Journal of Manufacturing Systems 83 (2025) 937–962

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

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

Technical paper

Intelligent disassembly scenario understanding for human behavior and intention recognition towards self-perception human-robot collaboration system

Jinhua Xiao [a,*], Bo Wang [b], Kaile Huang [c], Sergio Terzi [a], Wei Wang [d], Marco Macchi [a]

[a] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan 20133, Italy
[b] Wuhan University of Technology, School of Transportation and Logistics Engineering, Wuhan 430063, PR China
[c] Beijing Institute of Technology, School of Mechanical engineering, Beijing 100081, PR China
[d] School of Engineering Sciences, University of Skövde, Kaplansgatan 11, Skövde 54134, Sweden

ARTICLE INFO

Keywords:
Human robot collaboration
Intent recognition
Disassembly system
End-of-life product
ST-GCN algorithm

ABSTRACT

The recycling of end-of-life (EOL) products poses significant challenges due to inefficient and unsafe disassembly processes. To address this, we propose a novel self-perception human-robot collaboration (HRC) system that enhances disassembly efficiency and safety through multi-modal human intention recognition. Our core methodological innovation lies in the real-time fusion of three key perception modules: action recognition using a Spatial-Temporal Graph Convolutional Network (ST-GCN), disassembly tool detection based on an enhanced YOLO algorithm, and facial angle recognition for operator awareness inference. A dedicated dataset for retired power battery disassembly was constructed to support this research, encompassing human skeletal data for action recognition, labeled images for tool detection, and facial expression detection. The proposed system was validated on a physical HRC disassembly platform. Experimental results demonstrate a marked improvement, with our integrated intention recognition method achieving an accuracy of approximately 85 %, significantly outperforming traditional single-modality approaches. Furthermore, the HRC disassembly operation was completed in 238 s, which is 60 s (or 20 %) faster than purely manual disassembly. This work provides a robust and efficient HRC disassembly framework for intelligent disassembly scenario understanding, contributing to advancing circular manufacturing.

1. Introduction

Due to the intensification of policy incentives and the growing awareness of environmental protection and resource re-utilization in the related industrial applications, the recycling of End-Of-Life (EOL) products has increasingly expanded to further accelerate the requirements of sustainable and circular manufacturing [1]. However, it is poised to have a profound impact not only on the recycling industry but also on global supply chains, environmental sustainability, and resource management. By considering the complexity of EOL product recycling, effective disassembly is essential in circular manufacturing to recover valuable materials and reduce environmental impact. Furthermore, it will lead to an accumulation of EOL products by relying solely on manual disassembly, which might cause severe environmental pollution and resource waste [2]. However, market competitions cause a wide variety of EOL product structures, which might make it difficult for traditional industrial robots to dynamically handle complex disassembly operations with uncertainty.

In recent years, robot-assisted disassembly has been increasingly used due to its flexible operations and user-friendliness to perceive its surrounding environment. Similarly, human-robot collaboration (HRC) disassembly leverages the strengths of both humans and robots by assigning repetitive tasks to robots to reduce the workload of human workers, while humans can handle complex situations to address intelligent tasks with higher efficiency [3]. An appropriate HRC technology will provide explicit commands (i.e., voice prompts, gesture recognition, etc.) by controlling the robot that can recognize the human operator's intentions to provide the related assisted disassembly strategies [4]. Meanwhile, the disassembly system needs to recognize the human behavior intention to understand the disassembly operations and to

* Corresponding author.
E-mail address: jinhua.xiao@polimi.it (J. Xiao).

https://doi.org/10.1016/j.jmsy.2025.11.012
Received 26 March 2025; Received in revised form 6 November 2025; Accepted 12 November 2025

balance the disassembly tasks to improve the entire disassembly efficiency.

However, existing research on human intention recognition often overlooks the safety of disassembly operations, which poses certain disassembly risks to human operators during the actual disassembly process [5]. Moreover, the uncertain state of EOL products, including their structure and quality, during the recycling phase not only directly increases the complexity of disassembly tasks but also gives rise to the dynamic and non-standard nature of human operator actions. Unlike standardized assembly processes, disassembly under uncertainty requires human operators to frequently adjust their motions based on real-time product conditions, thereby triggering unpredictable changes in intent. Consequently, an intelligent disassembly system needs to accurately recognize human intention by capturing multimodal information such as human posture, disassembly tools, and facial data, enabling a comprehensive understanding of the dynamic disassembly behaviors of human operators. [6]. The intelligent disassembly system would then provide the feedback differently based on the specific human behavior intentions to accomplish the disassembly tasks, thereby combining HRC disassembly technologies to further improve the safety of collaborative disassembly operations [7]. In Section 2, the related literature has been explored to focus mainly on human intention recognition to understand the disassembly scenarios under the disassembly operations. In Section 3, an intention recognition method has been proposed to analyze the disassembly actions, tools, and face angles. The disassembly experiments have been demonstrated based on an example of retired EV batteries in Section 4. A deep discussion and comprehensive conclusion have been provided in Section 5 and Section 6, respectively.

2. Related works

2.1. EOL product structure disassembly

The widespread recycling of EOL products will inevitably involve the disassembly of EOL products under the safety and environmentally friendly disposal in a sustainable manufacturing of the circular economy [8]. The heterogeneity of EOL product structures designed by different manufacturers causes the complexity of robot-involved disassembly operations. To enhance the efficiency of disassembly, some scholars have explored the optimization of disassembly sequences for EOL product recycling. Zhang et al. [9] proposed a disassembly sequence planning based on a framework-subgroup structure for the disassembly graph. Xiao et al. [10] proposed a dynamic disassembly optimization approach based on a disassembly graph model to combine the forward-backward algorithms and Viterbi decoding. Baazouzi et al. [11] developed an optimization method for disassembly strategies by integrating disassembly sequence, disassembly depth, and optimal circular economy strategies. Xia et al. [12] proposed a HRC disassembly sequence planning method that integrates large language models with Dirichlet Bayesian networks, thereby enhancing the reliability and interpretability of sequential decision-making.

Furthermore, the related research has evolved to address multi-objective requirements, particularly for the safe disposal of waste electrical and electronic equipment. Disassembly sequence planning must not only consider efficiency but also prioritize the separation and treatment of hazardous components to minimize environmental and safety risks [13]. Zhang et al. [14] developed a mixed-integer linear programming models that harmonize human-robot collaboration with destructive disassembly to optimize both efficiency and resource recovery. Bahubalendruni et al. [15] proposed an automated planning based on CAD attributes to isolate toxins and maximize material reclamation. On the other hand, to support remanufacturing decision-making, some studies integrate disassembly sequences with component-specific attributes to generate disassembly paths that are not only efficient but also economically optimal. Kumar et al. [16] introduced a multi-level approaches incorporating bill of materials and stability matrices to generate economically optimal disassembly paths for remanufacturing. These studies have significantly enhanced the understanding and systematic optimization capabilities for disassembling complex EOL products.

To clearly demonstrate the disassembly process, the EOL product will focus on an example of EV battery pack structures of Tesla and BYD as shown in Fig. 1. As can be clearly observed from the Fig. 1, the disassembly processes for the two types of battery packs differ in terms of the steps involved, due to differences in module arrangement, types of mechanical fasteners (such as bolts and clips), and the application of sealants or adhesives. Furthermore, distinct disassembly tools and resources are required. It is precisely the structural heterogeneity among different EOL products that creates challenges for automating their recycling process.

From the perspective of automatic technological disassembly, Meng et al. [18] assessed the potential structure disassembly to balance the entire disassembly benefits by addressing uncertainty and safety issues. Lander et al. [17] conducted a comprehensive techno-economic assessment of the disassembly process, which identified product design and optimization strategies to save time and costs. Currently, key technologies for achieving intelligent disassembly include operation prediction, planning decision-making, and disassembly object detection. Due to the complexities and sustainability of AI-driven disassembly operations, Jan et al. [19] compared the automated disassembly operations with more feasibility for massive EOL product recycling. Wegener et al. [20] explored robotic disassembly combined with HRC to enhance the automation level of disassembly. Meanwhile, HRC disassembly not only addresses the flexibility of human operators to perform highly intelligent decision-making but also involves robot assistance to accomplish the highly intensive disassembly operations. However, while existing approaches have made strides in basic human-robot collaborative automation, they have yet to adequately address the dynamic interoperability between human cognitive capabilities and robotic precision under uncertain product structures, a capability that is crucial for enhancing the overall flexibility of collaborative workflows. Consequently, our research focuses on deepening the disassembly understanding of EOL products by intensively analyzing human involvement in HRC disassembly operations.

2.2. HRC disassembly safety

The HRC disassembly can be used to combine the strengths of both human operators and robots. A robot can provide the ability of high automation, precision, reliability, and adaptability for repetitive and tedious tasks, while human operators bring flexibility and intelligence to handle complex disassembly tasks [21]. Belhadj et al. [22] have explored the disassembly task allocation based on HRC operations, which are required to higher performances by humans for unique actions and the disassembly tasks. Laili et al. [23] researched the impact of timing strategies in human-machine interactions to discover the predictive timing strategies. Furthermore, it is necessary to pay more attention to the safety of HRC disassembly to ensure the safety of multi-agents and the environment. As shown in Table 1, this table summarizes key safety metrics in human-robot collaborative disassembly, highlighting problems, technical challenges, and representative methods for each aspect, based on recent research.

Furthermore, Simoes et al. [28] compared the robots with flexible linkages to apply less impact force during collisions, though they cannot fully prevent collisions with sharp end-effectors. Beyond physical safety measures, achieving true safety also depends on intelligent task planning and resource allocation. Safety-aware dynamic task planning is a key research focus to integrate hard safety constraints, such as maintaining safe human-robot distances and avoiding high-risk operations near personnel, directly into sequence optimization models [29]. Gao et al. [21] proposed a multi-agent strategy optimization method using

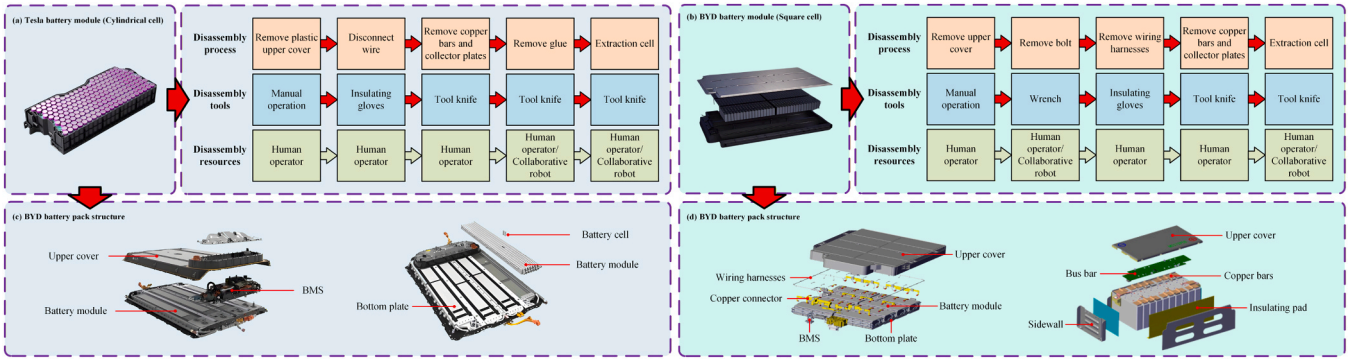

Fig. 1. The specific procedures of Tesla and BYD battery pack disassembly [17].

Table 1
The analysis of human operation safety metrics in HRC disassembly.

Safety metrics	Ref.	Problems description	Technical challenges	Representative algorithms
Collision Detection	Al et al. [24]	To prevent physical harm to human operators from direct contact with robots, especially sharp end-effectors.	– Difficulty in achieving zero false negatives while minimizing false alarms. – Sensing limitations in cluttered disassembly environments.	– Sensor-based Proximity Sensing. – Force/Torque Sensors – LiDAR-based Scanning
Trajectory Planning	PRAKASH et al. [25]	To generate robot paths that are not only efficient but also inherently safe, predictable, and non-threatening to humans.	– Balancing path optimality (time, energy) with stringent safety constraints (distance, speed). – Real-time re-planning in dynamic scenes.	– Potential Field Method – RRT (Rapidly-exploring Random Tree) – MPC (Model Predictive Control)
Task Planning	Asif et al. [26]	To optimally sequence disassembly operations and assign actions to humans or robot based on their capabilities and risks.	– Modeling the uncertainty of EoL product states and human actions. – Making sequential decisions that are both efficient and safe.	– POMDP (Partially Observable Markov Decision Process) – RL (Reinforcement Learning) – A* (A* search algorithm)
Resource Allocation	Lee et al. [18]	To dynamically assign resources and define the level of human-robot collaboration for each sub-task.	– Quantifying task complexity and risk in real-time. – Adapting allocation policies based on real-time human cognitive state.	– Game Theory Models – Auction-Based Methods – NSGA-II (Non-dominated Sorting Genetic Algorithm II)
Real-Time Interaction	Ma et al. [27]	To establish a natural and safe communication channel between human and robot for intuitive collaboration.	– Achieving low-latency, robust intention recognition without wearable sensors. – Ensuring interaction reliability under varying conditions.	– CNN (Convolutional Neural Network) – NLP (natural language processing) – AR (Augmented Reality)

partially observable deep reinforcement learning that simultaneously considers disassembly time, safety, and efficiency. Regarding resource allocation, beyond traditional task assignment, safety-oriented allocation of human and robotic resources is critically important [30]. This involves deciding whether a human or a robot should perform a task based on its complexity and risk level to maximize safety and efficiency [13]. It also includes determining the leading object in collaborative tasks to minimize the operator's physical load and risk [31]. By quantifying and optimizing safety metrics, the methods enable proactive risk avoidance at the system planning level, shifting from passive response to active prevention.

Unlike traditional industrial robots, the trajectory planning of collaborative robots should consider human safety as the most important precedence. Emrah et al. [32] proposed an intelligent system to explicitly detect the human presence in spatial reasoning for robotic activities. Xavier et al. [33] developed a smooth motion trajectory approach that ensures human safety by curbing the jerk, acceleration, and speed. Rahman et al. [34] improved the assistive control by introducing a novel control strategy based on gravity and load force characteristics. Furthermore, some scholars have explored real-time interaction to address the safety of HRC disassembly operations. Malik et al. [35] proposed digital twin platform to enable remote control to prevent robot operation accidents. Faisal et al. [36] considered the human posture using gesture recognition to control robots based on contact-based and vision-based detections. Ktistakis et al. [37] proposed a real-time human-robot collaboration framework integrating augmented reality and force-field guidance, which incorporates both safety policies and task strategies via a set of robot-assisted primitives. Similarly, vision-based recognition requires no wearable devices to recognize the HRC disassembly operations, which can improve the efficiency of HRC disassembly operations by combining both multi-mode and gesture recognition under human intention recognition [38]. While existing methods demonstrate notable effectiveness in mitigating physical risks, the overall systems remain reactive rather than predictive. Although HRC systems implement comprehensive pre-defined safety strategies, they fail to fully utilize the predictive potential inherent in human cognitive states and disassembly contexts. Therefore, our research will focus on human intention recognition to develop an HRC disassembly method suitable for the disassembly scenario understanding in the EOL product recycling.

2.3. *Human intention recognition*

Human intention recognition can be broadly discussed based on external visual action reasoning and facial recognition. Vision-based intention recognition currently focuses on determining robots and human behaviors to accurately understand human intentions, which can analyze various information with wide human actions [39]. Action-based recognition has significantly improved the performance of disassembly scenario understanding based on datasets by RGB-based

methods and skeleton-based methods [40]. The former can leverage prior knowledge from image recognition but requires substantial computational capability and storage. Similarly, the latter greatly reduces computational load and storage requirements, but it lacks useful contextual information for action recognition. Krestenitis et al. [41] accomplished high recognition rates by combining action recognition from different detection perspectives. Zhang et al. [42] conducted a comprehensive action recognition that mainly focuses on selecting appropriate image data to capture human action features. Therefore, it is necessary to recognize human behaviors with the superior performance of action features by various recognition methods as shown in Table 2.

In addition, intention recognition is to focus on action information with human involvement to improve its accuracy of HRC disassembly operations. Zhang et al. [48] considered both human operator behaviors and disassembly object information to classify the specific human intentions. Jang et al. [49] proposed the analysis approach of cognitive visual motion to capture the eye movements with a rich source of human intentions and behavior information. Singh et al. [50] proposed a model-based online intention recognition method to infer human intentions by combining the facial gaze. Schlenoff et al. [51] proposed a new intention recognition method based on the representation of state information in HRC environments by spatial relationships and fundamental direction information. Mohammadzadeh et al. [52] employed a head-mounted display to capture human motion data within a virtual reality environment, which was subsequently used to train a CNN-Transformer model, thereby enabling robots to effectively respond to human actions. The image recognition involves mapping patterns of any number of signal features to the desired postures, which can be used to learn the mapping based on LDA, SVM, and artificial neural networks (ANN) [53]. Kiguchi et al. [54] proposed an adaptive neuro-fuzzy modifier to learn the relationship between the measured root mean square (RMS) of the EMG signal and the estimation torque by human operators. Karayiannidis et al. [55] determined human intentions by measuring the contact force for the different limbs to generate different forces during various actions with sensor-wearing devices. Therefore, there are many challenges to deeply explore vision-based intention recognition between human operators and robots to understand the disassembly environment in the disassembly operations. While existing methods have made progress in predefined action recognition, they remain limited by passive response patterns and an inability to handle composite actions or multi-source data fusion efficiently. These gaps highlight the need for a real-time, multi-modal framework that integrates action, tool, and contextual cues. Based on the analysis above, although considerable progress has been made in the field of HRC disassembly, several key research gaps remain unaddressed, hindering the development of efficient, safe, and truly intelligent disassembly systems.

- Firstly, while action recognition has been extensively studied, existing work primarily focuses on classifying predefined actions rather than predicting human intent to enable anticipatory assistance from robots. Many current systems operate based on single-modality data, which often fails to distinguish between visually similar actions with different intentions in complex disassembly scenarios.
- Secondly, disassembly operation safety is often overlooked, but to solve the dynamic and uncertain disassembly that not only recognize operator intent but also perceive potential risks with unsafe postures or operator distraction.
- Thirdly, existing disassembly sequence planning methods often rely on static models and pre-defined workflows, which lack the adaptability to dynamic changes in human operator behavior and real-time environmental conditions that prevents efficient resource allocation and task reassignment.
- Furthermore, there is a lack of a real-time multi-modal data fusion framework capable of processing multi-source heterogeneous data (such as skeleton, visual and facial), which often compromises real-time performance for effective and safe HRC disassembly in EOL product recycling.

To bridge these gaps, it is necessary to propose a novel integrated framework for self-aware HRC disassembly in the complex disassembly scenarios, specifically designed to address the aforementioned limitations:

Table 2
The comparison of various recognition methods for human intentions.

Recognition methods	Ref.	Method description	Technical challenges	Representative algorithms
Vision-based (RGB)	Sharma et al. [43]	Captures RGB image via cameras and extracts motion features.	– High computational resource demands. – Susceptibility to background noise and lighting variations. – Dynamic occlusion issues. – Feature redundancy.	– Two-stream CNN (Spatiotemporal Feature Fusion). – I3D (3D Convolutional Networks).
Skeleton-based	Xin et al. [44]	Build human joint motion trajectories via skeletal key point tracking to analyze temporal action patterns.	– Lack of environmental context. – Difficulty capturing subtle motions. – Sensor dependency. – Poor adaptability to multi-object interaction scenarios	– ST-GCN (Spatio-Temporal Graph Convolutional Networks). – PoseC3D (3D keypoint encoding).
Multimodal Fusion	Shaikh et al. [45]	Combines RGB, skeleton, depth sensors, or other data sources to enhance robustness via feature-level/decision-level fusion.	– Synchronize multi-source data. – Inter-modal redundancy. – Real-time optimization challenges. – Heterogeneous data representation.	– MM-Action (Multimodal Action Dataset). – Transformer-based cross-modal attention.
Behavioral Sequence Modeling	Mazzia et al. [46]	Build probabilistic models based on temporal action relationships to infer intent with domain knowledge.	– Complexity in long-term dependency modeling. – Difficulty embedding knowledge. – Poor adaptability to dynamic scenes.	– LSTM/GRU temporal networks. – TCN (Temporal Convolutional Network).
Few-Shot Learning	Tu et al. [47]	Addresses data scarcity in industrial disassembly via meta-learning or transfer learning for intent recognition.	– Limited cross-task generalization. – Feature disentanglement challenges – Negative transfer risks	– Prototypical Networks. – MAML (Model-Agnostic Meta-Learning).

