et al., Ssd: Single Shot Multibox Detector, in: European Conference on Computer Vision, 2016, pp. 21–37.
[55] J.A. Kim, J.Y. Sung, S.H. Park, Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition, in: IEEE International Conference on Consumer Electronics, 2020, pp. 1–4.
[56] N. Chitraningrum, L. Banowati, D. Herdiana, et al., Comparison study of corn leaf disease detection based on deep learning YOLO-v5 and YOLO-v8, J. Eng. Technol. Sci. 56 (1) (2024) 61–70.
[57] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv:1804.02767, 2018.
[58] S. Liu, L. Qi, H. Qin, et al., Path Aggregation Network for Instance Segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[59] X. Li, W. Wang, L. Wu, et al., Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, Adv. Neural. Inform. Process. Syst. 33 (2020) 21002–21012.
[60] Z. Zheng, P. Wang, W. Liu, et al., Distance-IoU loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12993–13000.
[61] C. Feng, Y. Zhong, Y. Gao, et al., TOOD: Task-Aligned One-Stage Object Detection, in: Proceedings of the 2021 IEEE International Conference on Computer Vision, 2021, pp. 3490–3499.
[62] J. Brooke, SUS: a retrospective, J. Usability. Stud. 8 (2) (2013) 29–40.