Learning Human-to-Robot Dexterous Handovers for Anthropomorphic Hand

Haonan Duan, Peng Wang, *Member, IEEE*, Yiming Li, Daheng Li, Wei Wei

***Abstract*—Human-robot interaction plays an important role in robots serving human production and life. Object handover between humans and robotics is one of the fundamental problems of human-robot interaction. The majority of current work uses parallel-jaw grippers as the end-effector device, which limits the ability of the robot to grab miscellaneous objects from human and manipulate them subsequently. In this paper, we present a framework for human-to-robot dexterous handover using an anthropomorphic hand. The framework takes images captured by two cameras to complete handover scene understanding, grasp configurations prediction, and handover execution. To enable the robot to generalize to diverse delivered objects with miscellaneous shapes and sizes, we propose an anthropomorphic hand grasp network (AHG-Net), an end-to-end network that takes the single-view point clouds of the object as input and predicts the suitable anthropomorphic hand configurations with 5 different grasp taxonomies. To train our model, we build a large-scale dataset with 1M hand grasp annotations from 5K single-view point clouds of 200 objects. We implement a handover system using a UR5 robot arm and HIT-DLR II anthropomorphic robot hand based on our presented framework, which can not only adapt to different human givers but generalize to diverse novel objects with various shapes and sizes. The generalizability, reliability, and robustness of our method are demonstrated on 15 different novel objects with arbitrary handover poses from frontal and lateral positions, a system ablation study, a grasp planner comparison, and a user study on 6 participants delivering 15 objects from two benchmark sets.**

***Index Terms*—Handovers, anthropomorphic hand, human-robot interaction**

I. Introduction

GRASPING objects delivered by human givers is one of the most fundamental abilities of robots during human-robot interaction. Unlike robot-to-human handover just passes an object to a human, accomplishing human-to-robot handover can not only assist workers or disabled people in moving delivered objects but transfer those repetitive, low-skill, or ergonomically unfavorable tasks to robots [1]. The implementation of safe, fluent, and intelligent human-to-robot handovers needs to address several challenging problems, including effective grasp planning, reliable perception, and efficient motion planning and control [2].

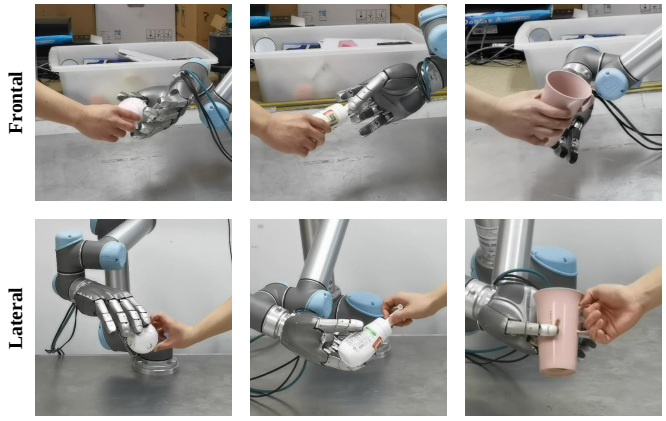


Fig. 1: Our handover system can adapt to numerous novel objects with diverse shapes and sizes that are delivered by human givers with arbitrary poses from different positions.

How to realize effective grasp planning that can generalize to a variety of objects with miscellaneous shapes and sizes delivered by human givers is one of the most crucial challenges of robust handover system design. One needs to consider the arbitrary positions and orientations of objects, as well as the unpredictable occlusions on objects by human hands during handover. Such reactive handover systems proposed by recent work [3]–[6] all utilize parallel-jaw gripper as their end-effectors. However, the study of the handover system using anthropomorphic hand remains a challenge.

The research on handover system for multi-finger hands from previous literature can be classified into motion-focused or grasp-focused. Motion-focused systems aim to obtain an adaptive motion planning method that is able to generate a fluent and natural trajectory during handover. [7] collect a database which is used for motion search during handover based on continuous observation. The method has low generalizability due to high dependence on the samples in the database. [8]–[10] build a handover scenario to validate their system based on Dynamic Movement Primitives [11]. Similarly, [12] present collaborative and assistive robots system by taking the Probabilistic Movement Primitives [13] for motion planning. [14] propose an online trajectory generator that enables robot to start motions as soon as the human giver delivers the objects. Although the methods mentioned

This work was supported in part by the National Natural Science Foundation of China under Grants (91748131, 62006229 and 61771471), in part by the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050106, and in part by the InnoHK Project. (*Corresponding author: Peng Wang.*)

Peng Wang is with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and with the CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong 999077, China (email: peng_wang@ia.ac.cn).

Haonan Duan, Yiming Li, Daheng Li, Wei Wei are with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