- Firstly, it is necessary to construct an integrated framework that synergistically combines three perceptual streams: (a) spatiotemporal action recognition using an ST-GCN model to accurately interpret human motion from skeletal data; (b) disassembly tool detection based on an improved YOLO algorithm to infer intent from contextual tool usage; and (c) facial angle recognition to estimate the operator's attention. The concurrent processing of these streams provides a rich, cross-validated understanding of operator intention, thereby reducing ambiguity.
- By incorporating facial angle and posture analysis (e.g., identifying leaning-forward behavior), the system can infer the operator's cognitive state and awareness of the robot's presence. This enables the system to anticipate potential unsafe interactions before they occur, allowing the robot to take proactive safety measures, such as reducing speed or pausing movement, thus addressing the current safety gap in HRC disassembly research.
- Accurately recognizing operator intent can provide a decision-making basis for intelligent dynamic resource allocation. For example, when the system identifies a specific disassembly intent, such as loosening a bolt, it can proactively activate an optimal resource allocation strategy where the robot, leveraging its high-precision capability, is assigned repetitive twisting operations, while the human operator focuses on adjacent tasks requiring flexibility and judgment.
- The proposed system is designed for real-time performance by selecting the ST-GCN and improved YOLO algorithms, along with the design of the lightweight facial recognition network ShuffleNetV2. Multi-modal data fusion does not introduce excessive latency, making the framework suitable for practical disassembly applications where timely responses are critical.

In the following sections, we will elaborate on the overall architecture of the HRC disassembly system, the specific implementation of each algorithm, and explain how the three types of data are fused to infer operator intent.

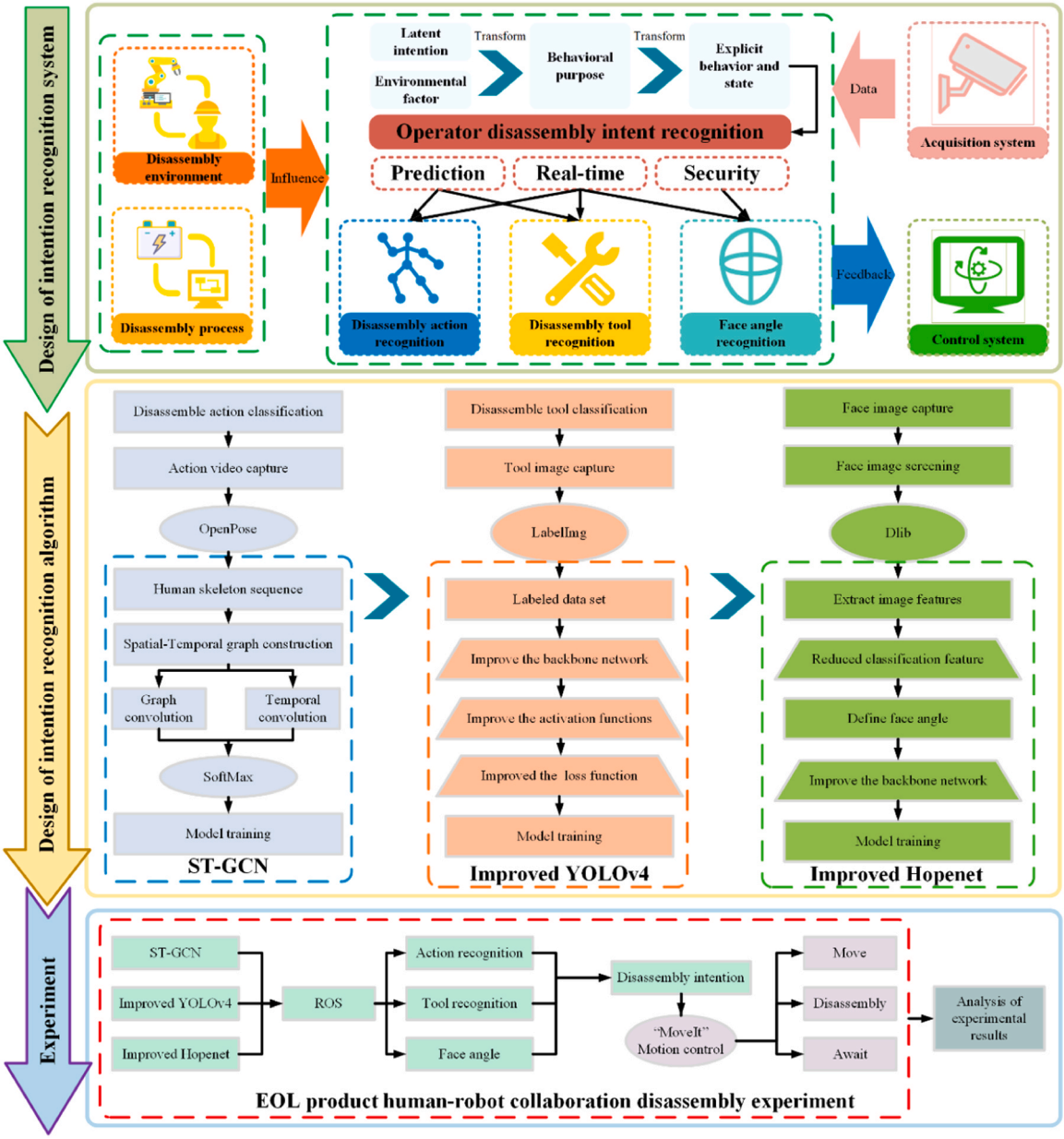

Fig. 2. The overall framework of human behavior and intention recognition system based on HRC disassembly.

3. Methodology

The disassembly and recycling of an EV battery pack can be regarded as a typical case study. This scenario involves achieving accurate recognition of human disassembly intentions and enabling intelligent collaborative disassembly by robots within a shared human-robot workspace, thereby providing support for HRC-based processing of various EOL products. However, this process faces inherent challenges. Specifically, the uncertain state of EOL products forces human operators to adopt differentiated disassembly strategies, causing their intention mapping to deviate from predefined procedures. Furthermore, traditional unimodal recognition systems cannot capture such context-dependent variations as they fail to distinguish between superficially similar actions with different intentions, making the collaborative process either unfeasible or hazardous. To overcome the challenges of intent recognition, an overall framework of disassembly intention recognition has been proposed to deeply explore the disassembly scenario understanding as shown in Fig. 2.

- Firstly, human action recognitions are essential for understanding disassembly intentions, such as a pickup motion for bolt disassembly or adhesive removal, which help robots anticipate and assist the disassembly tasks.
- Secondly, disassembly tool recognition can be used to accurately identify the disassembly actions by detecting the specific disassembly tools (i.e., a wrench for the bolts or a chisel for adhesive removal), while it can also infer the human action.
- Finally, disassembly safety can be considered to incorporate human face angle recognition, which determines human attention to distinguish between conscious and unconscious actions in human-robot interaction.

3.1. Disassembly action perception

Action recognition essentially involves classifying human motion data, which can be viewed as converting a video into multiple RGB frames to extract the features. The implementation of action recognition includes skeleton-based detection and RGB video-based detection. Firstly, a video segment can capture the action representation dataset to represent the action features, such as pose-based (skeleton points) or interest-point-based extraction. However, traditional behavior recognition features can typically be extracted first by classification using a classifier, such as the improved dense trajectories (IDT), while the optical flow field can capture trajectories in a video sequence to extract the features, including Histogram of Optical Flow (HOF), Histograms of Oriented Gradients (HOG), and Motion Boundary Histograms (MBH). Based on grayscale images and other features with dense optical flow, it is necessary to encode the features using the FV (Fisher Vector) and SVM classifier.

Furthermore, deep learning methods include skeleton-based detection and RGB video-based detection. Skeleton-based detection involves obtaining skeletal data from camera-based skeleton detection technologies or pose estimation algorithms to determine the behaviors. By comparing with the classical methods (i.e., Deformable Pose Traversal Convolution, etc.), the popular graph convolution methods (i.e., ST-GCN, RGB video-based detection, etc.) can be used to capture the spatiotemporal features from the set of RGB video frames. Moreover, the common architecture includes two-stream, 3D convolutional, and CNN-LSTM approaches, while the two-stream architecture offers higher accuracy but requires additional optical flow extraction with slower calculation. The 3D convolutional architectures are usually end-to-end to provide faster processing but slightly lower accuracy. Human action recognition needs to meet the related requirements for the recognition accuracy, algorithm robustness, and real-time performance. Based on the characteristics of human action recognition, the ST-GCN algorithm combined with deep learning will be a potential candidate for real-time action recognition.

3.1.1. Disassembly action collection and data pre-processing

During a complete human-robot collaborative disassembly process, human operators perform a diverse range of disassembly actions. Moreover, human operators may suffer from fatigue or other negative conditions during the disassembly operations, which might not only exhibit incorrect postures but also lean forward with an action that significantly causes the operation collisions. To ensure the safety of disassembly operations, it is important to define the leaning-forward behavior as a hazardous action by considering the operational actions for EOL products (i.e.: EV battery pack), including (1) removing end cover, (2) removing the components, (3) picking up disassembly tools, (4) loosening the bolt connection, (5) removing the nuts, (6) removing the adhesive, and (7) leaning-forward behavior with the action labels assigned as *operation 0, operation 1, operation 2, operation 3, operation 4, operation 5, and operation 6*, respectively. The testing videos related to each action can be recorded as the final recognition sample. For these seven categories, we collected 100 video samples per class under varying backgrounds and operator conditions, resulting in a total of 700 video samples. Each video was then trimmed into 5–6 s action segments containing approximately 150 consecutive frames, forming the raw dataset for subsequent processing.

As shown in Fig. 3, it illustrates the complete workflow of human intention recognition during disassembly operations. First, the raw video data captured from the disassembly process of an EOL product is used as input. These data then enter a critical preprocessing stage, where the OpenPose pose estimation algorithm is employed to extract human skeletal sequences, with a focus on key joints such as the neck, shoulders, and elbows. The refined skeletal data points are subsequently used as input to the ST-GCN model. The ST-GCN processes these structured spatiotemporal data to learn and classify complex human motion patterns. When the final action recognition confidence output from this process meets the task's required threshold, the result is utilized for fused intention recognition within the overall system. However, if the result falls below the threshold, it indicates an issue in the model's prediction process, potentially caused by the model encountering unfamiliar features. This necessitates enhancing dataset diversity by incorporating samples that cover different operator actions, environmental conditions, and edge cases, alongside the appropriate application of data augmentation techniques. By adapting the dataset to the model's existing architecture, we can preserve the generalization capability of the ST-GCN, maintaining its transferability to unfamiliar disassembly scenarios and preventing overfitting.

The OpenPose network structure defines an image of size w×h× 3 (a standard RGB image) as input, which can be handled through the 10 layers of the VGG-19 network to generate a set of feature maps. The main network consists of two large modules. The former is the Part Affinity Fields (PAF) module to generate a set of vectors. The latter is the confidence map module to train the detection of key body parts for each person. The confidence map represents the probability of each pixel for a specific body part. For the PAF network, the network can be divided into many stages T_p as a hyperparameter. The network only inputs the value F, while the network description of the first stage can be denoted as $\varnothing^1$ and its output can be defined as L^1.

$$L^1 = \varnothing^1(F) \tag{1}$$

When the network module runs for the second time, the output from the previous stage L^1 can be merged with the original input feature map F to start the next stage. However, sequential manners can represent a layer-wise propagation that effectively creates a layer PAF training module.

$$L^t = \varnothing^t\left(F, \quad L^{t-1}\right), \quad \forall 2 \le t \le T_p \tag{2}$$

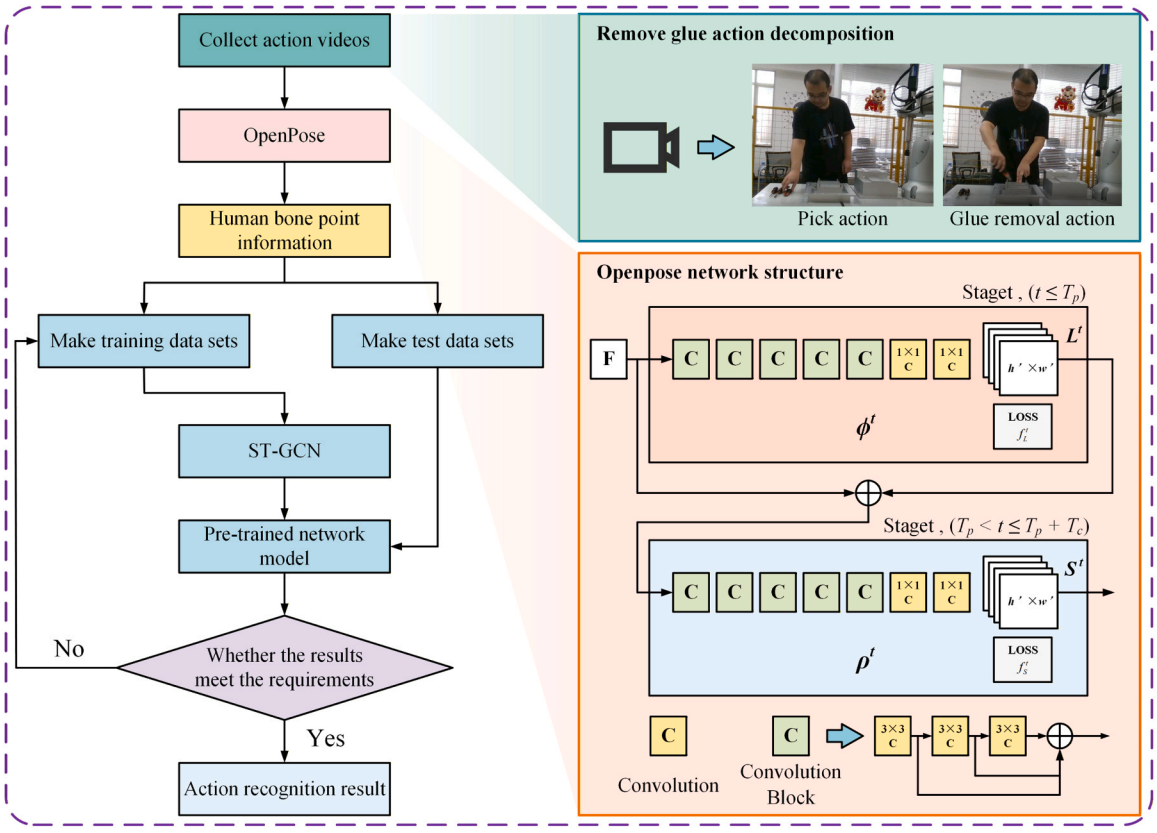

Fig. 3. The procedure of human intention recognition for disassembly operations.

Where T_p represents the layer number of the PAF training model. For the entire module, the formula can be expressed as follows:

$$S^{T_p} = \rho^t\left(F, \quad L^{T_p}\right), \quad \forall t = T_p \tag{3}$$

The PAF network module includes 5 convolution blocks and two 1 × 1 convolution layers, along with an L2 loss function. Each convolution block contains three 3 × 3 convolution layers. The output of each convolution layer can be transmitted not only to the next layer but also directly to the end of the convolution block. For the confidence map module, the intermediate architecture is completely the same, but there are two key differences: the confidence map module takes the PAF output from the previous module as input, which addresses the low precision for identifying the key human parts due to a lack of contextual determination. The recently trained PAF can be used to enhance the data through the next stage of training. Additionally, the current stage output can be regarded as input for the next stage, allowing the training of each confidence map (except for the first stage) to occur within a coherent framework. The iterative prediction architecture is inherited from Convolutional Pose Machines (CPM).

$$S^t = \rho^t\left(F, \quad L^{T_p}, \quad S^{t-1}\right), \quad \forall T_p < t \le T_p + T_c \tag{4}$$

Where, ρ^t represents the CNN inference at the stage t, while T_c indicates the total number of confidence map stages. By considering the modules, it is important to extract the confidence and association information from the video. However, it is necessary to employ the bipartite matching from graph theory to determine the partial association that connects the key points of human operators. Due to the vector of the PAF, the bipartite matching is highly accurate to merge into a complete skeleton for the operator.

3.1.2. Human action recognition based on ST-GCN

A graph structure can be constructed based on the skeletal sequence information from the action video as shown in Fig. 4(a). Based on the graph structure, each node related to a key point can represent the connections between joints. The joint coordinate vectors of the nodes are then used as input data for the ST-GCN to capture the spatial layout information of the skeleton. By considering a series of spatiotemporal graph convolution operations, it is necessary to extract high-level features from the input data. The output of the OpenPose serves as the input of the ST-GCN based on a batch of videos, which can represent the data using a 5-dimensional matrix $(\mathbf{N}, \mathbf{C}, \mathbf{T}, \mathbf{V}, \mathbf{M})$. where, **N** represents the number of videos, **C** denotes the features of the key points, **T** indicates the number of keyframes, **V** is the number of key points, and **M** represents the number of the human operator.

For the ST-GCN model, it is necessary to ensure consistent analysis for different nodes by maintaining uniform input data through batch normalization. The data can be processed by the 9 ST-GCN layers, the 3 layers output with 64-dimensional features, the next 128 dimensions, and the final 256 dimensions. A ResNet mechanism can address the gradient vanishing, while the 50 % dropout rate is implemented in each layer to enhance generalization and prevent overfitting. Pooling operations are used in the 4th and 7th layers to reduce spatial dimensions and extract higher-level features. To convert the graph data classification, it is necessary to handle a readout operation with fixed-dimensional vectors with a Softmax classifier, which enables the ST-GCN model with graph-structured data to achieve high-performance classification as shown in Fig. 4(b). Similarly, it is necessary to construct a spatiotemporal graph $G = (V, E)$ as shown in Fig. 4(c), where V denotes the node features and E denotes the edge features. Specifically, the node set V represents the features of the i-th key point in the t-th frame, where t iterates over all frames (From $t = 1$ to T) and i iterates over all key points within the same frame (from $i = 1$ to N). However, we can identify and track the same nodes across consecutive frames to build temporal information, which focuses on the spatial relationships between different key points within the same frame. By

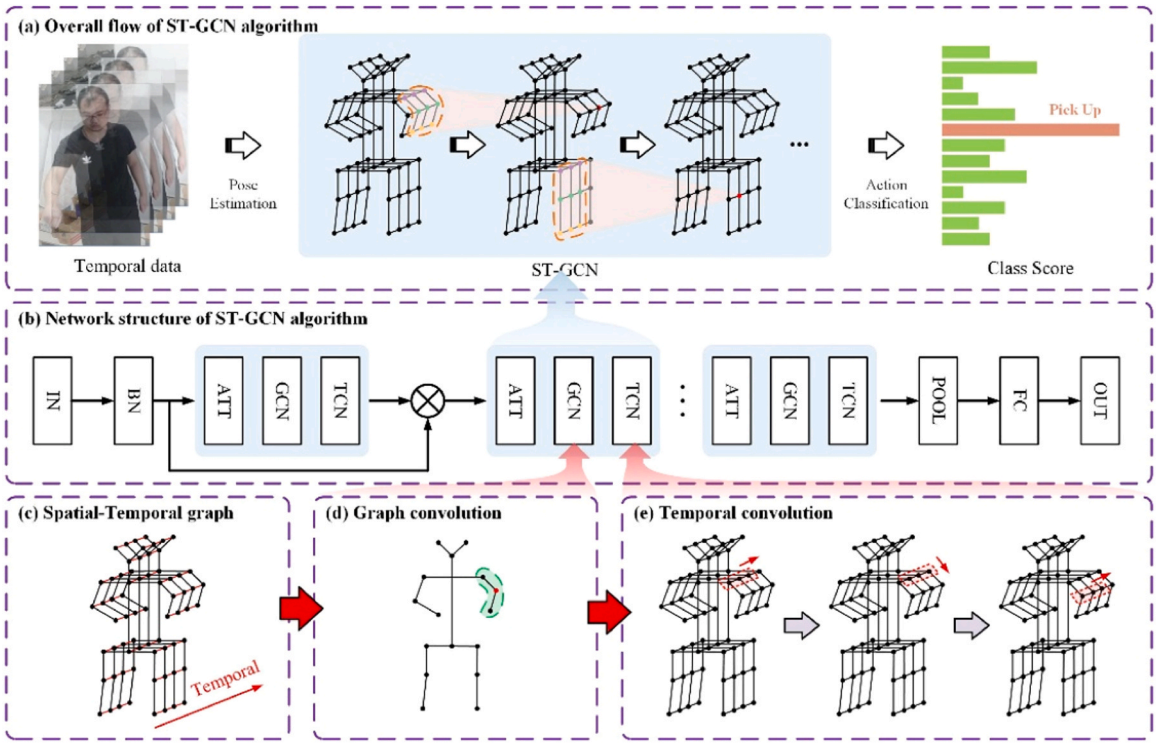

Fig. 4. The overall procedure of human action recognition based on ST-GCN.

capturing the relative positions and posture of key points, the intrinsic constraints and dynamic human body structure can be described.

$$E_s = \{v_{ti}v_{tj}|(i,j) \in H\} \tag{5}$$

Furthermore, the second subset focuses on the connections between the same key points across different frames. These temporal connections reveal the continuity and smoothness of human movement by tracking the trajectories of the key points.

$$E_F = \{v_{ti}v_{(t+1)i}\} \tag{6}$$

Based on the key point information of human bodies in consecutive frames from a video, it is necessary to construct a graph structure that integrates spatial and temporal information. Similarly, the ST-GCN model can compute the graph that extracts the features through a series of spatiotemporal convolutions from convolution networks.

$$f_{out}(x) = \sum_{k=1}^{K}\sum_{\omega=1}^{K} f_{in}(p(x,h,\omega)) \times w(h,\omega) \tag{7}$$

However, the sampling function p transforms into selecting specific nodes and their neighboring node sets for feature information extraction. The weight function w provides the weight matrix for dot-product calculations. To accommodate the unique data structure of the spatiotemporal graph, it is necessary to redefine the sampling function p and the weight function w. Specifically, the sampling function must consider the spatial and temporal connectivity between nodes to ensure that the spatiotemporal dependencies of node features can be captured.

