ANTICIPATORY HUMAN-ROBOT HANDOVER INTERACTION MODEL IN AN ASSISTIVE ROBOT DESIGN

SHUN GUI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

School of Design

Anticipatory Human-robot Handover Interaction Model in An Assistive Robot Design

Shun Gui

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

August 2024

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.

Shun Gui

Abstract

The current trend of population aging and an increasing number of individuals with disabilities has drawn significant attention from governments worldwide. This demographic shift has led to a shortage of caregivers, further exacerbating the issue, and raising concerns within society. In response to the needs of individuals with limited mobility in home settings, such as those who are bedridden or wheelchair-bound, there has been a surge in the development of assistive robotic technologies. Various types of assistive robot products have been introduced to address the demand for retrieving everyday objects in these scenarios. Despite significant advancements in robotics, developing effective assistive robots faces challenges in operating in unstructured environments. These settings pose obstacles with objects scattered in varying locations and orientations. Manipulating objects in such dynamic environments requires robust perception, adaptability, and intelligent decision-making. Another challenge is creating safe, reliable, and user-friendly human-robot interaction, achieved through integrating technologies like computer vision, manipulation, and human-robot interaction design. To meet these needs and overcome these challenges, I propose this research with the primary objective of developing a robotic system capable of grasping specific items based on user instructions and delivering them to the user's hand. To achieve this goal, I conduct research on robot recognition, grasping, and control technologies, as well as human-robot handover interaction design.

In Study 1, I perform some robot-to-human handover simulation experiments, which aim to investigate a range of issues pertaining to the robot-to-human handover scenario. The primary objective of these experiments is to gain insights into the genuine requirements of users and identify the key research considerations from both the robot's and the user's perspectives in this task. Through simulation experiments, I conclude some challenging but significant robot techniques and some key factors in this human robot interaction. The subsequent research focus on addressing these aspects.

In Study 2, I propose a 3D object detection algorithm called Recursive Cross-View (RCV) that can be rapidly applied to recognize various items in different robot scenarios. RCV leverages

the three-view principle, transforming 3D detection into multiple 2D detection tasks, using only a subset of 2D labels. RCV introduces a recursive paradigm where instance segmentation and cross-view 3D bounding box generation are performed recursively until convergence. Evaluations on the SUN RGB-D and KITTI datasets demonstrate that the proposed method outperforms existing image-based methods. To showcase the rapid applicability of RCV to new tasks, I implement it in two real-world scenarios: 3D human detection, and 3D hand detection. As a result, two new 3D annotated datasets are obtained, indicating that RCV can be considered as a (semi-)automatic 3D annotator. Furthermore, I deploy RCV on a real robot, achieving real-time 3D object detection at 7 frames per second on live RGB-D streams. Therefore, RCV can be used to recognize various objects for robots in robot-to-human handover scenarios.

In Study 3, I propose a novel 6-DoF robot grasp pose detection approach called GoalGrasp that circumvents the need for grasp pose annotations and training. It facilitates user-specified object grasping even in partially occluded scenes in robot-to-human handover scenarios. By combining 3D bounding boxes and human grasp priors, GoalGrasp introduces a new paradigm for grasp pose detection. Leveraging the RCV 3D object detector, which operates without 3D annotations, GoalGrasp achieves rapid 3D detection in new scenes. Through the integration of 3D bounding box information and human grasp priors, GoalGrasp achieves dense grasp pose detection. Experimental evaluation involving 18 common objects demonstrates the generation of dense grasp poses for 1000 scenes without grasp training, establishing a comprehensive grasp pose dataset. GoalGrasp demonstrates notably superior grasp pose stability compared to existing methods, as indicated by a novel stability metric. In user-specified robot grasping experiments, the method achieves an 94% grasp success rate. Moreover, in user-specified grasping experiments conducted under partial occlusion, the success rate reaches 92%.

In Study 4, I propose an anticipatory handover control model named Deep-MPC that aims to enhance robots' ability to anticipate system state during the handover process. The framework integrates a 3D hand detector (RCV), an online learning transition model, and a data-driven model predictive control (MPC) approach. The 3D hand detector detects hands,

providing visual input to the robotic system. To anticipate future states, Deep-MPC utilizes online learning from data collected during robot-environment interactions to infer forthcoming system states and optimize the robot's actions in real-time. The state transition module in Deep-MPC employs a neural network that takes states and actions as inputs, predicting the subsequent state. By performing multi-step predictions, comparing predicted states to the target state using a loss function, and optimizing actions through gradient backpropagation at each time step, Deep-MPC achieves effective action optimization. Deep-MPC can be viewed as an approach that establishes a human-robot interaction model from the robot's perspective, granting the robot human-like capabilities.

In Study 5, I integrate all proposed methods into a physical robot to execute robot-to-human handover interaction model design. Firstly, I explore the key factors from Study 1 in the robot-to-human handover interaction process, such as objects need to be grasped, robot motion speed, robot handover path, etc. These factors form the foundational elements for human-robot interaction. To determine the settings of all proposed factors, I conduct simulated experiments with participants who simulated individuals with mobility impairments, aiming to experience various interaction modes. Questionnaires are leveraged during the experiments to collect user feedback. Utilizing the gathered data, I develop a new robot-to-human handover interaction model. To validate the effectiveness of the interaction model, I conduct a validation experiment with new participants. Their feedback is collected, analyzed, and used to evaluate the performance of the model. The result demonstrates that the proposed interaction model achieves a good performance. This study proposes a new robot-to-human handover interaction model that partially fills a gap and provides insights for further developments in related robotic technologies.

This research explores the robot-to-human handover from two perspectives: robot techniques and human-robot interaction design. This research holds great significance in the field of robotics as it focuses on advancing automatic object grasping methods for robots and developing an interactive model for object handover between robots and humans. By addressing key research questions, this research aims to enhance the capabilities of robots in assisting individuals with limited mobility in retrieving objects and facilitating user-friendly

interactions. The outcomes of this research have significant implications for designing and implementing future human-robot handover interactions. By identifying crucial factors and leveraging the developed techniques, this research contributes to the advancement of robotic systems that can collaborate with humans in a user-friendly manner, fostering robot adoption and acceptance in domains such as healthcare, assistive robotics, and daily life assistance.

Publications arising from the thesis

- Gui, S., & Luximon, Y. (2023). Recursive Cross-View: Use Only 2D Detectors to Achieve 3D Object Detection without 3D Annotations. *IEEE Robotics and Automation Letters*, 8(10), p.6659-6666. doi: 10.1109/LRA.2023.3307282

- Gui, S. & Luximon Y. (2024). Anticipatory Control on Human-Following Robots Using Online Deep Model Predictive Control. *IEEE Transactions on Industrial Electronics*, p.1-10. doi: 10.1109/TIE.2024.3419209

- Gui, S., Gui, K., & Luximon, Y. (2024). GoalGrasp: Grasping Goals in Partially Occluded Scenarios without Grasp Training. arXiv preprint arXiv:2405.04783. (Under Review)

- Gui, S., Gui, K. & Luximon Y. (2024). Recursive Cross-View 2: Fully Oriented 3D Object Detection on Human-to-robot Handover. Under Review.

Acknowledgements

As I reach the culmination of this PhD journey, my heart is filled with profound gratitude and overwhelming emotions. The process of writing this thesis has been not only a quest for knowledge but also a profound personal growth experience. At this moment, I would like to take this opportunity to express my sincerest gratitude to all those who have supported, encouraged, and inspired me throughout my doctoral studies.

First and foremost, I would like to express my gratitude to my superior, Prof. Yan Luximon. Three years ago, by a fortuitous turn of events, I had the privilege of becoming her student. Her rigorous approach to research and insightful guidance have shaped my PhD journey, defined my research direction, and served as an exemplary role model for my life's journey. Under her mentorship, I have learned to conduct scholarly research from practical problems, greatly influencing my research taste.

I would like to extend my appreciation to my lab mates, whose presence has filled this journey with laughter and warmth, making my academic path less solitary. I am grateful to Ms. Xinyu Shi, Dr. Yuqian Wang, Dr. Shah Parth, Mr. Xiaokang Wei, Dr. Fang Fu, Dr. Jiaxin Zhang, Mr. Junjian Chen, and others, for their assistance in both my research and personal life, making my PhD a truly good journey.

Furthermore, I would like to express my gratitude to my family, especially my parents. Your unconditional love and silent support have been the solid foundation upon which I stand. When facing pressures and challenges, it is your understanding and encouragement that have given me the courage to persevere. The warmth of this family is an everlasting source of motivation for me.

Lastly, I would like to extend profound respect to all researchers, anonymous reviewers, and institutions that have been cited in this work or have provided data and resources. Without the collective efforts and contributions of the academic community, my research would have been difficult to materialize.

Although words cannot fully convey the depth of my gratitude, I hope that through these words, I can convey my appreciation to everyone who has left a mark during my doctoral studies and research. The road ahead is long, but with this experience and the support of all those who have accompanied me, I will stride confidently towards new horizons.

Contents

List of Tables

List of Figures

1. INTRODUCTION

This chapter presents the motivation, research objectives, research questions related to human-robot interaction model on robot-to-human handover, and interaction model design. The aim of this chapter is to describe the background and significance of the research.

1.1. Motivation and Research Background

The world is witnessing an unprecedented demographic shift, with an increasing number of elderly people and individuals with disabilities (Nationen 2022). This shift is particularly pronounced in China Mainland and Hong Kong, where the government has been paying increasing attention to the needs of these vulnerable groups (Miao et al. 2021). The motivation for this research stems from three aspects, supported by substantial data and policy initiatives.

1.1.1. Governmental Focus on the Needs of the Elderly and Disabled

The growing governmental focus on the needs of the elderly and disabled provides a strong impetus for this research. In China, the National Health Commission has implemented various policies and programs to support the aging population, recognizing the importance of addressing their specific requirements. For instance, the “Healthy China 2030 Initiative” (Tan et al. 2019) aims to provide comprehensive healthcare services, including long-term care and rehabilitation, to promote the well-being of the elderly. The "13th Five-Year Plan" (Kennedy et al. 2016) of China emphasizes the importance of improving the quality of life for

the elderly and disabled. Additionally, the Hong Kong government has launched the "Elderly Care Services Industry Scheme" (Pun et al. 2020) to promote the development of elderly care services. These policies reflect the governments' recognition of the importance of catering to the needs of these groups and provide a conducive environment for the development of technologies that can assist them.

In addition to these policies, the Chinese government has also been investing heavily in the development of assistive technologies. The "National Medium- and Long-Term Program for Science and Technology Development (2006–2020)" (State Council of the People's Republic of China 2006) identifies assistive technologies as a key area of focus, and the "Made in China 2025" initiative aims to make China a global leader in high-tech industries, including robotics (Wübbeke et al. 2016). These initiatives provide a strong impetus for the development of robotic technologies that can assist the elderly and disabled.

Furthermore, the Hong Kong government has also been proactive in promoting the use of technology in elderly care. The "Innovation and Technology Fund for Better Living" provides funding for projects that use innovative technologies to improve the quality of life for the public, including the elderly and disabled. The government has also established the "Gerontech and Innovation Expo and Summit" to promote the exchange of ideas and collaboration in the field of gerontechnology. These initiatives demonstrate the government's commitment to fostering the development and adoption of assistive technologies.

1.1.2. Societal Implications of Elderly and Disabled

The societal implications of an aging population and a shortage of caregivers underscore the urgent need to technological advancements in assistive robotics. In China, the proportion of elderly individuals (aged 60 and above) reached approximately 18.7% of the total population in 2020, with projections indicating a further increase to over 30% by 2050 (Bao et al. 2022). This rapid aging trend places immense pressure on the existing healthcare system and exacerbates the shortage of skilled caregivers. The gap between the demand and availability of caregivers is substantial, with an estimated shortage of over 10 million caregiving personnel. Similar challenges are observed in Hong Kong, where the aging population is

expected to account for more than 30% of the total population by 2041 (Kwok et al. 2017). The shortage of caregivers is not only a problem in China and Hong Kong but also a global issue. The World Health Organization estimates that there will be a global shortfall of 12.9 million healthcare workers by 2035 (Durrani 2016). This shortage is expected to be particularly acute in low- and middle-income countries, where the demand for healthcare services is growing rapidly due to population aging and the increasing prevalence of chronic diseases (Boniol et al. 2022). This global shortage of caregivers further underscores the importance of developing technological solutions to assist the elderly and disabled. These statistics highlight the urgent need for innovative solutions, such as assistive robotics, to provide essential care and support to bedridden and mobility-impaired individuals.

Moreover, the aging population and the shortage of caregivers have significant economic implications. The cost of healthcare is expected to rise dramatically as the demand for healthcare services increases. According to a report by the China Research Center on Aging, the cost of elderly care in China is projected to reach 26% of GDP by 2050, up from 7.33% in 2015 (Yang et al. 2021a). This escalating cost of elderly care poses a significant challenge to the sustainability of the healthcare system and underscores the need for cost-effective solutions, such as assistive technologies.

1.1.3. Basic Needs for Elderly and Disabled to Retrieve Items

The research is also motivated by the basic needs of bedridden or wheelchair-bound individuals, whose everyday lives are significantly impacted by their limited mobility. Accessing medication, food, and other necessities can pose significant challenges, as shown in Figure 1.1, and the development of assistive robotics offers promising solutions. In Hong Kong, the “Barrier-Free Access” (Famakin et al. 2018) policy focuses on creating an inclusive environment for individuals with disabilities, ensuring equal access to facilities and services. By aligning with these policies and addressing the basic needs of the target population, the primary objective of this research is to develop a robotic system capable of assisting individuals who are bedridden or wheelchair-bound in retrieving essential items for their daily activities. Furthermore, the research endeavors to establish a new interaction model

between the robot and the human, facilitating a smooth handover process between the two entities.

Figure 1.1 The person in wheelchair to take the object outside of their reach. (Generated by Stable Diffusion XL)

The motivation for this research is driven by the growing governmental focus on the needs of the elderly and disabled, the societal implication of an aging population and a shortage of caregivers, and the basic needs of bedridden or wheelchair-bound individuals. The availability of extensive data and the implementation of supportive policies in China Mainland and Hong Kong provide a solid foundation for this research, highlighting the pressing need for assistive robotics to improve the lives of these vulnerable populations.

1.1.4. Research Background

In this section, the research background of robotic system of handover and robot-to-human handover interaction model is presented.

The rapid advancements in technology, particularly in the field of robotics and artificial intelligence, present significant opportunities for addressing the challenges associated with an aging population and a shortage of caregivers (Czaja et al. 2022). Robots have been increasingly used in various fields, such as manufacturing, logistics, and healthcare, due to their ability to perform tasks accurately, consistently, and tirelessly. In the context of elderly care, robots can perform a wide range of tasks, such as assisting with mobility and providing companionship (Bardaro et al. 2022). The development of assistive robots has been a focus of research in many countries. For instance, Japan, a country with one of the highest proportions of elderly people in the world, has been a pioneer in the development of assistive robots. The Japanese government has launched the "Robot Revolution Initiative" to promote the development and adoption of robots, including assistive robots (Wright 2021), as shown in Figure 1.2. Similarly, the European Union has funded numerous projects under the "Horizon 2020" program to develop assistive technologies for the elderly and disabled (Zallio et al. 2022), as shown in Figure 1.2.

Despite the remarkable advancements made in the field of robotics, there remains a multitude of challenges that need to be effectively addressed in the development of assistive robots. One of the main challenges lies in developing robots that can operate effectively in unstructured environments, such as homes and hospitals (Holland et al. 2021). These environments present a myriad of obstacles and uncertainties, with objects scattered in varying locations and orientations. The ability of assistive robots to navigate and manipulate objects in such dynamic settings requires robust perception capabilities, adaptability, and intelligent decision-making. Overcoming these challenges entail the integration of advanced algorithms and sensor systems that can handle the complexity and variability of real-world environments. Another challenge is the development of robots that can interact with humans in a safe, reliable, and user-friendly manner (Obaigbena et al. 2024). This requires the integration of various technologies, such as computer vision, machine learning, and human-robot interaction.

Figure 1.2. Users' needs to human-robot handover robots. (a) Handover items to mobility impaired users (Toyota's Human Support Robot). Image source: https://bala93.github.io/ (12-Aug.-2024); (b) Handover items to elderly (The Robot House at the University of Hertfordshire). Image source: https://www.unialliance.ac.uk/2021/03/30/the-robot-house-at-the-university-of-hertfordshire/ (12-Aug.-2024)

The primary objective of this research is to develop a handover robotic system. The process of the robot fetching the requested items and delivering them to the user's hand, based on the user's instructions, can be described as follows:

1. User Instruction
2. Perception and Object Recognition.
3. Path Planning.
4. Object Grasping.
5. Navigation.
6. Handover Interaction.
7. Task Completion and Feedback.

By following this process, the robot can effectively fulfill the user's instructions, retrieve the requested item, and safely deliver it to the user's hand, providing assistance and convenience to individuals with mobility limitations.

This research primarily focuses on the most challenging aspects of the robot's capabilities in steps 2, and 4, as well as the human-robot handover interaction

in step 6. These specific areas require particular attention and innovation. In the context of developing robots that can fetch and hand over items to bedridden or wheelchair-bound users in unstructured environments such as homes or hospitals, it is crucial to address the limitations of existing machine learning or deep learning methods for object recognition and grasping. Many of these methods heavily rely on large amounts of manually annotated data and network training, making it difficult to apply them effectively in highly diverse real-world robotic grasping scenarios (Meng et al. 2021). The primary challenges can be summarized as follows:

- The lack of 3D annotated data for diverse objects hinders the robot's ability to accurately recognize the target object for grasping. Most existing datasets predominantly consist of 2D annotations, which do not provide the necessary depth information crucial for precise object recognition. Consequently, developing methods that can leverage limited or unannotated 3D data to improve object recognition in real-world scenarios becomes a significant challenge.

- The scarcity of 6D grasping pose annotations for new scenes and objects poses difficulties in training robots to accurately grasp target objects in various real-world situations, particularly when partial occlusions are present. The ability to generalize and adapt grasping strategies to handle occlusions and unseen object instances is vital for the success of assistive robots operating in complex and dynamic environments.

- The lack of research on real robot-to-human handover interaction models stems from the absence of deployable methods for robot handover on actual robots. This limitation hinders the conduct of authentic user experiments, thereby impeding the collection of user feedback essential for developing a robot-to-human handover interaction model from the user's perspective.

Addressing these challenges requires innovative approaches that reduce the reliance on large-scale annotated data and promote more robust and adaptive learning methods. Furthermore, developing robust grasp planning algorithms that can handle occlusions and

adapt to novel object instances will enhance the robot's ability to perform successful grasping tasks in diverse and unstructured environments. By focusing on these challenges and developing novel methodologies, the research can pave the way for assistive robots to excel in object recognition and grasping tasks, providing enhanced support and independence to individuals with mobility limitations.

Robot-to-human handover interaction model

Currently, there is limited research on the human-robot interaction models specifically focused on the task of robot delivering items to humans. One of the primary reasons for this gap is the inherent challenges associated with robots autonomously retrieving objects based on user instructions in real-world scenarios.

In the human-robot handover scenario, the user and the robot can be viewed as peer role. A human-robot peer role is a relationship between a human and a robot where both parties are considered to be equals in terms of status and social standing (Scholtz 2003). In this type of relationship, the robot is not viewed simply as a tool or machine, but rather as a companion or colleague that can interact with the human in a variety of ways. A human-robot peer role implies a certain level of reciprocity and mutual respect, where both the human and the robot have roles to play and contribute to the relationship. The robot is designed to interact with the human in a way that is responsive, adaptive, and intuitive, and may be programmed to learn and adapt to the human's preferences and needs. Meanwhile, users need to learn how to interact with robots as peers. In a human-robot handover task, when humans perceive a robot as a peer, they may treat it like a human. This means using similar language, gestures, and social norms during handover. Users also need to follow the cues from robots to establish a social interaction and make the handover more seamless. For example, if the robot looks at the object being handed over, the human can also direct their attention to the object. Providing feedback to the robot during handover can help improve future interactions. Humans can give positive feedback when the handover is successful or provide suggestions for improvement if something goes wrong. This can help the robot learn and adapt to the human's preferences and improve its performance over time.

The concept of a human-robot peer role has become increasingly important as robots become more advanced and integrated into our daily lives (Lim et al. 2021). As robots become more human-like in their behavior and appearance, there is a growing recognition that they can serve as more than just tools and can actually provide emotional and social support to humans. Overall, robots as peer role can provide users with a wide range of benefits and support, making them a valuable agent for improving quality of life. To this end, I propose a novel human-robot interaction (HRI) model from both technical and human perspective and achieve it in real world HRI tasks.

HRI is an emerging area of research that involves the study of interactions between humans and robots (Lim et al. 2021). This field is at the intersection of computer science, engineering, and psychology. The goal of HRI research is to create robots that can interact with humans in a natural and intuitive way. To describe human-robot peer role, Scholtz (Scholtz 2003) proposed an HRI model termed HRI Model – Peer Role. As HRI research continues to advance, there is growing interest in developing robots that can work collaboratively with humans in various settings, including homes, offices, and factories. This requires the development of new algorithms and control systems that can enable robots to adapt to different environments and tasks. To this end, I propose a novel HRI model with anticipatory ability based on HRI Model – Peer Role.

Predicting human behavior is crucial for robots that interact with people. Anticipating human actions can help robots to behave in a more natural and intuitive way, leading to a better human-robot interaction experience (Reily et al. 2018). In addition, predicting human behavior can also enhance the safety of the interaction, allowing robots to react in advance to avoid accidents or prevent dangerous situations (Lasota et al. 2017). However, predicting human behavior is not an easy task, as it involves understanding human cognition, perception, and decision-making processes. Nevertheless, recent advances (Rudenko et al. 2020) in machine learning and artificial intelligence have shown promising results in predicting human behavior, relying on data-driven methods that learn from past interactions. For example, human intent prediction is very essential in joint human-robot action, which can greatly smooth joint actions (Tong et al. 2022). Gaze and motion are

commonly used for human intent recognition in human-robot handover tasks (Choi et al. 2022, Belardinelli et al. 2022). Predicting the location of the human-robot handover in advance can speed up and smooth the handover process, especially in dynamic handover task. Lockwood et al. (2022) explored the trajectory, location, and timing in human-to-human handovers, which is expected to be applied to human-robot handovers, enabling robots to predict handover locations and plan actions in advance. Therefore, incorporating prediction capabilities into robots is becoming increasingly important, as it can unlock new possibilities for the way humans and robots interact, enabling more personalized and efficient experiences. Moreover, predicting human behavior can also help robots to adapt to different situations and contexts, making them more versatile and adaptable to real-world scenarios.

Furthermore, to validate the proposed human-robot interaction model for item handover and determine factors, I deploy the developed object recognition and grasping techniques along with the anticipatory human-robot interaction model onto a physical robot. Then I conduct simulated experiments to identify the crucial factors involved in the robot-to-human item handover process and determine various interaction modes. In general, this research investigates human-robot interaction from both the perspective of the robot (anticipating human behavior) and the perspective of the user (studying user preferences towards robot behavior). The goal is to develop a comprehensive human-robot interaction model that encompasses both parties, aligning with the concept of peer role.

1.2. Research Aims and Objectives

The primary aim of this research is to address the pressing societal challenges faced by individuals with limited mobility. These individuals often encounter significant barriers in accessing essential items necessary for their daily lives, leading to a loss of independence and reliance on caregivers. To alleviate these challenges and contribute to the field, this research aims to develop a fully automated robotic system capable of assisting this population through the efficient delivery of essential items. This research seeks to make significant advancements in two key areas: robotic grasping techniques and the development of a human-robot interaction model. The first focus area involves exploring innovative approaches to robotic

grasping, enabling the robot to securely and accurately handle a wide range of objects that individuals with limited mobility may require. This entails studying various grasping strategies, sensor integration, and dexterous manipulation techniques to ensure reliable and adaptive object retrieval. Additionally, the development of a new human-robot interaction model is crucial to enable item transfer between the robot and the individual. This model will facilitate effective communication and collaboration, allowing individuals to easily and intuitively interact with the robotic system to convey their needs and preferences. Designing user-friendly interfaces, adaptive control mechanisms, and personalized interaction modalities will be key aspects of this research, promoting comfort, trust, and efficiency in the human-robot interaction process.

To achieve these aims, I formulate the following research objectives:

- **Objective 1**: To figure out the challenging techniques and key factors in robot-to-human handover interaction model.

- **Objective 2:** To develop a new 3D object detection method that can be used in robot-to-human handover for various objects. By using this method, the robot can recognize user-specified objects.

- **Objective 3:** To develop a target-oriented 6-DoF grasp pose detection method that can be used to grasp user-specified objects for users in robot-to-human handover.

- **Objective 4:** To formulate a real-time and online anticipatory human robot interaction Model-Peer Role on robot-to-human handover. This robot control model explores the establishment of a human-robot interaction model from the robot's perspective.

- **Objective 5**: To develop a novel robot-to-human handover interaction model that can receive instructions from users and autonomously complete the recognition, grasping, and handover to meet the user’s needs for retrieving objects.

1.3. Scope and Research Questions

The research focuses on the automatic object grasping methods of robots and the interactive model of object handover between robots and humans. Firstly, I need to investigate the needs of individuals with limited mobility, such as those who are wheelchair-bound, regarding object retrieval. For example, what types of objects do they require assistance with the most? This question will guide the design of our future application scenarios. To enable real robots to quickly adapt to various object grasping scenarios, I propose novel robot recognition and grasp detection methods. Currently, most learning-based object perception and grasp detection methods rely on large-scale manually annotated data for model training (Meng et al. 2021). However, this approach is not suitable for rapid deployment in various human-robot interaction scenarios due to the time-consuming nature of manual annotation. Therefore, the research questions include whether robots can achieve perception in new scenarios without relying on extensive manually labeled 3D annotations, and whether grasp detection by robots can be achieved without the need for manual annotation data. By addressing these two questions, I propose robot recognition and grasp detection methods that can be rapidly applied to new scenarios. Once the technical aspects of robot grasping are implemented, I delve into the research of the interactive model for object handover between humans and robots. In the context of robot-to-human object transfer tasks, humans and robots can be considered as peers, leading to propose the Peer Role model for human-robot object handover. To accommodate variations in human behavior, I incorporate the robot's motion anticipation into this model. Another research question is how to incorporate anticipatory capabilities into the HRI handover model - Peer Role. After implementing all the technical methods on a real robot, I explore the most critical factors in the robot-to-human interaction process. This knowledge will be used for future robot-to-human robot designs. In summary, the research questions of this study are as follows:

- **RQ 1:** What are the challenging techniques and key factors in robot-to-human handover HRI?

- **RQ 2:** Can real-time robotic 3D object detection method in new scenes be achieved in the absence of 3D annotations?

- **RQ 3:** Can target-oriented 6-DoF grasp pose detection be achieved in robot-to-human handover tasks without grasping training?

- **RQ 4:** How to integrate anticipation into the HRI handover Model – Peer Role to form the anticipatory HRI Model – Peer Role?

- **RQ 5:** How to formulate a new robot-to-human handover interaction model?

1.4. Significances of the Study

The research holds significant importance in the field of robotics, focusing on the development of automatic object grasping methods for robots and the interactive model of object handover between robots and humans. By addressing several key research questions, this study aims to contribute to the advancement of robotic capabilities in assisting individuals with limited mobility in retrieving objects and enabling interactions between humans and robots.

One of the primary objectives of this research is to explore the specific needs of individuals who face challenges in performing actions such as wheelchair bound. By understanding their requirements for object retrieval, the research aims to provide valuable insights for the design of future robot applications in assisting such individuals. This knowledge will guide the development of tailored solutions that cater to their specific needs, enhancing their quality of life and promoting independence. Moreover, the study proposes novel approaches to robot recognition and grasp detection, addressing the limitations of existing methods that heavily rely on labor-intensive manual annotations. By investigating the feasibility of recognition without extensive human-labeled 3D annotations and grasp detection without manual annotation data, this research aims to accelerate the deployment of robots in various human-robot interaction scenarios. The proposed methods will enable real robots to swiftly adapt to new environments and efficiently grasp a wide range of objects, thereby enhancing their practicality and versatility in real-world applications. Furthermore, the research delves into the development of an interactive model for object handover between humans and robots. By considering humans and robots as peers, the Peer Role model is proposed,

integrating the anticipation by the robot. This model aims to improve the effectiveness and naturalness of object handover interactions, facilitating collaboration between humans and robots. Additionally, by incorporating foresight capabilities into the handover model, the research seeks to enhance the robot's anticipation and adaptability during the interaction process, leading to more intuitive and efficient handover experiences.

The outcomes of this research have significant implications for the design and implementation of future robot-to-human handover interactions. By identifying crucial factors in the robot-to-human handover interaction process and leveraging the developed techniques, this study will contribute to the advancement of robotic systems that can user-friendly collaborate with humans, promoting the adoption and acceptance of robots in various domains, such as healthcare, assistive robotics, and daily life assistance.

In summary, this study's significance lies in its potential to enhance the capabilities of robots in object grasping, object handover, and human-robot interaction, ultimately benefiting individuals with limited mobility and paving the way for more effective and natural collaboration between humans and robots.

1.5. Research Framework

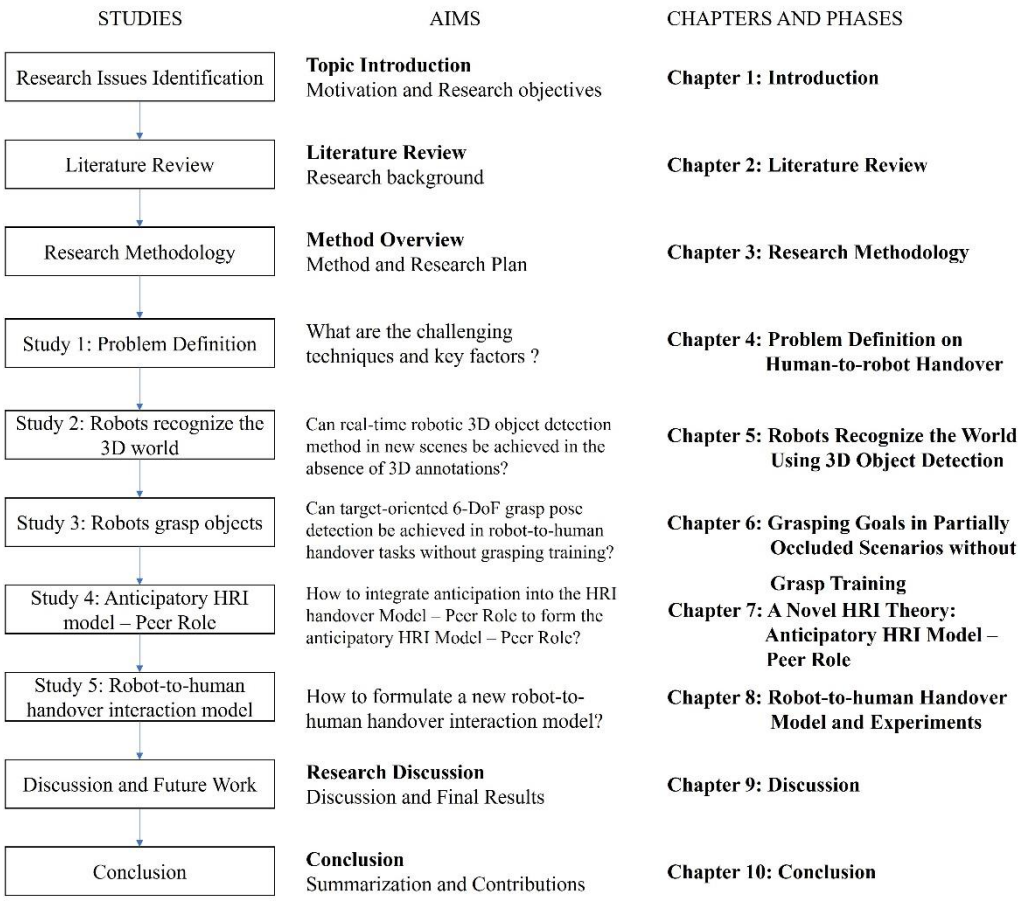


Figure 1.3. Outline of the thesis.

The research framework of thesis is structured as Figure 1.3. The thesis consists of ten chapters. Chapter 1, Introduction, presents motivation, research background, aims and objectives, and research significance.

Chapter 2, Literature Review, provides a comprehensive review of the existing literature from two perspectives: assistive robotics technologies for human-robot handover and models for human-robot interaction. It initially presents the advancements in assistive robots, followed by a review of the progress in robot recognition and grasp techniques for human-robot handover. Subsequently, it examines the literature on anticipatory models for human-robot interaction. By reviewing these studies, research gaps are identified and summarized, highlighting the specific challenges that this research aims to address.

Chapter 3, Methodology, outlines the overall research approach employed in this research. A mixed-methods methodology is utilized, incorporating both quantitative analysis of robot technologies and experimental data, as well as qualitative user-robot experiments and questionnaire surveys. The combination of quantitative and qualitative approaches ensures a comprehensive investigation from both technical and design perspectives.

Chapter 4, Problem Definition on Robot-to-human Handover, explores the challenging techniques and key factors in robot-to-human handover HRI. This chapter specifically addresses Objective 1 and Research Questions 1.

Chapter 5, Robots Recognize the 3D World Using 3D Object Detection, introduces a novel 3D object detection method called Recursive Cross-View (RCV). This method is designed to be quickly applicable to various robotic scenarios. It addresses the challenge of rapid identification of diverse objects by assistive robots in different environments, which is crucial for the efficient deployment of real-world robots. This chapter specifically addresses Objective 2 and Research Question 2.

Chapter 6, Grasping Goals in Partially Occluded Scenarios without Grasp Training, presents an efficient 6-DoF grasping detection method that enables fast and accurate grasping of user-specified objects in partially occluded scenes. This method serves as a prerequisite for successful robot-to-human handover. It builds upon the RCV method and incorporates human grasping prior knowledge, making it easily adaptable to new object grasping tasks. This chapter specifically addresses Objective 3 and Research Question 3.

Chapter 7, A Novel HRI Theory: Anticipatory HRI Model – Peer Role, introduces an anticipatory HRI model that endows robots with human-like interactive capabilities. This model enables robots to anticipate future human and environmental behaviors in real-time and optimize their own actions, accordingly, facilitating efficient robot-to-human handover. This chapter specifically addresses Objective 4 and Research Question 4.

Chapter 8, Robot-to-human Handover Model and Experiments, integrates the research findings from Chapters 4 to 7 into a real robot system and conducts robot-to-human interaction experiments. Through the analysis of experimental results, a user-friendly

human-robot interaction model is proposed, demonstrating favorable performance in the validation experiments. This chapter addresses Objective 5 and Research Question 5.

Chapter 9, Discussion, evaluates the extent to which this research addresses the research objectives and research questions. It also discusses the limitations of this research and suggests future directions for further research.

Chapter 10, Conclusion, provides a comprehensive summary of the contributions and significant findings of the research study.

2. LITERATURE REVIEW

This chapter offers a thorough review of the current body of literature, focusing on two key aspects: assistive robotics technologies for human-robot handover and models for human-robot interaction. The chapter begins by presenting the advancements made in the field of assistive robots, followed by a comprehensive review of the progress in robot recognition and grasp techniques specifically pertaining to human-robot handover scenarios. Additionally, it delves into an analysis of existing literature on anticipatory models for human-robot interaction. Through this extensive review, the chapter identifies and summarizes research gaps, emphasizing the specific challenges that this study aims to tackle.

2.1. The development of assistive robotics

Assistive robotics plays a pivotal role in addressing the challenges faced by elderly or disabled individuals, offering essential support, and enhancing their overall quality of life. Governments and policy-making institutions recognize the need for innovative solutions to support aging populations and individuals with disabilities. By complementing human caregivers, robots can alleviate the burden on healthcare systems and empower individuals to live more independently. Moreover, assistive robotics meets societal needs by filling the gap in caregiver shortages and combating loneliness and isolation. Robots designed for companionship and social interaction provide emotional support, engage in conversations, and stimulate cognitive abilities, promoting social inclusion and well-being. At the individual level, assistive robots empower individuals by promoting independence and autonomy,

assisting with daily tasks, and offering personalized support tailored to their needs. The significance of robots in meeting these needs has driven the vibrant development of assistive robotics. Increased investments and collaborations have led to rapid advancements in robot design, sensing technologies, and artificial intelligence, resulting in the introduction of innovative robots with a wide range of capabilities and functionalities, ultimately improving the lives of elderly and disabled individuals. Many companies and research institutions have introduced assistive robots.

In 2014, SoftBank Robotics introduced Pepper (SoftBank Robotics, 2014), as shown in Figure 2.1, which was designed to engage in social interactions with humans. It features a humanoid design, natural language processing capabilities, and facial recognition technology. Pepper can provide companionship, answer questions, and assist with various tasks, making it suitable for use in care homes, hospitals, and public spaces.

Figure 2.1. Pepper (SoftBank Robotics). Image source: https://en.wikipedia.org/wiki/Pepper_(robot) (12-Aug.-2024)

In 2015, The Fraunhofer Institute for Manufacturing Engineering and Automation (IPA) in Germany developed Care-O-bot 4 (Ackerman, 2015), as shown in Figure 2.2, which is a service robot designed to assist with various tasks in home and healthcare environments. It has a mobile base and a multi-fingered robotic hand for manipulation. Care-O-bot can perform tasks like serving meals, delivering medications, and tidying up. It has a user-friendly interface for interaction and can adapt to individual needs and preferences.

Figure 2.2. Care-O-bot (Fraunhofer IPA). Image source: https://www.care-o-bot.de/en/care-o-bot-4.html (12-Aug.-2024)

Figure 2.3. Jaco (Kinova Robotics). Image source: https://smanewstoday.com/columns/sma-challenges-eased-jaco-robotic-arm/ (12-Aug.-2024)

Jaco developed by Kinova Robotics (Kinova, 2020), as shown in Figure 2.3, was a robotic arm designed to assist individuals with limited arm mobility in performing daily tasks. It can be mounted on a wheelchair and is capable of gripping objects, manipulating items, and performing various activities of daily living, such as eating, drinking, and opening doors.

In 2008, Willow Garage developed a mobile robot equipped with two arms and a variety of sensors named PR2 (Garage, 2008), as shown in Figure 2.4. It is designed to perform complex manipulation tasks and assist with household chores. PR2 can navigate environments, pick up objects, fold laundry, fetch items, and even serve drinks. It has been widely used in research and development, particularly in the field of robotic manipulation.

Figure 2.4. PR2 (Personal Robot 2) (Willow Garage). Image source: https://robotsguide.com/robots/pr2 (12-Aug.-2024)

Mobile ALOHA (Fu et al. 2024), as shown in Figure 2.5, developed by Stanford University in 2024, is a groundbreaking robotic system that advances the capabilities of bimanual mobile manipulation through low-cost whole-body teleoperation. This innovative system builds upon Google DeepMind's existing ALOHA system, bringing mobility and dexterity to the forefront of robotic learning. Unlike most results in the field of imitation learning, which primarily focus on table-top manipulation, Mobile ALOHA extends its reach to mobile manipulation tasks that require both bimanual coordination and whole-body control. The system combines a low-cost mobile base with a whole-body teleoperation interface, allowing it to perform complex, long-horizon tasks.

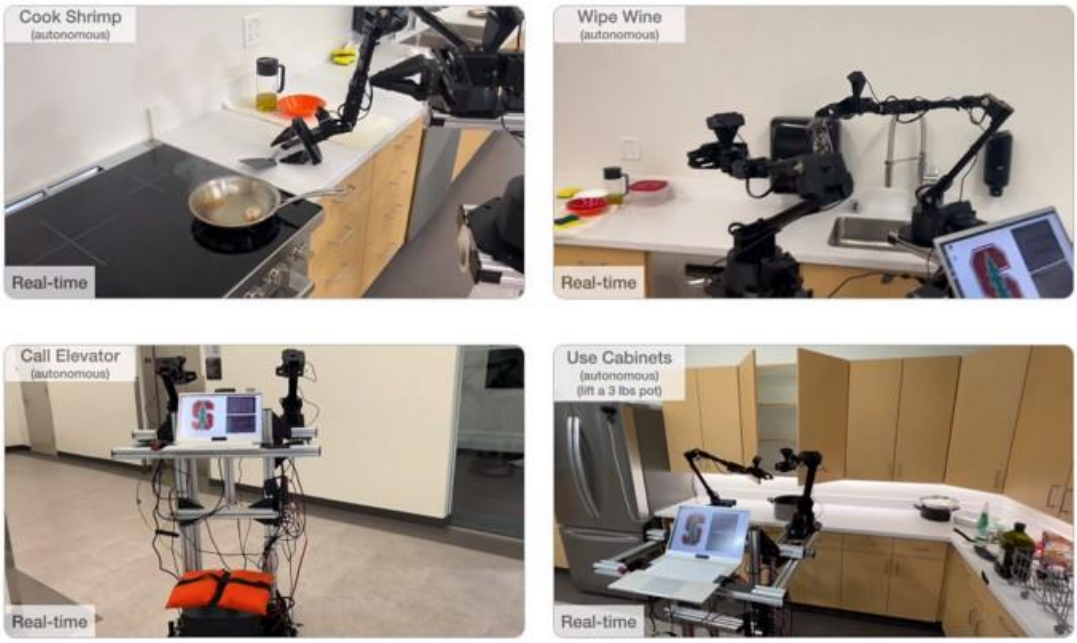


Figure 2.5. Mobile Aloha (Fu et al. 2024).

In the context of this research, I address the inherent challenges faced by existing robots in adapting to novel and unstructured environments, often necessitating labor-intensive data annotation or manual intervention (Jia et al. 2024). Moreover, these robots have overlooked a critical aspect—the user-friendly handover of objects between humans and robots. Their deficiencies lie in the insufficiency of capabilities to accurately identify and handle unfamiliar objects, thereby lacking the desired levels of adaptability and flexibility. Thus, my research is devoted to exploring and addressing these limitations, with a focus on enhancing the adaptability, perception, and dexterity of robots in order to enable successful human-robot handovers and improve their overall performance in diverse real-world scenarios.

2.2. Human-Robot Handover on Assistive Robots

In human-robot handover tasks, the robot can assume the role of either a giver or a receiver. In this study, I specifically focus on the robot's role as a giver, where it grasps objects and transfers them to humans. Currently, there is a considerable body of research dedicated to robot-to-human handover. Many studies (Lehotsky et al. 2023, Ovur et al. 2023, Sidiropoulos et al. 2021, Nowak et al. 2022) have been conducted to investigate various aspects of robot-to-human handover, with the aim of improving safety, efficiency, and naturalness in the handover process. These studies encompass diverse areas, including

human-robot interaction, computer vision, motion planning, and grasping techniques. Perovic et al. (2023) proposed an adaptive method, which combines Dynamic Movement Primitives (DMP) with Preference Learning (PL) to generate online trajectories that are reactive to human motion, for robot-to-human handovers under different scenarios. A DMP-based robot-to-human method was developed (Iori et al. 2023), which generates an online trajectory based on DMP, leading to the robot can adapt to human motion during handovers. To meet the individual preferences, a reinforcement learning-based method was proposed, which used the human-robot coefficiency score as reward to adapt and learn online the combination or robot interaction parameters that maximises such coefficiency. The experimental results shows that the human perceive comfort can be improved (Lagomarsino et al. 2023). A study (Qin et al. 2022) presents a taxonomy and a system for generating task-oriented handovers in robot-to-human interactions, demonstrating adaptability to various difficulty levels and receiving positive evaluations in terms of human comfort and task appropriateness compared to prior work. Ardon et al. (2021) addresses the limitations of existing approaches for estimating the effectiveness of object handovers, which are typically limited to users without arm mobility impairments and specific objects. The authors propose a method that generalizes handover behaviors to novel objects while considering the user's arm mobility levels and task context. Through an online study involving users with different arm mobility levels, they demonstrate that people's preferences for handover methods are correlated with their arm mobility capacities. Using a statistical relational learner (SRL) model, the proposed method achieves an average handover accuracy of 90.8% when generalizing handovers to novel objects. Meng et al. (2022) addresses the issue of comfort in transferring tools and objects to human hands by proposing a framework that utilizes advanced deep learning models to pre-generate handover target configurations based on object and tool characteristics. The experimental results demonstrate the robustness and efficiency of the framework in delivering various objects to human hands in convenient grasping poses, even when the hand moves to a new position. Christensen et al. (2022) introduced a novel approach for training object affordance segmentation models for robot-to-human handovers using a synthetic dataset, achieving performance levels comparable to methods trained on hand-labeled datasets.