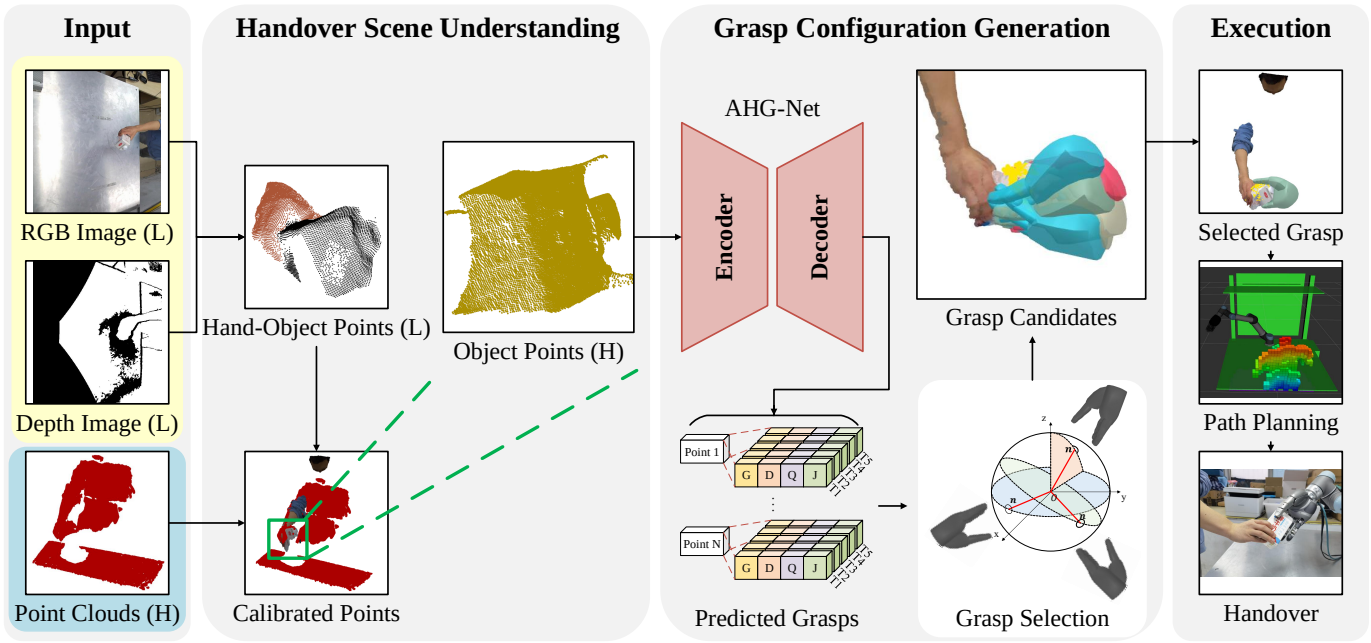


Fig. 2: The overview of our proposed handover framework. Given the RGB image and depth image captured by Azure Kinect (low-precision camera, denoted by L in figure) and point clouds captured by Ensenso N35 (high-precision camera, denoted by H in figure), it first obtains hand-object points by utilizing perceptual information from low-precision camera. Precise object points are then extracted with the help of hand-object points. AHG-Net takes visual processing results as input and predicts diverse grasps which are selected by approach directions to get grasp candidates. During execution, the candidates are selected for final handover completion.

above have addressed the safe, fluent, and natural motion planning problem, the system is incapable to generalize to numerous objects with different shapes and sizes, which limits the handover system application scenarios.

To promote the generalizability to diverse objects, grasp-focused systems pay attention to improving the system robustness by various types of perceptions. [15] utilize both RGB images and depth images captured by an RGB-D camera to detect humans and extract object clusters. Their method can generalize to several objects with different geometries, but the end-effector simply completes the opening and closing action with the help of force and time as the indicator of whether grasp the object successfully or not, which cannot bring out the dexterity of the multi-finger hand. [16] employ the information from the wrist force/torque sensor to enhance the generalization ability. The system performs well, however, tactile sensors are hard to equip or expensive in many cases.

In this paper, we present a framework for robot-to-human dexterous handover using an anthropomorphic hand (Fig. 2). The framework takes visual information captured by two cameras to complete handover scene understanding, grasp configurations generation, and handover execution. To enable the robot to generalize to diverse delivered objects with miscellaneous shapes and sizes, we propose an anthropomorphic hand grasp network (AHG-Net), an end-to-end network that takes the single-view point clouds of the object as input and predicts the suitable grasp configurations with 5 different grasp taxonomies. Unlike requiring complete geometric information of the object by previous work [17]–[19], we adopt the enormous progress of grasp proposals prediction for object single-view observation made by previous arts [20], [21]. We build a large-scale dataset containing 5K single-view point clouds of selected objects with over 1M hand grasp annotations to train our model. A handover system is implemented based on the presented framework by utilizing a UR5 robot arm and HIT-DLR II anthropomorphic hand.

We demonstrate the generalizability and robustness of our system with: (1) 15 novel objects handover with arbitrary poses from frontal and lateral positions; (2) system ablation study from input, prediction, and planning three aspects; (3) comparison of other grasp prediction networks; (4) user study on 6 participants with two object sets. The experimental results illustrate our handover system can adapt to different users and generalize to diverse novel objects.

In summary, our key contributions are:

- Present a framework for robot-to-human dexterous handover using anthropomorphic hand, which can robustly understand handover scene, generate reasonable grasp candidates and execute safe handover by taking captured perceptual information as input.
- Propose anthropomorphic hand grasp network (AHG-Net), an end-to-end network predicts precise grasp configuration for each taxonomy efficiently by taking the partial point cloud of a single object as input. A large-scale dataset with 200 household objects, 5K partial point clouds of a single object from cameras set at multiple angles, and over 1M anthropomorphic hand grasp annotations with selected five taxonomies.
- Implement the human-to-robot handover system that enables robot dexterously complete unknown object handovers by anthropomorphic hand. The generalizability, reliability, and robustness of our system are demonstrated

on experiments in simulation and real robot.

The paper is organized as follows. Related literature review is in Sec. II. The problem statement and handover framework are described in Sec. III. Sec. IV introduce our proposed anthropomorphic hand grasp network. The handover system design and implementation are illustrated in Sec. V. Sec. VI details how we generate our dataset. Sec. VII conducts the experiments and user study on the proposed system. Conclusion and future work are summarized in Sec. VIII.

II. Related Work

Grasp planner. The methods of generating grasp configurations on objects can be categorized into model-based and learning-based. Model-based approaches analyze feasible grasps based on the assumption of being aware of object full geometric information [22]–[24]. However, objects are usually not fully observed in real-world application scenarios, especially human hand occlusions during handover processing. Moreover, such methods are time-consuming due to searching-based algorithms to guarantee high-quality grasps. These dilemmas make model-based methods impossible to be deployed in the handover system to adapt to an unstructured environment and diverse objects. To this end, recent work adopts learning-based methods that predict grasps directly from partially observed objects by utilizing data-driven approaches, which makes tremendous progress in the robot community. [20], [21], [25]–[27] predict 6/7-DOF grasp configurations immediately on single-view RGB-D images or point clouds. These methods have robust generalization ability to environments and objects.

End-effector. Most researches mainly focus on robot grasp [20], [25], [26] or human-to-robot handover [3]–[5] of parallel-jaw gripper. Although such methods gain satisfactory performance, simple end-effectors incur the following predicaments: (1) they are only suitable for pick-place tasks; (2) they may grasp objects unstably because the feasible grasp areas are usually eccentric during handover, especially for objects with large shapes and sizes; (3) anthropomorphic hand can perform natural grasps on household objects which are designed based on the human hand. Currently, anthropomorphic hand interacts with objects attracts much interest [17], [19], [21], [28]–[31]. [19], [28] propose multi-finger hand grasp dataset, [29], [30] propose the methods by taking fully observed objects as input, [17], [21], [31] take single-view input and directly predict hand grasp poses.

Handover System. Human-robot interaction has been a primary concern within the robotics community, and how to enable robots to assist and even cooperate with humans in various production and life scenarios have always attracted researchers. Handover [2], human-robot collaboration [32] and wearable exoskeletons [33] all have the potential to help further improve the quality of human production and life. As one of the most fundamental capabilities in human-robot interaction, human-robot handover has evolved significantly over the past few decades. Many works improve the capabilities of handover systems from motion planning reliability to grasp generalizability. [7], [12], [14] address the problem that generates robot motion to accomplish handover fluently. [8]–[10] setup an application scenario to validate their motion plan methods. [15], [16] pay their attention to utilizing perception information to make the system generalize to diverse objects. Such methods obtain a desirable performance on some specific objects or environments but still cannot generalize to real-world handover scenarios. [34] propose a learning-based method that classifies how humans hold the object and generate the corresponding grasp to finish handover. The method attempts to enable the robot to adapt to several cases, however, it cannot adapt to numerous objects with different shapes and sizes. [3]–[5] present a reactive vision-based handover system that is able to grasp several objects during handover.

