A SYSTEMATIC VISION-BASED METHODOLOGY FOR HOLISTIC SCENE UNDERSTANDING IN HUMAN-ROBOT COLLABORATION

JUNMING FAN

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Industrial and Systems Engineering

A Systematic Vision-Based Methodology For Holistic Scene Understanding in Human-Robot Collaboration

Junming Fan

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

December 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Junming Fan (Name of student)

Abstract

The next generation of the industry has depicted a visionary blue-print of human-centricity in futuristic manufacturing systems. In the modern manufacturing sector, there has already begun a dramatic shift from the traditional mode of mass production towards mass personalization, driven by the increasing prevalence of personalization culture and customization requirements. The conventional approach for mass production has predominantly relied on the use of automated production lines, along with machines and robots that operate on preprogrammed routines. Although this method has demonstrated effectiveness in the era of mass production, the lack of intelligence and flexibility largely restrict its capacity to dynamically adjust to the frequently changing production schedule and specifications typical in mass personalization scenarios. To mitigate these limitations, human-robot collaboration (HRC) has emerged as an advanced manufacturing paradigm and is gaining traction as a promising solution to mass personalization since it can simultaneously leverage the consistent strength and repetitive precision of robots and the flexibility, creativity, and versatility of humans.

Over the past decade, a considerable amount of research efforts have been dedicated to HRC, addressing issues such as system architecture, collaboration strategy planning, and safety considerations. Among these topics, context awareness has drawn significant attention as it forms the bedrock of

critical functionalities such as collision avoidance and robot motion planning. Existing research works in context awareness have extensively concentrated on certain aspects of human recognition, such as activity recognition and intention prediction, due to the paramount importance of human safety in HRC systems. Nevertheless, there has been a noticeable shortage in addressing other vital components of the HRC scene, which can also substantially influence the collaborative working process. In order to fill this gap, this thesis aims to provide a systematic vision-based methodology for holistic scene understanding in HRC, which takes into account the cognition of HRC scene elements including *1) objects*, *2) humans*, and *3) environments*, coupled with *4) visual reasoning* to gather and compile visual information into semantic knowledge for subsequent robot decision-making and proactive collaboration. In this thesis, the four aspects are examined and potential solutions are explored to demonstrate the applicability of the vision-based holistic scene understanding scheme in HRC settings.

Firstly, a high-resolution network-based two-stage 6-DoF (Degree of Freedom) pose estimation model is constructed to enhance the object perception skill for subsequent robotic manipulation and collaboration strategy planning. Given the visual observation of an industrial workpiece, the first stage makes a coarse estimation of the 6-DoF pose to narrow down the solution space, and the second stage takes the coarse result along with the original image to refine the pose parameters to produce finer estimation results. In HRC scenarios, the workpieces are frequently manipulated by human hands, leading to another issue – the hand-object occlusion. Regarding this problem, an integrated hand-object 3D dense pose estimation model is designed with an explicit occlusion-aware training strategy aiming to mitigate the occlusion-related accuracy degradation (Chapter 3).

Then a vision-based human digital twin (HDT) modelling approach is explored in the HRC scenarios, hoping to serve as a holistic and centralized digital representation of human operator status for seamless integration into the cyber-physical production system (Chapter 4). The proposed HDT model is primarily composed of a convolutional neural network designed to concurrently monitor various aspects of hierarchical human status, including 3D human posture, action intention, and ergonomic risk assessment. Subsequently, based on the HDT information, a novel robotic motion planning strategy is introduced, which is focused on the adaptive optimization of the robotic motion trajectory, aiming to enhance the effectiveness and efficiency of robotic movements in complex environments. The proposed HDT modelling scheme provides an exemplary solution of how to model various human states from vision data with a unified deep learning model in an end-to-end manner.

Thirdly, a research endeavour is devoted to the perception of the HRC environment, for which a multi-granularity HRC scene segmentation scheme is proposed, along with a specifically devised semantic segmentation network with a bunch of advanced network designs (Chapter 5). Traditional semantic segmentation models mostly rely on a single-granularity semantic level. This formulation cannot adapt to different HRC situations where the requirements of semantic granularity are diversified such as a close-range collaborative assembly task versus a robotic workspace navigation case. Aiming to address this issue, the proposed model is designed to provide a hierarchical representation of the HRC scene which can dynamically switch between different semantic levels to flexibly accommodate the constantly changing needs of various HRC tasks.

Lastly, a vision-language reasoning approach is investigated to take a step further from visual perception to human-like reasoning and understanding of the HRC situation (Chapter 6). To address the inherent ambiguity of sole vision-based human-robot communication such as unclear reference of target objects or action intentions, linguistic data is introduced to complement visual data in the form of a vision-language guided referred object retrieval model. Based on the retrieved target object location, a large language model-based robotic action planning strategy is devised to adaptively generate executable robotic action code via natural form language interaction with the human operator. The incorporation of vision-language data demonstrates a viable pathway to achieve complex reasoning to enhance embodied robotic intelligence and maximize HRC working efficiency.

Keywords: Human-robot collaboration, holistic scene understanding, smart manufacturing, computer vision.

List of Publications

Journal Papers

1. **Fan, J.**, Zheng, P. (2023). A vision-language reasoning approach for ambiguity mitigation in human-robot collaborative manufacturing. *Journal of Manufacturing Systems*. (under review)

2. **Fan, J.**, Zheng, P., & Lee, C. K. (2023). A vision-based human digital twin modelling approach for adaptive human-robot collaboration. *Journal of Manufacturing Science and Engineering*, 145(12), 121002.

3. **Fan, J.**, Zheng, P., Li, S., & Wang, L. (2022). An integrated hand-object dense pose estimation approach with explicit occlusion awareness for human-robot collaborative disassembly. *IEEE Transactions on Automation Science and Engineering*. 21, 147-156.

4. **Fan, J.**, Zheng, P., & Li, S. (2022). Vision-based holistic scene understanding towards proactive human-robot collaboration. *Robotics and Computer-Integrated Manufacturing*, 75, 102304.

Conference Papers

1. **Fan J.**, Zheng P., & Lee C. K. (2022). A multi-granularity scene segmentation network for human-robot collaboration environment perception. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2105-2110).

2. **Fan J.**, Li S., Zheng P., & Lee C. K. (2021). A high-resolution network-based approach for 6D pose estimation of industrial parts. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)* (pp. 1452-1457).

Co-authored Papers

1. Zhang, X., **Fan, J.**, Peng, T., Zheng, P., Zhang, X., & Tang, R. (2023). Multimodal data-based deep learning model for sitting posture recognition toward office workers' health promotion. *Sensors and Actuators A: Physical*, 350, 114150.

2. Zhang, X., **Fan, J.**, Peng, T., Zheng, P., Lee, C. K. M., & Tang, R. (2022). A privacy-preserving and unobtrusive sitting posture recognition system via pressure array sensor and infrared array sensor for office workers. *Advanced Engineering Informatics*, 53, 101690.

3. Zheng, P., Li, S., **Fan, J.**, Li, C., & Wang, L. (2023). A collaborative intelligence-based approach for handling human-robot collaboration uncertainties. *CIRP Annals*.

4. Li, S., Zheng, P., **Fan, J.**, & Wang, L. (2021). Toward proactive human-robot collaborative assembly: A multimodal transfer-learning-enabled

action prediction approach. *IEEE Transactions on Industrial Electronics*, 69(8), 8579-8588.

Acknowledgement

The journey to this academic milestone was paved by the encouragement and assistance of many, to whom I owe profound gratitude.

My heartfelt appreciation goes to my chief supervisor, Dr. Pai Zheng. His invitation to embark on this advanced academic journey arrived at a crossroads in my life, when I was feeling confused and unsure. Dr. Zheng's unwavering professionalism and passion were the beacons that guided me through this scholarly expedition. His dedication to mentoring, providing insightful guidance, and fostering an understanding of the broader academic landscape was tremendously important in maintaining my focus and resilience throughout this journey. Despite moments of self-doubt that often beset me, Dr. Zheng's enduring confidence and counsel always inspire me to progress my work and stay the course of the big picture. Collaborating with Dr. Zheng has been both an immense honour and a deeply rewarding experience.

My sincere thanks go to my co-supervisor Dr. Carman K.M. Lee. She has kindly provided me with the chance to be supported by the AiDLab, which made it possible for me to commence my PhD project. Prof. Lee's kind advice, constructive feedback, and the provision of resources, including experimental equipment and facilities, have been invaluable to my research.

My profound gratitude goes to my external advisor Prof. Charlie C.L. Wang at the University of Manchester. Prof. Wang has kindly offered me the opportunity to conduct a period of visiting in his esteemed research group, an experience that contributed significantly to my academic growth and left me with cherished memories.

Many thanks to all colleagues of the Research Group of AI for Industrial Digital Servitization (RAIDS). Their multifaceted support and company have greatly enriched my research experience.

My heartfelt love and thanks go to my family for their understanding and unconditional support.

This research is supported by the Laboratory for Artificial Intelligence in Design (AiDLab) project 2-1 under the AIR@InnoHK project, the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU15210222 and PolyU15206723), and the Research Committee of The Hong Kong Polytechnic University under Collaborative Departmental General Research Fund (G-UAMS).

Contents

List of Figures

List of Tables

1 Introduction

The futuristic manufacturing paradigm is envisioned as a human-centric one, where humans and intelligent machines work in symbiosis with ultimate flexibility. Meanwhile, driven by the growing demand for personalized products and the need to cater to diverse customization requirements, the shift from traditional mass production to mass personalization also necessitates a flexible automation mechanism. To fill in this gap, human-robot collaboration (HRC) is emerging as a key approach, combining the precision of robots with the creativity and adaptability of humans, making it well-suited for the dynamic needs of futuristic manufacturing systems. This chapter commences with an examination of the background information of this study, particularly focusing on different aspects of visual perception in the HRC domain. Subsequently, the significance of the research, its scope, and the objectives are outlined. Finally, the structure of this thesis is presented.

1.1 Background

The landscape of manufacturing is undergoing a transformative phase, transitioning from traditional mass production methods to a more nuanced approach of mass personalization. This evolution aims to cater to the dynamic and rapidly evolving consumer needs and market trends. In an attempt to enhance operational efficiency and productivity, industrial robots have been extensively integrated into manufacturing environments. However,

their limited flexibility and problem-solving capabilities pose significant challenges. In response to these limitations, the concept of HRC has gained significant attention within the manufacturing sector. Recent studies [1] highlight HRC's potential to mitigate these challenges by synergizing the respective strengths of humans and robots, exploiting the precision and strength of robotic systems, complemented by the adaptability and creative problem-solving abilities inherent in human operators. HRC teams function within a shared workspace, engaging in cooperative tasks, as depicted in [2], thereby enhancing the overall efficacy of the manufacturing process.

In collaborative workspaces, humans possess an innate ability to interpret environmental cues, a skill not inherent in their robotic partners. Significant research efforts have been devoted to enhancing context awareness in HRC, as reported in various studies [3, 4]. The objective of these endeavours is to equip robots with the capacity for environmental perception and reasoning with considerable accuracy, thereby augmenting both productivity and safety in human-robot interactive settings. Although the application of computer vision techniques has been prevalent in enhancing the cognitive functions of robots, these methods have predominantly concentrated on human cognition such as gesture and activity recognition to facilitate direct communication and prevent collisions in shared environments [1, 5, 6, 7]. However, there remains a noticeable lack of attention towards other crucial scene elements, which can also significantly influence the dynamics of collaborative work.

On the other hand, the topic of vision-based scene understanding has been explored predominantly within the field of computer vision in recent years [8, 9, 10]. These studies have concentrated extensively on the reconstruction of scene layouts and the arrangement of objects in general daily scenes

such as the living room and kitchen. Nevertheless, it is notable that these investigations primarily focus on the environmental structural aspects, with little attention to the integration of human-related cognition in industrial contexts, let alone HRC scenarios.

To connect the dots scattered in previous research works, this study aims at providing a holistic perspective for vision-based scene understanding in HRC applications, including computer vision-based cognition of 1) *object*, 2) *human*, 3) *environment*, and 4) *visual reasoning* to gather and compile visual information into semantic knowledge for subsequent robot decision-making and proactive collaboration.

Concerning the four key facets of scene understanding, a substantial body of research works have been reported. However, the majority of these studies have been focused on general application contexts rather than specifically addressing the nuances of HRC scenarios, which are distinct in their inherent challenges and complexities stemming from the need to bridge the cognition and intelligence gap between humans and robots. Therefore, this thesis will delve into identifying the unique challenges for the cognition of HRC scene elements, based on which improvement strategies and potential solutions will be tailored to each identified aspect.

1.2 Research Scope

In the study on holistic scene understanding, the primary focus centres on the recognition of HRC scene elements including objects, humans, and environments, as well as the ability for abstract reasoning about these visual

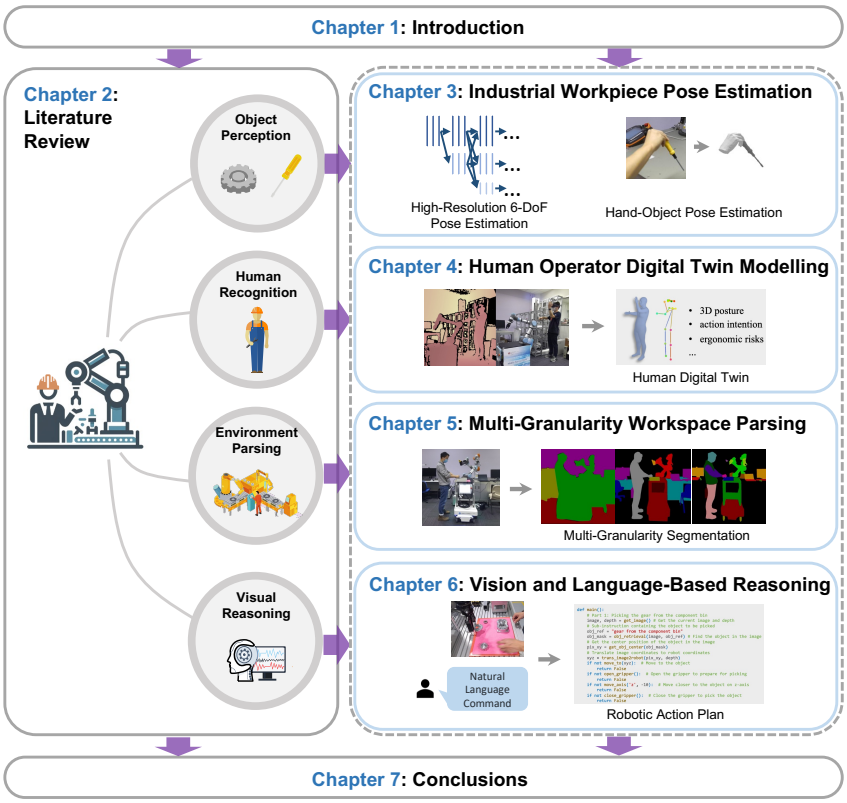


Fig. 1.1: Research scope of vision-based holistic scene understanding for HRC.

elements to achieve a comprehensive understanding of the HRC scene. The research scope and content organisation are shown in Fig. 1.1.

In the initial phase, the focus is on the perception of industrial objects within HRC contexts. It is crucial in HRC teamwork for the robot to be aware of the components being manipulated or the tools required to enable proactive decision-making for subsequent collaborative tasks. The initial efforts are directed towards the accurate estimation of the 6-DoF (Degree of Freedom) object pose, utilizing High-Resolution Networks (HRNet). Additionally, we further explore the joint 3D pose estimation of human hands and workpieces to address the prevalent issue of hand-object occlusion.

Subsequently, the attention will shift to the construction of a digital twin of the human operator, which aims to model the 3D dense posture of the

human body in real-time along with a series of other human factors such as action intentions and ergonomic risks. The primary research endeavours will be invested in building an end-to-end deep learning model that can simultaneously capture the multiple aspects of human status during the HRC working process. Based on the human body information, an adaptive robotic motion planning scheme will be explored to dynamically adjust the robotic movements in order to optimize human well-being without compromising the HRC system efficiency.

The third aspect pertains to environment parsing. Here, the emphasis is on representing the HRC scene in a hierarchical manner, which is crucial for providing dynamic semantic guidance and enabling flexible task fulfilment within collaborative environments. The research efforts will be devoted to the design and construction of such a hierarchical environment parsing model with multiple granularities of semantics.

Lastly, based on the scene information, a visual reasoning approach will be investigated, which takes into account both the visual context and linguistic cues in HRC scenarios. By endowing the collaborative robot with visual reasoning abilities, it can operate beyond the confines of predefined programming, reason in a human-like logical abstraction, and autonomously determine its actions based on the current status of HRC work and task specifications.

1.3 Research Objectives

Motivated by the limited exploration of HRC-oriented visual perception approaches, the primary aim of this research is to delve into four key facets of scene understanding in HRC, develop solutions to address existing limitations in these areas, and ultimately foster a holistic understanding of HRC scenarios. This ability of holistic scene understanding is pivotal for enhancing complex robotic reasoning and decision-making processes. Corresponding research objectives addressing the four aspects are stated as follows.

Objective 1: Development of a 6-DoF industrial object pose estimation method for HRC cases, with a special focus on the occlusion between workpieces and human hands.

An inevitable challenge in object recognition within HRC contexts is the occlusion resulting from hand-object interactions. Although the interplay between humans and objects is crucial, the explicit investigation of occluded hand-object perception remains underreported in the literature, especially regarding adaptive robot decision-making in close-proximity collaborations with partial hand-object occlusions. This study aims to first build a fundamental 6-DoF industrial object pose estimation model to enable adaptive robotic manipulation of workpieces in the HRC scenario. Then a further endeavour will be directed to explore the joint estimation of 3D hand-object pose in HRC cases with a specific consideration of the visual occlusion issue. Through real-time inference of hand-object poses, the intention is to effectively monitor the most vital and frequent situation—hand-object interactions—in HRC settings, thereby offering valuable insights for robotic actions.

Objective 2: Formulation of a vision-based reconstruction approach for human digital twin (HDT) modelling and adaptive robotic motion planning in HRC.

Human safety considerations make human perception a significant component of visual understanding in HRC. Despite extensive exploration of human perception in image processing, multisensory devices, and deep learning within HRC, the construction of a comprehensive human digital twin, especially from visual observations, remains underreported. As a centralized digital representation of various human data that can be easily and seamlessly integrated into a cyber-physical production system, HDT is of substantial importance, hence necessitating the exploration of related technologies. This study aims to propose an exemplary solution to construct a vision-based HDT that can concurrently monitor different facets of human states in an end-to-end manner with real-time performance, thereby facilitating subsequent adaptive robotic planning to proactively and timely respond to human movement and status fluctuations.

Objective 3: Establishment of a scene segmentation model for multi-granularity semantic understanding of the HRC environment.

Regarding visual environment parsing in HRC applications, the standard practice involves visual segmentation representing scene elements at a single-granularity semantic level. However, this approach falls short of detailed workspace modelling in advanced HRC systems, such as extremely flexible manufacturing floors. A more nuanced, hierarchical, and hybrid environmental representation is preferable. Consequently, this project seeks to introduce

a multi-granularity scene segmentation network, enabling the parsing and representation of HRC environments across varied semantic levels.

Objective 4: Exploration of a vision-language reasoning method for ambiguity mitigation in human-robot communication.

Achieving complex, human-like reasoning skills remains a main pursuit in artificial intelligence and robotics, particularly for human-robot collaborative manufacturing, to ensure reliable and effective collaboration between humans and robots. Preliminary research in HRC has utilized various methodologies, from mathematical to deep learning models, to achieve visual reasoning. However, these approaches often oversimplify reasoning tasks as direct mappings from visual or language cues to specific decisions or actions, neglecting the integration of knowledge, common sense, and vision-language observations. Regarding this issue, this study aims to propose a vision-language model that can intelligently locate the target object referred to by the language cue, and further leverage the outstanding reasoning capability and embedded world knowledge of Large Language Models (LLMs) to proactively and dynamically reason about the vision-language information and generate feasible action plans for collaborative robots to fulfil HRC tasks.

1.4 Thesis Structure

The rest content of this thesis is organized as follows:

Chapter 2 reviews previous works related to the four aspects of holistic scene understanding in the HRC context. Limitations and challenges that are not well addressed will be highlighted.

Chapter 3 first introduces the proposed high-resolution 6-DoF object pose estimation model and then extends the pose estimation to joint hand-object 3D pose estimation to cope with the notorious occlusion issue.

Chapter 4 presents the proposed vision-based human digital twin modelling approach for human operator monitoring in the HRC environment and demonstrates a feasible adaptive robotic motion planning strategy based on the real-time HDT information.

Chapter 5 depicts a multi-granularity scene segmentation model for HRC environment perception enhancement in hopes of providing a more flexible scene representation to accommodate fast changing HRC tasks.

Chapter 6 illustrates a vision-language reasoning approach that mainly consists of a vision-language-guided referred object retrieval model which can efficiently and intelligently locate the desired object specified by the human language instruction in the HRC scene. A large language model is then adopted to synthesize a reasonable robotic action plan based on the previously retrieved target object and the human language instruction.

Chapter 7 summarizes the accomplishments and contributions of this project and discusses future steps.

2

Literature Review

The literature review in this chapter delves into the key facets of HRC by exploring various methodologies for object perception, human recognition, environmental parsing, and visual reasoning within HRC systems. A critical analysis of the research gaps in Section 2.5 highlights challenges and potential research directions in the four aspects. This rigorous review aims to identify the opportunities and challenges for holistic scene understanding in HRC, ultimately guiding the direction for future research endeavours.

2.1 Object Perception for HRC

Objects, including workpieces, tools, etc., pervasively exist in HRC scenarios. It is fundamental during HRC assembly that the robot should be aware of where the ongoing assembly area is, what parts are still missing, and which tools should be used so that it can proactively make decisions about subsequent collaboration actions. This section mainly focuses on computer vision-based object perception methods that have been adopted in previous HRC-related works, since the topic of visual perception of general objects would be too vast to be covered in this review. The following discussion of vision-based object perception in HRC is divided into three key aspects: *identification*, *localization*, and *pose estimation*.

2.1.1 Object Identification

The most fundamental task of object perception in an HRC scenario is to identify what an object is and what attributes it has so that the robot can autonomously deduce the expected actions associated with the target object. Concretely, the task of object identification mainly has two aspects: 1) plainly classifying the objects into different categories such as wrenches, screws, gears, etc. and 2) elaborately identifying the affordances of objects based on the utilization or attribute such as grasping position, tool functionality, and so on. One example depicting the difference between classification and affordance is shown in Fig. 2.1. Object identification is beneficial during HRC because it allows the robot to autonomously understand which object serves what purpose and proactively carry out collaborative assistance without explicit programming or commanding. Existing works focusing on this task are listed in Table 2.1.

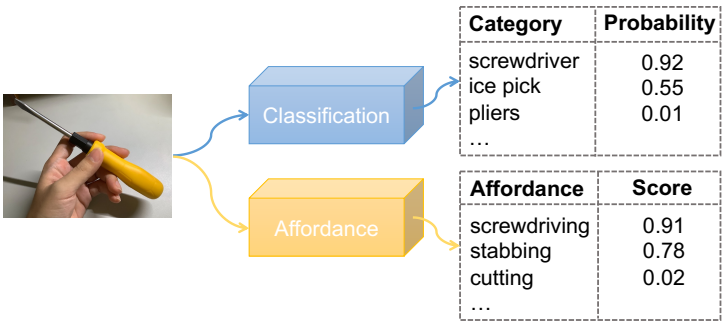


Fig. 2.1: A demonstration of the difference between classification and affordance identification.

1) Classification

Classification is a fundamental problem in computer vision and machine learning. Here the focus is mainly on the applications of object classification in HRC manufacturing. Ferreira et al. [11] reported a laser scanning-

based solution for spray coating in flexible robotic cells, which utilized laser scanning-based 3D reconstruction and K-Nearest Neighbors (KNN) for the classification of workpieces. In another work [12], the authors also adopted a laser range finder as the main sensor but utilized it to scan through the working space following a pre-defined path to generate grayscale images of the workstation, and further leveraged the invariant moments of Hu features along with KNN, Neural Networks (NN) and Support Vector Machine (SVM) to classify different objects. Although these methods can achieve very high precision owing to the laser sensor, the time-consuming nature of laser scanning largely restricts the application scenarios, and the KNN classifier with handcrafted features can be vulnerable when facing working environment fluctuations.

Table 2.1: Literature of object identification.

Category	Application	Key Elements	Source
Classification	Spray coating	Laser; KNN	[11]
	Pick and place	Laser; KNN	[12]
	Real-time recognition	LightNet; Multi-task	[13]
	Pick and place	Novel object; Multi-view CNN	[14]
	Pick and place	Novel object; ResNet50	[15]
	Welding	Pix2pix; AlexNet	[16]
	Robust recognition	CNN; SVM	[17]
Affordance	Pick and place	MobileNetv2; VGG16	[18]
	Pick and place	Canny; Hough line	[19]
	Grasping	CNN; Task-specific grasp	[20]
	Grasping	Seach-based learning	[21]
	Affordance detection	S-HMP; SRF	[22]
	Affordance detection	CNN	[23]
	Affordance detection	Encoder-decoder CNN	[24]

Being one of the most powerful classifiers, CNNs (Convolutional Neural Networks) are also commonly found to serve for object classification in recent manufacturing research. Zhi et al. [13] reported an approach for real-time object cognition in human-robot interaction applications and a lightweight CNN model called LightNet has been proposed to recognize 3D

objects by predicting the object class and orientation concurrently. Kasaei [14] proposed OrthographicNet which could handle 3D object recognition via multi-view CNN features and integrated human guidance to classify objects with novel categories. Dehghan et al. [15] also aimed at the novel object recognition problem and leverages human interactions to teach the robot novel object categories while training a ResNet-50 model. Feng et al. [16] made an attempt to apply CNNs to welding penetration status recognition utilizing pix2pix generative model for image denoising and AlexNet for image selection. Keller et al. [17] studied the influence of illumination for more robust object classification via CNN feature extractor and SVM classifier. These CNN-based approaches can accurately recognize industrial objects and maintain robustness to varied environments, but the notorious data-hungry issue might prevent them from being applied in many resource-constrained scenarios.

2) Affordance

The plain classification method is able to categorize target objects sufficiently enough for normal robotic applications, but for Proactive HRC objects should be identified in a more subtle way. The concept of affordance was first proposed in the field of perceptual psychology [25] and later introduced into robotics to represent the interactive properties of objects, such as where the grasping points are and what actions could be carried out with the objects. For example, D'Avella et al. [19] studied the problem of recognizing the picking points of objects in cluttered environments via Canny edge detector and Hough lines, while Nguyen et al. [18] reported a work to recognize the picking angle of objects in a robotic pick-and-place task, during which CNN models such as MobileNetv2 and VGG16 were adopted.

Kokic et al. [20] proposed a deep learning method for identifying which task an object could afford and how to grasp the object for specific tasks. Chatila et al. [21] adopted a search-based learning method to learn the affordance of environment objects such as grasp-ability for robotic interactions. Myers et al. [22] reported two approaches to learn the affordances of different tool parts including superpixel-based hierarchical matching pursuit (S-HMP) and structured random forests (SRF) to learn the affordances of different tool parts, while Nguyen et al. [23] also concentrated on this task but employed CNN models, which, according to the reported results, perform better than HMP and SRF-based methods. Thermos et al. [24] reported a solution to understand the affordances of daily objects that are being interacted with by humans, where an encoder-decoder CNN model was leveraged to jointly reason the affordance class and saliency map. These works for learning object affordance build a solid foundation for identifying objects beyond plain classification. Nevertheless, how to effectively learn the object affordance from unlabeled data during the robot operation process still remains a gap.

2.1.2 Object Localization

Another essential step of object perception is object localization, which means locating the objects of interest in the HRC environment and extracting their positions or coordinates in the image plane as shown in Fig. 2.2. It can be further converted to world coordinates if the cameras are calibrated. Existing works that leverage computer vision-based methods to tackle the object localization problem are classified into three categories based on the format of localized object position: 1) Detection, which represents the object positions with bounding boxes around the objects; 2) Segmentation, which localizes the objects based on their geometric information and output

pixel-level segmentation results; and 3) Others, which mainly rely on prior knowledge or geometrical information to locate the objects. Related works under this topic are listed in Table 2.2.

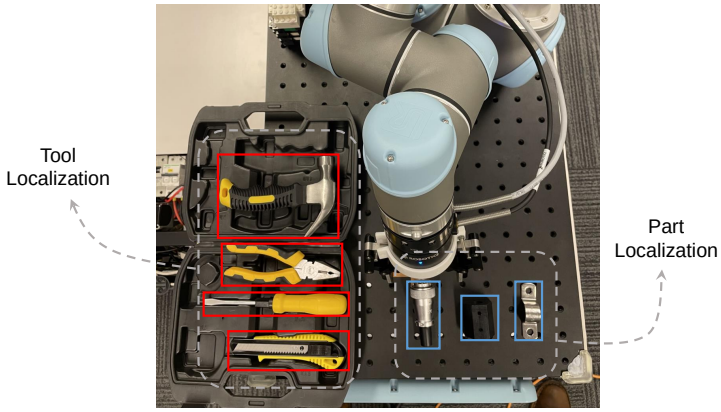


Fig. 2.2: Example of tools and parts localization.

1) Detection

Recently, CNN has emerged as the state-of-the-art method in object detection and many other computer vision tasks because of the ability to autonomously learn stronger feature representations over manually designed ones. One-stage CNN detection models, including You Only Look Once (YOLO) series [26, 32] and Single Shot Detector (SSD) series [28, 29], are prevailing in workpiece detection for their simplicity and efficiency. Two-stage models, specifically the R-CNN series, can gain better performance in applications with looser time constraints such as junior operator training [27] or with higher accuracy requirements such as detecting industrial components from a heavily cluttered background [30, 31].

Although CNNs are leading the trend in object detection, yet hand-crafted feature descriptors still have a place in industrial applications. CNN model can achieve better accuracy and robustness providing enough training data and can be easily accelerated via GPU (Graphics Processing Unit) because

Table 2.2: Literature of object localization.

Category	Application	Key Elements	Source
Detection	Workpiece	YOLOv3	[26]
	Assembly	Faster R-CNN	[27]
	Assembly	SSD	[28]
	Pick and place	Mobilenet SSD	[29]
	Pick and place	Mask R-CNN	[30]
	HRI	Mask R-CNN; Super-pixel	[31]
	Handover	YOLOv3	[32]
	Object handling	Haar-like features	[33]
	Workpiece	Marker based localization	[34]
	Robot system	LBP features	[35]
Segmentation	Object handling	Contour segmentation	[36]
	Assembly	Background substraction	[37]
	HRI	Color-based segmentation	[38]
	Welding	Sobel; Hough line	[39]
	Assembly	Edge and shape-based	[40]
Others	Object handling	3D CAD model	[41]
	PCB soldering	Shape matching	[42]
	Assembly	Distance-based rules	[43]
	Pick and place	Multi-sensory	[44]

of its parallel nature. However, in practical industrial applications, GPUs and large datasets are not always available, in which cases CNN generally performs poorly in terms of efficiency and reconfigurability, leaving space for hand-crafted features, which can provide an agile solution for novel scenarios by simply tweaking some parameters without heavy GPU training. Astanin et al. [33] aimed for a better detection rate of reflective metal workpieces for the application in flexible robotic cells, which is achieved via the multi-scale Viola-Jones detector integrated with Haar-like feature descriptor and decision trees. Castaman et al. [35] proposed an Unmanned Ground Vehicle (UGV) system with an LBP (Local Binary Pattern) feature descriptor applied to detect the wrench and valve. In another work [34] by Hsieh et al., printed markers are attached to target objects and the system only needs to detect the markers through multiple cameras to localize the targets. The main drawback of hand-

crafted feature-based detection methods is the limited detection performance, but they are the appropriate choice for proof-of-concept and applications in controlled environments.

In general, these detection models can produce excellent localization results for industrial objects owing to the simplified formulation of bounding box regression, which, however, is also regarded as a bottleneck since it cannot accurately describe the geometric appearance of different objects.

2) Segmentation

If the target object is easily separable from the background, simple image processing techniques such as background subtraction would suffice to locate and segment the object. Hoffmann et al. [36] exploited 3D depth information and level-set based contour segmentation method to extract the contour of tools held by a robot arm and locate the tooltips for the robot to further execute some subtle actions such as operating a hand drill or drawing with a pencil. In [37], Aliev et al. explored the segmentation of workpieces delivered by a MiR100 AGV under command from human operators via background subtraction. Jirak et al. [38] proposed a method to address object ambiguity during human-robot interaction by taking human pointing direction into consideration, and colour-based image segmentation is leveraged to segment the desired object.

Some other works leverage the prior information about the shape of the target object. Dinham et al. [39] adopted the Sobel edge detector and Hough Line Transform to segment welding seams based on the knowledge that they are most likely in the shape of lines, while Lee et al. [40] also detected

working-in-progress parts during the electric motor assembly process based on the Canny edge detector and shape information, and further transmitted the progress information to workers via Augmented Reality (AR) to achieve human-robot collaboration.

CNN-based methods have also been introduced into the segmentation task and made considerable progress in recent years. Back et al. [30] reported the adoption of Mask R-CNN in industrial component detection and segmentation with an RGB-D (Red Green Blue-Depth) data fusion and data synthesis strategy. Azagra et al. [31] presented a solution to incrementally teach the robot new objects via human-robot interaction, where Mask R-CNN is also employed for object segmentation. Although Mask R-CNN is equipped with the instance segmentation ability, the segmentation results only have relatively low resolution due to the model design and time efficiency concern. Another issue of segmentation approaches is that they can only extract the 2D silhouette but are unable to represent the 3D shape of the target object.

3) Others

Other works, which mainly rely on prior knowledge or geometrical information to locate the objects, do not fall into the aforementioned two categories and hence are classified as others. In some applications such as robotic product packaging and cooperative robotic soldering, the precise shapes or models of the target objects were provided [41, 42], in which cases image feature descriptors such as Speeded Up Robust Features (SURF) were combined with the object models to find the best matching target position in images.

As for the cases that do not require the exact positions but the rough occupancy of objects to achieve collision avoidance, depth sensors such as ToF (Time of Flight) cameras and LiDAR (Light Detection and Ranging) along with straightforward distance-based rules are proven to be sufficient [43, 44] without being overwhelmed by complex vision algorithms and computational overheads. The application of these methods is rather confined since the manual rule and algorithm process have to be redefined for specific scenarios.

2.1.3 Object Pose Estimation

In an HRC environment, the frequent human-robot interactions put a greater demand on the precision of object perception especially when the target objects are close to the human body. Object pose estimation serves as the missing puzzle piece towards autonomous robot manipulation since it could provide precise object postures in the form of a mapping between 3D object models and sensory observations. Literature of object pose estimation is listed in Table 2.3, and we further divide them into two categories based on the main input or feature source.

1) 2D Methods

2D RGB cameras, as the most available, applicable and affordable sensors, have encouraged a lot of attempts to tackle the 6-DoF pose estimation problem merely using 2D images as the input source, despite the 3D nature of the 6-DoF pose estimation task.

Table 2.3: Literature of object pose.

Category	Application	Key Elements	Source
2D	Assembly	3D edgelets	[45]
	Robot object handling	Shaver handles	[41]
	Assembly	Line2D; Canny	[46]
	Workpiece pose estimation	Line ending points; PnP	[47]
Point cloud	Robot system	Global descriptor	[48]
	Pick and place	Fast Point Feature Histogram	[49]
	Pick and place	PointNet++; YOLOv2	[50]
	Robot system	Mask R-CNN	[51]
	Robot system	PPF; ICP	[52]
	Assembly	OBB; ICP	[53]

The most straightforward approach to estimating 6-DoF pose from a 2D image is comparing and matching the 3D object model to 2D observation. Abdallah et al. [45] applied object pose estimation to inspect aeronautical assembly parts status by mapping the 3D CAD (Computer-Aided Design) model to observation images and ensure the presence and installing positions of parts. Meanwhile, Tsarouch et al. [41] leveraged the 3D CAD model of shavers to localize shaver handles during production lines for further robotic picking and placing.