In the context of human-robot object handover, the robot's ability plays a crucial role in ensuring successful interaction. The handover process consists of two distinct stages: the pre-handover stage and the physical handover stage. While the physical handover stage emphasizes the robot's grasp adjustment and failure handling abilities, the overall success of object handover heavily relies on the robustness of grasp strategy and motion planning during the pre-handover stage. As a result, current research in this field primarily focuses on enhancing these key abilities to enhance the efficiency and reliability of human-robot object handover. Next, I will review the research on robotics involved in robot-to-human handover.

2.2.1. 3D Objection Detection on Robotics

3D object detection is the process of identifying, categorizing, and creating 3D bounding boxes for objects within a given environment. Recent advancements in 3D sensors, annotated datasets, and deep learning methodologies have significantly propelled the field forward. This progress is particularly impactful for applications such as self-driving cars, robotic navigation, robotic manipulation, and interactions between humans and robots. Current approaches to 3D detection can be divided into several main types: image-based (Hu et al. 2023, Zhou et al. 2022, Liu et al. 2021c, Huang et al. 2018, He et al. 2019, Izadinia et al. 2017), projection-based (Chen et al. 2017, Fazlali et al. 2022), voxel-based (Hu et al. 2022, Mao et al. 2021, Shi et al. 2020), and point-based (You et al. 2022, Zhang et al. 2022a, Misra et al. 2021, Pan et al. 2021, Shi et al. 2019). The majority of existing research has concentrated on autonomous driving and indoor object detection, largely due to the availability of public datasets like KITTI (Geiger et al. 2012) and SUN RGB-D (Song et al. 2015). Typically, 3D object detection involves training a model to regress a set of 3D bounding boxes that represent the ground truth. But what happens when there is a need to detect a novel object in an unfamiliar scene?

A typical approach involves first generating a sufficient number of 3D annotations and then training a neural network. However, manually annotating 3D data from RGB-D sensors or LiDAR is both time-consuming and costly (Meng et al. 2021, Xu et al. 2022). For instance, the creators of the SUN RGB-D dataset employed and trained 18 oDesk workers, who collectively spent 2051 hours on data annotation (Song et al. 2015). Clearly, this method is

not efficient for rapid real-world applications that need results within a few hours. Many researchers have adopted a fully supervised learning framework, which includes data representation, feature learning, and classification and regression tasks. However, the success and dependability of these methods are highly dependent on the availability of accurate 3D annotations. To reduce the dependency on 3D annotations, alternative strategies such as weakly supervised (Meng et al. 2021, Xu et al. 2022, Peng et al. 2022a), semi-supervised (Zhang et al. 2022b, Zhao et al. 2020), and self-supervised (Liang et al. 2021) learning have been explored for 3D object detection. For example, Meng et al. (2021) introduced a weakly supervised framework that utilized BEV center-click annotations along with several hundred 3D labels to train a model. Nevertheless, this approach still requires some 3D annotations. So, is it possible to achieve 3D detection using only other readily available labels?

Although some 3D detection methods have shown impressive results in specific contexts, their capabilities are often limited to identifying objects with a vertical alignment (Chen et al. 2017, Lang et al. 2019, Qi et al. 2018, You et al. 2022, Qi et al. 2019, Qi et al. 2020, Zhang et al. 2022a, Mao et al. 2021, Misra et al. 2021, Pan et al. 2021). These methods typically generate bounding boxes that are vertically oriented, neglecting any potential roll or pitch, which restricts their overall utility. To predict fully oriented bounding boxes, one would need to increase the output dimensions, thereby complicating the detection process. Additionally, there is a scarcity of datasets that support this requirement. Fully oriented detection is crucial for a broader range of applications, such as robotic manipulation and human-robot interaction. Developing a 3D detection technique capable of identifying fully oriented objects and quickly adapting to new 3D sensors across various scenarios would be highly advantageous for practical applications.

2.2.2. Object Pose Estimation

Thanks to the deep learning technologies, there are a large number of studies that directly regress 6D pose parameters from RGB images based on much annotated data. For example, training a neural network in a supervised learning paradigm to predict the 6D pose of an object, given an image as input. Some methods (Do et al. 2018, Xiang et al. 2017, Lee et al.

2021) classify or regress 6D pose from a single view. Do et al. (2018) proposed a deep learning network named LieNet to segment objects in the image as well as estimate the 6D pose by regressing a Lie algebra-based rotation representation and a translation vector. PoseCNN (Xiang et al. 2017) directly regresses the relative position of the object to the camera and predict the rotation parameters by a quaternion vector, given an image as input. It is hard to estimate 6D pose from a single view, so some studies (Labbé et al. 2020, Merrill et al. 2022, Fu et al. 2021) use multi-view to do 6D pose estimation. CosyPose (Labbé et al. 2020) first utilized a single view-based method to propose several 6D pose hypotheses, then matched hypotheses across multi-view images to jointly predict 6D pose for each object. Merrill et al. (2022) introduced an object-level SLAM framework with continuous multi-view as input, to detect key points, and then estimate 6D pose. To further enhance the performance of 6D pose estimation, some studies (Deng et al. 2021, Li et al. 2018, Lin et al. 2022) introduced 3D CAD models to assist in pose estimation.

In some human-robot handover studies (Yang et al. 2021b, Tong et al. 2022), the estimation of object pos from images is low-level. Yang et al. (2021b) leveraged a semantic model to segment hand and object from the image with the help of depth information without further estimating the 6D pose of the object. The evaluation of the pose relies on the grasping network. Yang et al. (2021c) exploited Fast-SCNN (Poudel et al. 2019) to segment objects from an image and used REDE (Hua et al. 2021) to estimate the pose. In general, the study of the size and 6D pose of objects in human-robot object handover needs to be further enhanced, for example, some methods in CV can be introduced, which is very essential to improve the handover performance.

Depth maps or point clouds can provide additional 3D information, which is useful for object pose estimation. Some studies (Lin et al. 2022, Wang et al. 2019b, Sahin et al. 2018, Shi et al. 2021, Tian et al. 2020, Wang et al. 2019b, Lin et al. 2021, Chen et al. 2020a, Chen et al. 2021a) leveraged these data for 6D pose evaluation. SAR-Net (Lin et al. 2022) utilized both RGB images and point clouds to estimate the 3D size and 6D pose of an object. It aligned the observed point cloud and template shape to obtain the 3D rotation and yielded the 3D translation and 3D size by the symmetric correspondence of the point cloud. Furthermore,

SAR-Net was validated on a real robot to perform some grasping tasks. DenseFusion (Wang et al. 2019b) first processed RGB and depth data separately, followed by a dense fusion network to jointly extract features that were used to estimate the 6D pose. To further enhance the results, it adopted a pose refinement procedure in the end. FS-Net (Chen et al. 2021a) estimated 6D pose from a single-view RGB-D image. It proposed one shape-based network to estimate rotation, and one residual-based network to predict 3D translation and 3D size of the object. StablePose (Shi et al. 2021) extracted geometric features of patches and contextual features between patches, combined with geometric stability to estimate the 6D pose, given a point cloud as input. REDE (Hua et al. 2021) was an 6D pose estimator consuming RGB-D data, which regressed key points first and then leveraged a differentiable geometric to evaluate 6D pose.

Some studies (Trick et al. 2019) of human-robot handover also leveraged depth maps or point clouds to identify object poses to achieve the handover. Trick et al. (2019) exploited PointNet, PointNet++ and RandLA-Net to segment objects held in hand from multimodal data, i.e., RGB, point cloud and thermal, for safe robotic object handover. Poudel et al. (2019) also extracted the point cloud masked by the semantic label, then fed it to REDE to evaluate the 6D pose. Yang et al. (2021) extracted the point cloud of an object, and then leveraged 6-DOF GraspNet (Mousavian et al. 2019) to achieve pose estimation. Overall, the state-of-the-art 6D pose estimation methods could be used to enhance the performance of human-robot handover tasks, which will be a research trend.

2.2.3. Hand Pose Estimation

Hand pose estimation is very essential for object handover, whether in human-to-robot or robot-to-human handovers tasks. For robots to achieve object handover, recognition of hand position and pose is a perceptual prerequisite for the physical behavior of the handover. Here, I reviewed some learning-based methods for hand pose estimation, which are classified into 3 categories, namely, (a) hand pose estimation from images, (b) hand pose estimation from depth images or points clouds, and (3) hand pose estimation on multimodal. Then, I summarized some challenges for hand pose estimation in human-robot handover.

Estimating hand pose directly from RGB images is very economical, as RGB cameras are easily accessible. With the development of deep learning technologies, the mainstream image-based methods include convolutional neural networks (CNN), graph-convolutional neural networks (Graph CNN), and synthetic models. Zimmermann et al. (2017) first used CNN to estimate 3D hand pose from a monocular RGB images, in which CNN extracts image features and then combines camera parameters to predict the 3D pose. After this work, many CNN-based studies emerged (Cai et al. 2018, Chen et al. 2019a, Iqbal et al. 2018, Mueller et al. 2018, Tekin et al. 2019, Yang et al. 2019, Zimmermann et al. 2017, Fan et al. 2021, Panteleris et al. 2018). Spurr et al. (2021) proposed a two-stage model to achieve hand pose estimation, where an encoder was used to represent the images, then to predict 3D hand pose after supervised learning. To enhance performance, some studies (Simon et al. 2017, Chen et al. 2021b) leveraged multi-view images to predict hand pose. Some studies (Ge et al. 2019, Baek et al. 2019, Yang et al. 2020b) to predict 3D hand shape and pose simultaneously. Ge et al. (2019) leveraged Graph CNN to reconstruct a full 3D hand shape and pose. These well-designed studies are expected to be used in human-robot handover tasks to enhance the recognition of hand pose during handover. However, most current hand pose detections of human-robot handovers were low-level, focusing only on segmenting hand regions from images for subsequent handovers. Yang et al. (2021) trained a full convolution model, which leveraged the Feature Pyramid Network based on ResNet-50 (He et al. 2016), to segment hand from an image. Yang et al (2020a) leveraged 2D detector to locate the hand and the object from images. Rosenberger et al. (2020) leveraged RefineNet and ResNet to segment body and hand from an RGB image for saft human-to-robot handovers.

Depth images or point clouds can provide more 3D information, which can be used to directly to locate the hand position and pose. A depth map is a kind of 2.5D format data. Thanks to the strong deep learning technologies, many discriminative models were proposed in recent years. A model consumed one or several depth maps and then regresses to the 3D hand pose (Chen et al. 2020b, Fang et al. 2020a, Huang et al. 2020, Ren et al. 2021, Wan et al. 2018). Xiong et al. (2019) proposed the Anchor-to-Joint regression network (A2J) with depth images as input, in which anchor points that encode spatial context information are set compactly on the depth image as local regressors of the points to predict the 3D pose of the

hand joints. Yuan et al. (2019) utilized paired RGB and depth images to supervise an RGB-based model to mimic features from a model trained by fully labeled depth data, to help enhance the performance of 3D hand pose estimate in RGB images only conditions. Ren et al. (2022) proposed a self-supervised framework based on a cross-view fusion model and a graph convolution model, to predict 3D hand pose and mesh. On the other hand, in combination with the camera matrix, the depth map can be transformed into a 3D point cloud. In this case, the technology of 3D vision can be applied. Chen et al. (2019b) proposed a model named SO-HandNet. With the point cloud as input, the model can learn multi-level features and then fuse them to predict 3D hand pose, using a semi-supervised learning mode. Ge et al. (2018a, 2018b) leveraged PointNet (Qi et al. 2017a) to directly process the point cloud and regress the 3D hand pose. Malik et al. (2020) proposed a voxel-based network and combined with 3D convolutions to predict 3D hand pose.

The human-robot handover task could benefit from these studies by applying them to real handover tasks to achieve accurate detection of hand pose during human-robot handover. In recent human-robot handover studies (Yang et al. 2020a, Zhang et al. 2021a), researchers have explored a number of methods. Yang et al. (2020a) used PointNet++ (Qi et al. 2017b) to process the point cloud cropped by an 2D detector, and then classified human grasp to achieve human-robot handover. Zhang et al. (2021) exploited PointNet and PointNet++ to segment hands and objects for safe robotic handover.

In general, current approaches to 3D hand position and pose in human-robot handover tasks are relatively simple compared to studies in the field of computer vision, which results in poor hand pose estimation performance. Therefore, I need to consider how to apply relevant research in the field of CV to real handover tasks. However, I need to solve the problems of occlusion, real-time, noise, etc. in real tasks, which are extremely challenging.

2.2.4. Robotic Grasp Pose Detection

In human-robot interaction contexts, robots frequently depend on human instructions to carry out specific tasks. For example, home service robots often receive directives from users to achieve particular objectives, such as picking up a designated item (Xu et al. 2023). The

ability to grasp objects specified by users is crucial in these interactions, especially for individuals with limited mobility, such as the elderly or patients, where robots can offer essential support (Tröbinger et al. 2021).

Despite this, current research in robotic grasping (Fang et al. 2023, Mahler et al. 2017) tends to focus less on user-directed grasping. In many studies, the robot does not know in advance which object it will grasp (Chen et al. 2023). These studies usually involve identifying grasp poses for all objects in a scene and then selecting the optimal pose for execution. While this approach works well for tasks like bin picking (Cordeiro et al. 2022), it may not be suitable for user-driven human-robot interactions in everyday settings. Extending these methods to target-specific grasping tasks presents significant challenges, such as (1) the inability to classify generated grasp poses by object type and (2) the varied scenarios robots encounter, which often lack sufficient training data for reliable grasping. One possible solution is to manually create a large-scale dataset of categorized grasp poses. However, this is highly inefficient due to the labor-intensive and time-consuming nature of manually labeling 6D annotations (Chen et al. 2022a, Deng et al. 2020). Given the wide range of applications for robots, this method is impractical.

Some research on target-specific grasping (Murali et al. 2020, Liu et al. 2022) starts with object detection or instance segmentation (Xie et al. 2020) in images and then identifies the corresponding grasp poses in the associated point cloud to achieve user-specified grasping. However, these methods can be disrupted by nearby objects, as they may mistakenly include adjacent items during target detection. Additionally, their performance significantly drops when the target object is partially obscured. Indeed, occlusion is a major challenge for current robotic grasping methods (Yu et al. 2023a). In scenarios where target-driven grasping is required, partial occlusion of objects is often unavoidable. Therefore, addressing the issue of occlusion is a key challenge that our research aims to tackle.

Learning-based Grasp Pose Detection: Recent progress in deep learning has enabled the creation of data-driven systems for robotic grasping (Levine et al. 2018). However, these methods typically require extensive annotated datasets for training (Fang et al. 2023). This necessity has led to the development of several large-scale grasping datasets, including

GraspNet-1Billion (Fang et al. 2020b), 6-DOF GraspNet (Mousavian et al. 2019), EGAD! (Morrison et al. 2020), ACRONYM (Eppner et al. 2021), MetaGraspNet (Gilles et al. 2022), and Grasp-Anything (Vuong et al. 2023). These datasets feature both simulated and real-world data. However, the annotation process for these datasets is often labor-intensive and time-consuming (Chen et al. 2022, Deng et al. 2022), which can hinder the rapid deployment of learning-based methods in diverse real-world robotic applications. Fang et al. (2023) highlighted the limitations of the commonly used sim2real methods in the grasping community.

Moreover, learning-based robotic grasping methods often involve designing and training complex neural network models to predict grasp poses (Fang et al. 2023, Mahler et al. 2017, Sundermeyer et al. 2021, Zhao et al. 2021). For instance, Mousavian et al. (2019) used a variational autoencoder to generate a set of grasps, which were then evaluated and refined using a grasp evaluator model. Wang et al. (2021) introduced an end-to-end network called GSNet, which integrates a graspness model to predict grasp poses. Fang et al. (2020) proposed a grasp pose prediction network that separately learns the approaching direction and operation parameters using point cloud inputs. A significant challenge with learning-based methods is their difficulty in generalizing to new scenarios and objects. Often, new data must be collected, labeled, and the model retrained, which is time-consuming and inefficient for varied robotic applications. Our approach leverages simple grasping priors to enhance efficiency significantly.

Target-orientated Grasp Pose Detection: Many existing grasp pose detection methods work by taking a scene as input and generating multiple grasp poses, from which the robot selects the optimal one for execution (Fang et al. 2023, Mahler et al. 2017). While this approach is effective for tasks like bin picking, it may not be suitable for grasping user-specified objects in human-robot interaction scenarios. Recognizing this limitation, some research has focused on grasping user-specified objects (Murali et al. 2020, Liu et al. 2022, Sun et al. 2021, Li et al. 2022). Murali et al. (2020) used instance segmentation (Xie et al. 2020) to align grasps with target objects, while Liu et al. (2022) employed a semantic segmentation module to locate the target first and then predict the grasp poses. Most of these

studies have concentrated on 3-DoF grasping, which generates grasping poses within the camera plane, limiting their practicality. Additionally, they have not thoroughly evaluated their performance in scenarios where the target object is partially obscured.

Grasp Pose Detection in Partially Occluded Scenarios: One category of grasp pose detection methods is based on 6D pose estimation (Deng et al. 2020, Du et al. 2021, Cao et al. 2023, Zhang et al. 2021b, Yu et al. 2023b). Deng et al. (2020) proposed a self-supervised 6D object pose estimation for robotic grasping, while Zhang et al. (2021) presented a practical approach that uses 6D pose estimation along with corrective adjustments for protection. However, these methods have not shown satisfactory performance in scenarios where target objects are partially occluded, as 6D pose estimation methods are susceptible to occlusions (Hu et al. 2020). Similar issues arise in methods for detecting graspable rectangles (Mahler et al. 2017, Levine et al. 2018, Lenz et al. 2015a, Chu et al. 2018, Zhang et al. 2019, Cheng et al. 2023). Some methods (Mousavian et al. 2019, Ten et al. 2017, Liang et al. 2019, Shao et al. 2020) adopted a sampling-evaluation strategy, which involves sampling potential grasp candidates on point cloud data and then assessing their quality using a neural network. Additionally, some studies (Fang et al. 2023, Wang et al. 2021, Qin et al. 2020, Zheng et al. 2023) employed neural networks to regress grasp poses on point clouds. However, these point cloud-based methods struggle to generate accurate grasp poses when objects are partially occluded, especially when the optimal grasp regions are hidden. The nature of these methods limits their ability to generate grasp poses in regions where point cloud data is unavailable, and they often lack a comprehensive understanding of the grasped objects.

2.3. Human-Robot Interaction Model

A human-robot interaction model is a framework or approach aimed at achieving effective communication and collaboration between humans and robots. It involves the recognition and interpretation of human behavior, language, and intent, and responds through appropriate feedback and actions. The model typically incorporates elements such as perception, communication, action and control, learning and adaptation, social intelligence, trust and transparency, and user interface and interaction design. By integrating these elements, the human-robot interaction model strives to create intuitive, natural, and reliable

human-robot interaction experiences. Its applications span various domains, including healthcare, manufacturing, and everyday life, with the goal of enabling robots to become valuable companions, assistants, and partners.

Model for Interactive Human-Robot Interaction (MIHR) (Perez et al. 2020) is an advanced framework designed to facilitate seamless and effective communication between humans and robots. It integrates components such as perception, communication, action and control, learning and adaptation, social intelligence, trust and transparency, and user interface and interaction design. By incorporating these elements, MIHR aims to create a comprehensive model that enables robots to accurately perceive human actions, generate human-like speech, perform precise and safe physical actions, learn from user feedback, understand social cues, build trust through transparency, and provide intuitive interfaces.

In the field of human-computer interaction (HCI), the Norman's model (Norman et al. 1986) can be described as these seven stages: (1) formulation of the goal, (2) formulation of the intention, (3) specification of the action, (4) execution of the action, (5) perception of the system state, (6) interpretation of the system state, and (7) evaluation of the outcome. Figure 2.6 demonstrates the Norman's HCI model, where a number of intentions are inferred from the goals, then actions are derived, the state is perceived, and evaluated, and finally the intentions and actions are adjusted according to the evaluation results. Since HRI has some differences from HCI, Scholtz (2003) proposed some HRI models according to the Norman's HRI model, that are HRI Model – Supervisor Role, HRI Model – Operator Role, HRI Model – Mechanic Role, HRI Model – Peer Role, and HRI Model – Bystander Role.

According to (Scholtz 2023), a human-robot peer relationship refers to a connection between a human and a robot where both individuals are regarded as equal in terms of social status. In this kind of relationship, the robot is not seen as a mere instrument or device, but rather as a partner or co-worker capable of interacting with the human in diverse ways. In interaction tasks, humans and robots can be considered as peer role if they can be viewed as teammates to do something together (Groom et al. 2007). In this study, I choose human-robot handover task, in which the robot can be regard as an assistant to accomplish a task together with humans. Therefore, it can be classified as peer role to users. In this study, I will

extend HRI Model – Peer Role on time scale to formulate a novel HRI model, and then deploy it on HRI tasks.

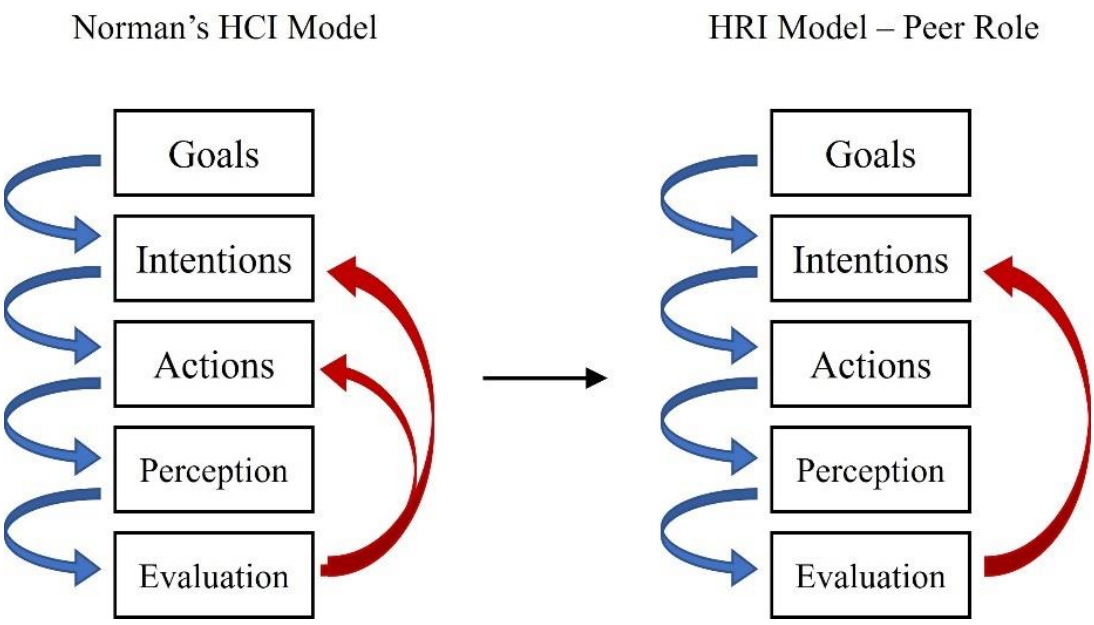


Figure 2.6. Norman's HCI model and HRI model – peer role.

2.4. Anticipatory Human-Robot Interaction

Anticipatory human-robot interaction refers to the ability of robots to anticipate and respond proactively to human actions and intentions. It involves predicting the future behavior of humans and adapting robot behaviors, accordingly, enhancing the efficiency, safety, and naturalness of human-robot collaboration. Anticipatory HRI relies on prediction of human actions. This involves understanding human intent, behavior patterns, and contextual cues to anticipate future actions. Various approaches, such as machine learning, computer vision, and probabilistic models, have been employed to predict human actions.

2.4.1. Human Action Prediction

Action recognition and prediction in video analysis involve inferring the present state and predicting the future state of human actions using computer vision and machine learning techniques. These tasks have gained significant attention due to their emerging applications in areas such as visual surveillance, autonomous driving, entertainment, and human-robot interaction. Action recognition focuses on categorizing and identifying human actions based

on complete action executions, while action prediction aims to forecast future actions based on incomplete executions. Researchers have dedicated efforts to develop robust frameworks and effective models for these tasks.

Ryoo (2011) introduced a novel approach for human activity prediction by utilizing probabilistic methods. It addresses the problem of inferring ongoing activities from videos containing only the onsets of the activities. The proposed methodology, including the formulation of the prediction problem and the development of the dynamic bag-of-words recognition approach, aims to enable early recognition of unfinished activities, particularly for surveillance systems. Gao et al. (2017) proposed a Reinforced Encoder-Decoder (RED) network for action anticipation, addressing the problem of detecting actions before they occur. Unlike existing methods that rely on single past frame representations, RED takes multiple history representations as input and learns to anticipate a sequence of future representations. The study (Furnari et al. 2020) addressed the problem of egocentric action anticipation by proposing the Rolling-Unrolling LSTM architecture. The method incorporates two LSTMs to model the past and predict the future, utilizes Sequence Completion Pre-Training to encourage sub-task focus, and employs the Modality ATTention (MATT) mechanism for efficient fusion of multi-modal predictions. Ke et al. (2019) proposes a novel time-conditioned method for efficient and effective long-term human action anticipation. The approach addresses the challenge of predicting future actions accurately at arbitrary time-horizons. By conditioning the anticipation network on time and incorporating attended temporal features and time-conditioned skip connections, the method efficiently anticipates both short-term and long-term actions. Ke et al. (2021) addresses the task of Assessing Future Moment of an Action of Interest (AFM-AI), which involves predicting whether a specific action will occur in the future and estimating the starting moment of that action. The proposed method utilizes a Deterministic Residual Guided Variational Regression Module (DR-VRM) that combines a Variational Regression Module (VRM) and a deterministic residual network. The VRM accounts for uncertainty and generates diverse predictions for the starting moment, while the deterministic network improves precision by leveraging residual information.

In HRI, there are also many studies focusing on the encounter of human actions, which can greatly improve the response speed and comfort of human-robot interaction. Aghapour et al. (2016) focuses on the challenge of coordinating human-robot interactions by making robots aware of human action plans. It explores the use of the behavioral systems modeling approach to address the problem of human action prediction. The paper aims to leverage this approach to improve coordination and performance in human-robot interactions by predicting human actions based on their (sub)optimal reasoning process and knowledge of the system's state. The study (Moon et al. 2021) explored addresses the challenge of accurately predicting human movements in physical human-robot interaction by considering individual user differences. It introduces a meta-learning framework that enables rapid adaptation of the prediction model to unseen users. The proposed model structure incorporates shared and adaptive parameters, allowing for user-specific predictions. Experimental results on a motion dataset demonstrate the effectiveness of the proposed method in predicting the movements of unseen users, outperforming existing baselines. Bandi et al. (2021) focuses on motion prediction and action recognition in a supermarket assistance scenario for human-robot interactions. It introduces a new small-scale dataset for this scenario and proposes two self-attention-based models that capture long-range correlations without relying on a predefined skeleton structure. The models are evaluated using specific feature encodings to enhance motion or trajectory features, achieving accurate prediction and recognition of actions. The effectiveness of the models is validated on the supermarket dataset and the NTU RGB+D benchmark dataset, aiming to enhance interactions between humans and robots in a supermarket setting. The study (Vianello et al. 2021) addresses the problem of predicting human postures in a collaborative scenario where a robot interacts physically with a human. The proposed method utilizes the distribution of the null space of the Jacobian and the weights of the weighted pseudo-inverse, learned from demonstrated human movements, to predict human joint velocities based on the current posture and robot end-effector velocity. The goal is to ensure that the predicted posture is coherent with the robot's action and considers individual differences and movement preferences. Yasar et al. (2021) introduces a novel sequence learning approach for human motion prediction in single and multi-agent settings, with a focus on enabling fluent human-robot collaboration. The proposed method learns a robust representation of human motion

and conditions future predictions based on a subset of past sequences. Wang et al. \cite{wang2021machine} developed a neural network that consumes skeleton data to predict the target's skeleton state half a second in advance, and the robot performs the following action accordingly, leading to improved following robustness. Chen et al. \cite{chen2019human} developed a new algorithm to predict the human's future position based on the current position and orientation using a human-walking model, to achieve smoother and faster human-following. However, these studies only predicted one future state and did not test the approach in occluded scenes.

Human intent recognition is very essential in joint human-robot action, which can greatly smooth joint actions (Tong et al. 2022). Gaze and motion are commonly used for human intent recognition in human-robot handover tasks (Choi et al. 2022, Belardinelli et al. 2022). Here, I reviewed some learning-based studies on handover intent detection. Choi et al. (2022) formulated a recurrent convolutional network with human gesture and gaze as input, which outputs a heatmap of a user's placement intentions. However, the method was applied to indirect human-robot handover, namely, human-placement-robot handover. After detecting human placement intent, the robot carries out preemptive motion planning to achieve smoother handover. Trick et al. (2019) trained a classifier for multimodal intention detection, with speech, gestures, gaze, and objects as input. The method was tested on a real 7-DoF robot and achieved a robustness human-robot handover. Wang et al. (2018) leveraged a wearable sensory system to collect gestures and muscle activities to predict handover intention using Hidden Markov Model. Some studies (Kshirsagar et al. 2020, Faibish et al. 2022, Newbury et al. 2022) also investigated robot gaze behaviors in human-robot handovers.

Predicting the location of the location of the human-robot handover in advance can speed up and smooth the handover process, especially in dynamic handover task. Lockwood et al. (2022) explored the trajectory, location, and timing in human-to-human handovers, which is expected to be applied to human-robot handovers, enabling robots to predict handover locations and plan actions in advance. Nemlekar et al. (2019) trained a model named Pro-MP by many human-robot handover demonstrations to estimate the object transfer point.

The method was tested on some real human-robot handovers, and a good performance. Simmering et al. (2019) formulated a hand tracking system to predict the handover point. Liu et al. (2021a) leveraged a binary cost function to estimate the location of handover. Combined with the torque cost and placement cost, the method could predict a handover position that is most comfortable for the human.

2.4.2. Robot Behavior Adaptation

Once human actions are predicted, robots need to adapt their behaviors to facilitate effective collaboration. This includes adjusting motion trajectories, planning appropriate actions, and providing timely feedback or assistance to meet human needs and preferences.

Mitsunaga et al. (2008) proposed an adaptation mechanism based on reinforcement learning for human-robot interactions, aiming to enable the robot to read subconscious comfort and discomfort signals from humans and adjust its behavior accordingly. The mechanism utilizes gazing at the robot's face and human movement distance as indicators of human comfort and discomfort. The study conducted with a humanoid robot and 12 subjects demonstrates that the proposed mechanism allows for autonomous adaptation to individual preferences in terms of interaction distances, gaze meeting, and motion speed and timing. The study (Chen et al. 2018) proposes an information-driven multi-robot behavior adaptation mechanism for human-robot interaction (HRI) by using facial expressions and identification information to understand human emotional intention. The mechanism utilizes information-driven fuzzy friend-Q learning (IDFFQ) to select the optimal policy of behavior, enabling robots to adapt their behaviors to human emotional intention for smoother HRI. Del et al. (2022) proposes a framework for enabling autonomous robots deployed in public spaces to adapt their behavior through online user interactions. The framework utilizes a Reinforcement Learning (RL) approach, specifically the Upper-Confidence-Bound Value-Iteration (UCBVI) algorithm, to maximize user engagement during interactions. The approach is tested in a public museum as a tour guide, and results show significant improvements in user engagement, with increased numbers of visited items and higher completion probabilities. Umbrico et al. (2020) proposes a cognitive approach for socially assistive robotics that integrates ontology-based knowledge reasoning, automated planning, and execution

technologies. The goal is to endow assistive robots with intelligent features to reason, understand health-related needs, and perform personalized assistive tasks. The approach addresses the challenges of realizing intelligent and continuous behaviors, robustness, flexibility, and adaptation in socially assistive robots. The study presents the cognitive approach, highlights the contribution of different knowledge contexts, demonstrates adaptation and personalization features through functioning traces, and provides an experimental assessment to validate the feasibility of the approach.

In addition to studies on robot behavior adaptation, there are several studies focusing on mutual behavior adaptation in HRI tasks. Nikolaidis et al. (2017) presents a computational formalism, the Bounded-Memory Adaptation Model, for mutual adaptation between a robot and a human in collaborative tasks. The model captures human adaptive behaviors under a bounded-memory assumption and is integrated into a probabilistic decision process. Human subject experiments demonstrate that the proposed formalism enhances the effectiveness of human-robot teams in collaborative tasks, compared to one-way adaptations of the robot to the human, while maintaining the human's trust in the robot. Van et al. (2021) focuses on the first stage of co-learning in human-robot teams, aiming to identify recurring behaviors that indicate co-adaptation. A computer simulation of an urban search and rescue task is used to study the interactions between a human participant and a virtual robot. The observations reveal patterns of interaction that facilitate behavior adaptation in the task and between the human and robot. The results demonstrate the feasibility of studying co-learning and suggest that participant adaptation improves robot learning and overall team learning.

2.4.3. Context Awareness

Contextual information, such as the environment, task requirements, and social norms, plays a crucial role in anticipatory HRI. Robots need to perceive and interpret contextual cues to understand the current situation and anticipate human actions in a meaningful way.

Quintas et al. (2019) aimed to investigate the impact of incorporating interaction workflows into an agent's information model and decision-making process. The framework developed captures the agent's expected behavior through descriptive scenarios and integrates them

into probabilistic planning and decision-making. The results demonstrate improved specificity without compromising precision and recall, indicating the plausibility of the proposed approach. This framework contributes to cognitive robotics by enhancing the usability of artificial social companions and overcoming limitations of static models in achieving natural interaction. Liu et al. (2021b) presents a context awareness-based collision-free human-robot collaboration system that ensures both human safety and assembly efficiency in a shared manufacturing environment. Yu et al. (2022) proposed a context awareness multi human-robot interaction (MHRI) system that enables multiple operators to interact with a robot. The system utilizes a monocular multi human 3D pose estimator based on convolutional neural networks to accurately estimate the positions of multiple individuals in real time, even in crowded scenes with occlusions. The identities of the individuals are recognized using action context and 3D skeleton tracking. The feasibility, effectiveness, safety, and collaborative efficiency of the MHRI system are demonstrated through multi human-robot interactive experiments and evaluated using HRI metrics. Lison et al. (2010) presents a framework using Markov Logic to construct rich belief models of a robot's environment. The framework captures relational structure and uncertainty, allowing beliefs to evolve dynamically over time. Beliefs are organized in distinct ontological categories and incorporate contextual information. The goal is to enhance the robot's awareness of its surroundings for natural interaction and integration with high-level cognitive functions. Liu et al. (2018) presents a context-aware safety system for human-robot collaboration in shared manufacturing environments. The system ensures both human safety and system efficiency by planning robotic paths that avoid collisions with humans while reaching target positions on time. It incorporates human pose recognition to further enhance efficiency.

2.5. Summary and Research Gap

I conduct a review of robot-to-human handover research from two perspectives: (1) robot technology and (2) human-robot interaction models. In terms of robot technology, in addition to advancements in assistive robots, I also review the use of 3D object detection and robot autonomous grasping techniques for robot perception. This study primarily focuses on utilizing 3D object detection to perceive various objects and employing robot autonomous

grasping techniques to successfully grasp the target items. Through the review of current literature, several research gaps were identified that significantly limit the real-world application of robots. In order to develop a grasping robot capable of adapting to various scenarios and objects, corresponding methods were proposed in this study.

Regarding human-robot interaction models, I review existing human-computer interaction models and the application of anticipatory control in human-robot interaction. As the objective of the research is to develop a robot capable of interactive handover with humans and proactively adapt to human behavior, I summarized the shortcomings of existing HCI and HRI models. Additionally, gaps in current research on anticipatory interaction were identified. Through the literature review of these aspects, several relevant research gaps were identified, as follows:

- **A study on the challenging techniques and key factors in robot-to-human handover HRI is currently missing from the current literature**. This research gap corresponds to **research objective 1** and **research question 1**.

  Currently, the research on robot-to-human handover predominantly focuses on exploring robot technologies such as object recognition and grasping. However, there is a significant gap in research from the perspective of both robot and the user, concerning challenging robot technologies and important factors in robot-to-human handover HRI. When it comes to robot-to-human handovers, the user experience plays a vital role in ensuring a seamless and intuitive interaction. It involves factors such as the perceived comfort, safety, and efficiency of the handover process, as well as the overall satisfaction of the human counterpart. To address this gap, it is essential to conduct research that specifically examines the user experience aspects of robot-to-human handovers.

- **A data-efficient and rapidly deployable 3D recognition method that can adapt to different scenarios on robot is still missing from the current literature.** This research gap corresponds to **research objective 2** and **research question 2**.

With the rapid development of artificial intelligence, 3D visual perception methods have made remarkable progress in various fields, such as object detection and semantic segmentation. However, most existing deep learning-based methods require a significant amount of annotated data to train sophisticated models. This approach is inefficient in real-world robot scenarios, as these scenarios are highly dynamic, and manual data annotation for each instance is impractical. Therefore, it is essential to propose a data-efficient and rapidly deployable 3D perception method that can adapt to different scenarios on robots.

- **A reliable and data-efficient target-oriented 6-DoF grasp pose detection method that can be deployed on robot is still missing from the current literature**. This research gap corresponds to **research objective 3** and **research question 3**.

  In the field of 6-DoF grasp pose detection, researchers typically require a substantial amount of annotated data, including 3D models of objects and corresponding grasp pose information. However, manual annotation of such datasets is time-consuming and labor-intensive, thus limiting the feasibility of these methods in real-world robotics applications. Additionally, the current research predominantly concentrates on bin-picking scenarios, wherein objects are grasped from a cluttered container. Nevertheless, in many practical applications, robots need to perform precise grasping based on different shapes, sizes, and object characteristics. Therefore, there is a pressing need for further exploration of object-oriented grasp scenarios.

- **A real-time and online anticipatory HRI Model - Peer Role on robot-to-human handover is still missing from the current literature.** This research gap corresponds to **research objective 4** and **research question 4.**

  While there has been some research on anticipatory HRI in various contexts, such as object handovers, there is a gap in the literature when it comes to explicitly considering the concept of peer roles in robot-to-human handover scenarios. A real-time and online anticipatory Human-Robot Interaction (HRI) model that

incorporates the concept of peer roles during robot-to-human handover scenarios. In these situations, there is a need for a model that enables the robot to anticipate and adapt its behavior based on the perceived role of the human counterpart in the handover interaction.

- **A study on the robot-to-human handover HRI model is currently missing from the current literature**. This research gap corresponds to **research objective 5** and **research question 5**.

  Currently, there is limited research on the human-robot interaction models specifically focused on the task of robot delivering items to humans. One of the primary reasons for this gap is the inherent challenges associated with robots autonomously retrieving objects based on user instructions in real-world scenarios. In this research, I conducted theoretical and experimental investigations to develop and evaluate a user-friendly human-robot interaction model in this scenario.

3. METHODOLOGY

3.1. General Methodology Description

To narrow down the previous research gaps and deal with the research questions, a research methodology is formulated, as shown in Figure 3.1. After identifying the research gaps following the literature review and in alignment with the research questions, I propose five distinct studies. Specifically, several role-play experiments are performed in Study 1 to identify the challenging techniques and key factors in robot-to-human handover tasks. Study 2 and Study 3 focus on the robot technologies involved in robot-to-human handover, namely 3D object detection and 6D grasp pose detection, respectively. Quantitative methods, including deep learning techniques and real robot experiments, are employed in both studies. Subsequently, the methods proposed in Study 2 and Study 3 are integrated into an HRI model, which is then applied in Study 4 to design an Anticipatory HRI Model. Deep learning methods and real robot experiments are primarily utilized in Study 4.

Upon completing Study 4, the research outcomes from Study 2, Study 3, and Study 4 are integrated into a physical robot. A real robot-to-human handover experiment (Study 5) is conducted, combining experimental procedures with questionnaires to acquire user experiences and feedback regarding the robot-to-human handover process. Quantitative statistical methods are employed to identify various factors within the handover interaction. This analysis aims to establish a new robot-to-human handover interaction model. Finally, validation experiments are conducted to validate the proposed handover interaction model, utilizing a combination of experimental procedures and questionnaires.

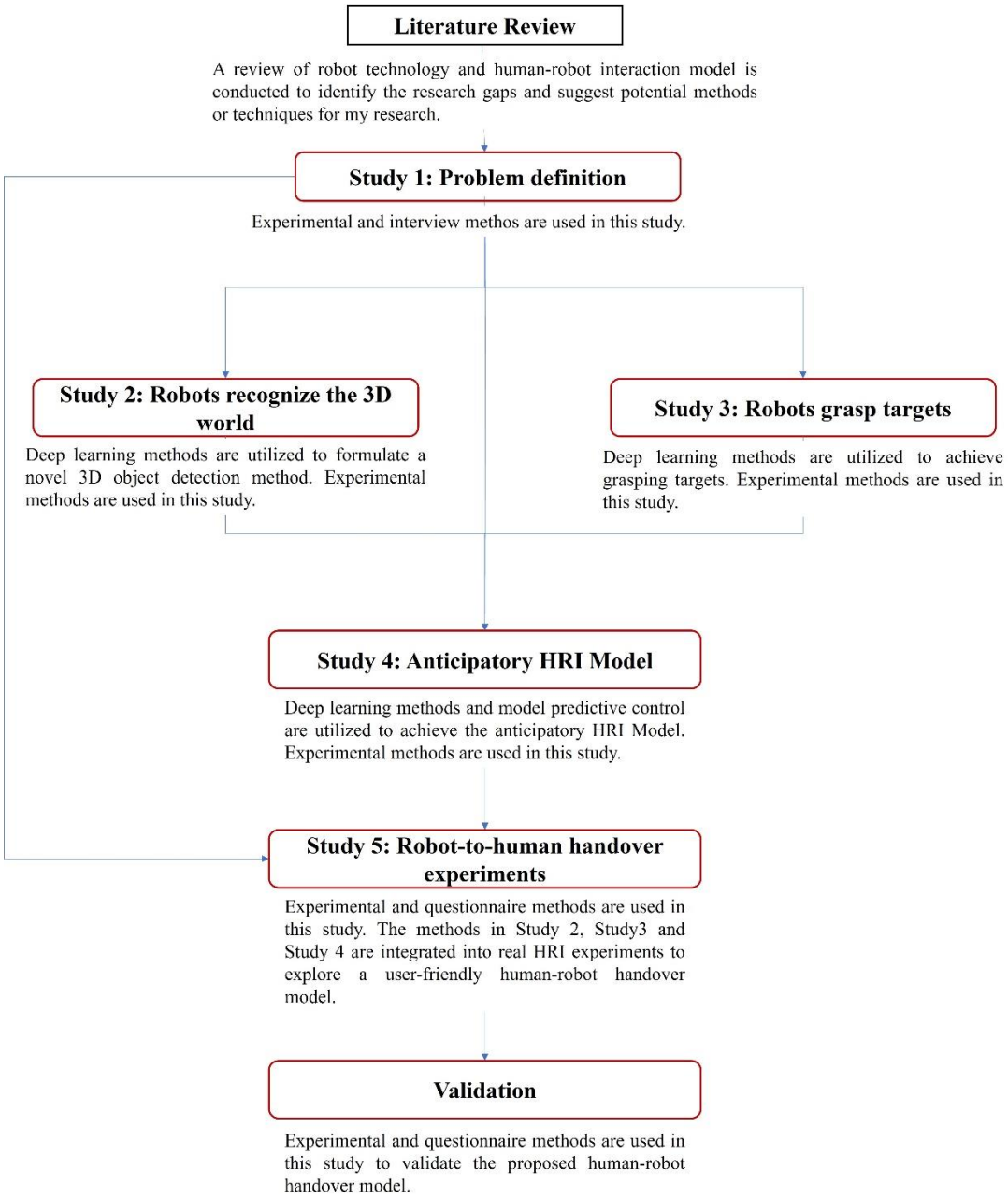


Figure 3.1. Research methodology.

In the Introduction section, I outlined the human-to-robot handover process, which consists of seven steps as illustrated in Figure 3.2. This process can be described as follows: the robot receives a command from the user, identifies the corresponding target object, plans a path, approaches, and grasps the object. After successfully grasping the object, the robot navigates back to the user and completes the handover, thus accomplishing the task. The five studies I proposed are strategically positioned within this process to address key challenges and enhance the robot's capabilities. Study 1 establishes the foundational research scope,

identifying the critical components and objectives of the handover process. Studies 2 and 3 focus on enabling the robot to effectively recognize and manipulate objects, incorporating advancements in computer vision and robotic manipulation to improve accuracy and reliability. Study 4 is dedicated to optimizing the robot's ability to transfer objects to the user, exploring techniques to enhance the fluidity and safety of the handover interaction. Study 5 delves into the exploration of various factors that influence the handover process, such as user preferences, environmental conditions, and robot behavior, to develop a comprehensive understanding of the dynamics involved.