Most closely related to our method is the work by [2], [3], [5]. They develop a vision-based handover system by utilizing a parallel-jaw gripper as the robot end-effector. Their approaches take RGB images and depth images as input and extract object clusters to generate feasible grasps on them. Compared with theirs, our system has the following significant differences: (1) our system utilizes an anthropomorphic robot hand as the end-effector, which can receive delivered objects dexterously and naturally with a variety of grasp configurations; (2) we propose AHG-Net by taking single-view object points as input and predict the precise grasps, which allows the system to adapt to novel objects with miscellaneous shapes and sizes; (3) our system permits users using any handover poses from different positions with no constraints; (4) our system detects hand and object simultaneously and applies skin segmentation to fast extract object points.

III. Framework and Problem Statement

This work aims to complete human-to-robot handover for unknown objects based on point clouds from single-view observation. As illustrated in Fig. 2, our framework is composed of four parts. It takes RGB and depth image captured by a low-precision camera and point cloud captured by a high-precision camera as input, then detects and segments hand and object to extract the precise object points in the handover scene understanding part for anthropomorphic hand grasp network to generate the grasp candidates. The robot will execute the handover based on predicted candidates. Crucial definitions are listed as follows:

Images: $\mathcal{I}_c^L$ and $\mathcal{I}_d^L$ denote the RGB image and depth image captured by the low-precision camera.

Bounding box: $\mathcal{B}_h$ and $\mathcal{B}_o$ denote the hand bounding box and object bounding box detected by the hand-object detector.

Segmentation mask: $\mathcal{M}_h$ and $\mathcal{M}_o$ denotes the hand mask and object mask segmented in the hand bounding box $\mathcal{B}_h$ and object bounding box $\mathcal{B}_o$ respectively.

Point clouds: $\mathcal{P}_h^L$ and $\mathcal{P}_o^L$ denotes the hand points and object points generated from low-precision camera perception. $\mathcal{P}_s^H$ denotes the scene points captured by the high-precision camera. $\mathcal{P}_o^H$ denotes the extracted precise object points for the anthropomorphic hand grasp network as input.

Hand grasp configuration: $\mathcal{H} = \{p, \theta, t\}$ denotes the anthropomorphic hand grasp configuration generated by our network. Specifically, $p \in \mathbb{R}^{1\times 7}$ denotes the hand wrist pose

in $SE(3)$, where first 3 elements represents translation in xyz and last 4 elements represents orientation quaternion in $wxyz$. $\theta \in \mathbb{R}^{1\times 20}$ denotes 20 hand joints. t denotes the hand taxonomy out of 5 diverse grasp taxonomies we select. Multifarious grasp taxonomies are distinguished by the different joint angles of their initial and final states.

IV. Anthropomorphic Hand Grasp Network

In this section, we present our proposed AHG-Net for predicting point-wise anthropomorphic hand dexterous grasp configuration by taking partially observed point cloud of a single object as input. The overall pipeline is illustrated in Fig. 3.

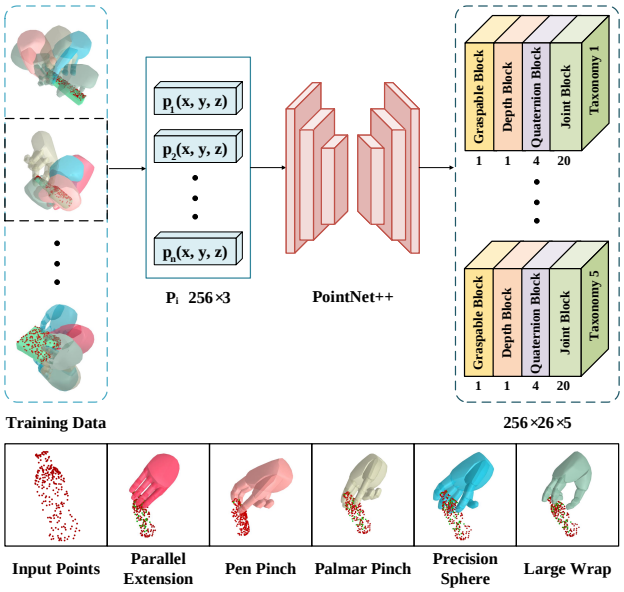


Fig. 3: The overview of proposed anthropomorphic hand grasp network. AHG-Net is trained on the synthetic dataset with over 1M hand grasp annotations from 5K single-view point clouds of 200 objects. Given a 256×3 partial point cloud of an object, our method predicts 5 grasp taxonomies on each point. Specifically, each grasp configuration $\mathcal{H}$ contains 26 parameters. 1 for point graspable determination, 1 for grasp depth, 4 for grasp orientation quaternion and 20 for hand joints.

A. Partial Point Encoding

Compared with parallel-jaw grippers, grasp detection of robot anthropomorphic hands is more challenging. Therefore, more requirements and assumptions are needed for the perceived information. At present, most methods for detecting human-like grasping must be aware of all information of the object geometry, that is, the object must be completely observed [18], [19], [29], [30], [35]. This is impossible in real-world scenarios, especially in human-to-robot object handover tasks. Some works make the efforts to reconstruct partially observed objects, or deploy feature fusion of multiple modalities to deal with the difficulty of appreciating exhaustive geometric information [28], [36]. However, these works generally demand rigorous requirements for experimental equipment that is laborious to reproduce and apply to robot tasks in the real world.

To this end, inspired by detecting grasp configurations on point clouds directly proposed by previous robot grasping work [20], [25], [37], we address predicting appropriate grasping configuration based on the partially observed object under an end-to-end learning-based framework. AHG-Net adopts PointNet++ [38] to extract object partial point feature. PointNet++ pays attention to the partial features of input points, which is capable to facilitate precise grasp prediction on partially observed objects. Specifically, our point-wise grasp detecting approach is based on the PointNet++ segmentation module, which contains three submodules to carry out point graspable determination, hand pose estimation, and hand joint prediction.