Some other works move further along this path by exploiting more hand-crafted features from 2D images. For instance, Hagelskjaer et al. [46] proposed a method to obtain precise 6-DoF pose by first applying Lind2D matching algorithm to obtain coarse pose and employing Canny edge detector and scene-specific spatial constraints to refine the pose results. He et al. [47] also reported a hand-crafted feature-based method, which exploits the straight lines and ending points of metal parts and resorts to the PnP (Perspective-n-Point) solver to retrieve the 6-DoF pose results.

2) 3D Methods

Despite the benefits of only using 2D RGB cameras, depth information still weighs heavily in the process of precise 6-DoF pose estimation. A commonly adopted approach of exploiting depth information is to transform the depth or RGB-D images into point clouds.

Luo et al. [48] proposed a robotic system for manufacturing automation which captures RGB-D images via Kinect camera and simply transforms into point clouds as the input to the processing algorithm. It then estimates the 6-DoF pose of objects through geometry-based global feature descriptors. Efforts have also been made to exploit point-based features [49], which adopted the Fast Point Feature Histogram to estimate object poses for picking and placing applications.

Post-refinement of the estimated pose is a frequently used technique to obtain a more accurate 6-DoF pose result. Bedaka et al. [52] reported an automatic path-planning robotic system that leverages Point Pair Features (PPF) and Iterative Closest Point (ICP) algorithm to estimate the 6-DoF poses of manufacturing objects, while Franceschi et al. [53] proposed a method that adopts the Oriented Bounding Box (OBB) to obtain rough pose estimations and ICP for refinement in the assembly process of bulky components such as a sidewall panel of an aeroplane.

Although CNN-based 6-DoF pose estimation methods still struggle with mediocre performance, CNNs are not uncommon in related applications. Zhang et al. [50] introduced the PointNet++ model to extract key points from the point cloud to facilitate further pose estimation tasks. In another

work, Nguyen et al. [51] proposed a robotic system for decaking 3D-printed parts, in which Mask R-CNN was utilized for object localization and point cloud segmentation, based on which the object pose could be obtained via simple point cloud-based calculations.

Despite better pose estimation accuracy facilitated by the additional depth information, a common deficiency of these approaches is the lack of explicit consideration for object occlusion, which could severely damage the pose estimation performance of existing methods.

2.2 Human Recognition for HRC

Human, as the most essential participant of HRC, has been regarded as the main research subject by numerous research works for almost all aspects of human recognition that one can think of. It is reasonable since intelligence would never be too much for robots to recognize humans in an HRC working scene regarding human safety and collaboration efficiency. Concretely, in this section, three facets of human recognition are examined: *human localization*, *human activity*, and *human pose*.

2.2.1 Human Localization

To achieve effective human-robot collaboration, the position or location of human operators in an HRC scenario should be firstly localized, so that the robot could proactively plan its collaborative actions without colliding with human bodies. During the review process, it was found that previous works

in HRC about human position recognition mainly focused on two topics: human body detection and face detection. The former topic hopes to make the human body position fully acknowledged by robots, while the latter wants to achieve more than just localization but further verifying the human identity. Table 2.4 indicates related previous works.

1) Human detection

Safety is the most important factor that one should consider in the designing process of an HRC system. Collision avoidance, as the fundamental level of safety requirement, could be achieved by detecting human bodies in an HRC scene through various approaches. Shariatee et al. [54] proposed a safe collaboration method for collaborative assembly workstations that leverages image processing techniques such as edge detection and morphological filtering to segment human and robot positions from RGB-D images obtained by a Kinect camera, and further measure the distance between them to calculate danger index. Tashtoush et al. [55] followed a similar path by monitoring the HRC workspace via a top-view Kinect RGB-D camera, while the difference is that this work leveraged a specifically designed background-foreground algorithm to detect human bodies. The main drawback of these hand-crafted localization methods is they are barely able to distinguish between the human body and other obstacles of similar size, thus can only be applied in controlled environments.

Ample research works have also explored the utilization of CNN models for human body detection. Liu et al. [56] proposed a safety system for HRC assembly that localizes human body positions via RGB-D camera and represents human and robot occupancy via OctMap, and further recognizes

Table 2.4: Literature of human localization.

Category	Application	Key Elements	Source
Human body detection	HRC workstation	Segmentation; Danger index	[54]
	HRC workstation	B-F algorithm; Kinect	[55]
	Collaborative assembly	OctMap; CNN	[56]
	Safe HRC	FMG; CNN	[57]
	Pedestrian detection	Mask R-CNN; LiDAR	[58]
	Human following	MobileNet-SSD	[59]
	HRI	OpenPose; SVM	[60]
	Safe HRC	YOLO; Bayesian DNN	[61]
Face detection	Surveillance	HMD; Haar feature	[62]
	Collaborative assembly	AWS DeepLens	[63]
	Human following	SSD; FaceNet; KCF	[64]
	Real-time HRI	SFPD; Multi-task	[65]

human actions through CNN models. Meanwhile, Anvaripour et al. [57] proposed a collision detection method for proximal human-robot cooperation utilizing a wearable Force Myography band as the data source and a CNN model as the classifier. The prevailing CNN-based detection models have also been widely adopted to achieve active human following or collision avoidance such as Mask R-CNN [58], MobileNet-SSD [59], and YOLO [61], while human pose estimation models such as OpenPose have also been leveraged as human body detectors to facilitate human-robot communication [60]. Unlike objects, human bodies have fewer variations in terms of geometric features, which can substantially facilitate data collection and model training, and further enable CNN-based methods to be the preferred choice when it comes to human detection if detailed human posture and body shape are not required.

2) Face Detection

As face is the most distinguishable area of the human body, some works in HRC also resort to face detection to retrieve human position information and get an opportunity to identify human operators.

Do et al. [62] reported an HRC surveillance application that allows human operators to control a mobile robot via a head-mounted display (HMD) to patrol the environment, detect human faces via the Haar feature, and send face images back to the HMD for human operators to recognize face identities. Lazaro et al. [63] attached an AWS (Amazon Web Services) Deeplens to a YuMi robot in the HRC assembly process to detect and recognize human faces for identification and unexpected leaving detection. Hwang et al. [64] aimed to achieve specific person following for mobile robots by employing SSD detector to locate humans and FaceNet model to identify the target human face. Meanwhile, Fiedler et al. [65] focused more on the running time of face detection and hope to achieve real-time HRI (Human-Robot Interaction) via a specifically designed SFPD (Simultaneous Face and Person Detection) model, which can be regarded as a multi-task extension of SSD. Although face detection has attracted less attention than body detection in the current HRC field, it can serve as the indispensable pre-process for further identity authentication especially when multi-human is involved.

2.2.2 Human Activity

A large body of existing works has been devoted to human activity recognition, which is convinced to play a pivotal role in HRC since human actions could exhibit a certain ambiguity that makes it hard for robots to act accordingly or proactively. To tackle this problem, some researchers put major efforts into the activity recognition part, which mainly studies the recognition of ongoing

activities such as the ones illustrated in Fig. 2.3, while others devote more to the prediction side, which considers more about the future action intentions. Table 2.5 shows involved works and their key elements.

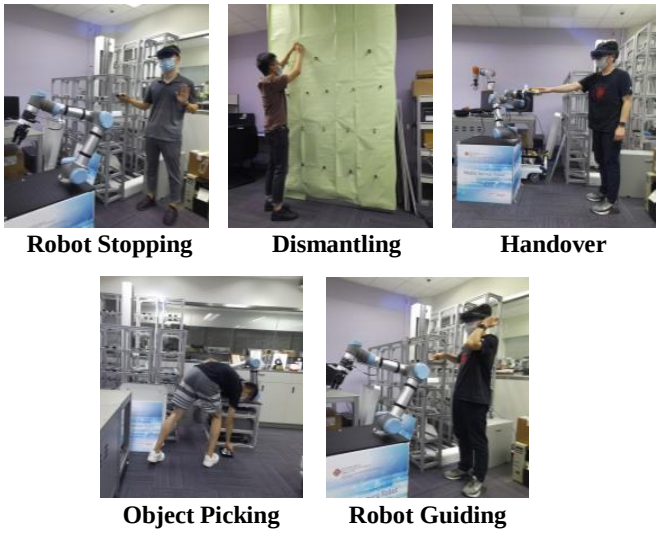


Fig. 2.3: HRC assembly activity example.

1) Recognition

The task of human activity recognition has attracted much attention in the HRC area. Concretely, the formulation of the task is that robots are supposed to understand the engaging activities of a person given past and present observations by camera or other sensors.

Some works only leverage RGB cameras as the data source. Wang et al. [4] reported a collaborative engine assembly case, during which a deep learning-based method was proposed to continuously analyze human operators' working actions such as grasping a screwdriver or plugging in a small part. Xiong et al. [66] also aimed at human activity recognition in the HRC engine assembly task and introduced a two-stream CNN model that consists of a spatial stream designed to extract spatial features from a single RGB frame and a temporal stream that takes optical flow maps as input to exploit

Table 2.5: Literature of human activity.

Category	Application	Key Elements	Source
Recognition	Collaborative assembly	AlexNet	[4]
	Collaborative assembly	Two-stream CNN	[66]
	Collaborative assembly	3D LRCN; 3D CNN; LSTM	[67]
	Companion robot	STJ-CNN; Skeleton	[68]
	Hand-object action	Hand skeleton; TCN	[69]
	Collaborative packaging	LSTM; VAE; DRL	[70]
	Safe HRC	3D CNN; 1D CNN; RGBD+Tactile	[71]
	Ergonomics in HRC	IMU; EMG; Random Forest	[72]
	Sustainable HRC	EMG; CNN; Fatigue	[73]
	Seamless HRC	Mixture-of-experts; Graphical attention	[74]
	Real-time action	DC-CNN; IMU	[75]
	Hands-free HRI	Eye gazing; head action; MR glasses	[76]
Prediction	Collaborative disassembly	CNN; LSTM	[77]
	Collaborative maintenance	Mask R-CNN; 3D CNN; LSTM	[78]
	Surveillance	CNN; Optical flow	[79]
	Collaborative assembly	CNN; VMM	[80]
	Collaborative manipulation	Gaussian mixture model	[81]
	Reactive robotic response	ATCRF; Object affordance	[82]
	Collaborative assembly	HMM	[83]
	Human following	Reinforcement learning	[84]
	Online HRI	CVAE; Kinect	[85]
	Human imitation	OpenPose; Motion GAN	[86]
	Human-robot handover	LSTM-RNN	[87]
	Collaborative manipulation	RNN; Object affordance	[88]
	Collaborative assembly	Kinect; RNN	[89]
	Human imitation	VAE; Motion embedding	[90]
	Social HRI	CVAE; LSTM	[91]
	Social HRC	EM-ART; Deep ART	[92]
	Ergonomic risk prediction	VGG16; TCN	[93]
	Attention estimation	Eye-tracking; ANN	[94]
	Attention estimation	Skeleton data; LSTM	[95]

temporal information. Meanwhile, Wen et al. [67] reported a method to recognize sub-assembly activities of the visual controller assembly process such as mainboard assembly and camera assembly, and a 3D LRCN (Long-term Recurrent Convolutional Network) model composed of 3D convolutions and LSTM (Long Short-Term Memory) module was proposed to achieve this target. Human skeleton joints have also been explored to facilitate human daily activity recognition for a companion robot in [68], which proposed a Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) model to parse human body parts and skeleton joints features. Sabater et al. [69] focused on the recognition of hand-object interactive activities based on hand skeleton extraction and TCN (Temporal Convolutional Neural Networks). Ghadirzadeh et al. [70] made an attempt to implement human-robot collaborative packaging in an unsupervised manner leveraging Deep Reinforcement Learning (DRL) incorporated with VAE (Variational Autoencoder) and LSTM modules for implicit human action representation. RGB data may have the advantage of convenience and availability, but multisensory data can provide more latent information that might also be vital for human action recognition.

Amin et al. [71] reported a study that implements safety monitoring in HRC via a mixed-perception method that inputs skeleton images into a 3D CNN model to obtain human activity results and leverages tactile sensor data for contact detection. In the work by Yoshikawa et al. [72], inertial measurement unit (IMU) and electromyography (EMG) were utilized to assist robots in evaluating ergonomic workload by recognizing human activities using rule-based method and Random Forest classifier, while [73] also adopted EMG to recognize human fatigue in a sustainable HRC workspace via CNN model. Islam et al. [74] work towards seamless HRC via a Multi-GAT (Multimodal

Graph Attention) model for human activity recognition that combines RGB, depth, skeleton, and physical sensor data via a graphical attention mechanism to capture multimodal correlation features. Real-time action cognition for potential human-robot application was considered by Alemayoh et al. [75], in which IMU data collected by smartphone is processed by the proposed DC-CNN (Double-channel CNN) model to obtain the action category. Park et al. [76] studied hands-free HRI via a multimodal interaction method that captures eye gazing and head orientation data by MR (Mixed Reality) glasses. Multisensory data-based approaches can achieve promising results, but the main problem is that these sensors normally have to be worn by humans, which may cause discomfort in long-term HRC working.

2) Prediction

Human activity recognition can suffice for reactive HRC applications, but Proactive HRC poses a greater challenge for timely response, for which future activity or motion prediction should be able to provide a feasible approach.

Some researchers resort to predicting the intended actions of human operators. Liu et al. [77] focused on intention prediction during a desktop disassembly process via a motion recognition and prediction network consisting of convolutional layers and LSTM layers which could predict the label of the intended action of the operator. Alati et al. [78] attempted to enable the robot assistant to infer human needs in a collaborative warehouse maintenance task through a human intention prediction method which also leverages 3D convolution and LSTM. Bibi et al. [79] explored the integration of Transformed Optical Flow Components (TOFCs) into CNN architecture to anticipate ongoing human interactions, while Zhang et al. [80] made an

attempt to predict human assembling actions during HRC motor assembly leveraging variable-length Markov modelling (VMM) and CNN models. The limitation of these intention prediction works is they can only provide future action labels, which is insufficient for robots to achieve autonomous collision avoidance and more subtle robot planning.

Motion trajectory prediction can empower robots to forecast human actions in a more delicate manner. Some earlier works [81, 82, 83] exploit probabilistic models to predict future human motion for collaborative assembly or robotic manipulation tasks, where GMM (Gaussian Mixture Model), CRF (Conditional Random Field), and HMM (Hidden Markov Model) are adopted or enhanced in these works to fulfil the task. Recently, more attention has been paid to skeleton-level trajectory prediction in HRC-related applications mainly through two types of methods: recurrent neural networks (RNN, LSTM, etc.) and deep generative models (VAE, GAN, etc.). RNN-based models are mainly adopted in proximal HRC working scenarios such as human-robot handover [87], collaborative object manipulation [88], and assembly [89], where the focus is to avoid colliding with human body by foreseeing human skeletal trajectory. Another line of works considered more about the multiple possibilities of human motion prediction for human imitation or HRI, relying on generative models such as GAN (Generative Adversarial Network) [86] and VAE [85, 90, 91] to generate multimodal trajectories of future human intention for proactive robot planning. Bayoumi et al. [84] considered a more concrete human-following situation for mobile robots, during which humans might be partially occluded by scene objects but the robot still needs to predict the human walking path through explicitly encoding occlusions into the reward function of reinforcement learning. Although the impressive performance of trajectory forecasting has been demonstrated in the literature,

a fundamental problem rarely mentioned is how to evaluate the quality of predicted trajectory in ongoing HRC missions, and prevent disastrous robotic actions promptly when the prediction is unreliable.

Prediction of other latent representations has also been explored in previous works. Lee et al. [92] proposed an episodic memory mechanism based on EM-ART and Deep ART models and applied it in learning the relations between human actions and emotional states to predict future human emotions in a social HRC case. Parsa et al. [93] aimed at the prediction of ergonomic risk of indoor object manipulation activities of human operators, which was implemented with VGG16 network as spatial feature extractor and encoder-decoder TCN (Temporal Convolutional Network) model for temporal information aggregation. Human attention estimation for human-robot interaction has also been intensively studied in previous works [94, 95], where eye-tracking or human skeleton data are collected as the main information source and Neural Networks such as ANN and LSTM are leveraged as the main models.

2.2.3 Human Pose Recognition

A vast body of works has explored the human activity recognition task, but less attention has been paid to human pose recognition, which inclines to explore the detailed posture of the human body on a finer granularity. There is ample literature demonstrating the utilization of human pose recognition in HRC scenarios, which are mainly divided into two categories, i.e., full body and hand. Despite being a part of the human body, hands serve different purposes from full body in most existing research works as they are the

most expressive organs. Therefore, we decided to discuss body and hand separately. The general categorization can be found in Table 2.6.

1) Body Pose

Human body pose is normally formulated as skeleton or joint maps inferred from sensor data to support fine-grained robot planning in HRC. Kinect cameras are often utilized to capture the human body and generate skeleton maps. The Kinect-shipped vision algorithms enabled some previous works in HRC disassembly or teleoperation applications [96, 97]. CNN models are also widely adopted in static body pose estimation. Liu et al. [98] leveraged PoseNet to estimate body joint locations to achieve collision-free HRC assembly. Van et al. [99] focused on the ergonomic adaption problem in HRC while utilizing OpenPose as the body pose estimator and joint angle-based rules for further ergonomic analysis. Another work [100] chose pressure sensors to recognize the standing postures of workers in an HRC manufacturing workspace via fusing CNN, KNN, and SVM classifiers based on Dempster–Shafer evidence.

It can be easily identified that previous works in the HRC domain mainly consider 2D human pose modelling in the form of sparse joint or skeleton maps, while the human awareness of robots could be significantly enhanced with a more complete 3D dense modelling of the human body.

2) Hand Gesture

Table 2.6: Literature of human pose.

Category	Application	Key Elements	Source
Body	Teleoperation	Kinect	[97]
	HRC disassembly	Kinect	[96]
	Human standing posture	CNN; KNN; SVM; Dempster-Shafer	[100]
	Ergonomics in HRC	OpenPose; Angle-based rules	[99]
	Safe HRC	OctMap; PoseNet	[98]
Hand	Teleoperation	HOG	[101]
	HRC surgery	HMM	[102]
	Robotic hand imitation	ANN	[103]
	Gesture-based control	Hu moment; Random Forest	[104]
	Gesture-based control	SoCJ feature; SVM	[105]
	Gesture-based control	HOG	[106]
	Gesture-based control	EEG; EMG	[107]
	Multimodal control	CNN; LSTM; Multi-modal	[108]
	Dual-hand gesture	RI-SSD; VGG19	[109]
	Gesture-based control	Faster R-CNN	[110]
	Safe HRC	Inception v3; OpenPose; Kinect	[111]
	Gesture-based control	3D SSD	[112]
	Space HRI	FF-SSD	[113]
	Dynamic gesture cognition	CNN; LSTM; Optical flow	[114]
	Dynamic gesture cognition	CNN; LSTM; Attention	[115]
	Cross-domain gesture cognition	Leap Motion; SVM; CNN	[116]
	Cross-subject gesture cognition	EMG; CNN; LSTM	[117]
	Gesture-based programming	R-FCN	[118]
	Service robot interaction	RGB-D; Inertial sensor; LSTM	[119]
	Surgical robot teleoperation	Leap Motion; LSTM-RNN	[120]

Hand gesture recognition is a prevalent topic in human-robot interaction and human-robot collaboration because it is intuitive, effective, and expressive serving as a robot-controlling interface.

Earlier works tended to employ hand-crafted feature-based solutions to recognize hand gestures. [101, 106] both relied on the Histogram of Oriented Gradient (HOG) to serve as the feature descriptor to facilitate subsequent hand gesture classification or tracking for further human-robot teleoperation or gesture-based robot control. Chen et al. [104] utilized Hu moment feature descriptor and random forest classifier to differentiate hand gestures as a remote robot controlling solution. Hendrix et al. [105] proposed a solution to recognize hand gestures and verified the control of a robotic manufacturing assistant in a limited-access scenario with the combination of the Shape of Connected Joints (SoCJ) feature and SVM classifier. HMM was adopted in [102] to facilitate instrument delivery for a surgical assistance robot, while ANN was leveraged in [103] to recognize hand gestures for robotic hand imitation. Unlike vision-based works, [107] resorted to EMG and Electroencephalography (EEG) for hand gesture recognition and robot control. Hand-crafted feature-based methods generally suffer from poor robustness as mentioned in earlier sections, therefore recent works present an inclination to shift towards deep learning solutions.

A large amount of deep learning-based works for hand gesture recognition have emerged in recent years. Liu et al. [108] proposed a solution towards HRC manufacturing that leverages multimodal fusion of speech, body motion and hand gesture based on CNN and LSTM models, which are also regarded as the core models for hand gesture recognition in [114, 115]. Gao et al. [109] reported an application in an astronaut-robot interaction system, where

dual hands detection and gesture recognition are implemented based on ResNet-Inception-Single Shot MultiBox Detector (RI-SSD), while other SSD variants were also adopted for hand gesture-based human-robot interaction such as 3D SSD [112] and FF-SSD (feature-map-fused single shot multibox detector) [113]. Region-based object detectors have also been applied in hand gesture recognition for human-robot collaborative controlling by Nuzzi and her group in [110, 118]. Leap Motion Controller is another prevalent choice in HRI application [116] and teleoperated surgery [120] since it has an integrated visual hand gesture recognition system and further action analysis could be achieved with extra deep learning models on top of the extracted gesture data. Other sensors such as EMG [117] and inertial sensor [119] were also leveraged to facilitate hand gesture recognition along with CNN, and LSTM models. Despite the better performance and robustness of recent works, a tough problem that remains is the self-occlusion of the human hand where some fingers might be occluded by the hand itself, making it hard to capture the information via visual sensors.

2.3 Environment Parsing for HRC

With object-level and human-level information obtained, robots could already perform collaborative actions in some relatively simple tasks such as tool or workpiece delivery in a fixed workstation. Nevertheless, to deal with more complex tasks such as navigating to places out of sight to fetch a specific object required in an HRC assembly process, robots should be equipped with the skill to perceive and model the whole working environment more comprehensively. In this section, existing works related to environment

parsing are summarized into three categories based on the utilized mapping representations, which are illustrated in Table 2.7.

Table 2.7: Literature of environment parsing.

Category	Application	Key Elements	Source
Scene graph	General robotics	DAG; RSG	[121]
	Scene description	GCN; RNN	[122]
	Safe HRC	Mask R-CNN; Fuzzy logic; Safety	[123]
	Safe HRC	Mask R-CNN; MSDN	[124]
2D map	HRI	GVG; Confidence tree	[125]
	Robot navigation	CNN; U-Net	[126]
	Robot navigation	SLAM; LSTM; Mask R-CNN	[127]
	HRI	3D CNN; Robot team	[128]
3D approach	Disassembly	CAD model	[129]
	Safe HRC	Kinect; Point clouds	[130]
	Teleoperation	VR; Kinect; PointNet	[131]
	Safe HRC	OctMap; PoseNet	[98]
	Safe HRC	OctMap; MDP; RL	[132]
	Safe HRC	Point clouds; RTLS	[133]
	Safe HRC	MR; Digital twin; Point clouds	[134]

2.3.1 Scene Graph

Among these representations, scene graph may be the most abstract one, which transforms the perception results of the environment into a topological graph structure. Blumenthal et al. [121] proposed a method called Robot Scene Graph (RSG), which leverages a Directed Acyclic Graph (DAG) to represent and manage 3D geometric entities for general robotic applications. Moon et al. [122] studied the generation of natural language description from environment images for further human-robot communication leveraging graph convolution networks (GCN) to extract local features from a 3D semantic graph map and LSTM to generate scene description. Hata et al. [123] reported a more specific application of scene graph for safe HRC in

a warehouse navigation case, in which Mask R-CNN is utilized to segment scene objects from images and subsequently encode the extracted object information into a scene graph for further fuzzy logic-based risk management, while Riaz et al. [124] considered a similar warehouse scenario for HRC safety analysis leveraging the proposed MSDN (Multi-level Scene Description Neural Networks) to generate scene graphs and region captions. Being a compact and efficient representation of the environment, scene graph is widely adopted in robotic applications, but the graph-based structure also undermines the ability to capture geometric relations between objects.

2.3.2 2D Map

To be able to represent detailed geometric relations of scene elements, 2D map is a natural choice following human practice, which normally takes the form of an overhead view. Liao et al. [125] employed occupancy grid mapping to generate a local map from laser range data for robot navigation based on generalized Voronoi graph (GVG) data representation and a confidence tree was introduced to fuse the classification results from different granularity layers to obtain the final place classification result. Hiller et al. [126] explored the residential environment modelling for autonomous robots leveraging existing occupancy grids as input to CNN classifier for patch-level door localization and U-Net for pixel-level door segmentation. In a work towards robot navigation under human instructions from Hu et al. [127], the semantic map generated via SLAM (simultaneous localization and mapping) technique was leveraged to represent the global map of the environment, Mask R-CNN was employed to detect scene objects, and LSTM was utilized to parse human instructions and provide constraints to the grounding process based on the map and scene elements. Dias et al. [128] leveraged occupancy

grids to represent the positions of a robot team and to serve as an interactive interface, through which a 3D CNN model was applied to learn from human demonstrations about robot controlling sequences. 2D map techniques are suitable for planar navigation in relatively simple environments, but the lack of height information hinders its application in more complex scenarios such as aeroplane cabins.

2.3.3 3D Approach

In some applications such as HRC assembly, delicate 3D information is required to represent the environment on a finer scale so that the robots could carry out more sophisticated operations without colliding with scene objects. Some works directly utilized the point cloud generated via RGB-D cameras to represent the environment [130, 131], while others adopted the representation of a voxel map, which could be regarded as a quantified 3D grid mapping of the original point cloud. Abou Moughlbay et al. [130] proposed a monitoring system for HRC production environment consisting of four Kinect RGB-D cameras, which are utilized to generate point clouds of the workspace, and then downsample the point clouds to voxel grids after filtering and trimming. Friedrich et al. [129] employed voxel maps to represent the recognized scene objects in an autonomous robot-space exploration task, during which the initial modelling is constructed via CAD models and later updated through vision data. A similar technique, OctMap was utilized in [98] to represent the 3D occupancy status of an HRC working space so that the robot could actively avoid collision with human operators and other objects. Liu et al. [132] aimed at collision-free robot planning for HRC manufacturing tasks, during which OctMap is also leveraged for workspace monitoring, while MDP (Markov Decision Process) and RL (Reinforcement Learning) techniques are

adopted for collision avoidance. Slovak et al. [133] aimed to develop a safe HRC shared workspace by employing point cloud and RTLS (Real-time Location System) technology to reconstruct the 3D environment. Choi et al. [134] proposed a safety measurement method for HRC system, utilizing 3D point cloud representation for the physical environment and synchronizing with a digital twin model in real-time for further distance measurement in the virtual space. 3D approaches contain the most abundant environmental information which could support finer-grained HRC action planning and execution, but generally requires more storage and computational resource, and might cost much more time to search for appropriate robot actions in large areas, which makes it less flexible to fit for different industrial practice.

2.4 Visual Reasoning for HRC

Table 2.8: Literature of visual reasoning.

Category	Application	Key Elements	Source
Visual cue	Collaborative assembly	Bayesian decision-making	[135]
	Collaborative assembly	ConvVAE; LSTM	[136]
	HRI	Siamese network; Spatial attention	[137]
	Collaborative assembly	Dual-input CNN	[138]
Visual & Language	HRI	Case-based reasoning	[139]
	HRI cooking support	CNN; HHMM	[140]
	HRC fault diagnosis	MDP	[141]
	Human-guided pickup	Hourglass network; RNN	[142]
	Human-guided pickup	Bi-LSTM; U-Net; Multi-head attention	[143]
	Explainable HRC	CNN; RNN; Logical reasoning	[144]
	Collaborative assembly	TC-VQA; CNN; Symbolic reasoning	[145]
	HRI	Multi-view VQA	[146]

The perception of objects, humans and the environment could provide a holistic understanding of an HRC working scene. To bridge the gap between scene understanding and proactive decision-making, a reasoning mechanism is necessary for robots when collaborating with human operators. In this section, we primarily focus on visual reasoning, which refers to reasoning about the latent meaning of visual cues or indications for future robot actions from visual observations of an HRC scene. Related works about visual reasoning are listed in Table 2.8, among which it is found that except for vision-only solutions, some works additionally introduced language information to compensate for the ambiguity caused by sole visual cues.

2.4.1 Visual Cue

Reasoning based on visual cues is a fundamental requirement of higher-level cognitive intelligence for collaborative robots, where some initial explorations have been conducted in previous works. Rahman et al. [135] reported an HRC scheme that could automatically reason about which sensing mode (human or robot) of assembly parts detection to take based on the confidence and cost of observations and regret-based Bayesian decision-making method. Murata et al. [136] proposed a method for HRC assembly that relies on ConvVAE and LSTM models to reason from the goal image and visual observation to determine which part should be delivered to the human for assembling operations.

Some works attempted to incorporate more human guidance during the visual reasoning process. Venkatesh et al. [137] tried to teach robots to localize novel objects by adding the human pointing cues to the object image, and leveraged the Siamese network and spatial attention mechanism to

accomplish the localization task. Sun et al. [138] proposed a dual-input CNN model which takes as input the assembly part image and workspace context image simultaneously to facilitate robot learning through human demonstration.

Visual information may have a certain level of ambiguity. For instance, when a human reaches out his hand towards robot collaborators to ask for something during assembly, the wanted object could either be a workpiece or a tool judging from mere visual observations. Thus, it is not uncommon to see that natural language information is included in the visual reasoning task.

2.4.2 Visual and Language Cue

It is natural to introduce human language as an additional reasoning cue for it is more accurate and compact. Earlier attempts mainly relied on mathematical models or knowledge-based models to implement the reasoning process with visual and language cues as supplementary information. Roncancio et al. [139] integrated object localization, human activity recognition, and speech recognition into a case-based reasoning system of a service robot, which mainly models prior knowledge via episodic memory mechanism. Hayes et al. [141] aimed at an HRC fault diagnosis task leveraging the Markov Decision Process as the policy model for a robot to generate the action policy based on visual observations and human queries and further generate a policy explanation in human language for better interpretability.

Recent works show an inclination to put more effort into data-driven deep learning models for end-to-end visual reasoning. Ahn et al. [142] studied the human-guided pickup task and proposed a Text2Pickup network, which

consists of an Hourglass network and an RNN, to locate the desired object for robots to pick up based on human language commands and workspace image observations, and additionally generate interactive questions for human to clarify when the initial command is vague. Venkatesh et al. [143] followed a similar task that requires the robot to reason about picking coordinates of objects from language and image input, but employed a different approach that leverages Bi-LSTM and multi-head attention to extract language features, which is subsequently combined with image features to be input to a U-Net model to generate the object coordinates.

Riley et al. [144] reported an attempt to tackle the task of explanatory Visual Question Answering (VQA) for HRC via the integration of CNN, RNN, logical reasoning, and inductive learning. Tan et al. [145] also considered the task of VQA, but proposed a new VQA task and dataset for task-oriented collaborative question answering (TC-VQA) for HRC gearbox assembly, and provided a baseline method that leverages deep learning-based object and hand detection, gesture recognition along with symbolic reasoning to generate answers. Qiu et al. [146] explored the VQA problem in a multiview setting for human-robot interaction, where the robot needs to autonomously choose a better viewpoint to obtain the necessary information to answer the questions correctly. The exploration of visual reasoning in HRC scenarios is still in its infancy, and currently the task setting and solution are rather naive and still far away from practical deployment.

2.5 Research Gaps

The above review of existing literature presents a brief glimpse of the current application status of vision techniques in HRC environments. Some limitations and research gaps have also been revealed through the review process and are summarized here.

2.5.1 Precise Object Modeling for Co-Manipulation

Despite the wide adoption of computer vision technologies such as object detection [27], and object classification [16] in robotic and industrial applications, there is a lack of discussion about precise object modelling under the topic of human-robot collaboration. In existing HRC assembly works, robots still mainly serve as assistants to human operators and leave the subtle assembly process to humans, partially because the uncertainty introduced by human partners prevents robots from obtaining the precise geometric pose information of the assembly parts. In such cases, real-time precise 6-DoF object pose estimation techniques can be particularly useful. Although there have already been some discussions about 6-DoF pose estimation for industrial parts [47], several limitations such as the dependency on object CAD models, weakness to occlusion, and computational inefficiency, severely hindered its application in HRC scenarios.

A major challenge emerges when dealing with occlusion, which is pervasive during HRC manufacturing, especially when human or robot agents are

handling objects. Hand-crafted feature-based 6-DoF pose estimation methods are generally fragile when facing severe occlusions, while stronger performing CNN models could achieve better results under occlusion, but still suffer from it. A possible alleviation to this problem is an explicit modelling of the occlusion area with extra constraints and priors potentially provided by the occluding human hands, which will be elaborated in the following content.

2.5.2 Finer-Scale Human Worker Body Reconstruction

Human perception-related works contributed quite a large portion to visual understanding in HRC, due to the highest priority of human safety. During the review process, it was found that traditional image processing, multi-sensory devices, and deep learning models have been substantially explored in HRC scenarios. Nevertheless, it still has a long way ahead to massive application because existing methods could only partially perceive the human body through wearable devices or only obtain the rough position via visual detection or skeleton recognition instead of fine-scale 3D geometric modelling. On the other hand, there is a recent trend towards dense human pose modelling, including dense body pose [147, 148] and dense hand pose [149, 150], in the computer vision community, which might be adopted for finer-scale human worker perception in Proactive HRC cases. Besides 3D human pose estimation, another remaining challenge is how to provide a more comprehensive human worker model that can capture multiple facets of human status in real-time to facilitate on-site robotic decision-making and motion planning. Regarding this research gap, a potential solution is

to construct a human digital twin of the human operators during the HRC working process which will be discussed in later chapters.

2.5.3 Hierarchical and Hybrid Workspace Modeling

Based on the review result of visual environment parsing in HRC applications, it is found that most researchers followed the routine of first recognizing the environment via vision algorithms and then representing scene elements with certain mapping techniques. The summarized scene representations (scene graph, 2D map, 3D representation) each have their own particular strength and weaknesses as mentioned in previous sections. However, none of them alone could suffice for comprehensive workspace modelling in future HRC systems such as extremely flexible manufacturing shop floors, where mobile robots need to be able to execute fine-grained collaborative production actions which require a delicate scene representation, as well as coarse-grained medium-to-long navigation tasks that demand responsive real-time route recommendation. A hierarchical and hybrid environment representation would be preferred in the above-mentioned HRC situation. The representation should possess multiple abstraction layers with different semantic levels to accommodate varied-grained applications.

2.5.4 Advanced Vision-Language Collaborative Reasoning

The capability of performing complex human-like reasoning is always the pursuit of artificial intelligence and robotics, which also stands true in HRC manufacturing to achieve truly reliable and seamless collaboration between humans and robots. Ample research works related to vision or vision-language-based reasoning have been conducted in HRC scenarios, leveraging various techniques ranging from mathematical models to deep learning models, but several shortcomings ought to be pointed out that pervasively exist in those works. One is that current research works mainly formulate the reasoning task as a naive mapping from visual or language cues to certain decisions or actions, without much consideration about the incorporation of prior knowledge and vision-language observations. Another issue is the limited consideration of the unique characteristics of applying reasoning techniques in HRC scenarios, which is the close and frequent interactions between humans and robots. Including human operators in the reasoning loop of the HRC system can better leverage human creativity and intelligence to achieve a more efficient collaborative reasoning scheme. However, introducing human intervention in every step of HRC process is highly impractical. The ability to autonomously determine when to ask for human assistance requires closer investigations.