Collectively, these studies contribute to the development of a robust robot-to-human handover interaction model, which aims to facilitate seamless and intuitive interactions between humans and robots. This model not only addresses the technical aspects of the handover process but also considers the human factors, ensuring that the interaction is user-friendly and adaptable to diverse scenarios. Through this research, I aim to advance the field of human-robot interaction, paving the way for more effective and harmonious collaborations between humans and robotic systems.

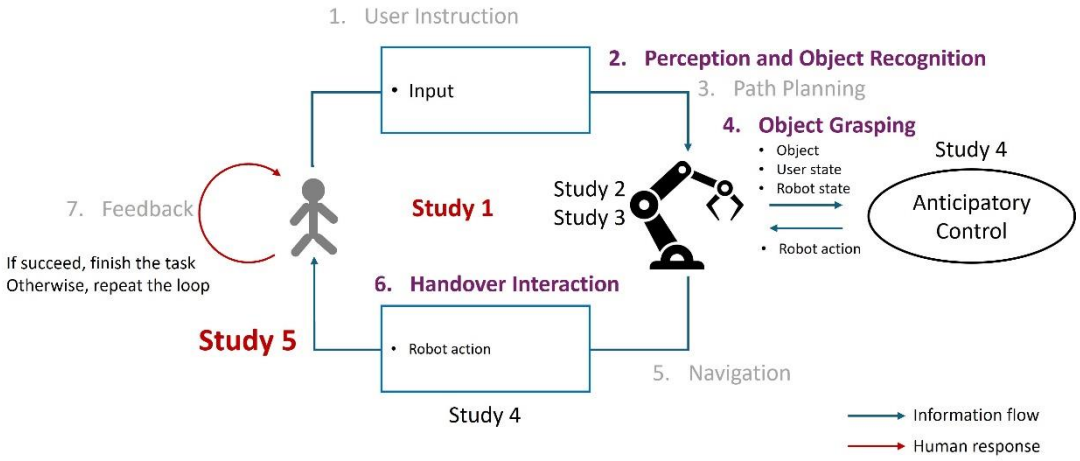


Figure 3.2. The relationship of the proposed five studies.

3.2. Study 1: Problem Definition on Robot-to-human Handover

In this research, I utilized a focused methodology to explore robot-to-human handover interactions, aiming to identify challenging technologies and key factors in the HRI model. Central to this approach were role-play experiments, where one participant was designated as the "user" and another as the "robot." This setup allowed us to simulate the handover process and examine the interaction dynamics from both perspectives.

The role-play experiments provided valuable insights into the users' needs and the technological challenges faced by robots. To complement these experiments, I also employed questionnaires, and interviews. The simulations helped refine robotic behaviors, while questionnaires gathered quantitative data on user experiences. Interviews offered qualitative insights, deepening the understanding of user expectations and satisfaction.

3.3. Study 2: Robots Recognize the World Using 3D Object Detection

I introduce a straightforward yet powerful 3D detection method called RCV, which operates on point clouds without requiring 3D annotations. By leveraging the relationship between 3D space and 2D planes, this method transforms 3D detection into multiple 2D detection tasks. Utilizing a recursive approach, RCV can perform instance segmentation and predict fully oriented 3D bounding boxes. RCV offers several advantages over existing methods. Firstly, it eliminates the need for 3D annotations, making it more practical for real-world applications. By relying on well-established 2D detectors, RCV only requires some 2D labels to achieve 3D detection and can benefit from the robustness and generalization capabilities of 2D detectors. Secondly, RCV is capable of predicting fully oriented bounding boxes, significantly broadening its range of applications. Thirdly, RCV can be rapidly deployed to new 3D sensors in various real-world environments. Additionally, once trained, RCV can serve as a 3D annotation tool, simplifying the manual labeling process or creating datasets for pretraining purposes.

3.4. Study 3: Grasping Goals in Partially Occluded Scenarios without Grasp Training

I present GoalGrasp, a simple yet effective 6-DOF robot grasp pose detection method that does not rely on grasp pose annotations and grasp training. The proposed approach enables user-specified object grasping in partially occluded scenes. By combining 3D bounding boxes and simple human grasp priors, GoalGrasp introduces a novel paradigm for robot grasp pose detection. First, I employ the 3D object detector proposed in Study 1, which requires no 3D annotations, to achieve rapid 3D detection in new scenes. Leveraging the 3D bounding box and human grasp priors, the method achieves dense grasp pose detection. The experimental evaluation involves 18 common objects categorized into 7 classes based on shape. Without grasp training, the method generates dense grasp poses for 1000 scenes, establishing a comprehensive grasp pose dataset. I compare the method's grasp poses to existing approaches using a novel stability metric, demonstrating significantly higher grasp pose stability. In user-specified robot grasping experiments, GoalGrasp achieves a 94% grasp success rate. Moreover, in user-specified grasping experiments under partial occlusion, the success rate reaches 92%.

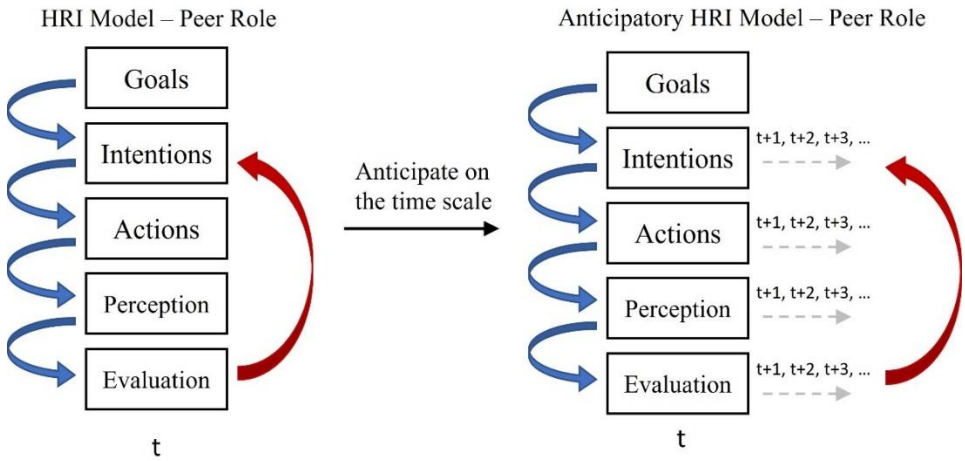


Figure 3.3. Anticipatory HRI Model – Peer Role.

3.5. Study 4: Anticipatory HRI Model – Peer Role

In interaction tasks, humans and robots can be considered as peer role if they can be viewed as teammates to do something together (Groom et al. 2007). In this study, I choose human-

robot handover task, in which the robot can be regarded as an assistant to accomplish a task together with humans. Therefore, it can be classified as HRI Model – Peer Role, as shown on the left-hand side of Figure 3.3. For example, when the goal is to take one object, the robot needs to derive intentions, and select and execute actions. In this case, direct observation is probably the perceptual input used for evaluation. Finally, the intentions are adjusted according to the evaluation results. In a human-robot handover task, the robot needs to detect the 3D position and pose of the object and hand, and then generate a grasping pose to complete the object handover task. After literature review, I found that applying anticipation to the scenario can effectively raise user satisfaction. Therefore, I extend the HRI Model – Peer Role on a time scale to form the Anticipatory HRI Model – Peer Role, as shown in Figure 3.3. Specifically, the robot assesses the current state of the system and then infers the possible future states of the system. By optimizing the actions that can be taken, the robot aims to expedite the system's transition to an ideal state.

To achieve this, I propose a solution called online deep model predictive control (Deep-MPC) and apply it to human-robot handover. Deep-MPC incorporates a 3D hand detector, an online learning transition model, and a data-driven MPC framework. Specifically, the 3D hand detector generates the target's 3D bounding box, while the transition model predicts future states, enabling anticipatory control. The data-driven MPC framework optimizes robot actions using the neural network of the transition model, and online learning occurs through autonomous interaction with the environment, eliminating the need for system modeling and controller design.

3.6. Study 5: Robot-to-human Handover Experiments

Due to the limited research on the robot-to-human handover interaction model, many factors affecting the interaction, such as the objects need to be grasped, handover speed, and robot trajectory planning, lack clear strategies. To address this gap and design a comprehensive robot-to-human handover interaction model, I integrate the proposed methods into a physical robot system. This robot-to-human handover system allows users to engage in experiments and enables the identification of these factors within the handover interaction model. In this study, experimental methods are employed to gather data and insights. To

collect feedback from users, questionnaire methods are utilized. Experimental methods are indispensable and crucial in human-robot interaction (HRI) research, particularly when aiming to achieve HRI in real-world settings. The utilization of experimental methods offers many advantages, prominently the ability to collect a substantial amount of real interaction data. Figure 3.4 demonstrates the methodology used in the Study 5.

Firstly, I invite individuals to simulate users with limited mobility and set different interaction modes for the robot. Through the users' experience of these different interaction modes, I collect their feedback using questionnaires. Once obtaining the feedback, I summarize the essential factors and develop a new handover interaction model. To validate this model, a verification experiment is conducted. New participants are invited to experience the handover interaction model, and their feedback is collected to validate the newly proposed interaction model.

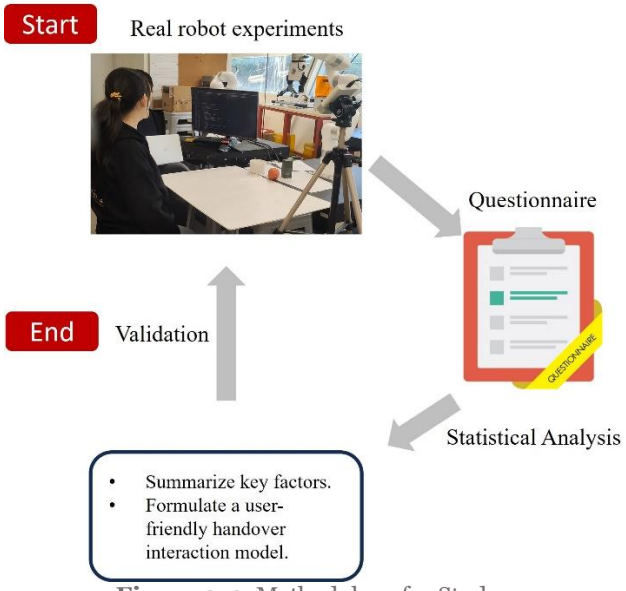


Figure 3.4. Methodology for Study 5.

3.7. Summary of the Methodology

The methodology framework, as shown in Figure 3.5, employed in this research encompasses five distinct studies to address the research gaps. Role-play experiments are performed in Study 1. Study 2 focuses on 3D object detection in the context of robot-to-human handover, utilizing quantitative methods such as deep learning techniques and real robot experiments. Study 3 delves into 6D grasp pose detection, employing similar quantitative approaches.

The methods proposed in Study 2 and Study 3 are integrated into an HRI model, which forms the basis for Study 4. In Study 4, an Anticipatory HRI Model is designed, incorporating deep learning methods and real robot experiments. Following the completion of Study 4, the research outcomes from Study 2, Study 3, and Study 4 are integrated into a physical robot. This setup facilitates Study 5, which involves conducting real robot-to-human handover experiments. A combination of experimental procedures and questionnaires is used to gather user feedback and experiences, allowing for the identification of various factors within the handover interaction. Statistical methods are employed to analyze the collected data and establish a user-friendly robot-to-human handover interaction model. Verification experiments are conducted, involving new participants who experience the handover interaction model. Their feedback and experiences are collected and analyzed to validate the effectiveness and usability of the newly proposed interaction model.

Through this comprehensive methodology framework, involving quantitative methods, real robot experiments, questionnaire surveys, and verification experiments, this research aims to contribute to the development of an assistive robot and a new robot-to-human handover interaction model.

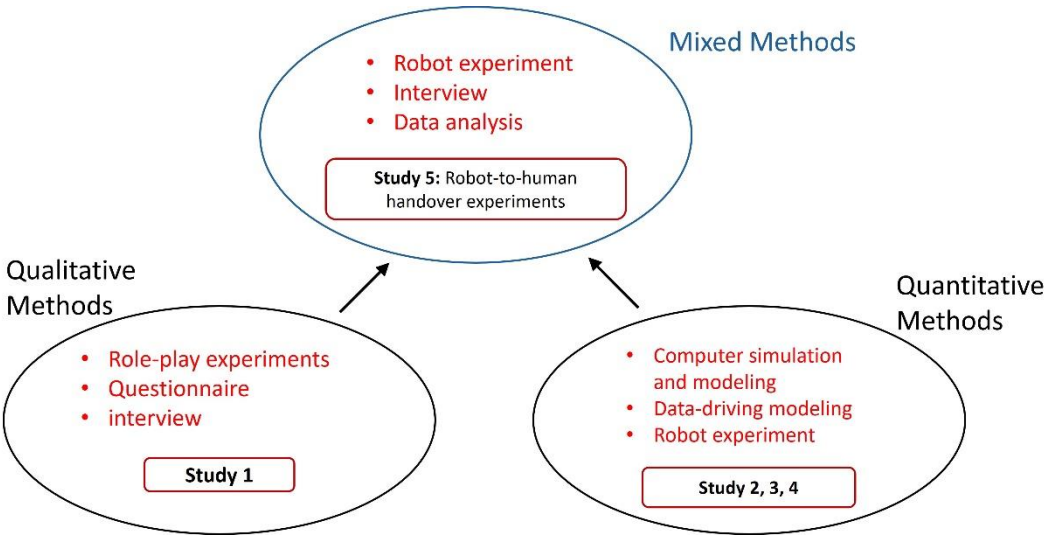


Figure 3.5. Methodology framework.

4. PROBLEM DEFINITION ON ROBOT-TO-HUMAN HANDOVER

In this chapter, I conduct role-play experiments to explore various issues that need to be studied in the robot-to-human handover scenario. The purpose of the simulation experiments is to understand the users' actual needs and identify the research issues involved in this task from both the perspectives of the robot and the user.

4.1. Introduction

The robot-to-human handover task involves both human and robot systems, encompassing various factors that require investigation. Therefore, I conduct simulation experiments to initially identify the key factors of concern for users and robots within this system, which would serve as the basis for subsequent research. In this simulation experiment, due to the lack of an actual robot system, I make a person simulate the role of the robot, interacting with individuals in the handover task. The performance of both the users and the simulated robot in the robot-to-human handover scenario is recorded during the execution of the simulation experiment. Additionally, interviews are conducted with the users to gather their requirements for the robot in specific situations and how they would like the robot to meet their needs. The purpose of this simulation experiment is to preliminarily identify the factors that need to be studied for the robot-to-human handover assistive robot from both the perspectives of the robot and the user. Although using a person to simulate the role of the

robot may introduce certain interference for the users and does not replicate a real robot-assisted scenario, it is only employed for the initial identification of my research and to elicit some requirements for the robot. A real robot handover scenario will be presented in Study 5.

4.2. Method

4.2.1. Robot-to-human Handover Simulation Experiments Design

The experimental scenario simulates the situation in a daily household setting where individuals who are sick need to retrieve everyday items, as shown in Figure 4.1. For the sick patients, their limited mobility often makes it difficult for them to fetch objects, typically requiring assistance from caregivers. In this scenario, a robotic system with grasping capabilities can play a significant role. It should be noted that the experimental scenarios assume that the participants have normal hand functionality, enabling them to grasp objects retrieved by the robot. In the simulation experiment, I aim to replicate such scenarios as closely as possible. Specifically, I invite five participants to simulate individuals who are sick in a home setting, while one participant assumed the role of the “robot”. The five simulated users are instructed to treat the simulated “robot” as a real one and interact with it accordingly. The main process of the simulation experiment involved the users informing the “robot” of the items they needed, and the “robot” retrieving and handing over the requested items. Throughout this process, the users are encouraged to make requests to the simulated “robot” and react to potential errors, such as picking up the wrong item or failed handovers. After several iterations of the simulation experiments, I conduct interviews with both the users and the simulated “robot” to understand their perspectives on the users' needs in this scenario, the role of the “robot”, and how they should interact with each other.

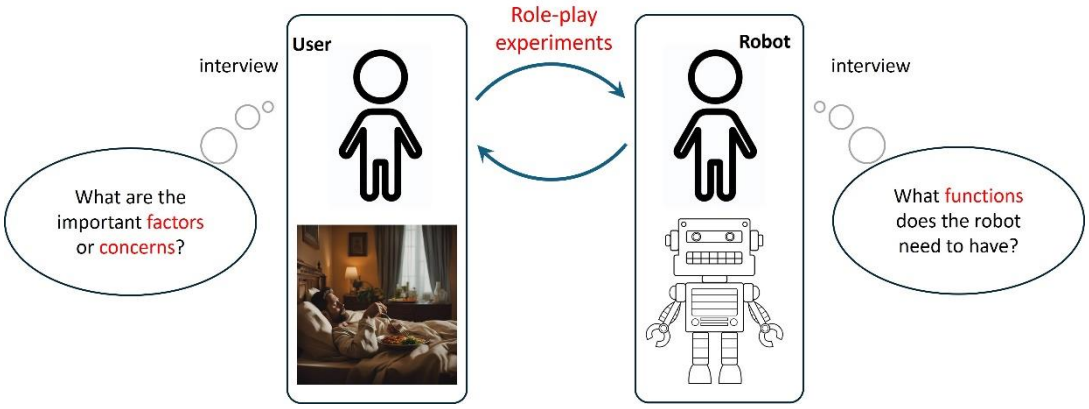


Figure 4.1. The role-play experiments.

4.2.2. Robot-to-human Handover Simulation Experiments Procedure

The experimental design aims to simulate a household environment where individuals with limited mobility due to illness require assistance in retrieving everyday items. This scenario is critical for evaluating the potential of robotic systems with grasping capabilities to aid such individuals, thereby reducing their dependency on caregivers.

Participants: five acting as individuals with limited mobility and one simulating the robotic system. The participants representing the sick individuals are instructed to interact with the simulated robot as if it is a real robotic assistant. The five participants representing individuals with limited mobility are aged between 25 and 30 years, comprising three females and two males. Each participant is briefed on their role and instructed to interact with the simulated "robot" as if they are genuinely experiencing mobility limitations due to illness. To enhance the realism of the simulation, the participants are asked to remain seated or in a stationary position throughout the experiment, mimicking the restricted mobility conditions. They are provided with a list of common household items they might need, such as a glass of water, a book, or a remote control, and are encouraged to request these items from the "robot" during the experiment. The participants are also instructed to react naturally to any errors made by the "robot", such as retrieving the wrong item or failing to complete a handover, to provide a comprehensive understanding of the interaction dynamics.

The role of the robotic system is assigned to one participant who is trained to simulate the actions of a "robot" with grasping capabilities. This participant is responsible for interpreting

the requests made by the individuals with limited mobility, retrieving the specified items, and handing them over. The participant simulating the robot is instructed to follow a predefined set of behaviors to ensure consistency across different iterations of the experiment. These behaviors included approaching the user, picking up the requested item, and attempting to hand it over in a manner that mimicked robotic movements. To maintain the integrity of the simulation, the participant acting as the robot is also briefed on how to handle errors. For instance, if an incorrect item is picked up, the simulated robot is to acknowledge the mistake and attempt to correct it. This aspect of the simulation is crucial for evaluating the error-handling capabilities and adaptability of the robotic system.

Procedure: The procedure for the role-play experiment is meticulously designed to replicate a realistic household environment where individuals with limited mobility require assistance in retrieving everyday items. The following steps outline the detailed procedure of the experiment.

Before the simulation began, all participants attended an orientation session. This session included an overview of the experiment's objectives, detailed instructions on the roles and responsibilities of each participant, a demonstration of the expected interaction protocols, and a Q&A segment to address any uncertainties and ensure all participants are comfortable with their roles. The experimental environment is set up to resemble a typical household setting. This included arranging common household items (e.g., a glass of water, a book, a remote control) in various locations within the room, ensuring that the participants acting as individuals with limited mobility are seated or in a stationary position to simulate restricted mobility conditions, and positioning the participant simulating the robot in a central location with clear access to all items. Each participant conducted at least 10 experiments, and at least 50 experiments in total.

The core of the simulation involved a series of tasks where the individuals with limited mobility requested items from the simulated robot. The procedure for each task is as follows: A participant acting as an individual with limited mobility would verbally request an item from the simulated robot. For example, "Can you please get me the book on the table?" The participant simulating the robot would acknowledge the request and proceed to locate and

retrieve the specified item. The simulated “robot” is instructed to follow a set of behaviors to ensure consistency, such as moving towards the item, grasping it, and returning to the requester. The simulated “robot” would attempt to hand over the retrieved item to the requester. This step is crucial for evaluating the handover process and the interaction dynamics between the robot and the user. If an error occurred (e.g., the wrong item was picked up or the handover failed), the simulated robot is instructed to acknowledge the mistake and attempt to correct it. The requester is encouraged to react naturally and provide feedback to the robot, such as, "This is not the book I wanted, can you please get the one with the blue cover?"

To provide a clear understanding, here is a specific example of a role-play experiment conducted in a home setting. In this scenario, the user is seated in a chair and requires the "robot" to fetch a cup. Upon realizing this need, the user verbally communicates to the "robot," saying, "Please bring me the cup." The "robot" first receives this instruction and, after clarifying the objective, searches the environment for the cup. Once the cup is located, the "robot" approaches it and reaches out to grasp it. After securing the cup, the "robot" returns to the user's side and informs the user that the cup has been retrieved. The user then instructs the "robot" to place the cup on the table. The "robot" complies by placing the cup on the table and retracting its hand. After determining that the task is complete, the "robot" returns to a designated standby location. The protocol of the experiment can be accessed in Figure 4.2.

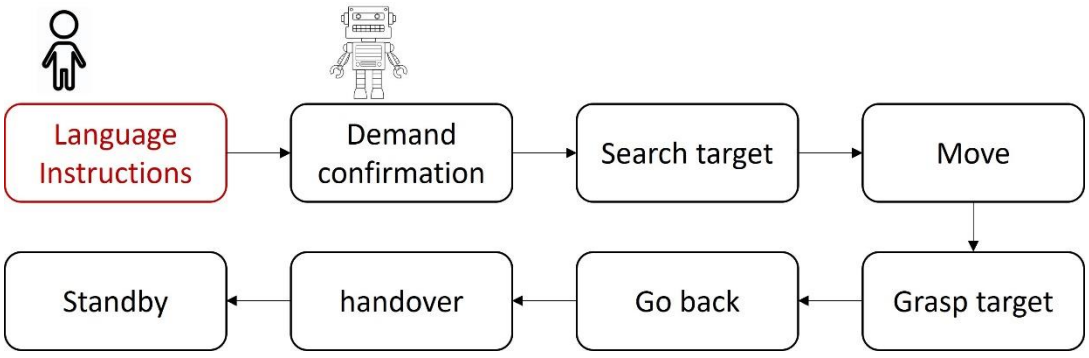


Figure 4.2. The protocol of role-play experiments.

4.3. Data Collection and Factors on Robot-to-human Handover Analysis

After completing the simulation tasks, post-experiment interviews are conducted with both the individuals with limited mobility and the participant simulating the robot. The interviews aimed to gather qualitative data on the users' needs and expectations from the robotic system, their perceptions of the robot's role and performance, their experiences during the interaction, including any difficulties or positive aspects, and suggestions for improvements in the robotic system.

I conclude the following procedures of the robot fetching the requested items and delivering them to the user's hand, based on the user's instructions, can be described as follows:

- **User Instruction**: The user communicates their request to the robot, either through voice commands or a user interface. For example, the user may say, "Robot, please fetch my book from the bedside table and hand it to me."

- **Perception and Object Recognition**: The robot utilizes its perception system, which may include cameras or sensors, to observe and analyze the surrounding environment. It identifies and recognizes the objects present, searching for the requested item based on its appearance or other distinguishing features.

- **Path Planning**: Once the target item is identified, the robot plans a suitable path to navigate through the environment, considering obstacles, furniture, and other potential hindrances. It calculates the optimal trajectory to reach the location of the requested item.

- **Object Grasping**: Using its robotic arm and gripper, the robot carefully grasps the desired item, taking into account its shape, size, and weight. It employs appropriate grasp planning algorithms to ensure a secure grip and minimize the risk of dropping or damaging the object.

- **Navigation**: The robot follows the planned path while carrying the item, avoiding any potential collisions or obstructions. It utilizes its localization and mapping capabilities to maintain an accurate understanding of its position within the environment.

- **Handover Interaction**: Upon reaching the user's location, the robot approaches the user and initiates a handover interaction. It may utilize visual or haptic cues to ensure a smooth and safe transfer of the object to the user's hand. The robot carefully releases its grip, allowing the user to securely receive the item.

- **Task Completion and Feedback**: Once the handover is successfully executed, the robot confirms the completion of the task, either through verbal acknowledgment or a visual indication. It may also seek feedback from the user to ensure their satisfaction and address any additional requests or needs.

Based on the simulation experiments, I identify the key research factors from both the perspectives of the robot and the user in the robot-to-human handover process. These findings will serve as the foundation for further research.

From the **perspective of the robot**, the robot needs to possess the following capabilities:

- **3D Object recognition**: In a robot-to-human handover scenario, once the robot receives instructions from the user, it needs to recognize the objects in the environment to locate the position of the desired item. With the rapid development of artificial intelligence, 3D visual perception methods have made remarkable progress in various fields, such as object detection and semantic segmentation. However, most existing deep learning-based methods require a significant amount of annotated data to train sophisticated models. This approach is inefficient in real-world robot scenarios, as these scenarios are highly dynamic, and manual data annotation for each instance is impractical. Therefore, it is essential to propose a data-efficient and rapidly deployable 3D perception method that can adapt to different scenarios on robots.

- **6-Dof grasp pose detection**: After recognizing the objects, the robot needs to generate the corresponding 6-DoF grasp pose and utilize this pose to successfully grasp the target object. In the field of 6-DoF grasp pose detection, researchers typically require a substantial amount of annotated data, including 3D models of objects and corresponding grasp pose information. However, manual annotation of such datasets is time-consuming and labor-intensive, thus limiting the feasibility of these methods in real-world robotics applications. Additionally, the current research predominantly concentrates on bin-picking scenarios, wherein objects are grasped from a cluttered container. Nevertheless, in many practical applications, robots need to perform precise grasping based on different shapes, sizes, and object characteristics. Therefore, there is a pressing need for further exploration of object-oriented grasp scenarios.

- **Object delivery motion control**: Once the robot successfully grasps the target object, it needs to proceed with the handover process, ensuring that the user can obtain the item. This process requires the robot to control its own movements based on the environment. It involves human-robot interaction, where the robot collaborates with the user to accomplish the transfer of the object. While there has been some research on anticipatory HRI in various contexts, such as object handovers, there is a gap in the literature when it comes to explicitly considering the concept of peer roles in robot-to-human handover scenarios. A real-time and online anticipatory Human-Robot Interaction (HRI) model that incorporates the concept of peer roles during robot-to-human handover scenarios. In these situations, there is a need for a model that enables the robot to anticipate and adapt its behavior based on the perceived role of the human counterpart in the handover interaction.

In addition to the mentioned three capabilities, the robot also needs to possess abilities such as command understanding, path planning, and obstacle avoidance. However, since these technologies have seen significant development and are relatively mature, I do not focus on researching them within the scope of this research. Instead, the focus is on the research of the three technologies.

From user's perspective, users pay attention to the following factors when interacting with the robot:

- **Objects needs to be grasped**: In a robot-to-human handover scenario, it is essential to consider the objects that need to be grasped and passed from the robot to the user. From the user's perspective, the desired items are those they specifically request for assistance. These objects can vary widely depending on individual needs and preferences. For example, an elderly or disabled individual may require the robot to hand over a medication bottle, a glass of water, a remote control, a book, or even personal belongings like a phone or a wallet. The successful handover of these objects plays a crucial role in ensuring the user's convenience, independence, and overall well-being. Understanding and addressing the specific objects that users desire to be handed over by the robot is central to designing an effective and user-centric handover system. By considering the user's perspective and incorporating their needs into the development process, the robot can provide personalized assistance and enhance the user's overall experience, promoting a more efficient and satisfactory human-robot interaction.

- **Success rate**: The success rate of a robot-to-human handover system refers to the ability of the robot to successfully deliver objects to users. From the user's perspective, the success rate holds significant importance as it directly impacts their experience and the effectiveness of the robot's assistance. A high success rate means that the robot consistently and reliably completes the handover process, ensuring that the requested items are safely and accurately transferred to the user. This reliability instills confidence in the user, assuring them that they can depend on the robot for their specific needs. A high success rate also signifies the efficiency and effectiveness of the handover system. It minimizes instances of dropped objects or failed transfers, which can cause inconvenience, frustration, and potential hazards for the user. Furthermore, a high success rate contributes to the user's sense of autonomy and independence. When the robot consistently delivers objects

successfully, users can rely on its assistance without hesitation, knowing that they can accomplish tasks and access necessary items without relying on human support.

- **Handover speed**: The move speed of a robot during the handover process refers to the rate at which the robot transfers objects into the user. A moderate and well-controlled move speed is essential to ensure a smooth and seamless handover. If the robot moves too slowly, it may lead to delays and frustration for the user, especially if they are waiting for an essential item or if they are in a time-sensitive situation. On the other hand, if the robot moves too quickly, it can create a sense of unease or discomfort for the user, potentially compromising safety during the handover process. The optimal move speed strikes a balance between promptness and user comfort. It allows the robot to swiftly deliver the requested items while ensuring a controlled and gentle transfer. This speed ensures that the user does not experience unnecessary delays and can rely on the robot's efficiency for their immediate needs.

- **Hand pose for receiving**: The hand posture of the user when receiving objects from a robot during the handover process is a crucial aspect to consider. The hand posture refers to the positioning and orientation of the user's hand as the robot transfers the item into their grasp. The user's hand posture plays a significant role in ensuring a successful and secure handover. By adopting an appropriate hand posture, the user facilitates a smooth transfer and minimizes the risk of dropping or mishandling the object. For example, having an open and stable hand with fingers slightly extended and ready to receive the item can enhance the success and ease of the handover process. Figure 4.3 shows some possible hand poses for receiving the item.

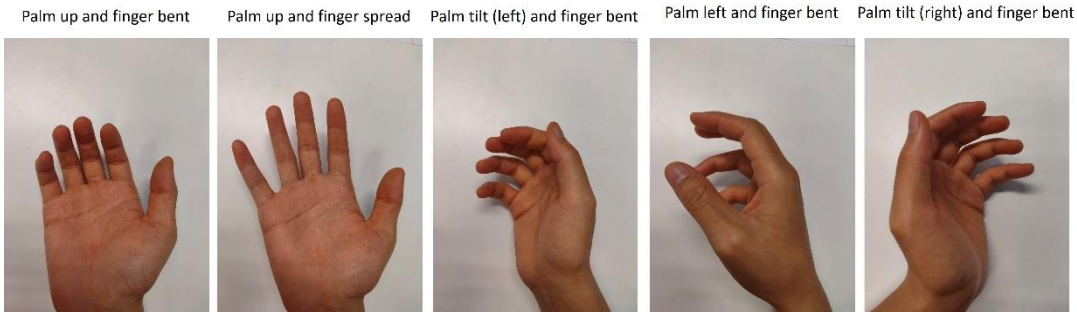


Figure 4.3. Some hand poses for receiving.

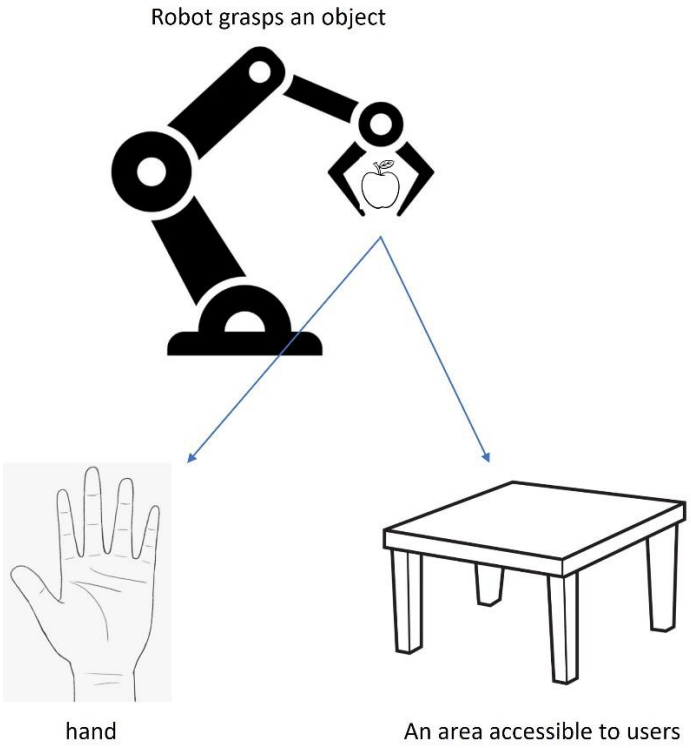


Figure 4.4. Receive modes.

- **Receive mode**: In a robot-to-human handover scenario, the user's grasping area refers to the region available for the user to grasp the object, which is determined by the robot after it completes the grasping action and passes the item to the user. After the robot successfully grasps the object, it carefully positions and presents the item within the reachable range of the user's hand. The specific location and orientation of the object within the user's grasping area may vary depending on factors such as the size and shape of the object, as well as the robot's hand design. Alternatively, the robot can place the object onto a fixed platform within the user's reach, as shown in Figure 4.4. This platform can be a tray, shelf, or any other stable surface where the object is securely deposited. By placing the item on the fixed platform, the robot enables the user to independently retrieve it at their convenience.

- **Robot path**: In a robot-to-human handover scenario, the motion path of the robot's manipulator arm plays a crucial role in transferring objects to the user. This path directly influences the efficiency of the handover process and the user's comfort. One

approach to enhance the handover experience is for the robot to mimic human-like motion path when transferring objects, as shown the Path 2 in Figure 4.5. By simulating human motion paths, the robot aims to create a more natural and intuitive interaction with the user. This involves carefully planning the path and motion of the robot's arm as it approaches, grasps, and delivers the object to the user. The path can be designed to follow smooth and fluid movements, resembling those of a human arm during a handover. Another path, as shown in Figure 4.4 as Path 1, involves the robot grasping the object and then returning to a fixed position before transferring it to the human's hand.

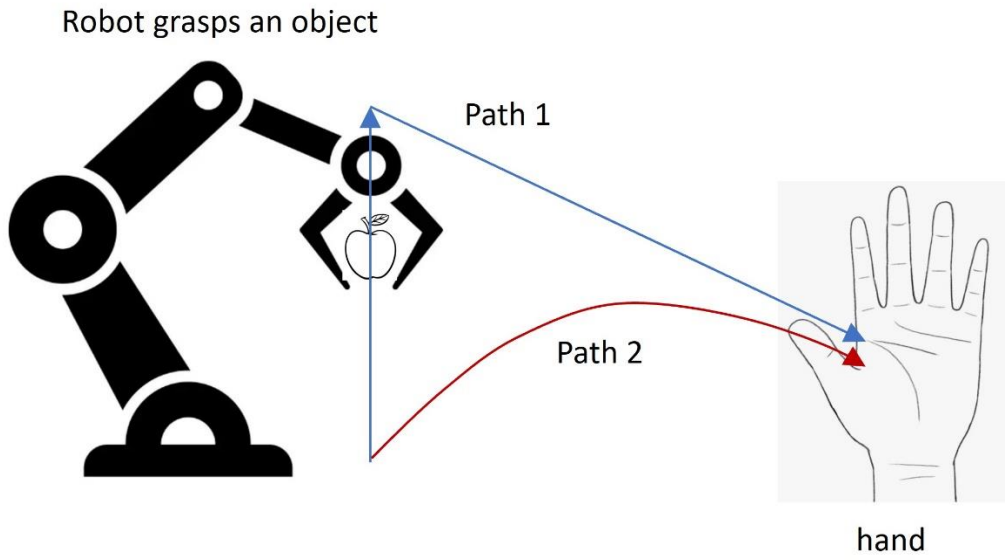


Figure 4.5. Robot path.

- **Trust, comfort, and safety**: In a robot-to-human handover scenario, users' perceptions of trust, comfort, and safety play a vital role in their overall experience with the handover robot. Users' trust in the handover robot is influenced by factors such as reliability, predictability, and accuracy of its actions. When the robot consistently and successfully transfers objects without errors or mishaps, users develop a sense of trust in its capabilities. User comfort during the handover process is crucial for a positive experience. The robot should consider ergonomic factors, such as positioning the object within the user's natural reach and providing appropriate support during the handover. Smooth and fluid motion trajectories,

gentle grasping and release actions, and a user-friendly interface contribute to user comfort. Users' perception of safety is paramount when interacting with a handover robot. They need to feel confident that the robot's actions won't cause harm or injury. Safety features such as collision avoidance, force sensing, and robust grasping mechanisms are essential to ensure safe handovers.

- **Robot's ability to cope with errors**: In the robot-to-human handover scenario, it is possible for the robot to make errors, such as picking up the wrong item or experiencing failed handovers. Users are highly concerned about these error situations and express the need for the robot to receive signals from the user, such as voice or gesture cues, to address these errors. For example, when the robot picks up the wrong item, the user should be able to inform the robot of the mistake, prompting the robot to put the incorrect item back in its original position and retrieve the correct item instead.

4.4. Discussions

This chapter conducts simulation experiments of robot-to-human interactions to identify the capabilities required of robots and the needs of users in this human-robot interaction task. Regarding the robot's capabilities, an analysis of the experimental data led to the identification of three primary functions: (1) 3D object recognition, (2) 6-DoF grasp pose detection, and (3) object delivery motion control. These capabilities drive Studies 2, 3, and 4. Following the completion of Studies 2, 3, and 4, a robotic system equipped with object recognition, grasping, and delivery capabilities is established. Subsequently, using this physical robotic system, research on robot-to-human handover interaction models is conducted, and a new robot-to-human handover interaction model is proposed based on user experiences in Study 5. Despite being simulation experiments, the research in this chapter provides motivation and direction for following studies.

I understand the concern regarding the use of humans to simulate robots, particularly in terms of differences in flexibility and degrees of freedom. While it is true that humans and robots differ significantly in these aspects, the primary objective of this experimental design

is not to replicate robotic behavior precisely, but rather to explore the interaction dynamics and identify key issues from both the user and robot perspectives. The role-play setup allows us to simulate the robot-to-human handover process in a controlled environment, providing valuable insights into user expectations and potential challenges that may arise during interactions. By focusing on the interaction process, I can identify critical factors that influence user satisfaction and robot performance, which are essential for defining the research questions and guiding the development of more effective robotic systems.

Although this approach has its limitations, it serves as a foundational step in understanding the complexities of human-robot interaction. The insights gained from these experiments will inform subsequent studies, where more sophisticated simulations or actual robotic systems can be employed to address the identified issues in greater detail. Thus, while acknowledging the limitations, role-play experiments are a reasonable and valuable method for defining the research problems in this context.

5. ROBOTS RECOGNIZE THE WORLD USING 3D OBJECT DETECTION

This chapter explores the method by which robots perceive the world using 3D object detection techniques. I propose a method that exhibits the characteristic of quickly adapting to various robotic tasks. Relying heavily on 3D annotations restricts the practical application of 3D object detection. To address this issue, I introduce a technique that eliminates the need for any 3D annotation while still being capable of predicting fully oriented 3D bounding boxes. This technique, named Recursive Cross-View (RCV), leverages the three-view principle to transform 3D detection into several 2D detection tasks, requiring only a portion of 2D labels. I propose a recursive framework where instance segmentation and 3D bounding box creation via Cross-View are performed iteratively until they converge. Specifically, the method uses a frustum for each 2D bounding box, followed by the recursive process that eventually produces a fully oriented 3D box along with its associated class and score. Note that the class and score are provided by the 2D detector. Evaluations on the SUN RGB-D and KITTI datasets show that this method surpasses existing image-based techniques. To demonstrate the method's adaptability to new tasks, I apply it to two real-world scenarios: 3D human detection, and 3D hand detection. Consequently, two new 3D annotated datasets are created, indicating that RCV can function as a (semi-) automatic 3D annotator. Additionally, I implement RCV on a depth sensor, achieving detection at 7 frames per second on a live RGB-D stream. RCV is the first 3D detection method to produce fully oriented 3D boxes without the use of 3D labels.

5.1. Introduction

3D object detection focuses on identifying, classifying, and generating 3D bounding boxes for objects within a scene. With the progress in 3D sensors, annotated datasets, and deep learning, 3D detection has seen significant advancements recently. It holds potential for various applications, including autonomous driving, robotic navigation, manipulation, and human-robot interaction.

In this study, I introduce a straightforward yet powerful 3D detection method called RCV, which operates without the need for 3D annotations. Figure 5.1 provides an overview of RCV. By leveraging the concept of three-view drawings, I transform 3D detection into multiple 2D detection tasks. Using a recursive approach, RCV can perform instance segmentation and predict fully oriented 3D bounding boxes. RCV has several benefits over existing methods. Firstly, it does not depend on 3D annotations, enhancing its real-world applicability. Thanks to advanced 2D detectors, RCV only needs a few 2D labels to achieve 3D detection and can inherit the robustness and generalization properties of 2D detectors. Secondly, RCV can predict fully oriented boxes, broadening its range of applications. Thirdly, RCV can be rapidly deployed to new 3D sensors in various real-world environments. Additionally, once trained, RCV can serve as a 3D annotation tool, simplifying manual labeling or creating datasets for pretraining.

Evaluations on the SUN RGB-D (Song et al. 2015) and KITTI (Geiger et al. 2012) datasets show that this method surpasses existing image-based techniques. Notably, our method is highly data-efficient, significantly outperforming existing image-based methods in the 3D detection of Pedestrians and Cyclists in KITTI using only 25% of the training data. In real-world experiments, RCV achieves 3D detection by training solely on 2D images and labels. Once trained, RCV can function as a (semi-) automatic 3D annotator, resulting in the creation of three new datasets. This method offers a practical solution by using some 2D annotations to achieve 3D detection, marking a significant contribution of this work.

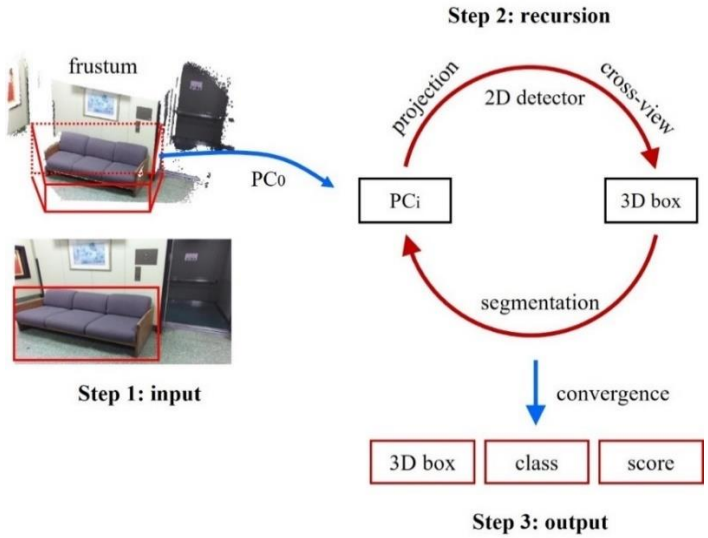


Figure 5.1. Overview of RCV. Step 1: execute 2D detector on an image and propose frustums on the point cloud. Step 2: perform recursion. Step 3: output. Note that, class and score are given by 2D detector. See Figure 5.4 for more details on the recursion.