B. Point Graspable Determination

Taking encoded partial points features as input, the first submodule of our network determines which points are graspable. Rather than immediately figuring out a specific point is graspable, our AHG-Net pulls off this process taxonomy-by-taxonomy. In other words, the network decides which taxonomies are applicable to each point. We apply weighted cross entropy loss to alleviate the influence on graspable classification caused by unbalancing distribution between graspable (positive) points and ungraspable (negative) points. The point graspable determination loss is defined by:

$$\mathcal{L}_g\left(p_i \mid t\right) = \mathcal{F}_g\left(\boldsymbol{g}(p_i \mid t), \hat{\boldsymbol{g}}(p_i \mid t)\right) \tag{1}$$

where $\mathcal{L}_g\left(p_i \mid t\right)$ denotes graspable determination loss of point p_i for taxonomy t, $\mathcal{F}_g$ denotes the function to calculate cross entropy, $\boldsymbol{g}(p_i \mid t)$ and $\hat{\boldsymbol{g}}(p_i \mid t)$ denotes the annotated and predicted graspable of point p_i for taxonomy t respectively.

C. Hand Pose Estimation

We convert 6-DOF hand pose in $SE(3)$ to $\{\boldsymbol{d}, \boldsymbol{q}\} \in \mathbb{R}^{1\times 5}$ due to the complication of high-dimension pose estimation. Therefore, the second submodule is designed to predict approaching depth $\boldsymbol{d}(p_i \mid t)$ along each point normal $\boldsymbol{n}^{p_i}$ and orientation quaternion $\boldsymbol{q}(p_i \mid t)$ for each taxonomy t. We utilize smooth L1 loss and quaternion loss for approaching depth and orientation quaternion estimation respectively. The quaternion loss is defined by:

$$\begin{aligned}
\mathcal{L}_q\left(p_i \mid t\right) &= \mathcal{L}_p\left(p_i \mid t\right) + \mathcal{L}_a\left(p_i \mid t\right) \\
\mathcal{L}_p\left(p_i \mid t\right) &= \mathcal{F}_p\left(\boldsymbol{q}(p_i \mid t), \hat{\boldsymbol{q}}(p_i \mid t)\right) \\
\mathcal{L}_a\left(p_i \mid t\right) &= \mathcal{F}_a\left(\boldsymbol{q}(p_i \mid t), \hat{\boldsymbol{q}}(p_i \mid t)\right)
\end{aligned} \tag{2}$$

where $\mathcal{L}_q\left(p_i \mid t\right)$, $\mathcal{L}_p\left(p_i \mid t\right)$ and $\mathcal{L}_a\left(p_i \mid t\right)$ denotes quaternion estimation loss, position loss and angle loss of point p_i for taxonomy t respectively, $\boldsymbol{q}(p_i \mid t)$ and $\hat{\boldsymbol{q}}(p_i \mid t)$ denotes the annotated and predicted quaternion respectively. Specifically, $\mathcal{L}_p$ is implemented by smooth L1 loss, $\mathcal{L}_a$ calculates the cosine angle between annotated and predicted rotation matrix obtained from orientation quaternion.

The total hand pose estimation loss is defined by:

$$\begin{aligned}\mathcal{L}_{hp}\left(p_i \mid t\right) &= \mathcal{L}_d\left(p_i \mid t\right) + \mathcal{L}_q\left(p_i \mid t\right) \\ &= \mathcal{L}_d\left(p_i \mid t\right) + \mathcal{L}_p\left(p_i \mid t\right) + \mathcal{L}_a\left(p_i \mid t\right) \\ &= \mathcal{F}_d\left(\boldsymbol{d}(p_i \mid t), \hat{\boldsymbol{d}}(p_i \mid t)\right) \\ &\quad + \mathcal{L}_p\left(p_i \mid t\right) + \mathcal{L}_a\left(p_i \mid t\right)\end{aligned} \tag{3}$$

where $\mathcal{L}_d\left(p_i \mid t\right)$ denotes regressed approaching depth loss of point p_i for taxonomy t, $\boldsymbol{d}(p_i \mid t)$ and $\hat{\boldsymbol{d}}(p_i \mid t)$ denotes the annotated and predicted approaching depth respectively.

D. Hand Joint Prediction

The third submodule predicts 20-DOF hand joint $\boldsymbol{j}(p_i \mid t)$ for each feasible taxonomy on the points. Mean square error (MSE) loss is employed for hand joint estimation. The hand joint prediction loss is defined by:

$$\mathcal{L}_j(p_i \mid t) = \mathcal{F}_j\left(\boldsymbol{j}(p_i \mid t), \hat{\boldsymbol{j}}(p_i \mid t)\right) \tag{4}$$

where $\mathcal{L}_j\left(p_i \mid t\right)$ denotes hand joint prediction loss of point p_i for taxonomy t, $\boldsymbol{j}(p_i \mid t)$ and $\hat{\boldsymbol{j}}(p_i \mid t)$ denotes the annotated and predicted hand joint respectively.

E. Total Loss

We set graspable weight to be $[1, 10]$ as illustrated in Sec. IV-B. The total loss of AHG-Net is formulated as follows:

$$\mathcal{L} = \sum_{p_i \in \mathcal{P}} \mathcal{L}_g(p_i \mid t) + \sum_{p_i \in \mathcal{P}_{pos}} \left(\mathcal{L}_{hp}(p_i \mid t) + \mathcal{L}_j(p_i \mid t)\right) \tag{5}$$

V. Handover System

Our implemented handover system diagram based on our proposed framework is shown in Fig. 4. The system spans six modules that are structured into two input modules for capturing handover scenes by a low-precision RGB-D camera and a high-precision depth camera, two vision perception processing modules for extracting high-precision single-view object points, the AHG-Net module for anthropomorphic hand grasp prediction and robot execution module for robot arm path planning and robot hand configuration accomplishing. Two input modules operate in parallel, other modules operate in sequence to further utilize processing results by the previous module.

A. Coarse Handover Points Generation Module

This module takes RGB image $\mathcal{I}_c^L$ and depth image $\mathcal{I}_d^L$ captured by the low-precision camera as the input to generate coarse handover points. The processing procedure of the module is provided in Fig. 5.