3 Industrial Workpiece Pose Estimation

In human-robot collaboration scenarios, industrial objects such as workpieces, tools, and other components are pervasively present. For instance, during the assembly process, it is crucial for the robot to be aware of the ongoing assembly area, identify the remaining missing parts, and recognize the required tools. This enables the robot to proactively make decisions about its subsequent collaboration actions. In this chapter, we mainly focus on the 6-DoF estimation of industrial workpieces since it is the foundation for subsequent robotic manipulation in the HRC process. Another major challenge, i.e., the mutual occlusion between the human hand and industrial workpieces, will also be discussed in the second part of this chapter.

The research work presented in this chapter is based on a conference paper presented at the 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE) [151], and a journal paper published in the IEEE Transactions on Automation Science and Engineering [152].

3.1 Introduction

Object perception, especially object pose estimation, is the most essential skill that a collaborative robot needs to possess in order to build a basic

understanding of the HRC scene objects, which are omnipresent in the HRC environment such as the assembly parts, working tools, et cetera. Traditional object pose estimation approaches mainly focus on matching corresponding key point pairs between observed 2D images and 3D object models via hand-crafted feature descriptors. However, key points are hard to discover from images when the parts are piled up in disorder or occluded by other distractors, e.g., human hands. Although the emerging deep learning-based methods are capable of inferring the poses of occluded parts, the accuracy is not satisfactory largely due to the loss of spatial resolution from multiple downsampling operations inside convolutional neural networks. To overcome this challenge, the first part of this chapter proposes a 6-DoF pose estimation model consisting of a pose estimator and a pose refiner, by leveraging the High-Resolution Network as the backbone network. Experiments are further conducted on a dataset of industrial parts to demonstrate its effectiveness.

To delve deeper into the occlusion issue of object pose estimation, especially the occlusion caused by human hands as is a frequent case in HRC scenarios, the pose estimation of human hands is further taken into account in the second part of this chapter. Explicit human-object perceptions are significant but remain little reported in the literature for adaptive robot decision-making, especially in the close proximity co-work with partial occlusions. Aiming to bridge this gap, this study proposes a vision-based 3D dense hand-object pose estimation approach for HRC cases. First, a mask-guided attentive module is proposed to better attend to hand and object areas, respectively. Meanwhile, explicit consideration of the occluded area in the input image is introduced to mitigate the performance degradation caused by visual occlusion, which is inevitable during HRC hand-object interactions. In addition, a 3D hand-object pose dataset is collected for a lithium-ion battery disassembly scenario in the

lab environment with comparative experiments carried out to demonstrate the effectiveness of the proposed method.

3.1.1 Object Pose Estimation

Pose estimation of industrial parts has a significant value in smart manufacturing and human-robot collaboration. For instance, robots need to be able to consistently recognize objects of interest to cope with uncertainties introduced by human collaborators or flexible production. 2D information-based methods approach this problem by image object detection or instance segmentation [153], which can only provide 2D location and shape for 2D tasks such as planar grasping. For more complex tasks such as human-robot handover, 3D information is indispensable since the objects can be presented at any location with any orientation in the workspace.

The problem of 3D location and shape estimation is normally formulated as 6-DoF pose estimation, which refers to the estimation of rotation and translation parameters between the observed image and 3D object model. Traditionally, hand-crafted feature-based methods were proposed to tackle this problem by matching key point pairs between image and 3D object model [154, 155], but these methods suffer from occlusion and textureless objects, in which cases it is difficult to find sufficient key points. Another line of work regard pose estimation as a template matching problem [156, 157], which could mitigate the problem of textureless objects, but occlusion still remains unsolved. Convolutional neural network is recently introduced into this field in recent studies [158, 159, 160], since it shows dominant performance in other computer vision tasks. Although CNN-based methods are capable of predicting object pose under occlusion, the overall accuracy is

not satisfactory. One possible explanation is that those methods borrow the backbone network design from classification models such as VGGNet [161] and GoogLeNet [162], which gradually reduces the feature map resolution by pooling operations, making it hard for the model to capture subtle differences of object poses.

Aiming at exploiting the advantage of high-resolution features, in this study, High-Resolution Network (HRNet) [163] is leveraged as the backbone network, upon which a model is constructed that directly predicts 6-DoF pose parameters from RGB-D data. The key idea behind HRNet is to keep a high-resolution branch as well as several gradually lower-resolution branches in parallel and fuse the features from different branches multiple times at certain positions of the network. This is from the commonly acknowledged insight that low-resolution feature maps represent semantic information better while high-resolution feature maps contain more precise spatial information. As for 6D pose estimation, this study suggests that it not only requires strong semantic representations to recognize the object appearance but also needs precise spatial features to distinguish small variations of object pose. In addition, an extra refining network is adopted to refine the predicted coarse 6-DoF pose. The strategy of pose refinement is widely adopted to improve the accuracy of pose estimation in previous work such as [154, 157, 159, 164]. While many of them utilize iterative optimization methods such as ICP algorithm [154], this study instead aims to predict the distance between coarse estimation and ground-truth pose by a CNN model, which takes the same network design as the coarse pose estimation network.

3.1.2 Hand-Object Pose Estimation

Human interventions and interactions are inevitable in HRC tasks because of the essential role of human operators in the HRC team, which makes it natural to additionally take the human hand into account during object perception. Ample research works have been devoted to the recognition of either human bodies or objects. For instance, human body information [165, 96] and hand gesture cues [166, 108] have been actively leveraged to recognize the human motion and intention via deep learning models in HRC applications. On the other hand, the identification and localization information of industrial objects have also been exploited via computer vision techniques in previous HRC studies [167, 168, 169] to facilitate adaptive robot manipulation and collaboration.

However, insufficient attention has been paid to jointly recognize the human hand and object, and to reconstruct their 3D dense geometries, which is believed to be crucial for robots to carry out interactive actions with human operators in close proximity. For instance, in a human-robot handover case, the ability to simultaneously exploit the 3D spatial relation of hand and object is preferred for the robot to plan the handover position and timing proactively. Meanwhile, this close-range human-robot co-manipulation in HRC can bring up another issue, which is the partial occlusion between the hand and object. For instance, the disassembly tool/part being handed over from human to robot is very likely to be partially occluded by the human hand from the robot view, which could cause the robot to misjudge the grasping position and fail to undertake the action accurately and safely.

Regarding the image occlusion issue, some recent efforts in the computer vision community have been made by leveraging generative models to synthesize the content of occluded areas in an input image [170]. These generative methods can produce visually plausible non-occluded images, but they are not suitable for subsequent recognition tasks since the generated images contain many artefacts that do not exist in natural images. Another line of works explored the pixel-level discrimination of occluder and occludee [171] to facilitate 2D instance segmentation. Although the core idea is ingenious, these works focus more on general daily objects and only 2D information, which renders it infeasible for direct application in 3D hand-object pose estimation.

Motivated by the aforementioned problems, this study additionally focuses on the simultaneous 3D dense reconstruction of hand-object poses from partially occluded observations in HRC activities. An integrated model for hand-object pose estimation is proposed with binary mask guidance for better hand and object attention separation, as well as an explicit occlusion-aware mechanism designed to minimize the reconstruction error caused by hand-object occlusion.

3.2 High-Resolution 6-DoF Pose Estimation of Industrial Parts

In this section, the proposed high-resolution network-based 6-DoF pose estimation method is explained in detail. Provided with the observed RGB-D image and 3D object model, the objective of 6-DoF pose estimation is to

infer the object pose parameters, which are normally presented as a SE(3) transformation (SE: Special Euclidean Group) consisting of a 3-DoF rotation $\mathbf{R}$ and a 3-DoF translation $\mathbf{t}$. With the estimated $\mathbf{R}$, $\mathbf{t}$ and the object's 3D model, the complete 3D information of the object can be well obtained. The overall architecture of the proposed model for 6-DoF pose estimation of industrial parts is illustrated in Fig. 3.1. The architecture mainly consists of three stages, i.e., industrial parts detection, coarse pose estimation, and pose refinement. The detection stage takes RGB images as input, and output parts bounding boxes and classification results. The detection stage can be regarded as a preprocessing step and the following pose estimation stages are actually agnostic to which specific detection model is used, so this study simply adopts Faster R-CNN [172] as the detector. Then in the coarse pose estimation stage, an HRNet-based pose estimation network is constructed to better distinguish small pose differences of industrial parts by taking advantage of high-resolution features. The coarse pose estimation network takes the cropped RGB-D patch of the detected part as input and estimates the rotation $\mathbf{R}$ and translation $\mathbf{t}$ respectively. The final stage is designed for pose refinement, which first generates a rendered RGB-D image based on the estimated coarse pose parameters and the object 3D model, then concatenates the rendered image and the cropped image together as the input, and estimates the pose deviations $\Delta\mathbf{R}$ and $\Delta\mathbf{t}$. By applying the predicted $\Delta\mathbf{R}$ and $\Delta\mathbf{t}$ to the coarse $\mathbf{R}$ and $\mathbf{t}$, the final 6-DoF pose estimation is obtained.

3.2.1 Industrial Part Detection

The first step of this work is to extract the regions of interest for industrial parts. Concretely, the observed image might contain multiple industrial parts,

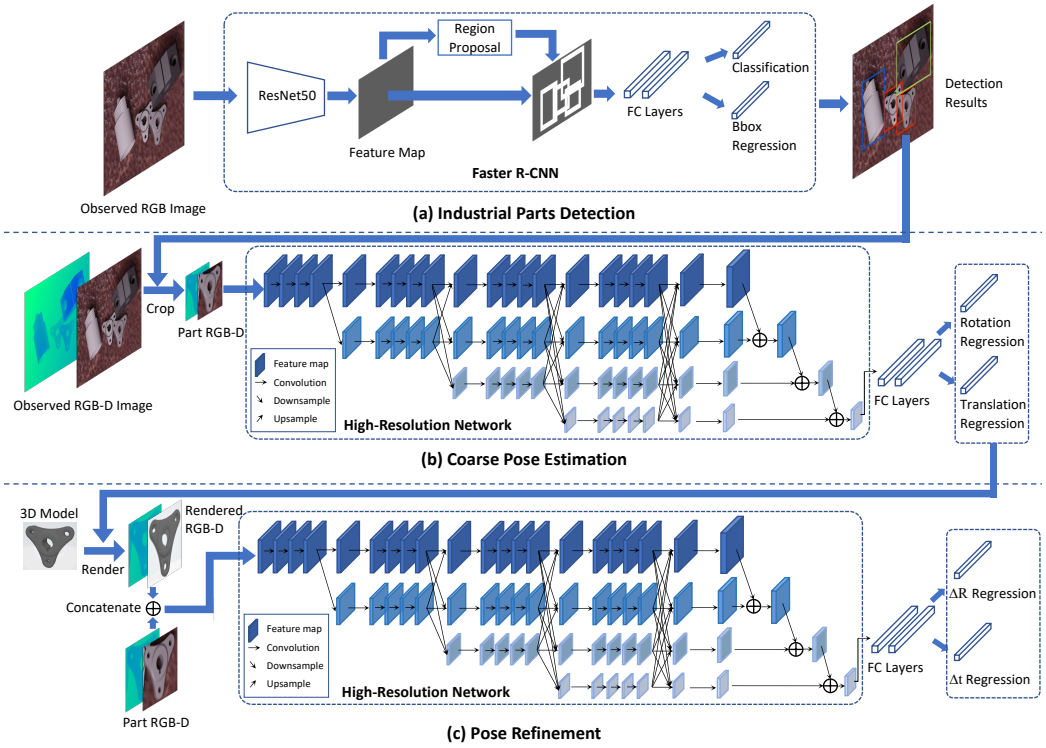


Fig. 3.1: Overall architecture of the high-resolution 6-DoF pose estimation model.

for which the individual regions and categories need to be extracted first to facilitate subsequent 6-DoF pose estimation. For a specific part in the image, the image patch will be cropped according to the detection results which are normally represented as bounding boxes. Then the following pose estimation processes only need to consider the cropped image area, which brings two benefits: 1) The removal of irrelevant image area could ease the model training process; 2) Better computational efficiency. Meanwhile, the part category given by the classification results will be used to decide which part model should be applied to the rendering process in the pose refinement stage. Therefore, a successful detector Faster R-CNN is employed to locate industrial parts in complex industrial scenarios, regardless of occlusion and textureless interference. Under this prerequisite, the work can pay more attention to the following two-stage pose estimation.

3.2.2 Coarse Pose Estimation

Instead of segmenting each object with extra branches [158], this study simply crops the part region from the observed image as input of the pose estimation model. Based on the cropped RGB-D image, the pose estimation model can directly regress the part pose parameters including the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ of the part to fully recover its pose in 3D space. The details are illustrated as follows. d image as input of the pose estimation model. Based on the cropped RGB-D image, a pose estimation model will directly regress the part pose parameters. This section illustrates the details of the pose estimation model as follows.

1) Depth Image

The introduction of depth image plays an essential role in the pose estimation model. Existing methods such as [173, 158] tend to tackle the 6-DoF pose estimation problem only from RGB images, which could bring uncertainties for the translation estimation. Without depth information, a traditional pose estimation model has to memorize the size of a specific object for coordinate transformation, which can be further utilized to transform each object from 2D image pixels to 3D space. This is not only difficult for the model to learn, but also prone to error when facing similar objects with different actual sizes, which is often the case for industrial parts such as bolts. Although Xiang et al. [158] have explored tackling this problem with the ICP refinement algorithm, it is infeasible for real industrial applications due to the time-consuming computation.

To avoid the above issue, this study simply attaches the depth image as an extra channel to the RGB image to form an RGB-D image, which is further input to the pose estimation model for processing.

2) High-Resolution Feature Extraction

A critical limitation that hinders current methods [174, 158] of 6-DoF pose estimation is that deep neural networks are prone to lose spatially feature representation after the gradually downsampling operations, e.g., VGGNet [161] and GoogLeNet [162]. Inspired by [163], this study adopts the backbone network design of High-Resolution Networks, which can ensure both spatially precise and semantically strong feature representation for 6-DoF pose estimation. The major differences between classification network design and high-resolution network design are shown as Fig. 3.2. While normal deep learning networks quickly decrease the feature map size, the high-resolution network maintains the spatial resolution throughout the process.

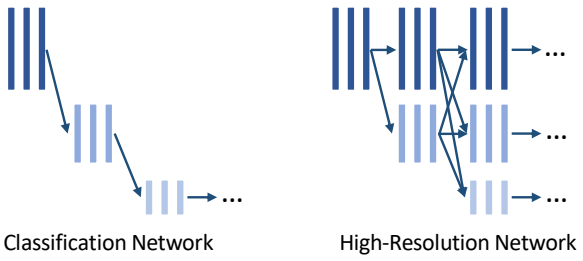


Fig. 3.2: Backbone comparison.

The input to the network has the shape $4 \times H \times W$, where 4 represents the 4-channel RGB-D image, H means image height and W means image width. The first two convolution layers have 3×3 kernels and the strides are 2, after which the feature map resolution is decreased to $\frac{H}{4} \times \frac{W}{4}$.

The main body of the network consists of several parallel branches with different spatial resolutions. As Fig. 3.1 (b) shows, the uppermost branch maintains the resolution $\frac{H}{4} \times \frac{W}{4}$ until the final fusion, while gradually lower-resolution branches are added to the network one-by-one with $\frac{1}{2}$ of resolution of the previous branch until there are four branches with different resolutions.

In the middle of the model, there are interleaved connections between the parallel branches every few convolutions. It is commonly acknowledged that low-resolution feature maps represent semantic information better while high-resolution feature maps contain more precise spatial information. The interleaved design is leveraged to better fuse and exchange information between multiresolution branches.

At the final part of the network, feature maps from higher-resolution branches are first downsampled by $\frac{1}{2}$ and concatenated with the ones from lower-resolution branches. This process repeats until all the features are squeezed into the final feature maps, which are further processed by 2 fully connected (FC) layers sequentially. Finally, the translation parameters are estimated by an FC layer with 3 neurons and the rotation parameters are estimated by another FC layer with 4 neurons.

3) 6-DoF Pose Estimation

With the definition of input format and network structure in previous parts, the model is ready to do forward inference. To be able to train the model, a loss function is required to represent the prediction error. In this part, the

pose parameterization is first introduced and then the loss function is defined based on the estimated pose parameters.

The 6-DoF pose is represented by a rotation $\mathbf{R}$ and a translation $\mathbf{t}$. Let $\mathbf{t} = (t_x, t_y, t_z)^T$ be the translation vector of the object, where t_x and t_y represent the object center in the image coordinates and t_z the average distance from the object to the camera. Here t_x and t_y are actually the pixel deviations from the left-top corner of the cropped image patch to the centre of the object for the convenience of implementation. The actual position could be easily obtained by combining this representation and the bounding box coordinates of the object from the detection stage. And the loss function for translation regression is defined as:

$$L_t(\hat{t}, t) = \begin{cases} 0.5(\hat{t} - t)^2 & \text{if } \left|\hat{t} - t\right| < 1 \\ \left|\hat{t} - t\right| - 0.5 & otherwise \end{cases}, \tag{3.1}$$

where $\hat{t}$ denotes the ground truth translation and t denotes the estimated translation. Notice that this is the smooth-L1 loss function [172], which is differentiable at 0. Following existing work [158], the rotation $\mathbf{R}$ is represented using a quaternion $\mathbf{q} = q_r + q_i\mathbf{i} + q_j\mathbf{j} + q_k\mathbf{k}$ as follows:

$$\mathbf{R} = \begin{bmatrix} 1 - 2\left(q_j^2 + q_k^2\right) & 2\left(q_i q_j - q_k q_r\right) & 2\left(q_i q_k + q_j q_r\right) \\ 2\left(q_i q_j + q_k q_r\right) & 1 - 2\left(q_i^2 + q_k^2\right) & 2\left(q_j q_k - q_i q_r\right) \\ 2\left(q_i q_k - q_j q_r\right) & 2\left(q_j q_k + q_i q_r\right) & 1 - 2\left(q_i^2 + q_j^2\right) \end{bmatrix}, \tag{3.2}$$

which is easier for the model to learn than naive rotation angles. And the loss function for rotation regression is defined as:

$$L_R(\hat{R}, R) = \frac{1}{N} \sum_{i \in N} \min_{j \in N} \left\| \hat{R}x_i - Rx_j \right\|^2, \tag{3.3}$$

where x_i denotes the i^{th} point of N points of the 3D object model, x_j the j^{th} point, $\hat{R}$ the ground truth rotation, and R the estimated rotation. The basic idea is to apply the ground-truth rotation and estimated rotation to a point of the object model and calculate the L2 distance. But for symmetrical objects, different rotation angles might result in the same appearance, which cannot be represented well by simply taking the L2 distance of applying rotation matrices to the same point. So this loss function instead measures the distance between a point with the estimated rotation and the closest point with the ground-truth rotation. The overall loss function is simply defined as the sum of the previous two loss functions:

$$L_{overall} = L_R + L_t \tag{3.4}$$

3.2.3 Pose Refinement

To improve the pose estimation accuracy, a pose refinement stage is introduced in this study. The goal is to predict the pose estimation error of the coarse pose estimation stage. To achieve this goal, the estimated pose from the previous stage is first applied to the 3D object model to obtain a rendered RGB-D image, which is then concatenated with the cropped RGB-D image same as the input of the coarse pose estimation stage to form an 8-channel tensor as the input of pose refinement model.

The backbone network of pose refinement takes the same design as the coarse pose estimation stage. Although this is not compulsory, this study utilizes the same backbone model for the convenience of implementation and also exploits the advantage of high-resolution feature representation. The output

format is also highly similar to that of the coarse pose estimation stage. The only difference is that the estimated target is the relative pose error $\mathbf{\Delta R}$ and $\mathbf{\Delta t}$ rather than absolute pose parameters. Applying the estimated pose error to the coarse pose, the final pose is obtained as follows:

$$\mathbf{R}_{\mathbf{final}} = \mathbf{\Delta R}\mathbf{R}, \tag{3.5}$$

$$\mathbf{t}_{\mathbf{final}} = \mathbf{t} + \mathbf{\Delta t}, \tag{3.6}$$

where $\mathbf{R}_{\mathbf{final}}$ and $\mathbf{t}_{\mathbf{final}}$ represent the final rotation and translation. The loss functions for pose refinement also take the same form as in coarse pose estimation but replacing $\mathbf{R}$ and $\mathbf{t}$ with $\mathbf{R}_{\mathbf{final}}$ and $\mathbf{t}_{\mathbf{final}}$ respectively.

3.3 Hand-Object Pose Estimation with Explicit Occlusion Awareness

The overall model architecture of the proposed joint hand-object 3D pose estimation model is shown in Fig. 3.3, in which the model is mainly divided into three parts: (a) mask-guided attentive feature extraction, (b) hand-object dense pose estimation, and (c) occlusion awareness. A monocular RGB image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, which contains the hand-object interaction is captured as the input to the proposed model. It is assumed that the hand-object area and the object type are already known since they are not the focus of this study and can be easily obtained via an object detection model. The image $\mathcal{I}$ is first input into the backbone network, which is designed based on ResNet50 [175], for hand and object feature extraction guided by hand and object binary masks respectively. Then in the hand-object pose estimation part, the extracted feature vectors are input to several fully connected layers to

predict the pose parameters including hand translation $\mathcal{T}_{hand}$, pose $\mathcal{P}_{hand}$, shape $\mathcal{S}_{hand}$, object rotation Θ_{obj}, and object translation $\mathcal{T}_{obj}$, which are subsequently applied to the object 3D model and MANO (hand Model with Articulated and Non-rigid defOrmations) hand model [176] to obtain the 3D geometric reconstructions. And the rendered ternary mask can be further generated by projecting the 3D reconstructions back to 2D image plane via differentiable rendering [177]. Meanwhile, the intermediate feature maps from the feature extraction stage are leveraged to construct an FPN-like (Feaure Pyramid Network) [178] subnetwork to predict the ternary mask in a segmentation fashion. Finally, the consistency between the predicted and rendered ternary masks is calculated and regarded as a constraint term $\mathcal{L}_{consist}$ in the training loss function which will be elaborated in more detail in the following sections.

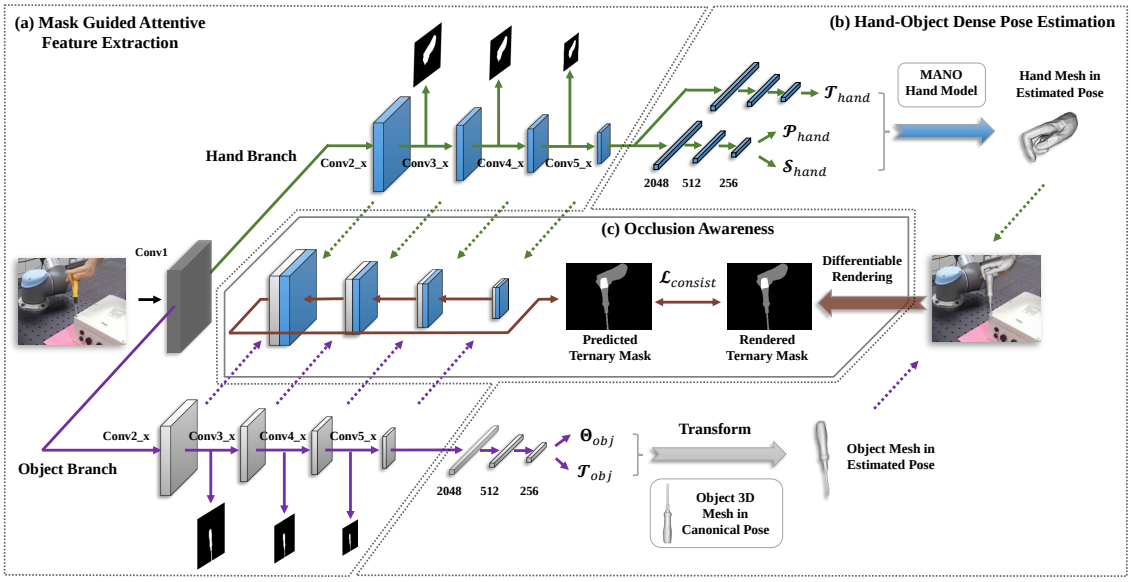


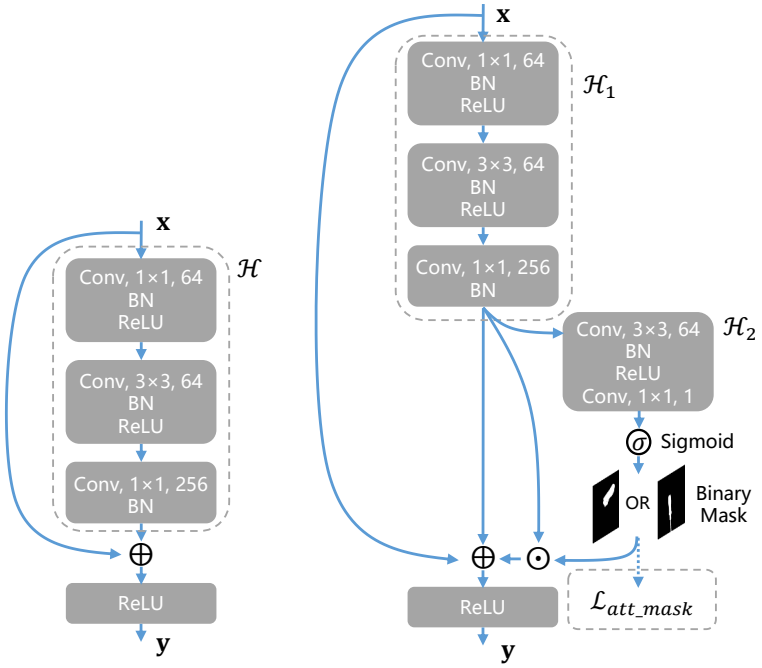
Fig. 3.3: Architecture of the proposed integrated hand-object dense pose estimation model.

3.3.1 Mask-Guided Attentive Feature Extraction

Feature extraction is an essential step to create a nonlinear mapping between the image space and the high-dimensional feature space in order to produce a compact and expressive feature representation $\mathcal{F} \in \mathbb{R}^N$ of the input image $\mathcal{I}$. It is widely accepted that deep learning models, especially CNNs, are highly effective for image feature extraction, thus we also follow the spirit to construct the feature extractor based on the renowned ResNet50 [175]. Previous works normally employ ResNet directly for hand-object feature extraction [179], which can bring certain ambiguity especially when the hand and object are partially occluded by each other. To make the model better attend to hand and object areas respectively, we decided to introduce hand and object binary masks $\mathcal{M} \in \mathbb{R}^{H \times W}$ as intermediate supervision signals during model training, which naturally leads to separated branches for hand feature $\mathcal{F}_{hand} \in \mathbb{R}^N$ and object feature $\mathcal{F}_{obj} \in \mathbb{R}^N$. The details of the proposed mask-guided attentive module and the backbone model structure are illustrated as follows.

1) Mask-Guided Attentive Residual Block

The original ResNet was designed to classify the most salient object in an image into its corresponding category, which may not be directly suitable for this task where there are two main subjects. Inspired by the recently prevailing spatial attention mechanism, we propose the mask-guided attentive residual (MAR) block to incorporate binary masks of the hand $\mathcal{M}_{hand}$ or object $\mathcal{M}_{obj}$ into the original residual block.



(a) Original bottleneck residual block. (b) Proposed MAR block.

Fig. 3.4: Comparison between the original residual block (a) and the proposed mask-guided attentive residual block (b).

The detailed comparison between the proposed MAR block and the original one is depicted in Fig. 3.4. Following He et al.'s definition [175], the original residual block can be defined as:

$$\mathbf{y} = \phi\left(\mathcal{H}\left(\mathbf{x}\right) + \mathbf{x}\right), \tag{3.7}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output of the residual block, $\mathcal{H}$ represents the in-between layers, and ϕ denotes the ReLU activation function. Hence, the proposed MAR block can be formulated as:

$$\mathbf{y} = \phi\left(\mathcal{H}_1(\mathbf{x}) \odot \left(\sigma\left(\mathcal{H}_2\left(\mathcal{H}_1(\mathbf{x})\right)\right)\right) + \mathcal{H}_1(\mathbf{x}) + \mathbf{x}\right), \tag{3.8}$$

where $\mathcal{H}_1$ represents the original operations as $\mathcal{H}$ in (3.7), $\mathcal{H}_2$ the extra layers for binary mask generation, σ the Sigmoid function, and $\odot$ the element-wise product. During training, the generated binary mask is compared with the

ground-truth mask through the binary cross-entropy loss, which is denoted as:

$$\mathcal{L}_{att\_mask} = -\frac{1}{HW} \sum_{i \in \mathcal{M}_{HW}} (y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)), \tag{3.9}$$

where y_i is the binary label of a pixel, p_i is the predicted probability of the pixel being foreground, $\mathcal{M}_{HW}$ represents the generated binary mask with H and W being the height and width respectively. The objective of this design is to impose extra supervision throughout the feature extraction process so that the learned feature can be more focused on the target area.

2) Parallel Feature Extraction Branches

With the MAR block, the backbone network naturally takes form in a branched shape as the goal is to enhance the model attention for hand or object exclusively in separate branches, which can also benefit the model training by reducing the ambiguity of the supervision signals. The detailed structure of the feature extraction network is illustrated in Table 3.1. The first column denotes the names of different layer groups, the second column represents the size of output (i.e., C for channels, H for height, and W for width) for each layer, and the third one shows the details of the model structure. The input to the network is an RGB image $\mathcal{I} \in \mathbb{R}^{3 \times H \times W}$ with $H = 270$ and $W = 480$ following the practice in [179]. The image is first processed by a 7×7 convolution layer and a 3×3 max pooling layer for low-level feature extraction. Then the model splits into two branches for hand and object respectively with the same structure, where the only difference is the intermediate mask supervision for each MAR block. The design for layer groups $Conv2\_x$ to $Conv5\_x$ generally follows the original ResNet50 with only the last residual block of each layer group replaced with the MAR block.

Table 3.1: Architecture of the Feature Extraction Network

Layers	**Size** $(C \times H \times W)$	**Model**	
Input	$3 \times 270 \times 480$		
Conv1	$64 \times 135 \times 240$ $64 \times 68 \times 120$	Conv, $7 \times 7, 64$, Stride 2 Max Pool, 3×3, Stride 2	
		(Hand Branch)	(Object Branch)
Conv2_x	$256 \times 68 \times 120$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 2$ $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix}_{MAR} \times 1$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 2$ $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix}_{MAR} \times 1$
Conv3_x	$512 \times 34 \times 60$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$ $\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix}_{MAR} \times 1$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$ $\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix}_{MAR} \times 1$
Conv4_x	$1024 \times 17 \times 30$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 5$ $\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix}_{MAR} \times 1$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 5$ $\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix}_{MAR} \times 1$
Conv5_x	$2048 \times 9 \times 15$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 2$ $\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix}_{MAR} \times 1$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 2$ $\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix}_{MAR} \times 1$
	$2048 \times 1 \times 1$	Average Pool	Average Pool

Finally, average pooling layers are utilized to convert the feature maps into feature vectors $\mathcal{F}_{hand} \in \mathbb{R}^N$ and $\mathcal{F}_{obj} \in \mathbb{R}^N$ with $N = 2048$.

3.3.2 Hand-Object Dense Pose Estimation

After feature extraction, the second part of the proposed model focuses on the estimation of the pose parameters for hand and object and the reconstruction of the 3D geometry.

1) 3D Hand Reconstruction

To simplify the problem of 3D hand reconstruction, we employ a parametric model MANO [176], which is able to generate a 3D hand mesh based on a set of pose and shape parameters. Following the practice in [179], the hand feature vector $\mathcal{F}_{hand} \in \mathbb{R}^{2048}$ is taken as input to several FC layers which subsequently regress the pose parameters $\mathcal{P}_{hand} \in \mathbb{R}^{15}$ that represent the 3D rotations of hand joints, and shape parameters $\mathcal{S}_{hand} \in \mathbb{R}^{10}$ for controlling the shape characteristics of different hands such as the length between joints. An additional FC layer branch is introduced to predict the spatial translation $\mathcal{T}_{hand} \in \mathbb{R}^3$. The process of the MANO model can be symbolized as:

$$(\mathcal{V}_{hand}, \mathcal{J}_{hand}) = MANO\left(\mathcal{P}_{hand}, \mathcal{S}_{hand}\right) + \mathcal{T}_{hand}, \tag{3.10}$$

where $\mathcal{V}_{hand} \in \mathbb{R}^{N\times 3}$ represents the 3D coordinates of the hand mesh vertices with $N = 778$, and $\mathcal{J}_{hand} \in \mathbb{R}^{M\times 3}$ the hand skeleton joints with $M = 21$. During training, the hand reconstruction supervision is imposed via calculating the l_2 loss between the estimated hand joints and the ground truth:

$$\mathcal{L}_{hand} = \frac{1}{M} \sum_{i \in M} \left\| J_i - \hat{J}_i \right\|^2, \tag{3.11}$$

where $J_i \in \mathcal{J}_{hand}$ represents a hand joint and $\hat{J}_i$ is the corresponding ground truth annotation. Meanwhile, two l_2 regularization loss terms are applied to the training including $\mathcal{L}_{\mathcal{P}_{hand}} = \sum_i \|P_i\|^2$ and $\mathcal{L}_{\mathcal{S}_{hand}} = \sum_i \|S_i\|^2$ to prevent extreme values for these parameters, which may cause unnatural hand geometry.

The output hand mesh is in the 3D camera coordinates, which already implicitly includes the camera extrinsic parameters in the pose parameters. To further reproject the 3D hand vertices into the 2D image plane, the perspective projection is employed based on the camera intrinsic matrix:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{3.12}$$

where (u, v) represent the pixel location in 2D image plane, (x, y, z) are the 3D coordinates of a vertex. (f_x, f_y, c_x, c_y) are the camera intrinsic parameters, where (f_x, f_y) are the horizontal and vertical focal lengths, and (c_x, c_y) are the center position of the camera view.

2) 6-DoF Object Pose Estimation

The pose estimation for object is less perplexing than its hand counterpart as we assume the workpieces and tools are rigid and their corresponding 3D mesh models are provided. The problem is commonly formulated as 6-DoF pose estimation as in [47, 180, 181]. Like the hand parameter regression, a similar FC branch is appended after the object feature vector $\mathcal{F}_{obj} \in \mathbb{R}^{2048}$ to predict the spatial rotation $\Theta_{obj} = (\alpha, \beta, \gamma) \in \mathbb{R}^3$ in the Euler angles form and translation $\mathcal{T}_{obj} = (t_x, t_y, t_z) \in \mathbb{R}^3$ of the target object. The rotation matrix

$\mathcal{R}_{obj} \in SO(3)$ can be easily constructed from the estimated spatial rotation angles, based on which the transformation from the canonical object mesh model to its estimated pose can be denoted as:

$$\mathcal{V}_{obj} = \mathcal{R}_{obj}\mathcal{V}_{can} + \mathcal{T}_{obj}, \tag{3.13}$$

where $\mathcal{V}_{obj} \in \mathbb{R}^{N\times 3}$ denotes the coordinates of object vertices in the estimated pose, while $\mathcal{V}_{can} \in \mathbb{R}^{N\times 3}$ represents the given object mesh model in the canonical pose. Then the loss for object pose estimation is similarly defined as a l_2 loss as the hand pose. Nevertheless, instead of the hand joints, the distance of the vertices is measured, since there is no joint point in the object model:

$$\mathcal{L}_{obj} = \frac{1}{N}\sum_{i\in N}\left\|V_i - \hat{V}_i\right\|^2, \tag{3.14}$$

where $V_i \in \mathcal{V}_{obj}$ is the ith object vertex and $\hat{V}_i$ the ground truth. The reprojection of object vertices to 2D space follows the same rule as in the case of hand.