5.2. A Novel 3D Object Detection Method: Recursive Cross-View

5.2.1. Conversions between 3D and 2D for Objects

Given the labor-intensive and costly nature of manually annotating 3D bounding boxes, there is a strong incentive to find alternative methods that allow 3D detection algorithms to be quickly adapted to new scenarios and objects. RCV draws inspiration from the principles of engineering drawing, where a 3D object can be comprehensively represented by three views (illustrated in the left-top subimage of Figure 5.2) and vice versa (shown in the right-top subimage of Figure 5.2). This relationship implies that a 3D object can be reconstructed using only three 2D views, which is the core concept behind RCV. But how do we derive a 3D bounding box in this context? In 3D object detection, the goal is to generate a 3D bounding box for each object rather than reconstructing every detail. A 3D bounding box can be obtained by determining the size and position of the 2D bounding boxes and then applying the three-view mechanism, as depicted in the left-bottom subimage of Figure 5.2. This process eliminates the need for 3D annotations. Another advantage of RCV is its ability to directly detect fully oriented bounding boxes without increasing detection complexity, as it does not rely on regression. To demonstrate this benefit, I create a "3D HAND" dataset, which includes annotations for fully oriented boxes.

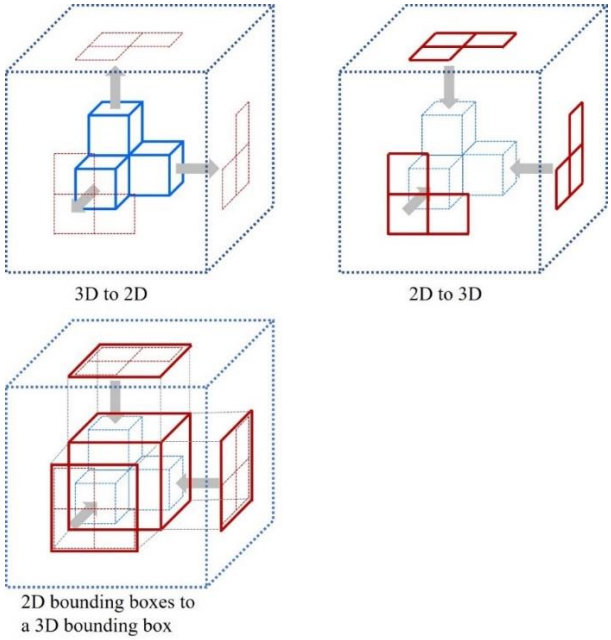


Figure 5.2. Conversion between 3D and 2D. The left-top and right-top subimages are three views, and the left-bottom subimage is the derivation of 3D bounding box from three views.

5.2.2. Perspective View

RCV is designed to incrementally remove points that do not belong to the target object while retaining volumetric regions occupied by the object but not captured by the depth camera. Ultimately, a bounding box is created for each object. With this approach, I outline the steps of RCV. The initial step, termed "Perspective View," involves processing raw data, such as RGB images and point clouds obtained from the depth camera. This step is akin to the method used by F-PointNets (Qi et al. 2018), where a frustum is derived based on the depth camera's projection matrix and the 2D bounding box. Figure 5.3 depicts this process, the initial step involves using a 2D object detection model on an RGB image to identify objects and their corresponding bounding boxes. The principle, as shown in Figure5.3, is to leverage the 2D detection to inform 3D space exploration. Each detected 2D bounding box is projected into 3D space, creating a frustum that extends along the depth axis. This frustum acts as a spatial filter, narrowing down the search area by focusing only on the region where the object is likely to be located. However, RCV employs fundamentally different concepts from F-PointNets beyond this operation. The reasons for starting with the perspective view include: (1) it is straightforward to capture RGB images from cameras, and (2) the perspective view covers a larger area than the orthographic view, which is beneficial for detecting objects in extensive scenes, such as in autonomous driving. After this step, a very coarse 3D bounding

box that encloses the point cloud within the frustum can be generated, though it is not always necessary. This depends on the specific scenario; for instance, I generate bounding boxes in the first step for the SUN RGB-D benchmark due to significant object occlusion and background clutter. However, coarse 3D boxes are not generated in the initial step for experiments conducted on the KITTI dataset. It is important to note that RCV utilizes YOLOv5 as the 2D detector throughout all steps.

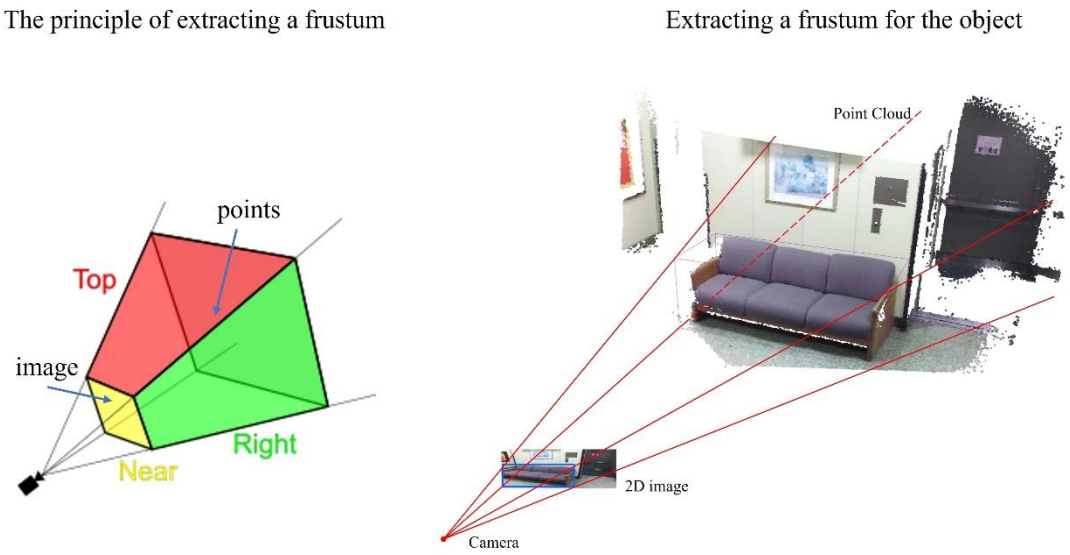


Figure 5.3. Perspective view. A 2D bounding box can be obtained from 2D detector, then a frustum can be derived.

5.2.3. Recursive Orthographic Cross-View

For each frustum derived from a perspective view, I utilize a divide-and-conquer strategy to detect objects concurrently. Subsequently, I recursively apply the orthographic Cross-View method to create the corresponding bounding box for each object, following the principle illustrated in the bottom-left subimage of Figure 5.2. Any pair of the three views can be used to generate a definitive box. Hence, in all experiments, I opt for the front-view and side-view. Figure 5.4 illustrates the recursive process. The point cloud shown at the top of Figure 5.4 is derived from the frustum created by a 2D box. I then project these points along orthogonal axes to produce two RGB images, as depicted in the second row of Figure 5.4, with red arrows indicating the projection directions. Following this, I employ YOLOv5 to detect objects in these two images, as shown in the second row of Figure 5.4. Consequently, points not identified as objects are discarded, and a more precise box is obtained by performing the

Cross-View, as seen in the point cloud in the middle of Figure 5.4. Recursively, I repeat these steps on the remaining point cloud: (1) calculating the projection axes, (2) projecting RGB images for both views, (3) conducting 2D detection to eliminate external points, and (4) obtaining a new box through Cross-View, until the process converges.

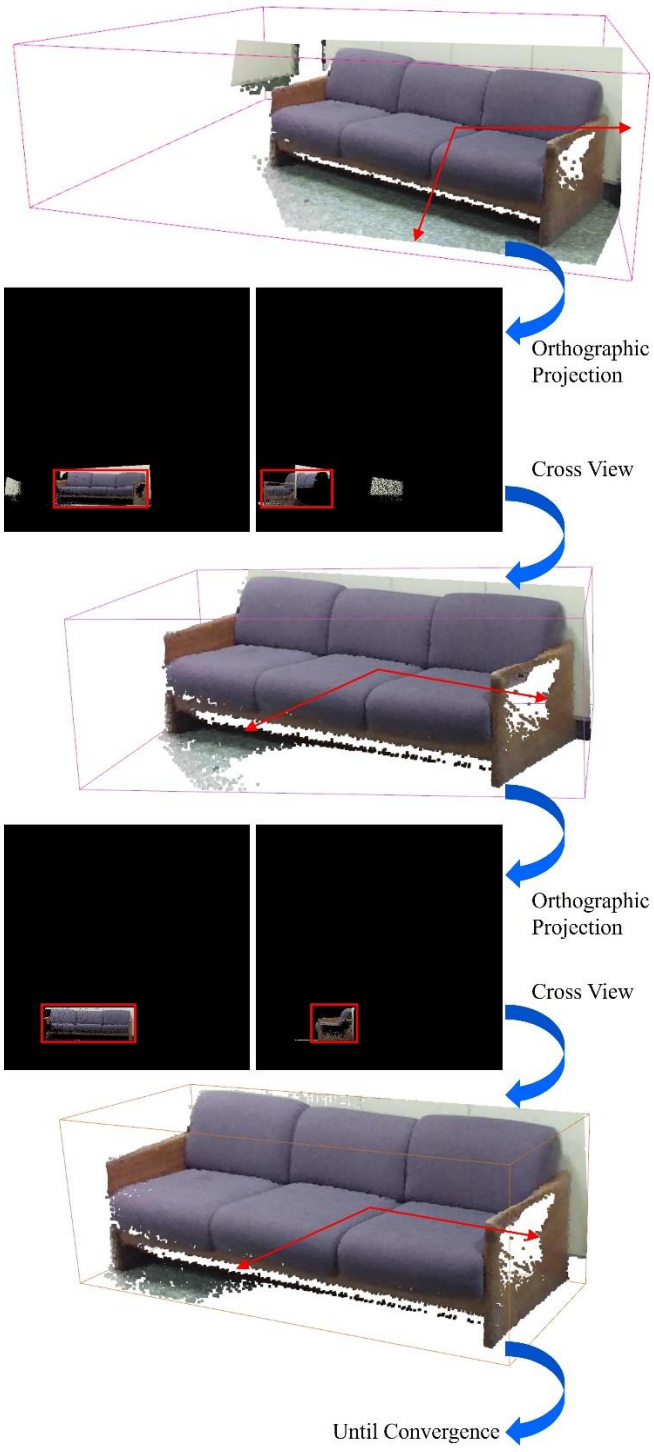


Figure 5.4. Recursively Cross-View. Red arrows indicate orthographic view direction, blue curved arrows indicate projection and Cross-View that generates a 3D bounding box.

Convergence Conditions Several criteria are established to end the recursion. One criterion involves the projection axes, represented by the red arrows in Figure 5.4. As the recursion progresses, the directions of these axes stabilize. The recursion halts when their variation falls below a specified threshold. Another criterion is the change in the 3D bounding box; the recursion stops when the difference between two successive boxes is less than a certain threshold. A third criterion is to empirically set a fixed number of recursion steps. The method for calculating the axes is detailed in the following section.

Pseudo-view Images The point cloud captured by the depth camera includes spatial (XYZ) and color (RGB) data. The orthographic projection maps each point to a corresponding pixel based on its spatial data while retaining its color information. This process results in the creation of projected images, which I refer to as 'pseudo-view' images.

Multi-object Detection in One Frustum Occlusion can lead to multiple objects being present within a single frustum. Therefore, RCV is designed to detect multiple objects in the pseudo-view images generated from the frustum's point cloud. Only detections that match the label of the 2D bounding box from the original image that proposed the frustum are retained. For instance, the topmost point cloud in Figure 5.4 originates from the frustum suggested by the 2D box in Figure 5.3, labeled "sofa." Consequently, I only detect multiple boxes labeled "sofa" in the projected images in the second row of Figure 5.4. Subsequently, one or more 3D boxes are generated through Cross-View, followed by a recursive process for each, as illustrated in Figure 5.4. Conversely, only one box is retained during subsequent detections, which is then used to refine the corresponding 3D bounding box.

3D Bounding Boxes For each point cloud depicted in Figure 5.4, I first determine the projection axis and then use the Cross-View method to derive the subsequent point cloud and bounding box. This process allows for the calculation of the transformation matrix (T_n^{n+1}) between two consecutive sets of point clouds and boxes, based on the coordinate system defined by the projection axes. Specifically, the middle point cloud in Figure 5.4 is projected along the projection axis (indicated by red arrows), and the resulting bottom point cloud and box are obtained through the Cross-View technique. Conversely, the bottom point cloud and box can be transformed back into the coordinate system of the middle point cloud using the

transformation matrix derived from the previous projection axis. As a result, I can map the final bounding box back to the original point cloud system using Eq. (5.1):

$$B_0 = \prod_{i=0}^{N-1} T_i^{i+1} B_N \tag{5.1}$$

where B_N, 4 by 8 matrix, is the box generated at the Nth step of the recursion. B_o, 4 by 8 matrix, is the box corresponding to the original point cloud system. T_i^{i+1} is a homogeneous transformation matrix with 4 by 4. Finally, I utilize non-maximum suppression (NMS) algorithm for all detected 3D bounding boxes, filtering the redundant detections.

5.2.4. Projection Axes

Camera Coordinate Axes Using the camera's coordinate axes as projection directions is a straightforward approach. These axes are applied to all point clouds extracted from the frustum, which is the initial step in the recursion process, as shown in the top point cloud of Figure 5.4. This method is effective because the point cloud within the frustum is likely to suffer from significant occlusion and a cluttered background. Performing projection and Cross-View without any transformation allows for quick preliminary detection.

Eigenvectors of the point cloud can provide a rough indication of its orientation, making them suitable as projection axes. However, a limitation exists: they may not accurately represent the orientation of an object if only a small portion of the point cloud is available..

Normal Vectors Normal vectors can effectively represent the orientation of an object, even when only a partial view is available. In our experiments, I use normal vectors as the projection axes. Specifically, I employ the K-Means algorithm to determine the primary normal vector of the point cloud.

5.3. Real-world 3D Object Detection Experiments on Robotics

I conduct four distinct experiments: (1) evaluating 3D detection performance on the SUN RGB-D dataset, (2) assessing data efficiency using the KITTI dataset, (3) testing a 3D annotation tool on my own data, and (4) performing real-time detection with a depth camera. As mentioned earlier, I do not develop a new neural network. Therefore, my approach

involves training YOLOv5 solely on the projected images, which is a straightforward process. It is important to note that I use the default hyperparameters for all experiments.

5.3.1. Obtaining 2D Annotations and Training Data

In this section, I outline the method for generating 2D bounding boxes for both public datasets and my own datasets. For public datasets, where 3D bounding boxes are already annotated, the 2D bounding boxes are derived through the orthogonal projection of these 3D bounding boxes, as illustrated in Figure 5.5. The resulting images and 2D bounding boxes are then used to train a 2D detector, which is subsequently employed to develop an RCV model. Although 3D bounding boxes are used to create the 2D bounding boxes, they are not directly involved in the training of the model. Next, I will explain the process of labeling 2D bounding boxes on our proprietary dataset without relying on any 3D bounding boxes.

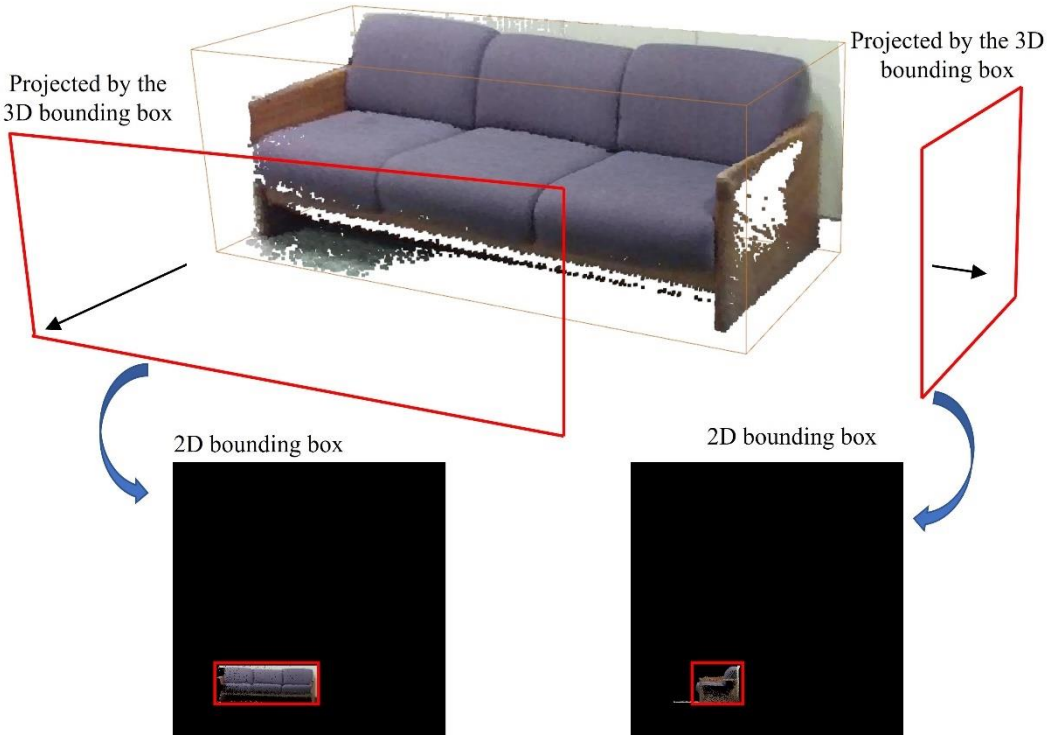


Figure 5.5. 2D bounding boxes projected by the 3D bounding box for SUN-RGBD dataset.

Manually annotating 3D bounding boxes is a highly challenging task. Here, I illustrate how to create 2D annotations for various scenarios or tasks using this method. As an example, I use an indoor 3D human dataset. Initially, an image and a point cloud are captured by a depth

camera. I then label 2D bounding boxes on the image, as shown in the first row of Figure 5.6. Next, the points within the frustum are retained and used for projection to generate 2D images, as seen in the second row of Figure 5.6. This process is repeated to produce two additional 2D images, as shown in the last row of Figure 5.6. I refer to these 2D images as 'pseudo-view' images. Annotating 2D bounding boxes on these images is straightforward, and these annotations are then used to train a 2D detector. Ultimately, this allows me to develop an RCV model capable of detecting 3D humans. In Section 5.3.4, I annotate 1,600 2D bounding boxes for 3D human detection and 530 2D bounding boxes for fully oriented 3D hand detection.

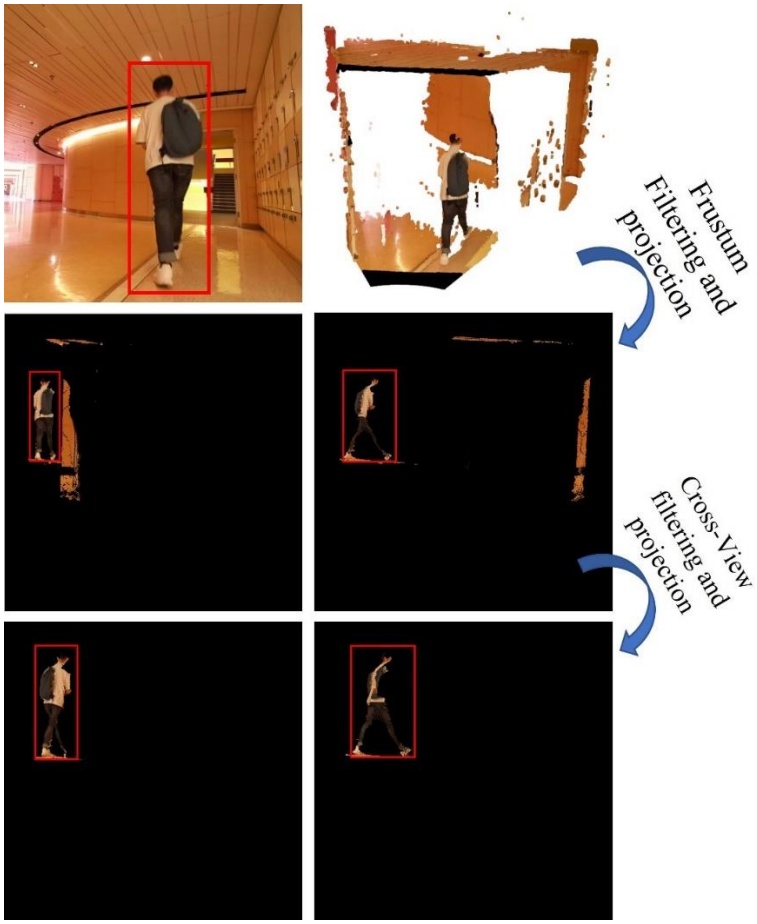


Figure 5.6. Manually 2D bounding box labeling method on our own dataset.

It is important to highlight the fundamental differences between the 2D annotation strategy and the conventional 3D annotation approach. Although it is possible to form a 3D box from two 2D boxes taken from orthogonal views, this does not necessarily ensure the quality of the resulting 3D box, as the correct projection axes may not be known. For instance, in Figure

5.4, the middle 3D box is not sufficiently accurate to serve as a 3D annotation. The purpose of 2D annotations is solely to enable a 2D detector to identify objects in the projection images, irrespective of the projection direction.

To achieve a high-quality 3D bounding box, I employ a recursive process that iteratively refines the 3D box until convergence. During each iteration, points outside the box are discarded, and new projection axes—essentially the orientation of the 3D box—are determined. Only the projection images from the first two steps in the 'Recursion' process are labeled, which is insufficient for generating a precise 3D annotation. Similarly, in datasets like SUN RGB-D and KITTI, 2D boxes are derived from annotated 3D boxes, but this method still does not capture the correct orientation of the object. Consequently, this approach does not directly utilize 3D annotations from these public datasets.

5.3.2. Experiments on SUN RGB-D

SUN RGB-D is an indoor 3D dataset comprising 5,285 training samples and 5,050 testing samples. I conduct a comparative experiment on this dataset, focusing on monocular 3D detection methods. The proposed method uses several images as input, including a raw RGB image and multiple 'pseudo-view' images generated through point cloud projection. Consequently, RCV shares similarities with some image-based 3D object detection methods. Monocular detection techniques have been developed to identify 3D objects by combining a single image with geometric features or 3D world priors. Additionally, some methods integrate a monocular image with depth maps to enhance 3D object detection accuracy.

To compare RCV with monocular detection methods, I evaluate RCV on SUN RGB-D for 10 out of 37 object categories (Nie et al. 2020). First, I convert all 3D objects into 2D images and 2D bounding boxes following the method described in Section 5.3.1, resulting in over 100,000 images for training and more than 70,000 images for validation. Table 5.1 provides further details on training the 2D detectors. In the experiment, I observe significant differences between the raw RGB images and the 'pseudo-view' images. Therefore, I train two separate 2D detectors: one for RGB images and another for 'pseudo-view' images. This setup is used for all experiments. Once the 2D detectors are trained, I can formulate RCV and use

it to detect 3D bounding boxes on the SUN RGB-D validation set. Table 5.2 shows that RCV outperforms all previous methods, achieving state-of-the-art performance on this benchmark without directly using 3D annotations. It is worth noting that IM3D (Zhang et al. 2021c) utilized additional data for training, so I do not compare our method with it.

Table 5.1: Settings of training YOLO for 10 out of 37 object categories (Nie et al. 2020) in SUN-RGBD. The first row is the setting of the first step detection model, and the second row is the setting of the recursive detection model.

	Model	Train no.	Val no.	Size	Device
1	YOLOv5x6	27044	5050	/	3090
2	YOLOv5x7	70924	69848	640	3090

5.3.3. Data Efficiency on KITTI

To showcase the data efficiency of our method, I conduct experiments on the KITTI dataset. Given that the Pedestrian (4,487 samples) and Cyclist (1,627 samples) categories are considerably smaller than the Car category (28,742 samples), these categories are selected as benchmarks for this experiment. This choice allows for an effective evaluation of the proposed method's capability to handle smaller datasets. The proposed method is trained using different proportions of the available training data: 80%, 50%, and 25%. The performance of the trained models is then assessed on the KITTI test set. Table 5.3 presents the 3D detection results for Pedestrian and Cyclist categories on the KITTI test set. The method significantly outperforms previous state-of-the-art monocular-based methods across all evaluated categories, even when using only 25% of the training data.

Table 5.2: 3D detection performance on SUN-RGB-D val. set for 10 out of 37 object categories (Nie et al. 2020). The metric is average precision with 3D IoU threshold 0.15. We compare our scores with previous state-of-the-art monocular detection method. Bold is used to highlight the best results. * means the method (IM3D) utilized extra data to train the model.

Method	Input	Label	Sink	Bed	Lamp	Chair	Desk	Dresser	Nightstand	Sofa	Table	Cabinet	mAP	Runtime
ImVoxelNet (Rukhovich et al. 2022)	Mono. + cam. Pose	2D+3D	45.12	79.17	13.27	63.07	31.20	**35.45**	38.38	60.59	**51.14**	19.24	43.66	0.14
T3DU (Nie et al. 2020)	Mono. + geo.	2D+3D	18.05	60.65	5.04	17.55	27.93	21.19	17.01	44.90	36.48	14.51	26.38	/
IM3D (Zhang et al. 2021c)	Mono. + extra data	2D+3D	33.81	**89.32**	11.90	35.14	**49.03**	29.27	41.34	**69.10**	57.37	**33.93**	45.21*	/
Perspective Net (Huang et al. 2019)	Mono.	2D+3D	41.35	79.69	13.14	40.42	20.19	/	/	62.35	44.12	/	/	/
Ours	Mono. + pseudo-view	2D	**65.44**	76.32	**22.48**	**70.66**	18.06	32.02	**56.19**	58.71	42.85	6.80	**44.95**	**0.12**

Table 5.3: 3D Detection performance of Pedestrian and Cyclist on the KITTI test set. Bold is used to highlight the best results.

Method	Data	Input	AP_{R40} [Easy / Mod / Hard]		AP_{R40} [Easy / Mod / Hard]	
			AP30@IoU = 0.5	AP_{BEV}@IoU = 0.5	AP_{3D}@IoU = 0.5	AP_{BEV}@IoU = 0.5
LPCG-Monoflex (Peng et al. 2022b)	100%	Image+LiDar	10.82 / 7.33 / 6.16	12.11 / 7.92 / 6.61	6.98 / 4.38 / 3.56	8.14 / 4.90 / 3.86
DEVIANT (Kumar et al. 2022)	100%	Image+depth	13.43 / 8.65 / 7.69	14.49 / 9.77 / 8.28	5.05 /3.13 / 2.59	6.42 / 3.97 / 3.51
DD3D (Park et al. 2021)	100%	Image+depth	13.91 / 9.30 / 8.05	15.90 / 10.85 / 8.05	7.52 / 4.79 / 4.22	9.20 / 5.69 / 5.20
PS-fld (Chen et al. 2022a)	100%	Image+LiDAR	16.95 / 10.82 / 9.26	19.03 / 12.23 / 10.53	11.22 / 6.18 / 5.21	12.80 / 7.29 / 6.05
OPA-3D (Su et al. 2023)	100%	Image+LiDar	15.65 / 10.49 / 8.80	17.14 / 11.01 / 9.94	5.16 / 3.45 / 2.86	6.01 / 3.75 / 3.56
MonoDTR (Huang et al. 2022)	100%	Image+LiDar	15.33 / 10.18 / 8.61	16.66 / 10.59 / 9.00	5.05 / 3.27 / 3.19	5.84 / 4.11 / 3.48
Ours	80%	Image+pseudo-view	40.19 /**31.89**/**28.32**	**52.26**/42.93/37.34	**20.02** /**13.93**/**12.48**	**28.51**/**21.82**/**18.94**
	50%		**40.85**/31.60/27.96	51.14/**44.11**/**38.39**	16.66/13.17/11.18	21.70/17.70/15.28
	25%		37.50/30.24/26.72	50.08/43.52/38.03	13.69/11.22/9.45	19.55/15.80/13.60

5.3.4. 3D Annotator Using RCV

To demonstrate that our method does not rely on any 3D annotations and can function as an automatic 3D annotator, I create two annotated datasets named "3D HUMAN" and "3D HAND" using RCV. All data is collected with an Azure Kinect DK. Following the 2D annotation method shown in Figure 5.6, I label 1,600 2D bounding boxes for "3D HUMAN" and 530 2D bounding boxes for "3D HAND." These 2D bounding boxes are labeled on 'pseudo-view' images. Annotations for the original images are not detailed as they are straightforward. After training, I develop two RCV models capable of producing 3D bounding boxes for humans and hands, respectively. "3D HUMAN" includes fully annotated humans in approximately 30 indoor scenes, as illustrated in Figure 5.7. It comprises around 1,500 frames of data, each containing an RGB image, a point cloud, and one or more 3D bounding boxes, totaling over 4,500 3D bounding boxes generated by RCV.

"3D HAND" features fully annotated hands from 8 participants, consisting of 1,500 frames of data. This dataset includes about 1,500 fully oriented 3D bounding boxes generated by RCV, as shown in Figure 5.8. I argue that these final datasets can be used to pretrain some 3D detection models after minor manual selection and adjustment. Therefore, I believe that using RCV as a preliminary 3D annotation tool is feasible. In the future, I plan to train some 3D detectors on my own datasets. Similarly, if one aims to achieve 3D object detection in different scenarios, the same steps can be followed using RCV. The process involves collecting data and labeling some 2D images, which is much simpler compared to 3D labeling.

5.3.5. 3D Detection on A Depth Camera

To demonstrate that the proposed method can perform real-time detection in practical scenarios, I deploy RCV on an Azure Kinect DK. Specifically, I utilize the hand detection model detailed in Section 5.3.4 to identify a hand and generate a fully oriented 3D bounding box. The system operates at a frequency of approximately 7Hz.

Figure 5.7. 3D boxes generated by RCV on '3D_HUMAN'.

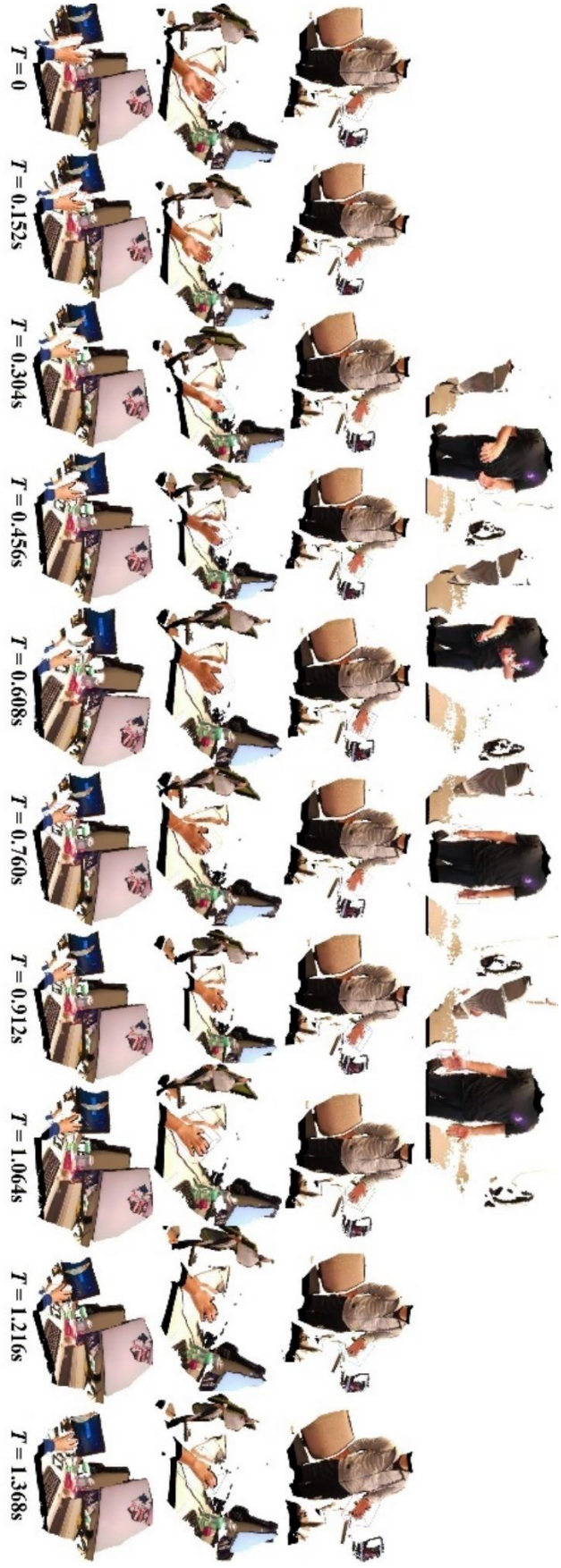


Figure 5.8. 3D boxes generated by RCV on ’3D_HAND’.

5.4. Discussions

5.4.1. Imitate 3D Labeling Process

In essence, the proposed method mimics the manual 3D labeling process described by Chen et al. (2017). In the manual process, an annotator initially places a rough bounding box and then iteratively rotates the box while manually identifying 2D bounding boxes from three different views to achieve a 3D annotation. Our method replicates this process by substituting 'rotate the box' with generating projection axes (as detailed in Section 5.2.4), 'manually detects 2D bounding boxes' with YOLOv5, and 'multiple times' with a recursive procedure. Ultimately, our method automates the entire 3D annotation process.

5.4.2. Automatic Labeling Pipeline and Datasets

The method enables (semi-) automatic generation of 3D annotations starting from a few 2D bounding boxes, which is a significant practical advantage for scenarios lacking any annotated data. Unlike some existing semi-automatic labeling pipelines, such as H2O (Kwon et al. 2021), which relies on the pre-trained DenseFusion model (Wang et al. 2019) for data labeling, our pipeline offers a distinct practical benefit. The resulting datasets can be utilized to train various existing 3D detectors, which we plan to explore in future work. Importantly, the 3D HAND dataset includes fully oriented 3D box annotations.

5.4.3. Limitations

In certain instances, the method might not converge, leading to unsuccessful object detection. This can occur due to: (1) Subpar performance of the trained 2D detector, which eliminates points associated with the object during each iteration, hindering system convergence, and (2) the 2D detector's inability to identify the object in the projected images, causing the detection process to halt prematurely. The experiments indicate that the first scenario is infrequent, whereas the second scenario is more prevalent. The effectiveness of our method is significantly dependent on the 2D detector's accuracy.

Additionally, I have noticed that the proposed method tends to underperform with larger objects. A plausible reason is that larger objects are less likely to have a substantial portion of their regions captured by the camera, which could negatively impact performance. This hypothesis, however, requires further experimental validation in future studies.

6. GRASPING GOALS IN PARTIALLY OCCLUDED SCENARIOS WITHOUT GRASP TRAINING

This chapter investigates robotic grasping methods, which is capable of grasping user-specified objects from a scene. To accomplish this, I introduce GoalGrasp, a straightforward yet powerful 6-DOF robot grasp pose detection technique that operates without the need for grasp pose annotations or training. This method facilitates user-specified object grasping even in partially occluded environments. By integrating 3D bounding boxes with basic human grasping principles, the proposed approach establishes a new framework for detecting robot grasp poses. Initially, I utilize the 3D object detector (RCV), which functions without 3D annotations, to swiftly detect objects in new scenes. Using the 3D bounding box and human grasp principles, the method performs dense grasp pose detection. The experimental assessment includes 18 common objects divided into 7 shape-based categories. Without any grasp training, the method produces dense grasp poses for 1000 scenes, creating an extensive grasp pose dataset. I evaluate our method's grasp poses against existing techniques using a new stability metric, revealing significantly enhanced grasp pose stability. In user- specified robot grasping trials, the method achieves a 94% success rate. Additionally, in user- specified grasping tests under partial occlusion, the success rate is 92%.

6.1. Introduction

In human-robot interaction scenarios, robots frequently depend on human instructions to execute specific tasks. For example, home service robots often receive directives from users to grasp particular objects (Xu et al. 2023). The capability to grasp user-specified items is crucial in human-robot interactions, especially for individuals with limited mobility, such as the elderly or patients, where robots can offer essential assistance (Tröbinger et al. 2021).

This study introduces a novel 'object-level' grasp pose generation method named GoalGrasp, designed to grasp user-specified objects even when partially occluded. Unlike traditional learning-based methods, this approach does not require grasp-specific training, thus eliminating the need for manual 6D grasp pose annotations. This significantly improves the method's efficiency in new scenarios. A key feature of this method is its 'object-level' granularity, ensuring grasp pose generation despite partial occlusion. To achieve this, I first utilize a 3D object detection method (RCV) from Study 1 to obtain 3D bounding boxes for objects in the scene. RCV's primary advantage over most existing 3D object detection methods is its independence from 3D annotations for training, facilitating 3D detection in new environments and with novel objects. For robotic tabletop grasping, I annotate approximately 200 2D bounding boxes per target object to ensure stable 3D object detection. Using RCV, I create a dataset of 1000 grasping scenes, with 3D bounding boxes generated for each object, all produced by RCV without manual annotation.

Reflecting on human grasping behavior, I observe that human grasping can also be considered 'object-level'. Humans often use simple heuristics for successful grasps. For instance, when grasping a box, we typically grasp two opposing faces, while for an apple, we grasp two points along its 'diameter'. We do not usually grasp a single vertex of a box or two adjacent faces. These simple heuristics can greatly simplify grasp pose generation, reducing the need for complex learning-based models and enabling rapid robotic object grasping. By leveraging the 3D bounding box, object category, and basic grasping heuristics, I propose a straightforward yet effective 'object-level' grasp pose generation method.

I conduct a series of experiments. First, I categorize the 18 objects in the collected dataset into seven major classes. For each class, I design specific grasp pose generation algorithms, allowing the creation of dense grasp poses for objects in 1000 scenarios. To assess the quality of these grasp poses, I introduce a novel stability metric and use it to compare our method's generated poses with those from state-of-the-art approaches. The results show that our method significantly outperforms existing techniques. To ensure fairness, the existing methods are not retrained. Since our method operates at the 'object-level', the generated grasp poses are linked to specific object categories, facilitating direct execution of user-specified object grasping. I deploy GoalGrasp on a real robot and conduct 500 user-specified grasping tasks, achieving a 94% success rate. Additionally, I perform 100 user-specified grasping tasks in occluded scenes, resulting in a 92% success rate.

6.2. Motivation and 'Object-level' Grasping

In Chapter 2, I discuss the main methods for robot grasp detection, which can be broadly categorized into three types: RGB-D-based, point cloud-based, and 6D pose estimation-based approaches. The RGB-D-based methods detect graspable rectangles from RGB-D images. The point cloud-based methods extract features from point clouds and regress feasible grasp poses for the point cloud scene. The 6D pose estimation-based methods first detect the 6D pose of the target object and then map the grasp pose to the detected 6D pose, resulting in feasible grasp poses (Kleeberger et al. 2020). However, these methods exhibit certain drawbacks in real-world applications of robot grasping, making it is challenging to swiftly apply them to grasp tasks involving new scenes or objects. For instance, applying point cloud-based grasp detection methods to tasks involving new scenes or objects requires the laborious task of re-labeling a significant amount of data for model retraining. My attempts to directly utilize such methods without retraining for new object grasping have yielded unsatisfactory results. Moreover, 6D pose estimation-based approaches necessitate the manual annotation of object 6D poses, which is highly tedious and time-consuming (Kleeberger et al. 2020). I argue that the need for manual annotation of grasp poses or 6D poses on every occasion is very inefficient in real-world robot grasping applications. To

circumvent these challenges, I propose a robot grasp detection method, eliminating the need for any grasp-specific training. This is the main motivation of this study.

Intuitively, humans have the ability to achieve stable grasping even with only partial visibility of an object. This ability is based on human perceptual capabilities, such as the recognition of object categories and an understanding of their spatial extent. The proposed method is directly inspired by this notion and leverages 3D object detection to enable robots to identify object categories and estimate their spatial occupancy, even in scenarios where objects are partially occluded. Subsequently, this information is combined with simple human grasping priors to generate 6D grasp poses. Upon reflecting on human grasping behavior, I observed that human grasping abilities can be characterized as operating at the 'object-level'. Humans often rely on simple heuristics to achieve successful grasps specific to each object. For instance, when grasping a box, a common heuristic is to grasp two opposing faces, while for an apple, the heuristic involves grasping two points along its 'diameter'. On the contrary, grasping a single vertex of a box or two adjacent faces are not commonly perceived as effective grasping strategies. I find that leveraging these intuitive grasping heuristics can significantly simplify grasp pose generation methods, reducing the reliance on complex learning-based models and facilitating rapid robotic object grasping. Figure 6.1 illustrates the procedure of the proposed method, which involves the detection of the target object's category and 3D bounding boxes. Then, I design the corresponding grasp pose generation strategy using the human grasping priors, followed by the mapping of the generated grasp poses to the scene. The proposed method offers several advantages over existing approaches. Firstly, it exhibits enhanced generalization capabilities, allowing it to adapt to new scenes or object grasping tasks without any grasp-specific retraining. This eliminates the need for manually annotated grasp pose labels, making our approach more efficient and versatile. Secondly, leveraging 3D bounding boxes, our method demonstrates robustness in scenarios where target objects may be partially occluded.

Benefiting from these characteristics, the method can be efficiently applied to new scenarios and achieve goal grasping, even in the presence of partial occlusion. Next, 3D object detection and grasp pose generation strategy are presented.

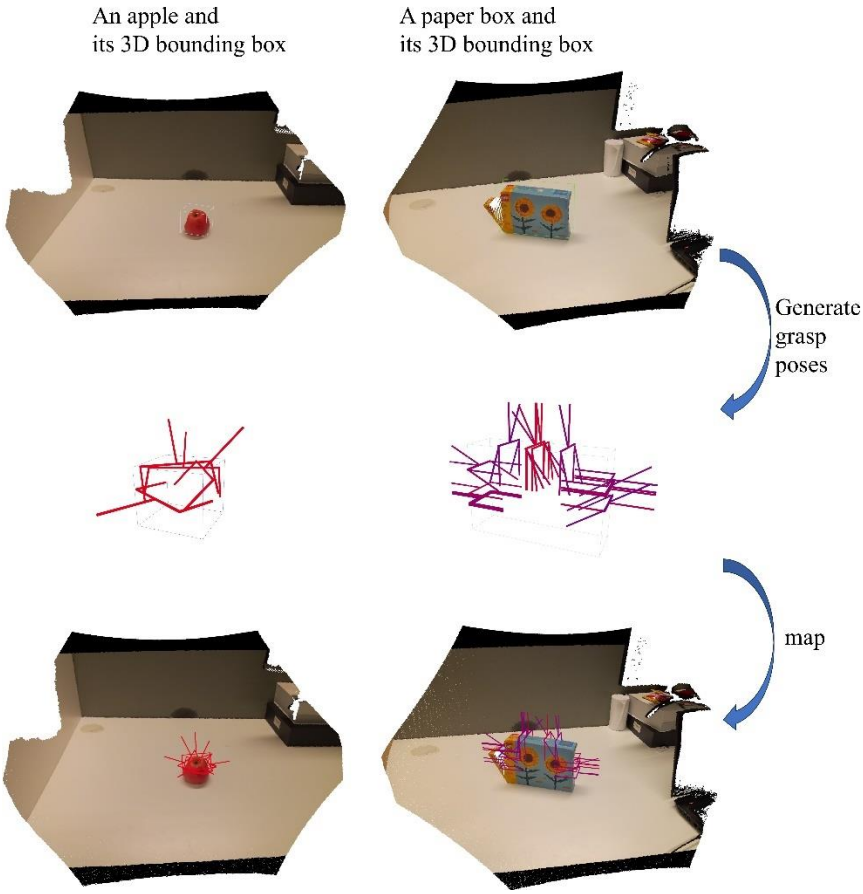


Figure 6.1. The procedure of generating grasp poses according to the object category and 3D bounding box.

6.3. 3D Object Detection on New Objects for Robots

The method I propose is based on utilizing 3D bounding boxes to detect the grasp poses of target objects, which requires to efficiently perform 3D object detection for various items in new scenes. However, the majority of existing 3D object detection methods heavily rely on extensive manual annotation of 3D labels. Clearly, manual annotation of 3D labels is both labor-intensive and time-consuming, making it impractical for achieving efficient robotic grasping. In contrast, in Study 2, I introduced a completely 3D label-free 3D object detection method that solely utilizes easily obtainable 2D bounding box labels to achieve 3D detection of novel objects. Next, I describe how to employ this method to accomplish 3D detection of tabletop objects.