By comparing with the traditional convolutional neural networks, the sampling function can be reflected through the size of the convolutional kernel to define the local region. For example, with a 3×3 convolutional kernel, the convolution to a specific pixel involves a weighted aggregation of its feature values with those of the surrounding neighboring pixels. In spatiotemporal graph convolution, the sampling function must be carefully designed to effectively extract feature information from nodes and their neighborhoods, which involves determining the neighborhood range of the nodes. The sampling process not only involves extracting features from neighboring nodes but also requires consideration of the connections and weight distribution between nodes. By redefining the sampling function, it is necessary to better capture the local features of nodes in the spatiotemporal graph. In this study, we set the neighboring nodes to represent the distance d as follows.

$$\mathrm{B}(v_{ti}) = \{v_{ti}|d(v_{ti},v_{tj}) \le \mathrm{D}\} \tag{8}$$

When $D = 1$, the sampling function p can represent the sampling of adjacent nodes.

$$\mathrm{P}(v_{ti}, \quad v_{tj}) = v_{tj} \tag{9}$$

As shown in Fig. 4(d), the directly connected neighboring nodes to the center node are considered to directly influence its feature representation. In the traditional neural networks, the weight function is typically implemented by indexing a tensor of dimensions that has a specific spatial order, which represents a convolution kernel to define the weights. Therefore, by considering the convolution operations on spatiotemporal graphs, it is necessary to arrange the weight function that can accommodate the characteristics of the graph structure. A new ST-GCN approach can be used to partition the neighboring nodes of a certain node in the spatiotemporal graph that assigns a unique label to each subset by effectively implementing the mapping.

$$l_{ti} : B(v_{ti}) \rightarrow \{0, \quad \ldots, \quad K-1\} \tag{10}$$

In addition, the weight function can be implemented by directly indexing a (c,k) tensor and the following formula, where $l_{ti}(v_{tj})$ denotes the label of v_{tj} in the subset when v_{ti} is used as the center node.

$$w(v_{ti}, \quad v_{tj}) = w'(l_{ti}(v_{tj})) \tag{11}$$

The traditional convolution can be updated to the spatiotemporal graph:

$$f_{out}(x) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(p(v_{ti}, \quad v_{tj})) \times w(v_{ti}, \quad v_{tj}) \tag{12}$$

Where $Z_{ti}(v_{tj})$ is the normalization term. By incorporating the previ-

ously mentioned sampling function and weight function, which can be obtained by the spatial graph convolution operation.

$$f_{out}(x) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \times w'(l_{ti}(v_{tj})) \tag{13}$$

Where, v_{tj} represents the neighboring nodes, and $l_{ti}(v_{tj})$ denotes the label associated with the neighboring nodes. For the spatiotemporal graph based on ST-GCN, the graph convolution can be further optimized by adding temporal constraints, while the neighboring nodes of the spatiotemporal graph are defined.

$$\mathrm{B}(v_{ti}) = \{v_{qj} | d(v_{tj}, \quad v_{ti}) \leq K, \quad |q-t| \leq \lfloor \Gamma/2 \rfloor\} \tag{14}$$

However, the ST-GCN proposes a new definition of neighboring nodes, which comprehensively considers both the spatial distance between nodes and their temporal continuity. By introducing temporal constraints, it is necessary to accurately identify the set of nodes that are adjacent to a given node in both spatial and temporal contexts. For the temporal graph convolution, a set of sampling and weight functions needs to be defined to focus on the characteristics of nodes to define the mapping function, which are partitioned based on their temporal proximity. The feature aggregation and weight distribution remain consistent with spatial graph convolution:

$$l_{ST}(v_{qi}) = l_{ti}(v_{tj}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K \tag{15}$$

Where $(q-t+\lfloor \Gamma/2 \rfloor) \times K$ represents the temporal label grouping. Through this approach, the sampling function and weight function can be effectively applied to temporal graph convolution to capture the dynamic changes of nodes by combining both GCN and TCN to extend the graph convolution approach.

$$f_{out}(x) = \Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}} f_{in} W \tag{16}$$

Where $\Lambda^{ii} = \sum_j (A^{ij} + I^{ij})$ represents the adjacency matrix of the spatiotemporal graph, I represents the identity matrix, $A+I$ represents self-loops to the spatiotemporal graph to ensure effective data transmission. W consists of weight vectors for multiple output channels. f_{in} is the input feature map with dimensions (c, V, T), where V is the number of nodes and T is the number of frames. For the partitioning strategy with multiple subsets, the adjacency matrix A is divided into multiple matrices A_j, namely $A + I = \sum_j A_j$. The graph convolution can be described.

$$f_{out}(x) = \sum_j A_j^{-\frac{1}{2}} A_j A_j^{-\frac{1}{2}} f_{in} W_j \tag{17}$$

Where $\Lambda^{ii} = \sum_k A_j^{ik} + \alpha$ represents the aggregation matrices. For human motion analysis, the joints often exhibit group movement, particularly in the linkage between different body parts, such as the wrist and elbow. Similarly, it is necessary to accurately describe and predict human behavior during the movement that enhances the accuracy of related scenario predictions, such as motion analysis and action recognition. Additionally, an attention mechanism provides each adjacency matrix with the internal connection relationships based on a learnable matrix M, which can modify $(A+I) \otimes M$ by performing element-wise multiplication with M. In the ST-GCN model, traditional convolutional layers perform temporal convolution operations using temporal data with the feature map, including the joint features (C), key frames (V), and joints (T), image convolution channel (C), the width (W), and the height (H). For image convolution, a kernel size of [w] × [1] covers w rows and moves with a stride of s. In temporal convolution, a kernel size of [k] × [1] processes one node and k consecutive key frames with a stride of 1. The model capability of temporal data can support tasks like motion analysis and action recognition as shown in Fig. 4(e).

Furthermore, the spatiotemporal modeling capability of ST-GCN provides a degree of tolerance to partial joint occlusions. This is achieved by leveraging global skeletal topology and temporal dependencies to infer actions, enabling the model to predict activities using information from adjacent frames and joint relationships even when certain joints are obscured. However, in highly cluttered environments, background clutter can interfere with the accuracy of skeletal point detection. This is particularly evident when occlusions cause failures in pose estimators OpenPose, leading to increased noise in the input data. Additionally, while dynamic occlusions may cause interruptions in action sequences, the graph structure of ST-GCN can partially compensate for missing information by utilizing data from neighboring nodes.

Based on the decomposed action dataset, the ST-GCN model can be trained. During training, data augmentation should be enabled using random_choose and random_move, with a window size of 150 frames. Simultaneously, a spatio-temporal graph convolutional network based on the OpenPose skeleton topology is employed, integrated with an edge importance weighting mechanism to enhance feature representation. The training process utilizes single-GPU parallel processing with a batch size of 32 samples. A step decay strategy is adopted, with an initial learning rate of 0.1, which is adjusted at the 20th, 30th, 40th, and 50th epochs. The entire training spans 50 epochs. This configuration ensures sufficient model convergence while maintaining training stability.

3.2. Disassembly tool detection

The technical approaches to object detection can be broadly categorized into those based on traditional machine learning paradigms and those leveraging deep learning. In computer vision, traditional methods often rely on hand-crafted feature extraction combined with classifiers, as seen in algorithms like HOG+SVM and Haar Cascade. In contrast, modern deep learning-based approaches utilize CNN to automatically learn hierarchical features from data, leading to architectures such as R-CNN, YOLO, and RetinaNet, which achieve superior accuracy and speed. For the HRC disassembly system, the capability for real-time multi-object recognition is essential. The YOLO algorithms have the advantages of their real-time performance and accuracy, which redefine object detection as a regression by a single CNN over the entire image to predict the probabilities and bounding boxes. It is remarkably 1000 times quicker than R-CNN in processing real-time video with less than 25 ms of latency to achieve high accuracy, while minimizing the deviations through its end-to-end deep learning approach.

3.2.1. Data preprocessing for tool detection

Before creating the disassembly tool dataset, it is crucial to identify the tools required for specific disassembly tasks, such as a wrench for bolt removal or the appropriate tool for adhesive clearance. Additionally, since various tools are often used together, the recognition algorithm must also identify the hands holding these tools by measuring the distance between the detected hand and tool bounding boxes. Therefore, the classification mainly includes the disassembly tools and hands, which analyze the overlapping detection from the YOLO algorithm. It is necessary to annotate the images of the tools by operator hand using LabelImg software. After processing the images, it is necessary to enhance a data augmentation package and the dataset into training, testing, and validation sets, which can be used to ensure the dataset quality and the algorithm effectiveness. As shown in Fig. 5, the YOLO algorithms consider the structure of the PASCAL VOC dataset, while the Annotations folder contains the corresponding annotation files (i.e.,.xml files) for each image. The Image sets store the text files listing the categorized image names, including predefined training and testing set splits. The JPEG images folder holds all original images in.JPEG format. The segmentation classification preserves segmentation categories for pixel-level annotations, while the segmentation object stores segmentation mask files. For object detection tasks, only the annotations and JPEG images folders are utilized to ensure high accuracy in the final recognition results, which is essential to collect many images. Therefore,

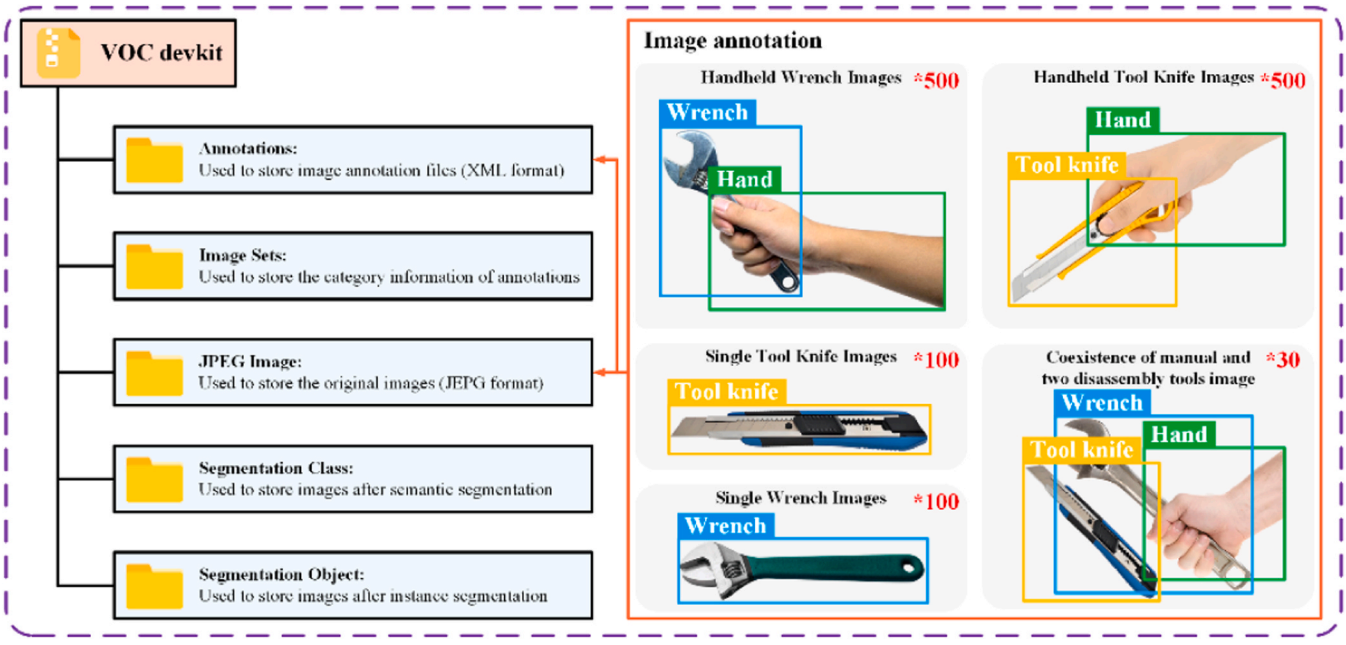

Fig. 5. Dataset catalog structure and image annotation for disassembly tools.

it is necessary to capture multiple videos of disassembly tools and extract frames from the videos for dataset creation.

In the experiments, we have gathered 500 images of a hand-held wrench, 500 images of a hand-held tool knife, 100 images of a single tool knife, 100 images of a single wrench, and 30 images featuring both hands and two disassembly tools, resulting in a total of 1230 images for the dataset. To enhance the robustness of the disassembly tool recognition, each image features a different background. Moreover, the training dataset intentionally incorporates images with overlapping and occluded scenarios, such as a human operator holding a tool, which closely mirrors real-world disassembly conditions. A key characteristic of these images is that portions of the tools are partially occluded by hands, other objects, or complex backgrounds. Including such images effectively enhances the model's detection robustness in cluttered or partially occluded environments. This approach enables the model to learn to identify key features and contextual cues even when visibility is significantly reduced to 50–70 %, thereby reducing false negatives and improving localization accuracy through exposure to diverse visual patterns.

3.2.2. *Vision-based disassembly tool detection*

The vision-based recognition algorithms (i.e., YOLOs) achieve end-to-end object detection by constructing a single convolutional neural network, which tackles the entire images as the input with the bounding boxes and their associated classification. Raw images are uniformly resized to 448 × 448 pixels from fixed-dimensional input vectors in the fully connected layers in the convolutional layers. YOLO produces a tensor of dimensions 7x7x30 for object detection. The 30-dimensional vector includes the positions and confidence of 2 bounding boxes. Each bounding box requires 4 values to indicate its position with x and y coordinates of the center point, totaling 8 values for 2 bounding boxes. The confidence of a bounding box can be represented:

$$Confidence = \Pr(Object) \times \mathrm{IOU}_{pred}^{truth} \tag{18}$$

Where $\Pr(Object)$ is the probability of an object within the bounding box. The loss function for YOLO is defined as follows:

$$\mathrm{Loss} = \lambda_{coord}\sum_{i}\left((x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2+(w_i-\hat{w}_i)^2+(h_i-\hat{h}_i)^2\right)$$
$$+\sum_{i}\lambda_{coord}\left((C_i-\hat{C}_i)^2\right)+\sum_{i}\lambda_{coord}\left((P_i-\hat{P}_i)^2\right) \tag{19}$$

Where (x_i, y_i) represent the predicted center coordinates of the bounding box, (w_i, h_i) are the predicted width and height, C_j is the confidence score for the predicted bounding box, and P_k is the predicted probability for class k. The λ terms are weights that control the relative importance of the coordinate, confidence, and class probabilities in the overall loss. This structure allows YOLO to effectively balance the various detections during the model training as shown in Fig. 6.

Furthermore, the real-time performance of practical production is a critical factor to enhance recognition accuracy, which often consume the computational resources of disassembly analysis. Throughout the object detection process, the YOLO effectively reduces feature map loss through a series of carefully designed optimizations. However, convolution and down-sampling, as a key step of network structure, inevitably result in the loss of important information for feature extraction. Convolution operations can capture the local features by sliding filters over input data to detect spatial relationships, while the resolution of feature maps can diminish. To preserve the feature maps and effectively reuse critical feature information, the CSP1-n residual block can be utilized, which consists of two main components with the backbone network as shown in Fig. 7(c). Firstly, input data can be analyzed to extract the initial feature via the main residual unit, while two 1 × 1 convolution operations adjust the number of feature channels to meet further processing needs. A 3 × 3 convolution operation is performed to capture and refine spatial features to enhance the depth and representation of the features. For the CSP1-n block, the outputs of the backbone network and the residual path are effectively fused to enrich the information of the feature layers that enable the network to learn more feature representations.

As shown in Fig. 7(a), the CSP1-n module replaces the CSP8 and CSP4 in the Backbone to enhance feature extraction and propagation capabilities. To further improve feature transfer across the layers, dense

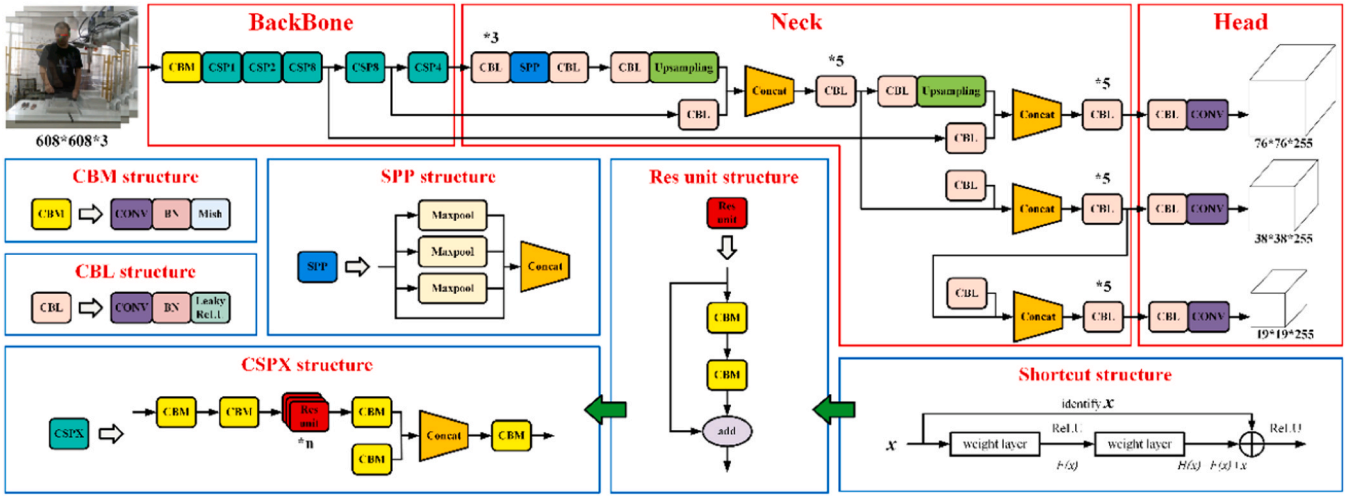

Fig. 6. The vision-based detection network structure of HRC disassembly.