3.3.3 Explicit Occlusion Awareness

The mutual occlusion between hand and object has always been a challenging issue in the hand-object pose estimation task because of the performance degradation caused by the loss of discriminant features in the occluded area. To mitigate this problem, an extra occlusion-aware mechanism is introduced, which can empower the model by explicitly predicting the ternary mask of the occlusion area via a FPN-like structure and further constraining it by comparing with the rendered ternary mask from the estimated hand-object pose.

1) Ternary Occlusion Mask Prediction

The ternary occlusion mask is proposed in this study as an extension of the binary mask. A binary mask generally represents the background pixels with 0 and foreground 1, while a ternary mask additionally represents the occlusion area with value 2. Let $\mathcal{M}_t \in \mathbb{R}^{H \times W}$ denotes the ternary mask, then we have

$$\mathcal{M}_t(i,j) = \begin{cases} 2 & \text{if } (i,j) \in P_{hand} \cap P_{obj} \\ 1 & \text{if } (i,j) \in P_{hand} \cup P_{obj} \text{ and} \\ & \quad (i,j) \notin P_{hand} \cap P_{obj} \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

where $\mathcal{M}_t(i,j)$ is the pixel at position (i,j). P_{hand} and P_{obj} represent the pixel set of hand and object areas, respectively.

Inspired by the FPN network [178], we extract the intermediate feature maps from the hand and object branches, concatenate them at different scales, and add convolution and upsampling operations to form a bottom-up pathway. Upsampling is mainly utilized to adapt the spatial sizes between different scales, while 1×1 convolution is leveraged to accommodate the channel size of feature maps. Finally, a 3×3 convolution layer is stacked upon the largest merged feature map to obtain the ternary mask prediction $\mathcal{M}_{t\_pred} \in \mathbb{R}^{H \times W}$ with $H = 68$ and $W = 120$. A cross-entropy loss $\mathcal{L}_{t\_pred}$, which is widely used in segmentation tasks, is adopted for model training.

2) Neural Rendering-Based Occlusion Consistency

With the estimated pose and mesh models, one can also obtain a rendered ternary mask $\mathcal{M}_{t\_render} \in \mathbb{R}^{H \times W}$ via differentiable rendering technique [177],

which is also called neural rendering when integrated into a neural network. The normal rendering process is non-differentiable because of the discrete nature of the rasterization operation, while differentiable rendering provides an approximation for the gradient of rasterization, thus making the whole rendering process differentiable and possible to be incorporated in the gradient-based optimization.

The ternary mask prediction can provide occlusion-aware supervision signals through the back-propagation only to the front part of the model but not the pose estimation layers, which can be addressed by rendering the 3D mesh models in estimated poses back to the image plane to generate the rendered ternary mask. Similarly, a cross-entropy loss $\mathcal{L}_{t\_render}$ is leveraged for training. Meanwhile, with the predicted and rendered ternary masks, an extra constraint loss term can be defined as:

$$\mathcal{L}_{consist} = \left\| \mathcal{M}_{t\_pred} - \mathcal{M}_{t\_render} \right\|^2, \tag{3.16}$$

which is supposed to constrain the two ternary masks to be consistent with each other, and the overall loss function for the model can be finally written as:

$$\begin{aligned} \mathcal{L}_{overall} = \mathcal{L}_{hand} + \mathcal{L}_{obj} + \\ \lambda_1 \mathcal{L}_{t\_pred} + \lambda_2 \mathcal{L}_{t\_render} + \\ \lambda_3 \mathcal{L}_{s\_hand} + \lambda_4 \mathcal{L}_{p\_hand} + \lambda_5 \mathcal{L}_{consist}, \end{aligned} \tag{3.17}$$

where λ_i denoting the corresponding weights of the loss terms which are empirically set as $\lambda_1 = \lambda_2 = 1 \times 10^{-5}$, $\lambda_3 = \lambda_4 = 1 \times 10^{-6}$, and $\lambda_5 = 1 \times 10^{-7}$.

3.4 Experimental Results

3.4.1 Evaluation of the Object Pose Estimation Model

The performance of the proposed model is evaluated on the new test set. Following [158], the average distance with symmetrical objects (ADD-S) is leveraged as the evaluation metric:

$$ADD - S = \frac{1}{N} \sum_{i \in N} \min_{j \in N} \left\| (\hat{R}x_i + \hat{t}) - (Rx_j + t) \right\|^2, \tag{3.18}$$

which basically takes the same idea as the loss function for rotation regression. While previous work normally choose a predefined distance threshold to calculate the percent accuracy, this study directly reports the average distance calculated by equation (3.18) for simplicity.

Table 3.2 presents the evaluation results comparison between the baseline model and the proposed model. The baseline model was proposed in [182], which had the best performance on the utilized dataset, and only the single-view model is adopted in the experiments of this work because the dataset only contains single-view data. The main differences between the baseline model and the proposed model lie in the backbone design and depth image usage. Concretely, the baseline model uses EfficientNet-B3 [183] as the backbone network, which is the state-of-the-art classification model on ImageNet, and the baseline model only takes RGB image as input without depth channel.

Table 3.2: Evaluation Results Comparison

Method	Backbone	Depth	Refinement	ADD-S
Baseline[182]	EfficientNet-B3	w/o	w/o	0.0636
		w/o	w/	0.0252
Proposed	HRNetV2-W32	w/	w/o	0.0442
		w/	w/	**0.0163**

As Table 3.2 depicts, the proposed model with refinement has the smallest ADD-S distance of 0.0163 on the test set. Compared with the baseline model, the proposed model performs better with or without the refinement stage, suggesting that the introduction of the high-resolution model design and depth image significantly improves the performance. Note that the reported results of the baseline model are different from the original paper because the experimental setup and the evaluation metric are different. In terms of evaluation time for a single sample, the baseline model requires 0.69s on average for the whole process but varies from 0.4s to 0.9s depending on the number of objects presented, while the proposed model is about 0.2s slower on average.

Fig. 3.5 demonstrates some examples of 6-DoF pose estimation results. For each pair, the left picture is the input image and the right one represents the rendered image according to the estimated 6-DoF pose parameters. The left two columns show some good examples, while the right two columns are several failure cases, which shows the model still has trouble estimating the rotation of symmetrical objects.

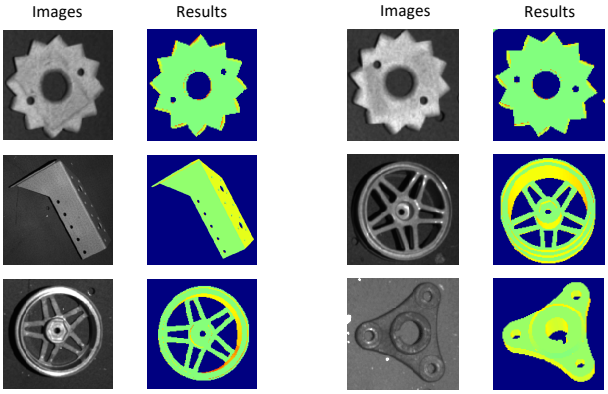


Fig. 3.5: Examples of 6-DoF pose estimation results.

3.4.2 Evaluation of the Hand-Object Pose Estimation Model

The prosperity of the electrical vehicle (E-V) industry in recent years poses a foreseeable challenge in the near future that, an enormous amount of ageing E-V battery modules will flush into the recycling market putting heavy pressure on the disassembly operations. While employing collaborative robots seems to be a viable solution, the limited cognitive capabilities for real-time interactions of human hands and disassembly workpieces largely hinder the HRCD (human-robot collaborative disassembly) efficiency.

To demonstrate the effectiveness of the hand-object pose estimation approach, a case study on the HRCD of Li-ion (Lithium-ion) battery module is conducted, as depicted in Fig. 3.6. In this context, hand-object interaction images are captured and organized into a dataset for the evaluation of the proposed method. Other comparative experiments on a public dataset are also carried out to further demonstrate the generalization and universality of the proposed approach.

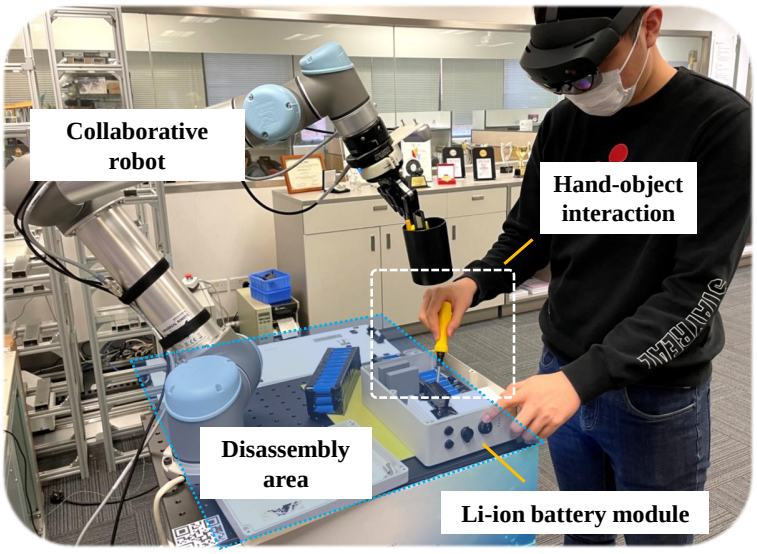


Fig. 3.6: Demonstration of the human-robot collaborative Li-ion battery module disassembly.

1) Human-Robot Collaborative Li-ion Battery Disassembly

Data Collection. During the simulated Li-ion battery disassembly task, a human operator is instructed to perform some of the disassembly steps, such as unfastening the screws, opening the module shell, cutting the connecting wires, etc., while the robot aims to mainly provide some assistance, such as delivering the disassembly tools, picking the disconnected parts and placing into baskets. Meanwhile, an industrial camera is installed to capture images of the working area in the disassembly process, and the hand-object interaction areas are further extracted as the input to the pose estimation model.

In total, 501 images were collected, including 6 types of different disassembly tools and parts: 1) screwdriver, 2) wire cutter, 3) hammer, 4) structural part, 5) small battery pack, and 6) long battery pack. The hand-object 3D

pose labels are annotated manually, with 400 samples randomly selected for model training and 101 samples for evaluation.

Experimental Setup. The model is implemented using Pytorch deep learning library and trained with an RTX3080 GPU for acceleration. The backbone layers are initialized with the weights from the pretrained ResNet50 model provided by Pytorch, while other layers are randomly initialized. Adam optimizer is employed for model training with initial learning rate 1×10^{-5}, batch size 2, and 1000 training epochs.

Evaluation Results. To evaluate the hand-object 3D pose estimation performance of the proposed method, the mean joint error is calculated for hand pose evaluation and the mean vertex error for object, following the practice in [179]. The two metrics have similar forms as (3.11) and (3.14) respectively.

The quantitative comparison result is illustrated in Table 3.3. The single-frame model from one of the state-of-the-art works [179] is leveraged as the baseline model. Three variants of the proposed model consisting of different components are included in the experiments. The *Branched model* denotes the branched feature extraction network but without the MAR block, which exhibits a moderate improvement over the baseline model. The *MAR block* contributes considerably to the decrease of the pose estimation error, while the *Occlusion awareness* mainly promotes the object pose estimation accuracy, which is as expected since the object is normally the one that is being occluded by hand in the data. The last row of the table shows the performance of the overall model, from which we can observe a considerable improvement over the baseline method by around 15% for object and 21%

for hand. In terms of inference speed, the model can achieve 39 FPS (frames per second) on an RTX3080 GPU, which is sufficiently fast for real-time applications.

Table 3.3: Pose Estimation Performance on The Collected Dataset

Method	Components	Object Pose Error (mm)	Hand Pose Error (mm)
Hasson *et al.* [179]	Baseline	45.45	45.81
Ours	Branched model	43.63	40.22
	Branched model + MAR block	41.60	36.01
	Branched model + MAR block + Occlusion awareness	**38.40**	**35.99**

The qualitative results on some of the collected images are presented in Fig. 3.7. The first row presents the input RGB images, the second row is the visualization of the estimated hand-object pose from the baseline method, the third one depicts the results from the model, and the final row is the projection overlay of 3D meshes to the image plane, from which it is easier to inspect the accuracy of the estimation results of the model. Although it is evident that the method can achieve better alignment between the estimated poses and the input images, the results are still far from perfect, especially when taking a closer look at the overlay images. One possible reason is that the dataset scale is relatively small with only several hundreds of samples, which may not be sufficient to fully cover all possible hand-object interaction angles and postures.

2) Experiments on the F-PHAB Dataset

To further demonstrate the generalization ability and universality of the proposed model, some additional experiments are conducted on a large-scale

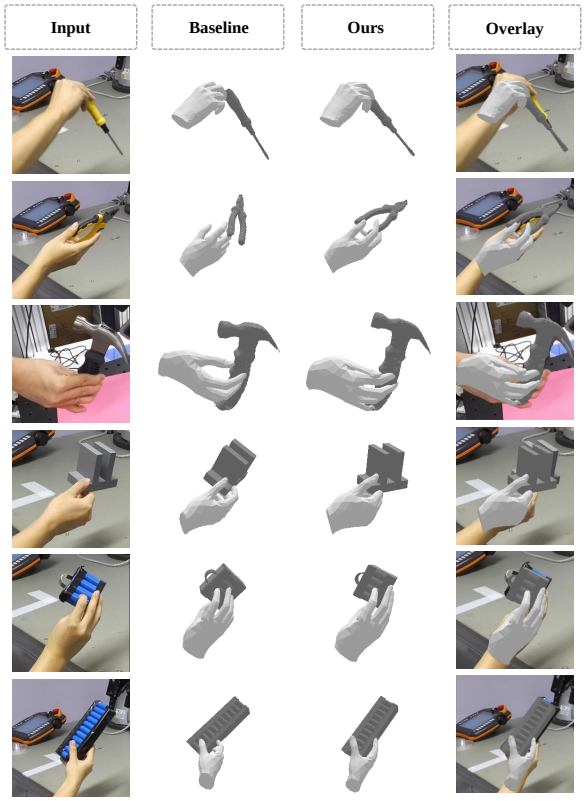


Fig. 3.7: Qualitative comparison on the test data between the baseline and the proposed model. We select one sample for each of the 6 object categories for demonstration.

public hand-object dataset F-PHAB (First-Person Hand Action Benchmark) [184], which was collected in a daily hand-object interaction scenario consisting of more than 100K images. The experimental results of our model compared with some state-of-the-art methods are shown in Table 3.4. The hand and object pose errors of some methods [185, 179, 186] are directly extracted from the published papers. For Doosti et al. [187], we modified their open source code to train the full model, since the experiments reported in their paper followed a different setting from ours. And the object pose error of [187] is not provided because their method only represents the object pose by bounding box coordinates, which cannot be directly compared with the posed object meshes in our method. It can be easily identified from the table that our model achieves the overall best accuracy in both hand and object pose estimation.

The computational cost and number of parameters are also appended in Table 3.4, which are calculated based on the open source implementations of these methods except for [186], of which no code is provided. It can be seen that our model is relatively bulkier and computationally heavier than the compared ones. This is not unexpected considering that our model utilizes two separate branches for hand and object respectively. Nevertheless, our model is still able to achieve a fast inference speed (39 FPS) with GPU acceleration as mentioned earlier, therefore we believe the extra computational cost should not be a major issue preventing it from real-time applications.

Table 3.4: Pose Estimation Performance and Model Complexity Comparison on The F-PHAB Dataset

Method	Object Pose Error (mm)	Hand Pose Error (mm)	FLOPs	Params
Tekin *et al.* [185]	24.89	15.81	13.62G	14.31M
Hasson *et al.* [179]	22.30	18.00	38.46G	11.99M
Huang *et al.* [186]	21.37	15.18	-	-
Doosti *et al.* [187]	-	15.62	8.28G	23.90M
Ours	**20.73**	**14.34**	96.50G	29.45M

3.4.3 Discussions

The experimental results on the collaborative Li-ion battery disassembly data and the public F-PHAB dataset both show consistent improvements of our proposed approach over previous works, in terms of 3D hand-object pose estimation. This is mainly attributed to better hand-object attention separation brought by the mask-guided attentive feature extraction model and enhanced occlusion robustness owing to the explicit occlusion awareness mechanism. Compared with existing works, our proposed method can not only produce the 3D hand-object pose estimation via an integrated model,

but also take a step further to gain insights from the inherent problems in the hand-object interaction scenarios, such as hand-object area ambiguity and mutual occlusion.

This simultaneous hand-object pose perception scheme is capable to enhance robot cognition skills, especially in HRCD scenarios to enable adaptive robot decision-makings and proactive collision avoidance. Nevertheless, some issues like limited data scale and lack of temporal information should be delved deeper in future explorations.

3.5 Chapter Summary

The HRC manufacturing paradigm puts collaborative robots on the manufacturing shop floor to work seamlessly alongside human operators. To equip those robots with the ability to understand the ongoing human hand-object interactions in an HRC environment, a high-resolution network-based 6-DoF object pose estimation model and an integrated hand-object 3D pose estimation approach were proposed in this chapter. The main contributions of this chapter can be summarized in threefold: 1) a two-stage coarse-to-refine 6-DoF pose estimation model with high-resolution feature exploration ability was proposed for industrial parts; 2) for hand-object pose estimation, a mask-guided attentive residual block was proposed in cooperation with the branched model structure to achieve finer hand-object attention separation during the feature extraction stage; 3) an FPN-like subnetwork was leveraged to predict the occlusion ternary mask which was compared with the rendered mask from the estimated hand-object pose to achieve explicit occlusion awareness to further reduce the pose estimation error caused by the occlusion.

The subsequent experimental results on both the public benchmarks and the Li-ion battery disassembly case demonstrated an obvious improvement over existing methods.

4 Human Operator Digital Twin Modelling

As mentioned in Chapter 2, human recognition is one of the most prevailing topics of vision application in previous HRC research works. It is not unanticipated since human plays the most important role in an HRC team. In order to enhance both human well-being and robotic flexibility within HRC, existing research efforts focused on human body perception but lack a holistic perspective of the human operator. A novel approach to addressing this challenge is the construction of an HDT, which serves as a centralized digital representation of various human data for seamless integration into the cyber-physical production system. However, the implementation of visual perception-based HDT remains underreported within the HRC realm. To this end, this chapter proposes an exemplary vision-based HDT model for highly dynamic HRC applications. The model mainly consists of a convolutional neural network that can simultaneously model the hierarchical human status including 3D human posture, action intention, and ergonomic risk. Then, on the basis of the constructed HDT, a robotic motion planning strategy is further introduced with the aim of adaptively optimizing the robotic motion trajectory.

The research content presented in this chapter is mainly based on a journal paper published in the Journal of Manufacturing Science and Engineering [188].

4.1 Introduction

The vision of Industry 5.0 has put additional emphasis on the transition from a technology-centred fashion to a sustainable, human-centric, and resilient industry [189]. Unlike the traditional purely profit-driven manufacturing paradigm, a human-centric approach prioritizes the well-being of human operators throughout the entire manufacturing process. In this context, human-robot collaboration has become increasingly prevailing in recent years because of the ability to unleash the full potential of human-centric creative problem solving with the assistance of robotic automation when encountered with flexible or uncertain situations [2, 190, 191].

To enhance human well-being during HRC, it is essential for the robot to be equipped with the ability to perceive the human body in an accurate and timely manner, to which abundant research efforts have been devoted over the years. Some previous researchers have adopted wearable sensors [192] or motion capture suits [193] to perceive human body posture for robotic applications. Although the pose capturing accuracy has been widely acknowledged, wearing additional equipment can cause human discomfort, especially after working long hours, and may not be practical in actual manufacturing shopfloors.

Another trend is to leverage computer vision techniques to parse human postures and actions via RGB or depth cameras in a non-intrusive way. For instance, Liu et al. [98] adopted RGB-D sensors and deep learning models to capture human skeleton pose and spatial occupancy to achieve real-time collision avoidance in an HRC system. Parsa et al. [93] proposed a spatial-temporal convolutional neural network to recognize human actions and

associated ergonomic risks from RGB-D video streams. However, most of these works only managed to consider a certain recognition task such as skeleton pose recognition, action recognition, etc., which leaves the holistic modelling of human operators largely unexplored. One promising approach to address this challenge is the creation of an HDT, which offers a unified digital representation of diverse human data that can be seamlessly integrated into the cyber-physical production system. The deployment of HDT can be advantageous for optimizing system performance and facilitating solution recommendations during HRC operations. Although the concept of HDT has been actively investigated by recent scholarly discourse [194, 195, 196], the practical implementation of HDT for HRC cases has received scant attention in the literature.

Therefore, this chapter presents a vision-based HDT model based on deep convolutional neural networks with concurrent consideration of multiple aspects of human status perception including 3D human posture, action intention, and ergonomic risk. This study intends to provide an exemplary implementation of HDT that can gather more holistic human information to enhance the adaptivity and flexibility of cobots (Collaborative Robots) during motion planning and action execution.

4.2 Vision-based HDT Modelling

Following the definition of HDT in [195], we extend it into the HRC context as "a virtual representation of a human worker that is used to optimize the collaboration between humans and robots". This HDT is an integrated model created from vision data and is used to facilitate the description, prediction,

and visualization of one or more characteristics of a human or class of humans as they perform within the human-robot scenarios with the aim to identify opportunities to improve worker well-being and ensure effective and safe collaboration between humans and robots.

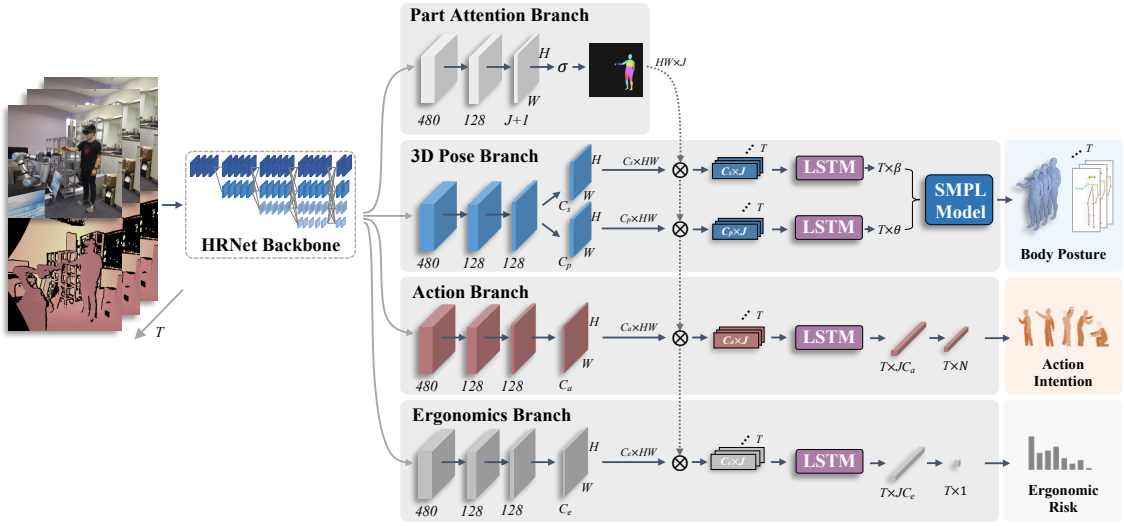


Fig. 4.1: The proposed HDT perception model for HRC

Specifically, based on visual observation of an HRC scene, the proposed vision-based HDT model is established to perceive and synchronize the human operator status which primarily focuses on three specific recognition aspects of the human digital twin including the 3D posture, action intention, and ergonomic risk. The concrete model is depicted in Fig. 4.1, which mainly consists of a cutting-edge CNN backbone and specifically designed functional branches. A short RGB-D sequence $\mathcal{V} = \{\mathcal{I}_t\}_{t=1}^{T}$, where $\mathcal{I}_t$ represents a single frame, of an HRC scene is captured as the input data. Although RGB information alone is sufficient for various recognition needs, the complimentary depth data can provide absolute distance values which are indispensable for the mapping between the HDT and the real world. Then an HRNet-W32 backbone [163] is leveraged to process $\mathcal{V}$ and extract intermediate feature maps $\mathcal{F}_{int} \in \mathbb{R}^{C \times H \times W}$ with $C = 480$, based on which several specifically purposed branches are further constructed: 1) part attention branch, 2) 3D pose

branch, 3) action branch, and 4) ergonomic branch, to fulfil different perception requirements of the HDT modelling process. For simplicity, we only illustrate the single frame case until the aggregation of sequential features right before LSTM modules, which then refine the sequential information and regress into different perception results. Through this unified model, the intended human status can be extracted all at once and updated to the HDT model in real-time. More details of the proposed model will be elaborated in the following subsections.

4.2.1 Body Part Attention

Human body only amounts to a small fraction of the entire image of an HRC scene, which can pose potential hindrances for the CNN model to focus on the relevant area and make reliable recognition assertions if not provided with explicit guidance. To this end, we propose the body part attention mechanism as a direct cue indicating the whereabouts of the human body and highlighting different body parts.

Based on the intermediate feature $\mathcal{F}_{int}$ extracted by the backbone network, two 3×3 convolution layers with 480 and 128 channels, respectively, are first applied to shrink the feature maps, which are then compressed into $(J+1) \times H \times W$ by another convolution layer. $J = 24$ stands for the number of body joints as well as body parts, the extra 1 channel represents the background, and $H = W = 56$ are the height and width of the feature maps. A softmax function σ is utilized to convert the feature maps to part attention mask $\mathcal{F}_{part} \in \mathbb{R}^{J \times H \times W}$, of which each pixel value indicates the likelihood of that pixel belonging to a certain body part. Finally, the part attention mask will be applied to the task-specific feature maps of the following branches to

constrain the features to be focused on the body part areas. $\mathcal{F}_{part}$ is reshaped to $HW \times J$ for the convenience of subsequent multiplications.

4.2.2 3D Human Pose Reconstruction

The most essential component of the proposed HDT is the recognition and reconstruction of human body posture and 3D mesh. Earlier approaches to 3D reconstruction normally rely on multi-perspective data and tend to be time-consuming, while in this work we adopted the SMPL (Skinned Multi-Person Linear Model) [197] human body model which can largely simplify the reconstruction process by blending and stretching a prior template human body mesh model towards the target posture according to the estimated body parameters. Let $\mathcal{M} \in \mathbb{R}^{6890\times3}$ represents the generated 3D body mesh, and $\mathcal{J} \in \mathbb{R}^{J\times3}$ represents 3D skeleton joint coordinates. Then the process of human mesh updating can be simply denoted as:

$$(\mathcal{M}, \mathcal{J}) = SMPL\,(\theta, \beta)\,, \tag{4.1}$$

where $\theta \in \mathbb{R}^{J\times3}$ denotes the estimated pose parameters, including the local rotation parameters for the body joints and the global rotation parameters. $\beta \in \mathbb{R}^{10}$ is the shape parameters drawing from the first 10 PCA (Principal Component Analysis) shape coefficients from the SMPL convention. We adopt the parameter settings and the gender-neutral model following previous practices [198, 199].

Getting back to the proposed model, the mentioned body parameters are estimated via the 3D pose branch, which follows a similar design strategy as the part attention branch in terms of the first several layers, but later splits into two sub-branches with individual feature maps $\mathcal{F}_p \in \mathbb{R}^{C_p \times H \times W}$ and $\mathcal{F}_s \in \mathbb{R}^{C_s \times H \times W}$, where $C_p = 128$ and $C_s = 64$, for better decoupling the pose and shape parameter regressions. $\mathcal{F}_p$ and $\mathcal{F}_s$ will be then reshaped and multiplied with the reshaped part attention mask $\mathcal{F}_{part}$ as mentioned in the previous section for better body area feature emphasis. Note that this process applies to each frame of the input video clip, resulting in two sequential features $\mathcal{F}_{pt} \in \mathbb{R}^{T \times C_p \times J}$ and $\mathcal{F}_{st} \in \mathbb{R}^{T \times C_s \times J}$, which will be exploited by two LSTM modules with hidden size 2048 for temporal coherence aggregation and further regressed to the corresponding body parameters θ_t and β_t for the whole sequence T. For training, three loss terms are employed to provide supervision on the body parameters, joint coordinates, and part attention map, respectively:

$$\mathcal{L}_{SMPL} = \sum_{t \in T} \left\| \theta_t - \hat{\theta}_t \right\|_2^2 + \sum_{t \in T} \left\| \beta_t - \hat{\beta}_t \right\|_2^2, \tag{4.2}$$

$$\mathcal{L}_{joint} = \sum_{t \in T} \left\| \mathcal{J}_t - \hat{\mathcal{J}}_t \right\|_2^2, \tag{4.3}$$

$$\mathcal{L}_{part} = -\sum_{t \in T} \sum_{i \in (J+1)} y_{i,t} \log(p_{i,t}). \tag{4.4}$$

$\mathcal{L}_{SMPL}$ and $\mathcal{L}_{joint}$ aims at measuring the difference between estimations θ_t, β_t, $\mathcal{J}_t$ and their corresponding ground truth labels $\hat{\theta}_t$, $\hat{\beta}_t$, $\hat{\mathcal{J}}_t$. Specifically, $\mathcal{L}_{SMPL}$ is formulated as the summation of two MSE (Mean Squared Error) loss functions over the temporal sequence T, which supervises the regression of SMPL model parameters θ_t and β_t for body joint rotations and body

shape coefficients based on the SMPL convention. Similarly, $\mathcal{L}_{joint}$ is defined following the same principle to constrain the learning of 3D skeleton joint coordinates. These two loss functions have been widely adopted in existing literature for SMPL-based 3D human pose estimation. $\mathcal{L}_{part}$ employs a pixel-wise cross-entropy loss—a standard loss for segmentation tasks—for part attention mask segmentation with $y_{i,t}$ denoting the one-hot vector of ground truth label for a pixel at frame t, and $p_{i,t}$ the prediction.

4.2.3 Action Recognition and Ergonomic Evaluation

To complement the HDT perception model, a higher semantic level is also indispensable when dealing with abstract reasoning and robotic motion optimization with regard to prioritized human-centricity, for which two representative tasks are taken into consideration in this part: human action intention recognition and ergonomic risk evaluation.

Human action recognition can enable the proactive adaptation of robotic planning to the human operator to achieve more natural and efficient HRC. In the proposed model, an extra action branch is constructed on top of the backbone network to predict the human action intention class. This branch adopts a similar structure to the 3D pose branch with the only difference lying in the final fully connected layers after LSTM, where the extracted spatial-temporal feature is first transformed into $T \times JC_a$ with $C_a = 64$, and then regressed to $T \times N$ with $N = 5$ types of actions, which is further converted via a softmax operation into the predicted probability distribution. During training, cross-entropy loss is utilized for supervision:

$$\mathcal{L}_{action} = -\sum_{t \in T} \sum_{x \in N} y_{x,t} \log(p_{x,t}), \tag{4.5}$$

where $y_{x,t}$ denoting ground truth label and $p_{x,t}$ is the predicted probability vector.

On the other hand, automatically evaluating the ergonomic risks of the human body is of great importance for robotic actions towards reducing occupational disease in a human-centric HRC environment. In this work, we adopted the REBA (Rapid Entire Body Assessment) [200] ergonomic assessment tool to rate the musculoskeletal disorder risk of a body posture on a scale of 1-15. Although it is possible to calculate the REBA scores directly from the body joints, the non-differentiable nature of the process renders it unrealistic to be directly embedded into a gradient-based optimization model. Therefore, we alternatively resort to regressing the REBA score via an extra neural network branch, which follows a similar design as the action branch but differs at the final fully connected layer that regresses the $T \times JC_e$ feature vector, where $C_e = 64$, into $T \times 1$. SmoothL1 [201] loss function is leveraged here for supervising since it has a smoother gradient transition at 0, which, as generally believed, leads to a better regression performance. The loss function can be formulated as:

$$\mathcal{L}_t = \begin{cases} 0.5(y_t - \hat{y}_t)^2 & \text{if } |y_t - \hat{y}_t| < 1 \\ |y_t - \hat{y}_t| - 0.5 & \text{otherwise} \end{cases} \tag{4.6}$$

$$\mathcal{L}_{reba} = \sum_{t \in T} \mathcal{L}_t, \tag{4.7}$$

where y_t denotes the regressed REBA score, and $\hat{y}_t$ the ground truth label. With all the loss terms presented, the overall loss function can be defined as the weighted sum of the loss terms:

$$\mathcal{L}_{overall} = \lambda_1 \mathcal{L}_{SMPL} + \lambda_2 \mathcal{L}_{joint} + \lambda_3 \mathcal{L}_{part} \\ \lambda_4 \mathcal{L}_{action} + \lambda_5 \mathcal{L}_{reba}, \tag{4.8}$$

where λ_i represents the weight for each loss term. In this work, we empirically set $\lambda_1 = 0.2$, $\lambda_2 = 1$, $\lambda_3 = 0.02$, and $\lambda_4 = \lambda_5 = 0.1$ based on experimental trials with the aim of prioritizing the posture reconstruction related terms since they are the most challenging part for learning.

4.3 HDT-based Adaptive HRC

4.3.1 Overview

Although the main focus of this work is the vision-based HDT modelling approach as elaborated above, it is also necessary to demonstrate the deployment and applicability of the HDT. The conceptual framework of the HDT-based adaptive HRC system is illustrated in Fig. 4.2. For an HRC scene in the physical space where the human operator is working with the robot in close range, RGB-D data will be captured and processed by the proposed perception model to obtain real-time human status and synchronize with the HDT in cyberspace. The updated HDT model is capable of providing

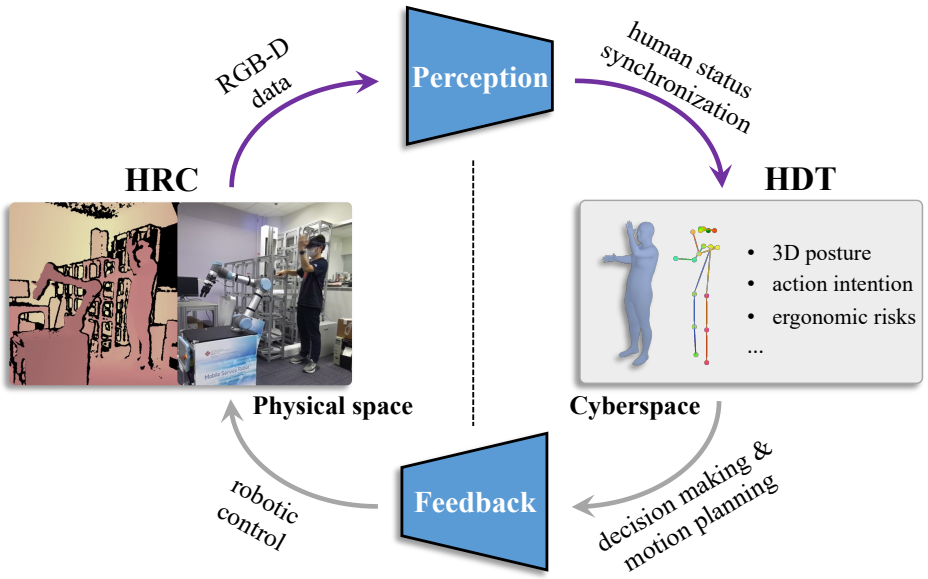


Fig. 4.2: Overview of the HDT-based adaptive HRC system

cognitive redundancy with a comprehensive understanding of the status of the human operator so that human safety and working efficiency can be optimized adaptively during robotic decision-making and motion planning. The optimized motion commands will then be delivered to the cobot via the feedback channel. Note that since this study mainly focuses on the human digital twin, other elements in the HRC environment are not included in the digital mock-up.

With the HDT perception model serving as the physical-to-cyber bridge, the feedback module operates as the backward passage from cyberspace to the physical counterpart. Unlike the normal definition of the bi-directional digital twin, in the HDT case, it is impractical to directly control the physical human state from cyberspace. Therefore, we resort to realizing the feedback module with adaptive robotic controlling based on the HDT information as an indirect way to influence human behaviour during HRC with considerations of optimized collaboration efficacy and human wellbeing.