To achieve 3D tabletop object detection in new scenes without relying on 3D annotations, I employ an Azure Kinect DK to gather RGB images and point cloud data. Initially, I manually

annotate 2D bounding boxes for RCV. Subsequently, I develop a 3D detector that operates without direct access to 3D annotations throughout the process. Figure 6.2 illustrates the process of 2D labeling and inferring 3D bounding boxes from the collected data. First, I manually label 2D bounding boxes on the raw RGB images captured by the Kinect (Figure 6.2, first row). These labeled bounding boxes are then utilized to train a 2D detector, enabling 2D object detection on the RGB images. Next, I filter out points located outside the 2D bounding boxes and project the remaining points orthogonally, resulting in two images (Figure 6.2, second row). Manual labeling is performed on these two images, generating two red 2D bounding boxes. Similarly, I filter out the points outside these newly labeled bounding boxes and project the remaining points orthogonally, producing two new images (Figure 6.2, fourth row). I annotate these two images with two red 2D bounding boxes. Importantly, all 3D bounding boxes are inferred based on the previously labeled 2D bounding boxes rather than relying on manual annotations. Subsequently, I utilize the 2D bounding boxes to train a 2D object detector (YOLOv5), which is then employed to develop an RCV model capable of detecting tabletop objects. For further details on the RCV model, please refer to Study 2.

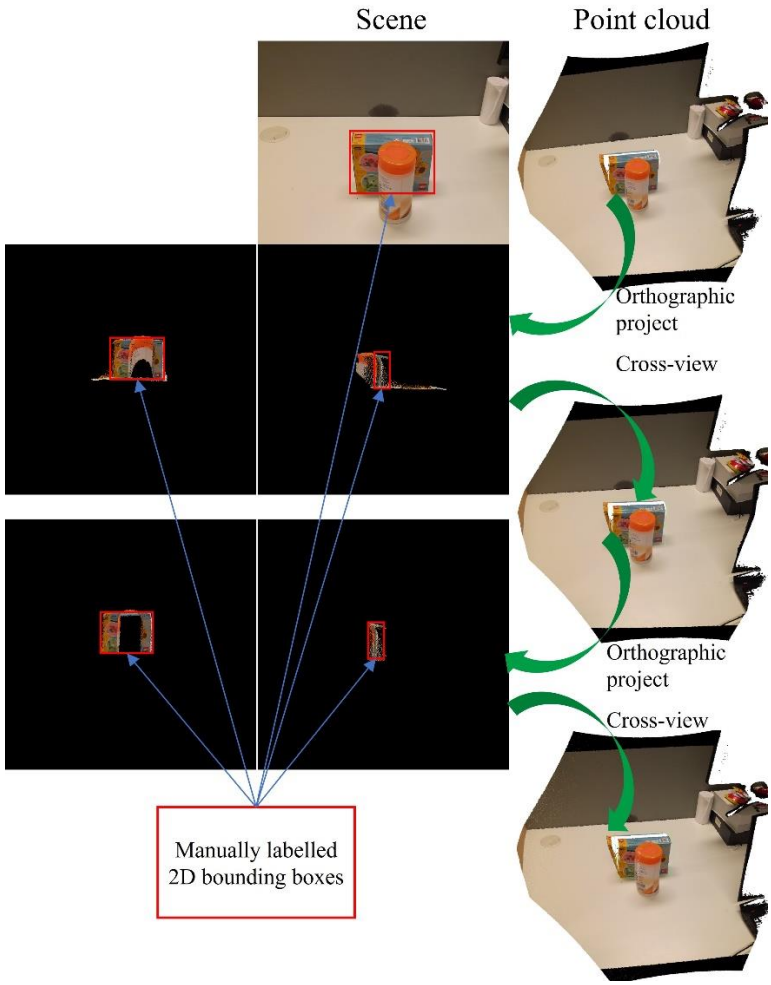


Figure 6.2. 2D labeling and inferring 3D bounding boxes for RCV on the collected data.

In this study, I specifically select 18 different objects and annotated approximately 200 2D bounding boxes for each object. The annotation process for each object took approximately half an hour with one annotator. Leveraging these 2D annotations, I realize 3D detection. The RCV method, relying solely on established 2D detection techniques, demonstrates excellent robustness. I utilize the RCV to construct a dataset comprising 500 scenes, encompassing 15 categories and 18 different objects, with over 5,000 3D bounding boxes. Some samples are demonstrated in Figure 6.3. Subsequently, I employ a grasp generation strategy to generate dense grasp poses for each object in the scenes.

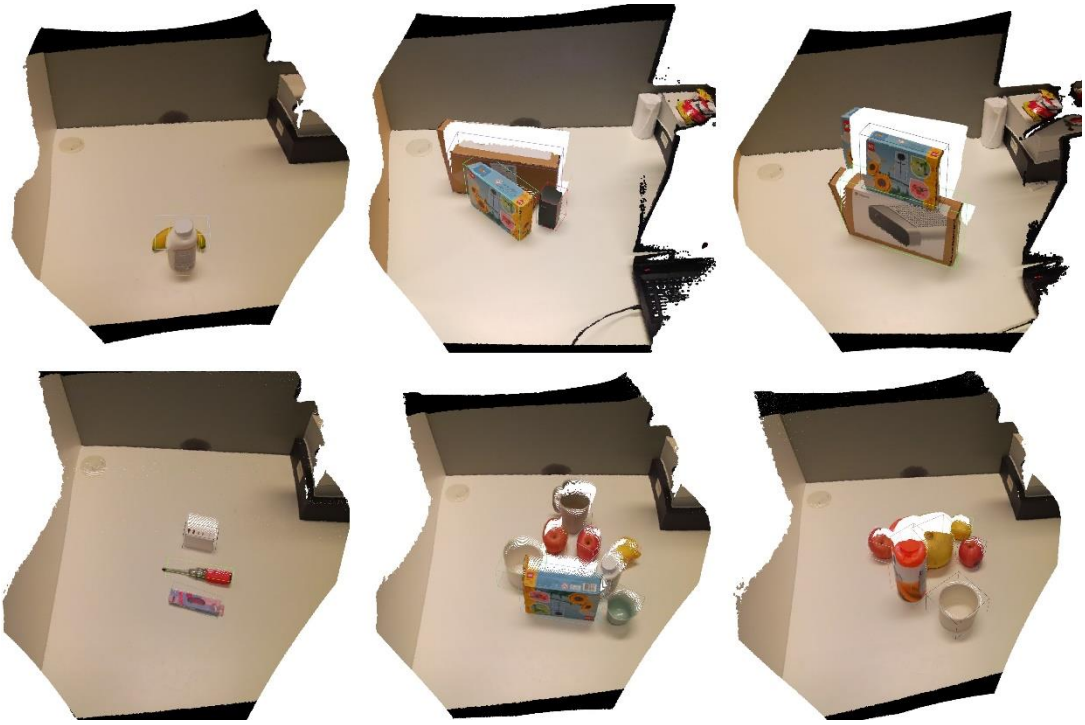


Figure 6.3. 3D-TABLETOP-OBJECT dataset with 15 categories: large box, small box, large cylinder, small cylinder, bowl, cucumber, banana, tape, screw, apple, lemon, grapefruit, pen, jar, mug.

6.4. Strategy for Robots to Grasp Objects

In this section, I introduce grasp pose generation strategies for objects in the 3D-TABLETOP-OBJECT dataset. Although these objects serve as examples, the strategies can be applied to many items not included in the dataset. I categorize these items based on their geometric characteristics into seven groups: box-shaped, spherical, cylindrical, curved, container, tool, and ring objects. First, I summarize human grasping priors specific to each category. Using this prior knowledge and the 3D bounding boxes in the object coordinate system, I then design a strategy to generate a dense set of grasp poses.

6.4.1. Box-shaped Objects

In everyday life, there are various box-shaped objects, such as packaging boxes and power banks. Regarding these objects with box-like shapes, human grasping prior knowledge can be summarized in the absence of considering the maximum gripper width. Specifically, the following priors apply: (1) The grasp position is located along the symmetric axis of each face, (2) the gripper's depth direction aligns closely with the normal orientation of each face, (3)

the gripper's width direction aligns with the corresponding edge direction, and (4) the grasp width and depth correspond to the respective edge lengths. For (1), the sampling trajectories (ST) of grasp positions are shown as the red lines in Figure 6.4(a). Mathematically, I express them as

$$ST_0 \leftarrow \left\{ \begin{Bmatrix} x = x_{mid} \\ y = 0 \\ z = t, t \in z_{range} \end{Bmatrix} \begin{Bmatrix} x = x_{mid} \\ y = y_{max} \\ z = t, t \in z_{range} \end{Bmatrix} \begin{Bmatrix} x = 0 \\ y = y_{min} \\ z = t, t \in z_{range} \end{Bmatrix} \begin{Bmatrix} x = x_{max} \\ y = y_{min} \\ z = t, t \in z_{range} \end{Bmatrix} \right.$$
$$\left. \begin{Bmatrix} x = x_{mid} \\ y = t, t \in y_{range} \\ z = 0 \end{Bmatrix} \begin{Bmatrix} x = x_{mid} \\ y = t, t \in y_{range} \\ z = z_{max} \end{Bmatrix} \begin{Bmatrix} x = 0 \\ y = t, t \in y_{range} \\ z = z_{mid} \end{Bmatrix} \begin{Bmatrix} x = x_{max} \\ y = t, t \in y_{range} \\ z = z_{mid} \end{Bmatrix} \right.$$
$$\left. \begin{Bmatrix} x = t, t \in x_{range} \\ y = y_{mid} \\ z = 0 \end{Bmatrix} \begin{Bmatrix} x = t, t \in x_{range} \\ y = y_{mid} \\ z = z_{max} \end{Bmatrix} \begin{Bmatrix} x = t, t \in x_{range} \\ y = 0 \\ z = z_{mid} \end{Bmatrix} \begin{Bmatrix} x = t, t \in x_{range} \\ y = y_{max} \\ z = z_{mid} \end{Bmatrix} \right\}$$

Algorithm 1 presents the grasp pose generation algorithm for box-shaped objects. For each element in ST_0, I sample multiple grasp points and derive the corresponding grasp pose (line 8-19 and 30), width (line 21-26), and depth (line 28). Figure 6.4(b) illustrates the generated grasping poses. It is important to note that these grasping poses do not account for factors such as the maximum gripper width and environmental constraints, which will be discussed later.

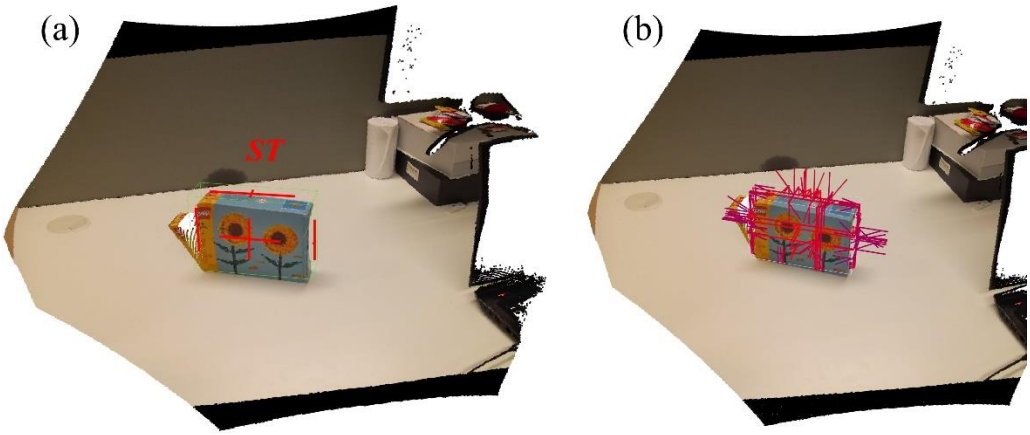


Figure 6.4. Grasp poses generation for box-shaped objects. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

6.4.2. Spherical Objects

Approximately spherical irregular objects, such as apples, are also commonly encountered in daily life. Human grasping prior knowledge for such objects can be summarized as follows:

(1) The grasp point lies on the surface of the enclosing sphere of the object, (2) the gripper's depth direction points from the grasp point towards the center of the sphere, (3) the gripper's width direction is determined by the cross product of the depth direction and the direction of gravity, and (4) the grasp width and depth correspond to the diameter and radius of the sphere, respectively. The sphere can be determined by the 3D bounding box, as shown in Figure 6.5(a), which is the sampling trajectory (ST) of grasp points. I mathematically express as

$$ST_1 \leftarrow \{(x - CT.x)^2 + (y - CT.y)^2 + (z - CT.z)^2 = R^2\}$$

where $[CT.x, CT.y, CT.z]$ are the center of the 3D box, R is the radius of the sphere, see line 4 in Algorithm 2.

Algorithm 2 describes the algorithm for generating grasp poses for spherical objects. For each sampling point, I derive the grasp pose (line 9-12), width (line 13), and depth (line 14). Figure 6.5(b) illustrates the generated grasp poses, with the number of samples set to 20 for better visualization.

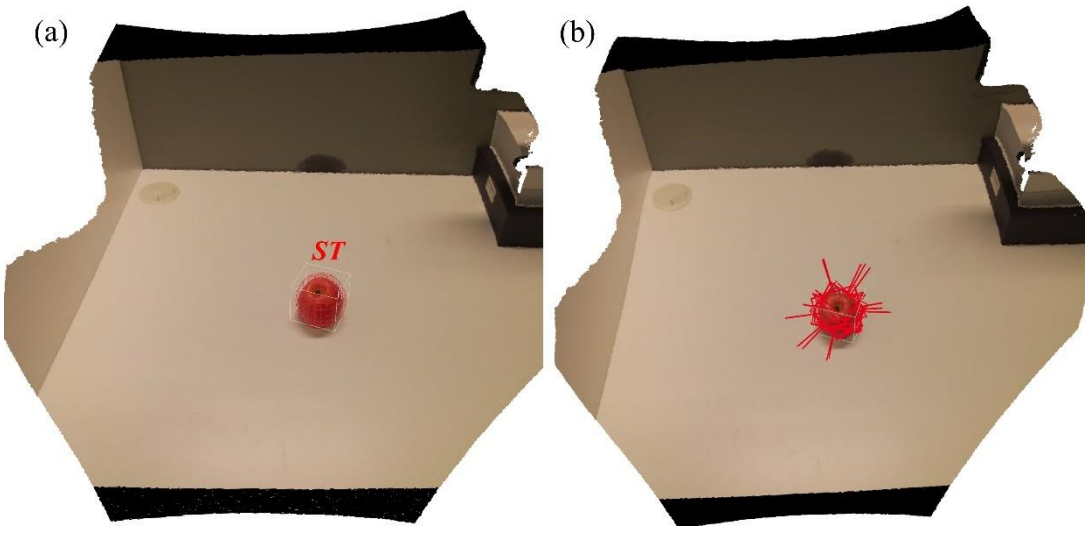


Figure 6.5. Grasp poses generation for spherical objects. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

Algorithm 1 Generate grasp poses for box-shaped objects

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
for st in ST_0 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    if $st.x == t$ **then**
      $X = [p.x,\ CT.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
      Transpose and normalize X
      $Y = X \times [1,\ 0,\ 0]$.T
    else if $st.y == t$ **then**
      $X = [CT.x,\ p.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
      Transpose and normalize X
      $Y = X \times [0,\ 1,\ 0]$.T
    else
      $X = [CT.x,\ CT.y,\ p.z] - [p.x,\ p.y,\ p.z]$
      Transpose and normalize X
      $Y = X \times [0,\ 0,\ 1]$.T
    end if
    if $Y == [\pm 1, 0, 0]$ **then**
      Grasp width $w = x\_length$
    else if $Y == [0, \pm 1, 0]$ **then**
      Grasp width $w = y\_length$
    else
      Grasp width $w = z\_length$
    end if
    Grasp depth $d = min(x\_length/2,\ y\_length/2,$
     $z\_length/2,\ gd)$
    $Z = X \times Y$
    $R = [X, Y, Z]$
    $T = p$

    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$

    Rotate R along Y with a random angle in $[-1/4\pi, 1/4\pi]$

    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
  end for
end for

Transfer g to the robot coordinate system.

return $g,\ W,\ D$

Algorithm 2 Generate grasp poses for spherical objects

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
$R = min(x\_length, y\_length, z\_length)/2$
for st in ST_1 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    $X = [CT.x,\ CT.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
    Transpose and normalize X
    $Y = X \times [0,\ 1,\ 0]$.T
    $Z = X \times Y$
    Grasp width $w = 2R$
    Grasp depth $d = R$
    $R = [X, Y, Z]$
    $T = p$

    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
  end for
end for
Transfer g to the robot coordinate system.
return $g,\ W,\ D$

6.4.3. Cylindrical Objects

Cylindrical objects, such as water bottles, are also common items. The grasp prior knowledge for such objects can be summarized as follows: (1) The grasp point lies on the surface of the cylinder, either at the center of the top or bottom, (2) the gripper's depth direction aligns closely with the normal orientation of each face, (3) the gripper's width direction is perpendicular to the height direction of the cylinder, and (4) the grasp width and depth correspond to the diameter and radius of the cylinder, respectively. Figure 6.6(a) demonstrates the sampling trajectories (ST) for grasp points, which can be mathematically expressed as

$$ST_2 \leftarrow \begin{cases} \begin{cases} (x - CT.x)^2 + (z - CT.z)^2 = R^2 \\ y = t, t \in y_{range} \end{cases}, & \begin{cases} x = x_{mid} \\ y = y_{min}, \\ z = z_{mid} \end{cases} & \begin{cases} x = x_{mid} \\ y = y_{max} \\ z = z_{mid} \end{cases} \end{cases}$$

Note that I perform some preprocessing steps to ensure that the y-axis of the object coordinate system aligns with the height direction of the cylindrical object. Algorithm 3 describes the algorithm for generating grasp poses for cylindrical objects. Regardless of whether the object is standing upright or lying down, the grasp pose (line 9-19), width (line 20), and depth (line 21) can be generated in the object coordinate system. Figure 6.6(b) presents the generated grasp poses. The grasp pose generation algorithm for other shaped objects, including curved objects, containers, tools, and circular objects can be accessed in Appendix A. By incorporating prior knowledge about object grasping, I can obtain grasp poses with high consistency in their distribution. The comparative experiments with other method are presented in Experiments.

Algorithm 3 Generate grasp poses for cylindrical objects

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
$R = min(x\_length, z\_length)/2$
for st in ST_2 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    if st is a point **then**
      $X = [CT.x,\ CT.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
      Transpose and normalize X
      Rotate [1, 0, 0] with a random angle as Y
      $Z = X \times Y$
    else
      $X = [CT.x,\ p.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
      Transpose and normalize X
      $Y = X \times [0,\ 1,\ 0]$.T
      $Z = X \times Y$
    end if
    Grasp width $w = 2R$
    Grasp depth $d = R$
    $R = [X, Y, Z]$
    $T = p$
    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
    if st is not a point **then**
      Rotate R along Y with a random angle in $[-1/4\pi, 1/4\pi]$
      $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
    end if
  end for
end for
Transfer g to the robot coordinate system.
return $g,\ W,\ D$

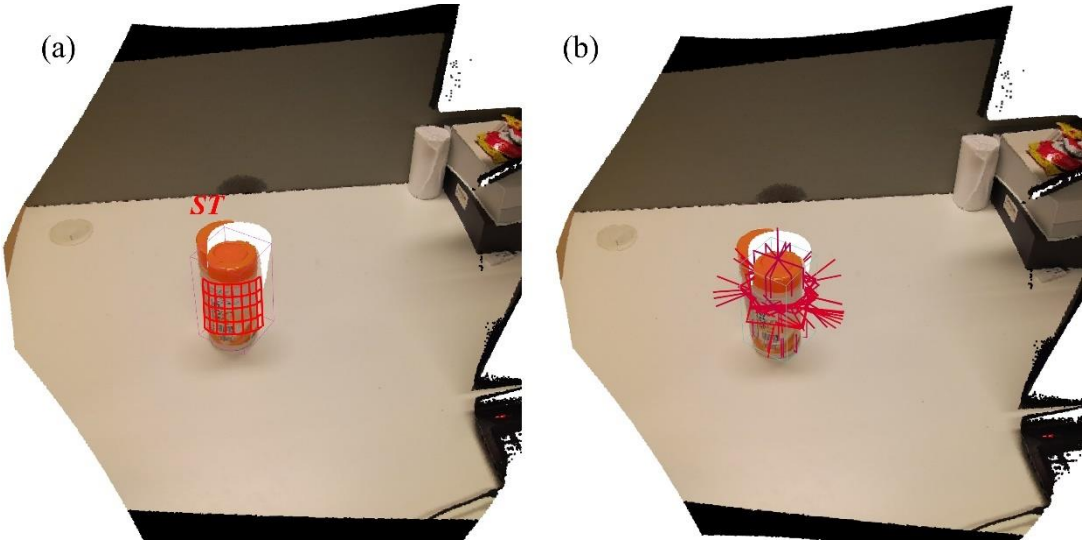


Figure 6.6. Grasp poses generation for cylindrical objects. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

6.4.4. Object Grasp Pose Filtering Metric

The grasp pose generation algorithm does not account for environmental constraints or interferences between objects, leading to the presence of infeasible grasp poses, such as those too close to other objects or colliding with the tabletop. In this section, I propose filtering metrics to eliminate these infeasible grasp poses. The first criterion is the orientation of the grasp pose, where valid poses are either horizontal or inclined downward. This aids in motion planning for the robot and prevents collisions with the tabletop. I evaluate this criterion using the cosine distance between the depth direction of the grasp pose and the direction of gravity, referred to as

$$dcos = 1 - \frac{X \cdot g}{|X||g|} < TH_0 \tag{6.1}$$

where X is the depth direction of the grasp pose, g is the direction of gravity, which is [0, 1, 0].T for our depth camera. TH_0 is the threshold.

The second criterion is that the grasp point must be positioned as a certain distance above the tabletop to avoid collisions between the robot and the tabletop. It can be represented as Equation (6.2).

$$\max(BBox[:,1]) - G.T[1,0] > TH_1 \quad (6.2)$$

where $BBox \in R^{8\times3}$ is the 3D bounding box of the object, $G.T \in R^{3\times1}$ is the location of the grasp pose. Note that the direction of gravity is defined as [0, 1, 0].T. TH_1 is the threshold.

The third criterion is the maximum width of the gripper. It is evident that if the generated grasp pose's width (w) exceeds the maximum gripper width (w_{max}), the grasp pose is considered invalid. This criterion can be represented as Equation (6.3)

$$\omega_{max} - \omega > 0 \quad (6.3)$$

The fourth criterion is to ensure that the grasp point is positioned at a distance greater than a certain threshold from other objects to avoid collisions between the robot and other objects. To determine the distance between the grasp pose and other objects, our method utilizes the 3D bounding boxes of other objects. I discretize a certain number of points on the surface of the bounding boxes and derive the minimum distance between the grasp pose and these points. If the minimum distance is greater than the predefined threshold, the grasp pose is considered valid. The criterion is shown as Equation (6.4).

$$min\{dist(G.T,\ BP_0),\ dist(G.T,\ BP_1),\ \ldots,\ dist(G.T,\ BP_{n-1})\ \ dist(G.T,\ BP_n)\} > TH_2 \quad (6.4)$$

where *dist* represents the Euclidean distance, $G.T \in R^{3\times1}$ indicates the location of grasp pose, $BP \in R^{1\times3}$ denotes a point on the 3D bounding boxes, and n signifies the number of sampled points. TH_2 is the threshold.

I present four fundamental criteria here for filtering out invalid grasp poses. However, additional criteria can be extracted to ensure that the generated grasp poses meet the specific requirements of the given scenario.

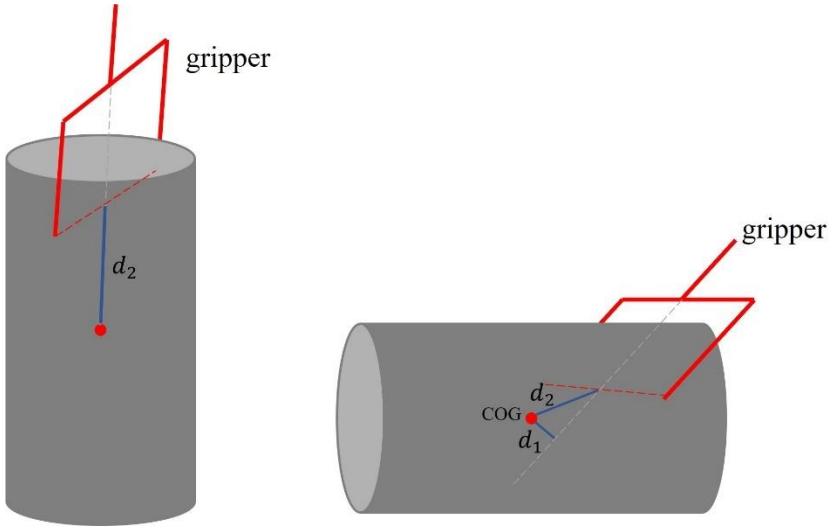


Figure 6.7. The illustration of the novel stability metric.

6.4.5. Object Grasp Pose Evaluation Metric

Using the proposed grasp pose generation and filtering methods, I can obtain a dense set of feasible grasp poses. However, their performance lacks a quantifiable measure. In this section, I introduce a quantifiable stability score to evaluate the poses. In (Fang et al. 2023), the concept of the center of gravity (COG) for objects was used to assess the stability of grasp poses. Specifically, the normalized perpendicular distance (denoted as d1) between the gripper plane and the object's COG is defined as the stability score. However, this metric can be insufficient in certain cases, as shown in the left subplot of Figure 5.7, where d1 is zero, but the grasp appears unstable. I believe this instability is due to the distance between the touch point and the COG, denoted as d2. Therefore, I propose incorporating both d1 and d2 to comprehensively evaluate the stability of grasp poses. The right subplot of Figure 6.7 illustrates a grasp pose's d1 and d2. Equation (6.5) demonstrates the proposed evaluation metric.

$$M = \alpha \left[1 - \frac{d_1}{l_{diag}/2}\right] + (1-\alpha)\left[1 - \frac{d_2}{l_{diag}/2}\right] \tag{6.5}$$

where α is a weight coefficient, l_{diag} denotes the length of the diagonal of the 3D bounding box of the object. It is utilized to normalize d_1 and d_2 to the range [0, 1]. d_1 and d_2 are also

derived using the 3D bounding box. Specifically, I consider the center of the 3D bounding box as the COG and then compute d_1 and d_2.

6.4.6. Generate Grasp Poses for A Scene

In this section, I introduce a comprehensive algorithm for generating grasp poses in a given scene by integrating the previously discussed grasp pose generation algorithms, filtering metrics, and evaluation metrics. The process begins with the use of a pre-trained 3D detection model, RCV, to detect objects in the scene, providing 3D bounding boxes, labels, confidences, and point clouds for each detected object. Next, I apply the corresponding grasp pose generation algorithm based on the object's label to create dense grasp poses. These poses are then filtered using Equations (6.1)-(6.4) to remove invalid ones. Subsequently, the proposed evaluation metrics are used to assess the grasp poses. To account for the quality of the 3D bounding box, I multiply the evaluation scores by the confidence associated with each 3D bounding box. This adjusted score serves as the final evaluation score for each grasp pose. Algorithm 4 outlines the process of generating grasp poses for a given scene.

Algorithm 4 Generate grasp poses for a scene

Require: Trained RCV; img, $PC \in \mathbb{R}^{n\times 3}$ and $rgb \in \mathbb{R}^{n\times 3}$ of the scene, empty buffer **G**.

```
boxes, labels, confs, obj_pcs = RCV(img, PC, rgb)
// obj_pcs is a set of segmented points for each objects.
for i ← 0 to len(boxes)-1 do
    Use labels[i] to retrieve the corresponding grasping pose generation algorithm, denoted as GA.
    g, W, D = GA(boxes[i], obj_pcs[i])
    Filter g, W, D using Eq.(2)-(5).
    Utilize Eq.(6) to measure the grasp poses, and obtain the scores, denoted as s.
    s = confs[i]×s
    G← G ⋃ [g, W, D, s]
end for
return G
```

6.5. Real-world Robot Grasp Pose Detection Experiments

In this section, I present four experiments conducted to evaluate the effectiveness of the proposed method. The experiments include (1) grasp pose detection without any grasp-

specific training, (2) a comparative experiment of grasp pose quality against existing methods, (3) goal-oriented grasping implementation on a physical robot, and (4) goal-oriented grasping in partially occluded scenarios.

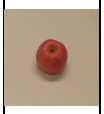
6.5.1. Generating Dense Grasp Poses for 1000 Scenes

The proposed method enables object grasping without requiring any grasp training, which is highly valuable in the diverse and dynamic scenarios of robotic grasping. This approach not only facilitates the rapid deployment of robotic grasping but also eliminates the need for labor-intensive annotation of 6D grasp poses. To validate the effectiveness of the proposed method, I generate dense grasp poses for all objects in the 3D-TABLETOP-OBJECT dataset established in Section 6.3, using Algorithm 4. This dataset includes 18 types of objects, which I categorize into 7 shape-based groups, each corresponding to a specific grasp pose generation algorithm, as shown in Table 6.1. By setting certain sampling parameters, I generate 300 to 500 grasp poses per object, resulting in approximately 2 million grasp poses for the entire dataset.

However, the proposed method cannot guarantee that all generated grasp poses are feasible. Certain conditions can render grasp poses invalid, primarily due to (1) poor quality of the generated 3D bounding boxes and (2) generated grasp poses being too close to nearby objects. I believe performance can be further improved by annotating more 2D labels for the 3D detector or implementing additional filtering metrics. Figure 6.8 showcases some scenes with the grasp poses, including multiple and occluded objects. Even when objects are partially occluded, with an occlusion area of approximately 50%, the method can still generate effective grasp poses. Next, I will compare the generated grasp poses with those produced by existing methods using the evaluation metric (Equation (6.5)).

Table 6.1: 3d-tabletop-grasp dataset. '/' indicates that the object is larger than the size of the gripper, leading to the experiment cannot performed.

Shapes	Objects	Success Rate

Box-shaped		1.0
		/
		1.0
		0.939
		0.939
Spherical Objects		1.0
		/
Cylindrical Objects		/
		0.971
		0.912
		0.906
		0.938
Curved Objects		0.970
Containers		0.909
		0.939

		0.909
Tools		0.882
Ring Objects		0.912

6.5.2. Comparing Grasping Poses

In this experiment, I compare the grasp poses generated by the proposed method with the grasp poses detected by AnyGrasp (Fang et al. 2023) on the 3D-TABLETOP-GRASP dataset. I chose to compare on this dataset because my focus is on achieving robotic grasping without any grasp-specific training, which is a departure from the majority of existing grasp research that heavily relies on training. As a result, I directly apply existing methods to new scenes without retraining, simulating a scenario where grasp training is not required. This setting ensures a certain level of fairness in comparing the method to training-based approaches. However, comparing the method with existing approaches in new scenes presents a challenge in selecting appropriate evaluation metrics. When comparing training-based methods, different models are trained on the same dataset, allowing for the output of confidence for each grasp pose on the test set, thus enabling direct comparison with annotations. However, the proposed method cannot follow this framework. Therefore, I introduce the stability metric (M), as defined in Equation (6.5), as the comparative metric. Unlike the confidence generated by neural networks, this metric evaluates the stability of grasp poses based on structural properties such as object size and shape. As a result, it can be applied to measure grasp poses generated by any method.

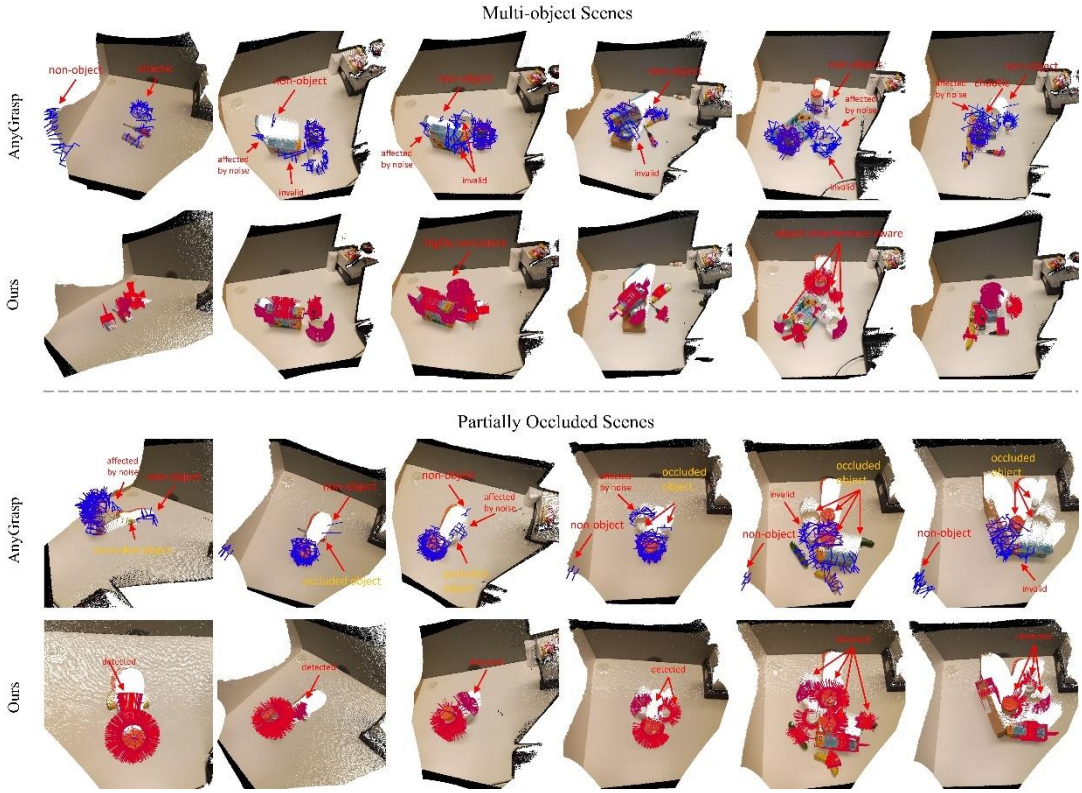


Figure 6.8. The comparison of our method and AnyGrasp on multiple objects and partially occluded objects. We note the problem that AnyGrasp is prone to when generating grasping poses for items in the scene. Our method avoids these problems to a certain extent.

The proposed method directly generates grasp poses for each object in the scene and evaluates the stability of each grasp pose using the 3D bounding box of the object. However, AnyGrasp cannot directly compute this value because the grasp poses generated by AnyGrasp are not classified, meaning it is not known which object each grasp pose corresponds to. Therefore, I manually perform statistics on the generated grasp poses and combined them with the 3D bounding boxes generated by our method to calculate the stability of the grasp poses. Specifically, I generate 100 grasp poses using AnyGrasp in each scene and then select up to three objects as statistical objects. For the selected objects, I rank the top five grasp poses and calculate the stability metric **M**. Then, I multiply **M** by a binary coefficient, denoted as β, where β is set to 1 if the grasp pose is determined by the human to successfully grasp the object, and 0 otherwise. Finally, I calculate the mean value as the stability metric. Similarly, in our method, I select the same object in the same scene and utilize the same methodology to calculate the stability metric for the grasp poses.

Table 6.2 presents the grasp stability values of the proposed method and AnyGrasp for 110 objects in 60 randomly selected scenes. The data demonstrate the grasp poses generated by the proposed method exhibit significantly higher stability compared to those generated by AnyGrasp. I observe that AnyGrasp performs better in single-object scenes compared to multi-object scenes, possibly because it is more susceptible to interference from surrounding objects in multi-object scenarios. In occluded scenes, AnyGrasp struggles to generate effective grasp poses for occluded objects, indicating its inability to handle occlusion. On the other hand, the proposed method consistently demonstrates excellent performance across single-object, multi-object, and occluded scenarios. Given that the method can handle various types of scenes without the need to any grasp-specific training, this has significant implications for object grasping in human-robot interaction.

To provide a visual comparison, I visualize 12 scenes in Figure 6.8. To facilitate viewing, I reduce the number of grasp poses generated by the proposed method. The grasp poses generated by the method demonstrate higher consistency and performance.

Table 6.2: Stability metric in different scenarios. We compare our scores with Anygrasp. For each scenario, we select up to 3 items for testing. Bold is used to highlight the best results.

Single-object											
AnyGrasp	0.689	0.655	0.720	0.887	0.311	0.245	0.262	0.320	0.274	0	0.715
GoalGrasp	**0.886**	**0.822**	**0.830**	**0.893**	**0.704**	**0.841**	**0.981**	**0.881**	**0.672**	**0.995**	**0.874**
Single-object										Mean	Variance
AnyGrasp	0.187	0.413	0.098	0.495	0	0.607	**0.699**	0.524	0.135	0.412	0.071
GoalGrasp	**0.998**	**0.911**	**0.686**	**0.897**	**0.975**	**0.894**	0.651	**0.890**	**0.852**	**0.857**	**0.011**
Multi-object											
AnyGrasp	0.645 0.537 0.461	0.820 0.721 0.825	0.845 0.867 0.772	0.291 0 0.835	0.519 0.205 0.465	0 0.403 0.402	0.102 0.024 0.070	0.052 0.422 0	0.281 0.158 0.658	0.317 0 0.470	0.413 0.444 0.606
GoalGrasp	**0.930** **0.859** **0.863**	**0.894** **0.888** **0.900**	**0.888** **0.898** **0.892**	**0.826** **0.709** **0.957**	**0.819** **0.909** **0.683**	**0.841** **0.972** **0.677**	**0.997** **0.998** **0.890**	**0.841** **0.981** **0.717**	**0.986** **0.861** **0.897**	**0.996** **0.996** **0.882**	**0.852** **0.842** **0.981**
Multi-object										Mean	Variance
AnyGrasp	0.102 0.119 0	0.468 0.669 0.305	0 0 0.274	0.091 0.866 0.177	0 0.725 0.262	0.438 0 **0.914**	0.211 0.450 0.141	0.201 **0.700** 0.458	0.512 0 0	0.362	0.084
GoalGrasp	**0.845** **0.838** **0.998**	**0.878** **0.890** **0.885**	**0.911** **0.700** **0.675**	**0.850** **0.897** **0.893**	**0.808** **0.895** **0.689**	**0.832** **0.992** 0.895	**0.807** **0.845** **0.689**	**0.872** 0.640 **0.833**	**0.903** **0.911** **0.892**	**0.866**	**0.008**
Occluded-object											
AnyGrasp	0 / /	0 / /	0 / /	0.129 / /	0 / /	0.166 / /	0 / /	0 / /	0.491 0.241 /	0.304 0 /	0 / /
GoalGrasp	**0.969** / /	**0.886** / /	**0.715** / /	**0.700** / /	**0.897** / /	**0.901** / /	**0.827** / /	**0.939** / /	**0.678** **0.847** /	**0.806** **0.921** /	**0.882** / /
Occluded-object										Mean	Variance
AnyGrasp	0 / /	0 / /	0 0 /	0 0 /	0.149 / /	0.125 0 /	0.575 0 /	0.412 0 0	0.262 0 0	0.086	0.027
GoalGrasp	**0.886** / /	**0.892** / /	**0.893** **0.890** /	**0.708** **0.888** /	**0.692** / /	**0.917** **0.881** /	**0.893** **0.889** /	**0.892** **0.700** **0.914**	**0.892** **0.705** **0.879**	**0.846**	**0.008**

6.5.3. Goal-oriented Grasping for A Scene

In this section, I deploy GoalGrasp onto a real robot to accomplish the task of grasping user-specified targets. Specifically, the robot utilizes a depth sensor to capture scene data, and GoalGrasp is employed to generate grasp poses for each target. Upon receiving a grasping target instruction from the user, the robot executes the corresponding grasp for the specified target. In the experiments, I utilize an Ufactory xArm7 and a xArm Gripper as the robot platform, along with a Microsoft Azure Kinect DK as the depth sensor. The experimental setup is illustrated in Figure 6.9. It is worth noting that the depth sensor is deployed at an inclined position above the scene to simulate the perspective of a real home-service robot. A computer equipped with an NVIDIA 3090 GPU is used to run the system.

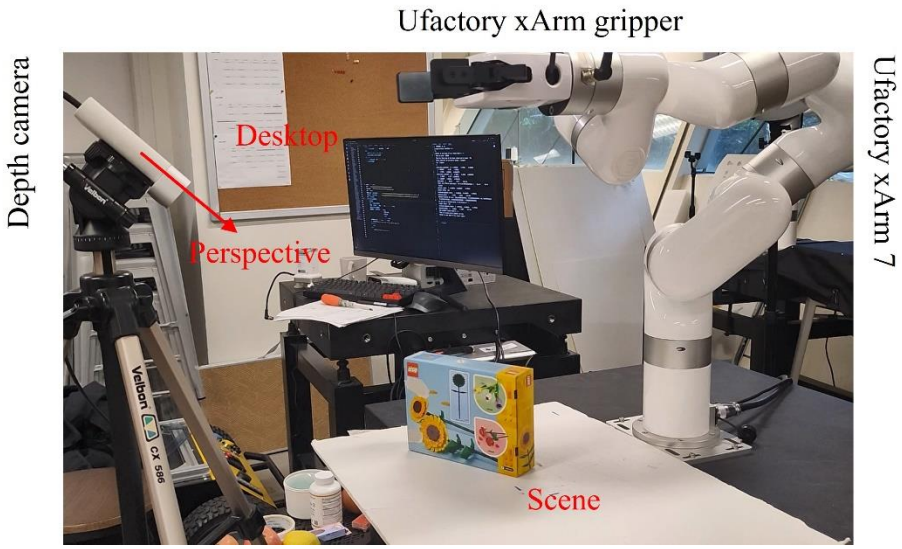


Figure 6.9. Experimental settings.

I do not conduct experiments comparing AnyGrasp with our method in this experimental setup because AnyGrasp cannot achieve target-orientated grasping, meaning it cannot complete the task of grasping user-specified objects. This target-orientated grasping is the focus of my work. Indeed, there are other studies on target-orientated robot grasping, but I also do not compare them with the proposed method for the following reasons: (1) these studies mainly focus on top-down 3D grasping rather than 6D grasping, and (2) they do not demonstrate strong generalization capabilities to new scenes, making it difficult to guarantee their performance in novel scenarios.

GoalGrasp can operate in two modes: (1) First, it generates grasp poses for all objects in the scene and then waits for instructions. Once the instruction is received, the robot performs the grasping. After completing the grasp, it goes back to the waiting state for further instructions. This mode avoids the need to detect grasp poses for the target upon receiving an instruction, reducing the user's waiting time. However, the initial process of generating grasp poses for all objects in the scene can be time-consuming (about 1s). (2) The robot waits for the user's instruction. Upon receiving the instruction, it only detects the grasp poses for the target in the scene and performs the grasp. After completing the grasp, it goes back to the waiting for further instructions. This mode distributes the detection time across each grasping operation and can handle cases where object positions change. Table 6.3 presents the detection times for both modes. I select a scene with four objects for testing purposes.

AnyGrasp performs four grasp detections, and after each detection, a human manually removes the corresponding target object. For GoalGrasp, I specify the grasping targets in the same order and conduct grasping trials, recording the time taken for detection in both two modes. The proposed method demonstrates a time for grasp detection approximately 50% of that of AnyGrasp, indicating excellent real-time performance.

In the experiments, I conduct 500 grasping trials in both single-object and multi-object scenes (with a maximum of 8 objects), and the robot achieved a success rate of 94%. Here, I define the success rate as the ratio of successful grasps to the total number of grasping attempts. The success rates for each object are listed in Table 6.1. The method demonstrates a high success rate, which validates the effectiveness of GoalGraps. Figure 6.10 illustrates the process of GoalGrasp generating a grasp pose for each object in the scene and the subsequent execution of the robot's grasping action based on user instructions.

Figure 6.10. User-specified grasping experiments, in which the object is determined by the user.

Table 6.3: Stability metric in different scenarios. We compare our scores with Anygrasp. For each scenario, we select up to 3 items for testing. Bold is used to highlight the best results.