Hand Object Detector. There has been a lot of work on detecting or segmenting human hands and objects. However, the vast majority of related work is aimed at only one of the two tasks, either only human hands or objects are detected. Some work using multiple networks to detect human hand and object separately [4], or only detect one of two (e.g., human hand) and assume the remaining part to the other (e.g., object) within the effective area [3], which can perform well

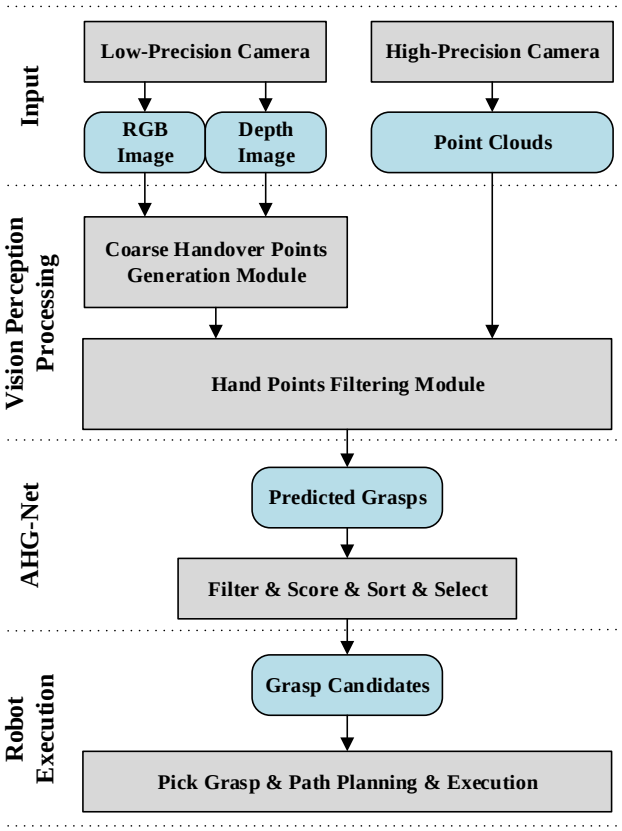


Fig. 4: Our handover system diagram. The system consists of four parts and spans six modules. The input module contains two cameras. Low-precision camera (Azure Kinect) captures RGB image $\mathcal{I}_c^L$ and depth image $\mathcal{I}_d^L$, at the the same time, high-precision camera (Ensenso N35) captures handover scene point clouds $\mathcal{P}_s^H$. $\mathcal{I}_c^L$ and $\mathcal{I}_d^L$ are first fed into coarse handover points generation module to generate hand and object points $\mathcal{P}_h^L$ and $\mathcal{P}_o^L$. Then hand points filtering module takes $\mathcal{P}_h^L$, $\mathcal{P}_o^L$ and $\mathcal{P}_s^H$ to extract precise object points $\mathcal{P}_o^H$, which is utilized by AHG-Net to predict grasp proposals. Such grasps are filtered by approach directions, scored by predicted graspable values as confidence and sorted to select feasible grasps. Robot execution module pick collision-free grasp from grasp candidates and finish subsequent path planning for completing handover.

with suitable camera mounting position and reasonable hand occlusion, but these methods can hardly reflect the hand-object interaction in handover. Due to the arbitrary nature of human hand pose and occlusion during handover, it inevitably affects the actual usage.

We adopt the hand object detector proposed in [39] to obtain the hand and object bounding box during the handover. Based on our camera mounting position, we crop the half RGB image $\mathcal{I}_c^L$ captured by the low-precision camera as the handover detecting region of interest (ROI). It will not exceed the preset ROI no matter what handover pose we use from our testing. The redundant hand bounding boxes are filtered by the distance between the center of the hand and object bounding boxes. As demonstrated in Fig. 6, our final detection result only contains one bounding box $\mathcal{B}_h$ and $\mathcal{B}_o$ for hand and object respectively. The extracted hand and object images $\mathcal{I}_{c(h)}^L$ and $\mathcal{I}_{c(o)}^L$ are used for subsequent process.

Hand Skin Detection and Segmentation. Most learning-

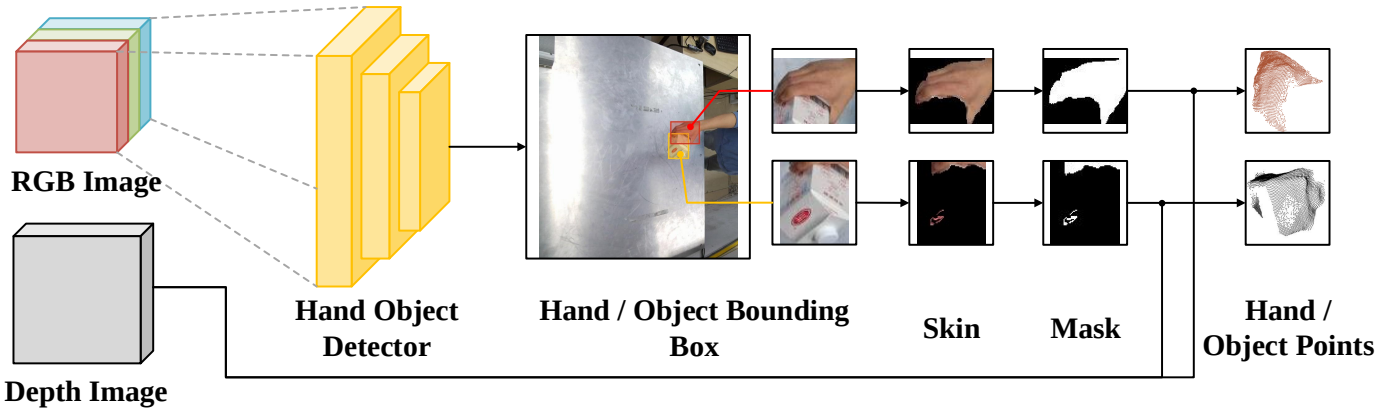


Fig. 5: The overview of coarse handover points generation procedure. The module takes RGB image $\mathcal{I}_c^L$ and depth image $\mathcal{I}_d^L$ captured by the low-precision camera as input. $\mathcal{I}_c^L$ is fed into a hand-object detector network to obtain the hand and object bounding box. The hand bounding box $\mathcal{B}_h$ and object bounding box $\mathcal{B}_o$ are cropped from $\mathcal{I}_c^L$ to attain corresponding RGB image $\mathcal{I}_{c(h)}^L$ and $\mathcal{I}_{c(o)}^L$. A skin detector is applied on $\mathcal{I}_{c(h)}^L$ and $\mathcal{I}_{c(o)}^L$ to carry out the hand segmentation. Finally two mask images $\mathcal{M}_h$ and $\mathcal{M}_o$ are projected into depth image $\mathcal{I}_d^L$ frame to get hand and object points.

Fig. 6: The qualitative results of hand-object detector. After obtaining the detection results inferred by the detector, we calculate the distance between object bounding box and hand bounding box to filter out the idle hand.

based segmentation algorithms at present are implemented based on Fully Convolutional Networks (FCN) [40]. However, the inference speed of FCN is unsatisfactory due to the upsampling layers, and the arbitrary pose of the human hand during the handover also makes the network segmentation effect hard to be guaranteed.