4.3.2 Adaptive Robotic Motion Control

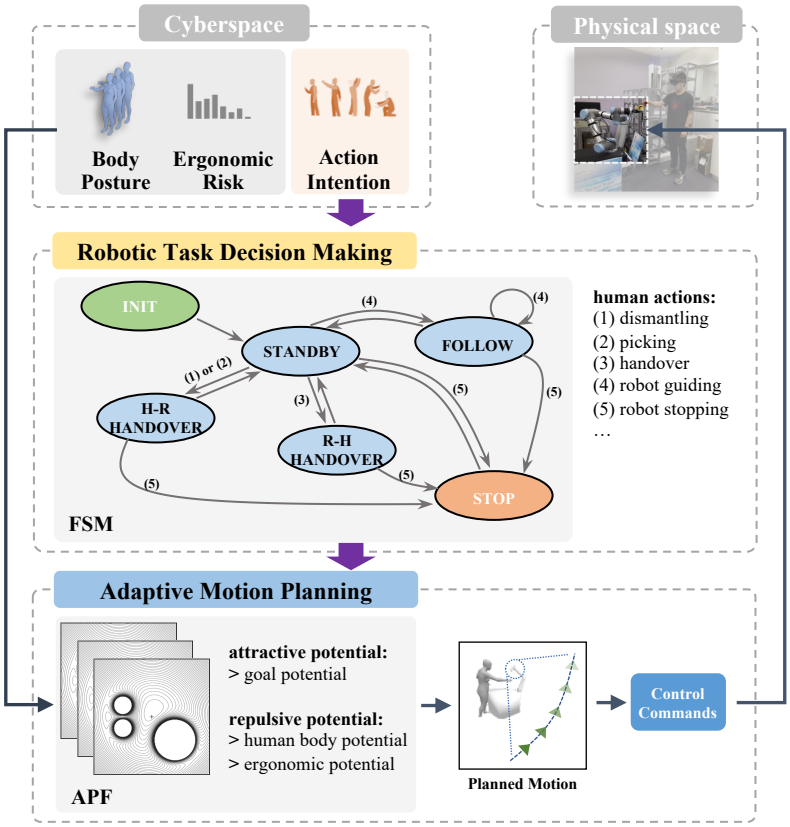


Fig. 4.3: The HDT-based adaptive robotic motion control

The HDT-based adaptive robotic motion control strategy is depicted in Fig. 4.3. First, the decision-making step is responsible for the selection of robotic action objectives, such as picking up a workpiece or handing it over to the human, along with the associated task specifications, such as the handover position. In this process, the predicted human action intention is leveraged as the main cue to decide the robot task for proactive assistance, for which a Finite State Machine (FSM) [202] is employed to map the human operator action intentions to different robotic action states as illustrated in Fig. 4.3. After booting up into the initial state, the robot will first enter the standby state, which is the default state to return to immediately after the completion of other robot actions. One-directional arrows indicate the transition condition between different states, while bidirectional arrows mean that the robot will

switch back to the previous state after successful execution without further commands. The decided current robotic task will be transmitted along with the body posture and ergonomic status into the motion planning part.

Subsequently, the adaptive motion planning step will try to generate a viable robotic motion trajectory to fulfil the designated robot task while satisfying constraints including collision avoidance and ergonomically friendliness. An Artificial Potential Fields (APF) based motion planner [203] is adopted in this step. Although the APF method is a rather simple motion planner and may not be suitable for all complex scenarios, it can be effective for certain situations such as collision avoidance, and is especially suitable to incorporate different aspects of human information as potential fields without greatly increasing the algorithm complexity. The task-associated goal position for the end-effector is encoded as an attractive potential, which can be formulated as:

$$\mathcal{U}_a = \frac{k_a}{2} \left\| \boldsymbol{q}_g - \boldsymbol{q} \right\|^2, \tag{4.9}$$

where q is the robot configuration, q_g the goal configuration, and k_a the scaling factor. The human body is naturally treated as an obstacle and represented as a repulsive potential for collision avoidance, while the interaction points with high ergonomic risk are regarded as virtual obstacles that the robot should try to avoid. The repulsive potential for obstacle i can be represented as:

$$\mathcal{U}_{r,i} = \begin{cases} 0 & d_i(\boldsymbol{q}) > c \\ \frac{k_{r,i}}{2(d_i(\boldsymbol{q})-c)^2} & 0 \leq d_i(\boldsymbol{q}) \leq c \end{cases}, \tag{4.10}$$

$$\mathcal{U}_r = \sum_i \mathcal{U}_{r,i}, \tag{4.11}$$

where $d_i(\boldsymbol{q})$ is the distance to the obstacle, c the maximum obstacle avoidance distance, and $k_{r,i}$ the weight factor for different obstacles. To apply the potentials to a robot manipulator, we choose a set of control points $p_1, ..., p_N$ with the first N-1 points representing the robot joint links and the last point for the end-effector. The end-effector will be influenced by both the attractive and repulsive potentials which result in the total potential $\mathcal{U}_t = \mathcal{U}_a + \mathcal{U}_r$, while other points will only be affected by $\mathcal{U}_r$. The combined force fields will be utilized as the reference velocities:

$$\dot{\boldsymbol{q}} = -\sum_{i=1}^{N-1} J_i^T(\boldsymbol{q}) \nabla \mathcal{U}_r\left(p_i\right) - J_N^T(\boldsymbol{q}) \nabla \mathcal{U}_t\left(p_N\right), \tag{4.12}$$

where $J_i(\boldsymbol{q})$ is the Jacobian matrix associated with p_i. Finally, the planned motion will be translated into corresponding control commands and sent to the robot controller for execution.

4.4 Experimental Results

To demonstrate the utility and capability of the proposed HDT model, experiments are conducted for a simulated HRC disassembly scenario in our lab environment, during which serval participants are designated to work alongside a UR5 robot arm and repetitively carry out several possible types of disassembly actions. An Azure Kinect RGB-D camera is deployed to capture the HRC scene data, which will be sent to a GPU server to establish and update the HDT, based on which adaptive robotic motion will then be planned and performed to provide assistance to the human operator. In the following experiments, the proposed perception model for HDT modelling will first be evaluated and compared with some baseline approaches, and

then the HDT-based adaptive robotic control strategy will be demonstrated in a simulation environment.

4.4.1 HDT Modelling for HRC Disassembly Scenario

1) Data Collection and Experimental Settings

For the evaluation of the HDT perception model performance, RGB-D data of the disassembly scenario were collected via the Azure Kinect camera and trimmed into fix-length (T=16) clips, each of which contains one of 5 types of human action including 1) dismantling, 2) part picking, 3) robot handover, 4) robot guiding, and 5) robot stopping. We chose these classes because they are the most typical and representative human actions in an HRC disassembly scenario. After removing invalid or low-quality data, there were 939 clips as the overall experiment dataset, which was further split into a training set with 751 clips and a testing set with 188. For data annotation, the action intention ground truth labels were first manually annotated, then the human pose labels were generated by iteratively fitting the SMPL model to a given human image [204]. Ergonomic risk labels were calculated according to the REBA ergonomic assessment steps while ignoring joint load since we do not have force data.

The HDT perception model is programmed using the prevailing PyTorch deep-learning library with an Nvidia RTX3090 GPU leveraged for hardware acceleration. Adam optimizer is employed with learning rate 5×10^{-5}, batch size 2, and 100 training epochs.

2) Evaluation Results

Table 4.1: Evaluation results of the HDT perception model

Method	MPJPE (mm)	Action Intention Accuracy	Ergonomic Score Mean Error
HMR [198]	67.19	-	-
PARE [199]	55.04	-	-
ST-GCN [205]	-	94.24%	-
MTL [206]	-	95.15%	0.83
Ours	**48.98**	**98.54%**	**0.66**

To quantitatively evaluate the performance of the HDT perception model, MPJPE (Mean Per Joint Position Error) is leveraged for the measurement of the reconstructed body pose error, accuracy for action recognition, and mean error for ergonomic risk score regression. After training the proposed model and compared models on the training dataset, evaluation was conducted on the testing set and the mentioned metrics were calculated and reported, which are shown in Table 4.1. Since there is no previous work that simultaneously addresses these three aspects, we chose several baseline methods regarding different tasks. For pose estimation, HMR (Human Mesh Recovery) [198] mainly leverages a ResNet-50 model to estimate human pose parameters, while PARE (Part Attention REgressor) [199] considers the part attention similar to ours but does not involve temporal information. ST-GCN [205] employs GCN (Graph Convolutional Network) to recognize human action directly from skeleton joint data, and MTL (Multi-Task Learning) [206] extends the GCN to additionally regress REBA ergonomic scores. Compared to the baseline methods, the proposed model achieves better performance on all three recognition tasks. In terms of human pose reconstruction, we believe it owes to the enhanced feature extraction ability of the backbone

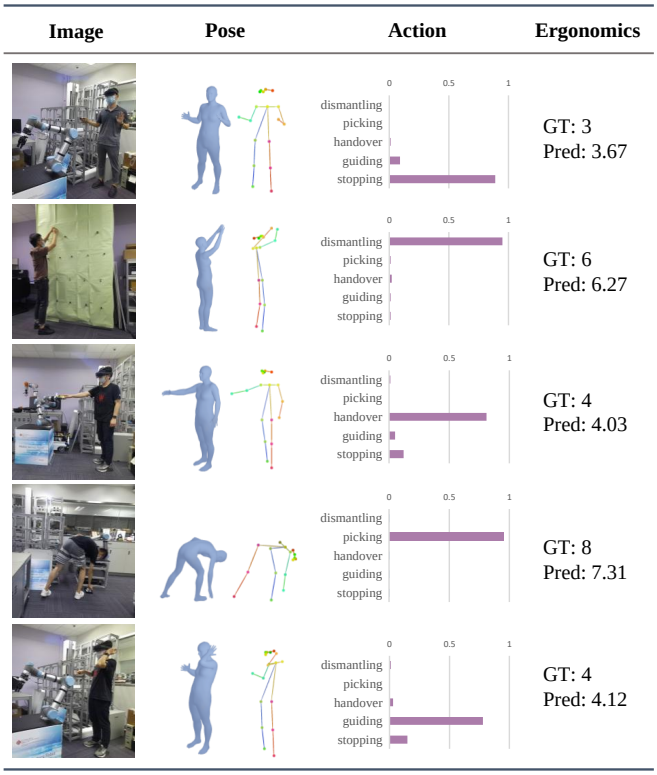


Fig. 4.4: Qualitative examples of the HDT perception model

HRNet, the more focused attentional guidance provided by the body part attention mechanism, and the temporal coherence information. As for action and ergonomic score prediction, the sole dependency on skeleton data of the baseline model weighs against its ability to exploit other heuristic features during human-robot interactions, while the proposed model can better leverage the overall visual observations and thus achieves better accuracy. Fig. 4.4 presents several example results from the HDT model to qualitatively illustrate in a visualized manner in order to provide an intuitive impression of its performance. Note that we just randomly chose one sample for each human action category, so there are no orderly or sequential relations between these examples. But one can still easily compose or imagine a possible sequence of tasks with these actions: the human operator first 1) dismantles some parts from the workpiece, then uses gestures to 4) guide the robot to come closer, and signs the robot again to 5) make it stop when close enough;

subsequently, the human 2) picks up a part and 3) handovers to the robot, then 4) guides it again to move away. And further actions can be added to the model to adapt it to more types of HRC tasks.

4.4.2 HDT-based Adaptive Cobot Motion Control

To demonstrate the HDT-based adaptive robot control for HRC scenarios, an experiment is further conducted in a simulation environment as illustrated in Fig. 4.5, in which a human operator is working in close range with a UR5 robot arm to fulfil some manipulation tasks. To simplify the repetitive experiments, videos of a human operator carrying out different actions in the real world were recorded offline, from which the HDT was established following the proposed modelling approach. Frame-wise human data were extracted from the HDT and imported into the PyBullet simulation environment, while a UR5 was positioned alongside the human operator attempting to reach some predefined goal points based on different human actions. The human sequence was replayed over again for each simulation episode.

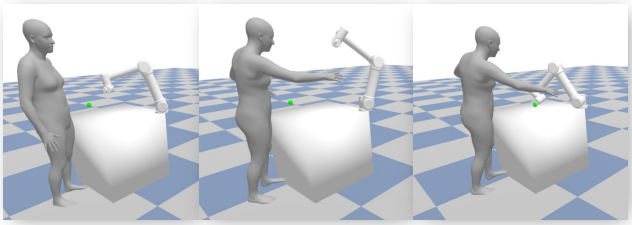


Fig. 4.5: Illustration of the adaptive cobot motion control

Since the human body is constantly moving around, the objective of the cobot motion planning would be to reach the target point with adaptive collision avoidance with the human body while avoiding high ergonomic risk interaction points. We first iterated through human sequence data to identify frames with predicted REBA scores higher than 8 (high risk according to

the REBA assessment) and stored the corresponding human hand positions in these frames as high-risk points. Then the APF-based motion planner as described in Section 4.3.2 was applied to plan and control the UR5 robot. We ran the simulation for 1000 rounds and the robot was able to reach the designated targets without collision for 959 times. Although this experiment seems rather naive, it depicts the fundamental application of the HDT model in HRC scenes and exhibits its effectiveness.

4.4.3 Discussions

Based on the proposed HDT perception model, comparative experiments were first conducted on the collected HRC data for the evaluation of perception performance, which evidently shows better accuracy in all three tasks over the baseline models. This is mainly attributed to the end-to-end structure that integrates the three functionalities into a unified neural network, which is beneficial for exploiting and reusing more robust visual feature representations with extra support from the enhanced backbone model and part attention mechanism. Then a simplified case of the HDT-based cobot motion control was demonstrated in a simulated close-range HRC scenario, and a success rate of 95.9% was obtained from the trials. This simple glimpse can already reflect the huge potential of the HDT model in futuristic human-centric manufacturing systems, especially in cases where fine-grained human information is demanded. Nevertheless, some issues such as limited data sources, limited perception tasks, and lack of more complex case studies still exist and should be investigated more closely in future endeavours.

4.5 Chapter Summary

One of the trends in the next generation of industry is to shift away from a purely profit-driven manufacturing approach towards a more human-centric one. HRC is seen as a natural choice to achieve this, as it has the potential to unleash the combined strength of humans and robots. To equip the robot with a comprehensive understanding of its human partner beyond standalone recognition tasks, this chapter proposes a vision-based HDT modelling approach that addresses multiple human perception aspects with a unified deep learning model in an end-to-end manner. The main contributions are summarized as follows: 1) a specifically designed deep learning architecture was proposed to simultaneously perceive human 3D posture, action intention, and ergonomic risk to accomplish the HDT modelling; 2) an adaptive robotic motion control strategy based on the proposed HDT model was presented to demonstrate the fundamental application of the HDT model in HRC scenarios.

5

Multi-Granularity Workspace Parsing

With object-level and human-level information obtained, robots could already perform collaborative actions in some relatively simple tasks such as tool or workpiece delivery in a fixed workstation. Nevertheless, to deal with more complex tasks such as navigating to places out of sight to fetch a specific object required in an HRC assembly process, robots should be equipped with the skill to perceive and model the whole working environment more comprehensively. Existing robotic systems normally adopt a single-granularity semantic segmentation scheme for environment perception, which lacks the flexibility to be implemented in various HRC situations. To fill the gap, this chapter presents a multi-granularity scene segmentation network. Inspired by modern network designs, we construct an encoder network with two ConvNext-T backbones for RGB and depth respectively, and a decoder network consisting of multi-scale supervision and multi-granularity segmentation branches.

The research content of this chapter is based on a conference paper presented at the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) [207].

5.1 Introduction

The modern trend of mass personalization in the manufacturing industry has incited tremendous interest and inclination to adopt human-robot collaboration in the manufacturing shop floor for the complementary strength of human and robot teams, and the flexibility to swiftly adapt to diverse individualized production demands [2].

To achieve autonomous navigation and adaptive collaboration in a shared space with human operators, the robot should be equipped with an advanced cognition system that can constantly perceive the surrounding environment. Earlier robotic systems mainly rely on raw sensory data such as force and depth values to construct the robotic perception and controlling strategy [208], while a recent trend is to empower the robotic perception skills by incorporating semantics into the scene perception process [209].

A commonly used technology is semantic segmentation, which leverages the visual observation of the environment as input and segments it into different semantic regions at a pixel level. A large body of work has been devoted to the semantic segmentation task in the computer vision community over the last decades, ranging from image processing-based methods to the recently prevailing deep learning-based approaches [210]. However, most existing works only consider the scene segmentation task in a single granularity, which adopts a uniform criterion for the segmentation of all scenes regardless of the perception distance and intended objectives. This single-granularity scheme cannot suffice for volatile situations that a robot often encounters in a human-robot environment. For example, a human body can be segmented as a whole for collision avoidance when the robot is navigating at a distance,

but in a close-range co-assembly case, the hands and arms of a human should be distinguished for delicate robotic interactions with human hands.

Therefore, this work aims to extend the scene segmentation with a multi-granularity task formulation including three levels of granularity: area level, entity level, and part level. With this multi-granularity scene representation, the cobot can adaptively alternate its attention among different granularity levels according to its current situation and further analyze the environmental information to facilitate subsequent decision-making and motion planning.

In this work, an RGB-D camera is utilized to capture the HRC scene since it is more affordable and delivers reasonable performance in indoor scenes. A multi-granularity scene segmentation network is proposed, which takes the form of the prevalent encoder-decoder structure. The encoder network, which is developed based on the ConvNext backbone [211], fuses the RGB and depth information into a unified feature representation, upon which the decoder network is constructed leveraging multi-level refinement and multi-task strategies that can simultaneously produce multi-granularity segmentation results. A simulated case in a human-robot collaborative battery module disassembly scenario is studied to demonstrate the effectiveness of the proposed model, and comparative experiments are carried out on a public dataset NYU-Depth V2 [212] to illustrate the generality.

5.1.1 Environment Perception in HRC

As a prerequisite of human-robot collaborative systems, vision-based environment perception has been extensively investigated in the literature [191]. Some works directly employ raw sensory data such as RGB-D camera or

LiDAR to ensure safety for human-robot teams. Liu et al. [98] presented a collision avoidance strategy which constructed an OctMap from the depth sensory data of the environment to enable collision-free human-robot collaboration. A point cloud-based scene perception approach was leveraged by Choi et al. [134] to synchronize the physical environment status with a digital twin model of the human-robot collaborative workspace to measure the safety distance in the virtual realm.

Another trend is to leverage semantics in the environment perception process. Butler et al. [213] reported an interactive scene segmentation scheme that additionally introduces human aids into the robotic scene perception process to increase object segmentation performance. To achieve safe robotic navigation with natural language instructions in a complex indoor environment, Hu et al. [127] leveraged a 2D map generated by SLAM to represent the global environment and adopted the Mask R-CNN model to realise instance-level scene segmentation for local robotic observations. The limitation of these works is only a single level of semantics is considered, which is not flexible enough to adapt to multi-granularity HRC activities.

5.1.2 RGB-D Semantic Segmentation

RGB-D cameras are widely used in robotic perception systems as they can provide 3D environmental information while being more affordable than LiDAR. Many efforts have been devoted to investigating the RGB-D information-based semantic segmentation task. One line of works attempts to project the RGB-D data into 3D space and carry out semantic segmentation based on 3D point cloud or voxel data [214, 215]. However, a major issue is current neural networks are generally inefficient when processing 3D volumetric data,

which renders it impractical to deploy in time-sensitive robotic applications. A more feasible way is processing the RGB and depth images separately and then fusing the features together [216, 217, 218]. Chen et al. [218] proposed the SA-Gate model which adopted an intertwined fusion strategy between the RGB and depth encoder branches. Seichter et al. [217] focused more on the efficient design of the segmentation encoder and decoder networks to facilitate robotic applications. Nevertheless, these works still mainly rely on the ResNet backbone for feature extraction, which has already been outperformed on many vision tasks by some modern networks such as Swin Transformer [219] and ConvNext [211]. Another line of work that is closely related to this work is the multi-task semantic segmentation [220, 221, 222]. Although various pixel-level prediction tasks such as edge map, surface normal, and object part segmentation have been considered in these works, it has been rarely mentioned to implement semantic segmentation in a multi-granularity manner, which we believe has great potential in HRC environments.

5.2 Multi-Granularity Segmentation Network for HRC Scenes

The volatile nature of robot tasks in an HRC environment renders a single-granularity scene perception scheme rather fragile due to the lack of flexibility and versatility. Thus, we employ a multi-granularity scene segmentation model to enhance the cognitive capability of the robot. In this section, the multi-granularity segmentation criterion for the demonstrative HRC scenario

will first be disclosed, then the proposed multi-granularity scene segmentation network architecture will be described in detail.

5.2.1 Multi-Granularity Segmentation Criterion

A typical semantic segmentation task normally adopts a segmentation criterion solely based on the entity level such as a robot or a person, and remains unchanged for images taken from different perspectives and distances as in some renowned public datasets [223, 212]. To enhance the perception flexibility of collaborative robots, we expand the segmentation criterion for the HRC disassembly case into three levels, i.e., area level, entity level, and part level, the detailed definition of which is depicted in Fig. 5.1.

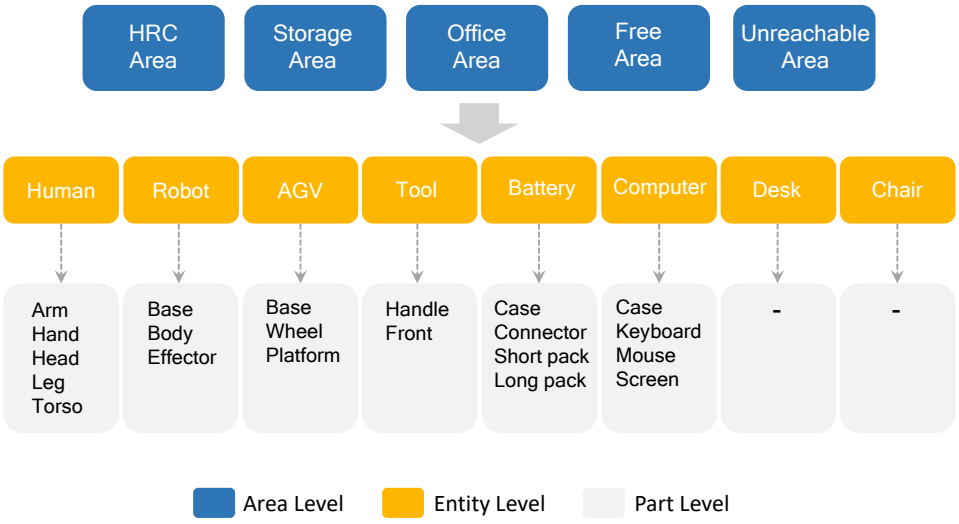


Fig. 5.1: Multi-granularity segmentation criterion.

The area level is employed to handle coarse-grain robotic tasks such as navigating to a specific work area. Here we heuristically define 5 types of areas in an HRC scenario.

The entity level in this work is aligned with the traditional semantic segmentation criterion which segments different entities based on their semantic

categories. This is reserved for general perception purposes of autonomous robots.

The part level, on the other hand, is defined in a finer grain, which splits an entity into its constituent parts based on different functionalities and possibilities to interact with robotic end effectors. This level can be beneficial to some delicate tasks such as human-robot co-assembly, where the recognition of detailed parts such as product components and human hands is a prerequisite.

5.2.2 Model Architecture

The overall architecture of the proposed multi-granularity segmentation model (MGS-Net) is depicted in Fig. 5.2. The general model design follows the encoder-decoder structure. The encoder consists of two branches of networks, which are adopted from the ConvNext model [211], for RGB and depth information, respectively. The decoder part is inspired by the ESANet (Efficient Scene Analysis Network) [217] following a lightweight design, and we further extend it with multiscale refinement and multi-granularity segmentation designs.

1) RGB-D Encoder

The purpose of the encoder network is to extract RGB and depth features and aggregate them at different stages so that the complementary information in the RGB and depth maps can be better exploited. The RGB and depth branches follow the same philosophy of ConvNext [211], the main difference is the incorporation of the Fusion Module.

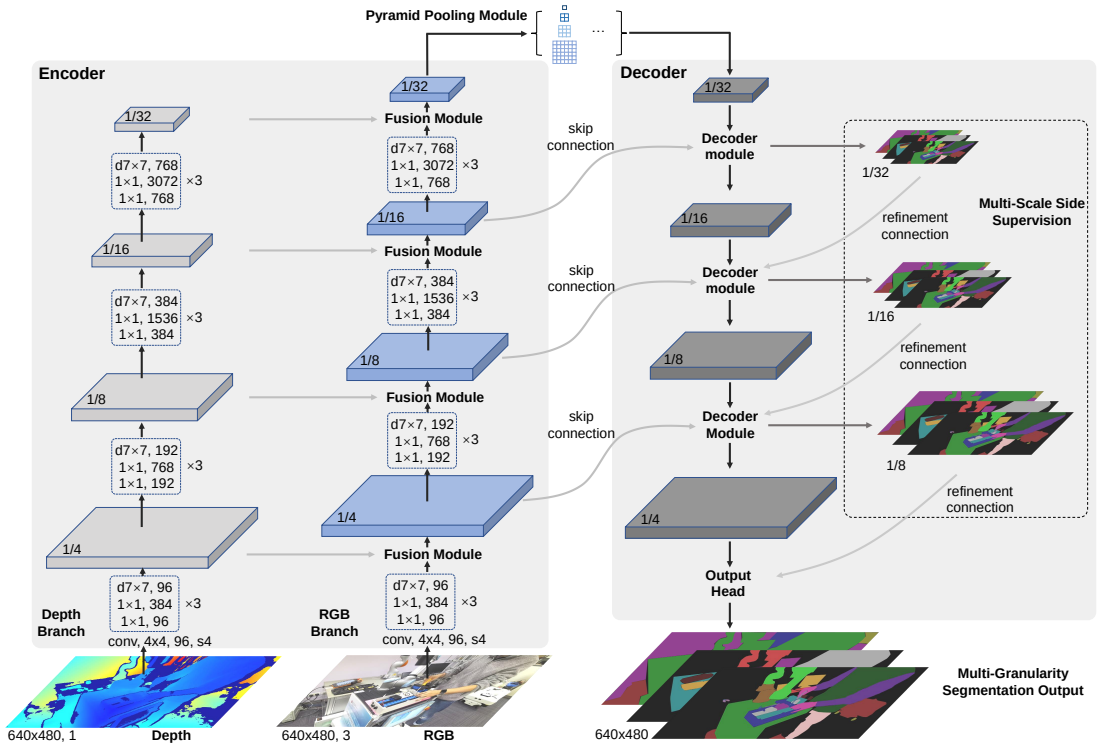
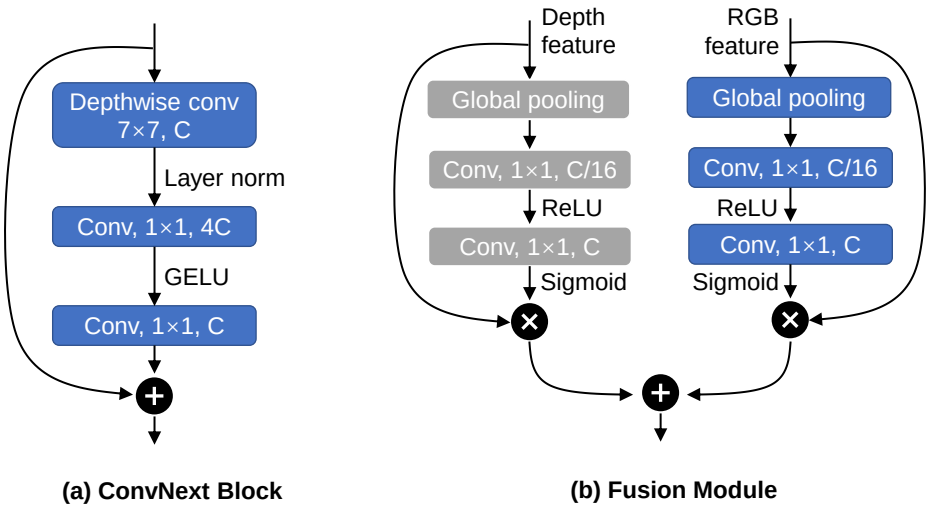


Fig. 5.2: The structure of the proposed multi-granularity segmentation network (MGS-Net).

The ConvNext is a modernized version of ResNet with some micro and macro design principles borrowed from Visual Transformer models. With a similar computational cost as ResNet, ConvNext performs better by a considerable margin. In this work, we employ the ConvNext-T variant for simplicity. Fig. 5.3(a) shows the details of a ConvNext block, of which the major difference is the introduction of the depthwise convolution layer, Layer Normalization, and GELU (Gaussian Error Linear Units) activation function.

At each stage of the backbone network, we fuse depth features into the RGB branch via the Fusion Module, which leverages the channel attention module proposed in [224] to enable an adaptive and learnable fusion mechanism. The detailed structure is illustrated in Fig. 5.3(b).



(a) ConvNext Block **(b) Fusion Module**

Fig. 5.3: The structure of ConvNext block and Fusion Module.

After the feature extraction of the encoder part, we additionally employ the Pyramid Pooling Module [225] to process the features with different pooling scales, which is believed to be able to aggregate global and local context information and has been proved to be beneficial for segmentation performance by previous research works.

2) Multi-Granularity Decoder

The decoder network mainly consists of three consecutive Decoder Modules, which gradually decode and enlarge the feature maps, and an Output Head, which recovers the feature map scale to match the input image size and output the final segmentation results.

The Decoder Module is depicted in Fig. 5.4(a), in which the first part contains a 3×3 convolution layer and a factorized Residual Block [226] composed of several 3×1 and 1×3 convolutions, which is meant for better computational efficiency. Then the flow branches into a main path and a side output path. The main path consists of a nearest upsampling operation and a depthwise

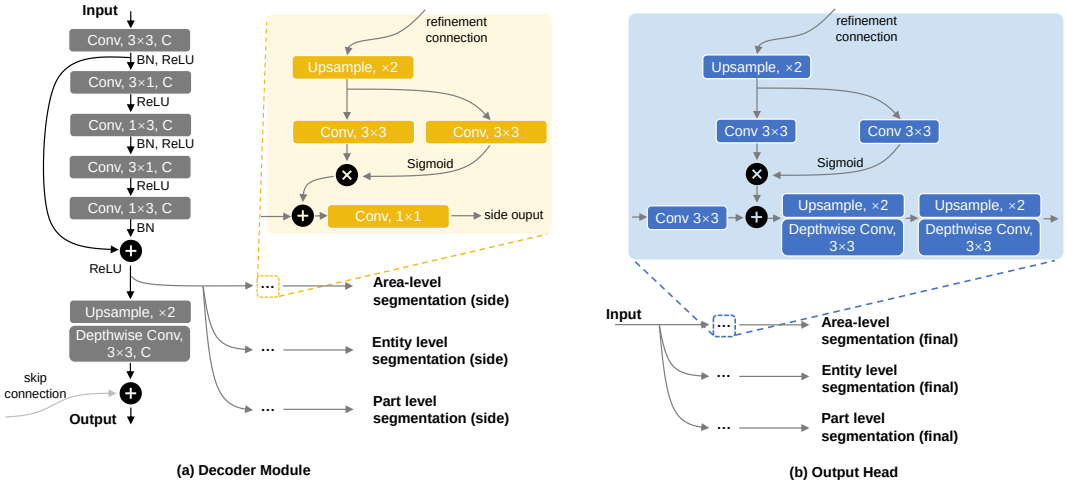


Fig. 5.4: The Decoder Module and Output Head of the decoder network.

convolution layer, and the skip connection from the encoder is additionally attached to the output feature map. On the other hand, the side output path is intended to provide multiscale segmentation supervision by producing side segmentation results on different decoder stages. For each stage, the side path will output three levels of segmentation results simultaneously with the same network structure, which mainly consists of a nearest upsampling, a simplified spatial attention module, and an output 1×1 convolution. The upsampling and spatial attention are utilized to incorporate the smaller-scale segmentation results to produce refined segmentation. Note that for the first Decoder Module, there is no refinement connection since it is already at the smallest scale. The detail is shown in yellow in the figure.

The final stage of the decoder is the Output Head which will recover the feature map to the input size so that a pixel-wise segmentation result can be generated. The main feature map is first processed by a 3×3 convolution, then the previous side segmentation result is merged into the main feature in a similar way as in the Decoder Module. After two consecutive upsampling stages, the final segmentation results are generated. The three branches for multi-granularity segmentation all follow similar network structures with the

only difference lying in the final output channels according to the categories of different granularity levels defined in Section 5.2.1.

For the segmentation supervision, we adopt the Cross-Entropy loss with additional weight terms for each class based on the number of pixels present in the ground truth segmentation map. The weighted Cross Entropy loss can be formulated as:

$$\mathcal{L}_{WCE} = -\sum_{i \in \mathcal{C}} w_i \cdot y_i \log(p_i), \tag{5.1}$$

where $\mathcal{C}$ is the number of classes, w_i the weight of the class, y_i the one-hot vector of ground truth label, and p_i the prediction. We apply this loss function to all the segmentation supervisions of different scales and different granularities, including 4 scales $\times$ 3 granularities. The overall loss function is simply the summation of these loss terms.

5.3 Experimental Results

In this section, we first carry out some experiments on an HRC battery module disassembly environment. Human-robot collaborative disassembly is regarded as a viable solution to address the increasing labour demand and safety issues of the disassembly and recycling of end-of-life Li-ion batteries. Current cobots normally adopt a single-granularity perception system, which cannot suffice for the flexibility demand of HRCD environments. Thus, we simulated this collaborative battery disassembly case to demonstrate the effectiveness of the proposed multi-granularity segmentation model.

To illustrate the generality and universality of the proposed model, further experiments on the publicly available NYUv2 dataset are also presented

by means of comparative studies with previous state-of-the-art models and ablation studies for evaluating different components of the network.

5.3.1 Implementation Details

The proposed model is implemented via Pytorch and accelerated by a Nvidia RTX3080Ti GPU. The backbone part of the encoder is initialized with the pretrained weights provided by [211]. Other layers are randomly initialized. AdamW optimizer with initial learning rate 1e-4 is leveraged along with the cosine annealing scheduler with warm restart, where $T_0 = 5$, $T_{mult} = 3$. The model is trained for 500 epochs with batch size 4. Other common training techniques such as data augmentation including random resizing, crop, and flipping are also adopted following the common practice.

5.3.2 Human-Robot Collaborative Disassembly Case

The first part of the experiments is conducted in a simulated Li-ion battery disassembly scenario, where a human operator is carrying out disassembly operations with assistance from a robot collaborator. The data collection setup and experimental results are shown in the following content.

1) Data Collection

While the human-robot team is working on the disassembly task, an RGB-D camera is placed at different view angles and different distances around

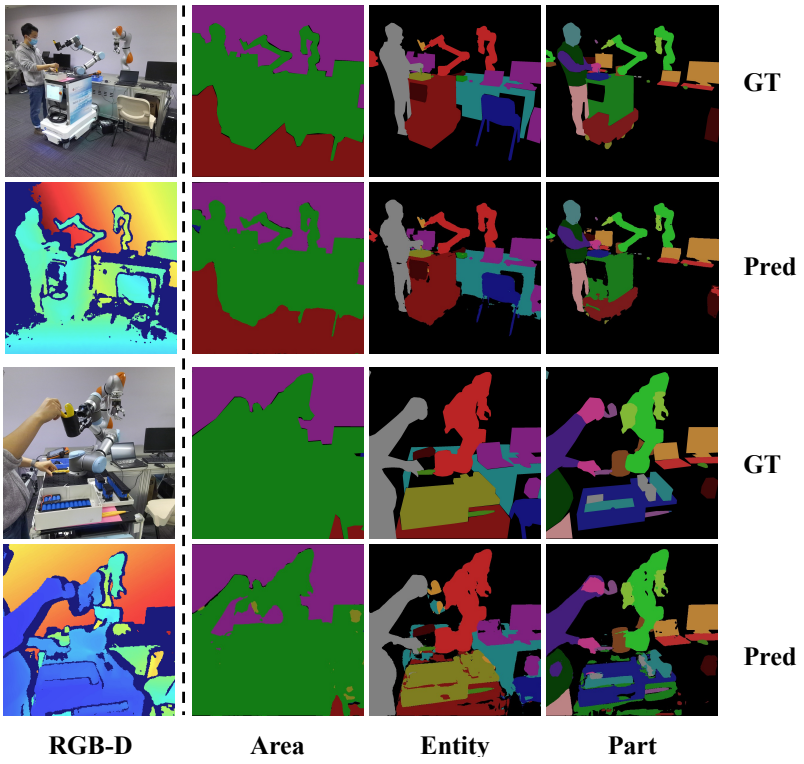


Fig. 5.5: Qualitative results in the HRCD case.

the working area to simulate the possible views of a mobile cobot. Several hundreds of images were captured, but we only managed to annotate 40 of them due to limited manpower and the non-trivial nature of manual annotation of the three levels of segmentation labels. But we believe this quantity of samples should suffice for the demonstration of the proposed model since only a single HRCD scene is considered.