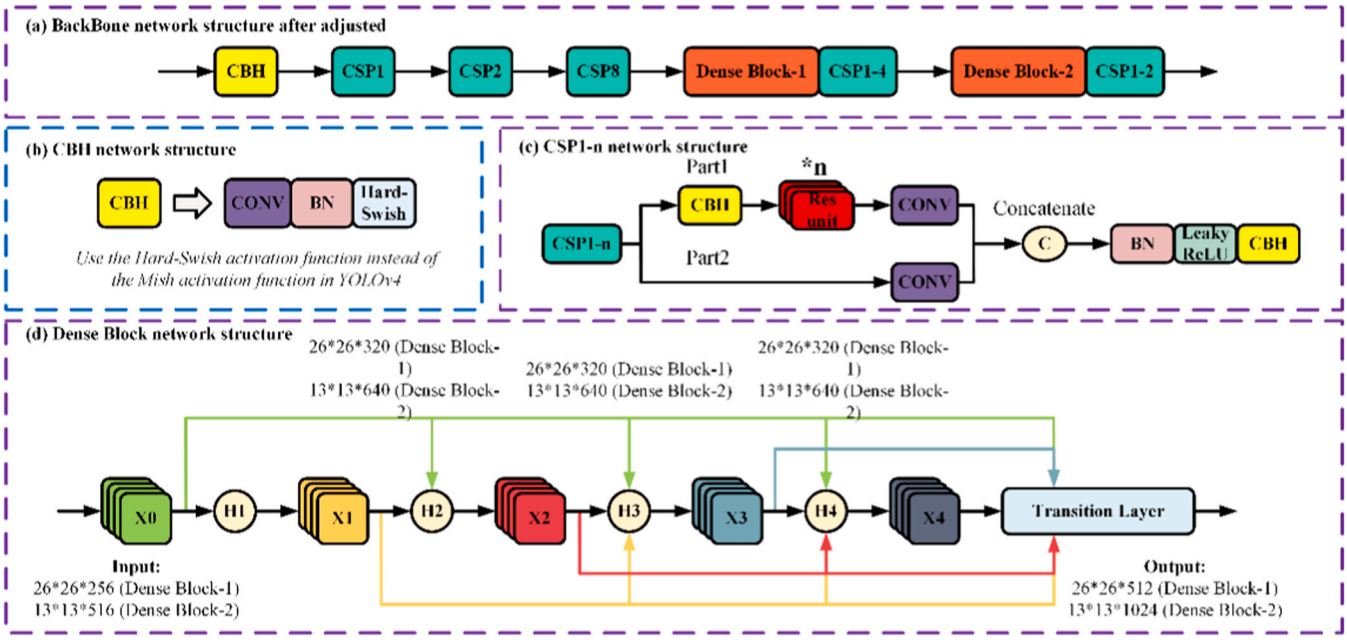

Fig. 7. feature recognition and object detection network structure.

connection blocks are introduced by replacing the last two residual blocks (CSP8 and CSP4) in the original CSPDarknet53 with Dense-CSP1–4 and Dense-CSP1–2. Additionally, the network structure is streamlined by removing redundant CSP-n blocks, which retains better performance to enhance the computational speed for the Dense Block and network parameters as shown in Fig. 7(d). For the Dense Block-1, feature propagation and layer concatenation can be conducted with a resolution of 26 × 26 by integrating multi-layer feature information, which can be represented by a forward-propagated feature layer with dimensions of 26 × 26 × 512. Similarly, for the Dense Block-2, feature propagation and layer concatenation can be performed with a resolution of 13 × 13 that yields a forward-propagated feature layer with a resolution of 13 × 13 × 1024. The configuration of Dense-CSPDarknet53 ensures that image information of network training can be effectively inherited by subsequent feature layers. Experimental results indicate that after adjusting the backbone network, accuracy improved by 10.6 %. To enhance recognition performance during the disassembly operations, the *Hard-Swish* activation function has been used to improve computational speed with *CSPDarkNet*.

$$HardSwish(x) = x \times HardSigmoid(x)$$

$$= x \times \frac{ReLU(x+3)}{6} = x \times \begin{cases} 0, & x \le -3 \\ 1, & x \ge 3 \\ \frac{x}{6} + \frac{1}{2}, & -3 \le x \le 3 \end{cases} \tag{20}$$

Compared to traditional activation functions, *Hard-Swish* maintains excellent nonlinear characteristics while simplifying the computation process, which demonstrates the notable accuracy improvements on the datasets to provide more reliable detection results for detection tasks. Furthermore, the localization algorithm within the loss function significantly impacts the average accuracy of the model detection results by selecting an appropriate loss function for detection performance. The YOLO algorithm commonly utilizes CIOU as its loss function for training and optimizing the model, as it comprehensively considers the overlap, distance, and size differences to describe the similarity between the

predicted and the truth boxes. Similarly, EIOU addresses this limitation by independently calculating the length and width of both the predicted and the truth boxes, allowing for a more accurate reflection of their matching.

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp}$$

$$= 1 - IOU + \frac{\rho^2(b, \quad b^{gt})}{c^2} + \frac{\rho^2(w, \quad w^{gt})}{c_w^2} + \frac{\rho^2(h, \quad h^{gt})}{c_h^2} \tag{21}$$

Where c_w and c_h refers to the width and height of the smallest enclosing box that covers both the predicted box and the truth box. The EIOU loss function highlights its superiority in detection tasks, including three core components between the predicted box and the truth box with the overlap loss (L_{IOU}), the center distance loss (L_{dis}), and the width and height loss (L_{asp}). Meanwhile, the loss function L_{IOU} and L_{dis} presents the loss and center distance with the classic strategy to ensure precise alignment of the predicted box and the truth box. For the width and height loss, EIOU abandons the relatively complex methods compared with CIOU to directly minimize the absolute differences in the width and height between the predicted box and the truth box, while the simplified design not only enhances computational efficiency but also converges more quickly when optimizing the aspect ratio of the predicted box.

During the training of the YOLOv5 model, the hyperparameters were finely tuned based on the detection task to ensure high performance and stability in object recognition. The training was conducted for 300 epochs with a batch size of 32 and an image size of 640 × 640 pixels. GPU parallel computing was employed, along with 8 data-loading worker threads to optimize data throughput efficiency. The initial learning rate was set to 0.01, using a step decay strategy, with a final learning rate of 0.01. This was combined with a momentum of 0.937 and a weight decay of 0.0005 to enhance gradient stability and generalization capability. A warm-up phase of 3 epochs was implemented to smoothly transition through the initial training instability. Data augmentation strategies were also enabled, including HSV color space adjustments, translation, scaling, and Mosaic augmentation, with RandAugment and random erasing introduced to improve model robustness against occlusion and lighting variations. The loss function weights were configured as follows: bounding box regression 7.5, classification 0.5, and distribution focal loss 1.5. Label smoothing was also applied to mitigate overfitting. The overall design aims to balance the training speed, accuracy, and adaptability, laying a foundation for subsequent real-time object detection applications.

The final training results can be described as shown in Fig. 8. The YOLO model exhibited favorable convergence characteristics and a clear performance improvement trend over the 300 training epochs. As observed from the loss curves, the bounding box loss, classification loss, and distribution focal loss for both the training and validation sets decreased rapidly within the first 100 epochs before gradually stabilizing, indicating that the model effectively learned key features of the object detection task. It is noteworthy that the validation loss remained slightly higher than the training loss, which is consistent with expectations and reflects the model's generalization ability on unseen data. In terms of performance metrics, the precision and recall curves showed a steady upward trend and reached a relatively stable level by the mid-training phase. The mean average precision metrics mAP50 and mAP50–95, performed particularly well, both continuously improving throughout the training process and eventually reaching excellent levels above 0.85 and 0.65, respectively.

3.3. Human operator facial recognition

The angle of facial recognition indicates its deviation from the camera, which captures the work area into disassembly tools, disassembly objects, and robot zones. However, it is necessary to infer the rightward and leftward turns with the interaction of the robot, while a downward or centered position shows focus on the workpiece. Facial angle recognition can be used to determine whether actions are consciously motivated to improve intent recognition accuracy by reducing errors from similar actions. However, there are several methods for face angle recognition, including PNP-based and deep learning-based head pose estimation. The former utilizes feature points on a 2D plane and their corresponding coordinates in 3D space to establish a mapping relationship for estimating head pose. The latter relies heavily on accurate facial keypoint detection, leading to significant testing errors. These models take an RGB facial image as input and

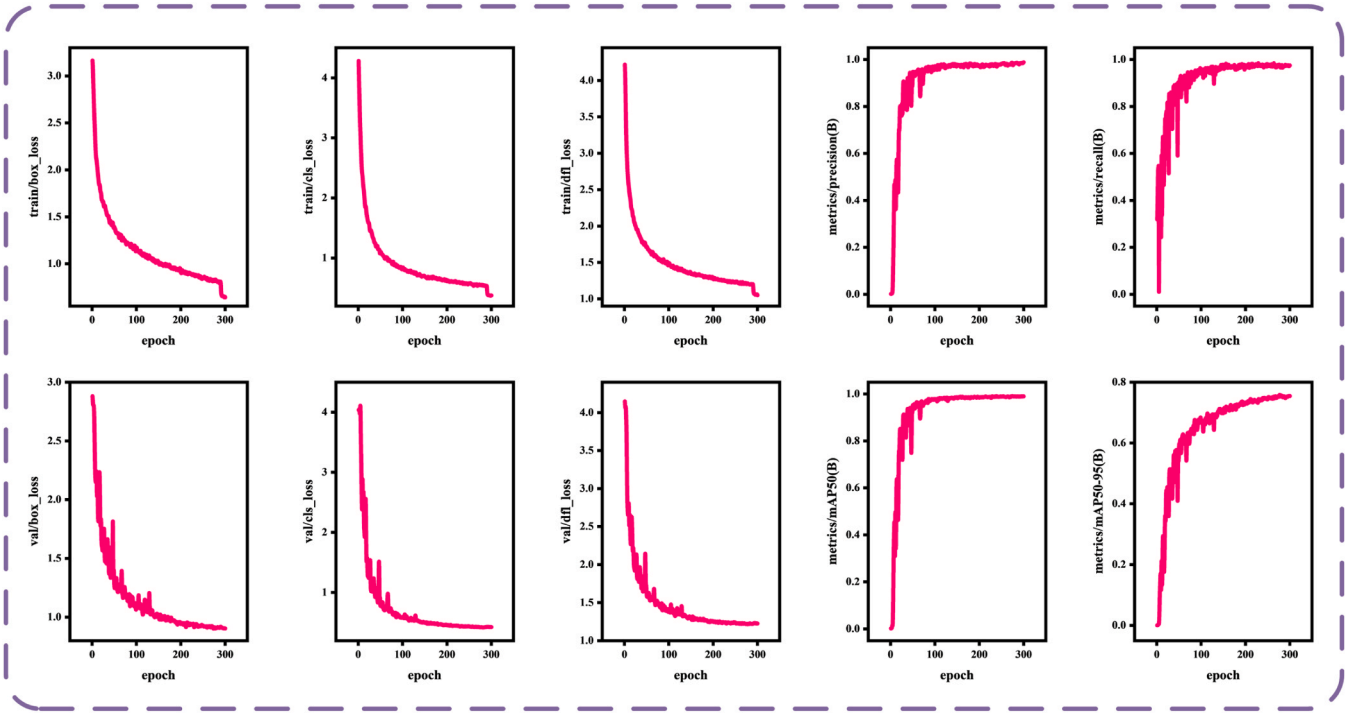

Fig. 8. Parameter variations during training of YOLO models.

Table 3
The feature comparison of various head-pose estimation algorithms.

Estimation algorithm	Related features and benefits	Application constraints and pitfalls
Keypoint-based	To train annotated keypoint data. To provide coordinate information.	To cause a severe imbalance in training samples. To generate the calculation errors from 2D to 3D
Deep Learning	To directly regress the Euler angles To design a simple model.	To converge the model training To annotate the data.

output three values to represent the head Euler angles. As shown in Table 3, compared to traditional head pose estimation algorithms, deep learning methods do not depend on facial keypoints, resulting in higher accuracy and better performance.

3.3.1. *Data preprocessing for human facial recognition*

The face angle recognition algorithm utilizes a multi-loss convolutional neural network to combine the classification and regression that can be used to predict head pose angles (yaw, pitch, and roll) without facial keypoint detection. To enhance model robustness, we collected a primary dataset of 1000 images featuring diverse individuals and backgrounds. The data collection was designed to capture the variability and patterns of facial features by various types of shots, including upper-body, head, and full-body captures taken in different environments. Furthermore, it is necessary to remove any unsuitable images from the dataset that affect the recognition performance with the overexposed or too dark backgrounds. The remaining face images were then resized to 64 × 64 pixels and saved in the same folder, which impacts both the accuracy and speed of training. The larger images may yield more accurate results, but also increase training time. After preprocessing the dataset, it is necessary to train a face detection model for angle recognition, which utilizes the Dlib model library, which employs a method based on HOG (Histogram of Oriented Gradients) features and a cascade classifier. HOG features effectively describe the edge and texture information within images, while the cascade classifier is a multi-layer system where each layer acts as a weak classifier, culminating in a strong classifier. Similarly, the algorithm uses a five-layer cascade classifier to detect the facial features across images of various scales. Before training the model, the doubled size of the images can be defined to enhance detection capabilities for smaller facial features. During this training process, it is necessary to extract the HOG features involved addressing two main issues: scale and position. For scale, it is necessary to ensure that it can be compared to the training images. Regarding position, the target could be located anywhere within the image to identify regions of interest. Even if the target size matches that of the training set, it is necessary to employ the OpenCV method to accurately locate the target within the image.

Furthermore, it is necessary to set the size of the scanning window for extracting HOG features with a configuration of 80 × 80 pixels. This fixed-size window, combined with image scaling (image pyramid), allows us to detect the facial features at different scales as shown in Fig. 9. Constructing an image pyramid generally involves several steps: (1) obtaining the image; (2) scaling the image using arbitrary scale parameters; (3) employing Gaussian blur to smooth the image with the noise. Moreover, it is necessary to define a trainer for face detection. The important parameters need to be set appropriately within the trainer to fit the training set better. By testing various data sets, the optimal value can be determined, while it can be verified that a smaller value leads to a more precise optimization; however, it also increases training time.

3.3.2. *DL-based human facial angle recognition*

In addition, a straightforward approach can be used to detect face angles using deep learning, mainly involving training a multi-loss convolutional neural network. However, the network integrates both classification and regression to predict three head pose angles: yaw, pitch, and roll. As shown in Fig. 10, the architecture of facial recognition can be designed by combining the ResNet50 as the backbone for feature extraction. Therefore, it is important to classify and regress the data sets of the face angles with a multi-output network model.

In addition, it can be completely defined as the recognition angles for head rotation from −99° to 99°. To effectively classify the continuous facial angles, it can employ a binning approach to separate them into 66 discrete intervals, each with a step size of 3°, resulting in 66 distinct classifications. To reduce the dimensionality of average pooling before

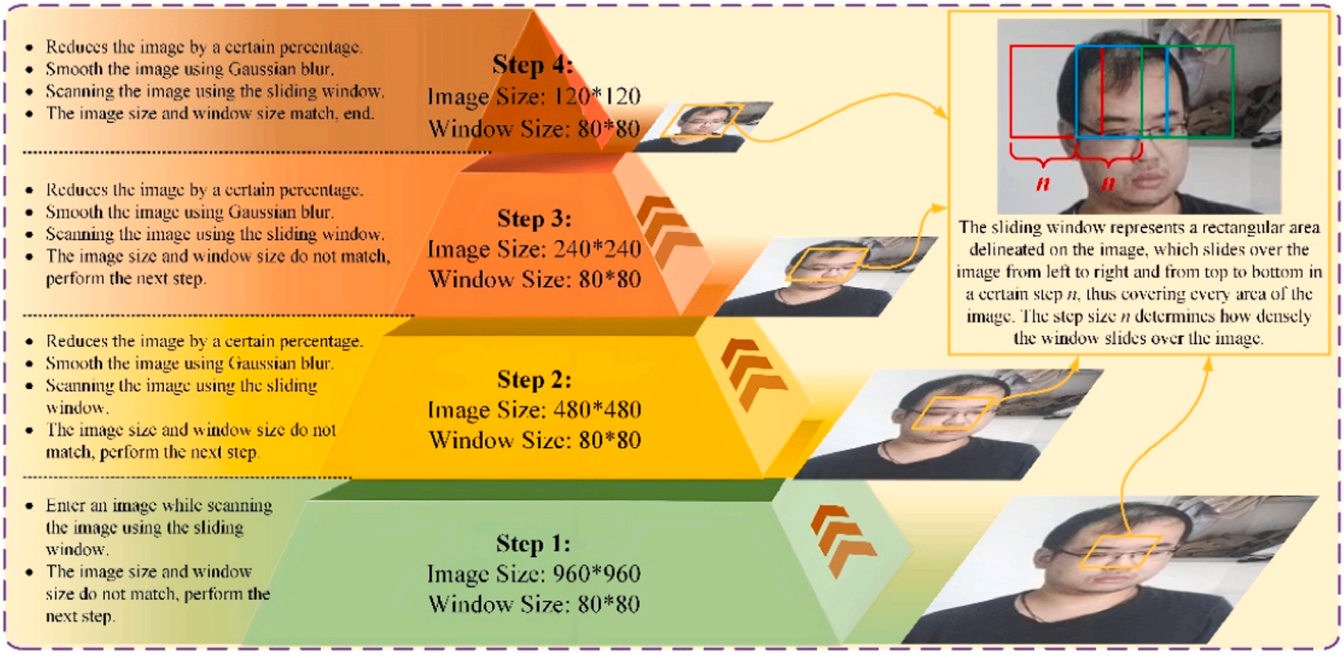

Fig. 9. The image pyramid procedure of facial feature recognition.

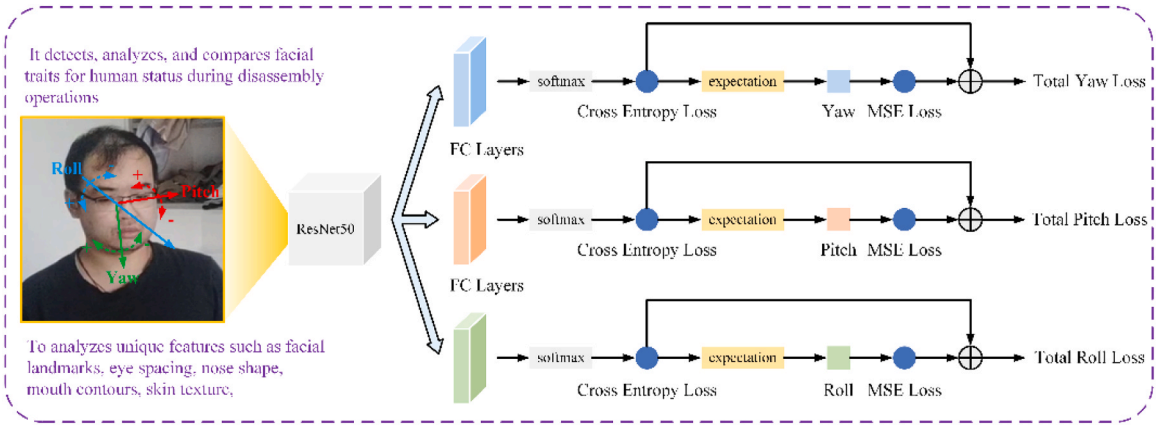

Fig. 10. The network architecture for face angle detections.

the output of three fully connected (FC) layers, the architecture yields a 66-dimensional output, where each dimension represents the score associated with the Euler angles. Each branch of the network specifically predicts one of the Euler angles: yaw, pitch, or roll. The scores are then transformed into probabilities using a Softmax function, which considers the calculation of cross-entropy loss as follows:

$$H_{\text{Loss}} = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \times \log(\hat{y}_i) + (1-y_i) \times \log(1-\hat{y}_i)] \tag{22}$$

Where $\hat{y}_i$ is the predicted value and y_i is the real value. The ground truth (GT) refers to the labels derived from the BIWI dataset, where the Euler angles can be extracted from the rotation matrices associated with head poses. Specifically, it can be determined that the 66 bins (0−65) related to the given angle can identify the bin location of the Euler angles. For example, if the Euler angle is 0°, it would fall into the central bin labeled as 33. Subsequently, a weighted sum of the predicted probabilities can be calculated for each bin, which yields the estimated Euler angles. Finally, the Mean Squared Error (MSE) loss can be calculated by comparing the estimated angle.

$$\text{MSE} = \frac{1}{N}\sum_i^n (\hat{y}_i - y_i)^2 \tag{23}$$

Finally, a weighted combination of two loss functions can be calculated as follows:

$$\text{L} = \lambda_1 \times H_{\text{Loss}} + \lambda_2 \times \text{MSE}(\text{y}, \quad \hat{y}) \tag{24}$$

Where λ_1 and λ_1 represent the weight parameters that can be adjusted according to the specific conditions. Although the aforementioned methods can accurately identify the face angles, the recognition speed is relatively slow and makes them unsuitable for practical engineering applications. Euler angles can represent the roll, the yaw, and the pitch with different colors. Meanwhile, the pitch angle involves rotation around the X-axis (nodding), where upward is negative and downward is positive. The yaw angle corresponds to the rotation around the Y-axis (shaking the head), with left being positive and right negative. The workspace is divided into three zones, including the robot zone, the object zone, and the tool zone. For instance, a face turned to the right suggests greater attention to the tools, while a left turn indicates focus on the robot. Additionally, since precise angle values are unnecessary, larger angular intervals can simplify classification, which enables the output of categorical labels instead of specific angles to facilitate the determination of left and right turning ranges in subsequent experiments. Furthermore, the large ResNet50 architecture has a lightweight backbone network with 1:1 ratio between input and output channels to reduce memory access costs (MAC), which avoids the excessive group convolution and recognizes the element-wise operations like ReLU and AddTensor. However, ShuffleNetV2 can be selected as the backbone to offer significantly lower computational demands compared to ResNet and ResNeXt, which allows ShuffleNet to accommodate more feature maps within given computational resources with smaller networks that may lack sufficient channel capacity. As shown in Fig. 11, ShuffleNetV2 enhances the foundational framework of ShuffleNetV1, which employs a Channel Split strategy at the start of each unit to separate the input feature channels into two parts. Unlike ShuffleNetV1, the 1 × 1 convolutions are not group convolutions as the Channel Split uses a Concat operation to maintain channel consistency, while the Channel Shuffle operation facilitates information exchange. Similarly, the ShuffleNetV2 also retains the Depthwise Convolution layer for efficient feature extraction, allowing for the construction of networks with varying depths to accommodate diverse tasks.