Instead of training an FCN for the hand segmentation task, we simply realize this by deploying the skin detection via converting the hand and object RGB images $\mathcal{I}_{c(h)}^L$ and $\mathcal{I}_{c(o)}^L$ into YCbCr color space [41]. We tune the threshold in YCbCr color space to extract the pixels belonging to the hand and get the hand mask in both two images. Though the performance is affected when the color the of object and human hand are similar to each other, it obtains rational skin detection results efficiently with a reasonable parameter set in most cases. The human hand mask in object image $\mathcal{I}_{c(o)}^L$ is inverted to obtain the object mask. Hand and object masks $\mathcal{M}_h$ and $\mathcal{M}_o$ are projected into depth image $\mathcal{I}_d^L$ to generate hand and object points $\mathcal{P}_h^L$ and $\mathcal{P}_o^L$ with camera intrinsic parameters.

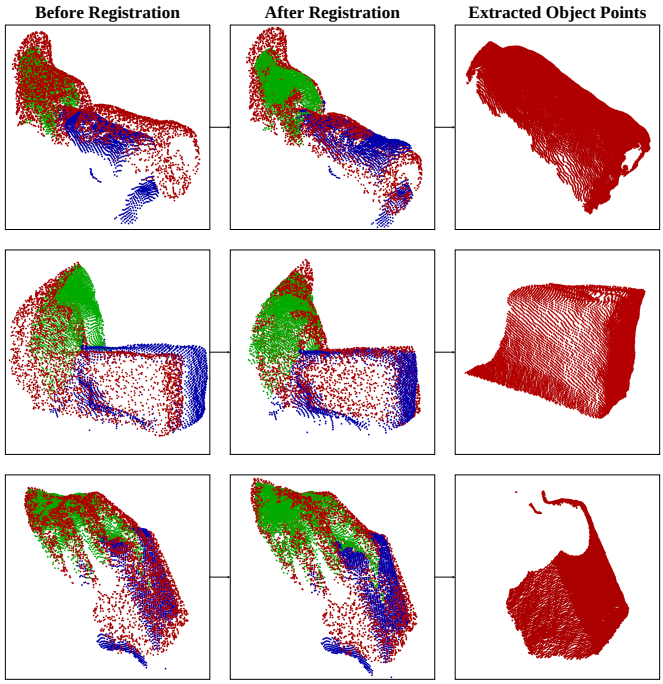


Fig. 7: The process of the precise object points $\mathcal{P}_o^H$ extraction. Scene points $\mathcal{P}_s^H$, hand points $\mathcal{P}_h^L$ and object points $\mathcal{P}_o^L$ are colored in red, green and blue respectively. The first column is the point clouds before registration, the second column is the point clouds after apply ICP registration, the third column is the extracted $\mathcal{P}_o^H$.

B. Hand Points Filtering Module

To extract high-precision single-view object points $\mathcal{P}_o^H$, this module first calibrates two scene points $\mathcal{P}_s^L$ and $\mathcal{P}_s^H$ captured from two cameras by transforming them from camera frame to world frame. To avoid the effect of calibration error, we utilize ICP registration [42] to employ a fine transformation on the point clouds captured by the low-precision camera. We

apply two KD-Tree models with a radius of 3 cm on $\mathcal{P}_s^H$ by taking $\mathcal{P}_h^L$ and $\mathcal{P}_o^L$ separately as point querying. The former one aims to filter out the object points from $\mathcal{P}_s^H$ that are very close to hand $\mathcal{P}_h^L$, since predicted grasps on these points have high probability of collision with the human hand. The latter one will extract the object points $\mathcal{P}_o^H$ from $\mathcal{P}_s^H$. To speed up querying, we sample 256 points on both $\mathcal{P}_h^L$ and $\mathcal{P}_o^L$ by farthest point sampling. Extracted object points $\mathcal{P}_o^H$ are fed into the AHG-Net module for predicting and selecting grasps. The process of the precise object points $\mathcal{P}_o^H$ extraction is shown in Fig. 7.

C. AHG-Net Module

By taking object points $\mathcal{P}_o^H$ as input, AHG-Net predicts the grasp configurations on each point. To avoid the robot motion planning failure during execution, we first filter all the generated grasps. Some unreasonable grasps will not be considered as grasp candidates based on the angle between the approach direction and the unit vector of the coordinate axes. Specifically, we calculate the angle between the grasp approach direction and the opposite direction of y-axis, those grasps with an angle less than 85 degrees are removed. This is to filter out those grasps generated on the human side, which may cause failure in path planning or lead to a safety hazard to the human giver during the handover based on the robot end effector mounting position and handover safety considerations. Selected grasp candidates will be sorted based on the predicted value of the graspable as the confidence score, and the results will be passed to the robot execution module.

D. Robot Execution Module

As illustrated in Algo. 1, given the set of grasp candidates, the module loops each grasp, recovers grasp configuration to the hand mesh $\mathcal{M}_h$ ($\mathcal{M}_h$ denotes recovered robot hand mesh in this subsection, which has a conflict with the symbol of segmented hand mask in Sec. III.) and determines whether the grasp will have any collision with human hand by computing Signed Distance Field (SDF) between hand mesh $\mathcal{M}_h$ and human hand point P_h^L. Once the current grasp is checked as a collision-free configuration, the RRT-Connect algorithm [43] is applied to find a feasible path from the robot home position to the position 15 cm back along the grasp approach direction. We specify that the algorithm must find the path within 1 second, otherwise the grasp will be skipped for the next one and current path planning will be recorded as a failure case. Once the path planning fails on 3 grasps, the current handover will be considered as a failure. The robot arm will follow the planned path to reach the object, and go along the approach direction of 15 cm in a straight line. Robot hand $\mathcal{R}_h$ will be closed according to the predicted joint angle g_i to complete the grasp. During executing grasp, the human giver will determine if the obvious collision and penetration will occur on robot hand $\mathcal{R}_h$ and delivered object O. The robot arm finally moves to the home position to finish the handover process. The object falling flag $\mathcal{F}_o$ is the last success decision for the handover process.