2) Results

We split the annotated samples into two subsets with 32 samples for training and 8 for testing. After the training of the multi-granularity segmentation model, we evaluate the model performance based on the commonly used mIoU (mean Intersection-over-Union) metric, which basically measures the overlap between the predicted segmentation result and the ground truth, to show how well the model performs on the testing data. The proposed

model achieves 84.90, 75.07, and 69.47 in terms of mIoU on the three levels (Area, Entity, and Part) of granularity proposed in this work, and some qualitative results are illustrated in Fig. 5.5, in which the images are cropped to square for the convenience of illustration. Although the produced multi-granularity segmentation result seems reasonable in general, some noise and errors still widely exist in boundary areas and small structures, which we believe are mainly because of the limited dataset size. As a part of a robotic system, the inference speed of the model is also an imperative concern. The Pytorch implementation of the proposed model can achieve 62 FPS with GPU acceleration, which can satisfy the laboratory demonstration purpose. For more time-sensitive robotic applications, more techniques such as pruning, quantization, etc., can be further exploited in future research works. Here we do not compare with other methods since we believe the dataset size is too small to make a fair comparison, which is also the reason we decided to additionally employ the public dataset for comparative evaluation.

5.3.3 Experiments on the NYU-Depth V2 Dataset

The NYU-Depth V2 dataset [212] is a frequently utilized benchmark for evaluating RGB-D semantic segmentation algorithms. The dataset contains 1,449 indoor scene RGB-D samples with pixel-level semantic labelling, which, according to the proposed multi-granularity criterion, is mainly defined on an entity level with 40 classes of entities. We follow the original split of the dataset, which includes 795 samples for training and 654 for testing. As the dataset only contains single-granularity semantic annotations, we made some adjustments to the proposed model by retargeting the area level and part level branches to predict edge maps and normal maps instead as in [222].

These two tasks serve as extra supervision to facilitate the model training but are not evaluated as they are not the main focus of this work.

1) Comparative Results

The evaluation results of the proposed MGS-Net compared with some recent state-of-the-art methods are listed in Table 5.1. We adopt three common metrics, including PixAcc (pixel accuracy), mAcc (mean accuracy), and mIoU, which are widely used in previous works. Based on these metrics, we can see that the proposed model is on par with the top performers. One reason of this achievement is the ConvNext backbone. With only a slightly higher computational cost than ResNet50, the ConvNext backbone shows better feature extraction ability than some bulkier models such as ResNet-101 and ResNet-152. The adoption of other modern network designs, such as multi-scale supervision, multi-level refinement connection, multi-task prediction, etc., also largely contribute to the performance, which will be depicted and discussed in more detail in the ablation study part.

Table 5.1: Experimental results on NYUv2 dataset compared with state-of-the-art methods.

Method	Backbone	PixAcc	mAcc	mIoU
RefineNet [227]	ResNet-152	74.40	59.60	47.60
MTI-Net [222]	HRNet48-V2	75.30	62.90	49.00
PADNet [221]	ResNet-50	75.20	62.30	50.20
ESANet [217]	ResNet-50	-	-	50.53
Zig-Zag [228]	ResNet-152	77.00	64.00	51.20
ShapeConv [216]	ResNext-101	76.40	63.50	51.30
SA-Gate [218]	ResNet-101	**77.90**	-	52.40
MGS-Net (ours)	ConvNext-T	77.41	**66.45**	**52.86**

2) Ablation Study

Table 5.2 lists the results of the ablation study, during which we mainly consider 5 components: backbone, optimizer, multi-scale supervision (MS), multi-task prediction (MT), and refinement connection (Refine). It can be identified that each component has contributed to the overall performance improvement. While the ConvNext backbone contributes to the improvement as expected, the AdamW optimizer with cosine annealing scheduler also makes a significant impact on the performance by enabling a faster convergence and smoother training process. The multi-scale supervision provides a substantial improvement, which we believe is because the extra constraints force the model to learn more from multi-scale features. The refinement connection is actually a part of the multi-scale supervision strategy which serves as a bridge between different scales of side supervision in a refinement manner. The multi-task prediction, or multi-granularity in the original design, presents moderate improvement by providing extra supervision signals from two extra tasks. In general, the experiments on the public dataset have clearly demonstrated the effectiveness and generality of the MGS-Net.

Table 5.2: Ablation study of the model components on NYUv2 dataset.

Backbone	Optimizer	MS	Refine	MT	mIoU
ResNet50	Adam				47.74
ResNet50	Adam	✓			49.27
ConvNext-T	Adam	✓			50.27
ConvNext-T	AdamW+CosAnneal	✓			51.98
ConvNext-T	AdamW+CosAnneal	✓	✓		52.01
ConvNext-T	AdamW+CosAnneal	✓	✓	✓	**52.86**

5.3.4 Discussions

This chapter introduces a novel multi-granularity scene segmentation network designed to enhance environment perception in HRC systems. Tested in a simulated battery disassembly scenario, the network demonstrated superior

segmentation performance across different granularity levels: area, entity, and part. This adaptability is crucial for dynamic HRC environments. The network's architecture, integrating ConvNext backbones with multi-level refinement and multi-task strategies, proved effective, achieving high segmentation accuracies with mIoU scores of 84.90%, 75.07%, and 69.47% across the respective granularities. These results highlight its capability to provide detailed and contextually relevant segmentation results essential for effective HRC. However, the study faces limitations, particularly the small size of the training dataset and the need for computational optimization for real-time applications. Future work could focus on expanding the dataset, enhancing processing speeds through advanced neural network techniques, and incorporating richer sensory inputs to improve the model's generalizability and efficiency in diverse industrial environments.

5.4 Chapter Summary

Motivated by the lack of a flexible environment perception scheme in current HRC systems, we proposed a multi-granularity scene segmentation model in this work, aiming to simultaneously segment the environment into different granularity of semantics to accommodate the constantly changing needs during cobot operations. By incorporating a bunch of modern network design strategies, the proposed MGS-Net has achieved prominent results in the collaborative battery disassembly case and demonstrated comparative performance with state-of-the-art methods on the NYUv2 dataset. The main contributions of this chapter can be summarized as follows: 1) an RGB-D segmentation network MGS-Net was proposed leveraging modern network designs such as the ConvNext backbone, multi-scale supervision,

multi-granularity prediction, et cetera, 2) the multi-granularity segmentation criterion was defined in an HRCD scenario and the feasibility of the proposed model was demonstrated based on this criterion, and 3) the model was further evaluated on the NYUv2 dataset and achieved comparable results to state-of-the-art methods.

6

Vision and Language-Based Collaborative Reasoning

The perception of object, human and environment could provide a holistic understanding of an HRC working scene. To bridge the gap between scene understanding and proactive decision-making, a reasoning mechanism is necessary for robots when collaborating with human operators. The ultimate goal of vision-based scene understanding is for the robot to proactively reason and decide what to do next based on the holistic scene information. In the smart manufacturing context, the robot collaborator in an HRC team should be able to autonomously perceive the ongoing production status and flexibly adapt to different operations without explicit pre-programming. This ability of abstract reasoning is the missing puzzle piece towards holistic scene understanding. In this chapter, we will delve deeper into a vision and language-based reasoning approach and explore the potential of integrating vision-language cues and Large Language Models into the robotic reasoning functionality, and further refine and smooth the reasoning process by including human guidance in the loop.

The research work of this chapter is extracted from a journal manuscript submitted to the Journal of Manufacturing Systems [229].

6.1 Introduction

The emerging paradigm of HRC in recent years has been regarded as the most promising avenue to achieve mass personalization [1] and can drastically reshape the manufacturing landscape by combining the best of both human and robotic merits: the creativity and critical thinking of human operators, coupled with the precision and efficiency offered by robotic counterparts [2, 230]. While this synergy has the potential to deliver unparalleled productivity, the seamless and flexible collaboration between humans and robots continues to pose significant challenges, especially in terms of effective communication and task understanding [231, 232].

Existing HRC approaches mainly rely on visual perception to autonomously recognize the collaborative environment [191, 233, 234] since vision data contain rich semantic information that enables robots to detect and interact with humans and objects in their vicinity. However, sole vision-based methods often struggle to fully comprehend the inherent ambiguity that prevalently exists in human-robot communication, such as unclear human gestures, leading to misinterpretations and subsequent compromises of efficiency [6, 235]. This naturally leads to an exploration of seeking complementary modalities of data to enhance the communication channel.

Human language, as a more precise representation of human intentions, has attracted abundant research interest in the HRC field, especially in guiding robots based on varied language commands [236, 237]. Nevertheless, many of these works normally oversimplify the task formulation by classifying the language instructions into predefined categories and preprogramming robots to respond in predetermined ways, which can considerably restrict

the flexibility and extendability of possible input language commands, and more severely, prevent the robot from understanding the nuances embedded in human language, thus cannot effectively cope with the ambiguity during human-robot communication.

Recently, the remarkable advancements in large language models have obtained enormous attention in the research community because of their capability of understanding and generating human-like text [238, 239], providing a potential avenue to complement visual cues and alleviate communication ambiguity in HRC. However, while these LLMs have been applied and tested in daily contexts, their potential within the HRC manufacturing scenarios, especially how to associate and cooperate with visual perception systems, remains largely underreported. It is, therefore, necessary to explore how to leverage the unprecedented linguistic capability of LLMs and integrate them into the HRC perception and reasoning system to reduce uncertainty and improve HRC efficiency.

In response to these issues, this chapter proposes a vision-language reasoning approach for ambiguity mitigation in human-robot collaborative manufacturing scenarios. A novel referred object retrieval model is first proposed, which aims at finding the goal object in visual observations based on the specifications of paired language commands. Meanwhile, a human-guided refinement strategy is introduced to refine the referred object retrieval performance by requesting human operators to click on the image. The retrieval model is expected to reduce the object-reference ambiguity during HRC, enabling the robot to accurately understand the object being referred to by the human operator. Then an LLM-based robotic action planner is designed to generate

feasible robot action sequences with reference to the language commands as well as the retrieved object locations.

6.2 Vision-Language Reasoning for Ambiguity Mitigation

As mentioned above, pure vision-based clues may exhibit a certain level of vagueness when it comes to reasoning about the human's intended robotic action due to inherent visual ambiguity. Therefore, to alleviate the communication ambiguity, the human language modality is introduced to provide explicit instructions about robotic action targets.

The overview of the proposed approach is demonstrated in Figure 6.1. Specifically, the target reference expression will first be extracted from the full language command by an LLM, the implication cue of which will be leveraged by the proposed referred object retrieval model to locate the target object in the image observation for the robot to interact with.

While in most cases this should suffice to deliver reasonable object geometric locations, the model may struggle for some samples because of the probabilistic and blackbox nature of neural network models. Thus a backup plan will also be proposed to remedy low-confidence segmentation results by asking human operators for additional clicking input to provide further location information.

The extracted object location will then be fed along with the language command into the LLM-based robot planning module to generate corresponding robotic action sequences in the form of executable codes that evoke APIs (Application Programming Interface) for primitive robotic skills such as moving to a certain coordinate or opening the gripper. The vast world knowledge and common sense reasoning ability embedded in the LLM model will fuel the planner to creatively combine primitive robotic skills and yield feasible solutions to address the designated tasks.

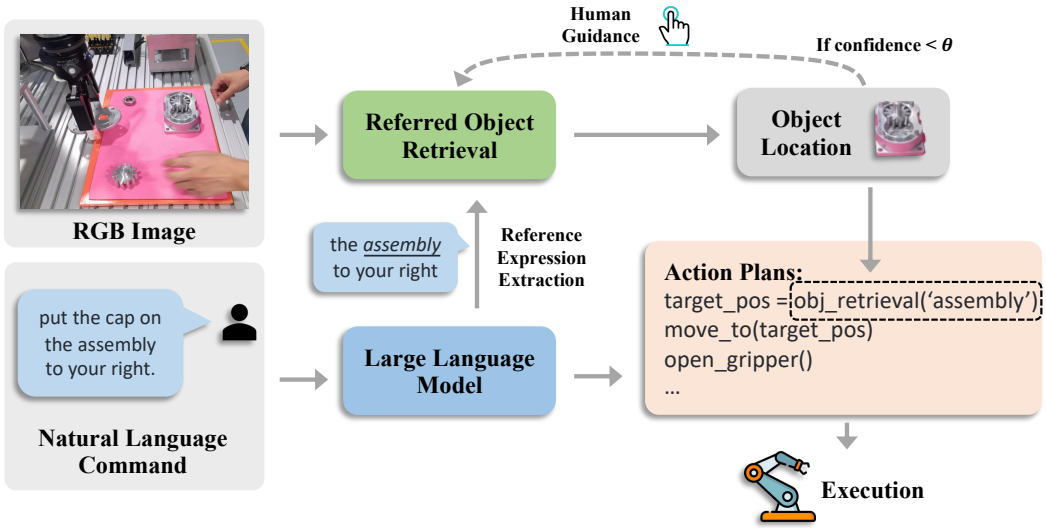


Fig. 6.1: Overview of the proposed vision-language reasoning approach for HRC

6.2.1 Ambiguity-aware Referred Object Retrieval

The identification and localization of the target object is the most important component of the overall vision-language reasoning approach as the human commands are designed to center around an assembly tool or part in the HRC scene. Given an image $\mathcal{I}$ of the HRC scene and a reference expression $\mathcal{E}$ in the natural language form, the referred object retrieval model aims at identifying and segmenting the target object from $\mathcal{I}$ based on the reference cue in $\mathcal{E}$. The explicit indication of the requested target through human

language serves as an effective device to fight against the first ambiguity source in sole visual cue. Another major source of ambiguity resides in the calculation process of the neural network, regarding which a confidence score s of the output object mask $\mathcal{M}$ is additionally produced to represent the uncertain or ambiguous level of the generated mask responding to $\mathcal{E}$. Human assistance will be requested if s is below a threshold to provide a single click on the target object which will be incorporated into the input to refine the segmentation result. The referred object segmentation model and the human-guided refinement strategy are depicted in Figure 6.2 and will be described in more detail in the following content.

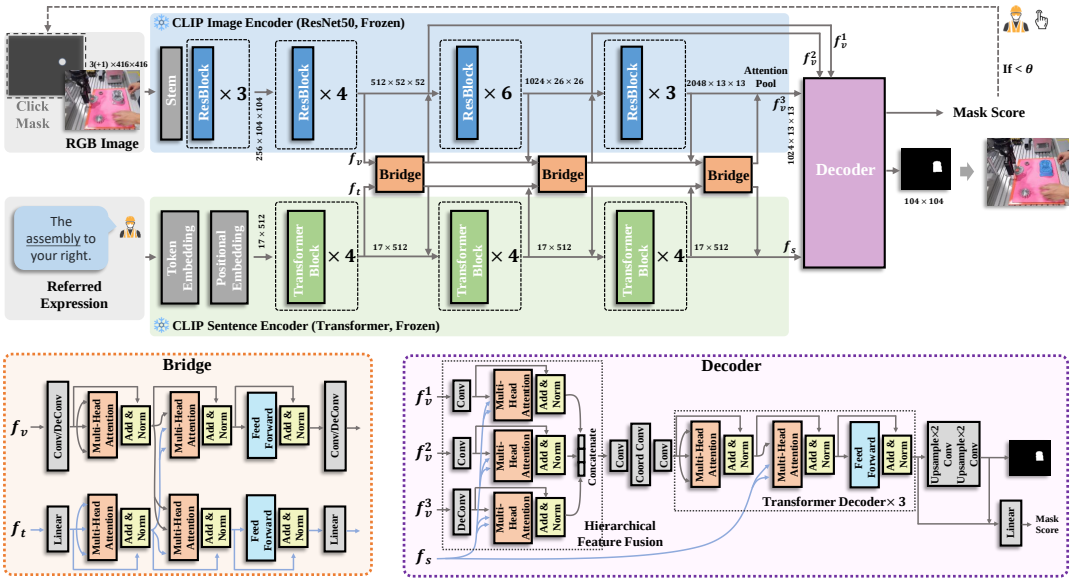


Fig. 6.2: Architecture of the proposed referred object retrieval model for HRC

1) Vision-Language Encoders

We mainly employ the CLIP (Contrastive Language-Image Pretraining) [240] model as the image and text encoder for the outstanding vision-language alignment ability provided by its web-scale pre-trained model. The captured HRC scene image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, where $H = W = 416$, is first input into the CLIP ResNet50 image encoder to extract the visual features. The visual

features from the last 3 stages, including $f_v^1 \in \mathbb{R}^{C_1 \times \frac{H}{8} \times \frac{W}{8}}$, $f_v^2 \in \mathbb{R}^{C_2 \times \frac{H}{16} \times \frac{W}{16}}$, and $f_v^3 \in \mathbb{R}^{C_3 \times \frac{H}{32} \times \frac{W}{32}}$, with $C_1 = 512, C_2 = C_3 = 1024$, will be extracted and fed into the decoder module for further processing.

On the other hand, the input reference expression $\mathcal{E} \in \mathbb{R}^L$, where $L = 17$ is the maximum length of the expression, is processed by a Transformer model [241] with the pre-trained weights from CLIP to extract the text feature $f_t \in \mathbb{R}^{L \times C}$. The Transformer utilizes a lower-case byte pair encoding (BPE) [242] representation for text, which is segmented to different sentences by `[SOS]` and `[EOS]` tokens. The activation of the final Transformer layer at the `[EOS]` token is extracted and further converted into the sentence-level feature representation $f_s \in \mathbb{R}^C$.

2) Cross-Modal Feature Fusion

Following the practice in [243], we fix the pre-trained weights of the CLIP encoders while introducing extra trainable Bridge modules to fuse the intermediate visual f_v and text features f_t to achieve cross-modality feature fusion. In this way, the strong capability of the pre-trained CLIP model could be preserved, the target domain vision-language feature interaction could be additionally learned by the Bridge modules, and the training process could be tremendously accelerated since most parameters are stored in the CLIP models.

The Bridge module takes in the visual feature f_v and text feature f_t and outputs the corresponding $f_v^{'}$ and $f_t^{'}$ with fused cross-modal information. Concretely, f_v will first be transformed to a predefined size $C_p \times H_p \times W_p$, where $C_p = 64, H_p = \frac{H}{16}, W_p = \frac{W}{16}$, by a stride-2 convolutional layer or

deconvolutional layer depending on whether the input height and width are higher than H_p and W_p. Meanwhile, the text feature f_t is also resized to match the predetermined size $L \times C_p$ via a linear layer. Then a Multi-Head Attention (MHA) [244] module along with the residual connection and layer normalization will be applied to the vision and text features respectively. The MHA is defined as:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1 \ldots, \text{head}_h)W^O, \tag{6.1}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{6.2}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(QK^T/\sqrt{d_k}\right)V, \tag{6.3}$$

where Q, K, V stand for query, key, and value respectively with dimensions $d_k = 64$. W^O, W_i^Q, W_i^K, W_i^V are the weight matrices of the linear projections inside MHA. $h = 6$ is the number of parallel heads utilized in this module. The first MHA module in the Bridge functions as the self-attention mechanism since the input Q, K, V are all from the same source. The second one serves as the cross-modal attention as Q comes from visual features while K, V come from text features. Residual connection and layer normalization are employed for each module following the standard Transformer block design. Then a fully connected feed-forward network is applied, upon which a final de-/convolutional and linear layer is constructed to recover the vision and text features to their original size.

3) Segmentation Decoder

As mentioned above, the multi-scale visual features f_v^1, f_v^2, f_v^3 and sentence feature f_s are sent to the decoder for segmentation generation. The detailed structure of the decoder is illustrated in Figure 6.2. Similar to the Bridge

module, the multi-scale visual features will first be transformed to be of the same spatial size $H_p \times W_p$ via de-/convolutional layers, followed by MHA modules as described in the previous part with f_s as query and key for extra vision-language feature fusion. Then the three branches of features are fused together by concatenation, after which a 1×1 convolution is first applied, closely followed by a coordinate convolution [245] which means a 2D spatial coordinate feature with the shape $2 \times \frac{H}{16} \times \frac{W}{16}$ will be attached to the fused feature maps before processed by an additional 3×3 convolutional layer.

Subsequently, several Transformer decoder blocks are borrowed to continue processing the fused visual feature f_v and sentence feature f_s. The fixed sine spatial positional encoding is first imposed, and the following self-attention, cross-attention, and feed-forward modules are similar to the ones adopted in the Bridge module. After the Transformer decoder, the final stage will consider how to project the features back to the image plane to obtain the mask prediction $\mathcal{M}$. The output cross-modal feature from the Transformer decoder can be denoted as $f_c \in \mathbb{R}^{C_1 \times \frac{H}{16} \times \frac{W}{16}}$, which will go through a combination of $2\times$ upsampling and 3×3 convolution for twice, resulting in an upscaled feature map $f_{up} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4}}$. The final target object mask prediction $\mathcal{M} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4}}$ can be further obtained by applying a sigmoid function to f_{up}. For training, a binary cross-entropy loss will be leveraged to guide the learning process of the mask prediction:

$$\mathcal{L}_{mask} = -\frac{1}{\hat{H}\hat{W}} \sum_{i \in \mathcal{M}} (y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)), \tag{6.4}$$

where y_i denotes the pixel label, p_i the probability of a pixel belonging to the target area, and $\hat{H} = \frac{H}{4}, \hat{W} = \frac{W}{4}$.

Meanwhile, to be able to represent the ambiguity or uncertainty of the predicted target mask, we further introduce a small branch to generate a mask score. The final feature map f_{up} will first be resized to match f_c, and later concatenated with f_c to be further regressed by a linear layer and a sigmoid activation function to obtain a score ranging from 0 to 1, depicting the estimated intersection percentage of the predicted mask with the ground truth mask. L2 loss is adopted for the mask score prediction:

$$\mathcal{L}_{score} = \|s - \hat{s}\|^2, \tag{6.5}$$

where s is the predicted mask score, while $\hat{s}$ is the ground truth intersection score. Since we already have the ground truth mask during training, $\hat{s}$ can be easily obtained by calculating the intersection-over-union of the predicted mask and the ground truth mask on the fly without requiring manual annotation. The overall loss function is the weighted sum of these two loss terms and can be written as:

$$\mathcal{L}_{overall} = \mathcal{L}_{mask} + \lambda \mathcal{L}_{score}, \tag{6.6}$$

where λ is empirically set as 0.1 so that the mask prediction task can be paid more attention during optimization.

6.2.2 Human-Guided Refinement Strategy

The model mentioned above should be able to effectively locate the target object based on the reference expression. However, due to the inherent probabilistic nature and random fluctuations of neural network models, the model may not perform well for some samples, in which cases a fail-safe

mechanism should exist. Regarding this issue, a human-guided refinement strategy is proposed by including human intervention in the model inference loop to improve the referred object segmentation accuracy.

The rationale behind the extra bother of the mask score prediction is the requirement of an indicator for the mask ambiguity or uncertainty that can inform the human operator whether intervention is needed. When the estimated score s is below a certain threshold $\theta = 0.5$, the human operator will be asked to provide extra information about the target goal in the form of a click on the image. A binary click mask $\mathcal{M}_{click} \in \mathbb{R}^{H \times W}$ will be generated based on the click coordinate consisting of a foreground circle centring around the click with a diameter of 5 pixels. The click mask will then be concatenated with the original input image $\mathcal{I}$ to form a new visual input into the referred object retrieval model and a fresh round of inference will be carried out to produce a refined estimation of the referred target mask. During training, it is not practical to seek human assistance for every training iteration, thus we simulate human click by randomly sampling some points inside the ground truth mask area and feed the model with or without the click mask with a probability of 0.5. For the cases where there is no click mask input, the extra input channel is set as an empty background.

For actual deployment in HRC manufacturing scenarios, this is a rather natural and reasonable strategy that offers the human operator a channel to correct algorithm mistakes without interrupting the production process.

6.2.3 LLM-Based Reasoning for Robot Planning

Based on the object location information provided by the referred object retrieval model, the object-referring ambiguity can already be effectively counteracted through the joint consideration of vision and language cues. Nonetheless, the full robotic task designated by the human language command still requires further comprehension and deduction in order to be translated into executable robotic action sequences, for which LLM is introduced as an adaptive robotic action planner. The proposed planner leverages the most advanced large language models like GPT-4 (Generative Pre-trained Transformer) to facilitate the generation of precise, context-aware robotic action plans, thereby enhancing the effectiveness and reliability of the collaborative tasks undertaken in the manufacturing environment.

The LLM-based planner is tasked with two essential functions: 1) extracting reference expressions from the natural language commands, and 2) planning the robotic actions based on the full command, primitive robotic skills, and target object locations. The first functionality is rather straightforward to implement with simple prompts demonstrating the reference expression extraction task with a few examples, while the second requires a more complicated construction of prompts.

1) Prompt Composition

The LLM-based robot action planner operates by first processing the input prompts, which are meticulously structured to contain essential components including domain knowledge of the dis-/assembly task, robotic primitive skill functions, utility functions, third-party libraries, few-shot examples,

Fig. 6.3: A demonstration of the prompt composition.

and the human instruction, as shown in Figure 6.3. Based on the prompt design strategies from [246, 247], we additionally incorporate manufacturing domain knowledge, such as the component information or assembly order of a certain product, into the input prompt so that the LLM can better adapt to specific industrial scenarios. The robotic primitive skill functions contain the APIs to the code implementations of essential robotic actions, which will be described in more detail in the next section. Utility functions refer to some helpful functions such as extracting the center of an object mask or transforming the image-plane coordinate into the robot coordinate. Third-party libraries include some widely used libraries such as NumPy, OpenCV, and PyTorch, while the few-shot examples provide some demonstrative input-output pairs for the LLM to achieve in-context learning [248]. Finally, the human language instruction is inserted to complete the full prompt.

2) Robotic Primitive Skills

As mentioned above, the most essential and indispensable part of the prompt is the robotic primitive skills, each associated with specific APIs to facilitate direct control and execution of tasks by the robot. The functionality descriptions of the APIs should also be included in the prompt so that the LLM can understand how to leverage them. Table 6.1 demonstrates examples of some of the primitive skills. Each of these skills and APIs is implemented through the Python interface of the official robotic driver. Based on the most fundamental primitive robotic actions such as `move_to()` or `open_gripper()`, more complex skills can be autonomously constructed. For example, the pick-and-place skill could be wraped into `pick_place()` by combining `move_to()`, `open_gripper()`, and `close_gripper()`. The actual implementation of these skills largely depends on the robot driver so that the LLM does not have to delve into the motion trajectory planning level. One advantage of the abstraction of robotic skills is the decoupling of robotic action planning and the actual implementation of robotic action control, which makes it more versatile and adaptive to different models of robots. Here we only include a minimum set of primitive skills for a robot arm to conduct pick-and-place tasks in order to illustrate the workflow of leveraging LLM as a robotic action planner. Note that other skills can always be added to the list so that the LLM can generate action plans for more complex scenarios and tasks.

Primitive Skills	**API**
move to position	`move_to(coord)`
move along axis	`move_axis(axis, dist)`
open gripper	`open_gripper()`
close gripper	`close_gripper()`
call the referred object retrieval model	`obj_retrieval(image, obj_ref)`

Table 6.1: Robotic Primitive Skills and Corresponding APIs

6.3 Experimental Results

To demonstrate the utility and effectiveness of the proposed vision-language reasoning approach, experiments are carried out in a simulated HRC assembly scenario in the lab environment, in which the human-robot team is designated to assemble a gear pump module. The human operator will guide a UR5 robot arm via language command to accomplish certain sub-tasks such as fetching an assembly part. An Azure Kinect RGB-D camera is positioned above the working station to monitor the whole assembly area. To verify the performance of the proposed method, we first evaluate the referred object retrieval model against existing works on the HRC scenario data and a public dataset, and then the LLM robotic action planner is assessed based on some specifically synthesized language commands that mimic possible human instructions.

6.3.1 Experiments for Referred Object Retrieval

1) Evaluation on the HRC Assembly Case

Data Collection. To evaluate the proposed referred object retrieval model, a dataset consisting of paired images and reference expressions should be first set up. Specifically, during the assembling process of the gear pump module, the Kinect camera will capture images of the assembly working station which covers the area of different parts or tools that may be required by a certain assembling step. The human operator is mainly responsible for guiding the robot arm to execute different sub-tasks through language instructions. In order to build a dataset with enough data variance, the images of the HRC

assembly scenes are first captured, each of which will then be manually allocated 2 to 3 handcrafted reference expressions to its contained objects. The corresponding segmentation masks for these referred objects are also manually annotated. In total, we collected 463 data pairs, 370 of which are utilized as training data, while the rest are regarded as the testing set.

Experimental Settings. The referred object retrieval model is implemented using PyTorch, which is the most widely employed deep learning library for its flexibility and user-friendliness. For hardware acceleration, a Nvidia RTX3090 GPU is leveraged to train and evaluate the model. The pre-trained weights of the CLIP image and sentence encoders are borrowed from the original CLIP paper [240]. The pre-trained weights of the Bridge and Decoder modules are partially adopted from [243] for the common layers, while the additional layers are randomly initialized. Adam optimizer with initial learning rate 1×10^{-4} and batch size 32 is adopted to train the model.

Table 6.2: Referred Object Segmentation Performance on the HRC Dataset

Method	Components	mIoU
Xu *et al.* [243]	-	79.04
Ours	w/ mask score + w/o click input	79.31
	w/o mask score + w/ click input	82.91
	w/ mask score + w/ click input	**83.18**

Evaluation Results. To evaluate the performance of the proposed referred object retrieval model, we adopt the mIoU metric, which is a standard metric normally utilized to measure segmentation model performance. In the referred object segmentation case, it calculates the overlap between the estimated target object segmentation mask and the ground truth mask over

the testing set. The quantitative evaluation results are illustrated in Table 6.2, which compares the proposed model with some previous approaches and includes a brief ablation study. Compared to the baseline model [243], the proposed model with mask score prediction and click input performs the best, which demonstrates the effectiveness and superiority of the proposed model. To validate the effect of the two main components including the mask score prediction branch and the click input strategy, we also include the evaluation results of two variants of the proposed model–one excludes the click input and the other additionally excludes the mask score prediction–both showing inferior results than the full model depicted in the last row of the table. While the mask score can only marginally improve the accuracy, the click input strategy has a considerably higher impact on the model performance since the human click is in fact quite strong prior knowledge and rather inefficient if human intervention is required for every case. The mask score prediction is thus designed to minimize the ratio of human intervention by automatically evaluating the mask quality to achieve a balance between accuracy and efficiency.

Some qualitative examples are illustrated in Figure 6.4 containing input-output pairs from the proposed model to provide an intuitive visualization of the model performance. As shown in the figure, the object retrieval model can successfully comprehend the reference text and provide the target object segmentation mask in the HRC scenario. One sample to be noted is in the last row, in which we deliberately feed a nonexistent reference expression with an incorrect location description. Since the target object is not in the referred location, the model is confused and yields responses in both the target object and the referred location, which illustrates that the proposed model can

indeed understand both the referred object and the referred location instead of simply responding to the target object.

2) Evaluation on the RefCOCO dataset

Dataset Introduction. To further illustrate the universality of the proposed model to different scenarios and scalability to large-scale data, we additionally conduct comparative experiments on a public dataset RefCOCO [249]. It is a prevailing benchmark dataset for referring object segmentation, encompassing 19,994 images paired with 142,210 referring expressions for 50,000 unique objects derived from the MSCOCO dataset [250]. It is systematically partitioned into four subsets: 120,624 for training, 10,834 for validation, 5,657 for test A, and 5,095 for test B. Every image incorporates at least two objects, underscoring the dataset's comprehensive and nuanced construction for evaluating reference segmentation models.

Table 6.3: Comparative Experiments on the RefCOCO Dataset

Method	mIoU		
	val	**test A**	**test B**
Ding *et al.* [251]	65.65	68.29	62.73
Wang *et al.* [252]	70.47	73.18	66.10
Xu *et al.* [243]	71.06	74.11	66.66
Ours	**77.23**	**79.53**	**74.88**

Evaluation Results. Table 6.3 lists the comparative results of the proposed model with some existing approaches. As mentioned above, the RefCOCO dataset has three subsets besides the training set. Therefore, the evaluated metrics for these three subsets are all provided following the common practice. The evaluation metric values of the compared methods are directly borrowed from the best results reported in the corresponding published papers. As

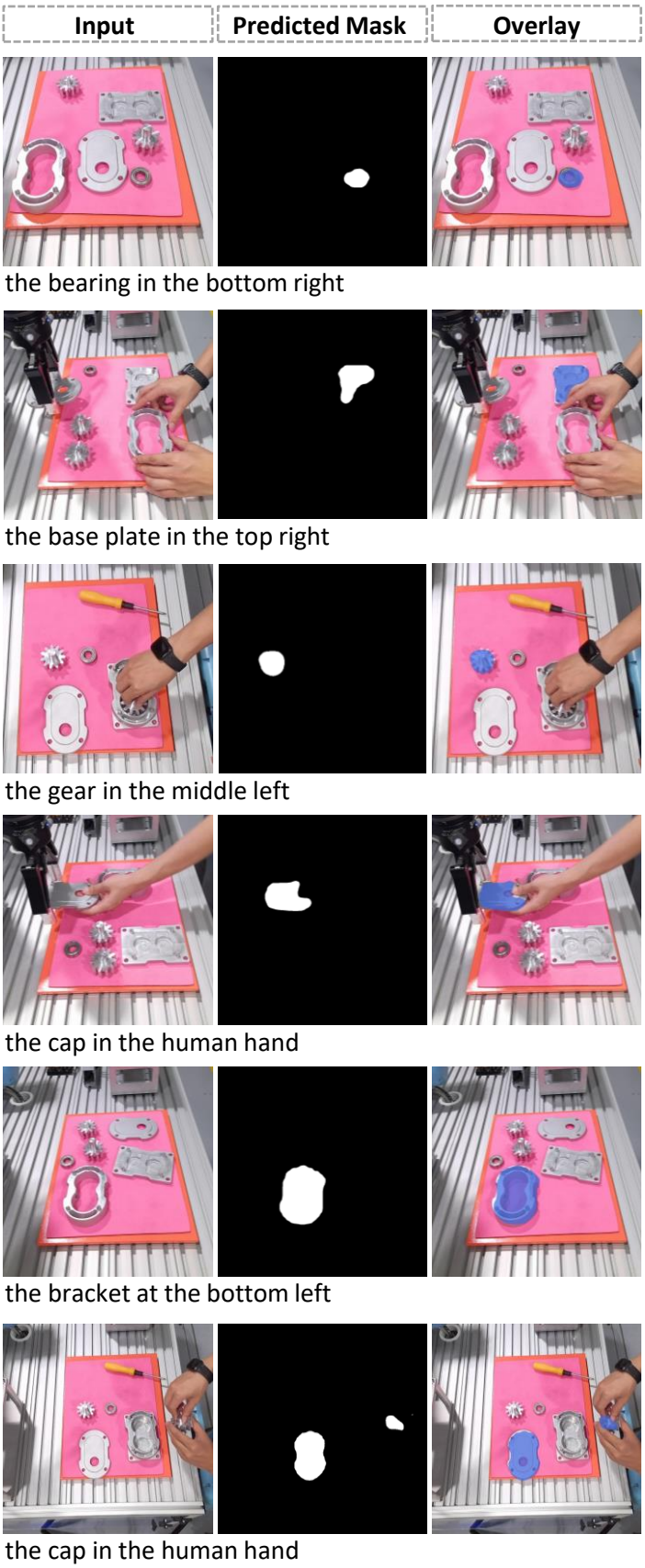


Fig. 6.4: Qualitative samples of referred object segmentation.

for the proposed method, the evaluation results are obtained from the full model with all the contributed components. From the listed experimental results, we can observe a considerable improvement in the proposed model over other methods. We mainly attribute this achievement to the adoption of the click input strategy which provides a strong prior for the approximate whereabouts of the target object. While this might seem a bit unfair to other compared methods at first glance, it is in fact quite practical and effortless to implement in HRC scenarios where interactions between humans and robots already pervasively exist in various manufacturing stages.