By considering the fundamental structure and functionality, the network structure of ShuffleNetV2 can be outlined as shown in Table 4. The structure of ShuffleNet is similar to ResNet, which is separated into several stages (i.e., ResNet has four stages while ShuffleNet has three). In each stage, the ShuffleNet units replace the original Residual blocks, which form the core of the ShuffleNet algorithm. Furthermore, it is necessary to compare the recognition efficiency of the improved algorithm and the original algorithm by recognizing the same-sized video and recording the recognition time. The results showed that the improved algorithm was 40 % faster than the original one.

Based on the operational characteristics of HRC disassembly, the workbench is typically divided into distinct zones, namely the robot zone, the disassembly zone, and the tool zone. Accordingly, the system can map the operator's head orientation into three discrete categories, identified as left-turn, center, and right-turn. This design significantly simplifies detection requirements, as the system only needs coarse-grained classification labels rather than high-precision angular regression. This strategy indirectly enhances model robustness by increasing its tolerance to sensor noise, illumination variations, and partial occlusions, such as those caused by safety helmets or temporary obstructions, since minor angular deviations do not lead to misclassification. By employing this coarse-grained classification method, the system maintains intent recognition accuracy while effectively mitigating visual interference common in industrial settings, thereby ensuring both the smooth operation and safety of the HRC disassembly process.

Furthermore, it is necessary to have specifically optimized the hyperparameters for the training process of the facial angle detection CNN model to enhance its training performance. The number of training epochs was set to 5, with a batch size of 16 and a base learning rate of 0.001. The optimization process utilized the Adam optimizer, and a layered learning rate strategy was applied to the network parameters: the parameters of the fully connected layer were updated using 5 times the base learning rate, the parameters of the backbone network layers

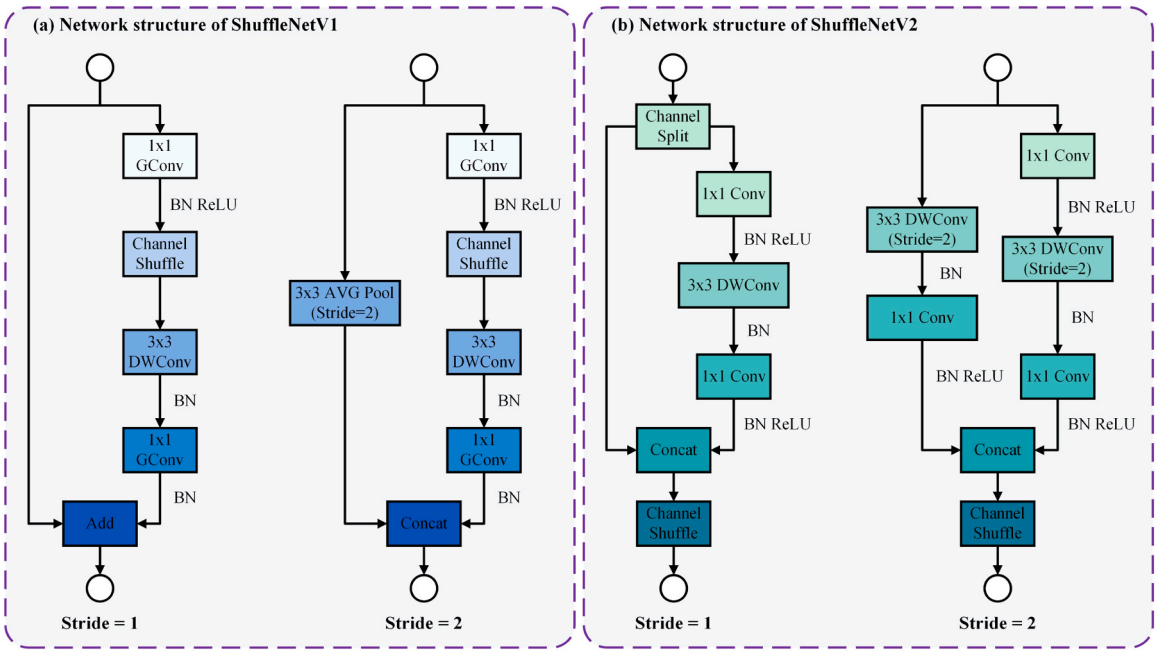

Fig. 11. The network architecture of ShuffleNetV1 and ShuffleNetV2.

Table 4
The hyperparameters of the ShuffleNetV2 network architecture.

Layer types	Output sizes	*K*	stride	Loop	Output channels			
					0.5x	1x	1.5x	2x
Image	224 × 224				3	3	3	3
Conv1 MaxPool	112 × 112	3 × 3	2	1	24	24	24	24
	56 × 56	3 × 3	2					
Stage2	28 × 28		2	1	48	116	176	244
	28 × 28		1	3				
Stage3	14 × 14		2	1	96	232	352	488
	14 × 14		1	7				
Stage4	7 × 7		2	1	192	464	704	976
	7 × 7		1	3				
Conv5	7 × 7	1 × 1	1	1	1024	1024	1024	2048
GlobalPool	1 × 1	7 × 7						
FC					1000	1000	1000	1000
FLOPs					41 M	146 M	299 M	591 M

used the base learning rate, while the parameters of some lower-level convolutional layers and batch normalization layers were kept frozen. In terms of loss function design, a combination of cross-entropy loss for angle classification and mean squared error loss for continuous angle regression was used. The contributions of these two loss components were balanced using a weighting coefficient of 0.001. The input images underwent resizing to 240 pixels, followed by a random 224 × 224 crop, tensor conversion, and normalization based on ImageNet dataset statistics. During training, loss information was output every 100 iterations, and a model snapshot was saved upon the completion of each training epoch. The model architecture was based on transfer learning from pre-trained models such as ResNet50 or AlexNet, with weight initialization using the parameters pre-trained on ImageNet.

3.4. Multi-modal Information Fusion and Decision Logic

Based on the aforementioned independent methods for disassembly action perception, tool detection, and facial angle recognition, the HRC disassembly system can effectively fuse these heterogeneous perceptual streams through a sequence-level fusion strategy to generate a robust, context-aware estimation of the human operator's intent. This fusion is not merely parallel processing but a synergistic integration, where the outputs of each module are cross-validated to resolve ambiguity and form a consistent understanding of the disassembly task progress. A conceptual summary of this decision logic is illustrated in Fig. 12, which aims to categorize human intent into three fundamental types: Preparation, Operation, and Hazard, each triggering a distinct robotic collaboration strategy.

Preparation-type intent signifies that the operator is engaged in preliminary actions required for subsequent disassembly subtasks. A typical characteristic of this intent is the directional focus of the operator's attention. The facial angle recognition module consistently detects a gaze direction towards the tool area, indicating a conscious effort to locate and select the appropriate tool. Simultaneously, the action recognition module typically identifies actions such as reaching or picking up, while the tool detection module confirms that a specific tool is in the operator's hand. Once a preparation intent is inferred, the robotic system transitions into a proactive support mode. This may include pre-emptively fetching and positioning the next required component or preparing its end-effector for the upcoming collaborative task, thereby minimizing idle time and enhancing workflow continuity.

Operation-type intent indicates the active execution of core disassembly operations. At this point, the operator's attention is primarily focused on the EOL product itself, with the facial angle recognition

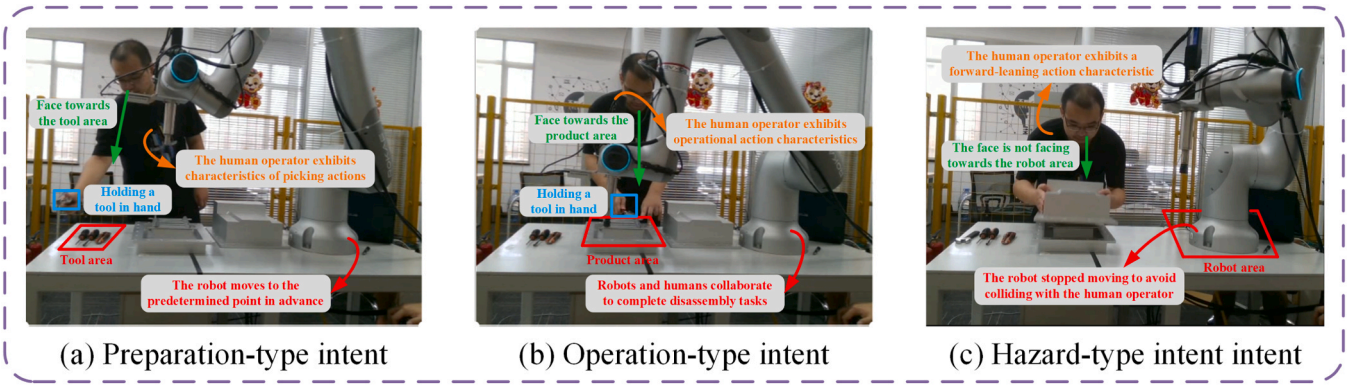

Fig. 12. The human operator intent inference and response with combination rules.

module consistently reporting a central gaze direction. The action recognition module classifies definitive disassembly actions, such as loosening a bolt or prying a component. The tool detection module concurrently verifies the correct use of the corresponding tool for the current task. The fusion of signals such as focused attention, definitive action, and contextual tool use provides clear confirmation of the ongoing operation. Accordingly, the robot executes its pre-planned collaborative routine. This often involves providing complementary assistance, such as applying a stable counter-force, physically supporting the workpiece to prevent displacement, or operating a screwdriver precisely on a predefined bolt, all aimed at reducing the cognitive and physical load on the human operator.

Hazard-type intent is crucial for maintaining a safe collaborative workspace and is primarily triggered by the detection of abnormal postures or loss of attention. The action recognition module plays a key role by identifying potentially hazardous body configurations, notably recognizing a forward-leaning posture that brings the operator too close to the robot's operational envelope. This risk factor is significantly amplified when the facial angle recognition module shows that the operator's gaze is not directed towards the robot, suggesting a potential lack of awareness of its movement. Once a hazard intent is inferred, regardless of the tool detected, it triggers an immediate, safety-oriented response from the robotic system. This response is governed by a priority safety protocol, which can be dynamically adjusted from initiating protective speed reduction to a complete halt of movement, thereby avoiding potential collisions until the operator resumes a safe posture and situational awareness.

For example, the intent to prepare for loosening a bolt is not inferred merely from the action of picking up an object, but from the combination of that action with the simultaneous visual detection of a wrench in the operator's hand and, crucially, the confirmed direction of the operator's gaze towards the tool area. This multimodal consensus distinguishes it from the intents behind other similar actions. Conversely, the intent to actually perform the loosening operation is confirmed by the fusion of the nut-loosening action, the sustained presence of the wrench, and focused attention on the workpiece. This logical framework, validated in our subsequent experiments, provides a foundation for intuitive and safe human-robot collaboration by enabling the robot to perceive not just actions, but their contextual meaning and safety implications.

To prevent safety incidents caused by misclassification, the system incorporates a safety-aware error handling mechanism. In cases where intent is incorrectly classified, for example, when sensor noise leads to a "lifting" intention being misinterpreted as a "forward" intention, the fusion algorithm compares the current output with recent historical data to detect anomalies. Should discrepancies persist, the system triggers a low-confidence protocol that defaults to a safety-first control strategy. Specifically, the robot immediately switches to a protective operational mode where its speed is reduced by 50 %, compliance is increased through impedance control, and audiovisual alerts are issued to the operator for manual confirmation. For critical errors, such as detecting a "fall" intent with high confidence but lacking corresponding tool or facial data, the system initiates an emergency stop and locks robotic movement until manual verification is provided.

4. Experimental analysis

4.1. Experimental platform settings

The HRC disassembly experiments for EV battery were conducted on a dedicated disassembly platform, involving the ROKAE CR12 robot with a maximum 12 kg load, a working radius of 1400 mm, and a repeat positioning accuracy of 0.03 mm. The robot equipped with a DDK AFC3000 electric screwdriver as its end effector provides a maximum torque of 40 N·m and a maximum speed of 250 rpm. A depth camera with the D435i can capture environmental information with the camera positioned approximately 1.7 m above the ground and about 1.5 m horizontally from the human operator. The depth camera connects to a PC via USB 3.0 and outputs RGB images with a resolution of 1920 × 1080 at a frame rate of 30 fps, which is equipped with a GTX 1060 graphics card. Disassembly tools (i.e., wrench and utility knife, etc.) are arranged in the right area of the test bench, while the disassembly object can be placed in the central area of the HRC platform. To ensure the electric screwdriver operates correctly, the experiment utilizes a PLC (Programmable Logic Controller) to control its working status, which is connected to the PLC via the control I/O ports using Step7 software with specific I/O allocations as shown in Table 5.

The video capturing, intent recognition, and control monitoring can be integrated into ROS (Robot Operating System). As shown in Fig. 13, the video is captured using a depth camera to acquire RGB and depth data of the work area, while simultaneously collecting information

Table 5
PLC I/O allocation for screwdriver control device for disassembly operations.

No.	Signal Name	Description	NO.	Signal Name	Description
I0.1	STOP	Emergency Stop	Q0.1	REJECT	Unqualified
I0.2	RESET	Stop Unlocking Action	Q0.2	ACCEPT	Qualified
I0.3	REVERSE	Reverse Rotation	Q0.3	ABNORMAL	Abnormality During Operation
I0.4	START	Start loosening action	Q0.4	READY	Ready for Operation
I0.5	BYPASS	Pause Operation	Q0.5	BUSY	In Operation
I0.6	SELF CHECK	Sensor Self-diagnosis	Q0.6	BYPASS	Axis Disconnection
I0.7	ALL OFF	All Output Signals OFF	Q0.7	ALL OFF	All Output Signals OFF

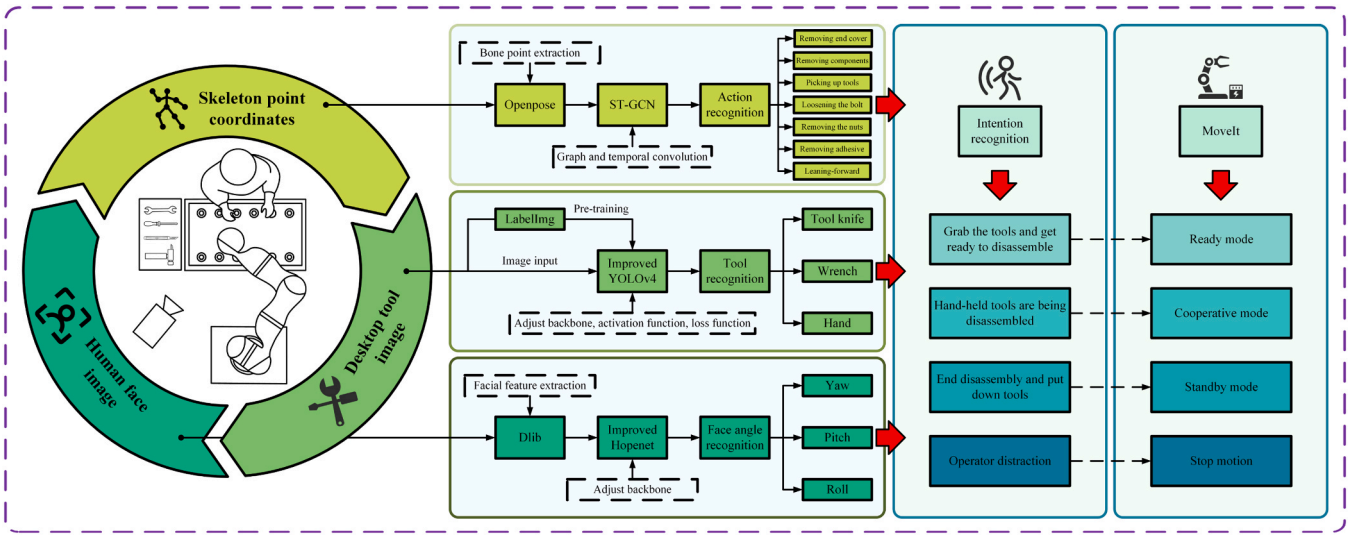

Fig. 13. The overall framework of human intent recognition based on the HRC disassembly system.

about the human operator and the disassembly objects. The computer then processes the data in parallel channels to infer the human operator's intention. This involves the use of an ST-GCN model for action recognition, an improved YOLO algorithm for tool detection, and a lightweight CNN for facial angle recognition. Once the intention recognition algorithm identifies the specific human intention, the result is output to the control monitoring module. Based on the recognized intention and integrated with robot information (such as the collaborative robot and its end-effector), this module executes various disassembly actions. The MoveIt package is primarily utilized for motion

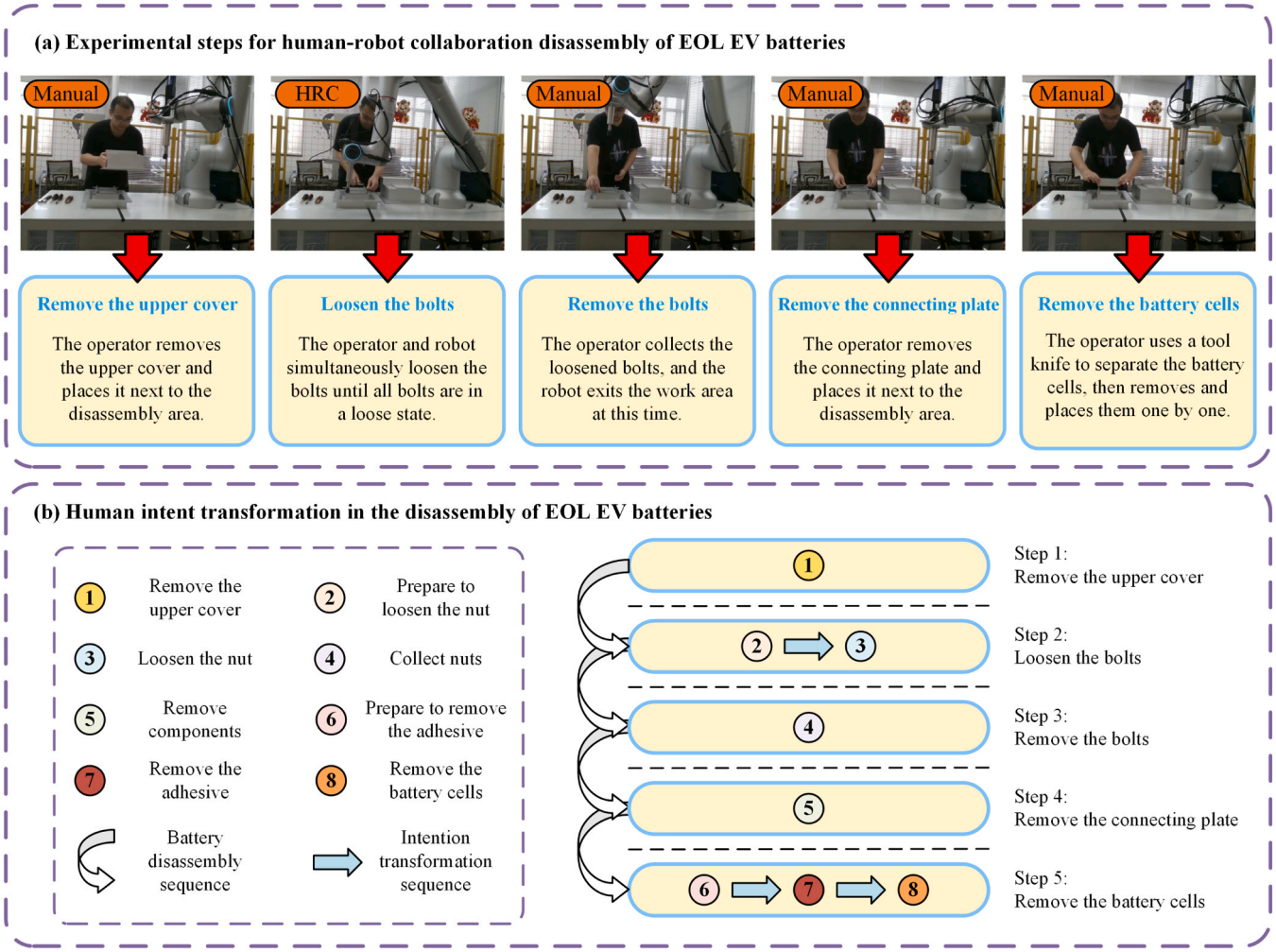

Fig. 14. The detailed experimental steps and human operator intention evolution chain based on HRC disassembly operations.

planning and trajectory generation, thereby controlling the robot to complete the disassembly tasks. The core component, *move_group*, integrates various elements to offer a comprehensive API with data interaction between *move_group* and user interfaces. The Actions, Services, or other methods enable the users to utilize interfaces like *move_group_interface* (C++), *moveit_commander* (Python), and the RViz plugin for disassembly tasks such as path planning, execution, and kinematics. The ROS parameter server stores critical parameters and variables in ros_master for Move_group, categorized into three types: the robot model description, the configuration information, the additional motion planning.