Methods	Modes	Single detection (s)				Mean
		1	2	3	4	
AnyGrasp	Mode 2	0.509	0.468	0.466	0.475	0.480
GoalGrasp	Mode 1	1.030	**0**	**0**	**0**	0.258
	Mode 2	**0.204**	0.198	0.201	0.209	**0.203**

6.5.4. Grasping Partially Occluded Objects

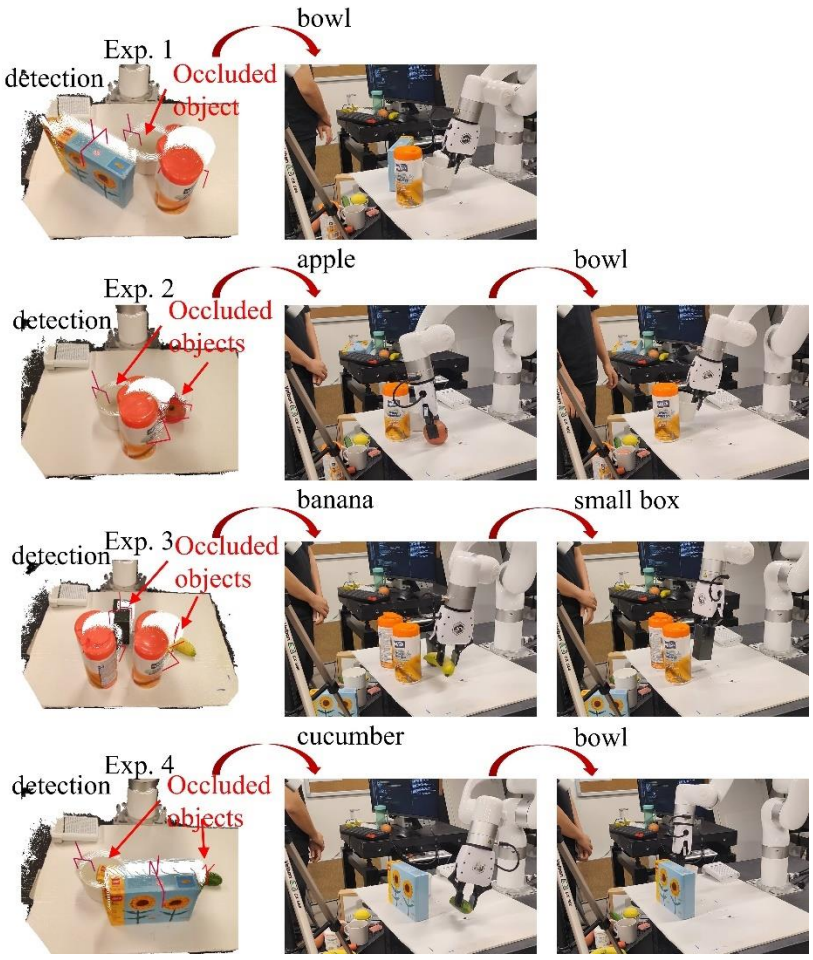


Figure 6.11. User-specified grasping experiments in partially occluded scenarios.

In this experiment, I validate that GoalGrasp can successfully grasp the target even when it is partially occluded (approximately 50%). To the best of our knowledge, existing research has not achieved reliable grasping of target objects in partially occluded scenes. However, for home-service robots, this ability is essential. I employ the same setup as the previous experiment to evaluate this capability. Experimentally, I perform 100 grasping trials on

occluded objects and achieve a grasp success rate of 92% with the robot. The method demonstrates high success rates in both single-object, multi-object, and partially occluded scenes, which indicates the robustness across different scenes. Figure 6.11 showcases several scenes of the robot executing grasping tasks, providing perspectives from both the robot's perspective and a third-party viewpoint.

In scenes with multiple objects and partial occlusion, there is a possibility of collisions during robot grasping attempts. To mitigate this, I employ Equation (6.4) to filter out grasp poses that are too close to other objects. As shown in the fourth row and fifth column of Figure 6.8, the white container's grasp poses, close to the box, are eliminated. Despite these measures, collisions may still occur during actual grasping, which remain a primary cause of failures. In future work, I plan to integrate obstacle avoidance algorithms to further enhance the grasping performance. Another potential reason for failures could be errors in the depth camera and the calibration between the depth camera and the robot's coordinate system. These errors increase the likelihood of failure, particularly when grasping small objects, despite seemingly feasible grasp poses in the point cloud.

6.6. Discussions

6.6.1. Object Coordinate System

In the Section 6.4, I design specific grasping strategies for each type of object. First, leveraging the object's 3D bounding box and point cloud data, I infer a sampling trajectory of grasp points denoted as ST. It is important to note that ST is defined in the object's coordinate system rather than the world coordinate system. Subsequently, I employ the corresponding grasp pose generation algorithm to obtain the grasp poses. However, it is essential to address potential variations in object orientations, such as cylindrical objects being either upright or lying flat. To establish a unified grasp pose generation algorithm capable of accommodating diverse object orientations, I introduce a preprocessing step to partially align the object's coordinate system. Taking the example of cylindrical objects, I align the y-axis of the object's coordinate system with the height direction of the cylinder. This alignment enables the direct application of the grasp pose generation algorithm. The

grasp point sampling trajectories I establish for each object category already consider the object's coordinate system. By aligning ST with the object's coordinate system, the corresponding grasp pose generation algorithm can be directly applied. The preprocessing methods primarily rely on 3D bounding box and point cloud analysis. For instance, in the case of cylindrical objects, the height direction can be determined as the longest edge of the 3D bounding box.

6.6.2. 3D Object Part Detection

In the current setup, I perform 3D detection of the entire object and subsequently generate grasp poses. However, I find that for certain objects, it is possible to only detect the graspable portions instead of the entire object. For instance, in the case of a screw, I can limit the 3D detection to only the handle portion. This setting is beneficial for objects that exceed the gripper's size, as it simplifies the grasp generation algorithm. For smaller-sized objects, it remains reasonable to detect the entire object.

6.6.3. Robots Grasp New Objects

I demonstrate grasp pose generation algorithms for 18 objects belonging to 7 different shapes, without relying on any grasp-specific training. Although the number of included object categories is limited, the method can easily be applied to new objects, scenes, and sensors. By following the same technical pipeline of data collection, annotating 2D bounding boxes, training 2D detectors, and designing grasp pose generation algorithms based on prior knowledge, I can adapt the method to new scenarios. Leveraging mature 2D detection techniques, I achieve stable detection with only approximately 200 training samples. The manual annotation of 2D bounding boxes for a single object takes approximately half an hour. When the shape of a new object falls into one of the 7 predefined categories, the corresponding grasp pose generation algorithm can be directly utilized. However, when the shape of a new object does not belong to any of these 7 categories, a new grasp pose generation algorithm needs to be designed.

6.6.4. Object-level Grasping Pose Detection

Many existing robotic grasping studies are sampling-based, where grasp points are sampled from point clouds to generate grasp poses. Additionally, to avoid collisions, grasp poses in contact with the point cloud are discarded. This approach can be considered 'point-level' grasp pose detection, but it fails to distinguish between points belonging to the object and noise, leading to the elimination of some grasp poses in contact with noise points. Moreover, this method performs poorly when objects are occluded and cannot sample grasp points in such cases. I believe that this 'point-level' approach lacks a holistic understanding of the grasping object. In addition to the aforementioned issues, I also observe in experiments that this method generates grasp poses even for objects that are not graspable, such as lying boxes. In contrast, the proposed method is 'object-level', which addresses these problems to some extent. Some comparisons in Figure 6.8 illustrate this phenomenon.

6.7. Conclusion

In this study, I propose a novel 6-DoF grasp pose detection method that achieves dense grasp pose detection without the need for grasp pose annotations or grasp training. The method operates at the 'object-level', allowing for the classification of generated grasp poses and enabling user-specified or target-oriented grasping. An important advantage of the proposed approach is its ability to maintain stable grasp detection even when the grasping target is partially occluded, differentiating it from existing methods. Extensive experiments are conducted to validate our method's rapid adaptability to new scenes and objects, the stability and consistency of generated grasp poses, and the high success rate of robot grasping user-specified targets, even in the presence of partial occlusions.

The development of GoalGrasp opens new possibilities for enhancing robotic grasping capabilities without the need for extensive training or grasp pose annotations. However, the method still requires further exploration in certain aspects. For instance, incorporating spatial localization information from 3D detection to grasp specific positions of objects would help overcome the challenge of multiple identical objects in the scene. Additionally, I can investigate the integration of large language models into the grasping system, allowing users

to directly specify the robot's grasping target using natural language instructions. These avenues of research have the potential to enhance the capabilities and usability of our grasp pose detection method.

7. A NOVEL HRI THEORY: ANTICIPATORY HRI MODEL – PEER ROLE

In human-robot interaction scenarios, humans and robots share an environment and expect to make some decisions that move the environment towards desired states. This environment has the attributes of 'state' and 'time', where 'state' represents many components, such as humans, robots, and various information, and 'time' refers to the dynamic variation of the 'state'. In this chapter, I investigate a novel HRI model with anticipatory ability. This model allows for the prediction of future states of the system and accelerates the system's progression towards the desired state by optimizing the potential actions that can be taken. To achieve that, I propose a solution called online deep model predictive control (Deep-MPC) and apply it to robot-to-human handover tasks.

7.1. Introduction

Robots have become increasingly important in various industrial and service fields, with applications ranging from industrial robots to home service robots. In certain scenarios, robots need to collaborate with humans to complete tasks, such as human-robot handovers. In these tasks, humans and robots can be considered peers. The Literature Review defines a human-robot peer relationship as one where both parties are regarded as equals in terms of social status. In this relationship, the robot is not merely an instrument or device but a

partner or co-worker capable of diverse interactions with humans. When humans and robots are viewed as teammates working together, they can be considered peers (Groom et al. 2007). In this study, I focus on the human-robot handover task, where the robot acts as an assistant to accomplish tasks alongside humans, thus classifying both as peers.

In robot-to-human handovers, the robot must recognize the human's state, such as the position and motion of their hands, to take appropriate actions. Models that select actions based on the robot's current observations of the human's state are commonly used in human-robot interaction (HRI). However, these models often suffer from delays, causing the robot's actions to lag behind the current state of the system. Additionally, these models lack the ability to autonomously learn and adapt to user behavior. To address these limitations, I propose an anticipatory human-robot interaction model. This model aims to predict future states and optimize actions accordingly, enabling the robot to proactively interact with humans and adapt to their behavior. With the concept of HRI – Peer Role in mind, I extend the HRI Model – Peer Role on a time scale to formulate a novel model called the Anticipatory HRI Model – Peer Role, which is then deployed in HRI tasks.

Anticipatory control has been explored in some human-robot tasks, which has been reviewed in the Literature Review. However, there is limited research focused specifically on the human-robot handover task, and the existing studies often lack the capability to for fully online learning and optimization. This limitation hinders the efficient application of human-robot handover in various robotics scenarios. To overcome that, I propose a solution called online deep model predictive control (Deep-MPC) and apply it to robot-to-human handover tasks. Several studies have explored the integration of deep learning and Model Predictive Control (MPC) methods. One of the most relevant research areas to this study is learning system dynamics using neural networks (Hewing et al. 2020, Hafner et al. 2019a, Hafner et al. 2019b). Hansen et al. (2022) introduced the TD-MPC method, which employs neural networks to learn system dynamics and combines model-based and model-free reinforcement learning to achieve optimal control strategies. Hafner et al. (2019a, 2019b) used the cross-entropy method (Rubinstein et al. 1997) or reinforcement learning algorithms to optimize actions based on neural network-based dynamics. However, the application of

fully online neural network dynamics with gradient backpropagation in real robots is rarely addressed in existing research. The proposed approach, in contrast, is particularly well-suited for real-world robotic tasks. Additionally, Lucia et al. (2020) used deep neural networks (DNNs) to approximate the optimal control policy for MPC problems offline, subsequently applying this approximation for online control. Similarly, Karg et al. (2020) utilized DNNs to approximate the optimal control law of MPC, replacing MPC to achieve real-time control. Hoeller et al. (2020) introduced a Deep Model Predictive Control (DMPC) approach, where an MPC policy acted as an actor to interact with the environment, collecting data for training a critic model represented by a neural network. Elnour et al. (2022) employed a neural network to learn the dynamics of sports facilities and applied MPC methods to address the corresponding problem. Lenz et al. (2015b) proposed a variant of DeepMPC, which required manual data collection and offline training of the dynamics model. In contrast, the proposed method achieves fully online learning and control without any manual intervention, making it well-suited for real-world applications.

With the advancement of robot learning technology, numerous studies have focused on enhancing robot behavior through interaction with the environment (Ibarz et al. 2021). These studies cover tasks such as robot manipulation (Kroemer et al. 2021), navigation (Bansal et al. 2020), and quadruped locomotion (Wu et al. 2023). Some research (Hansen et al. 2022) has also utilized Model Predictive Control (MPC) for task learning, but their performance was only evaluated in simulation tasks. Enabling robots to learn task execution in real-world scenarios is a critical aspect of robotics research.

In this study, I propose Deep-MPC, which incorporates a 3D hand detector, an online learning transition model, and a data-driven MPC framework. Specifically, I use the 3D hand detector from Study 2 for hand detection, providing visual input for the robotic system. To achieve state anticipation, I introduce a Deep Model Predictive Control (Deep-MPC) approach, a data-driven method that leverages online learning from data collected during robot-environment interactions to predict future states and optimize current actions. Deep-MPC employs a neural network as the state transition module, taking states and actions as inputs to predict subsequent states. The method performs predictions for H steps, calculates

the loss function by comparing these states to the target state, and optimizes actions at each time step through gradient backpropagation.

7.2. Anticipatory Control on Robot-to-human Handover

7.2.1. Deep Model Predictive Control

Model Predictive Control (MPC) is a type of advanced control strategy that uses a mathematical model of the system being controlled to predict future behavior and determine the optimal control actions to achieve desired objectives. Generally, MPC can be described as Equation (7.1).

$$\min_{a_i} \sum_{i=t+1}^{t+H} F(s_i, \quad \hat{s}_{i+1}) \tag{7.1}$$

s.t.

$$s_{i+1} = T(s_i, \quad a_i)$$

where F is the cost function that measures the distance between s_i and $\hat{s}_{i+1}$. s_i is the state at time i, and $\hat{s}_{i+1}$ is the goal state at time i. H is the horizon. T is the transition function of the system. a_i is the action adopted by the system at time i.

MPC possesses the capability of anticipatory control, as it rollouts the future states H-steps ahead and utilizes them to optimize the current actions. To achieve that, one needs to formulate the mathematical or physical model of system transitions, which can sometimes be difficult or even infeasible. In this paper, I propose Deep-MPC, which is an online data-driven MPC that enables learning from scratch. In other words, a robot can learn how to finish a task without any prior knowledge of system transitions by using Deep-MPC. Figure 7.1 demonstrates the overview of Deep-MPC. Here, a neural network is leveraged to learn the transition function, that is T, see Section 7.2.2 for more details. In Deep-MPC, a robot captures environmental data using an RGB-D sensor, which is subsequently processed by RCV, a 3D object detector. RCV yields the status of objects in relation to the robot, providing spatial information for motion planning and control. Then, the robot rollouts H steps using T. At each step, the robot samples an action (a_i), which is then fed into T along with s_i. As a

consequence, an anticipatory state s_{i+1} can be generated by T. Repeatedly, an anticipatory sequence of states can be obtained, as shown in Figure 7.1. Compared with expected states, $\hat{s}$ in Figure 7.1, by loss function (F), I can obtain the total loss of anticipatory states, which is leveraged to optimize actions. Note that the expected states are manually specified by a human operator. The state (s_i) observed by the robot is generated from the center point of the 3D bounding box detected by RCV.

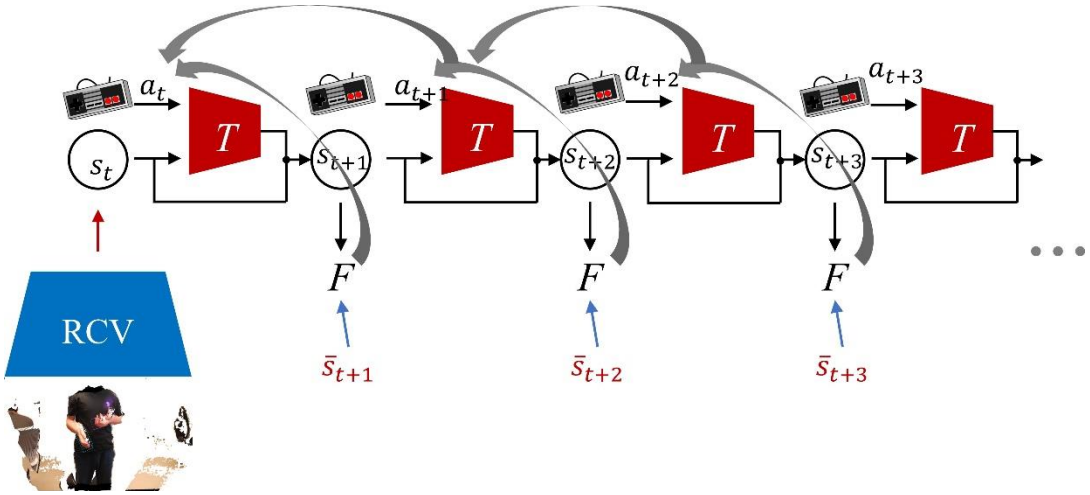


Figure 7.1. Overview of Deep-MPC. The red arrow indicates state perception, the black arrows represent forward data flow, and gray curved arrows denote gradient back-propagation. *T* is the transition model formulated by a neural network.

Once the anticipatory sequence of states ($s_i\{i = t + 1, \ldots t + H\}$) and actions ($a_i\{i = t + 1, \ldots t + H\}$) is obtained, a gradient-based optimization method can be applied to optimize the actions, as the transition model (T) is implemented as a neural network. The flow of gradient is shown as gray curved arrows in Figure 7.1. The update law is shown in Equation (7.2).

$$a_i \leftarrow a_i - \eta \frac{\partial \sum_{n=t+i+1}^{t+H} l_n}{\partial a_i} \tag{7.2}$$

where η is the learning rate, and $l_n = F(s_n, \hat{s}_n)$ is the loss at step n. Specifically, a_i affects only the loss generated in the subsequent time steps, that is from $t+i+1$ to $t+H$. In this manner, all actions can be planned to decrease the total loss.

In general, the control law for Deep-MPC can be summarized as follows:

$$\min_{a_i} \sum_{i=t+1}^{t+H} ||s_i - \hat{s}_i||^2 \tag{7.3}$$

s.t.

$$\begin{cases} a_i = P(s_i) \\ s_{i+1} = T_\theta(s_i, \quad a_i, \quad \delta t) + s_i \\ a_i \leftarrow a_i - \eta \dfrac{\partial \sum_{n=t+i+1}^{t+H} l_n}{\partial a_i} \quad for\ 0\ to\ H \\ \theta \leftarrow \theta - \beta \dfrac{\partial \left|\left|s_{i+1} - (T_\theta(s_i, \quad a_i, \quad \delta t) + s_i)\right|\right|^2}{\partial \theta} \quad for\ every\ k\ steps \end{cases}$$

where $s_i \in \mathbb{R}^{1\times3}$ represents the position (x, y, z) of the target in the robot coordinate system. $a_i \in \mathbb{R}^{1\times3}$ represents the action adopted by the robot, and $\delta t \in \mathbb{R}^{1\times1}$ represents the gap between s_i and s_{i+1}. The output is the next state s_{i+1}. Next, each component of Deep-MPC is introduced in detail.

7.2.2. Transition Model for Robots

In MPC, the transition model (T_θ) is employed to anticipate the succeeding state. In some previous studies, researchers formulated a neural network that directly predicts next state based on the current state and action, as shown in the left-hand side section in Figure 7.2 and Equation (7.4).

$$s_{i+1} = T_\theta(s_i, \quad a_i) \tag{7.4}$$

where θ represents the parameters of a neural network. By contrast, I propose a new structure, see the right-hand side section in Figure 7.2, which utilizes a neural network to predict the state's rate of change, given the current state, action, and interval. This is inspired by the state equation of a system, see Equation (7.5).

$$\begin{cases} \dot{s} = \boldsymbol{A}s + \boldsymbol{B}a \\ y = \boldsymbol{C}s + \boldsymbol{D}a \end{cases} \tag{7.5}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are coefficient matrices. s, a, and y refer to the state, action and output of the system, respectively. Specifically, the differential equation expresses the state variables as

time derivatives, describing how the system's state changes with respect to time. The response of the system, given initial states and inputs, can be derived. Inspired by this, I employ a neural network, specifically a MLP with ReLU as the activation function, which does not include any recurrent units. This neural network is used to simulate the differential equation and predict the next state, as depicted in Figure 7.2 and described by Equation (7.6).

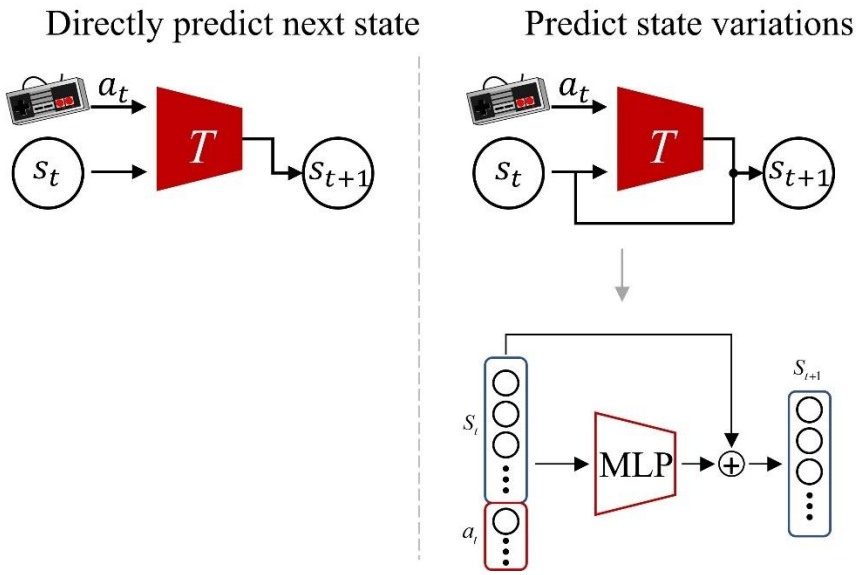


Figure 7.2. Transition models. The left-hand side represents a direct transition model that predicts the next state based on the current state and action, while the right-hand side represents a transition model that predicts state variations. A multi-layer perceptron (MLP) with ReLU as activation function is applied.

$$s_{i+1} = T_\theta(s_i, \quad a_i, \quad \delta t) + s_i \tag{7.6}$$

where δt is the interval between s_i and s_{i+1}. In experiments, I observe several advantages of this architecture, including faster convergence rates, and reduced prediction errors resulting from data scarcity in early robot task learning. Similarly, this architecture can be viewed as a form of skip connection (He et al. 2016), but I only implement skip connections for a subset of the input.

I adopt an online mode to train the transition model, as illustrated in Figure 7.3. The update law is shown in Equation (7.7).

$$\theta \leftarrow \theta - \beta \frac{\partial \left\| s_{i+1} - (T_\theta(s_i, \quad a_i, \quad \delta t) + s_i) \right\|^2}{\partial \theta} \tag{7.7}$$

The robot executes actions in the real-world environment and collects state transition data (s_i, a_i, δ_i, s_{i+1}), which is then used to train the transition model. At each step, the robot captures environmental data using an RGB-D sensor, which is subsequently processed by RCV, which yields the status of objects in relation to the robot. For further details on the training process, please refer to the next section.

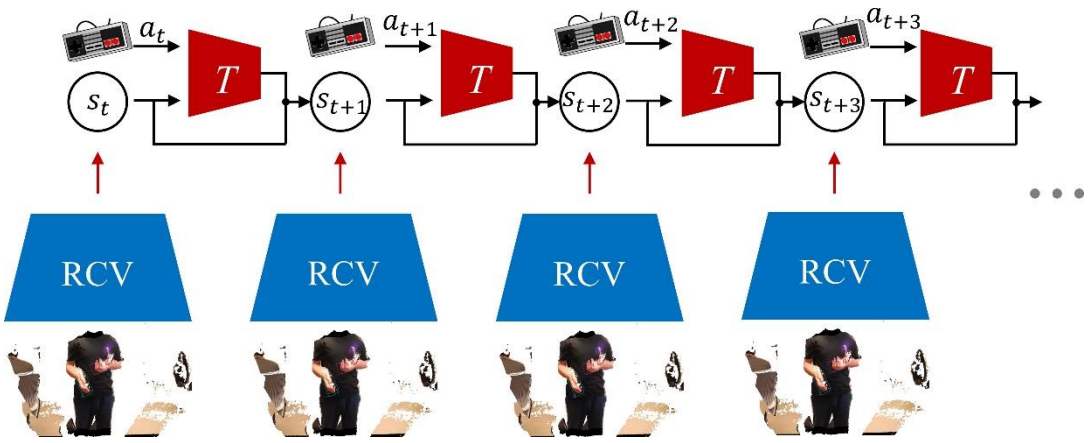


Figure 7.3. Online training of the transition model.

7.2.3. Anticipatory Control for Robots

In this section, I introduce a comprehensive anticipatory control algorithm, detailed in Algorithm 5, which can be applied to robots using Deep-MPC and the transition model. The robot is equipped with a 3D hand detector (RCV) trained to detect hands in a 3D space. Additionally, the robot is fitted with sensors capable of capturing real-time RGB and point cloud data from the environment. The robot performs the following operations:

- **Executing Deep-MPC** based on the current state perceived by RCV. In each rollout, a proportional controller computes a coarse action. After predicting H steps ahead, a gradient descent algorithm optimizes all actions to minimize the distance between the predicted states and the target states.

- **Executing the first action**, then perceive the environment to obtain the next state. Note that I use receding MPC, which involves using the transition model to predict future states and optimize actions over a finite time horizon; however, only the first action is executed.

- **Collecting transition data** of the robot and update the transition model.

In the initial operation, I use a simple proportional controller to generate initial robot actions, which accelerates the convergence rate of action optimization compared to random sampling or training a policy network. This approach enhances the efficiency of robot learning by enabling the optimization process to quickly converge to an effective solution. Importantly, this proportional controller does not rely on any knowledge of the system dynamics, allowing the robot to independently learn how to perform the task. All operations are executed online, enabling the robot to collect data and optimize its model simultaneously. This allows the robot to learn task completion rapidly by interacting with the real environment as much as possible. This method is significant for robotic applications as it reduces the need for complex modeling processes and increases the robot's level of intelligence.

Algorithm 5 Anticipatory Control on Mobile Robots

Require: Create an empty buffer D, a proportional controller P, anticipatory horizon H, expected states $\{\hat{s}_0^{H-1}\}$.
Require: Initialize neural network T_θ with parameters θ, and a mobile robot with a RGB-D sensor.
Require: Trained RCV model.
  while not ended **do**
    Capture environmental data using an RGB-D sensor and note as *env*.
    $s_0 \leftarrow \text{RCV}(env)$
    for $i \leftarrow 0$ to L **do**
      // Deep-MPC
      for $h \leftarrow 0$ to $H-1$ **do**
        $a_{i+h} \leftarrow P(s_{i+h})$ {// P controller as a coarse Actor.}
        $s_{i+h+1} \leftarrow T_\theta(s_{i+h}, a_{i+h}, \delta t) + s_{i+h}$
        $l_{i+h+1} \leftarrow ||\hat{s}_h - s_{i+h+1}||^2$
      end for
      for $step \leftarrow 0$ to n **do**
        Update $\{a_i^{i+H-1}\}$ using Eq.(2)
      end for
      // Environment Interaction
      Perform a_i on the mobile robot.
      Capture environmental data (*env*) using an RGB-D sensor.
      $s_{i+1} \leftarrow \text{RCV}(env)$
      Add experience to buffer $D \leftarrow D \bigcup (s_i, a_i, \delta t, s_{i+1})$
      // Transition Learning
      for every k steps **do**
        Perform a gradient descent step on $||s_{t+1} - (T_\theta(s_t, a_t) + s_t)||^2$ with respect to θ.
      end for
    end for
  end while

7.3. Real-world Robot Anticipatory Control Experiments

7.3.1. Real-world Robot Platform

I assemble a physical robot system as an experimental platform, as shown in Figure 7.4. The system consists of an UFACTORY xArm 7 with an UFACTORY gripper, an RGB-D sensor (Azure Kinect DK), and a desktop with a Nvidia 3090 GPU. The Deep-MPC can achieve a control rate of around 6 Hz.

Figure 7.4. The experimental platform. The proposed Deep-MPC is deployed on the platform and is performed in robot-to-human handover tasks.

7.3.2. Comparison of Two Transition Models

In Figure 7.2, two transition models are discussed: one that directly predicts the next state and another that predicts the rate of change of the state. Here, I compare the performance of these two models. Since Deep-MPC employs online learning, where the robot updates the model while interacting with the environment, it is crucial for the transition model to converge quickly with minimal early training errors. I collect 2,000 state transition datasets for training. It is important to note that all hyperparameters are configured identically for both models. Figure 7.5 illustrates the convergence speed and training errors of the two models, clearly showing that the proposed model demonstrates superior performance, making it more suitable for robot online learning scenarios.

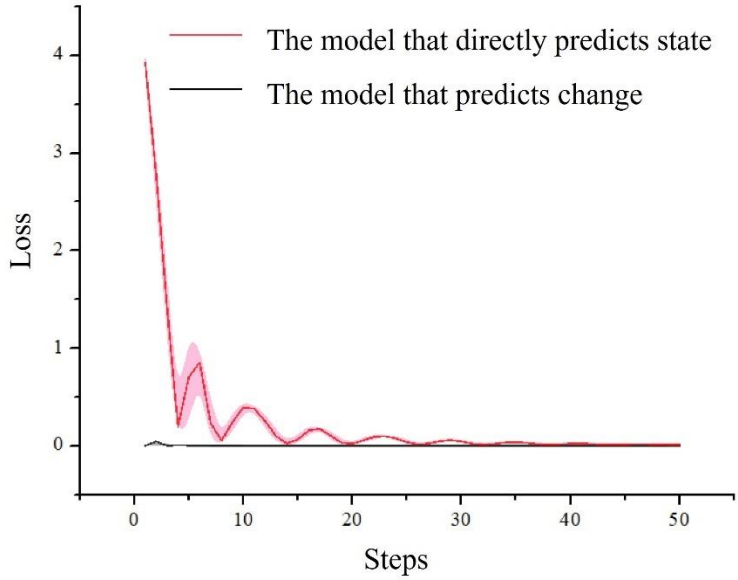


Figure 7.5. Convergence speed and errors of two transition models.

7.3.3. Robot-to-human Handover Experiments Using Anticipatory Control

The most important control variables in Equation (7.3) are the learning rate η of the action (a_i), the horizon H. I conduct multiple experiments to determine the optimal values of these variables, as shown in Table 7.1.

Table 7.1: Control variables of Deep-MPC.

Variables	η	H	Epoch for Deep-MPC	β	k
values	0.75	5	5	0.001	20

In the robot-to-human handover experiments, Deep-MPC is deployed in the system to control the robot's action. The RGB-D camera captures real-time scene data, and RCV is utilized to detect the position (x, y, z) of the user's hand from the scene. This hand position information is then fed into Deep-MPC. Deep-MPC operates in an online learning mode, collecting the system's dynamics data and utilizing it to train a dynamic model. By leveraging a data-driven MPC algorithm, Deep-MPC optimizes the robot's action to achieve effective action control. Specifically, it aims to transfer the object grasped by the robot to the user's hand during the handover task.

In this experiment, I compare the performance of a PI controller and the Deep-MPC controller. The experiment involved allowing the experimenter's hand to move freely within a certain range. I conducted 10 experiments for each control method, and Deep-MPC demonstrated superior motion characteristics. Under the same parameter conditions, it was able to complete the handover process more quickly, highlighting its adaptability to dynamic environments. The experimental results are presented in Table 7.2.

Table 7.2: Handover time.

Method	PI controller	Deep-MPC
Time	10 s	7 s

7.4. Discussion

7.4.1. Detection Interval

The proposed method achieves an average detection interval of approximately 0.167 seconds (6Hz), which means that the time interval between s_t and s_{t+1} in Figure 7.2 is 0.167s. However, in the presence of occlusion, the robot may lose track of the target, resulting in an elongated interval between s_t and s_{t+1}. In order to alleviate the decline in the performance of the transitional model caused by the difference in detection intervals due to occlusion, I also used the detection interval as one of the inputs to the transitional model.

7.4.2. Anticipatory Ability of Deep-MPC

Deep-MPC demonstrates better dynamic motion performance compared to the PI controller in the robot-to-human handover task. This is due to its ability to adapt to user actions, allowing for faster adjustments of the robot's actions to accommodate the user's behavior. Its anticipatory capability primarily stems from the MPC algorithm and real-time gradient-based optimization. When the robot observes the current system state, it uses the learned dynamic model within the MPC framework to predict future system states and optimizes future states by optimizing the actions using gradient descent algorithms. This is the essence of the proposed new human-robot

interaction model. It exhibits stronger adaptability compared to interaction models that solely consider the current state, and it can adapt to user behavior patterns.

In the experiment, I specifically focus on comparing the duration of the handover process between the two controllers. I chose this metric as the basis for comparison because it is a parameter that many users express concern about. Quantifying other metrics such as safety, comfort, and similar factors proved to be challenging. Therefore, I opt to prioritize the duration of handover as a tangible and measurable criterion for evaluating the performance of the controllers.

7.4.3. Human-robot Interaction from A Robot Perspective

Deep-MPC is an anticipatory human-robot interaction model that endows robots with the ability to predict the future states of the environment and optimize current actions based on these predictions. This capability aims to simulate the behavior observed in human-human interactions, where one party often anticipates the actions of the other and optimizes its own behavior accordingly. Deep-MPC mimics this ability by establishing a human-robot interaction model from the perspective of the robot, thus granting the robot human-like capabilities. In an intuitive explanation, Deep-MPC can be seen as an approach that enables the robot to anticipate and optimize its actions in a manner similar to how humans interact with each other.

8. ROBOT-TO-HUMAN HANDOVER MODEL AND EXPERIMENTS

In this chapter, I explore the robot-to-human handover model and propose a novel interaction model. Benefit from the outcome of Study 2 to 4, I assemble a physical robot-to-human handover robotic system. This robot-to-human handover system allows users to engage in experiments and enables the identification of the handover interaction model. Firstly, some important factors in this interaction process are identified in Study 1. Then, I invite individuals to simulate users with limited mobility and set different interaction modes for the robot. Through the users' experience of these different interaction modes, I collect their feedback using questionnaires. Once obtaining the feedback, I summarize the essential factors and develop a robot-to-human handover interaction model. To validate this model, a validation experiment is conducted. Participants are invited to experience the handover interaction model, and their feedback is collected to validate the newly proposed interaction model.

8.1. Introduction

In recent years, the development of assistive robotics has gained significant attention, particularly in the context of enhancing the quality of life for individuals with temporary or permanent mobility impairments. One critical aspect of assistive robotics is the robot-to-human handover, a process where a robot delivers an object to a human user. This interaction is not only a fundamental task in human-robot collaboration but also a complex one that

requires careful consideration of safety, efficiency, and user comfort. Robot-to-human handover involves several key components: the robot must accurately perceive the human's position and intent, plan a safe and efficient trajectory, and execute the handover in a manner that is intuitive and comfortable for the human. This process becomes even more crucial when the human user is in a vulnerable state, such as being ill or injured, where their ability to move or react may be compromised. In such scenarios, assistive robots can play a vital role in providing support and improving the user's autonomy and quality of life.

In this study, I focus on developing and evaluating a robot-to-human handover model tailored for users with temporary mobility impairments. I construct a real-world robotic system to conduct robot-to-human handover experiments, gathering user experience data throughout the process. Based on these experiments, I propose a novel interaction model for robot-to-human handover. The proposed interaction model has been preliminarily validated through a series of experiments, demonstrating its potential effectiveness in real-world applications.

8.2. Method

8.2.1. Robot-to-human Handover Experiments Design

Based on the outcomes from Study 2 to 4, I integrate a real robot-to-human handover robot in this study to investigate the impact of various factors explored in Study 1 in this interaction. A combined approach involving questionnaires and experiments is employed. The overall research approach involved inviting multiple participants to experience different operation modes of the robot while collecting feedback data during the experimental process. The experiment setting is demonstrated in Figure 8.1. It shows the experimental setup for the robot-to-human handover interaction study. This system includes a robotic arm with a gripper, a sensor to detect human and objects, and a control algorithm for smooth and safe handovers. In the experiment, participants interact with the robot to experience the handover process. The setup consists of the robot, the object to be handed over, and the participant.

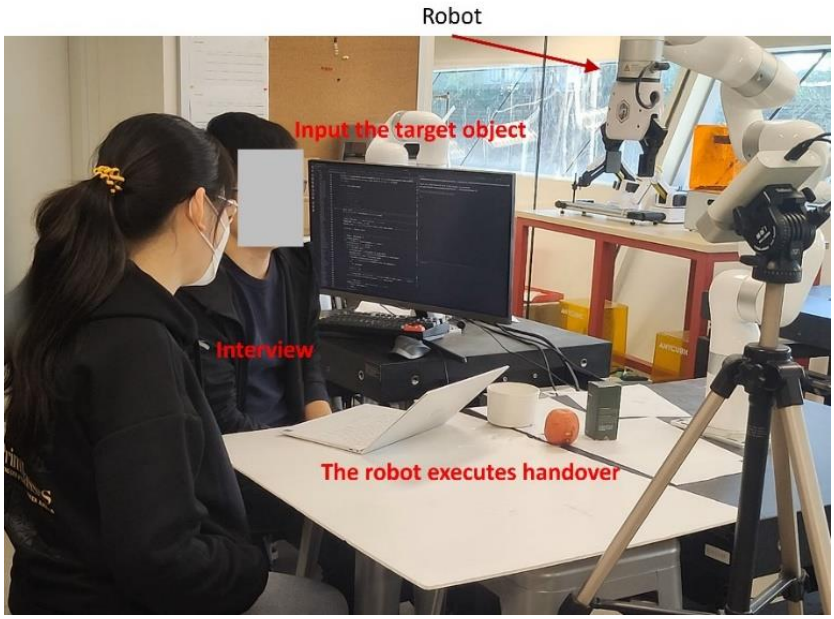


Figure 8.1. Experimental setting.

The experimental scenario simulates the situation in a daily household setting where individuals who are sick. For the sick patients, their limited mobility often makes it difficult for them to fetch objects, typically requiring assistance from caregivers. In this scenario, a robotic system with grasping capabilities can play a significant role. It should be noted that the experimental scenario assumes that the participants have normal hand functionality, enabling them to grasp objects retrieved by the robot. In the actual experiment, multiple healthy individuals are recruited to participate in a simulated experiment. Considering that most individuals have experienced illness, this simulation experiment is deemed reasonable.

To begin with, a corresponding questionnaire is meticulously designed, taking into consideration the factors mentioned in Study 1, along with some open-ended questions to capture more nuanced feedback. The factors from Study 1 that are incorporated into the questionnaire include the types of objects that need to be grasped during the robot-to-human handover, the speed at which the handover occurs, the path taken by the robot during the handover, and the modes in which humans receive the objects. Participants are asked to experience different settings for each of these factors. For instance, they interacted with the robot under various handover speeds, ranging from slow to fast, to determine their preferred speed. They also experience different handover paths to assess which path felt more natural

and comfortable. The questionnaire is structured to gather data. For each factor, participants are asked to rate their preferences. For example, questions included:

- What items would you like a robot to help you get when you are sick and bedridden at home?

- After trying out different speeds, which speed setting do you like best?

In addition to these questions, open-ended questions are included to capture detailed feedback and suggestions. Examples of open-ended questions are:

- Do you have any suggestions for improving the handover path?

- What kind of robot appearance do you prefer?

However, this experiment has some limitations. The main limitation is that the current robot used in the experiment is stationary, meaning that the robot for grasping and transferring objects is fixed on a tabletop. As a result, participants are unable to experience the robot autonomously fetching objects from a distance and navigating to the user, which may affect the realism of the user experience. This limitation arises from the current focus of research on object recognition, grasping techniques, and the study of interaction models during the object transfer process. To mitigate the impact of the lack of mobility on participants' experience, I inform them during the experiment that this specific aspect is omitted, and I provide them with relevant videos showcasing mobile robots to increase their awareness of this aspect.

8.2.2. Robot-to-human Handover Experiments Procedure

Participants: In this robot-to-human handover experiment, I recruit a total of 20 participants (10 males and 10 females) to experience the real robot-to-human handover. The age range of the participants was between 25 and 35 years. It should be noted that the experimental scenario assumes that the participants have normal hand functionality, enabling them to grasp objects retrieved by the robot.

During the experiment, participants were asked to interact with the robotic system in a controlled environment, where they received objects handed over by the robot. This setup allowed to systematically observe and analyze the handover interactions, focusing on key metrics such as safety, and user satisfaction. The data collected from these interactions provided valuable insights into the performance and usability of the robotic system, guiding further refinements and validations of the proposed handover model.

Procedure: Initially, 20 participants aged between 25 and 35 years are recruited for the study, with an equal distribution of 10 males and 10 females. Upon arrival, participants are given a detailed briefing about the experiment's objectives and procedures. They are informed about the different settings they would experience and the types of feedback they would be asked to provide. At the beginning of the experiment, participants will be asked if they are familiar with the application of assistive robots. Additionally, a video is played for participants to introduce how assistive robots assist users in retrieving items. The purpose of this step is to provide participants with a preliminary understanding of robot-to-human handover scenarios.

Participants are then asked to perform a series of handover tasks under different experimental conditions. These conditions varied based on several factors: the types of objects to be grasped, the speed of the handover, the path taken by the robot, and the modes in which participants received the objects. Specifically, participants receive different types of objects to assess the robot's adaptability and the ease of grasping. The speed of the handover is adjusted across trials, ranging from slow to fast, to determine the optimal speed for user comfort and efficiency. The robot's handover path is varied, including direct linear paths and more complex curved trajectories, to evaluate which path is most intuitive and comfortable. Additionally, participants are asked to receive objects in different modes.

After completing the handover tasks, participants are asked to fill out a detailed questionnaire. Following the completion of the questionnaire, they are given the opportunity to provide any additional feedback or ask questions about the experiment.

8.3. Data Collection and Analysis

In this section, the experimental results and data analysis are presented and analyzed. After experiencing the different operation modes of the robot, their feedback on various interaction factors of the robot is collected.

8.3.1. Objects need to be Retrieved by Users

Figure 8.2 illustrates the categories of items that the 20 participants identified as requiring the robot's assistance for object retrieval. It can be observed that there is a general demand for water, electronic device, drug, and food. For the subsequent validation experiments, I select drug, food, and fruit as the test items.

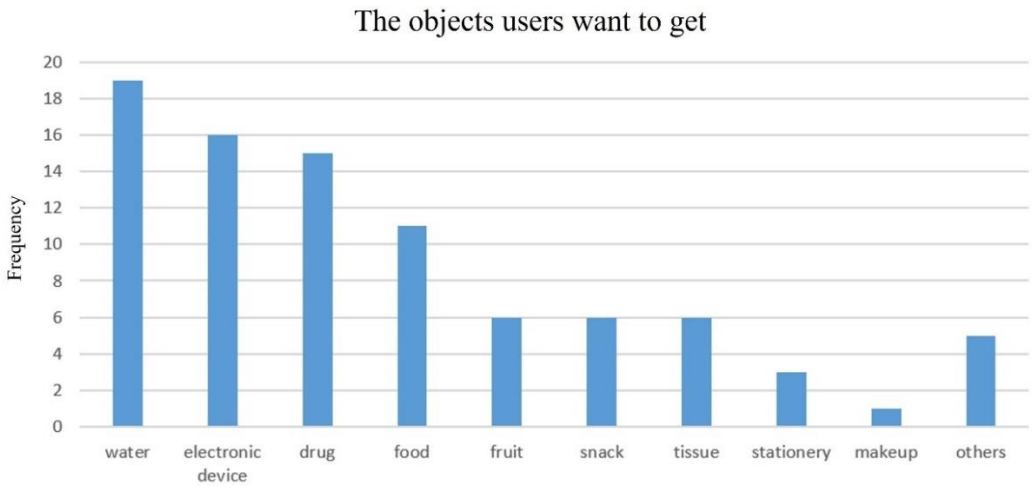


Figure 8.2. The objects users want to get.

8.3.2. Robot-to-human Handover Speed

Regarding the handover speed, I provide five options: low speed (1-5cm/s), medium-low speed (5-10cm/s), medium speed (10-15cm/s), medium-high speed (15-20cm/s), and high speed (20-25cm/s). However, from a design perspective, it is customary to include higher speed intervals for users to experience and collect feedback. Considering safety concerns, I prioritize ensuring the users' absolute safety during the experiments. As a result, I do not include higher speed intervals. Although most users selected the 20-25 cm/s speed interval, I cannot guarantee whether users would have chosen speeds greater than 25 cm/s.

Consequently, the statistical results may be questioned due to the lack of experiments with higher speeds. It is important to note that the experiments are conducted using the UFactory xArm 7 robot, a relatively large robot lacking active force control. Therefore, higher speeds could pose a significant risk to users if any bugs or malfunctions occurred. During the experiments, there are indeed instances of unexpected robot behavior. These safety considerations and incidents with the xArm 7 robot necessitated the cautious approach in setting the speed intervals.