Algorithm 1 Robot execution process

INPUT: Grasp candidates $\mathcal{G}_c = \{g_1, g_2 ... g_n\}$, extracted human hand points $\mathcal{P}_h^L$, Delivered object O
OUTPUT: Handover success determination s
define $t = 0$
for all $g_i \in \mathcal{G}_c$ **do**
  $\mathcal{M}_h \leftarrow$ recover_hand_mesh(g_i)
  if collision_check($\mathcal{M}_h$, $\mathcal{P}_h^L$) is **True**:
    continue
  else:
    if path_planning(g_i) is **False**:
      $t \leftarrow t + 1$
      if $t \geq 3$:
        return $s \leftarrow$ **False**
      else:
        continue
    else:
      $\mathcal{R}_h \leftarrow$ execute(g_i)
      if collision_check($\mathcal{R}_h$, O) is **True**:
        return $s \leftarrow$ **False**
      else:
        $\mathcal{F}_o \leftarrow$ move_to_home_position($\mathcal{R}_h$)
        if $\mathcal{F}_o$ is **True**:
          return $s \leftarrow$ **False**
        else:
          return $s \leftarrow$ **True**

VI. Dataset Generation

In this section, we provide a brief view of our grasp configuration on the partial point cloud generation pipeline.

A. Objects

Inspired by [17], we select 200 miscellaneous objects with numerous categories, shapes, and scales from the dataset. Unlike the purpose of statically grasping objects placed on the plane, we need to accomplish human-to-robot object handover, the objects we choose are safe for both the human giver and the robot receiver.

B. Single Object Grasp Generation

We adopt the approach-based sampling scheme summarized in [45] to generate the grasp configuration for each object. Specifically, we randomly sample a vertex on object model mesh, and let the hand approaches the object along the backward direction of vertex normal. During the approaching, grasp attempts and uniform rotation are taken simultaneously. We sample 512 vertices on each object model mesh as the full observed point cloud of the object. Then the approach-based sampling method is applied for each sampled point. The combination of predefined depths, angles and 5 taxonomies [44], [46] (Fig. 8) on each point are evaluated on the physical simulator MuJoCo [47]. For each taxonomy, the hand closes from the initial joint to the final joint until any contact occurs between hand and object. After applying a slight shake and filtering the object falling or unstable cases, proper grasp configurations are recorded as grasp annotations for the object.

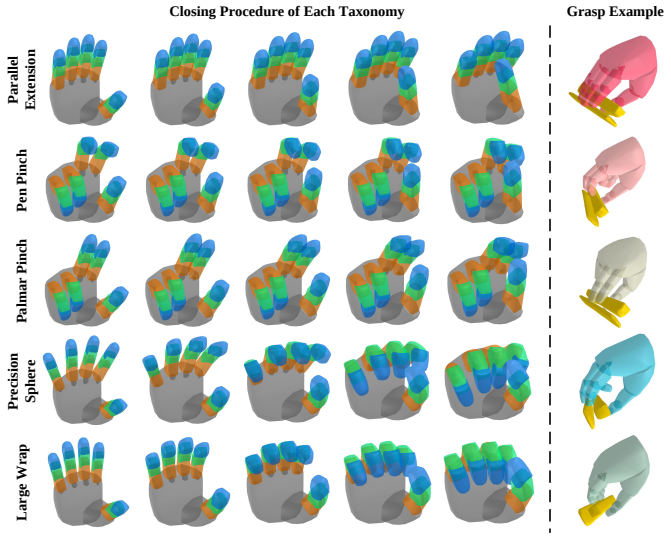


Fig. 8: The closing procedure of each taxonomy from initial joint to final joint. The most right column is the grasp example of the corresponding taxonomy. Five grasp taxonomies are selected based on the standard grasp taxonomies [44] and the characteristics of the anthropomorphic hand. Selected taxonomies can be divided into thumb abduction and thumb adduction from the thumb movement, while they can be also divided into power and precision determined by if all movements of the object have to be evoked by the arm. The specific description of each taxonomy is Parallel Extension (thumb adduction, precision), Pen Pinch (thumb abduction, precision), Palmar Pinch (thumb abduction, precision), Precision Sphere (thumb abduction, precision), Large Wrap (thumb abduction, power).

C. Single Object Partial Point Cloud Grasp Generation

We utilize BlenderProc [48] to generate 5K single object scenes. One object is arbitrarily selected from the dataset and a random pose is applied to the object. The gravity is disabled to simulate objects held by a human. The scene is captured by 4 depth cameras mounted in different views. The partial point cloud P_i is obtained from the captured depth images. The generated partial point cloud of the object is to simulate the situation where the object is occluded by the human hand during the handover. We then transform the corresponding grasp annotations generated in the previous step based on the object pose applied during scene rendering in BlenderProc. A KD-Tree point querying algorithm with 2 cm as the radius is applied to register the grasp annotation to the partial point cloud of the object. Finally, we randomly choose one grasp configuration for each taxonomy on each point. Generated dataset is visualized in Fig. 9.

VII. Experiments

In this section, we first introduce our system setup, AHG-Net implementation detail and evaluation metrics. We then evaluate our AHG-Net in simulation environment and handover system on the real robot. The experiment results not only demonstrate our proposed AHG-Net is capable of predicting robust and judicious grasp proposals by taking a single-view partial point cloud of object as input, but illustrate the reasonableness of our system construction and the adaptability to the arbitrary pose of human hand and random occlusion of delivered objects in handover.

Fig. 9: Generated anthropomorphic hand grasp dataset for a single object. We pick at most 5 annotations for visualization. The first column is the entire grasp annotations for the object. Object meshes are colored yellow, while the red points on the objects are the observed partial points of the objects. The last five columns are grasp annotations visualized for each taxonomy.

A. System Setup

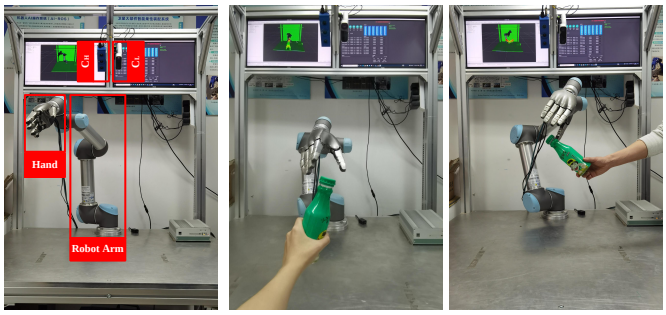


Fig. 10: The hardware setup of our handover system. Key hardware are marked in four red boxes. "C_H" denotes Ensenso N35 for high precision camera, "C_L" denotes Azure Kinect for low precision camera. Two cameras capture the scene from the backside at a 45-degree viewpoint. "Hand" denotes HIT-DLR II anthropomorphic robot hand. "Robot Arm" denotes UR5 robot arm. The handover experiments are conducted from frontal and lateral two delivering positions.