In addition, it is necessary to primarily utilize the MoveIt packages for motion planning by determining the state of the robotic arm, which includes reading its current joint angles, positions, and velocities. Next, a planning request can be used to outline the disassembly tasks, which perform motion planning based on the current state and disassembly operations to generate a feasible trajectory for robotic movement. The motion planner operates by automatically calculating the robot's path and control commands based on its kinematic model, environmental information, and movement requirements. For the path planning, the planner searches for the viable paths within the robot motion space to generate a series of waypoints typically represented in Cartesian space with the end effector postures. During the trajectory planning, these waypoints are converted into joint angles and time-parameterized to create a smooth trajectory to avoid sudden movements or vibrations, which is sent to the robot controller for execution to control each joint movement at specified time points.

4.2. Disassembly experiments for the EV battery as an example

The designed experiment involves the HRC disassembly of an EV battery model as an example to accomplish the disassembly sub-tasks: 1) Remove the upper cover; 2) Loosen the bolts; 3) Remove the bolts; 4) Remove the connecting plate; 5) Remove the battery cells. Fig. 14 systematically elucidates the complete disassembly process of an EOL EV battery in a laboratory environment, while also explaining the real-time transformation of the human operator's intent embedded within each physical operation sequence. First, as shown in Fig. 14(a), the overall procedure can be delineated into five distinct, sequential steps. The process begins with the operator manually removing the top cover, thereby exposing the internal battery structure for subsequent disassembly. This is followed by a core phase of HRC, where the operator and the collaborative robot simultaneously loosen the bolts on the connecting plate, with the robot's involvement significantly enhancing the efficiency of this stage. Subsequently, the operator manually collects all the loosened bolts, while the robot retracts from the work area to ensure safety. The process continues with the manual removal of the connecting plate and concludes with the separation and removal of individual battery cells using tools.

Concurrently, based on our analysis of the human operator's intent, the entire disassembly process follows a well-defined intent evolution chain. This chain, numbered chronologically from ① to ⑧ in Fig. 14(b), is not merely a simple sequence but a continuous cascade where each intent logically precipitates the next, precisely synchronized with the physical operation sequence. The process initiates with the operator's intent to remove the top cover (①), a conscious decision that starts the task. Upon completion, this intent naturally transitions into the intent to prepare for loosening nuts (②), signifying preparation for the subsequent collaborative operation. The action of grasping a tool immediately triggers the intent to loosen the nuts (③), which defines the collaborative goal. Crucially, the recognition of this executive intent (③) serves as the key signal prompting the collaborative robot to engage, enabling it to position itself synchronously with the human and initiate the bolt-loosening operation. Thereafter, the operator's focus shifts to the intent of collecting the loosened nuts (④), a post-execution activity, while the robot, upon completion of the previous executive intent, retracts accordingly. The workflow then progresses as the operator forms the intent to remove the connecting plate (⑤), an action performed manually. This is followed by the operator forming the intent to prepare for adhesive removal (⑥) in preparation for the final separation task. This preparatory intent seamlessly evolves into the intent to remove the adhesive (⑦), another executive action. The entire disassembly sequence culminates with the intent to remove the battery cells (⑧), marking the completion of the core objective. The recognition of these distinct intents enables the robot to provide anticipatory assistance, thereby creating a fluid and efficient collaborative workflow.

All environmental conditions in the experiment, such as the initial state of the robot, the positions of tools and EOL products, and lighting parameters, were maintained consistently. Each operator was fully instructed regarding the specific procedural steps to be performed. All actions during the experiment were natural operational behaviors, carried out without artificial acceleration or intentional delay, thus ensuring the randomness of the experimental process. Fig. 15 summarizes all possible detection results from the three core perception modules in the HRC disassembly system, along with the combinatorial rules used for inferring human intention based on these results. Fig. 15(a) presents the output of the human action recognition algorithm, which primarily includes seven types of actions: Remove upper cover, Pick up, Loosening, Remove nut, Remove component, Remove adhesive, and Forward. Fig. 15(b) displays the output of the facial angle recognition algorithm, consisting of three facial orientation categories: Turn left, Center, and Turn right. Fig. 15(c) shows the output of the disassembly tool detection module, which includes four possible results: Tool knife, Wrench, Hand, and Nothing (no tool detected). Thereby, by combining different detection results, the operational intention of the human operator can be sufficiently inferred. Considering experimental safety and operational fluency, in addition to the three main categories mentioned above, several additional actions, such as Collect nuts, Fall, and Wander along with their corresponding combinations of detection results have been defined.

Three experimenters conducted individual intention recognition experiments for single actions. Each experimenter performed the actions of removing adhesive, loosening nuts, removing components, and removing the upper cover 20 times. One operator performed the tasks in an environment using only the ST-GCN action recognition for intention recognition, another in an environment using only the YOLO tool detection for intention recognition, and the final operator in an environment using the integrated multimodal fusion algorithm for intention recognition. The results indicate that the intention recognition accuracy for loosening nuts is the highest at approximately 91 %, while the accuracy for removing the upper cover is the lowest at 78 %, resulting in an average recognition accuracy of 84 %. The lower accuracy for removing the upper cover can be attributed to the heavy weight of the cover, which often requires a bending motion that may be misclassified as a forward-leaning action. Additionally, the actions for removing components and the upper cover share similarities, leading to potential misrecognition. The angle recognition results demonstrate that the adjusted algorithm can quickly and accurately identify the angle of the face, which relies on a face detection model, including image blur, insufficient lighting, or other disturbances. To ensure better image quality of the disassembly process, it is necessary to pre-process the images by enhancing the clearer images and lighting conditions to provide better training examples.

4.3. Performance evaluation for HRC disassembly

To comprehensively validate the performance of each perception module and the entire integrated system, we first conducted experimental evaluations on the action recognition and target detection models separately. However, it is crucial to emphasize that the facial angle detection model was intentionally omitted from an independent experimental evaluation. This decision was made because facial

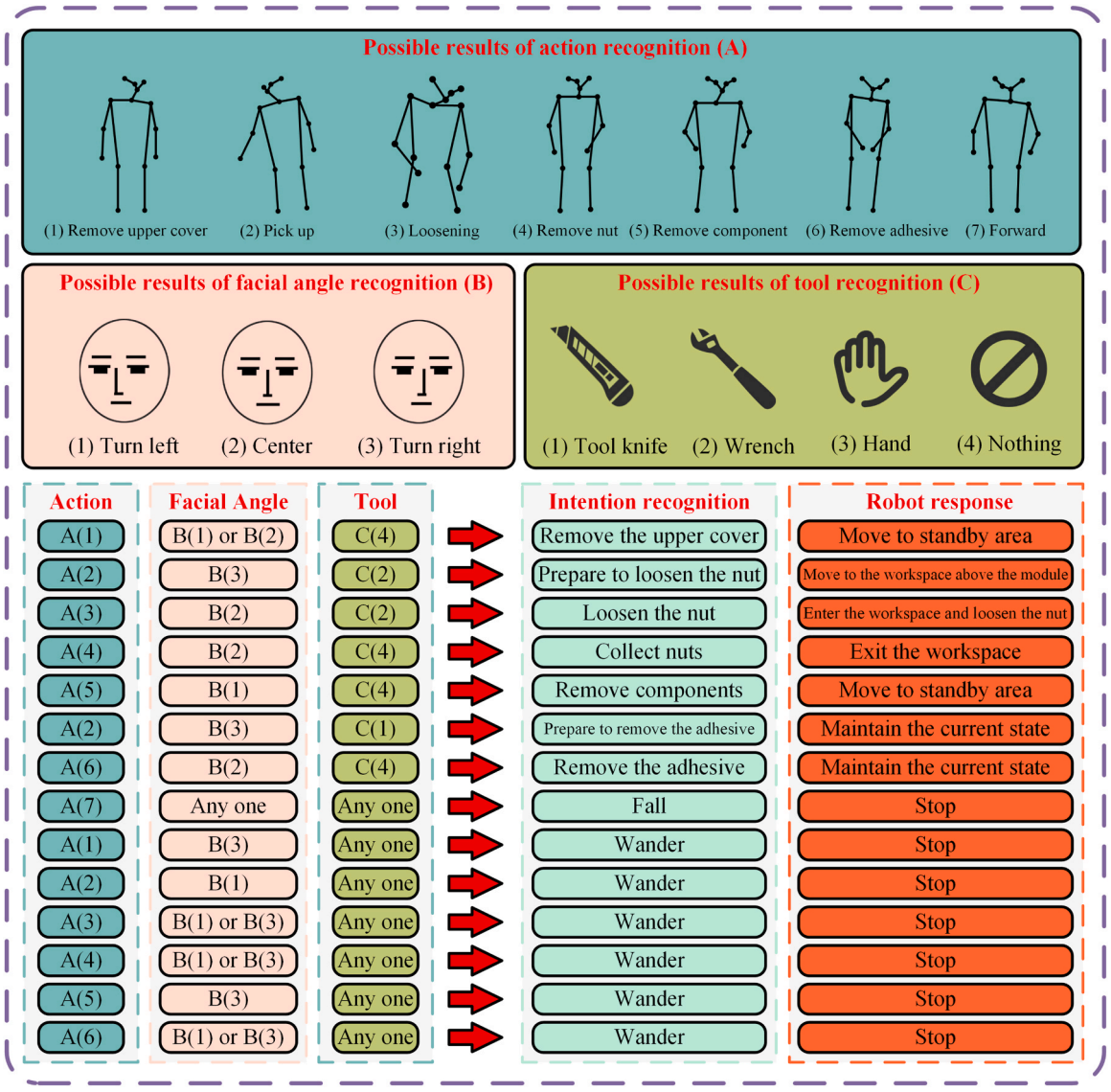

Fig. 15. The experimental results of human action recognition, facial angle recognition, and tool detection based on the disassembly tasks.

orientation information, by itself, possesses limited semantic value for inferring specific human actions or disassembly intents. For instance, a face turned to the right could indicate attention directed towards the tool area, towards the robot, or merely a transient movement; without the contextual correlation of simultaneously occurring actions and tool usage, its interpretation remains ambiguous. Therefore, the effectiveness of the facial angle recognition module was not assessed in isolation but was instead intrinsically validated through its contribution to the overall intention recognition system.

As shown in Fig. 16, it describes the recognition performance of the action recognition model based on the ST-GCN architecture during the experimental process. It revealed that some behaviors were composite actions, meaning that a single behavioral intention could encompass two distinct actions. For instance, as shown in Fig. 16(a), the removal of the upper cover required a forward-leaning posture due to the weight of the cover. This indicates that some actions, while they may appear as a single behavior, can involve multiple sub-actions that are not always easily distinguished by a basic recognition system. When the subsequent action was not recognized, the output remained consistent with the previously recognized action, which could distinguish the overlapping actions or identify the transitions. However, despite the presence of some overlapping actions, the algorithm can still accurately identify individual actions, demonstrating its ability to learn effective features of different actions and correctly distinguish the characteristic differences between them. The results demonstrate the algorithm's performance in recognizing human actions with composite behaviors and overlapping intentions, which proves the effectiveness of separating and identifying individual actions within complex action sequences to improve the action recognition in more realistic disassembly scenarios.

Furthermore, during the development of the action recognition component, we initially experimented with the PyTorchVideo framework, which combined object detection (Faster R-CNN) and action classification (SlowFast) for operator action recognition. However, the object detection module within PyTorchVideo utilized the built-in Detectron2 implementation of Faster R-CNN, which proved to be relatively slow and operated in a non-continuous manner, leading to processing delays. We then attempted to replace the native Faster R-CNN with the faster YOLOv4 model. Nevertheless, a significant incompatibility arose: YOLOv4 processes video frame-by-frame, while the SlowFast model requires batches of 25 frames for analysis. Furthermore, the SlowFast model itself has considerable computational latency. This mismatch, combined with the inherent slowness of SlowFast, resulted in severe detection lag during practical experiments. Therefore, after comprehensive consideration, the ST-GCN architecture was selected as

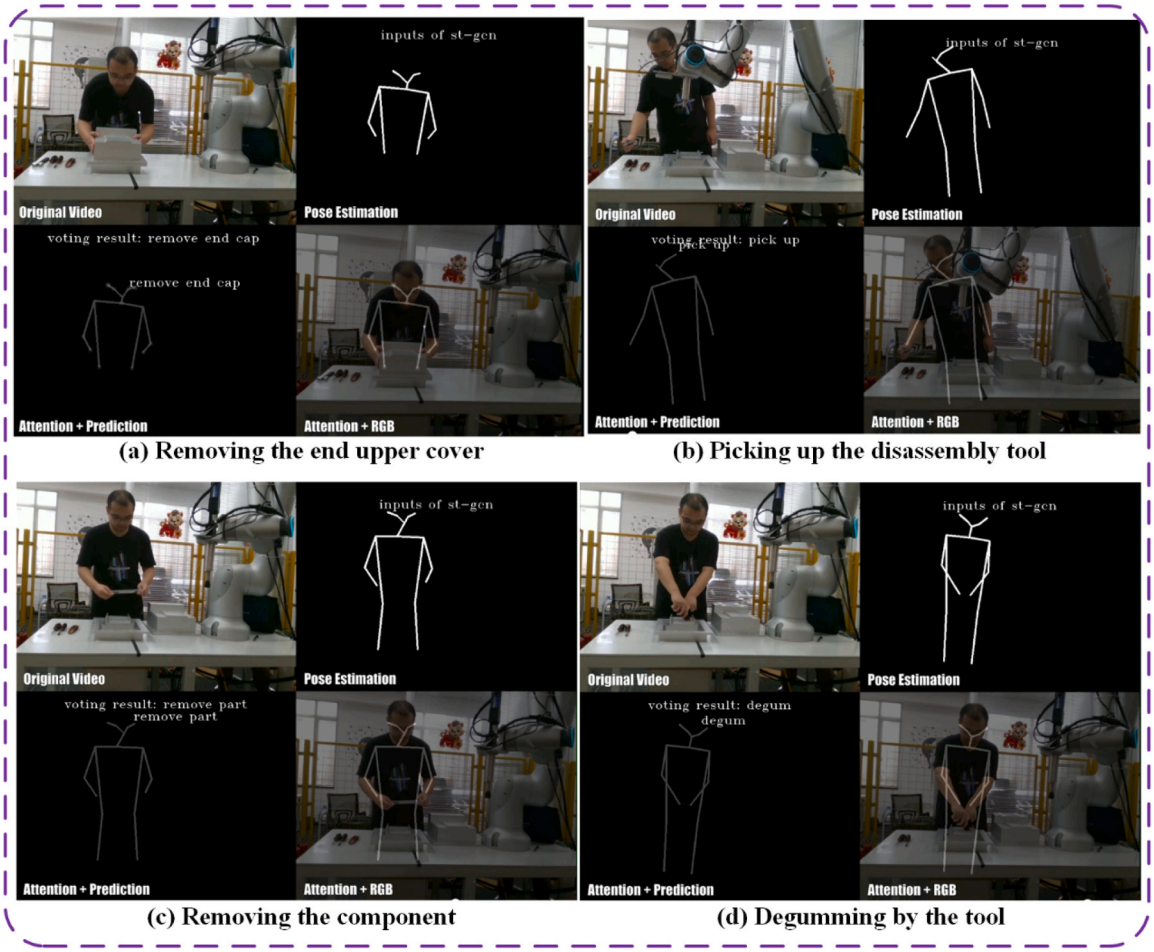

Fig. 16. The results of human action recognition during the specific disassembly operation, involving removing the upper cover, picking up the tool, removing the components, and degumming by the tool.

it provides smoother performance and is better suited for intent recognition in EOL EV battery disassembly tasks. As shown in Fig. 17(a), the average disassembly success rate using only action recognition methods was 76 %, while the average successful disassembly rate of integrated intent recognition was 84 %. The data clearly shows a more accurate identification of human intent by combining action recognition, object detection, and facial angle detection. To further evaluate the effectiveness of the adjusted YOLOv4 model, a comparative experiment was conducted between the original and modified algorithms using only the object detection method. This experiment focused on adhesive removal and nut removal, as tool detection is particularly critical for recognizing these specific operations.

As shown in Fig. 17(b) indicates that changing the activation function yields a slight accuracy improvement of disassembly action recognition. Meanwhile, the improved CSPDarkNet network alone can greatly boost accuracy at the reduced detection speed. The overall algorithm (D-CSPDarknet53 +H-swish) offers the improvements of both accuracy and detection speed compared to the original algorithm. To fully analyze the performance of different detection and prediction models based on various disassembly operations, it is necessary to train the CNN-LSTM model using the experimental disassembly data using multiple regression predictions. As shown in Fig. 17(c) and (d), the prediction and actual results show that the error values are all within 5 %, and the average recognition rate for using only object recognition methods is 72.5 %, which is sufficient to show the effectiveness of the target detection method.

In the subsequent experiments, it is necessary to conduct a comprehensive evaluation of the entire disassembly operation of the EV battery as an example. In the context of HRC disassembly, any errors in action recognition might lead to the robot's failures in battery disassembly or incorrect action classifications. As shown in Fig. 18, three operators conducted 20 trials each under three experimental conditions: "action recognition only", "object detection only", and "integrated intention recognition". Model performance was evaluated focusing on five key intent categories: Prepare to loosen the nut, Loosen the nut, Collect nuts, Fall, and Wander. This selection was based on their critical impact on the success and safety of the overall disassembly process. The intents Prepare to loosen the nut, Loosen the nut, and Collect nuts were chosen due to their direct relevance to the initiation and termination of robotic assistance during crucial human-robot collaboration steps. Accurate recognition of this intent sequence is essential, as any failure in this core collaborative phase would lead to the failure of the entire process, whereas purely manual steps are less prone to errors. Additionally, the safety-related intents Fall and Wander were included to assess the system's capability in mitigating operational risks. Confusion matrices were generated from the experimental recordings to evaluate model performance for these five intents.