The experimental results, as shown in Figure 8.3, demonstrate that the high speed is preferred by users. However, it is important to note that most participants emphasize the need for the robot to ensure safety. I employ the chi-square test to validate the significance of user selections. The results revealed a p-value of 0.002, which is less than 0.05. These findings indicate a significant difference in user choices. In the subsequent validation experiments, I set the robot's speed to 20-25cm/s, primarily considering safety requirements.

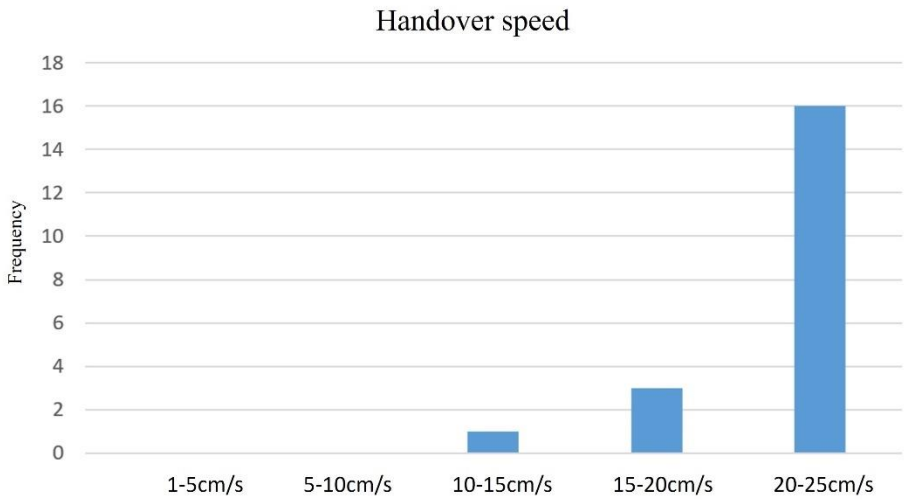


Figure 8.3. Robot-to-human handover speed.

8.3.3. Robot-to-human Handover Robot Movement Path

For the pathway of the robot's object transfer, I define two paths as shown in Figure 4.4. The experimental results, as shown in Figure 8.4, reveal that most participants chose the second path, as they perceive it to be more natural and efficient. Three individuals express

indifference towards the handover path taken by the robot, while only one participant opt for the first path. I employ the chi-square test to validate the significance of user selections. The results reveal a p-value of 0.009, which is less than 0.05. These findings indicate a significant difference in user choices. In the subsequent validation experiments, the second path is used.

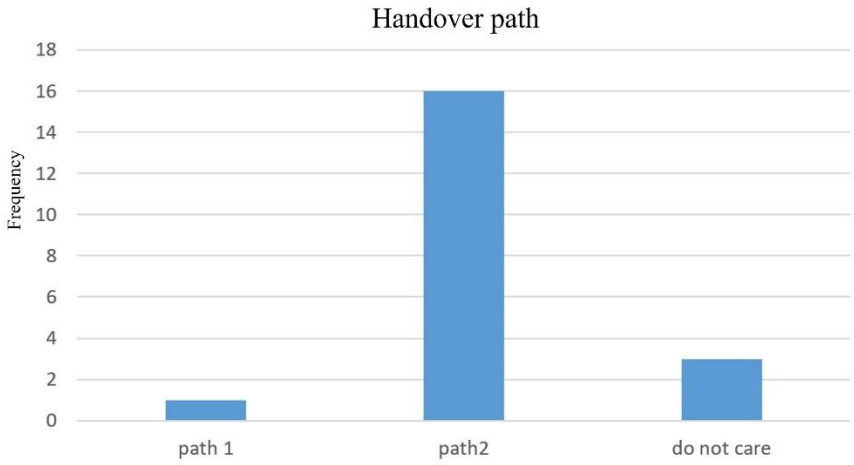


Figure 8.4. Robot-to-human handover path.

8.3.4. Receive modes Adopted by Users in Robot-to-human Handover

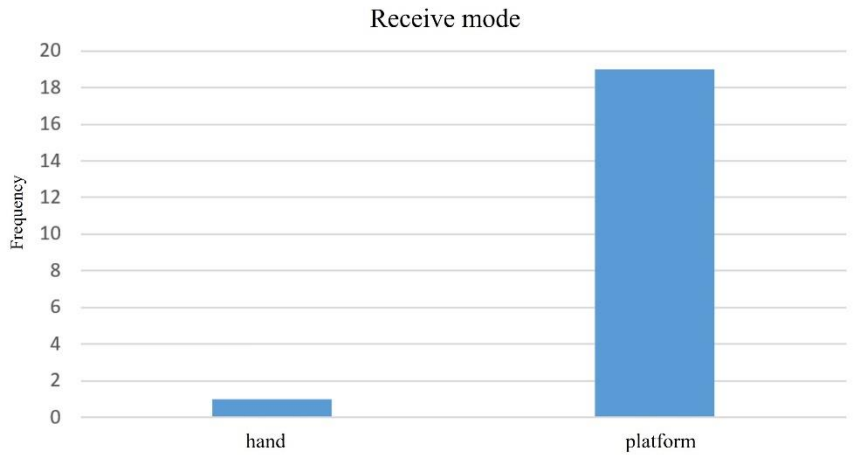


Figure 8.5. Users receive mode.

Regarding how people receive objects handed over by the robot, I implement two modes: one where the user directly grasps the object from the robot's gripper and another where the robot places the object on a fixed platform for the user to pick up, as shown in Figure 4.4. The experimental results, as shown in Figure 8.5, indicate that most participants prefer the mode

where the robot places the object in a designated area. This preference primarily stems from the perceived safety and reduced waiting time associated with this mode. I employ the chi-square test to validate the significance of user selections. The results revealed a p-value of 0.001, which is less than 0.05. These findings indicate a significant difference in user choices. For the subsequent validation experiments, I select this mode.

8.3.5. The Weight of Factors on Robot-to-human Handover Interaction

During the experiment, the participants experience various factors in the robot-to-human handover interaction mode. After their experience with these factors, they allocate importance weights to indicate the perceived significance of these factors in the interaction mode. I collect data from 20 participants and calculate the average importance weights for four factors: **success rate** (the number of successful trials divided by the total number of trials), **speed**, **hand posture**, and **grasp area**, as shown in Figure 8.6. The experimental results reveal that participants place significant emphasis on the success rate compared to other factors, indicating their concern about whether the robot can successfully hand over the object to them. The majority of participants express their willingness to actively participate in the object transfer process to ensure a higher success rate. For instance, they adjust their hand posture, grasp area, and paid attention to the robot's status feedback to ensure the successful completion of the task. Therefore, in the subsequent validation experiments, priority will be given to ensuring a high success rate in the robot's object transfer.

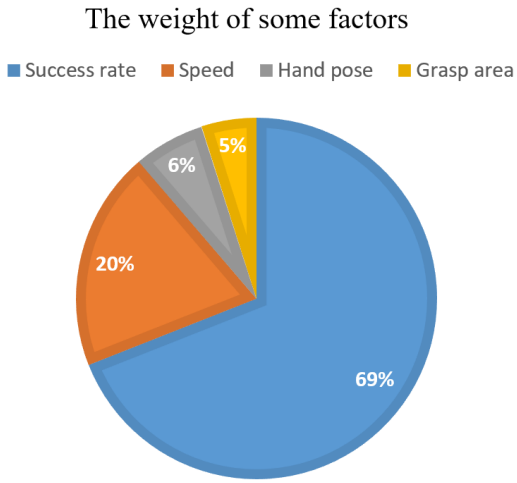


Figure 8.6. The weight of four factors.

8.4. Robot-to-human Handover Interaction Model and Validation Experiments

Based on the analysis of the experimental results, I develop a novel robot-to-human handover interaction model, as shown in Table 8.1 and Figure 8.7. Here, each considered factor is set to the mode that is most acceptable to the users.

Figure 8.7 presents the proposed robot-to-human handover interaction model. The blue arrows represent the flow of information within the interaction model. Users communicate specific items they want the robot to fetch through verbal commands. The verbal commands consist of three pieces of information: the item, its location, and the target location. The robot receives the user's instructions and utilizes a large language model to comprehend the user's intent, followed by performing recognition and motion planning accordingly. It is assumed that the robot can accurately perceive the user, objects, and the surrounding environment. Building upon the research outcomes of Study 2 and 3, the robot accomplishes object recognition and grasping tasks. Subsequently, the robot executes the handover process using the anticipatory control method developed in Study 4. By utilizing information from the object, user, and robot, the anticipatory control method performs online motion optimization and generates the robot's actions. Throughout the handover interaction, the robot communicates various information to the user, such as the object, motion speed, and trajectory. These factors are explored in Study 5. After receiving the robot's information, the user provides feedback. If the task is successful, the interaction concludes. However, if the task fails, the interaction resumes, and the user can inform the robot of the task through verbal commands once again.

This comprehensive robot-to-human handover interaction model encompasses both robot technology and a user-centered human-robot interaction model design, as outlined in this research.

Table 8.1: The proposed robot-to-human handover interaction model.

Factors	Settings

Objects	Water, drug, electronic devices
Success rate	Top priority
Handover speed	20-25 cm/s
Hand pose	Not be considered
Hand grasp area or platform	Platform
Robot path	Path 2 (human-like path)

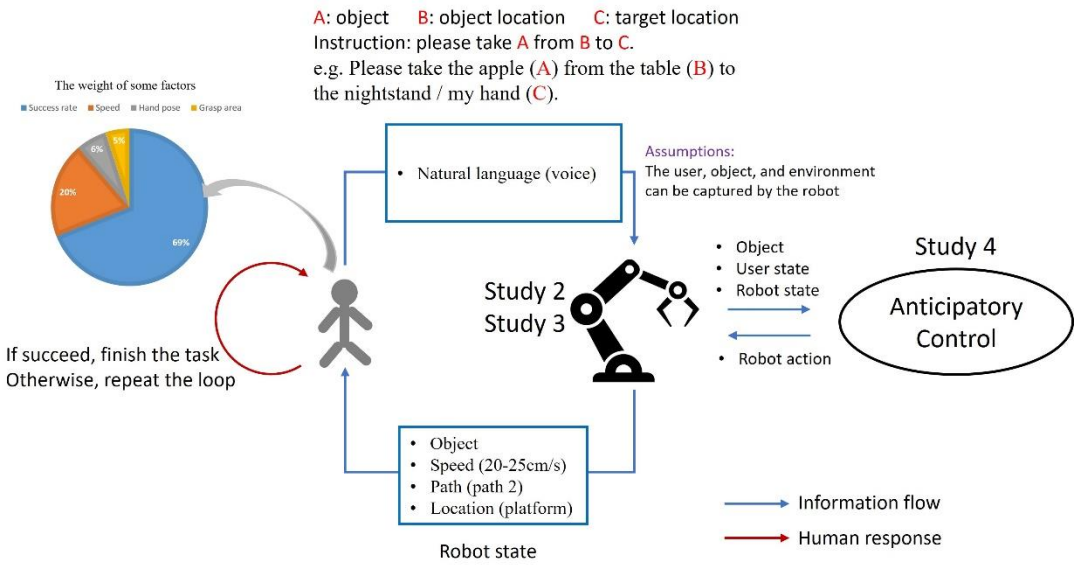


Figure 8.7. Robot-to-human handover interaction model (Study 5).

To validate this interaction model, I invite 7 participants to conduct validation experiments, allowing them to experience this interaction model and collect their feedback on trustworthiness, comfort, safety, and satisfaction. I quantify the four indicators on a scale of 1 to 5, where 5 represents "very satisfied", 4 represents "satisfied", 3 represents "neutral", 2 represents "dissatisfied", and 1 represents "very dissatisfied". Specifically, 7 participants are involved in experiencing the proposed robot-to-human interaction model. I collect their interaction experiences separately, and the experimental results are shown in Figure 8.8. The average scores provided by the users for trust is 4.14, indicating a high level of trust in the robot's capabilities and reliability. For comfort, the average score is 3.71, suggesting a moderate level of comfort experienced by the users during the handover interactions.

Regarding safety, the average score is 4, indicating that users perceive the robot's actions as safe and reliable. Lastly, the average score for satisfaction is 4, indicating a high level of satisfaction with the overall robot-to-human handover experience.

These results suggest that the proposed robot system and human-robot interaction model is able to establish a sense of trust and safety among the users, while also providing a satisfactory experience. The moderate level of comfort suggests that there may be room for improvement in enhancing the user's comfort during the handover interactions. These findings contribute to the understanding of the user's perspective in human-robot handover scenarios and can guide future improvements and developments in this field.

Figure 8.8. The results of validation experiments.

8.5. Discussion

8.5.1. Limitations of Simulation Experiments

Due to the focus of my research on robot object recognition, grasping detection and execution techniques, and the human-robot interaction model during robot-to-human handover, other aspects such as robot navigation, speech recognition commands, and the physical appearance of the robot have not been prioritized. This necessitates compromises in the interactive

experiments, such as omitting direct observation of the robot's navigation and instead informing the participants about this process through alternative means, such as videos. Indeed, this selection may impact the participants' experience, but incorporating features like navigation into the system would introduce excessive complexity. Therefore, I have chosen to use alternative methods during the experiment to raise participants' awareness of the robot's movement as much as possible, thereby compensating for the limitation.

8.5.2. Other Factors on Robot-to-human handover Interaction Model

In addition to the aforementioned factors that affect human-robot interaction, I also sought feedback from the participants regarding other aspects during the experiments, such as robot appearance, gripper design, and interaction modalities. The experimental results reveal that the participants prefer lightweight robot designs, as it made them feel safer. Regarding interaction modalities, the majority of participants express the importance of multiple modes of interaction. They wish for the robot to be capable of receiving their instructions through speech, text, mobile apps, etc., and emphasize the accuracy of the robot's understanding of their intentions. These experimental findings can serve as fundamental principles for further enhancing robot design in the future.

8.5.3. Other Settings on Robot-to-human handover Interaction Model

Although a new interaction model is summarized in Section 8.4, it does not imply that other options are meaningless. For instance, regarding whether the robot should directly deliver the object to the user or place it in an accessible area, while most people may prefer placing the object in a designated area, there will still be individuals who prefer direct handover or situations where there is no available designated area around the user. Therefore, the robot should also be capable of the second mode. Similarly, for other factors, the robot should possess the ability to offer multiple selectable modes, allowing it to adjust its behavior to better adapt to different users and scenarios.

9. DISCUSSION

This chapter discusses the results of each study to address the research questions proposed in Chapter 1. By combining the outcomes from Studies 2 to 4, a robot-to-human handover system is proposed. Based on this system, all research questions are explored, and several conclusions are drawn. Furthermore, the proposed robot system is investigated from both a technical and user experience perspective, and the research findings are further discussed. Finally, the limitations of my research and future work are also discussed in this chapter.

9.1. Discussion of Research Questions

In the Introduction, after discussing the research background, I present the research questions. With these research questions in mind, I conduct this research to attempt to provide solutions and answers to these questions. Now, after conducting five studies, have these questions been addressed, and to what extent have the research objectives been achieved? First, let's review the research questions, as outlined below:

Research Questions:

- **RQ 1:** What are the challenging techniques and key factors in robot-to-human handover HRI?

- **RQ 2:** Can real-time robotic 3D object detection method in new scenes be achieved in the absence of 3D annotations?

- **RQ 3:** Can target-oriented 6-DoF grasp pose detection be achieved in robot-to-human handover tasks without grasping training?

- **RQ 4:** How to integrate anticipation into the HRI handover Model – Peer Role to form the anticipatory HRI Model – Peer Role?

- **RQ 5:** How to formulate a robot-to-human handover interaction model?

9.1.1. Understand the challenging techniques and key factors in robot-to-human handover HRI (RQ 1)

The development of assistive robots for fetching everyday objects for users involves an interactive system between humans and robots. In this system, humans and robots collaborate to achieve the task of object transfer, fulfilling the needs of humans. However, most existing research primarily focuses on robot perception and grasping techniques, while neglecting the comprehensive examination of this system from both the user's and robot's perspectives. In this research, I initially conduct simulation experiments where humans simulate the role of the robot in collaboration with users to complete the tasks. By obtaining feedback from both the users and simulated robots, I aim to identify challenging technologies and key factors that are significant in this context. These identified technologies and factors will guide the subsequent research.

It is important to note that the feedback collected from simulation experiments may introduce certain inaccuracies when compared to real-world experiments. Thus, I validate these factors in real-world experiments and develop a user-friendly HRI model in Study 5.

9.1.2. Can real-time robotic 3D object detection method in new scenes be achieved in the absence of 3D annotations? (RQ 2)

Thanks to the rapid development of deep learning techniques, robots have greatly improved their ability to understand scenes. However, many deep learning-based methods heavily rely on large-scale datasets annotated by humans. This problem becomes even more severe when dealing with 3D data, making it challenging to quickly leverage existing techniques to achieve

corresponding functionalities in real robot tasks. Yet, real-world robot scenarios are characterized by diversity and dynamic changes, which greatly reduce the application efficiency of robots in real environments. One fundamental capability of a robot is real-time object perception in 3D scenes, which is essential for robot-to-human handover tasks. The robot needs to identify objects in the scene based on user instructions. Therefore, a real-time 3D perception method that does not rely on 3D manual annotations and can quickly adapt to new scenes and objects needs to be proposed. However, this remains an unresolved problem for existing methods.

In this research, I propose a novel approach to address this goal. The proposed method is completely independent of 3D manual labels and can achieve rapid 3D recognition of various scenes and objects. Furthermore, experimental results demonstrate that the proposed method can be deployed on a robot to achieve real-time 3D object perception. As a result, this problem has also been addressed. However, the proposed method still has some limitations that can be further improved. For instance, the current approach determines the projection direction based on the distribution and geometric shape of the point cloud, which can introduce significant errors when objects are occluded. This limitation can be addressed by training a neural network model to predict the orientation of objects, which can be considered as future work. Additionally, the method needs further optimization to achieve higher real-time performance. Improved real-time performance would better adapt to dynamic scenes and a wider range of real-world robot scenarios.

9.1.3. Can target-oriented 6-DoF grasp pose detection be achieved in robot-to-human handover tasks without grasping training? (RQ 3)

Similarly, the development of 6-DoF grasp pose detection in robotics has also benefited greatly from the advancements in deep learning techniques. However, these methods also heavily rely on large-scale datasets and complex network training, which inevitably reduce their application efficiency in various robotic grasping scenarios. Moreover, existing methods have paid little attention to target-oriented grasp tasks in occluded scenes, making it challenging for them to adapt to user-specified object grasping in human-robot interactions. So, how to achieve user-specified target grasping? In this research, I propose a novel method

that leverages the 3D object detection approach introduced in Study 1. This method can provide 6-DoF grasp poses for user-specified targets without the need for annotated data or grasp training. Additionally, the proposed method can handle partial occlusion, demonstrating its advantages. As a result, this problem has also been addressed.

In Study 2, I present grasp pose generation algorithms for 18 objects across 7 distinct shapes, without relying on specific grasp training. While the number of object categories included is limited, the proposed method can readily adapt to new objects, scenes, and sensors. By following a consistent technical pipeline involving data collection, 2D bounding box annotation, training of 2D detectors, and designing grasp pose generation algorithms based on prior knowledge, I can extend the method to new scenarios. Leveraging established 2D detection techniques, I achieve reliable detection performance using only approximately 200 training samples. Manual annotation of 2D bounding boxes for a single object typically takes around half an hour. When a new object's shape falls within the predefined set of 7 categories, the corresponding grasp pose generation algorithm can be directly applied. However, if the shape of a new object does not fit into any of these 7 categories, a new grasp pose generation algorithm would need to be developed.

Most existing research in the field of grasping focuses primarily on either parallel jaw grippers or multi-fingered grippers, with limited studies considering both types of grippers. Multi-fingered grippers offer significantly higher degrees of freedom compared to parallel jaw grippers, posing substantial challenges for learning-based approaches. The proposed method, through the design of corresponding grasp pose generation algorithms for multi-fingered grippers, holds promise for extending its applicability to the realm of multi-fingered grippers as well. This will be one of the future research directions.

9.1.4. How to integrate anticipation into the HRI handover Model – Peer Role to form the anticipatory HRI Model – Peer Role? (RQ 4)

Unlike the various factors discussed in the robot-to-human handover model in RQ 2, this problem considers the robot's ability during the robot-to-human handover interaction. I aim to empower the robot with the capability to anticipate future states of the system during this

interaction. These system states can include the robot's state, the human's state, and potential disruptive states in the environment, such as obstacles that may appear unexpectedly. The goal is to simulate the ability of humans to anticipate the future states of each other and adjust their behaviors accordingly during human-to-human interactions. By posing this question, I aim to imbue the robot with such abilities, enabling it to enhance the intelligence of the interaction.

Furthermore, I emphasize the peer role between the human and the robot in this interaction process. This means that the human and the robot collaborate as partners, working together to accomplish the task of robot-to-human handover. The peer role relationship is supported by the experimental results in Study 5, where the majority of participants express their willingness to adjust their behavior to coordinate with the robot's actions to ensure the smooth completion of the task. The experimental results from Study 4 demonstrate that the anticipatory approach can expedite the robot-to-human handover process, providing some evidence of the effectiveness of the proposed method.

However, testing other aspects of performance, such as resistance to environmental disturbances, in the handover task is challenging due to limitations in the experimental setup, such as the fixed perspective of the robot. As a result, to some extent, this question has been addressed. Further research is needed in future work to explore this question more comprehensively and investigate other aspects of performance.

9.1.5. How to formulate a robot-to-human handover interaction model? (RQ 5)

There has been limited research on the robot-to-human handover interaction model, primarily due to limitations in robot technology. Existing technologies make it difficult for robots to identify and deliver grasped objects based on user requirements, making it challenging to study the various factors that influence this interaction model in detail. Users are unable to experience a real robot, making it difficult to study the robot-to-human interaction model from their perspective. However, with the advancements achieved through this research, I have partially realized the functionality of this robot, enabling further

exploration of the robot-to-human handover interaction model. I have identified several interaction factors within this human-robot interaction model and investigated the optimal values for each factor through experiments and questionnaires, as shown in Table 8.1. Hence, this question has also been addressed in this study. However, I have only explored some factors at present, and there are still many more factors that need to be studied, which will be part of my future work.

Although a novel interaction model is summarized in Section 8.4, it should not be interpreted as rendering other options meaningless. For example, when considering whether the robot should directly hand over an object to the user or place it in an accessible area, while most individuals may prefer the object to be placed in a designated area, there will still be instances where users prefer direct handover or where no designated area is available nearby. Hence, the robot should also possess the capability to accommodate the second mode. Likewise, for other factors, the robot should be equipped with the ability to provide multiple selectable modes, enabling it to adjust its behavior and better adapt to different scenarios.

Building upon the discussion of RQ1, overall, the ultimate technological goal for assistive robots is to adapt to diverse user needs and provide multiple selectable interaction modes to accommodate users' requirements as much as possible. Developing such a robot aligns with my long-term research objective.

9.1.6. Realization of Research Objectives

I propose 5 research objectives in the Introduction, as follows:

- **Objective 1**: To figure out the challenging techniques and key factors in robot-to-human handover interaction model.

- **Objective 2:** To develop a new 3D object detection method that can be used in robot-to-human handover for various objects. By using this method, the robot can recognize user-specified objects.

- **Objective 3:** To develop a target-oriented 6-DoF grasp pose detection method that can be used to grasp user-specified objects for users in robot-to-human handover.

- **Objective 4:** To formulate a real-time and online anticipatory human robot interaction Model-Peer Role on robot-to-human handover. This robot control model

- **Objective 5**: To develop a novel robot-to-human handover interaction model that can receive instructions from users and autonomously complete the recognition, grasping, and handover to meet the user's needs for retrieving objects.

Based on the previous discussion regarding the research questions, most of these objectives have been achieved. Specifically, (1) I propose some challenging techniques and key factors in robot-to-human handover interactions. (2) I introduce an efficient 3D perception method for robots. (3) I develop a reliable target-oriented 6-DoF robot grasp detection method. (4) I propose a real-time and online anticipatory HRI Model-Peer Role for human-robot interaction. (5) I propose a novel robot-to-human handover interaction model.

To validate the effectiveness of these methods, I integrate all the research into real robot scenarios and conduct simulation experiments, thus demonstrating the efficacy of the proposed approaches. Therefore, these objectives have been largely achieved.

9.2. Limitations and Future Work

9.2.1. The lack of mobility

One limitation of this research is the lack of mobility in the proposed robotic system. This limitation may have implications for the interaction experiments and could potentially impact the user experience. The inability of the robot to move restricts the range of interactions that can be explored and limits the diversity of scenarios that can be studied. This lack of mobility may result in a less dynamic and realistic interaction environment, potentially affecting the overall user experience and the generalizability of the research results. The primary reason for this limitation is that the technology for mobile navigation is relatively more mature compared to the 3D object perception and grasp detection techniques

in robotics. Therefore, the focus of this research is not on mobile navigation, and as a result, it is not included in the system. To minimize the impact of this limitation in the robot experiments, participants are informed about relevant mobile robot usage scenarios through videos before the experiments. This is done to provide them with a basic understanding of robot mobility operations and assisted mobility. Additionally, emphasis is placed on the research focus of the experiment, which is the recognition of grasp techniques and the handover interaction technology, to direct participants' attention more towards these aspects of the experience.

In future work, it is essential to address the lack of mobility by incorporating mobile capabilities into the robot system. The proposed techniques from this research can be integrated into a mobile robot with autonomous navigation capabilities to achieve a complete mobile operation and handover interaction process. This would result in a fully functional robot system capable of both mobility and handover interactions. The main challenge of this work lies in the integration of software and hardware systems, which involves interdisciplinary backgrounds and various engineering issues.

9.2.2. Simulation Experiment

Another limitation of this research is that all experiments are conducted in a simulated environment, and no experiments are performed in a home setting. This limitation primarily arises from the inability of the robot system to move, which prevents its deployment in real-life scenarios. In addition, the current experiments are conducted on simulated sick people, and more appropriate populations such as the real elderly or wheelchair users should be considered to further enhance the breadth of this study. Conducting experiments with these specific user populations will be a focus of future work.

In the current experiments, I conduct simulated experiments where participants imagined themselves in a home environment with limited mobility, and the robot assists them in retrieving objects. Following these simulated experiments, I derive a corresponding human-robot interaction model. In future research, this interaction model will serve as a reference

for subsequent experiments, which will be conducted with different user populations and in different scenarios, to meet the specific needs of diverse users and situations.

From a robotics perspective, a versatile assistive robot should possess strong adaptability to different scenarios, meaning that the robot needs to modify its interaction approach based on the specific user and context. In this regard, conducting simulated experiments initially to establish preliminary interaction patterns and subsequently optimizing these patterns and providing multiple interaction options to cater to different user needs is a reasonable approach. Therefore, in future work, I aim to further enhance the robot's adaptability to different scenarios, offering a variety of interaction modes for users to choose from, thereby augmenting the robot's versatility.

10. Conclusion

This chapter summarizes the contributions of this research and presents the conclusions drawn from the study.

The major contributions of this research are as follows:

- **Proposed techniques for 3D object perception**: The Recursive Cross-View (RCV) method proposed in this research demonstrates the ability to quickly adapt to diverse robotic tasks. A major challenge in 3D object detection is the heavy reliance on 3D annotations, which limits its real-world application. To overcome this challenge, the RCV method is introduced, which does not require any 3D annotation while still being able to predict fully oriented 3D bounding boxes. Instead, it leverages the three-view principle to transform 3D detection into several 2D detection tasks, requiring only a portion of 2D labels. I propose a recursive framework where instance segmentation and 3D bounding box creation via Cross-View are performed iteratively until they converge. Specifically, the method uses a frustum for each 2D bounding box, followed by the recursive process that eventually produces a fully oriented 3D box along with its associated class and score. Note that the class and score are provided by the 2D detector. Evaluations on the SUN RGB-D and KITTI datasets show that this method surpasses existing image-based techniques. To demonstrate the method's adaptability to new tasks, I apply it to two real-world scenarios: 3D human detection, and 3D hand detection. Consequently,

two new 3D annotated datasets are created, indicating that RCV can function as a (semi-) automatic 3D annotator. Additionally, I implement RCV on a depth sensor, achieving detection at 7 frames per second on a live RGB-D stream. This practical deployment highlights the method's applicability in real-time robotic applications.

- **Proposed techniques for 6-DoF grasp detection**: The introduced technique, GoalGrasp, offers a straightforward yet powerful solution for detecting 6-DoF robot grasp poses without the need for grasp pose annotations or training. This method facilitates user-specified object grasping even in scenes with partial occlusions. By combining 3D bounding boxes and human grasp priors, GoalGrasp introduces a new paradigm for grasp pose detection. The approach employs the RCV 3D object detector, which operates without 3D annotations, enabling swift 3D detection in unfamiliar environments. Utilizing the information from 3D bounding boxes and human grasp priors, GoalGrasp performs dense grasp pose detection. Tests conducted on 18 commonly used objects reveal that GoalGrasp can generate dense grasp poses for 1000 scenes without any grasp training, thereby creating an extensive grasp pose dataset. When compared to existing methods using a novel stability metric, GoalGrasp shows markedly higher stability in grasp poses. In experiments involving user-defined robot grasping, the method achieves a remarkable 94% success rate. Additionally, in scenarios with partial occlusion, the success rate remains high at 92%.

- **Proposed techniques for anticipatory handover**: To equip robots with the capability to predict system states during handover tasks, I introduce a method called Deep-MPC. This approach combines a 3D hand detector, an online learning transition model, and a data-driven Model Predictive Control (MPC) strategy. The 3D hand detector is used to identify hands, supplying visual data to the robotic system. To forecast future states, the Deep Model Predictive Control (Deep-MPC) method leverages online learning from interactions between the robot and its environment, allowing it to predict upcoming states and optimize the robot's current actions. The state transition module within Deep-MPC employs a neural network

that takes in states and actions to predict the next state. By making predictions over H steps, comparing these predictions to the desired state using a loss function, and refining actions through gradient backpropagation at each time step, the method ensures effective action optimization. Deep-MPC can be seen as a technique that builds a human-robot interaction model from the robot's perspective, endowing the robot with human-like predictive abilities.

- **Robot-to-human handover interaction model**: The proposed model aims to enhance the robot-to-human handover experience for users. The development of the model involves several steps. Firstly, some critical factors in the handover process, such as objects need to be grasped, speed, and path, are identified. These factors serve as the foundation for optimizing the handover interaction. To refine the model, individuals simulate users with limited mobility, and the robot is configured with different interaction modes. User feedback is collected through questionnaires, which helps evaluate and identify the strengths and weaknesses of each mode. Based on the gathered feedback, a novel handover interaction model is developed. To validate the effectiveness of the proposed model, a validation experiment is conducted. Seven participants engage with the handover interaction model, and their feedback is collected and analyzed. This interaction model fills a gap in the field of robot-to-human interactions and provides initial guidance for the development of related robotic technologies.

This research holds significant importance in the field of robotics by focusing on the development of automatic object grasping methods for robots and the interactive model of object handover between robots and humans. By addressing key research questions, this research aims to advance robotic capabilities in assisting individuals with limited mobility in retrieving objects and enabling interactions between humans and robots.

One of the primary objectives of this research is to understand the specific needs of individuals facing mobility challenges such as being bedridden or wheelchair-bound. By gaining insights into their requirements for object retrieval, this study guides the design of future robot applications tailored to their needs, enhancing their quality of life and promoting

independence. The study also proposes novel approaches to robot perception and grasp detection, addressing the limitations of existing methods that heavily rely on labor-intensive manual annotations. By exploring perception without extensive human-labeled 3D annotations and grasp detection without manual annotation data, this research accelerates the deployment of robots in human-robot interaction scenarios. Furthermore, the study focuses on developing an interactive model for object handover between humans and robots. The Peer Role model is proposed, treating humans and robots as peers, and integrating robot anticipation of human hand motions. This model aims to improve the effectiveness and naturalness of object handover interactions, enabling collaboration between humans and robots. Additionally, by incorporating foresight capabilities into the handover model, the research enhances the robot's anticipation and adaptability during interactions, resulting in more intuitive and efficient handover experiences.

The outcomes of this research have significant implications for designing and implementing future robot-to-human handover interactions. By identifying crucial factors and leveraging the developed techniques, this study contributes to the advancement of robotic systems that can collaborate with humans in a user-friendly manner, fostering robot adoption and acceptance in domains such as healthcare, assistive robotics, and daily life assistance.

References

Ackerman, E. (2015), Care-O-bot 4 Is the Robot Servant We All Want but Probably Can't Afford, {https://spectrum.ieee.org/care-o-bot-4-mobile-manipulator}

Aghapour, E., & Farrell, J. A. (2016, July). Human action prediction for human robot interaction. In 2016 American Control Conference (ACC) (pp. 5407-5412). IEEE. doi: 10.1109/ACC.2016.7526517.

Ardón, P., Cabrera, M. E., Pairet, E., Petrick, R. P., Ramamoorthy, S., Lohan, K. S., & Cakmak, M. (2021). Affordance-aware handovers with human arm mobility constraints. IEEE Robotics and Automation Letters, 6(2), 3136-3143. doi: 10.1109/LRA.2021.3062808

Bandi, C., & Thomas, U. (2021, December). Skeleton-based action recognition for human-robot interaction using self-attention mechanism. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1-8). IEEE. doi: 10.1109/FG52635.2021.9666948

Bao, J., Zhou, L., Liu, G., Tang, J., Lu, X., Cheng, C., ... & Bai, J. (2022). Current state of care for the elderly in China in the context of an aging population. Bioscience trends, 16(2), 107-118. https://doi.org/10.5582/bst.2022.01068

Bardaro, G., Antonini, A., & Motta, E. (2022). Robots for elderly care in the home: A landscape analysis and co-design toolkit. International Journal of Social Robotics, 14(3), 657-681. https://doi.org/10.1007/s12369-021-00816-3

Baek, S., Kim, K. I., & Kim, T. K. (2019). Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1067-1076).

Bansal, S., Tolani, V., Gupta, S., Malik, J., & Tomlin, C. (2020, May). Combining optimal control and learning for visual navigation in novel environments. In Conference on Robot Learning (pp. 420-429). PMLR.

Belardinelli, A., Kondapally, A. R., Ruiken, D., Tanneberg, D., & Watabe, T. (2022, October). Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 9806-9813). IEEE. doi: 10.1109/IROS47612.2022.9982249.

Boniol, M., Kunjumen, T., Nair, T. S., Siyam, A., Campbell, J., & Diallo, K. (2022). The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal'health coverage?. BMJ global health, 7(6), e009316. https://doi.org/10.1136/bmjgh-2022-009316

Cai, Y., Ge, L., Cai, J., & Yuan, J. (2018). Weakly-supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European conference on computer vision (ECCV) (pp. 666-682).

Cao, H., Dirnberger, L., Bernardini, D., Piazza, C., & Caccamo, M. (2023). 6IMPOSE: Bridging the reality gap in 6D pose estimation for robotic grasping. Frontiers in Robotics and AI, 10, 1176492. https://doi.org/10.3389/frobt.2023.1176492

Chen, D., Li, J., Wang, Z., & Xu, K. (2020a). Learning canonical shape space for category-level 6d object pose and size estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11973-11982).

Chen, L., Lin, S. Y., Xie, Y., Tang, H., Xue, Y., Lin, Y. Y., ... & Fan, W. (2019a). Tagan: Tonality-alignment generative adversarial networks for realistic hand pose synthesis. In BMVC.

Chen, K., Cao, R., James, S., Li, Y., Liu, Y. H., Abbeel, P., & Dou, Q. (2022a, October). Sim-to-real 6d object pose estimation via iterative self-training for robotic bin picking. In European Conference on Computer Vision (pp. 533-550). Cham: Springer Nature Switzerland.

Chen, L., Lin, S. Y., Xie, Y., Lin, Y. Y., & Xie, X. (2021b). Mvhm: A large-scale multi-view hand mesh benchmark for accurate 3d hand pose estimation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 836-845).

Chen, W., Jia, X., Chang, H. J., Duan, J., Shen, L., & Leonardis, A. (2021a). Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1581-1590).

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1907-1915).

Chen, X., Yang, J., He, Z., Yang, H., Zhao, Q., & Shi, Y. (2023). QwenGrasp: A Usage of Large Vision Language Model for Target-oriented Grasping. arXiv preprint arXiv:2309.16426. https://doi.org/10.48550/arXiv.2309.16426

Chen, X., Wang, G., Guo, H., & Zhang, C. (2020b). Pose guided structured region ensemble network for cascaded hand pose estimation. Neurocomputing, 395, 138-149. https://doi.org/10.1016/j.neucom.2018.06.097

Chen, Y., Tu, Z., Ge, L., Zhang, D., Chen, R., & Yuan, J. (2019b). So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6961-6970).

Chen, Y. N., Dai, H., & Ding, Y. (2022). Pseudo-stereo for monocular 3d object detection in autonomous driving. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 887-897).

Cheng, H., Wang, Y., & Meng, M. Q. H. (2023). Anchor-based multi-scale deep grasp pose detector with encoded angle regression. IEEE Transactions on Automation Science and Engineering, 21(3), pp. 3130-3142. doi: 10.1109/TASE.2023.3275771

Chu, F. J., Xu, R., & Vela, P. A. (2018). Real-world multiobject, multigrasp detection. IEEE Robotics and Automation Letters, 3(4), 3355-3362. doi: 10.1109/LRA.2018.2852777

Choi, A., Jawed, M. K., & Joo, J. (2022, May). Preemptive motion planning for human-to-robot indirect placement handovers. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 4743-4749). IEEE. doi: 10.1109/ICRA46639.2022.9811558.

Christensen, A. D., Lehotský, D., Jørgensen, M. W., & Chrysostomou, D. (2022, December). Learning to segment object affordances on synthetic data for task-oriented robotic handovers. In The 33rd British Machine Vision Conference. British Machine Vision Association.

Cordeiro, A., Rocha, L. F., Costa, C., Costa, P., & Silva, M. F. (2022, April). Bin picking approaches based on deep learning techniques: A state-of-the-art survey. In 2022 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) (pp. 110-117). IEEE. doi: 10.1109/ICARSC55462.2022.9784795.

Czaja, S. J., & Ceruso, M. (2022). The promise of artificial intelligence in supporting an aging population. Journal of Cognitive Engineering and Decision Making, 16(4), 182-193. https://doi.org/10.1177/155534342211299

De Magistris, G., Caprari, R., Castro, G., Russo, S., Iocchi, L., Nardi, D., & Napoli, C. (2021, December). Vision-based holistic scene understanding for context-aware human-robot interaction. In International Conference of the Italian Association for Artificial Intelligence (pp. 310-325). Cham: Springer International Publishing.

Del Duchetto, F., & Hanheide, M. (2022). Learning on the Job: Long-Term Behavioural Adaptation in Human-Robot Interactions. IEEE Robotics and Automation Letters, 7(3), 6934-6941. doi: 10.1109/LRA.2022.3178807

Deng, X., Mousavian, A., Xiang, Y., Xia, F., Bretl, T., & Fox, D. (2021). PoseRBPF: A Rao–Blackwellized particle filter for 6-D object pose tracking. IEEE Transactions on Robotics, 37(5), 1328-1342. doi: 10.1109/TRO.2021.3056043

Deng, X., Xiang, Y., Mousavian, A., Eppner, C., Bretl, T., & Fox, D. (2020, May). Self-supervised 6d object pose estimation for robot manipulation. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3665-3671). IEEE. doi: 10.1109/ICRA40945.2020.9196714

Do, T. T., Pham, T., Cai, M., & Reid, I. (2018). Real-time monocular object instance 6d pose estimation. In British

Machine Vision Conference 2018. British Machine Vision Association.

Durrani, H. (2016). Healthcare and healthcare systems: inspiring progress and future prospects. Mhealth, 2.

Du, G., Wang, K., Lian, S., & Zhao, K. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. Artificial Intelligence Review, 54(3), 1677-1734. https://doi.org/10.1007/s10462-020-09888-5

Elnour, M., Himeur, Y., Fadli, F., Mohammedsherif, H., Meskin, N., Ahmad, A. M., ... & Hodorog, A. (2022). Neural network-based model predictive control system for optimizing building automation and management systems of sports facilities. Applied Energy, 318, 119153. https://doi.org/10.1016/j.apenergy.2022.119153

Eppner, C., Mousavian, A., & Fox, D. (2021, May). Acronym: A large-scale grasp dataset based on simulation. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6222-6227). IEEE. doi: 10.1109/ICRA48506.2021.9560844

Famakin, I., & Leung, M. Y. (2018). A Comparison of Barrier-Free Access Designs for the Elderly Living in the Community and in Care and Attention Homes in Hong Kong. In Proceedings of the 21st International Symposium on Advancement of Construction Management and Real Estate (pp. 11-18). Springer Singapore. https://doi.org/10.1007/978-981-10-6190-5_2

Fan, L., Rao, H., & Yang, W. (2021). 3d hand pose estimation based on five-layer ensemble cnn. Sensors, 21(2), 649. https://doi.org/10.3390/s21020649

Faibish, T., Kshirsagar, A., Hoffman, G., & Edan, Y. (2022). Human preferences for robot eye gaze in human-to-robot handovers. International Journal of Social Robotics, 14(4), 995-1012. https://doi.org/10.1007/s12369-021-00836-z

Fazlali, H., Xu, Y., Ren, Y., & Liu, B. (2022). A versatile multi-view framework for lidar-based 3d object detection with guidance from panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17192-17201).

Fang, L., Liu, X., Liu, L., Xu, H., & Kang, W. (2020a). Jgr-p2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16 (pp. 120-137). Springer International Publishing. https://doi.org/10.1007/978-3-030-58539-6_8

Fang, H. S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., ... & Lu, C. (2023). Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. IEEE Transactions on Robotics, 39(5), (pp. 3929-3945). doi: 10.1109/TRO.2023.3281153

Fang, H. S., Wang, C., Gou, M., & Lu, C. (2020b). Graspnet-1billion: A large-scale benchmark for general object grasping. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11444-11453).

Furnari, A., & Farinella, G. M. (2020). Rolling-unrolling lstms for action anticipation from first-person video. IEEE transactions on pattern analysis and machine intelligence, 43(11), 4021-4036. doi: 10.1109/TPAMI.2020.2992889

Fu, Z., Zhao, T. Z., & Finn, C. (2024). Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv preprint arXiv:2401.02117. https://doi.org/10.48550/arXiv.2401.02117

Fu, J., Huang, Q., Doherty, K., Wang, Y., & Leonard, J. J. (2021, September). A multi-hypothesis approach to pose ambiguity in object-based slam. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 7639-7646). IEEE. doi: 10.1109/IROS51168.2021.9635956

Garage, W. (2008), PR2, https://rasc.usc.edu/robots/humanoid/pr2/

Gao, J., Yang, Z., & Nevatia, R. (2017). RED: Reinforced Encoder-Decoder Networks for Action Anticipation. In Procedings of the British Machine Vision Conference 2017. British Machine Vision Association.

Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., & Yuan, J. (2019). 3d hand shape and pose estimation from a single rgb image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10833-10842).

Ge, L., Cai, Y., Weng, J., & Yuan, J. (2018a). Hand pointnet: 3d hand pose estimation using point sets. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8417-8426).

Ge, L., Ren, Z., & Yuan, J. (2018b). Point-to-point regression pointnet for 3d hand pose estimation. In Proceedings of the European conference on computer vision (ECCV) (pp. 475-491).

Geiger, A., Lenz, P., & Urtasun, R. (2012, June). Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition (pp. 3354-3361). IEEE.

Gilles, M., Chen, Y., Winter, T. R., Zeng, E. Z., & Wong, A. (2022, August). Metagraspnet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis. In 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) (pp. 220-227). IEEE. doi: 10.1109/CASE49997.2022.9926427

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. Interaction studies, 8(3), 483-500. https://doi.org/10.1075/is.8.3.10gro

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019a). Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603. https://doi.org/10.48550/arXiv.1912.01603

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019b). Learning latent dynamics for planning from pixels. In International conference on machine learning (pp. 2555-2565). PMLR.