6.3.2 Experiments for LLM-Based Robot Planning

The second part of the experiments focuses on demonstrating the effectiveness of the LLM-based robotic planning strategy. To be able to quantitatively evaluate the robot action planning effectiveness, we predefined 3 subtask instructions extracted from the HRC gear pump assembly procedure: 1) fetch a gear pump case from the storage area, 2) place a gear into the case, and 3) pick a case cover and put onto the assembly, representing the most frequent cases in the assembling process. For each of these tasks, the LLM-generated robotic action plans will be evaluated by human experts based on the correctness of function calls and the feasibility of the action sequence. The LLM generation process is repeated 20 times for each subtask, and the success rate is utilized as the evaluation metric. In terms of LLMs, we adopt GPT-3.5 and GPT-4 [253] via OpenAI API, and LLaMA-2 (Large Language Model Meta AI) [254] for comparative study. The comparison results are shown in Table 6.4, from which we can observe an obvious superiority of the GPT-4 model, which is not unexpected given its tremendously larger amounts of parameters. However, the close-source nature of GPT models renders it infeasible to be

applied in off-line environments and time-sensitive production scenarios, which leaves abundant opportunities for open-source models such as LLaMA and compressing and accelerating technologies.

Table 6.4: Results of the LLM-based Robot Planning

Model	Task 1	Task 2	Task 3	Average
LLaMA2-7b	7/20	6/20	3/20	26.7%
GPT-3.5	17/20	19/20	15/20	85.0%
GPT-4	19/20	20/20	17/20	93.3%

6.3.3 Discussions

The experimental results of our proposed referred object retrieval model exhibit consistent improvements on our HRC gear pump assembly dataset and the RefCOCO dataset. While the pretrained CLIP encoders, cross-modal fusion, and decoder design each make a solid contribution to the model performance, the main merit of our approach is the better leveraging of the human-robot symbiotic environment via the human-guided refinement strategy, which is naturally available in an HRC scenario. Compared with existing methods that already incorporate human language instructions to locate target objects, our proposed model simply moves a step further by requesting the human operator to click on the image when the confidence of the segmentation result is unsatisfactory. On the other hand, the brief experimental verification of the LLM-based robotic action planner also makes a preliminary demonstration of how to integrate LLMs into a human-robot collaborative manufacturing system. Though being premature for actual deployment in production environments for issues such as network latency and computational cost, the tremendous potential of LLMs will undoubt-

edly empower futuristic HRC systems with more advanced intelligence and flexibility.

6.4 Chapter Summary

The forthcoming industrial transition has prioritized human centricity over purely profit-oriented drives, in which context HRC is considered a promising solution. Motivated by the visual information ambiguity and insufficient exploitation of human language cues in previous HRC systems, a vision-language-based reasoning approach is proposed in this chapter. The main contributions can be summarized in twofold: 1) Proposed a referred object retrieval model for spotting target object location based on vision-language input fueled by pretrained CLIP encoders, cross-modal fusion, Transformer-like decoder design, and human-guided refinement strategy. 2) Explored LLM-based robotic action planning strategy for generating robotic action plans for HRC scenarios in the form of executable code, featured with a specially and meticulously designed prompt structure containing task-specific domain knowledge and robotic primitive skills. Further experimental studies for these two major aspects have been conducted and demonstrated favorable results of the proposed approach. Nevertheless, some limitations still exist such as the extra computational cost of the human-guided refinement strategy and the network and response latency of the LLM-based robotic planner. To further investigate these challenges, some potential future research directions include: 1) incorporating other modalities of data into human-robot interactions to reduce the task complexity of object indication understanding; 2) developing light-weight LLM models for specific industrial scenarios instead of general purpose to shrink the model parameters scales;

and 3) compressing or distillating opensource LLM models for offline and time-sensitive HRC cases.

7 Conclusions

This research has marked a leap forward as a systematic exploration of vision-based approaches for holistic scene understanding in HRC. The four aspects of scene understanding have been systematically investigated to form a holistic perspective which paved the way for subsequent collaborative robot decision-making in HRC systems. The objectives illustrated in the early chapters have been met by: 1) developing a high-resolution 6-DoF pose estimation model for industrial components and incorporating explicit occlusion awareness to refine hand-object pose estimation, 2) advancing human operator digital twin modelling for improved human-robot interaction, 3) innovating multi-granularity workspace parsing for more nuanced environment perception in HRC, and 4) pioneering vision-language reasoning for ambiguity mitigation, thus facilitating a more intuitive and efficient human-robot collaboration. This chapter summarizes the key contributions, discusses the research limitations, and outlines future work directions in Sections 7.1, 7.2, and 7.3, respectively.

7.1 Contributions

The main goal of this project is to investigate the research gaps and establish solutions for vision-based holistic understanding in HRC scenarios, specifically in four aspects: object perception, human recognition, environment parsing, and visual reasoning. At the beginning of this project, a systematic

survey was conducted around the four aspects of vision-based holistic scene understanding. Then the following chapters each contributed to a specific facet, the contributions of which are listed as follows.

Contribution 1: A high-resolution network-based 6-DoF pose estimation model of industrial parts has been proposed aiming to facilitate HRC disassembly, which later extended to the investigation of joint hand-object reconstruction motivated by the frequent mutual occlusion.

To equip collaborative robots with the ability to understand the objects and ongoing human hand-object interactions in an HRC environment, the initial endeavour of this project has been devoted to the 6-DoF pose estimation of industrial workpieces. A novel high-resolution network-based 6-DoF pose estimation model for industrial parts was first designed to improve the accuracy and success rate of robotic manipulation. This model, composed of a coarse pose estimation stage followed by a refinement stage, utilizes the High-Resolution Network as the backbone to extract high-resolution feature representations. The rotation and translation parameters are first roughly estimated in the first stage and then refined in the subsequent stage. Empirical evaluations demonstrate that this model outperforms baseline models on an industrial parts dataset. Motivated by the prevalent hand-object occlusion issue during HRC, an integrated hand-object pose estimation model was further proposed. This model employs a mask-guided attentive residual block within a branched model structure, facilitating the hand-object attention separation during feature extraction. Additionally, an FPN-like subnetwork was introduced to predict the occlusion ternary mask, which is then compared with a rendered mask from the estimated hand-object pose to provide explicit occlusion awareness, thereby reducing pose estimation errors

caused by hand-object occlusions. The subsequent experimental results on both the Li-ion data disassembly data and a public hand-object pose dataset demonstrated an obvious improvement over existing methods.

Contribution 2: A vision-based human digital twin modelling approach has been proposed to serve as an exemplary implementation of onsite human operator digitalization and monitoring for HRC.

Aiming to provide robots with a comprehensive understanding capability of their human partners beyond standalone recognition tasks and also fill the gap of lacking practical solutions to the vision-based HDT, this research work proposes a vision-based HDT modelling approach that integrates multiple perspectives of human perceptions into a unified deep learning model which can operate in an end-to-end manner in the onsite HRC scenarios. The primary contributions can be summarized in twofold: 1) A specifically designed deep learning architecture is tailored to concurrently assess 3D human posture, action intention, and ergonomic risk, thus facilitating effective HDT modelling. 2) Based on the real-time updated HDT information, an adaptive robotic motion control strategy is developed and demonstrated in the HRC contexts to serve as an example of the potential applications of the HDT model in HRC cases.

Contribution 3: A multi-granularity scene segmentation model has been studied aiming to provide multi-granularity semantics for flexible HRC tasks.

In response to the absence of a flexible environment perception scheme in current HRC systems, this study introduced a multi-granularity scene segmen-

tation model, aiming to segment the HRC environment into various semantic granularities, thereby being able to adapt to the dynamically changing requirements prevalent in HRC situations. By integrating a series of modern network design strategies, the proposed model has achieved prominent results in the collaborative battery disassembly case and demonstrated comparative performance with state-of-the-art methods on the NYUv2 dataset. The main contributions can be summarized as follows: 1) an RGB-D segmentation network MGS-Net was proposed leveraging modern network designs such as the ConvNext backbone, multi-scale supervision, multi-granularity prediction, et cetera; 2) a multi-granularity segmentation criterion was established in an HRC scenario and the feasibility of the proposed model was demonstrated based on this criterion; 3) the model performance was evaluated on the NYUv2 dataset and demonstrated comparable results to state-of-the-art methods.

Contribution 4: A vision and language-guided reasoning method has been presented to reduce the uncertainty of human-robot communication in HRC scenarios.

To further enhance the robotic embodied intelligence with abstract and logical reasoning ability, the final step of this thesis was dedicated to the exploration of visual reasoning technologies. Motivated by the ambiguity inherent in visual information and the limited exploitation of human linguistic cues, a vision-language-based referred object retrieval model was first proposed. Then a large language model was also employed to boost the natural-form human-robot communication. The main contributions can be summarized in twofold: 1) This study proposed a referred object retrieval model for spotting target object location based on vision-language input. This model leverages

a combination of pre-trained CLIP encoders and cross-modal feature fusion, along with a Transformer-like decoding mechanism. It further incorporates a human-guided refinement process to include humans in the inference loop, which is a unique strength that can be leveraged in HRC contexts. 2) The application of large language models in robotic action planning specific to HRC contexts was explored. A carefully crafted prompt structure was studied, which incorporates both domain-specific knowledge relevant to the task at hand and primitive robotic skills. This structure ensures that the generated plans are both relevant and practical for real-world HRC applications.

7.2 Limitations

Limitation 1: The dependency on annotated data for model training.

Although the employment of deep learning models in the holistic understanding of HRC scenes has illustrated excellent performance during preliminary studies, the requirement of a large dataset with ground truth annotations can impede the deployment of the proposed models since it may not be always fulfilled. On the one hand, in manufacturing scenarios, it is impractical to collect a large-scale dataset since each factory and production line is highly customized and can be tremendously different from each other. On the other hand, even with a large-scale dataset, the manual annotation would demand too much human labour. A possible solution is to explore a high-fidelity simulation environment that could synthesize almost indefinite variations of data to fuel the training of deep learning models. Another direction is to develop semi- or unsupervised training techniques to hopefully reduce the dependency on annotated data.

Limitation 2: The throughput latency and heavy computational cost of deep learning-based scene understanding models.

The second issue regarding the adoption of deep learning models in the holistic scene understanding scheme is the heavy reliance on computational resources and power, and the high latency associated with it. With the parameter scales growing larger, deep learning models seem to be able to consistently deliver better performance, especially with the most recent Transformer-based models. However, this phenomenon also gradually shifts deep learning models away from being able to run on traditional computing hardware in the manufacturing site. While many try to deploy the models in cloud servers, the inevitable network latency also poses another challenge. Therefore, compression and acceleration technologies should be paid more attention when applying the proposed models in the actual production environment.

Limitation 3: The lack of consideration of multi-modal data source.

This study mainly discussed visual data-based, such as RGB image and depth image, approaches for holistic scene understanding in HRC due to the affordability and availability of vision sensors. Although human language has also been introduced as a complementary modality in the last part of the thesis, it is worthwhile to devote more effort to investigating the integration of more types of modalities such as acoustic data and wearable sensors. The incorporation of multiple modalities of data sources can potentially introduce a novel perspective to address existing technological challenges that have proved to be difficult to tackle with visual data alone, thereby improving the

cognition skills of collaborative robots and further boosting the HRC working efficacy.

Limitation 4: The limited incorporation of holistic scene understanding into the collaborative robot controlling scheme.

This thesis closely contemplated the four aspects of holistic scene understanding and proposed several approaches to address the unique challenges in each of these aspects. Nevertheless, the proposed methods are still largely confined within the perception and scene understanding scope, while the incorporation of the perceived results with collaborative robot control strategies has only been superficially scratched without in-depth consideration. More explorations could be made with regard to integrating the scene understanding skills into the adaptive controlling or robot learning process in order to achieve more flexible collaboration and improve the overall HRC productivity.

Limitation 5: The lack of in-depth theoretical analysis regarding the rationale of customized deep learning model architecture design.

This thesis investigated several deep learning architectures for different visual understanding objectives and proposed customized and modified models to enhance scene understanding performance for further improvement in HRC efficacy. Although intuitive insights behind these model design choices were provided, a deeper theoretical examination to validate the effectiveness of the approaches is lacking. More analyses and discussions from a principle level, such as cognitive science, could be explored to fill in this gap, and thereby consolidating the theoretical underpinnings of the research works.

7.3 Future Research Directions

The concept of seamless collaboration between human and robotic agents depicts a splendid vision for futuristic manufacturing. While this thesis has made a trivial contribution to the field by devising potential solutions to each of the four aspects of HRC scene understanding, the ultimate comprehension of the holistic scene perspective of HRC remains a topic with significant room for advancement. This section outlines several promising future research directions that are critical to the progression of HRC.

(1) Semi-/unsupervised object pose estimation with minimum data reliance and sim-to-real transfer.

As stated earlier, one major issue that weighs heavily against the current 6-DoF object pose estimation approaches is the dependency on large-scale labelled data. The present formulation of deep learning-based 6-DoF pose estimation is in fact ill-posed since it is impossible to collect object images from all possible view angles. Several possible techniques could be taken better advantage of in future research: 1) The exploration and exploitation of semi-/unsupervised model training schemes for reduced manual annotation demand. 2) The adoption of sim-to-real transfer techniques for better learning performance on simulation-synthesized data. While these technologies are still nowhere near maturity, the combination of them is definitely a promising line of work to explore.

(2) Dismountable modular human digital twin model with dynamic plug-and-play design for flexible customization of different HRC scenarios.

The proposed HDT model in this thesis managed to integrate three aspects of human operator body status into an end-to-end deep learning perception model including human 3D posture, human action intention, and ergonomic risk of the current human posture. However, after the training process of the model, it is impossible to modify the perception functionalities such as adding or removing a branch without going over the training process altogether. On the other hand, a dismountable modular design of the HDT model can greatly facilitate the flexibility of the model to be applied in more HRC scenarios. To tackle this issue, possible future research directions include: 1) leveraging contrastive learning strategies to separate the feature extraction pretraining stage and the downstream specific task alignment training stage, and 2) adopting Transformer structures to modularly train a series of decoders for different functionalities while sharing the same encoder to improve inference efficiency. It is envisioned that in the future a library of different general human perception modules might be shared and exchanged and can be easily plugged into different HDT systems via mere graphical interfaces without programming.

(3) HRC scene parsing with multiple views of data and advanced neural representations.

The multi-granularity scene segmentation model proposed in Chapter 5 focuses on how to perceive multiple levels of semantics from a single view of camera observation in one go. While being efficient, the incompleteness of the visual scene information would inevitably result in failures in some corner cases, which is unacceptable for serious manufacturing scenarios. In order to obtain a more comprehensive and accurate representation of the HRC environment, future directions could be: 1) leveraging multiple-view

camera data to update the digital representation of the HRC scene in real-time via 3D reconstruction or SLAM (Simultaneous Localization And Mapping) technologies for multiple fixed cameras or single mobile camera, respectively, and 2) exploring implicit neural representations of the 3D HRC scene to encode the 3D geometries in a more efficient fashion.

(4) Lightweight reasoning approach with multimodal data sources and unified scene element description.

As discussed in Section 6, the final leap from scene perception to scene understanding is visual reasoning, which resembles the human abstract reasoning process based on human sensory information. Although current vision-language models and large language models have exhibited unprecedented intelligence and understanding capability, the computational resources required to train these models are astronomical. And the unavoidable latency issue is also an urgent factor preventing these large models from being widely deployed in real production environments. One future direction therefore naturally should be how to compress and accelerate the large models without compromising their human-like intelligence. On the other hand, when adopting large vision and language models in the holistic scene understanding perspective, how to efficiently and adequately incorporate the perceived scene element information into the reasoning process is also a promising direction.

(5) Human cognition process-inspired principle perspective for improved vision and cross-modal scene understanding.

As mentioned in the limitation section, the research works presented in this thesis focused more on how to intuitively modify existing deep learning models to better fit into specific HRC scene characteristics. However, the absence of a higher principle level considerations renders these customizations less attractive in terms of scientific solidity. In this regard, more explorations in the future could be devoted into the biological human cognition process in order to draw more reasonable and plausible insights and provide more solid theoretical understandings.

References

[1] Eloise Matheson, Riccardo Minto, Emanuele GG Zampieri, Maurizio Faccio, and Giulio Rosati. "Human–Robot Collaboration in Manufacturing Applications: A Review". In: *Intel. Syst. Contr. Aut.* 8.4 (2019), p. 100 (cit. on pp. 2, 122).

[2] Lihui Wang, R Gao, József Váncza, et al. "Symbiotic human-robot collaborative assembly". In: *CIRP Ann.* 68.2 (2019), 701–726 (cit. on pp. 2, 84, 104, 122).

[3] Hongyi Liu and Lihui Wang. "Collision-free human-robot collaboration based on context awareness". In: *Rob. Comput. Integr. Manuf.* 67 (2020), p. 101997 (cit. on p. 2).

[4] Peng Wang, Hongyi Liu, Lihui Wang, and Robert X Gao. "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration". In: *CIRP Ann.* 67.1 (2018), 17–20 (cit. on pp. 2, 27, 28).

[5] Sandra Robla-Gómez, Victor M Becerra, José Ramón Llata, et al. "Working together: A review on safe human-robot collaboration in industrial environments". In: *IEEE Access* 5 (2017), 26754–26773 (cit. on p. 2).

[6] Hongyi Liu and Lihui Wang. "Gesture recognition for human-robot collaboration: A review". In: *Int. J. Ind. Ergonom.* 68 (2018), 355–367 (cit. on pp. 2, 122).

[7] Zanwu Xia, Qujiang Lei, Yang Yang, et al. "Vision-based hand gesture recognition for human-robot collaboration: A survey". In: *2019 5th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE. 2019, 198–205 (cit. on p. 2).

[8] Dahua Lin, Sanja Fidler, and Raquel Urtasun. "Holistic scene understanding for 3d object detection with rgbd cameras". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2013, 1417–1424 (cit. on p. 2).

[9] Muzammal Naseer, Salman Khan, and Fatih Porikli. “Indoor scene understanding in 2.5/3d for autonomous agents: A survey”. In: *IEEE Access* 7 (2018), 1859–1887 (cit. on p. 2).

[10] Giovanni Pintore, Claudio Mura, Fabio Ganovelli, et al. “State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments”. In: *Comput. Graphics Forum*. Vol. 39. Wiley Online Library. 2020, 667–699 (cit. on p. 2).

[11] Marcos Ferreira, António Paulo Moreira, and Pedro Neto. “A low-cost laser scanning solution for flexible robotic cells: spray coating”. In: *Int. J. Adv. Manuf. Technol* 58.9-12 (2012), 1031–1041 (cit. on pp. 12, 13).

[12] Andry Maykol Pinto, Luís F Rocha, and A Paulo Moreira. “Object recognition using laser range finder and machine learning techniques”. In: *Rob. Comput. Integr. Manuf.* 29.1 (2013), 12–22 (cit. on p. 13).

[13] Shuaifeng Zhi, Yongxiang Liu, Xiang Li, and Yulan Guo. “Toward real-time 3D object recognition: A lightweight volumetric CNN framework using multitask learning”. In: *Comput. Graphics* 71 (2018), 199–207 (cit. on p. 13).

[14] S Hamidreza of Kasaei. “OrthographicNet: A Deep Transfer Learning Approach for 3D Object Recognition in Open-Ended Domains”. In: *IEEE/ASME Trans. Mechatron.* (2020) (cit. on pp. 13, 14).

[15] Masood Dehghan, Zichen Zhang, Mennatullah Siam, et al. “Online object and task learning via human robot interaction”. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, 2132–2138 (cit. on pp. 13, 14).

[16] Yunhe Feng, Zongyao Chen, Dali Wang, Jian Chen, and Zhili Feng. “DeepWelding: A deep learning enhanced approach to GTAW using multisource sensing images”. In: *IEEE Trans. Ind. Inf.* 16.1 (2019), 465–474 (cit. on pp. 13, 14, 44).

[17] Ingo Keller and Katrin S Lohan. “On the Illumination Influence for Object Learning on Robot Companions”. In: *Front. Rob. AI* 6 (2020), p. 154 (cit. on pp. 13, 14).

[18] Van-Thanh Nguyen, Chao Lin, Chih-Hung G Li, Shu-Mei Guo, and Jenn-Jier James Lien. “Visual-guided robot arm using self-supervised deep convolutional neural networks”. In: *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2019, 1415–1420 (cit. on pp. 13, 14).

[19] Salvatore D’Avella, Paolo Tripicchio, and Carlo Alberto Avizzano. “A study on picking objects in cluttered environments: Exploiting depth features for a custom low-cost universal jamming gripper”. In: *Rob. Comput. Integr. Manuf.* 63 (2020), p. 101888 (cit. on pp. 13, 14).

[20] Mia Kokic, Johannes A Stork, Joshua A Haustein, and Danica Kragic. "Affordance detection for task-specific grasping using deep learning". In: *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE. 2017, 91–98 (cit. on pp. 13, 15).

[21] Raja Chatila, Erwan Renaudo, Mihai Andries, et al. "Toward self-aware robots". In: *Front. Rob. AI* 5 (2018), p. 88 (cit. on pp. 13, 15).

[22] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. "Affordance detection of tool parts from geometric features". In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2015, 1374–1381 (cit. on pp. 13, 15).

[23] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. "Detecting object affordances with convolutional neural networks". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, 2765–2770 (cit. on pp. 13, 15).

[24] Spyridon Thermos, Gerasimos Potamianos, and Petros Daras. "Joint Object Affordance Reasoning and Segmentation in RGB-D Videos". In: *IEEE Access* 9 (2021), 89699–89713 (cit. on pp. 13, 15).

[25] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014 (cit. on p. 14).

[26] Xiang Li, Jintao Wang, Fang Xu, and Jilai Song. "Improvement of YOLOv3 Algorithm in Workpiece Detection". In: *2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE. 2019, 1063–1068 (cit. on pp. 16, 17).

[27] Kung-Jeng Wang, Diwanda Ageng Rizqi, and Hong-Phuc Nguyen. "Skill transfer support model based on deep learning". In: *J. Intell. Manuf.* 32.4 (2021), 1129–1146 (cit. on pp. 16, 17, 44).

[28] George Andrianakos, Nikos Dimitropoulos, George Michalos, and Sotirios Makris. "An approach for monitoring the execution of human based assembly operations using machine learning". In: *Procedia CIRP* 86 (2019), 198–203 (cit. on pp. 16, 17).

[29] Eugen Solowjow, Ines Ugalde, Yash Shahapurkar, et al. "Industrial Robot Grasping with Deep Learning using a Programmable Logic Controller (PLC)". In: *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2020, 97–103 (cit. on pp. 16, 17).

[30] Seunghyeok Back, Jongwon Kim, Raeyoung Kang, Seungjun Choi, and Kyoobin Lee. "Segmenting unseen industrial components in a heavy clutter using rgb-d fusion and synthetic data". In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, 828–832 (cit. on pp. 16, 17, 19).

[31] Pablo Azagra, Javier Civera, and Ana C. Murillo. “Incremental Learning of Object Models From Natural Human–Robot Interactions”. In: *IEEE Trans. Autom. Sci. Eng.* 17.4 (2020), pp. 1883–1900 (cit. on pp. 16, 17, 19).

[32] Patrick Rosenberger, Akansel Cosgun, Rhys Newbury, et al. “Object-Independent Human-to-Robot Handovers Using Real Time Robotic Vision”. In: *IEEE Robot. Autom. Lett.* 6.1 (2021), pp. 17–23 (cit. on pp. 16, 17).

[33] Sergey Astanin, Dario Antonelli, Paolo Chiabert, and Chiara Alletto. “Reflective workpiece detection and localization for flexible robotic cells”. In: *Rob. Comput. Integr. Manuf.* 44 (2017), 190–198 (cit. on p. 17).

[34] Yi-Hsuan Hsieh, Pei-Chi Huang, Qixing Huang, and Aloysius K Mok. “LASSO: Location Assistant for Seeking and Searching Objects”. In: *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE. 2019, 94–100 (cit. on p. 17).

[35] Nicola Castaman, Elisa Tosello, Morris Antonello, et al. “RUR53: an unmanned ground vehicle for navigation, recognition, and manipulation”. In: *Adv. Robotics* 35.1 (2021), 1–18 (cit. on p. 17).

[36] Heiko Hoffmann, Zhichao Chen, Darren Earl, et al. “Adaptive robotic tool use under variable grasps”. In: *Robot. Auton. Syst.* 62.6 (2014), 833–846 (cit. on pp. 17, 18).

[37] Khurshid Aliev and Dario Antonelli. “Analysis of cooperative industrial task execution by mobile and manipulator robots”. In: *International Scientific-Technical Conference MANUFACTURING*. Springer. 2019, 248–260 (cit. on pp. 17, 18).

[38] Doreen Jirak, David Biertimpel, Matthias Kerzel, and Stefan Wermter. “Solving visual object ambiguities when pointing: an unsupervised learning approach”. In: *Neural. Comput. Appl.* 33.7 (2021), 2297–2319 (cit. on pp. 17, 18).

[39] Mitchell Dinham and Gu Fang. “Autonomous weld seam identification and localisation using eye-in-hand stereo vision for robotic arc welding”. In: *Rob. Comput. Integr. Manuf.* 29.5 (2013), 288–301 (cit. on pp. 17, 18).

[40] Hwaseop Lee, Yeeyeng Liau, Siku Kim, and Kwangyeol Ryu. “A framework for process model based human-robot collaboration system using augmented reality”. In: *IFIP International Conference on Advances in Production Management Systems*. Springer. 2018, 482–489 (cit. on pp. 17, 18).

[41] Panagiota Tsarouchi, Stereos-Alexandros Matthaiakis, George Michalos, Sotiris Makris, and George Chryssolouris. “A method for detection of randomly placed objects for robotic handling”. In: *CIRP J. Manuf. Sci. Technol.* 14 (2016), 20–27 (cit. on pp. 17, 19, 21).

[42] Xiang Li, Xing Su, and Yunhui Liu. “Cooperative robotic soldering of flexible PCBs”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, 1651–1656 (cit. on pp. 17, 19).

[43] Rafiq Ahmad and Peter Plapper. “Safe and automated assembly process using vision assisted robot manipulator”. In: *Procedia CIRP* 41 (2016), 771–776 (cit. on pp. 17, 20).

[44] Vladimir Kuts, Tauno Otto, Toivo Tähemaa, Khuldoon Bukhari, and Tengiz Pataraia. “Adaptive industrial robots using machine vision”. In: *ASME 2018 International Mechanical Engineering Congress and Exposition*. Vol. 52019. ASME. 2018, V002T02A093 (cit. on pp. 17, 20).

[45] Hamdi Ben Abdallah, Igor Jovančević, Jean-José Orteu, and Ludovic Brèthes. “Automatic inspection of aeronautical mechanical assemblies by matching the 3D CAD model and real 2D images”. In: *J. Imaging* 5.10 (2019), p. 81 (cit. on p. 21).

[46] Frederik Hagelskjær, Thiusius Rajeeth Savarimuthu, Norbert Krüger, and Anders Glent Buch. “Using spatial constraints for fast set-up of precise pose estimation in an industrial setting”. In: *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2019, 1308–1314 (cit. on p. 21).

[47] Zaixing He, Zhiwei Jiang, Xinyue Zhao, Shuyou Zhang, and Chenrui Wu. “Sparse template-based 6-D pose estimation of metal parts using a monocular camera”. In: *IEEE Trans. Ind. Electron.* 67.1 (2019), 390–401 (cit. on pp. 21, 44, 69).

[48] Ren C Luo and Chia-Wen Kuo. “Intelligent seven-DoF robot with dynamic obstacle avoidance and 3-D object recognition for industrial cyber–physical systems in manufacturing automation”. In: *Proc. IEEE* 104.5 (2016), 1102–1113 (cit. on pp. 21, 22).

[49] Daniel Wahrmann, Arne-Christoph Hildebrandt, Christoph Schuetz, Robert Wittmann, and Daniel Rixen. “An autonomous and flexible robotic framework for logistics applications”. In: *J. Intell. Robot. Syst.* 93.3 (2019), 419–431 (cit. on pp. 21, 22).

[50] Y Zhang, C Zhang, R Nestler, M Rosenberger, and G Notni. “Efficient 3D object tracking approach based on convolutional neural network and Monte Carlo algorithms used for a pick and place robot”. In: *Photonics and Education in Measurement Science 2019*. Vol. 11144. International Society for Optics and Photonics. 2019, p. 1114414 (cit. on pp. 21, 22).

[51] Huy Nguyen, Nicholas Adrian, Joyce Lim Xin Yan, et al. “Development of a Robotic System for Automated Decaking of 3D-Printed Parts”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, 8202–8208 (cit. on pp. 21, 23).

[52]Amit Kumar Bedaka, Joel Vidal, and Chyi-Yeu Lin. “Automatic robot path integration using three-dimensional vision and offline programming”. In: *Int. J. Adv. Manuf. Technol* 102.5 (2019), 1935–1950 (cit. on pp. 21, 22).

[53]Paolo Franceschi, Nicola Castaman, Stefano Ghidoni, and Nicola Pedrocchi. “Precise Robotic Manipulation of Bulky Components”. In: *IEEE Access* 8 (2020), 222476–222485 (cit. on pp. 21, 22).

[54]Morteza Shariatee, Hooman Khosravi, and Ehsan Fazl-Ersi. “Safe collaboration of humans and SCARA robots”. In: *2016 4th International Conference on Robotics and Mechatronics (ICROM)*. IEEE. 2016, 589–594 (cit. on pp. 24, 25).

[55]Tariq Tashtoush, Luis Garcia, Gerardo Landa, et al. “Human-Robot Interaction and Collaboration (HRI-C) Utilizing Top-View RGB-D Camera System”. In: *Int. J. Adv. Comput. Sci. Appl.* 12.1 (2021) (cit. on pp. 24, 25).

[56]Hongyi Liu, Yuquan Wang, Wei Ji, and Lihui Wang. “A context-aware safety system for human-robot collaboration”. In: *Procedia Manuf.* 17 (2018), 238–245 (cit. on pp. 24, 25).

[57]Mohammad Anvaripour and Mehrdad Saif. “Collision detection for human-robot interaction in an industrial setting using force myography and a deep learning approach”. In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE. 2019, 2149–2154 (cit. on p. 25).

[58]Fan Bu, Trinh Le, Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. “Pedestrian planar LiDAR pose (PPLP) network for oriented pedestrian detection based on planar LiDAR and monocular images”. In: *IEEE Robot. Autom. Lett.* 5.2 (2019), 1626–1633 (cit. on p. 25).

[59]Redhwan Algabri and Mun-Taek Choi. “Deep-learning-based indoor human following of mobile robot using color feature”. In: *Sensors* 20.9 (2020), p. 2699 (cit. on p. 25).

[60]Guntitat Sawadwuthikul, Tanyatep Tothong, Thanawat Lodkaew, et al. “Visual Goal Human-Robot Communication Framework with Few-Shot Learning: a Case Study in Robot Waiter System”. In: *IEEE Trans. Ind. Inf.* (2021) (cit. on p. 25).

[61]Lei Shi, Cosmin Copot, and Steve Vanlanduit. “A bayesian deep neural network for safe visual servoing in human–robot interaction”. In: *Front. Rob. AI* 8 (2021), p. 165 (cit. on p. 25).

[62]Ha Manh Do, Craig Mouser, Meiqin Liu, and Weihua Sheng. “Human-robot collaboration in a mobile visual sensor network”. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, 2203–2208 (cit. on pp. 25, 26).

[63] Olatz De Miguel Lázaro, Wael M Mohammed, Borja Ramis Ferrer, Ronal Bejarano, and Jose L Martinez Lastra. "An Approach for adapting a Cobot Workstation to Human Operator within a Deep Learning Camera". In: *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. Vol. 1. IEEE. 2019, 789–794 (cit. on pp. 25, 26).

[64] Chih-Lyang Hwang, Ding-Sheng Wang, Fan-Chen Weng, and Sheng-Lin Lai. "Interactions Between Specific Human and Omnidirectional Mobile Robot Using Deep Learning Approach: SSD-FN-KCF". In: *IEEE Access* 8 (2020), 41186–41200 (cit. on pp. 25, 26).

[65] Marc-André Fiedler, Philipp Werner, Aly Khalifa, and Ayoub Al-Hamadi. "SFPD: Simultaneous Face and Person Detection in Real-Time for Human–Robot Interaction". In: *Sensors* 21.17 (2021), p. 5918 (cit. on pp. 25, 26).

[66] Qianqian Xiong, Jianjing Zhang, Peng Wang, Dongdong Liu, and Robert X Gao. "Transferable two-stream convolutional neural network for human action recognition". In: *J. Manuf. Syst.* 56 (2020), 605–614 (cit. on pp. 27, 28).

[67] Xianhe Wen and Heping Chen. "3D long-term recurrent convolutional networks for human sub-assembly recognition in human-robot collaboration". In: *Assembly Autom.* (2020) (cit. on pp. 28, 29).

[68] Hazem Abdelkawy, Naouel Ayari, Abdelghani Chibani, Yacine Amirat, and Ferhat Attal. "Spatio-Temporal Convolutional Networks and N-Ary Ontologies for Human Activity-Aware Robotic System". In: *IEEE Robot. Autom. Lett.* 6.2 (2020), 620–627 (cit. on pp. 28, 29).

[69] Alberto Sabater, Iñigo Alonso, Luis Montesano, and Ana C. Murillo. "Domain and View-Point Agnostic Hand Action Recognition". In: *IEEE Robot. Autom. Lett.* 6.4 (2021), pp. 7823–7830 (cit. on pp. 28, 29).

[70] Ali Ghadirzadeh, Xi Chen, Wenjie Yin, et al. "Human-Centered Collaborative Robots With Deep Reinforcement Learning". In: *IEEE Robot. Autom. Lett.* 6.2 (2021), pp. 566–571 (cit. on pp. 28, 29).

[71] Fatemeh Mohammadi Amin, Maryam Rezayati, Hans Wernher van de Venn, and Hossein Karimpour. "A mixed-perception approach for safe human–robot collaboration in industrial automation". In: *Sensors* 20.21 (2020), p. 6347 (cit. on pp. 28, 29).

[72] Taizo Yoshikawa, Viktor Losing, and Emel Demircan. "Machine learning for human movement understanding". In: *Adv. Robotics* 34.13 (2020), 828–844 (cit. on pp. 28, 29).

[73] Chiuhsiang Joe Lin and Rio Prasetyo Lukodono. "Sustainable Human–Robot Collaboration Based on Human Intention Classification". In: *Sustainability* 13.11 (2021), p. 5990 (cit. on pp. 28, 29).

[74] Md Mofijul Islam and Tariq Iqbal. "Multi-Gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition". In: *IEEE Robot. Autom. Lett.* 6.2 (2021), 1729–1736 (cit. on pp. 28, 29).