Fig. 18(a) presents the model's performance when inferring human intention based solely on object detection. This method exhibits significant limitations, as the mere presence of a tool often fails to unambiguously indicate the ongoing action. This ambiguity is reflected in severe misclassifications, particularly between intents involving the same tool, such as confusing "Prepare to loosen the nut" with "Loosen the nut". Furthermore, the model demonstrates considerable difficulty in accurately recognizing states not primarily defined by tool usage, such as "Collect nuts" and the safety-critical state "Fall". Fig. 18(b) shows the

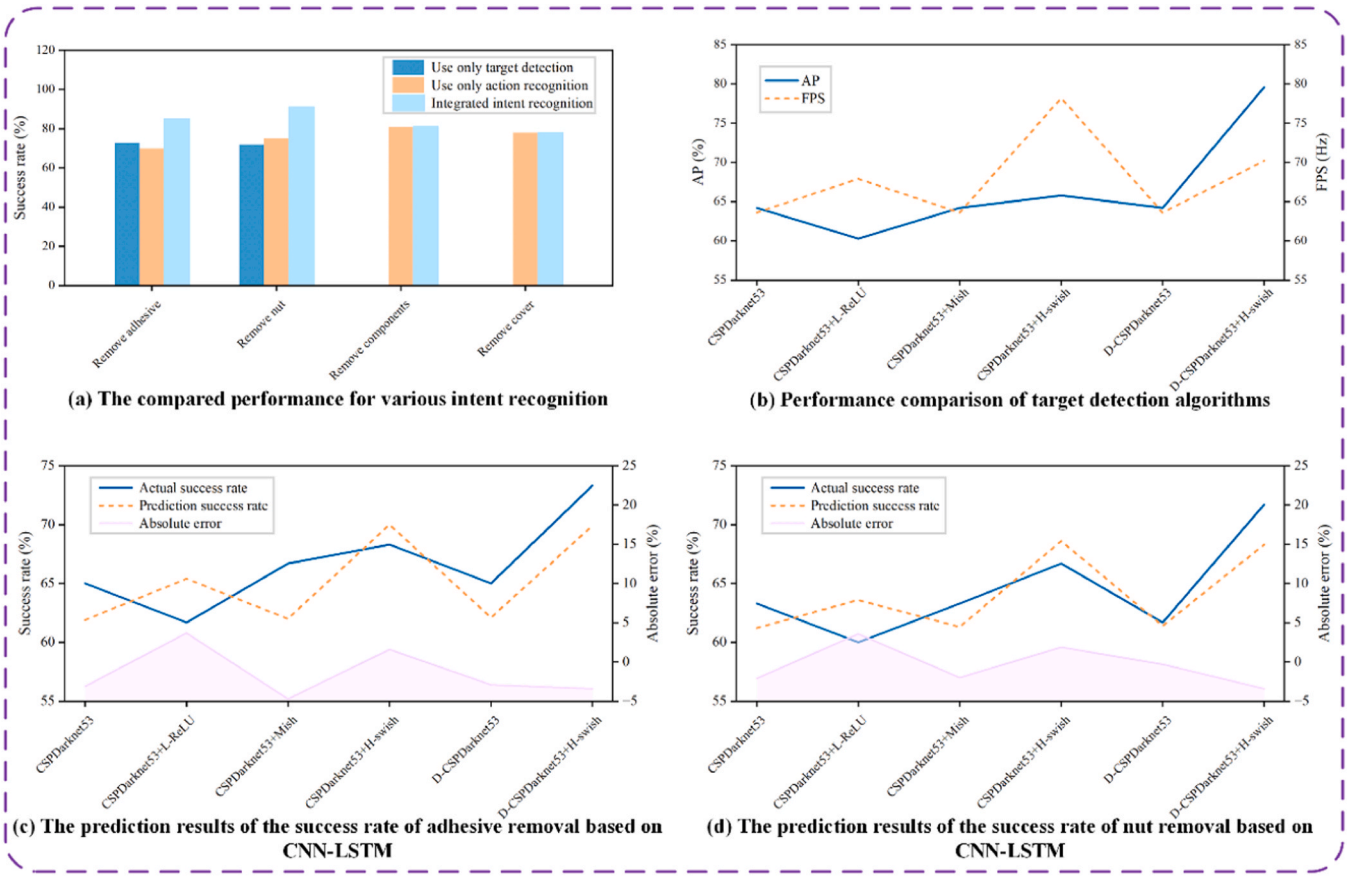

Fig. 17. The performance comparison for various disassembly actions (a) and target detection algorithms (b); and the prediction results of the successful rates for adhesive removal (c) and the nut removal (d) based on CNN-LSTM.

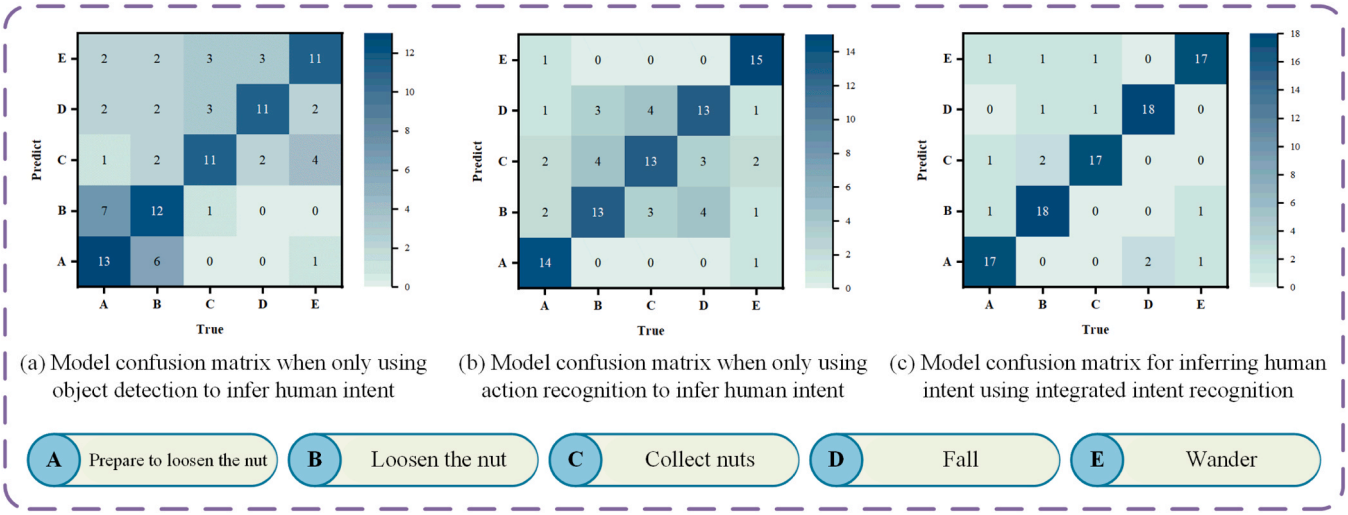

Fig. 18. The performance evaluation of intent recognition models in the HRC disassembly scenario.

results when intention is inferred solely through action recognition using the ST-GCN model, where overall accuracy improves markedly. Compared to the object-detection-only approach, the diagonal elements of the confusion matrix, representing correct classifications, are significantly more pronounced across all categories. This indicates that skeletal motion data provides a more reliable basis for distinguishing most intents.

However, challenges persist, particularly in differentiating actions that are visually or kinematically similar. Fig. 18(c) displays the confusion matrix for the integrated intention recognition method, which achieves the most robust classification performance. The fusion of object context information with action kinematics effectively resolves the ambiguities that plagued the unimodal models. The most notable improvement lies in the model's enhanced ability to differentiate the preparatory intent "Prepare to loosen the nut" from the executive intent "Loosen the nut", a distinction which is crucial for timely robotic

assistance. Additionally, the recognition accuracy for the safety-critical state "Fall" is significantly higher, underscoring the importance of multimodal perception for ensuring operational safety. The experimental results indicate that the proposed integrated intention detection method achieves success in 17 out of 20 EOL EV battery disassembly trials, yielding a disassembly accuracy of 85 %. In contrast, the accuracy of disassembly operations using only action recognition is 65 %, while that using only object detection is 55 %. For HRC disassembly experiments, effective planning of robotic decision-making can significantly reduce the occurrence of intention recognition errors, thereby preventing disassembly failures caused by inaccuracies in a single recognition method.

Finally, it is necessary to conduct a pure manual disassembly experiment for EV batteries to record the total time for the entire disassembly process by comparing the HRC disassembly method. As shown in Figs. 19(a) and 19(c), the time required for the steps Remove the upper cover, Remove the bolts, Remove the connecting plate, and Remove the battery cells is similar between the HRC and purely manual disassembly modes, as these operations are performed solely by the human operator in both cases. However, a significant time difference exists in the Loosen the bolts step. In the HRC mode, the robot, enabled by precise intention recognition, performs the bolt-loosening operation synchronously with the human operator, substantially reducing the time required for this step. Figs. 19(b) and 19(d) present the time statistics and standard deviation (visually magnified threefold in the chart) for each step under the two modes. The standard deviation for the Loosen the bolts step is notably larger than for other steps in both modes, which is attributable to the inherent nature of the task. When loosening multiple sets of bolts, factors such as the operator's reaction delay or variations in operation introduce variability, resulting in a more dispersed time distribution. In contrast, the standard deviation for the Loosen the bolts step in HRC disassembly is smaller compared to the purely manual approach. This reduction is because the robot operates based on a fixed program, leading to more consistent performance and thereby decreasing the dispersion in time for this specific phase.

To quantitatively evaluate the significant efficiency advantage of the

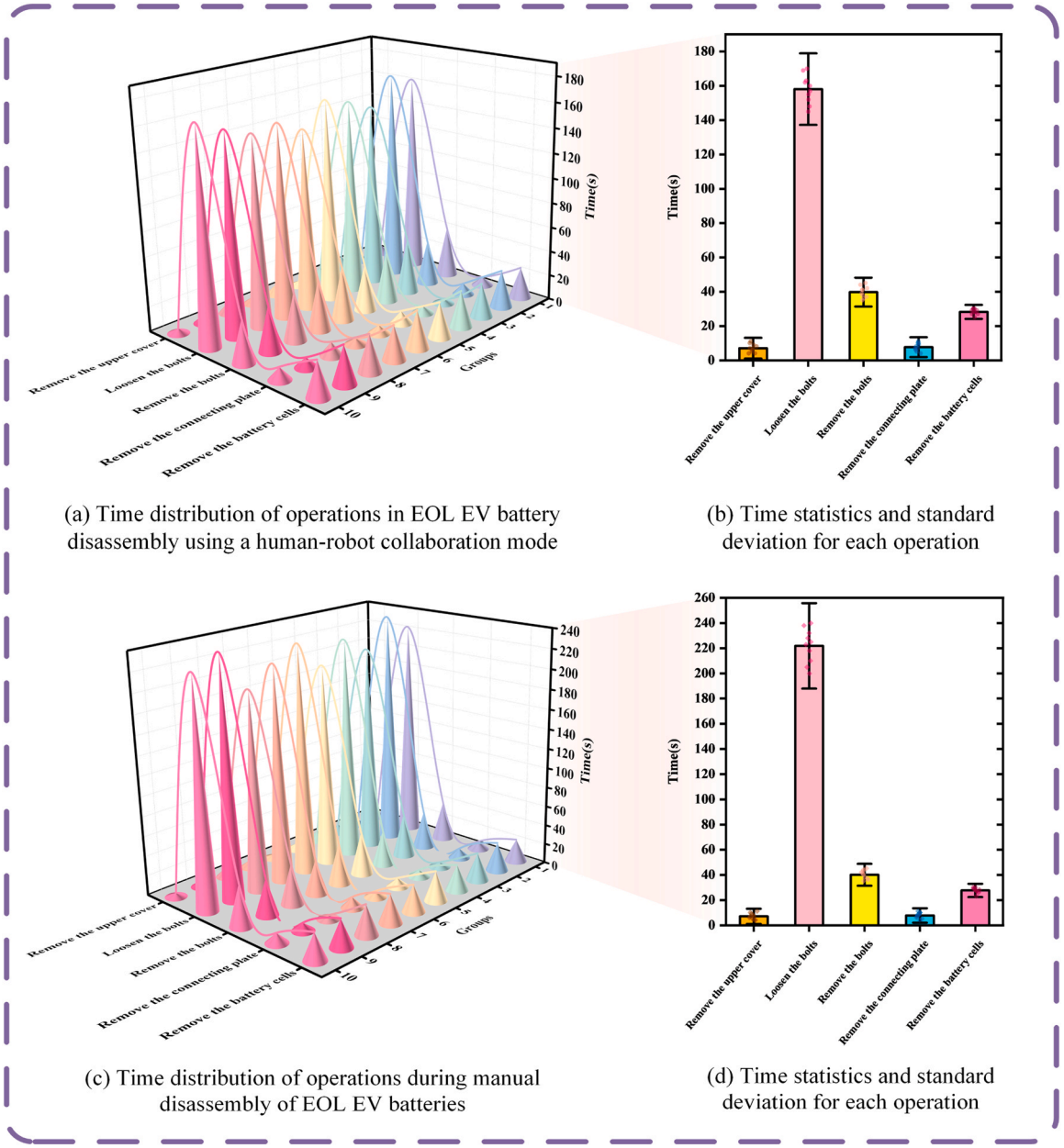

Fig. 19. The comparative analysis of temporal efficiency between HRC and manual disassembly modalities.

HRC disassembly mode over the traditional purely manual disassembly mode, a detailed statistical analysis was conducted on the overall disassembly times for both modes, with the results summarized in Table 6. The analysis shows that the HRC mode had an average disassembly time of 238.300 s with a standard deviation of 5.056 s, indicating high stability and repeatability. In contrast, the purely manual mode required an average of 298.700 s, with a standard deviation of 10.188 s, reflecting not only lower efficiency but also greater variability in operation time. Subsequently, an independent samples *t*-test was performed to test the null hypothesis that "there is no difference in the overall mean disassembly time between the HRC mode and the manual mode". The test yielded a t-value of −16.794 with 13.180 ° of freedom. The most critical indicator, the p-value, was calculated to be less than 0.001. Based on the commonly adopted significance level of $\alpha = 0.05$, this p-value provides extremely strong evidence to reject the null hypothesis, indicating a statistically significant difference in the overall mean times between the HRC disassembly mode and the manual mode. This conclusion is further supported by the 95 % confidence intervals: the HRC mode's interval was (235.166, 241.434) seconds, while the manual mode's interval was (292.386, 305.014) seconds. The complete lack of overlap between these two intervals further confirms the statistically significant difference in mean times between the two modes. The analysis results indicate that HRC disassembly of the EV battery is nearly one minute faster than pure manual disassembly. Additionally, the experiment design provides a limited number of components with only eight sets of bolts and nuts, which are significantly fewer than those of actual battery disassembly. Therefore, in practical disassembly scenarios involving an EV battery as an example, the HRC disassembly approach has an obvious advantage for disassembly operations that is more than one minute faster than manual disassembly.

To thoroughly analyze the advantages of our proposed multimodal fusion intention recognition method, a comprehensive comparison with existing approaches is necessary, as shown in Table 7. Although direct quantitative comparison is infeasible due to differences in datasets and experimental setups, this qualitative analysis effectively demonstrates the superior capabilities of our integrated framework. Current mainstream research primarily adopts dual modal fusion approaches, typically combining skeleton-based action recognition and object detection. For example, Zhang et al. [56] integrated ST-GCN with an improved YOLOX model, employing rule-based reasoning to fuse action and component information. While effective and emphasizing real-time performance, the authors themselves noted the limited confidence of relying solely on action recognition, indicating the need for richer contextual information to stabilize intention inference. Similarly, Liu et al. [57] adopted a multi-task framework combining ST-GCN-LSTM and YOLOv3 for HRI, demonstrating effectiveness across datasets but overlooking critical human factors such as facial attention cues, which are essential for predicting operator state and safety risks. Li et al. [58] focused on enhancing cross-scenario skeleton recognition through domain adaptation, while another study by Liu et al. [59] achieved progress in skeleton feature fusion and fine-grained assembly action recognition, while their methodology remains confined to single-modality skeleton data.

From a comparative perspective, our proposed multimodal fusion framework integrates action, tool usage, and facial angle perception. This holistic design enables effective integration of human factors signals and task context, forming a comprehensive scene-level intention and safety inference system. We argue that facial angle recognition is a critical component, allowing the system to infer not only what action is being performed but also to assess the operator's awareness and focus. This capability is vital for proactive safety intervention. Consequently, our framework provides a more robust, context-aware, and safety-oriented solution for intelligent disassembly scenario understanding in EV battery recycling.

5. Discussion

Considering the challenges of the implementation of HRC disassembly intent recognition for EOL products, the authors present a human behavior and intent recognition approach to solve the multi-scenario disassembly understanding in the paper. The proposed method provides the following advantages:

- First, regarding dataset construction, we developed a multi-modal annotated dataset specifically designed for disassembling retired power batteries, encompassing skeletal sequences, tool images, and facial orientation data. Through rigorous annotation protocols and diverse scenario collection, the dataset ensures richness and representativeness. It not only addresses the absence of benchmark data in EOL product disassembly but also provides valuable methodological references for other industrial human-robot interaction scenarios.
- Second, in algorithm development, we designed a real-time multi-modal fusion framework integrating three core modules: action recognition based on ST-GCN, tool detection using an improved YOLO algorithm, and lightweight facial angle recognition. By employing feature-level and decision-level fusion strategies, the framework achieves complementary verification of cross-modal information, effectively resolving intent ambiguity inherent in unimodal systems while maintaining real-time computational efficiency.
- Third, for system validation, we implemented complete system verification on a physical HRC disassembly testbed. The platform integrates industrial-grade collaborative robots and multi-source vision sensors, capable of simulating the dynamic characteristics of real disassembly environments. Empirical testing with EV battery packs demonstrated the system's adaptability in complex scenarios, providing a solid foundation for industrial application and broader implementation.
- Finally, regarding evaluation, we established a multi-dimensional assessment framework covering accuracy and efficiency metrics. Experiments show that the multi-modal intent recognition approach achieved approximately 85 % accuracy, significantly outperforming unimodal baselines. The HRC disassembly process demonstrated 20 % time reduction compared to manual operations with improved stability, providing crucial benchmarks for performance standardization of intelligent disassembly systems.

Due to the complexity of the EOL product disassembly environment and the diversity of disassembly tasks, there are still potential research points that can be further improved. The future work can mainly focus on the following aspects:

- The HRC disassembly primarily focuses on the development and application of human intent recognition while not imposing high demands on robot capabilities. Furthermore, relatively simple tasks of disassembly operations can be designed according to the real-world disassembly environments with uncertainties, including disassembly structure, unexpected component conditions, and operator behaviors [60]. Therefore, it is necessary for more

Table 6
The statistical analysis of the overall disassembly time for the two modes.

Disassembly mode	Mean	Standard deviation	Confidence interval (95 %)	T-value	Degree of freedom	P-value
HRC disassembly	238.300	5.056	(235.166, 241.434)	−16.794	13.180	< 0.001
Manual disassembly	298.700	10.188	(292.386, 305.014)			

Table 7
The comparison of human intent recognition methods for HRC disassembly.

Ref.	Application scenario	Recognition methods	Intention inference strategy	Main achievements
Zhang et al. [56]	Decelerator HRC assembly	– ST-GCN (skeleton) – Improved YOLOX (component)	Fusion inference based on predetermined rules	Effectively identify the operator's intention and accelerate HRC assembly
Liu et al. [57]	General HRI	– ST-GCN-LSTM (skeleton) – YOLOv3 (object)	Real time fusion inference of dual modal data streams	Real time interaction between robots and humans has been achieved
Li et al. [58]	Aircraft bracket HRC assembly	– ST-GCN (skeleton) – Improved ResNet	Embedded data fusion layer in the framework and knowledge transfer	Through the dynamic decision-making of robots, HRC assembly can be achieved
Liu et al. [59]	HRC assembly	– Multi-scale and multi-stream GCN (skeleton)	Deep extraction of action features	Achieve good performance on the NTU 60 dataset
Our proposed method	EOL EV battery HRC disassembly	– ST-GCN (skeleton) – Improved YOLO (tool) – Improved CNN (Facial angle)	Real time fusion inference of multimodal information based on rules	Identify the disassembly intention and the cognitive state of the operator, achieve active assistance of the robot, and improve efficiency and safety.

advanced algorithms and adaptive systems to handle diverse and unpredictable scenarios effectively, which could explore the possibility of enabling robots to perform more complex operations beyond simple tasks. Additionally, it is necessary to further explore HRC disassembly to perform disassembly tasks, which could enhance efficiency and flexibility of large-scale or complex disassembly operations, involving improving scalability and optimizing workflows of HRC disassembly.