Hansen, N., Wang, X., & Su, H. (2022). Temporal difference learning for model predictive control. arXiv preprint arXiv:2203.04955. https://doi.org/10.48550/arXiv.2203.04955

He, T., & Soatto, S. (2019, July). Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 8409-8416). https://doi.org/10.1609/aaai.v33i01.33018409

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hewing, L., Wabersich, K. P., Menner, M., & Zeilinger, M. N. (2020). Learning-based model predictive control: Toward safe learning in control. Annual Review of Control, Robotics, and Autonomous Systems, 3, 269-296. https://doi.org/10.1146/annurev-control-090419-075625

Hoeller, D., Farshidian, F., & Hutter, M. (2020, May). Deep value model predictive control. In Conference on robot learning (pp. 990-1004). PMLR.

Holland, J., Kingston, L., McCarthy, C., Armstrong, E., O'Dwyer, P., Merz, F., & McConnell, M. (2021). Service robots in the healthcare sector. Robotics, 10(1), 47. https://doi.org/10.3390/robotics10010047

Hu, Y., Fang, S., Xie, W., & Chen, S. (2023). Aerial monocular 3d object detection. IEEE Robotics and Automation Letters, 8(4), 1959-1966. doi: 10.1109/LRA.2023.3245421

Hu, J. S., Kuai, T., & Waslander, S. L. (2022). Point density-aware voxels for lidar 3d object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8469-8478).

Hu, Y., Fua, P., Wang, W., & Salzmann, M. (2020). Single-stage 6d object pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2930-2939).

Hua, W., Zhou, Z., Wu, J., Huang, H., Wang, Y., & Xiong, R. (2021). Rede: End-to-end object 6d pose robust estimation using differentiable outliers elimination. IEEE Robotics and Automation Letters, 6(2), 2886-2893. doi: 10.1109/LRA.2021.3062304

Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., & Zhu, S. C. (2018). Holistic 3d scene parsing and reconstruction from a single rgb image. In Proceedings of the European conference on computer vision (ECCV) (pp. 187-203).

Huang, W., Ren, P., Wang, J., Qi, Q., & Sun, H. (2020, April). Awr: Adaptive weighting regression for 3d hand pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 11061-11068). https://doi.org/10.1609/aaai.v34i07.6761

Huang, S., Chen, Y., Yuan, T., Qi, S., Zhu, Y., & Zhu, S. C. (2019). Perspectivenet: 3d object detection from a single rgb image via perspective points. Advances in neural information processing systems, 32.

Huang, K. C., Wu, T. H., Su, H. T., & Hsu, W. H. (2022). Monodtr: Monocular 3d object detection with depth-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4012-4021).

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. The International Journal of Robotics Research, 40(4-5), 698-721. https://doi.org/10.1177/02783649209878

Iori, F., Perovic, G., Cini, F., Mazzeo, A., Falotico, E., & Controzzi, M. (2023). DMP-Based Reactive Robot-to-Human Handover in Perturbed Scenarios. International Journal of Social Robotics, 15(2), 233-248. https://doi.org/10.1007/s12369-022-00960-4

Iqbal, U., Molchanov, P., Gall, T. B. J., & Kautz, J. (2018). Hand pose estimation via latent 2.5 d heatmap regression. In Proceedings of the European conference on computer vision (ECCV) (pp. 118-134).

Izadinia, H., Shan, Q., & Seitz, S. M. (2017). Im2cad. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5134-5143).

Jia, F., Ma, Y., & Ahmad, R. (2024). Review of current vision-based robotic machine-tending applications. The International Journal of Advanced Manufacturing Technology, 131(3), 1039-1057. https://doi.org/10.1007/s00170-024-13168-9

Karg, B., & Lucia, S. (2020). Efficient representation and approximation of model predictive control laws via deep learning. IEEE Transactions on Cybernetics, 50(9), 3866-3878. doi: 10.1109/TCYB.2020.2999556

Ke, Q., Fritz, M., & Schiele, B. (2019). Time-conditioned action anticipation in one shot. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9925-9934).

Ke, Q., Fritz, M., & Schiele, B. (2021). Future moment assessment for action query. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 3219-3228).

Kennedy, S., & Johnson, C. K. (2016). Perfecting China, Inc.: China's 13th five-year plan. Rowman & Littlefield.

Kinova (2020), {https://www.kinovarobotics.com/product/gen2-robots}

Kleeberger, K., Bormann, R., Kraus, W., & Huber, M. F. (2020). A survey on learning-based robotic grasping. Current Robotics Reports, 1, 239-249. https://doi.org/10.1007/s43154-020-00021-6

Kroemer, O., Niekum, S., & Konidaris, G. (2021). A review of robot learning for manipulation: Challenges, representations, and algorithms. Journal of machine learning research, 22(30), 1-82.

Kshirsagar, A., Lim, M., Christian, S., & Hoffman, G. (2020). Robot gaze behaviors in human-to-robot handovers. IEEE Robotics and Automation Letters, 5(4), 6552-6558. doi: 10.1109/LRA.2020.3015692

Kumar, A., Brazil, G., Corona, E., Parchami, A., & Liu, X. (2022, October). Deviant: Depth equivariant network for monocular 3d object detection. In European Conference on Computer Vision (pp. 664-683). Cham: Springer Nature Switzerland.

Kwon, T., Tekin, B., Stühmer, J., Bogo, F., & Pollefeys, M. (2021). H2o: Two hands manipulating objects for first person interaction recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10138-10148).

Kwok, C. L., Lee, C. K., Lo, W. T., & Yip, P. S. (2017). The contribution of ageing to hospitalisation days in Hong Kong: a decomposition analysis. International Journal of Health Policy and Management, 6(3), pp.155-164. doi: 10.15171/ijhpm.2016.108

Lagomarsino, M., Lorenzini, M., Constable, M. D., De Momi, E., Becchio, C., & Ajoudani, A. (2023). Maximising Coefficiency of Human-Robot Handovers through Reinforcement Learning. IEEE Robotics and Automation Letters, 8(8), pp. 4378-4385. doi: 10.1109/LRA.2023.3280752

Lasota, P. A., Fong, T., & Shah, J. A. (2017). A survey of methods for safe human-robot interaction. Foundations and Trends® in Robotics, 5(4), 261-349. http://dx.doi.org/10.1561/2300000052

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12697-12705).

Labbé, Y., Carpentier, J., Aubry, M., & Sivic, J. (2020). Cosypose: Consistent multi-view multi-object 6d pose estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16 (pp. 574-591). Springer International Publishing.

Lehotsky, D., Christensen, A., & Chrysostomou, D. (2023, October). Optimizing Robot-to-Human Object Handovers using Vision-based Affordance Information. In 2023 IEEE International Conference on Imaging Systems and Techniques (IST) (pp. 1-6). IEEE. doi: 10.1109/IST59124.2023.10355704

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International journal of robotics research, 37(4-5), 421-436. https://doi.org/10.1177/0278364917710318

Lenz, I., Lee, H., & Saxena, A. (2015a). Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5), 705-724. https://doi.org/10.1177/0278364914549607

Lenz, I., Knepper, R. A., & Saxena, A. (2015b). DeepMPC: Learning deep latent features for model predictive control. In Robotics: Science and Systems (Vol. 10, p. 25).

Lee, T., Lee, B. U., Kim, M., & Kweon, I. S. (2021). Category-level metric scale object shape and pose estimation. IEEE Robotics and Automation Letters, 6(4), 8575-8582. doi: 10.1109/LRA.2021.3110538

Li, Y., Wang, G., Ji, X., Xiang, Y., & Fox, D. (2018). Deepim: Deep iterative matching for 6d pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 683-698).

Li, T., An, J., Yang, K., Chen, G., & Wang, Y. (2022, December). An Efficient Network for Target-oriented Robot Grasping Pose Generation in Clutter. In 2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA) (pp. 967-972). IEEE. doi: 10.1109/ICIEA54703.2022.10005947

Liang, H., Jiang, C., Feng, D., Chen, X., Xu, H., Liang, X., ... & Van Gool, L. (2021). Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3293-3302).

Lim, V., Rooksby, M., & Cross, E. S. (2021). Social robots on a global stage: establishing a role for culture during human–robot interaction. International Journal of Social Robotics, 13(6), 1307-1333. https://doi.org/10.1007/s12369-020-00710-4

Liang, H., Ma, X., Li, S., Görner, M., Tang, S., Fang, B., ... & Zhang, J. (2019, May). Pointnetgpd: Detecting grasp configurations from point sets. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 3629-3635). IEEE. doi: 10.1109/ICRA.2019.8794435

Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., & Li, Y. (2021). Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3560-3569).

Lin, H., Liu, Z., Cheang, C., Fu, Y., Guo, G., & Xue, X. (2022). Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6707-6717).

Liu, H., Wang, Y., Ji, W., & Wang, L. (2018). A context-aware safety system for human-robot collaboration. Procedia Manufacturing, 17, 238-245. https://doi.org/10.1016/j.promfg.2018.10.042

Liu, D., Wang, X., Cong, M., Du, Y., Zou, Q., & Zhang, X. (2021a). Object transfer point predicting based on human comfort model for human-robot handover. IEEE Transactions on Instrumentation and Measurement, 70, 1-11. doi: 10.1109/TIM.2021.3089227

Liu, H., & Wang, L. (2021b). Collision-free human-robot collaboration based on context awareness. Robotics and Computer-Integrated Manufacturing, 67, 101997. https://doi.org/10.1016/j.rcim.2020.101997

Liu, Y., Yixuan, Y., & Liu, M. (2021c). Ground-aware monocular 3d object detection for autonomous driving. IEEE Robotics and Automation Letters, 6(2), 919-926. doi: 10.1109/LRA.2021.3052442

Liu, Z., Wang, Z., Huang, S., Zhou, J., & Lu, J. (2022, October). Ge-grasp: Efficient target-oriented grasping in dense clutter. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1388-1395). IEEE. doi: 10.1109/IROS47612.2022.9981499

Lison, P., Ehrler, C., & Kruijff, G. J. M. (2010, September). Belief modelling for situation awareness in human-robot interaction. In 19th International Symposium in Robot and Human Interactive Communication (pp. 138-143). IEEE. doi: 10.1109/ROMAN.2010.5598723

Lockwood, K., Bicer, Y., Asghari-Esfeden, S., Zhu, T., Furmanek, M., Mangalam, M., ... & Tunik, E. (2022, June). Leveraging submovements for prediction and trajectory planning for human-robot handover. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments (pp. 247-253). https://doi.org/10.1145/3529190.3529220

Lucia, S., Navarro, D., Karg, B., Sarnago, H., & Lucia, O. (2020). Deep learning-based model predictive control for resonant power converters. IEEE Transactions on Industrial Informatics, 17(1), 409-420. doi: 10.1109/TII.2020.2969729

Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., ... & Xu, C. (2021). Voxel transformer for 3d object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3164-3173).

Malik, J., Abdelaziz, I., Elhayek, A., Shimada, S., Ali, S. A., Golyanik, V., ... & Stricker, D. (2020). Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7113-7122).

Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., ... & Goldberg, K. (2017). Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv:1703.09312. https://doi.org/10.48550/arXiv.1703.09312

Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., & Van Gool, L. (2021). Towards a weakly supervised framework for 3d point cloud object detection and annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8), 4454-4468. doi: 10.1109/TPAMI.2021.3063611

Meng, C., Zhang, T., & lun Lam, T. (2022, October). Fast and Comfortable Interactive Robot-to-Human Object Handover. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3701-3706). IEEE. doi: 10.1109/IROS47612.2022.9981484

Merrill, N., Guo, Y., Zuo, X., Huang, X., Leutenegger, S., Peng, X., ... & Huang, G. (2022). Symmetry and uncertainty-aware object slam for 6dof object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14901-14910).

Miao, J., & Wu, X. (2021). Subjective wellbeing of Chinese elderly: A comparative analysis among Hong Kong, Urban China and Taiwan. Ageing & Society, 41(3), 686-707. doi:10.1017/S0144686X19001272

Mitsunaga, N., Smith, C., Kanda, T., Ishiguro, H., & Hagita, N. (2008). Adapting robot behavior for human--robot interaction. IEEE Transactions on Robotics, 24(4), 911-916. doi: 10.1109/TRO.2008.926867

Misra, I., Girdhar, R., & Joulin, A. (2021). An end-to-end transformer model for 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2906-2917).

Moon, H. S., & Seo, J. (2021). Fast user adaptation for human motion prediction in physical human–robot interaction. IEEE Robotics and Automation Letters, 7(1), 120-127. doi: 10.1109/LRA.2021.3116319

Morrison, D., Corke, P., & Leitner, J. (2020). Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. IEEE Robotics and Automation Letters, 5(3), 4368-4375. doi: 10.1109/LRA.2020.2992195

Mousavian, A., Eppner, C., & Fox, D. (2019). 6-dof graspnet: Variational grasp generation for object manipulation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2901-2910).

Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., & Theobalt, C. (2018). Ganerated hands for real-time 3d hand tracking from monocular rgb. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 49-59).

Murali, A., Mousavian, A., Eppner, C., Paxton, C., & Fox, D. (2020, May). 6-dof grasping for target-driven object manipulation in clutter. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6232-6238). IEEE. doi: 10.1109/ICRA40945.2020.9197318

Nationen, V. (2022). World population prospects 2022: Summary of results. UN.

Newbury, R., Cosgun, A., Crowley-Davis, T., Chan, W. P., Drummond, T., & Croft, E. A. (2022, August). Visualizing robot intent for object handovers with augmented reality. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1264-1270). IEEE. doi: 10.1109/RO-MAN53752.2022.9900524

Nemlekar, H., Dutia, D., & Li, Z. (2019, May). Object transfer point estimation for fluent human-robot handovers. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 2627-2633). IEEE. doi: 10.1109/ICRA.2019.8794008

Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., & Zhang, J. J. (2020). Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 55-64).

Nikolaidis, S., Hsu, D., & Srinivasa, S. (2017). Human-robot mutual adaptation in collaborative tasks: Models and experiments. The International Journal of Robotics Research, 36(5-7), 618-634. https://doi.org/10.1177/0278364917690593

Norman, D., & Draper, S. W. (1986). User centered design: new perspectives on human-computer interaction. New Jersy: L.

Nowak, J., Fraisse, P., Cherubini, A., & Daures, J. P. (2022, May). Assistance to older adults with comfortable robot-to-human handovers. In 2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO) (pp. 1-6). IEEE. doi: 10.1109/ARSO54254.2022.9802960

Obaigbena, A., Lottu, O. A., Ugwuanyi, E. D., Jacks, B. S., Sodiya, E. O., & Daraojimba, O. D. (2024). AI and human-robot interaction: A review of recent advances and challenges. GSC Advanced Research and Reviews, 18(2), 321-330. https://doi.org/10.30574/gscarr.2024.18.2.0070

Ovur, S. E., & Demiris, Y. (2023). Naturalistic robot-to-human bimanual handover in complex environments through multi-sensor fusion. IEEE Transactions on Automation Science and Engineering. 21(3), pp. 3730-3741, doi: 10.1109/TASE.2023.3284668

Pan, X., Xia, Z., Song, S., Li, L. E., & Huang, G. (2021). 3d object detection with pointformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7463-7472).

Panteleris, P., Oikonomidis, I., & Argyros, A. (2018, March). Using a single rgb frame for real time 3d hand pose estimation in the wild. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 436-445). IEEE. doi: 10.1109/WACV.2018.00054

Park, D., Ambrus, R., Guizilini, V., Li, J., & Gaidon, A. (2021). Is pseudo-lidar needed for monocular 3d object

detection?. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3142-3152).

Perovic, G., Iori, F., Mazzeo, A., Controzzi, M., & Falotico, E. (2023). Adaptive Robot-Human Handovers with Preference Learning. IEEE Robotics and Automation Letters, 8(10), pp. 6331-6338, doi: 10.1109/LRA.2023.3306280

Pérez, J., Aguilar, J., & Dapena, E. (2020). MIHR: A human-robot interaction model. IEEE Latin America Transactions, 18(09), 1521-1529. doi: 10.1109/TLA.2020.9381793

Peng, L., Yan, S., Wu, B., Yang, Z., He, X., & Cai, D. (2022a). Weakm3d: Towards weakly supervised monocular 3d object detection. arXiv preprint arXiv:2203.08332. https://doi.org/10.48550/arXiv.2203.08332

Peng, L., Liu, F., Yu, Z., Yan, S., Deng, D., Yang, Z., ... & Cai, D. (2022b, October). Lidar point cloud guided monocular 3d object detection. In European conference on computer vision (pp. 123-139). Cham: Springer Nature Switzerland.

Poudel, R. P., Liwicki, S., & Cipolla, R. (2019). Fast-scnn: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502. https://doi.org/10.48550/arXiv.1902.04502

Pun, J. W., & Elliott, L. (2020). A systematic review to assess the impact of the Elderly Health Care Voucher Scheme (EHCVS) and the feasibility to fully adopt in Hong Kong elder care services.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 652-660).

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, 30.

Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum pointnets for 3d object detection from rgb-d data. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 918-927).

Qi, C. R., Litany, O., He, K., & Guibas, L. J. (2019). Deep hough voting for 3d object detection in point clouds. In proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9277-9286).

Qi, C. R., Chen, X., Litany, O., & Guibas, L. J. (2020). Imvotenet: Boosting 3d object detection in point clouds with image votes. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4404-4413).

Qin, M., Brawer, J., & Scassellati, B. (2022, January). Task-oriented robot-to-human handovers in collaborative tool-use tasks. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). https://doi.org/10.1109/RO-MAN53752.2022.9900599

Qin, Y., Chen, R., Zhu, H., Song, M., Xu, J., & Su, H. (2020, May). S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In Conference on robot learning (pp. 53-65). PMLR.

Quintas, J., Martins, G. S., Santos, L., Menezes, P., & Dias, J. (2018). Toward a context-aware human–robot interaction framework based on cognitive development. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(1), 227-237. doi: 10.1109/TSMC.2018.2833384

Ren, P., Sun, H., Huang, W., Hao, J., Cheng, D., Qi, Q., ... & Liao, J. (2021). Spatial-aware stacked regression network for real-time 3d hand pose estimation. Neurocomputing, 437, 42-57. https://doi.org/10.1016/j.neucom.2021.01.045

Ren, P., Sun, H., Hao, J., Wang, J., Qi, Q., & Liao, J. (2022). Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 20555-20565).

Reily, B., Han, F., Parker, L. E., & Zhang, H. (2018). Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction. Autonomous Robots, 42, 1281-1298. https://doi.org/10.1007/s10514-017-9692-3

Rosenberger, P., Cosgun, A., Newbury, R., Kwan, J., Ortenzi, V., Corke, P., & Grafinger, M. (2020). Object-independent human-to-robot handovers using real time robotic vision. IEEE Robotics and Automation Letters, 6(1), 17-23. doi: 10.1109/LRA.2020.3026970

Rukhovich, D., Vorontsova, A., & Konushin, A. (2022). Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2397-2406).

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). Human motion trajectory prediction: A survey. The International Journal of Robotics Research, 39(8), 895-935. https://doi.org/10.1177/0278364920917446

Ryoo, M. S. (2011, November). Human activity prediction: Early recognition of ongoing activities from streaming videos. In 2011 international conference on computer vision (pp. 1036-1043). IEEE. doi: 10.1109/ICCV.2011.6126349

Sahin, C., & Kim, T. K. (2018). Category-level 6d object pose recovery in depth images. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops (pp. 0-0).

Scholtz, J. (2003, January). Theory and evaluation of human robot interactions. In 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the (pp. 10-pp). IEEE. doi:

10.1109/HICSS.2003.1174284

Shao, L., Ferreira, F., Jorda, M., Nambiar, V., Luo, J., Solowjow, E., ... & Bohg, J. (2020). Unigrasp: Learning a unified model to grasp with multifingered robotic hands. IEEE Robotics and Automation Letters, 5(2), 2286-2293. doi: 10.1109/LRA.2020.2969946

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10529-10538).

Shi, S., Wang, X., & Li, H. (2019). Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 770-779).

Shi, Y., Huang, J., Xu, X., Zhang, Y., & Xu, K. (2021). Stablepose: Learning 6d object poses from geometrically stable patches. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15222-15231).

Sidiropoulos, A., Karayiannidis, Y., & Doulgeri, Z. (2021, May). Human-robot collaborative object transfer using human motion prediction based on cartesian pose dynamic movement primitives. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3758-3764). IEEE. doi: 10.1109/ICRA48506.2021.9562035

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1145-1153).

Simmering, J., Meyer zu Borgsen, S., Wachsmuth, S., & Al-Hamadi, A. (2019). Combining static and dynamic predictions of transfer points for human initiated handovers. In Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11 (pp. 676-686). Springer International Publishing. https://doi.org/10.1007/978-3-030-35888-4_63

SoftBank Robotics (2014), {https://us.softbankrobotics.com/pepper}

Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 567-576).

Spurr, A., Dahiya, A., Wang, X., Zhang, X., & Hilliges, O. (2021). Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 11230-11239).

State Council of the People's Republic of China. (2006). The national medium-and long-term program for science and technology development (2006–2020).

Su, Y., Di, Y., Zhai, G., Manhardt, F., Rambach, J., Busam, B., ... & Tombari, F. (2023). Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3d object detection. IEEE Robotics and Automation Letters, 8(3), 1327-1334. doi: 10.1109/LRA.2023.3238137

Sun, M., & Gao, Y. (2021). Gater: Learning grasp-action-target embeddings and relations for task-specific grasping. IEEE Robotics and Automation Letters, 7(1), 618-625. doi: 10.1109/LRA.2021.3131378

Sundermeyer, M., Mousavian, A., Triebel, R., & Fox, D. (2021, May). Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13438-13444). IEEE. doi: 10.1109/ICRA48506.2021.9561877

Surís, D., Liu, R., & Vondrick, C. (2021). Learning the predictability of the future. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12607-12617).

Tan, X., Zhang, Y., Shao, H. (2019). Healthy China 2030, a breakthrough for improving health. Glob Health Promot. 2019 Dec;26(4):96-99. doi: 10.1177/1757975917743533.

Tekin, B., Bogo, F., & Pollefeys, M. (2019). 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 (pp. 4511-4520).

Ten Pas, A., Gualtieri, M., Saenko, K., & Platt, R. (2017). Grasp pose detection in point clouds. The International Journal of Robotics Research, 36(13-14), 1455-1473. https://doi.org/10.1177/0278364917735594

Tian, M., Ang, M. H., & Lee, G. H. (2020). Shape prior deformation for categorical 6d object pose and size estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16 (pp. 530-546). Springer International Publishing.

Tong, T., Setchi, R., & Hicks, Y. (2022). Context change and triggers for human intention recognition. Procedia Computer Science, 207, 3826-3835. https://doi.org/10.1016/j.procs.2022.09.444

Trick, S., Koert, D., Peters, J., & Rothkopf, C. A. (2019, November). Multimodal uncertainty reduction for intention recognition in human-robot interaction. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 7009-7016). IEEE. doi: 10.1109/IROS40897.2019.8968171

Tröbinger, M., Jähne, C., Qu, Z., Elsner, J., Reindl, A., Getz, S., ... & Haddadin, S. (2021). Introducing garmi-a service robotics platform to support the elderly at home: Design philosophy, system overview and first results. IEEE

Robotics and Automation Letters, 6(3), 5857-5864. doi: 10.1109/LRA.2021.3082012

Umbrico, A., Cesta, A., Cortellessa, G., & Orlandini, A. (2020). A holistic approach to behavior adaptation for socially assistive robots. International Journal of Social Robotics, 12(3), 617-637. https://doi.org/10.1007/s12369-019-00617-9

Van Zoelen, E. M., Van Den Bosch, K., & Neerincx, M. (2021). Becoming team members: Identifying interaction patterns of mutual adaptation for human-robot co-learning. Frontiers in Robotics and AI, 8, 692811. https://doi.org/10.3389/frobt.2021.692811

Vianello, L., Mouret, J. B., Dalin, E., Aubry, A., & Ivaldi, S. (2021). Human posture prediction during physical human-robot interaction. IEEE Robotics and Automation Letters, 6(3), 6046-6053. doi: 10.1109/LRA.2021.3086666

Vuong, A. D., Vu, M. N., Le, H., Huang, B., Huynh, B., Vo, T., ... & Nguyen, A. (2023). Grasp-anything: Large-scale grasp dataset from foundation models. arXiv preprint arXiv:2309.09818. https://doi.org/10.48550/arXiv.2309.09818

Wan, C., Probst, T., Van Gool, L., & Yao, A. (2018). Dense 3d regression for hand pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5147-5156).

Wang, W., Li, R., Diekel, Z. M., Chen, Y., Zhang, Z., & Jia, Y. (2018). Controlling object hand-over in human–robot collaboration via natural wearable sensing. IEEE Transactions on Human-Machine Systems, 49(1), 59-71. doi: 10.1109/THMS.2018.2883176

Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., & Savarese, S. (2019a). Densefusion: 6d object pose estimation by iterative dense fusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3343-3352).

Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., & Guibas, L. J. (2019b). Normalized object coordinate space for category-level 6d object pose and size estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2642-2651).

Wang, C., Fang, H. S., Gou, M., Fang, H., Gao, J., & Lu, C. (2021). Graspness discovery in clutters for fast and accurate grasp detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15964-15973).

Wu, P., Escontrela, A., Hafner, D., Abbeel, P., & Goldberg, K. (2023, March). Daydreamer: World models for physical robot learning. In Conference on Robot Learning (pp. 2226-2240). PMLR.

Wübbeke, J., Meissner, M., Zenglein, M. J., Ives, J., & Conrad, B. (2016). Made in china 2025. Mercator Institute for China Studies. Papers on China, 2(74), 4.

Wright, J. (2021). Comparing public funding approaches to the development and commercialization of care robots in the European Union and Japan. Innovation: The European Journal of Social Science Research, 1-16. https://doi.org/10.1080/13511610.2021.1909460

Xiang, Y., Schmidt, T., Narayanan, V., & Fox, D. (2017). Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199. https://doi.org/10.48550/arXiv.1711.00199

Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou, J. T., & Yuan, J. (2019). A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 793-802).

Xie, C., Xiang, Y., Mousavian, A., & Fox, D. (2020, May). The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation. In Conference on robot learning (pp. 1369-1378). PMLR.

Xu, K., Zhao, S., Zhou, Z., Li, Z., Pi, H., Zhu, Y., ... & Xiong, R. (2023, May). A joint modeling of vision-language-action for target-oriented grasping in clutter. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11597-11604). IEEE. doi: 10.1109/ICRA48891.2023.10161041

Xu, X., Wang, Y., Zheng, Y., Rao, Y., Zhou, J., & Lu, J. (2022). Back to reality: Weakly-supervised 3d object detection with shape-guided label enhancement. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8438-8447).

Yang, L., & Yao, A. (2019). Disentangling latent hands for image synthesis and pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9877-9886).

Yang, W., Paxton, C., Cakmak, M., & Fox, D. (2020a, October). Human grasp classification for reactive human-to-robot handovers. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 11123-11130). IEEE. doi: 10.1109/IROS45743.2020.9341004

Yang, J., Chang, H. J., Lee, S., & Kwak, N. (2020b). Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16 (pp. 122-139). Springer International Publishing.

Yang, S., Wang, D., Li, W., Wang, C., Yang, X., & Lo, K. (2021a, October). Decoupling of elderly healthcare demand and expenditure in China. In Healthcare (Vol. 9, No. 10, p. 1346). MDPI. https://doi.org/10.3390/healthcare9101346

Yang, W., Paxton, C., Mousavian, A., Chao, Y. W., Cakmak, M., & Fox, D. (2021b, May). Reactive human-to-robot handovers of arbitrary objects. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3118-3124). IEEE. doi: 10.1109/ICRA48506.2021.9561170

Yang, Y., Lin, L., Zhang, Y., Zhou, Z., Wang, Y., & Xiong, R. (2021c, December). A Human-Robot Collaboration System for Object Handover. In 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO) (pp. 358-363). IEEE. doi: 10.1109/ROBIO54168.2021.9739333

Yasar, M. S., & Iqbal, T. (2021). A scalable approach to predict multi-agent motion for human-robot collaboration. IEEE Robotics and Automation Letters, 6(2), 1686-1693. doi: 10.1109/LRA.2021.3058917

You, Y., Ye, Z., Lou, Y., Li, C., Li, Y. L., Ma, L., ... & Lu, C. (2022). Canonical voting: Towards robust oriented bounding box detection in 3d scenes. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1193-1202).

Yu, X., Xu, C., Zhang, X., & Ou, L. (2022). Real-time multitask multihuman–robot interaction based on context awareness. Robotica, 40(9), 2969-2995. doi:10.1017/S0263574722000017

Yu, H., Lou, X., Yang, Y., & Choi, C. (2023a, October). IOSG: Image-Driven Object Searching and Grasping. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3145-3152). IEEE. doi: 10.1109/IROS55552.2023.10342009

Yu, S., Zhai, D. H., & Xia, Y. (2023b). Robotic Grasp Detection Based on Category-Level Object Pose Estimation With Self-Supervised Learning. IEEE/ASME Transactions on Mechatronics. 29(1), pp. 625-635. doi: 10.1109/TMECH.2023.3287635

Yuan, S., Stenger, B., & Kim, T. K. (2019). 3D hand pose estimation from RGB using privileged learning with depth data. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

Zallio, M., & Ohashi, T. (2022). The evolution of assistive technology: a literature review of technology developments and applications. Human Factors in Accessibility and Assistive Technology, 37, 85.

Zhang, H., Lan, X., Bai, S., Zhou, X., Tian, Z., & Zheng, N. (2019, November). Roi-based robotic grasp detection for object overlapping scenes. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4768-4775). IEEE. doi: 10.1109/IROS40897.2019.8967869

Zhang, Y., Müller, S., Stephan, B., Gross, H. M., & Notni, G. (2021a). Point cloud hand–object segmentation using multimodal imaging with thermal and color data for safe robotic object handover. Sensors, 21(16), 5676. https://doi.org/10.3390/s21165676

Zhang, H., Liang, Z., Li, C., Zhong, H., Liu, L., Zhao, C., ... & Wu, Q. J. (2021b). A practical robotic grasping method by using 6-D pose estimation with protective correction. IEEE Transactions on Industrial Electronics, 69(4), 3876-3886. doi: 10.1109/TIE.2021.3075836

Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., & Liu, S. (2021c). Holistic 3d scene understanding from a single image with implicit representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8833-8842).

Zhang, Y., Chen, J., & Huang, D. (2022a). Cat-det: Contrastively augmented transformer for multi-modal 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 908-917).

Zhang, Z., Ji, Y., Cui, W., Wang, Y., Li, H., Zhao, X., ... & Pu, S. (2022b). Atf-3d: Semi-supervised 3d object detection with adaptive thresholds filtering based on confidence and distance. IEEE Robotics and Automation Letters, 7(4), 10573-10580. doi: 10.1109/LRA.2022.3187496

Zhao, N., Chua, T. S., & Lee, G. H. (2020). Sess: Self-ensembling semi-supervised 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11079-11087).

Zhao, B., Zhang, H., Lan, X., Wang, H., Tian, Z., & Zheng, N. (2021, May). Regnet: Region-based grasp network for end-to-end grasp detection in point clouds. In 2021 IEEE international conference on robotics and automation (ICRA) (pp. 13474-13480). IEEE. doi: 10.1109/ICRA48506.2021.9561920

Zheng, L., Ma, W., Cai, Y., Lu, T., & Wang, S. (2023). GPDAN: Grasp Pose Domain Adaptation Network for Sim-to-Real 6-DoF Object Grasping. IEEE Robotics and Automation Letters. 8(8), pp. 4585-4592. doi: 10.1109/LRA.2023.3286816

Zhou, Z., Du, L., Ye, X., Zou, Z., Tan, X., Zhang, L., ... & Feng, J. (2022). SGM3D: Stereo guided monocular 3D object detection. IEEE Robotics and Automation Letters, 7(4), 10478-10485. doi: 10.1109/LRA.2022.3191849

Zimmermann, C., & Brox, T. (2017). Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE international conference on computer vision (pp. 4903-4911).

Appendix A.

The Grasp Pose Generation Algorithm of Curved Objects

For curved objects similar to bananas, the grasp prior knowledge can be summarized as follows: (1) The grasp point lies along the curved path of the object, (2) the grasp depth direction is perpendicular to the plane containing the curve, (3) the grasp width direction is along the normal line at the grasp point, and (4) the grasp width and depth are determined by the dimensions of the 3D bounding box. I utilize a quadratic curve to approximate the shape of the object's curve, as shown in Figure A1(a). This quadratic curve can be derived based on the distribution of the point cloud and the 3D bounding box. I mathematically express the curve as:

$$ST_3 \leftarrow \left\{\begin{cases} z = a(x - x_0)^2 + bx + z \\ y = y_{max} \end{cases}, \quad \begin{cases} x = a(x - x_0)^2 + bx + c \\ y = y_{min} \end{cases}\right\}.$$

Algorithm 6 depicts the algorithm for generating grasp poses for curved objects. Figure A1(b) demonstrates the generated poses.

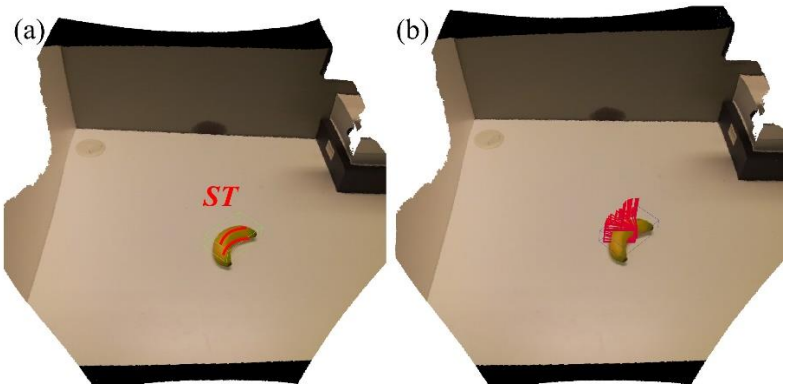


Figure A1. Grasp poses generation for curved objects. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

Algorithm 6 Generate grasp poses for curved objects

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
Derive a, b, c in ST_3 using the distribution of points.
for st in ST_3 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    $X = [p.x,\ CT.y,\ p.z] - [p.x,\ p.y,\ p.z]$
    Transpose and normalize X
    $Y = [-st'(p),\ 0,\ 1]$
    Transpose and normalize Y
    $Z = X \times Y$
    Grasp width $w = mid(x\_length, y\_length, z\_length)$
    Grasp depth $d = 2/3 y\_length$
    $R = [X, Y, Z]$
    $T = p$
    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
  end for
end for
Transfer g to the robot coordinate system.
return $g,\ W,\ D$

The Grasp Pose Generation Algorithm of Containers

In the context of circular containers, such as bowls, the approach incorporates the following grasp prior knowledge: (1) The distribution of grasp points aligns with the circular rim of the container, (2) the grasp depth direction is perpendicular to the plane defined by the container

rim, (3) the grasp width direction is perpendicular to the container wall, and (4) the grasp depth and width are contingent upon the dimensions of the object. Figure A2(a) visually represents the trajectory of sampled grasp points, designated as:

$$ST_4 \leftarrow \left\{ \begin{cases} (x - CT.x)^2 + (z - CT.z)^2 = R^2 \\ y = y_{min} \end{cases} \right\}.$$

Algorithm 7 depicts the algorithm for generating grasp poses for containers. Figure A2(b) demonstrates the generated poses.

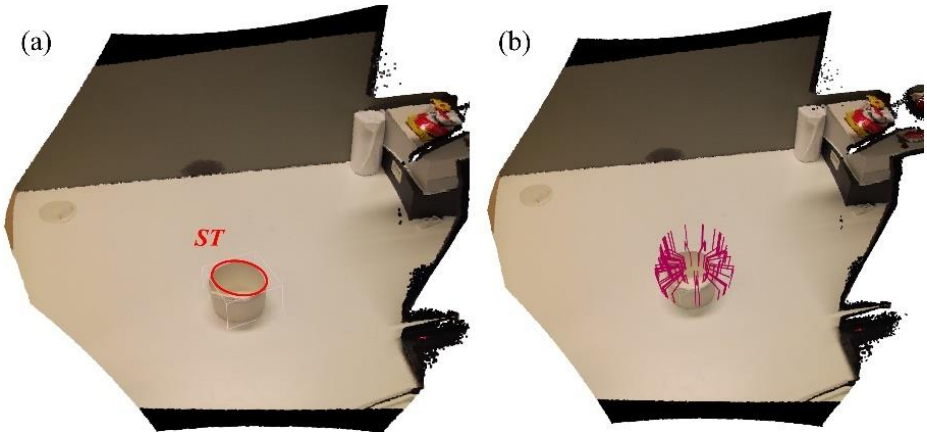


Figure A2. Grasp poses generation for containers. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

Algorithm 7 Generate grasp poses for containers

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
for st in ST_4 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    $X = [p.x, CT.y, p.z] - [p.x, p.y, p.z]$
    Transpose and normalize X
    $Y = [CT.x,\ p.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
    Transpose and normalize Y
    $Z = X \times Y$
    Grasp width $w = 0.5min(x\_length, y\_length, z\_length)$
    Grasp depth $d = 0.5y\_length$
    $R = [X, Y, Z]$
    $T = p$
    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
  end for
end for
Transfer g to the robot coordinate system.
return $g,\ W,\ D$

The Grasp Pose Generation Algorithm of Tools

In the case of tools, specifically screws, the derived grasp prior knowledge can be summarized as: (1) The distribution of grasp points predominantly occurs along the handle portion, (2) the grasp depth direction is oriented perpendicular to the sampling point, (3) the grasp width direction aligns orthogonally with the length of the object, and (4) the grasp depth and width are determined based on the dimensions of the object's 3D bounding box. Illustratively, Figure A3(a) depicts the trajectory of sampled points, denoted as:

$$ST_5 \leftarrow \left\{ \begin{matrix} (x - CT.x)^2 + (z - CT.z)^2 = R^2 \\ y = t, t \in y_{range} \end{matrix} \right\}$$

Algorithm 8 depicts the algorithm for generating grasp poses for tools. Figure A3(b) demonstrates the generated poses.

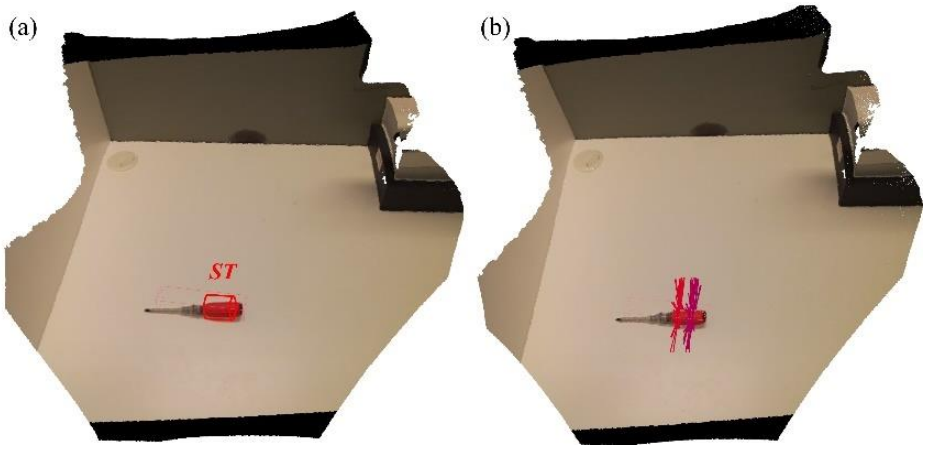


Figure A3. Grasp poses generation for tools. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

Algorithm 8 Generate grasp poses for tools

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
Derive y_range through the distribution of points
for st in ST_5 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    $X = [CT.x,\ p.y,\ CT.z] - [p.x,\ p.y,\ p.z]$
    Transpose and normalize X
    $Y = X \times [0,\ 1,\ 0]$.T
    $Z = X \times Y$
    Grasp width $w = 2R$
    Grasp depth $d = R$
    $R = [X, Y, Z]$
    $T = p$
    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
    Rotate R along Y with a random angle in $[-1/4\pi, 1/4\pi]$

    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
  end for
end for
Transfer g to the robot coordinate system.
return $g,\ W,\ D$

The Grasp Pose Generation Algorithm of Ring Objects

In the context of ring objects, exemplified by adhesive tape, the derived grasp prior knowledge can be summarized as follows: (1) The distribution of grasp points primarily lies along the circular ring of the object, (2) the grasp depth direction is oriented perpendicular to the plane encompassing the circular ring, (3) the grasp width direction aligns orthogonally with the side surface of the circular object, and (4) the grasp depth and width are determined based on the dimensions of the 3D bounding box. Illustratively, Figure A4(a) depicts the trajectory of sampled points, denoted as:

$$ST_6 \leftarrow \left\{\begin{cases}(x-CT.x)^2+(z-CT.z)^2=R^2\\ y=y_{min}\end{cases}, \quad \begin{cases}(x-CT.x)^2+(z-CT.z)^2=R^2\\ y=y_{max}\end{cases}\right\}$$

Algorithm 9 depicts the algorithm for generating grasp poses for ring objects. Figure A4(b) demonstrates the generated poses.

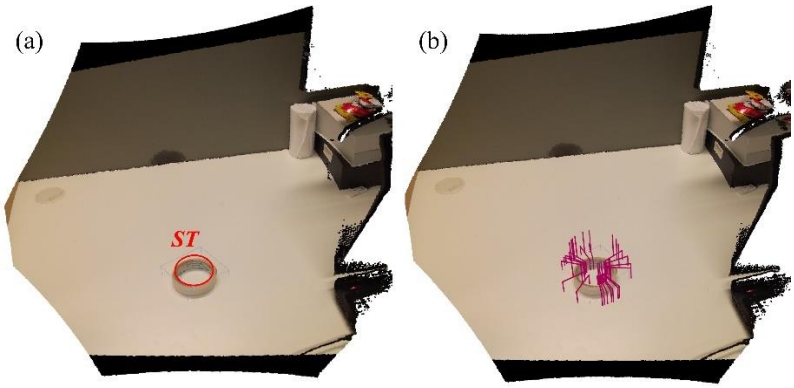


Figure 4A. Grasp poses generation for ring objects. (a) Sampling trajectory (ST) for grasp points. (b) Generated grasp poses without filtering.

Algorithm 9 Generate grasp poses for tapes

Require: 3D bounding box B, point cloud of object PC, empty buffer g, empty buffer W, empty buffer D, sampling quantity N, maximum gripper depth gd.

Require: $x\_length, y\_length, z\_length$ of B. x_min, x_max, y_min, y_max, z_min, z_max are the minimum and maximum values of B in the coordinate system of the object.

$CT.x = (x\_min + x\_max)/2$
$CT.y = (y\_min + y\_max)/2$
$CT.z = (z\_min + z\_max)/2$
for st in ST_6 **do**
  // generate grasp poses
  for $i \leftarrow 0$ to N **do**
    Randomly sample a point p on st
    $X = [p.x, CT.y, p.z] - [p.x, p.y, p.z]$
    Transpose and normalize X
    $Y = [CT.x, p.y, CT.z] - [p.x, p.y.p.z]$
    Transpose and normalize Y
    $Z = X \times Y$
    Grasp width $w = 0.5min(x\_length, y\_length, z\_length)$
    Grasp depth $d = 0.5y\_length$
    $R = [X, Y, Z]$
    $T = p$
    $g \leftarrow g \bigcup \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix}$ $W \leftarrow W \bigcup w$ $D \leftarrow D \bigcup d$
  end for
end for
Transfer g to the robot coordinate system.
return g, W, D

Appendix B.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

To	Luximon Yan (School of Design)		
From	Ng Po Ling, Delegate, Departmental Research Committee		
Email	bobo-pl.ng@	Date	26-Oct-2021

Application for Ethical Review for Teaching/Research Involving Human Subjects

I write to inform you that approval has been given to your application for human subjects ethics review of the following project for a period from 01-Sep-2021 to 01-Sep-2024:

Project Title:	Human-Robot Interaction Reinforcement Learning based on Facial Emotion Recognition
Department:	School of Design
Principal Investigator:	Luximon Yan
Project Start Date:	01-Sep-2021
Project type:	Human subjects (non-clinical)
Reference Number:	HSEARS20211015005

You will be held responsible for the ethical approval granted for the project and the ethical conduct of the personnel involved in the project. In case the Co-PI, if any, has also obtained ethical approval for the project, the Co-PI will also assume the responsibility in respect of the ethical approval (in relation to the areas of expertise of respective Co-PI in accordance with the stipulations given by the approving authority).

You are responsible for informing the PolyU Institutional Review Board in advance of any changes in the proposal or procedures which may affect the validity of this ethical approval.

Ng Po Ling

Delegate

Departmental Research Committee (on behalf of PolyU Institutional Review Board)