[75] Tsige Tadesse Alemayoh, Jae Hoon Lee, and Shingo Okamoto. "New Sensor Data Structuring for Deeper Feature Extraction in Human Activity Recognition". In: *Sensors* 21.8 (2021), p. 2814 (cit. on pp. 28, 30).

[76] Kyeong-Beom Park, Sung Ho Choi, Jae Yeol Lee, et al. "Hands-Free Human–Robot Interaction Using Multimodal Gestures and Deep Learning in Wearable Mixed Reality". In: *IEEE Access* 9 (2021), 55448–55464 (cit. on pp. 28, 30).

[77] Zitong Liu, Quan Liu, Wenjun Xu, et al. "Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing". In: *Procedia CIRP* 83 (2019), 272–278 (cit. on pp. 28, 30).

[78] Edoardo Alati, Lorenzo Mauro, Valsamis Ntouskos, and Fiora Pirri. "Help by predicting what to do". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, 1930–1934 (cit. on pp. 28, 30).

[79] Shafina Bibi, Nadeem Anjum, Tehmina Amjad, Graeme McRobbie, and Naeem Ramzan. "Human Interaction Anticipation by Combining Deep Features and Transformed Optical Flow Components". In: *IEEE Access* 8 (2020), 137646–137657 (cit. on pp. 28, 30).

[80] Jianjing Zhang, Peng Wang, and Robert X Gao. "Hybrid machine learning for human action recognition and prediction in assembly". In: *Rob. Comput. Integr. Manuf.* 72 (2021), p. 102184 (cit. on pp. 28, 30).

[81] Jim Mainprice and Dmitry Berenson. "Human-robot collaborative manipulation planning using early prediction of human motion". In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2013, 299–306 (cit. on pp. 28, 31).

[82] Hema S Koppula and Ashutosh Saxena. "Anticipating human activities using object affordances for reactive robotic response". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.1 (2015), 14–29 (cit. on pp. 28, 31).

[83] Hongyi Liu and Lihui Wang. "Human motion prediction for human-robot collaboration". In: *J. Manuf. Syst.* 44 (2017), 287–294 (cit. on pp. 28, 31).

[84] AbdElMoniem Bayoumi, Philipp Karkowski, and Maren Bennewitz. "Learning foresighted people following under occlusions". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, 6319–6324 (cit. on pp. 28, 31).

[85] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. "Anticipating many futures: Online human motion prediction and generation for human-robot interaction". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, 4563–4570 (cit. on pp. 28, 31).

[86] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, et al. "Teaching robots to predict human motion". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, 562–567 (cit. on pp. 28, 31).

[87] Xuan Zhao, Sakmongkon Chumkamon, Shuanda Duan, Juan Rojas, and Jia Pan. "Collaborative human-robot motion generation using LSTM-RNN". In: *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE. 2018, 1–9 (cit. on pp. 28, 31).

[88] Philipp Kratzer, Niteesh Balachandra Midlagajni, Marc Toussaint, and Jim Mainprice. "Anticipating human intention for full-body motion prediction in object grasping and placing tasks". In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2020, 1157–1163 (cit. on pp. 28, 31).

[89] Jianjing Zhang, Hongyi Liu, Qing Chang, Lihui Wang, and Robert X Gao. "Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly". In: *CIRP Ann.* 69.1 (2020), 9–12 (cit. on pp. 28, 31).

[90] Judith Bütepage, Ali Ghadirzadeh, Özge Öztimur Karadağ, Mårten Björkman, and Danica Kragic. "Imitating by generating: Deep generative models for imitation of interactive tasks". In: *Front. Rob. AI* 7 (2020), p. 47 (cit. on pp. 28, 31).

[91] Boris Ivanovic, Karen Leung, Edward Schmerling, and Marco Pavone. "Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach". In: *IEEE Robot. Autom. Lett.* 6.2 (2020), 295–302 (cit. on pp. 28, 31).

[92] Won-Hyong Lee and Jong-Hwan Kim. "Hierarchical emotional episodic memory for social human robot collaboration". In: *Auton. Robot.* 42.5 (2018), 1087–1102 (cit. on pp. 28, 32).

[93] Behnoosh Parsa, Ekta U Samani, Rose Hendrix, et al. "Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks". In: *IEEE Robot. Autom. Lett.* 4.4 (2019), 3153–3160 (cit. on pp. 28, 32, 84).

[94] Partha Chakraborty, Sabbir Ahmed, Mohammad Abu Yousuf, et al. "A Human-Robot Interaction System Calculating Visual Focus of Human's Attention Level". In: *IEEE Access* 9 (2021), pp. 93409–93421 (cit. on pp. 28, 32).

[95] Xiang Shi, You Yang, and Qiong Liu. "I Understand You: Blind 3D Human Attention Inference From the Perspective of Third-Person". In: *IEEE Trans. Image Process.* 30 (2021), 6212–6225 (cit. on pp. 28, 32).

[96] Quan Liu, Zhihao Liu, Wenjun Xu, et al. “Human-robot collaboration in disassembly for sustainable manufacturing”. In: *Int. J. Prod. Res.* 57.12 (2019), 4027–4044 (cit. on pp. 33, 34, 53).

[97] Daniel Kruse, Richard J Radke, and John T Wen. “A sensor-based dual-arm tele-robotic manipulation platform”. In: *2013 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE. 2013, 350–355 (cit. on pp. 33, 34).

[98] Hongyi Liu and Lihui Wang. “Collision-free human-robot collaboration based on context awareness”. In: *Rob. Comput. Integr. Manuf.* 67 (2021), p. 101997 (cit. on pp. 33, 34, 37, 39, 84, 106).

[99] Marike K van den Broek and Thomas B Moeslund. “Ergonomic Adaptation of Robotic Movements in Human-Robot Collaboration”. In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2020, 499–501 (cit. on pp. 33, 34).

[100] Guan Li, Zhifeng Liu, Ligang Cai, and Jun Yan. “Standing-Posture Recognition in Human–Robot Collaboration Based on Deep Learning and the Dempster–Shafer Evidence Theory”. In: *Sensors* 20.4 (2020), p. 1158 (cit. on pp. 33, 34).

[101] Hao Zhong, Juan P Wachs, and Shimon Y Nof. “A collaborative telerobotics network framework with hand gesture interface and conflict prevention”. In: *Int. J. Prod. Res.* 51.15 (2013), 4443–4463 (cit. on pp. 34, 35).

[102] Mithun G Jacob, Yu-Ting Li, and Juan P Wachs. “Surgical instrument handling and retrieval in the operating room with a multimodal robotic assistant”. In: *2013 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2013, 2140–2145 (cit. on pp. 34, 35).

[103] Mehmet Celalettin Ergene, Akif Durdu, and Halil Cetin. “Imitation and learning of human hand gesture tasks of the 3D printed robotic hand by using artificial neural networks”. In: *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE. 2016, 1–6 (cit. on pp. 34, 35).

[104] Lei Chen, Haiwei Yang, and Pei Liu. “Intelligent Robot Arm: Vision-based dynamic measurement system for industrial applications”. In: *International Conference on Intelligent Robotics and Applications (ICIRA)*. Springer. 2019, 120–130 (cit. on pp. 34, 35).

[105] Rose Hendrix, Parker Owan, Joseph Garbini, and Santosh Devasia. “Context-Specific Separable Gesture Selection for Control of a Robotic Manufacturing Assistant”. In: *IFAC-PapersOnLine* 51.34 (2019), 89–96 (cit. on pp. 34, 35).

[106] Xuexiang Zhang and Xuncheng Wu. “Robotic control of dynamic and static gesture recognition”. In: *2019 2nd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM)*. IEEE. 2019, 474–478 (cit. on pp. 34, 35).

[107] Joseph DelPreto, Andres F Salazar-Gomez, Stephanie Gil, et al. “Plug-and-play supervisory control using muscle and brain signals for real-time gesture and error detection”. In: *Auton. Robot.* 44.7 (2020), 1303–1322 (cit. on pp. 34, 35).

[108] Hongyi Liu, Tongtong Fang, Tianyu Zhou, and Lihui Wang. “Towards robust human-robot collaborative manufacturing: multimodal fusion”. In: *IEEE Access* 6 (2018), 74762–74771 (cit. on pp. 34, 35, 53).

[109] Qing Gao, Jinguo Liu, Zhaojie Ju, and Xin Zhang. “Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation”. In: *IEEE Trans. Ind. Electron.* 66.12 (2019), 9663–9672 (cit. on pp. 34, 35).

[110] Cristina Nuzzi, Simone Pasinetti, Roberto Pagani, Franco Docchio, and Giovanna Sansoni. “Hand gesture recognition for collaborative workstations: A smart command system prototype”. In: *International Conference on Image Analysis and Processing (ICIAP)*. Springer. 2019, 332–342 (cit. on pp. 34, 36).

[111] Osama Mazhar, Benjamin Navarro, Sofiane Ramdani, Robin Passama, and Andrea Cherubini. “A real-time human-robot interaction framework with robust background invariant hand gesture detection”. In: *Rob. Comput. Integr. Manuf.* 60 (2019), 34–48 (cit. on p. 34).

[112] Xing Li. “Human–robot interaction based on gesture and movement recognition”. In: *Signal Process. Image Commun.* 81 (2020), p. 115686 (cit. on pp. 34, 36).

[113] Qing Gao, Jinguo Liu, and Zhaojie Ju. “Robust real-time hand detection and localization for space human–robot interaction based on deep learning”. In: *Neurocomputing* 390 (2020), 198–206 (cit. on pp. 34, 36).

[114] Wenjin Zhang, Jiacun Wang, and Fangping Lan. “Dynamic hand gesture recognition based on short-term sampling neural networks”. In: *IEEE/CAA J. Autom. Sin.* 8.1 (2020), 110–120 (cit. on pp. 34, 35).

[115] Osama Mazhar, Sofiane Ramdani, and Andrea Cherubini. “A Deep Learning Framework for Recognizing Both Static and Dynamic Gestures”. In: *Sensors* 21.6 (2021), p. 2227 (cit. on pp. 34, 35).

[116] Bi-Xiao Wu, Chen-Guang Yang, and Jun-Pei Zhong. “Research on transfer learning of vision-based gesture recognition”. In: *Int. J. Autom. Comput.* 18.3 (2021), 422–431 (cit. on pp. 34, 36).

[117] Paras Gulati, Qin Hu, and S Farokh Atashzar. “Toward Deep Generalization of Peripheral EMG-Based Human-Robot Interfacing: A Hybrid Explainable Solution for NeuroRobotic Systems”. In: *IEEE Robot. Autom. Lett.* 6.2 (2021), 2650–2657 (cit. on pp. 34, 36).

[118] Cristina Nuzzi, Simone Pasinetti, Roberto Pagani, et al. “MEGURU: a gesture-based robot program builder for Meta-Collaborative workstations”. In: *Rob. Comput. Integr. Manuf.* 68 (2021), p. 102085 (cit. on pp. 34, 36).

[119] Laura Fiorini, Federica G Cornacchia Loizzo, Alessandra Sorrentino, et al. “Daily gesture recognition during human-robot interaction combining vision and wearable systems”. In: *IEEE Sensors J.* (2021) (cit. on pp. 34, 36).

[120] Wen Qi, Salih Ertug Ovur, Zhijun Li, Aldo Marzullo, and Rong Song. “Multi-sensor Guided Hand Gestures Recognition for Teleoperated Robot using Recurrent Neural Network”. In: *IEEE Robot. Autom. Lett.* (2021) (cit. on pp. 34, 36).

[121] Sebastian Blumenthal, Herman Bruyninckx, Walter Nowak, and Erwin Prassler. “A scene graph based shared 3D world model for robotic applications”. In: *2013 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2013, 453–460 (cit. on p. 37).

[122] Jiyoun Moon and Beomhee Lee. “Scene understanding using natural language description based on 3D semantic graph map”. In: *Intell. Serv. Robot.* 11.4 (2018), 347–354 (cit. on p. 37).

[123] Alberto Hata, Rafia Inam, Klaus Raizer, Shaolei Wang, and Enyu Cao. “AI-based safety analysis for collaborative mobile robots”. In: *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE. 2019, 1722–1729 (cit. on p. 37).

[124] Hassam Riaz, Ahmad Terra, Klaus Raizer, Rafia Inam, and Alberto Hata. “Scene Understanding for Safety Analysis in Human-Robot Collaborative Operations”. In: *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE. 2020, 722–731 (cit. on pp. 37, 38).

[125] Yiyi Liao, Sarath Kodagoda, Yue Wang, Lei Shi, and Yong Liu. “Place classification with a graph regularized deep neural network”. In: *IEEE Trans. Cognit. Dev. Syst.* 9.4 (2016), 304–315 (cit. on pp. 37, 38).

[126] Markus Hiller, Chen Qiu, Florian Particke, Christian Hofmann, and Jörn Thielecke. “Learning topometric semantic maps from occupancy grids”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, 4190–4197 (cit. on pp. 37, 38).

[127] Zhe Hu, Jia Pan, Tingxiang Fan, Ruigang Yang, and Dinesh Manocha. “Safe navigation with human instructions in complex scenes”. In: *IEEE Robot. Autom. Lett.* 4.2 (2019), 753–760 (cit. on pp. 37, 38, 106).

[128] Adhitha Dias, Hasitha Wellaboda, Yasod Rasanka, et al. “Deep Learning of Augmented Reality based Human Interactions for Automating a Robot Team”. In: *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE. 2020, 175–182 (cit. on pp. 37, 38).

[129] Christian Friedrich, Akos Csiszar, Armin Lechler, and Alexander Verl. “Fast robot task and path planning based on cad and vision data”. In: *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE. 2017, 1633–1638 (cit. on pp. 37, 39).

[130] Amine Abou Moughlbay, Héctor Herrero, Raquel Pacheco, Jose Luis Outón, and Damien Sallé. “Reliable workspace monitoring in safe human-robot environment”. In: *International Joint Conference SOCO’16-CISIS’16-ICEUTE’16*. Springer. 2016, 256–266 (cit. on pp. 37, 39).

[131] Tianyu Zhou, Qi Zhu, and Jing Du. “Intuitive robot teleoperation for civil engineering operations with virtual reality and deep learning scene reconstruction”. In: *Adv. Eng. Inform.* 46 (2020), p. 101170 (cit. on pp. 37, 39).

[132] Quan Liu, Zhihao Liu, Bo Xiong, Wenjun Xu, and Yang Liu. “Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function”. In: *Adv. Eng. Inform.* 49 (2021), p. 101360 (cit. on pp. 37, 39).

[133] Juraj Slovák, Markus Melicher, Matej Šimovec, and Ján Vachálek. “Vision and RTLS Safety Implementation in an Experimental Human—Robot Collaboration Scenario”. In: *Sensors* 21.7 (2021), p. 2419 (cit. on pp. 37, 40).

[134] Sung Ho Choi, Kyeong-Beom Park, Dong Hyeon Roh, et al. “An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation”. In: *Rob. Comput. Integr. Manuf.* 73 (2022), p. 102258 (cit. on pp. 37, 40, 106).

[135] SM Mizanoor Rahman, Zhanrui Liao, Longsheng Jiang, and Yue Wang. “A regret-based autonomy allocation scheme for human-robot shared vision systems in collaborative assembly in manufacturing”. In: *2016 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE. 2016, 897–902 (cit. on pp. 40, 41).

[136] Shingo Murata, Wataru Masuda, Jiayi Chen, et al. “Achieving Human–Robot Collaboration with Dynamic Goal Inference by Gradient Descent”. In: *International Conference on Neural Information Processing (ICONIP)*. Springer. 2019, 579–590 (cit. on pp. 40, 41).

[137] Sagar Gubbi Venkatesh, Raviteja Upadrashta, Shishir Kolathaya, and Bharadwaj Amrutur. “Teaching Robots Novel Objects by Pointing at Them”. In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2020, 1101–1106 (cit. on pp. 40, 41).

[138] Yi Sun, Weitian Wang, Yi Chen, and Yunyi Jia. “Learn How to Assist Humans Through Human Teaching and Robot Learning in Human-Robot Collaborative Assembly”. In: *IEEE Trans. Syst. Man Cybern.: Syst.* (2020) (cit. on pp. 40, 42).

[139] Catalina Roncancio, Jose L Rodríguez, Eduardo Zalama, et al. “Improvement in service robot’s interaction through case based reasoning”. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2010, 1–7 (cit. on pp. 40, 42).

[140] Ryosuke Kojima, Osamu Sugiyama, and Kazuhiro Nakadai. “Audio-visual scene understanding utilizing text information for a cooking support robot”. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, 4210–4215 (cit. on p. 40).

[141] Bradley Hayes and Julie A Shah. “Improving robot controller transparency through autonomous policy explanation”. In: *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2017, 303–312 (cit. on pp. 40, 42).

[142] Hyemin Ahn, Sungjoon Choi, Nuri Kim, Geonho Cha, and Songhwai Oh. “Interactive text2pickup networks for natural language-based human–robot collaboration”. In: *IEEE Robot. Autom. Lett.* 3.4 (2018), 3308–3315 (cit. on pp. 40, 42).

[143] Sagar Gubbi Venkatesh, Anirban Biswas, Raviteja Upadrashta, et al. “Spatial Reasoning from Natural Language Instructions for Robot Manipulation”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2021, pp. 11196–11202 (cit. on pp. 40, 43).

[144] Heather Riley and Mohan Sridharan. “Integrating non-monotonic logical reasoning and inductive learning with deep learning for explainable visual question answering”. In: *Front. Rob. AI* 6 (2019), p. 125 (cit. on pp. 40, 43).

[145] Hui Li Tan, Mei Chee Leong, Qianli Xu, et al. “Task-Oriented Multi-Modal Question Answering For Collaborative Applications”. In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, 1426–1430 (cit. on pp. 40, 43).

[146] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. “Multi-View Visual Question Answering with Active Viewpoint Selection”. In: *Sensors* 20.8 (2020), p. 2281 (cit. on pp. 40, 43).

[147] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. “Densepose: Dense human pose estimation in the wild”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, 7297–7306 (cit. on p. 45).

[148] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. “Learning 3D Human Shape and Pose from Dense Body Parts”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2020) (cit. on p. 45).

[149] Liuhao Ge, Zhou Ren, Yuncheng Li, et al. “3d hand shape and pose estimation from a single rgb image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, 10833–10842 (cit. on p. 45).

[150] Jun Lv, Wenqiang Xu, Lixin Yang, et al. “HandTailor: Towards High-Precision Monocular 3D Hand Recovery”. In: *The 32nd British Machine Vision Conference (BMVC)*. 2021 (cit. on p. 45).

[151] Junming Fan, Shufei Li, Pai Zheng, and Carman K.M. Lee. “A high-resolution network-based approach for 6D pose estimation of industrial parts”. In: *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. 2021, pp. 1452–1457 (cit. on p. 49).

[152] Junming Fan, Pai Zheng, Shufei Li, and Lihui Wang. “An integrated hand-object dense pose estimation approach with explicit occlusion awareness for human-robot collaborative disassembly”. In: *IEEE Transactions on Automation Science and Engineering* (2022) (cit. on p. 49).

[153] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask r-cnn”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969 (cit. on p. 51).

[154] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. “Model globally, match locally: Efficient and robust 3D object recognition”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 998–1005 (cit. on pp. 51, 52).

[155] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. “3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints”. In: *International Journal of Computer Vision* 66.3 (2006), pp. 231–259 (cit. on p. 51).

[156] Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, et al. “Gradient response maps for real-time detection of textureless objects”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.5 (2011), pp. 876–888 (cit. on p. 51).

[157] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, et al. “Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes”. In: *Asian Conference on Computer Vision*. Springer. 2012, pp. 548–562 (cit. on pp. 51, 52).

[158] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. “PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes”. In: *Robotics: Science and Systems (RSS)*. 2018 (cit. on pp. 51, 57, 58, 60, 73).

[159] Mahdi Rad and Vincent Lepetit. “Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3828–3836 (cit. on pp. 51, 52).

[160] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. “Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1521–1529 (cit. on p. 51).

[161] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations*. 2015 (cit. on pp. 52, 58).

[162] Christian Szegedy, Wei Liu, Yangqing Jia, et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9 (cit. on pp. 52, 58).

[163] Jingdong Wang, Ke Sun, Tianheng Cheng, et al. “Deep high-resolution representation learning for visual recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020) (cit. on pp. 52, 58, 86).

[164] Chen Wang, Danfei Xu, Yuke Zhu, et al. “Densefusion: 6d object pose estimation by iterative dense fusion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3343–3352 (cit. on p. 52).

[165] Marco Costanzo, Giuseppe De Maria, Gaetano Lettera, and Ciro Natale. “A Multimodal Approach to Human Safety in Collaborative Robotic Workcells”. In: *IEEE Transactions on Automation Science and Engineering* (2021), pp. 1–15 (cit. on p. 53).

[166] Weitian Wang, Rui Li, Yi Chen, Yi Sun, and Yunyi Jia. “Predicting Human Intentions in Human-Robot Hand-Over Tasks Through Multimodal Learning”. In: *IEEE Transactions on Automation Science and Engineering* (2021), pp. 1–15 (cit. on p. 53).

[167] Patrick Rosenberger, Akansel Cosgun, Rhys Newbury, et al. “Object-independent human-to-robot handovers using real time robotic vision”. In: *IEEE Robotics and Automation Letters* 6.1 (2020), pp. 17–23 (cit. on p. 53).

[168] Ruiya Li, Duc Truong Pham, Jun Huang, et al. “Unfastening of hexagonal headed screws by a collaborative robot”. In: *IEEE Transactions on Automation Science and Engineering* 17.3 (2020), pp. 1455–1468 (cit. on p. 53).

[169] Huixu Dong, Dilip K. Prasad, and I-Ming Chen. “Object Pose Estimation via Pruned Hough Forest With Combined Split Schemes for Robotic Grasp”. In: *IEEE Transactions on Automation Science and Engineering* 18.4 (2021), pp. 1814–1821 (cit. on p. 53).

[170] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. "Human De-occlusion: Invisible Perception and Recovery for Humans". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3691–3701 (cit. on p. 54).

[171] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. "Deep Occlusion-Aware Instance Segmentation with Overlapping BiLayers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4019–4028 (cit. on p. 54).

[172] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Advances in Neural Information Processing Systems*. 2015 (cit. on pp. 55, 60).

[173] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. "Deepim: Deep iterative matching for 6d pose estimation". In: *Proceedings of the European Conference on Computer Vision*. 2018, pp. 683–698 (cit. on p. 57).

[174] Alex Kendall, Matthew Grimes, and Roberto Cipolla. "Posenet: A convolutional network for real-time 6-dof camera relocalization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2938–2946 (cit. on p. 58).

[175] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778 (cit. on pp. 62, 64, 65).

[176] Javier Romero, Dimitrios Tzionas, and Michael J Black. "Embodied hands: Modeling and capturing hands and bodies together". In: *ACM Transactions on Graphics (ToG)* 36.6 (2017), pp. 1–17 (cit. on pp. 63, 68).

[177] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. "Neural 3d mesh renderer". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3907–3916 (cit. on pp. 63, 71).

[178] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2117–2125 (cit. on pp. 63, 71).

[179] Yana Hasson, Bugra Tekin, Federica Bogo, et al. "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 571–580 (cit. on pp. 64, 66, 68, 77–80).

[180] Chenrui Wu, Long Chen, Zaixing He, and Junjie Jiang. "Pseudo-Siamese Graph Matching Network for Textureless Objects' 6D Pose Estimation". In: *IEEE Transactions on Industrial Electronics* (2021) (cit. on p. 69).

[181] Hui Zhang, Zhicong Liang, Chen Li, et al. “A Practical Robotic Grasping Method by Using 6D Pose Estimation with Protective Correction”. In: *IEEE Transactions on Industrial Electronics* (2021) (cit. on p. 69).

[182] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. “CosyPose: Consistent multi-view multi-object 6D pose estimation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 574–591 (cit. on pp. 73, 74).

[183] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114 (cit. on p. 73).

[184] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. “First-person hand action benchmark with rgb-d videos and 3d hand pose annotations”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 409–419 (cit. on p. 79).

[185] Bugra Tekin, Federica Bogo, and Marc Pollefeys. “H+ o: Unified egocentric recognition of 3d hand-object poses and interactions”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4511–4520 (cit. on pp. 79, 80).

[186] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. “HOT-Net: Non-autoregressive transformer for 3D hand-object pose estimation”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 3136–3145 (cit. on pp. 79, 80).

[187] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. “Hope-net: A graph-based model for hand-object pose estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6608–6617 (cit. on pp. 79, 80).

[188] Junming Fan, Pai Zheng, and Carman KM Lee. “A Vision-based Human Digital Twin Modelling Approach for Adaptive Human-Robot Collaboration”. In: *Journal of Manufacturing Science and Engineering* (2023), pp. 1–11 (cit. on p. 83).

[189] Maija Breque, Lars De Nul, and Athanasios Petridis. “Industry 5.0: towards a sustainable, human-centric and resilient European industry”. In: *Luxembourg, LU: European Commission, Directorate-General for Research and Innovation* (2021) (cit. on p. 84).

[190] Shufei Li, Pai Zheng, Sichao Liu, et al. “Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives”. In: *Robotics and Computer-Integrated Manufacturing* 81 (2023), p. 102510 (cit. on p. 84).

[191] Junming Fan, Pai Zheng, and Shufei Li. “Vision-based Holistic Scene Understanding Towards Proactive Human-Robot Collaboration”. In: *Robotics and Computer-Integrated Manufacturing* Accepted (2022) (cit. on pp. 84, 105, 122).

[192] Weitian Wang, Rui Li, Zachary Max Diekel, et al. “Controlling object hand-over in human–robot collaboration via natural wearable sensing”. In: *IEEE Transactions on Human-Machine Systems* 49.1 (2018), pp. 59–71 (cit. on p. 84).

[193] Lorenzo Vianello, Jean-Baptiste Mouret, Eloïse Dalin, Alexis Aubry, and Serena Ivaldi. “Human posture prediction during physical human-robot interaction”. In: *IEEE Robotics and Automation Letters* 6.3 (2021), pp. 6046–6053 (cit. on p. 84).

[194] Baicun Wang, Huiying Zhou, Geng Yang, Xingyu Li, and Huayong Yang. “Human Digital Twin (HDT) Driven Human-Cyber-Physical Systems: Key Technologies and Applications”. In: *Chinese Journal of Mechanical Engineering* 35.1 (2022), pp. 1–6 (cit. on p. 85).

[195] Michael E Miller and Emily Spatz. “A unified view of a human digital twin”. In: *Human-Intelligent Systems Integration* (2022), pp. 1–11 (cit. on p. 85).

[196] Wei Shengli. “Is human digital twin possible?” In: *Computer Methods and Programs in Biomedicine Update* 1 (2021), p. 100014 (cit. on p. 85).

[197] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. “SMPL: A Skinned Multi-Person Linear Model”. In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16 (cit. on p. 88).

[198] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. “End-to-end recovery of human shape and pose”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7122–7131 (cit. on pp. 88, 98).

[199] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. “PARE: Part attention regressor for 3D human body estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11127–11137 (cit. on pp. 88, 98).

[200] Sue Hignett and Lynn McAtamney. “Rapid entire body assessment (REBA)”. In: *Applied ergonomics* 31.2 (2000), pp. 201–205 (cit. on p. 91).

[201] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015) (cit. on p. 91).

[202] Michalis Foukarakis, Asterios Leonidis, Margherita Antona, and Constantine Stephanidis. “Combining finite state machine and decision-making tools for adaptable robot behavior”. In: *International Conference on Universal Access in Human-Computer Interaction*. Springer. 2014, pp. 625–635 (cit. on p. 94).

[203] Puze Liu, Kuo Zhang, Davide Tateo, et al. “Regularized Deep Signed Distance Fields for Reactive Motion Generation”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022, pp. 6673–6680 (cit. on p. 95).

[204] Shanyan Guan, Jingwei Xu, Michelle Z He, et al. “Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022) (cit. on p. 97).

[205] Sijie Yan, Yuanjun Xiong, and Dahua Lin. “Spatial temporal graph convolutional networks for skeleton-based action recognition”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018 (cit. on p. 98).

[206] Behnoosh Parsa and Ashis G Banerjee. “A multi-task learning approach for human activity segmentation and ergonomics risk assessment”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2352–2362 (cit. on p. 98).

[207] Junming Fan, Pai Zheng, and Carman KM Lee. “A multi-granularity scene segmentation network for human-robot collaboration environment perception”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 2105–2110 (cit. on p. 103).

[208] Stefan Escaida Navarro, Stephan Mühlbacher-Karrer, Hosam Alagi, et al. “Proximity perception in human-centered robotics: A survey on sensing systems and applications”. In: *IEEE Transactions on Robotics* (2021) (cit. on p. 104).

[209] Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. “Semantic information for robot navigation: A survey”. In: *Applied Sciences* 10.2 (2020), p. 497 (cit. on p. 104).

[210] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. “Recent progress in semantic image segmentation”. In: *Artificial Intelligence Review* 52.2 (2019), pp. 1089–1106 (cit. on p. 104).

[211] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al. “A ConvNet for the 2020s”. In: *arXiv preprint arXiv:2201.03545* (2022) (cit. on pp. 105, 107, 109, 114).

[212] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. “Indoor segmentation and support inference from rgbd images”. In: *European Conference on Computer Vision (ECCV)*. Springer. 2012, pp. 746–760 (cit. on pp. 105, 108, 116).

[213] Daniel J Butler, Sarah Elliot, and Maya Cakmak. “Interactive scene segmentation for efficient human-in-the-loop robot manipulation”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 2572–2579 (cit. on p. 106).

[214] Loic Landrieu and Martin Simonovsky. “Large-scale point cloud semantic segmentation with superpoint graphs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, 4558–4567 (cit. on p. 106).

[215] Christopher Choy, JunYoung Gwak, and Silvio Savarese. “4d spatio-temporal convnets: Minkowski convolutional neural networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3075–3084 (cit. on p. 106).

[216] Jinming Cao, Hanchao Leng, Dani Lischinski, et al. “ShapeConv: Shape-aware Convolutional Layer for Indoor RGB-D Semantic Segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 7088–7097 (cit. on pp. 107, 117).

[217] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. “Efficient rgb-d semantic segmentation for indoor scene analysis”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 13525–13531 (cit. on pp. 107, 109, 117).

[218] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, et al. “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation”. In: *European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 561–577 (cit. on pp. 107, 117).

[219] Ze Liu, Yutong Lin, Yue Cao, et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*. 2021, pp. 10012–10022 (cit. on p. 107).

[220] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. “Unified perceptual parsing for scene understanding”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 418–434 (cit. on p. 107).

[221] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. “Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 675–684 (cit. on pp. 107, 117).

[222] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. “Mti-net: Multi-scale task interaction networks for multi-task learning”. In: *European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 527–543 (cit. on pp. 107, 116, 117).

[223] Bolei Zhou, Hang Zhao, Xavier Puig, et al. “Semantic understanding of scenes through the ade20k dataset”. In: *International Journal of Computer Vision* 127.3 (2019), pp. 302–321 (cit. on p. 108).

[224] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7132–7141 (cit. on p. 110).

[225] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2881–2890 (cit. on p. 111).

[226] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation". In: *IEEE Transactions on Intelligent Transportation Systems* 19.1 (2017), pp. 263–272 (cit. on p. 111).

[227] Guosheng Lin, Fayao Liu, Anton Milan, Chunhua Shen, and Ian Reid. "Refinenet: Multi-path refinement networks for dense prediction". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.5 (2019), pp. 1228–1242 (cit. on p. 117).

[228] Di Lin and Hui Huang. "Zig-zag network for semantic segmentation of RGB-D images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.10 (2019), pp. 2642–2655 (cit. on p. 117).

[229] Junming Fan and Pai Zheng. "A vision-language reasoning approach for ambiguity mitigation in human-robot collaborative manufacturing". In: *Journal of Manufacturing Systems* (2024) (cit. on p. 121).

[230] Shufei Li, Ruobing Wang, Pai Zheng, and Lihui Wang. "Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm". In: *J. Manuf. Syst.* 60 (2021), 547–552 (cit. on p. 122).

[231] Lihui Wang. "A futuristic perspective on human-centric assembly". In: *Journal of Manufacturing Systems* 62 (2022), pp. 199–201 (cit. on p. 122).

[232] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, et al. "Progress and prospects of the human–robot collaboration". In: *Autonomous Robots* 42 (2018), pp. 957–975 (cit. on p. 122).

[233] Yongshi Liang, Pai Zheng, and Liqiao Xia. "A visual reasoning-based approach for driving experience improvement in the AR-assisted head-up displays". In: *Advanced Engineering Informatics* 55 (2023), p. 101888 (cit. on p. 122).

[234] Shufei Li, Pai Zheng, Shibao Pang, Xi Vincent Wang, and Lihui Wang. "Self-organising multiple human–robot collaboration: A temporal subgraph reasoning-based method". In: *Journal of Manufacturing Systems* 68 (2023), pp. 304–312 (cit. on p. 122).

[235] Yue Yin, Pai Zheng, Chengxi Li, and Lihui Wang. “A state-of-the-art survey on Augmented Reality-assisted Digital Twin for futuristic human-centric industry transformation”. In: *Robotics and Computer-Integrated Manufacturing* 81 (2023), p. 102515 (cit. on p. 122).

[236] Sichao Liu, Lihui Wang, and Xi Vincent Wang. “Multimodal data-driven robot control for human–robot collaborative assembly”. In: *Journal of Manufacturing Science and Engineering* 144.5 (2022), p. 051012 (cit. on p. 122).

[237] Haodong Chen, Ming C Leu, and Zhaozheng Yin. “Real-time multi-modal human–robot collaboration using gestures and speech”. In: *Journal of Manufacturing Science and Engineering* 144.10 (2022), p. 101007 (cit. on p. 122).

[238] Yiheng Liu, Tianle Han, Siyuan Ma, et al. “Summary of chatgpt/gpt-4 research and perspective towards the future of large language models”. In: *arXiv preprint arXiv:2304.01852* (2023) (cit. on p. 123).

[239] Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223* (2023) (cit. on p. 123).

[240] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. “Learning transferable visual models from natural language supervision”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763 (cit. on pp. 126, 136).

[241] Alec Radford, Jeffrey Wu, Rewon Child, et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9 (cit. on p. 127).

[242] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units”. In: *54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL). 2016, pp. 1715–1725 (cit. on p. 127).

[243] Zunnan Xu, Zhihong Chen, Yong Zhang, et al. “Bridging Vision and Language Encoders: Parameter-Efficient Tuning for Referring Image Segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 17503–17512 (cit. on pp. 127, 136–138).

[244] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on p. 128).

[245] Rosanne Liu, Joel Lehman, Piero Molino, et al. “An intriguing failing of convolutional neural networks and the coordconv solution”. In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 129).

[246] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. “ChatGPT for Robotics: Design Principles and Model Abilities”. In: *arXiv preprint arXiv:2306.17582* (2023) (cit. on p. 133).

[247] Siyuan Huang, Zhengkai Jiang, Hao Dong, et al. “Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions with Large Language Model”. In: *arXiv preprint arXiv:2305.11176* (2023) (cit. on p. 133).

[248] Tanmay Gupta and Aniruddha Kembhavi. “Visual programming: Compositional visual reasoning without training”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14953–14962 (cit. on p. 133).

[249] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. “Referitgame: Referring to objects in photographs of natural scenes”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 787–798 (cit. on p. 138).

[250] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. “Microsoft coco: Common objects in context”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755 (cit. on p. 138).

[251] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. “Vision-language transformer and query generation for referring segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 16321–16330 (cit. on p. 138).

[252] Zhaoqing Wang, Yu Lu, Qiang Li, et al. “Cris: Clip-driven referring image segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11686–11695 (cit. on p. 138).

[253] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL] (cit. on p. 140).

[254] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL] (cit. on p. 140).