- The disassembly process addresses mainstream EOL product disassembly structures currently, while not considering specialized disassembly models with unique designs. Future research could focus on developing the HRC disassembly plans to accommodate a wider range of EOL product models with special configurations. Additionally, this study does not examine the disassembly sequence for further exploration in future work. The optimal disassembly sequence can be used to determine the most efficient disassembly tasks, which could improve the overall efficiency of the disassembly process [61]. Furthermore, task allocation between humans and robots could be optimized to enhance the interactions with streamlined workflows and reduce downtime, which offers valuable directions for improving HRC disassembly tasks.
- Furthermore, this study did not optimize the motion trajectory of the robot in terms of its control aspects. Future research could focus on refining the movement trajectories of the robot disassembly to the specific disassembly environment of EOL products. By optimizing motion paths, the robot could operate more efficiently to reduce the unnecessary movements that can be used to improve the overall performance [62]. Furthermore, it is a potential trend to enhance the safety and precision of the disassembly process, particularly when dealing with delicate components or complex product structure disassembly. Therefore, the disassembly control optimizations would contribute to more effective and reliable HRC disassembly tasks for future complex intelligent decision-making systems.
- Finally, Augmented Reality (AR) and Virtual Reality (VR) technologies show considerable potential for applications in industrial disassembly, forming a highly complementary relationship with the multimodal visual perception framework proposed in this paper. In disassembly guidance, AR can be employed through mobile or wearable devices to overlay virtual sequences, tool usage instructions, and hazard warnings directly into the operator's real-world view, offering intuitive visual assistance. In remote collaboration scenarios, the integration of VR and AR can establish an immersive teleoperation environment. At the level of human-robot interaction, AR/VR technologies can further extend the interactive capabilities of the visual perception framework introduced in this study. The deep integration of AR/VR with human-robot collaborative disassembly is expected to significantly advance the intelligence of disassembly operations.

6. Conclusion

This paper proposed a human-robot collaborative intent recognition method to improve the disassembly efficiency of End-Of-Life (EOL) products. By integrating action recognition, object detection, and face angle detection, the paper optimizes traditional intent recognition methods to address the complexity of EOL product disassembly environment. Similarly, the targeted datasets are created, while the target and face angle detection algorithms are optimized to improve detection speed and accuracy. The experimental results show that the proposed intent recognition method outperforms the accuracy of traditional algorithms by successfully enabling efficient HRC disassembly of EOL products. This work presents an innovative solution for automating EOL product disassembly and offers significant practical application value. However, our study still has certain limitations, including a relatively small dataset scale, a limited number of experimental participants, and the lack of extreme condition testing. In future work, we will focus on expanding the dataset, enhancing the experimental design, and exploring integration with immersive AR/VR-guided disassembly systems.

CRediT authorship contribution statement

Jinhua Xiao: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sergio Terzi:** Writing – review & editing. **Kaile Huang:** Writing – original draft, Data curation. **Bo Wang:** Visualization, Software. **Marco Macchi:** Writing – review & editing. **Wei Wang:** Writing – review & editing.

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.

Acknowledgements

This work has been supported by the project "European Lighthouse to Manifest Trustworthy and Green AI" (ENFIELD) from the European Union's Horizon Europe Research and Innovation Program under grant agreement No. 101120657.

Data availability

Data will be made available on request

References

[1] Gunji BM, Pabba SK, Rajaram IRS, Sorakayala PS, Dubey A, Deepak BBVL, et al. Optimal disassembly sequence generation and disposal of parts using stability graph cut-set method for End of Life product. Sadhana Acad Proc Eng Sci 2021;46. https://doi.org/10.1007/s12046-020-01525-9.

[2] Guo X, Fan C, Zhou M, Liu S, Wang J, Qin S, et al. Human–Robot Collaborative Disassembly Line Balancing Problem With Stochastic Operation Time and a Solution via Multi-Objective Shuffled Frog Leaping Algorithm. IEEE Trans Autom Sci Eng 2023:1–12. https://doi.org/10.1109/TASE.2023.3296733.

[3] Chen Y, Liao HY, Behdad S, Hu B. Human activity recognition in an end-of-life consumer electronics disassembly task. Appl Erg 2023;113:104090. https://doi.org/10.1016/j.apergo.2023.104090.

[4] Poschmann H, Brüggemann H, Goldmann D. Disassembly 4.0: A Review on Using Robotics in Disassembly Tasks as a Way of Automation. ChemIngTech 2020;92:341–59. https://doi.org/10.1002/cite.201900107.

[5] Amirnia A, Keivanpour S. A context-aware real-time human-robot collaborating reinforcement learning-based disassembly planning model under uncertainty. Int J Prod Res 2024;62:3972–93. https://doi.org/10.1080/00207543.2023.2252526.

[6] Liu Z, Liu Q, Xu W, Liu Z, Zhou Z, Chen J. Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing. Procedia CIRP 2019;83:272–8. https://doi.org/10.1016/j.procir.2019.04.080.

[7] Liu W, Liang X, Zheng M. Task-Constrained Motion Planning Considering Uncertainty-Informed Human Motion Prediction for Human-Robot Collaborative Disassembly. IEEE/ASME Trans Mechatron 2023;28:2056–63. https://doi.org/10.1109/TMECH.2023.3275316.

[8] Schäfer J, Singer R, Hofmann J, Fleischer J. Challenges and Solutions of Automated Disassembly and Condition-Based Remanufacturing of Lithium-Ion Battery Modules for a Circular Economy. Procedia Manuf 2020;43:614–9. https://doi.org/10.1016/j.promfg.2020.02.145.

[9] Zhang H, Zhang Y, Wang Z, Zhang S, Li H, Chen M. A novel knowledge-driven flexible human–robot hybrid disassembly line and its key technologies for electric vehicle batteries. J Manuf Syst 2023;68:338–53. https://doi.org/10.1016/j.jmsy.2023.04.005.

[10] Xiao J, Anwer N, Li WD, Eynard B. Dynamic Bayesian Network-based Disassembly Sequencing Optimization for Electric Vehicle Battery. CIRP J Manuf Sci Technol 2022;38:824–35. https://doi.org/10.1016/j.cirpj.2022.07.010.

[11] Baazouzi S, Rist FP, Weeber M, Birke KP. Optimization of disassembly strategies for electric vehicle batteries. Batteries 2021;7. https://doi.org/10.3390/batteries7040074.

[12] Xia L, Hu Y, Pang J, Zhang X, Liu C. Leveraging Large Language Models to Empower Bayesian Networks for Reliable Human-Robot Collaborative Disassembly Sequence Planning in Remanufacturing. IEEE Trans Ind Inform 2025;21:3117–26. https://doi.org/10.1109/TII.2024.3523551.

[13] Asif S, Callari TC, Khan F, Eimontaite I, Hubbard EM, Bahraini MS, et al. Exploring tasks and challenges in human-robot collaborative systems: A review. Robot Comput Integr Manuf 2026;97:103102. https://doi.org/10.1016/j.rcim.2025.103102.

[14] Zhang W, Li Y, Wang K, Xu W, Gao L. A green and efficient disassembly line balancing with human-robot collaboration and destructive disassembly. Robot Comput Integr Manuf 2026;97:103081. https://doi.org/10.1016/j.rcim.2025.103081.

[15] Bahubalendruni MVAR, Varupala VP. Disassembly Sequence Planning for Safe Disposal of End-of-Life Waste Electric and Electronic Equipment. Natl Acad Sci Lett 2021;44:243–7. https://doi.org/10.1007/s40009-020-00994-0.

[16] Anil Kumar G, Bahubalendruni MVAR, Prasad VSS, Sankaranarayanasamy K. A multi-layered disassembly sequence planning method to support decision making in de-manufacturing. Sadhana Acad Proc Eng Sci 2021;46. https://doi.org/10.1007/s12046-021-01622-3.

[17] Lander L, Tagnon C, Nguyen-Tien V, Kendrick E, Elliott RJR, Abbott AP, et al. Breaking it down: A techno-economic assessment of the impact of battery pack design on disassembly costs. Appl Energy 2023;331:120437. https://doi.org/10.1016/j.apenergy.2022.120437.

[18] Lee ML, Behdad S, Liang X, Zheng M. Task allocation and planning for product disassembly with human–robot collaboration. Robot Comput Integr Manuf 2022;76:102306. https://doi.org/10.1016/j.rcim.2021.102306.

[19] Hellmuth JF, DiFilippo NM, Jouaneh MK. Assessment of the automation potential of electric vehicle battery disassembly. J Manuf Syst 2021;59:398–412. https://doi.org/10.1016/j.jmsy.2021.03.009.

[20] Wegener K, Andrew S, Raatz A, Dröder K, Herrmann C. Disassembly of electric vehicle batteries using the example of the Audi Q5 hybrid system. Procedia CIRP 2014;23:155–60. https://doi.org/10.1016/j.procir.2014.10.098.

[21] Gao J, Wang G, Xiao J, Zheng P, Pei E. Partially observable deep reinforcement learning for multi-agent strategy optimization of human-robot collaborative disassembly: A case of retired electric vehicle battery. Robot Comput Integr Manuf 2024;89:102775. https://doi.org/10.1016/j.rcim.2024.102775.

[22] Belhadj I, Aicha M, Aifaoui N. Product disassembly planning and task allocation based on human and robot collaboration. Int J Inter Des Manuf 2022;16:803–19. https://doi.org/10.1007/s12008-022-00908-y.

[23] Laili Y, Tao F, Pham DT, Wang Y, Zhang L. Robotic disassembly re-planning using a two-pointer detection strategy and a super-fast bees algorithm. Robot Comput Integr Manuf 2019;59:130–42. https://doi.org/10.1016/j.rcim.2019.04.003.

[24] Al GA, Martinez-Hernandez U. Safe multi-channel communication for human–robot collaboration. Robot Comput Integr Manuf 2026;97:103109. https://doi.org/10.1016/j.rcim.2025.103109.

[25] Prakash J, Altnji S, Thiyagarajan K, Nanda JS, Biswas A, Chowdhury AR. Human-Aware Robot Collaborative Task Planning Using Artificial Potential Field and DQN Reinforcement Learning. IEEE Access 2025;13:140889–99. https://doi.org/10.1109/ACCESS.2025.3595995.

[26] Asif ME, Rastegarpanah A, Stolkin R. Robotic disassembly for end-of-life products focusing on task and motion planning: A comprehensive survey. J Manuf Syst 2024;77:483–524. https://doi.org/10.1016/j.jmsy.2024.09.010.

[27] Ma Y, Tang D, Zhu H, Cai Q, Zhang Z, Wang L, et al. Probing AR-assisted seamless HRC assembly for industry 5.0: Multi-modal mutual cognition and LLM-driven knowledge reasoning. Robot Comput Integr Manuf 2026;97. https://doi.org/10.1016/j.rcim.2025.103112.

[28] Simões AC, Pinto A, Santos J, Pinheiro S, Romero D. Designing human-robot collaboration (HRC) workspaces in industrial settings: A systemic literature review. J Manuf Syst 2022;62:28–43. https://doi.org/10.1016/j.jmsy.2021.11.007.

[29] Khan AA, Andrev M, Murtaza MA, Aguilera S, Zhang R, Ding J, et al. Saf Aware Task Plan via Large Lang Models Robot 2025.

[30] Cui J, Forssberg E. Mechanical recycling of waste electric and electronic equipment: A review. J Hazard Mater 2003;99:243–63. https://doi.org/10.1016/S0304-3894(03)00061-X.

[31] Yao B, Li X, Ji Z, Xiao K, Xu W. Task reallocation of human-robot collaborative production workshop based on a dynamic human fatigue model. Comput Ind Eng 2024;189:109855. https://doi.org/10.1016/j.cie.2023.109855.

[32] Edis EB. Constraint programming approaches to disassembly line balancing problem with sequencing decisions. Comput Oper Res 2021;126:105111. https://doi.org/10.1016/j.cor.2020.105111.

[33] Xavier DM, Silva NBF, Branco KRLJC. Path-following Algorithms Comparison using Software-in-the-Loop Simulations for UAVs. J Intell Robot Syst Theory Appl 2022;106:63. https://doi.org/10.1007/s10846-022-01764-4.

[34] Rahman SMM, Ikeura R, Hayakawa S, Sawai H. Design and control of a power assist system for lifting objects based on human operator's weight perception and load force characteristics. IEEE Trans Ind Electron 2011;58:3141–50. https://doi.org/10.1109/TIE.2010.2087291.

[35] Malik AA, Brem A. Digital twins for collaborative robots: A case study in human-robot interaction. Robot Comput Integr Manuf 2021;68:102092. https://doi.org/10.1016/j.rcim.2020.102092.

[36] Faisal MAA, Abir FF, Ahmed MU, Ahad MAR. Exploiting domain transformation and deep learning for hand gesture recognition using a low-cost dataglove. Sci Rep 2022;12:1–15. https://doi.org/10.1038/s41598-022-25108-2.

[37] Ktistakis S, Gimeno L, Laftissi FZ, Hoss A, De Donno A, Meboldt M. Robot assistance primitives with force-field guidance for shared task collaboration. Robot Comput Integr Manuf 2025;96:103061. https://doi.org/10.1016/j.rcim.2025.103061.

[38] Xu J, Li Y, Xu L, Peng C, Chen S, Liu J, et al. A Multi-Mode Rehabilitation Robot with Magnetorheological Actuators Based on Human Motion Intention Estimation. IEEE Trans Neural Syst Rehabil Eng 2019;27:2216–28. https://doi.org/10.1109/TNSRE.2019.2937000.

[39] Kong Y, Fu Y. Human Action Recognition and Prediction: A Survey, 130. Springer US; 2022. https://doi.org/10.1007/s11263-022-01594-9.

[40] Wang C, Yan J. A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. IEEE Access 2023;11:53880–98. https://doi.org/10.1109/ACCESS.2023.3282311.

[41] Krestenitis M, Passalis N, Iosifidis A, Gabbouj M, Krestenitis M, Passalis N, et al. Hum Action Recognit Using Recurr BagFeatures Pooling cite this Version Hal Id Hal03265177 Hum Action Recognit Using Recurr BagFeatures Pooling 2021.

[42] Zhang S, Wei Z, Nie J, Huang L, Wang S, Li Z. A Review on Human Activity Recognition Using Vision-Based Method. J Health Eng 2017;2017. https://doi.org/10.1155/2017/3090343.

[43] Sharma S, Singh S. Vision-based hand gesture recognition using deep learning for the interpretation of sign language. Expert Syst Appl 2021;182:115657. https://doi.org/10.1016/j.eswa.2021.115657.

[44] Xin W, Liu R, Liu Y, Chen Y, Yu W, Miao Q. Transformer for Skeleton-based action recognition: A review of recent advances. Neurocomputing 2023;537:164–86. https://doi.org/10.1016/j.neucom.2023.03.001.

[45] Shaikh MB, Chai D, Islam SMS, Akhtar N. Multimodal fusion for audio-image and video action recognition. Neural Comput Appl 2024;36:5499–513. https://doi.org/10.1007/s00521-023-09186-5.

[46] Mazzia V, Angarano S, Salvetti F, Angelini F, Chiaberge M. Action Transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recognit 2022;124:108487. https://doi.org/10.1016/j.patcog.2021.108487.

[47] Tu NA, Abu A, Aikyn N, Makhanov N, Lee MH, Le-Huy K, et al. FedFSLAR: A Federated Learning Framework for Few-shot Action Recognition. Proc - 2024 IEEE Winter Conf Appl Comput Vis Work WACVW 2024. 2024. p. 270–9. https://doi.org/10.1109/WACVW60836.2024.00035.

[48] Zhang X, Yi D, Behdad S, Saxena S. Unsupervised Human Activity Recognition Learning for Disassembly Tasks. IEEE Trans Ind Inform 2024;20:785–94. https://doi.org/10.1109/TII.2023.3264284.

[49] Jang YM, Mallipeddi R, Lee M. Identification of human implicit visual search intention based on eye movement and pupillary analysis. Use Model UseAdapt Inter 2014;24:315–44. https://doi.org/10.1007/s11257-013-9142-7.

[50] Singh R, Miller T, Newn J, Velloso E, Vetere F, Sonenberg L. Combining gaze and AI planning for online human intention recognition. Artif Intell 2020;284:103275. https://doi.org/10.1016/j.artint.2020.103275.

[51] Schlenoff C, Kootbally Z, Pietromartire A, Franaszek M, Foufou S. Intention recognition in manufacturing applications. Robot Comput Integr Manuf 2015;33:29–41. https://doi.org/10.1016/j.rcim.2014.06.007.

[52] Kamali Mohammadzadeh A, Alinezhad E, Masoud S. Neural-Network-Driven Intention Recognition for Enhanced Human–Robot Interaction: A Virtual-Reality-Driven Approach. Machines 2025;13. https://doi.org/10.3390/machines13050414.
[53] Yao G, Lei T, Zhong J. A review of Convolutional-Neural-Network-based action recognition. Pattern Recognit Lett 2019;118:14–22. https://doi.org/10.1016/j.patrec.2018.05.018.
[54] Kiguchi K, Hayashi Y. A study of EMG and EEG during perception-assist with an upper-limb power-assist robot. Proc IEEE Int Conf Robot Autom 2012:2711–6. https://doi.org/10.1109/ICRA.2012.6225027.
[55] Karayiannidis Y., Smith C., Kragic D. Mapping human intentions to robot motions via physical interaction through a jointly-held object. IEEE RO-MAN 2014 - 23rd IEEE Int Symp Robot Hum Interact Commun Human-Robot Co-Existence Adapt Interfaces Syst Dly Life, Ther Assist Soc Engag Interact 2014:391–7. ⟨https://doi.org/10.1109/ROMAN.2014.6926284⟩.
[56] Zhang Y, Ding K, Hui J, Lv J, Zhou X, Zheng P. Advanced Engineering Informatics Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly. Adv Eng Inform 2022;54:101792. https://doi.org/10.1016/j.aei.2022.101792.
[57] Liu C, Li X, Li Q, Xue Y, Liu H, Gao Y. Neurocomputing Robot recognizing humans intention and interacting with humans based on a multi-task model combining ST-GCN-LSTM model and YOLO model. Neurocomputing 2021;430:174–84. https://doi.org/10.1016/j.neucom.2020.10.016.
[58] Li S, Zheng P, Fan J, Wang L. Toward Proactive Human – Robot Collaborative Assembly: A Multimodal. IEEE Trans Ind Electron 2022;69:8579–88. https://doi.org/10.1109/TIE.2021.3105977.
[59] Liu D, Huang Y, Liu Z, Mao H, Kan P. Technical paper A skeleton-based assembly action recognition method with feature fusion for human-robot collaborative assembly. J Manuf Syst 2024;76:553–66.
[60] Yu W, Lv J, Zhuang W, Pan X, Wen S, Bao J, et al. Rescheduling human-robot collaboration tasks under dynamic disassembly scenarios: An MLLM-KG collaboratively enabled approach. J Manuf Syst 2025;80:20–37. https://doi.org/10.1016/j.jmsy.2025.02.015.
[61] Zheng C, Du Y, Xiao J, Sun T, Wang Z, Eynard B, et al. Semantic map construction approach for human-robot collaborative manufacturing. Robot Comput Integr Manuf 2025;91:102845. https://doi.org/10.1016/j.rcim.2024.102845.
[62] Liao HY, Chen Y, Hu B, Behdad S. Optimization-Based Disassembly Sequence Planning Under Uncertainty for Human–Robot Collaboration. J Mech Des 2023;145:1–9. https://doi.org/10.1115/1.4055901.