For this handover task, a desk-mounted UR5 robot arm is used. The robot has 6 degrees of freedom, a maximum reach of 850 mm, and a HIT-DLR II anthropomorphic hand as its end effector. The maximum playload of the arm is 5 kg with the end effector 1.5 kg weight accounted. An Azure Kinect RGB-D camera and an Ensenso N35 camera are mounted on the top of the robot and capture scenes at a 45-degree viewpoint. Specifically, Azure Kinect RGB-D camera is used as the low-precision camera, captured RGB image $\mathcal{I}_c^L$ and depth image $\mathcal{I}_d^L$ are processed to generate coarse handover points to extract the high-precision object points $\mathcal{P}_o^H$ from the scene points $\mathcal{P}_s^H$ captured by Ensenso N35 camera. We crop half RGB image $\mathcal{I}_c^L$ and depth image $\mathcal{I}_d^L$ captured by Azure Kinect RGB-D camera as effective region for coarse handover points generation, while scene points $\mathcal{P}_s^H$ captured by Ensenso N35 located in 50cm $\times$ 50cm $\times$ 50cm cube are cropped for object points extraction. We calibrate two cameras separately and fine-tune their extrinsic parameters to make sure scene points $\mathcal{P}_s^L$ and $\mathcal{P}_s^H$ in the world frame overlap perfectly. All the above modules share a desktop computer with one NVIDIA RTX 2070 GPU. ROS is used to control the robot arm which

adopts MoveIt! [49] for path planning based on the generated OctoMap [50] from scene points $\mathcal{P}_s^H$ to ensure the safety of human giver during handover. HIT-DLR II hand control runs on another computer. The handover system hardware setup is shown in Fig. 10.

B. AHG-Net Implementation Details

We sample 256 points for each single-view object points as input. The network is trained on one NVIDIA RTX 3070 GPU for 40 epochs with 8 batch size and Adam optimizer. Start learning rate is set to 0.001, and decreased by a factor of 2 for every 10 epochs.

C. Evaluation Metrics

We first evaluate our AHG-Net in simulation environment, and then validate the handover system on the real robot. The evaluation metrics are slightly different between the two platforms according to previous work [2]–[4], [17], [20]. S and R denote the metric is used in simulation and real robot respectively.

Penetration (S) is used to measure the severity of the collision between the predicted hand mesh and the object model. We only use this metric in simulation environment. The penetration depth and volume are both taken considered for evaluation.

Success Rate (S & R) is used to measure the quality of the grasps. When evaluated in simulation environment, the number of successes is accounted for once if the predicted grasp can stably hold the object for a constant time horizon. During test on the real robot, the following cases will be considered as a failure attempt:

- No path planning solution found within 3 times.
- Obvious collision and penetration will occur when the robot hand approaches the object.
- Object drops from the robot hand when the robot arm moves back to the home position.

Time Cost (R) represents the efficiency of the system to complete the handover, which is a crucial indicator. Human givers do not want to keep the handover posture for a too long time, which will cause their tiredness.

D. AHG-Net in Simulation

Our simulation experiments are carried out by utilizing a physical simulator Mujoco [47]. We evaluate our AHG-Net both on the validation and test dataset which contains 4000 and 2000 different object partial point clouds respectively. Our experiment setup is detailed as follows: as Sec. VI-C describes, we utilize BlenderProc [48] to generate object partial point clouds by arbitrarily selecting an object with random pose and capturing the scene by 4 depth cameras mounted in different views. We first pass the point cloud into AHG-Net for forward inference to obtain predicted grasp proposals. Then the top 5 grasp configurations $\mathcal{H}_{w/o}^5$ are selected by sorting the total grasps via predicted graspable value as confidence score, which is used for evaluating our method by ignoring the taxonomy consideration. We also select top 5 grasp configurations $\mathcal{H}_w^5 = \{\mathcal{H}_{tax1}^5, \mathcal{H}_{tax2}^5, \mathcal{H}_{tax3}^5, \mathcal{H}_{tax4}^5, \mathcal{H}_{tax5}^5\}$ for each taxonomy by using same sorting indicator as before, so that we can evaluate the grasp performance of each taxonomy. Object is loaded in Mujoco with disabled gravity and applied the same pose as captured in BlenderProc during dataset generation. Robot hand joints are set to the predicted grasp configuration. The gravity is enabled after the above processes are done, the grasp is considered as a successful one if the object does not drop down within 5 seconds by applying a slight shake on the hand wrist. Fig. 11 illustrates the experimental results in Mujoco.

Fig. 11: Qualitative experimental results for AHG-Net in Mujoco simulator.

TABLE I: Grasp Result in Simulation of Each Taxonomy and Totals

	Success Rate (%)		Penetration D (e^{-3} cm)		Penetration V (e^{-8} cm^3)	
	Val	Test	Val	Test	Val	Test
Parallel Extension	79.98	74.06	6.20	7.02	3.37	5.92
Pen Pinch	89.21	83.38	9.19	12.04	20.92	71.43
Palmar Pinch	85.11	80.71	9.06	11.19	3.52	6.36
Precision Sphere	81.75	82.66	6.33	7.08	6.35	8.17
Large Wrap	75.24	77.76	5.74	6.43	6.37	8.60
w/o Taxonomy	84.27	80.13	6.81	8.21	24.18	47.59

As the simulation experiment results are shown in Tab. I, not only does our method perform on the validation and test dataset with a slight difference, which indicates that our network has satisfactory generalizability. But our model performs well whether the taxonomy is considered or not. It also demonstrates that the grasping success rate of our method does not depend on individual taxonomy, but each one performs robustly.

E. Handover System on Robot

To evaluate the robustness and reliability of our handover system, we select 15 novel objects with different geometric shapes and sizes that are not in our dataset as shown in Fig. 12. During handover, we let the human giver uses arbitrary pose and occlusion on the delivered object to avoid the high success rate caused by repeated use of handover poses that may be beneficial for camera capture or robotic arm path planning. Each object will be delivered 20 times for both the frontal and lateral positions. Any case that occurs as described in Sec. VII-C will be considered as a failure attempt. The time to successful handover, the number of grasp attempts and

TABLE II: Handover Result on Robot with Two Positions.

	Object ID		1	2	3	4	5
Position	Frontal	Time (s)	19.12 ± 4.75	19.53 ± 6.09	18.22 ± 7.94	18.75 ± 4.99	19.00 ± 10.31
		Success Rate	85%	90%	75%	80%	60%
		Number of Attempts	1.29	1.50	1.47	1.44	1.83
	Lateral	Time (s)	19.46 ± 2.56	19.60 ± 4.14	19.61 ± 2.91	19.77 ± 2.73	20.04 ± 2.86
		Success Rate	80%	95%	80%	90%	70%
		Number of Attempts	1.63	1.26	1.50	1.61	1.57

	Object ID		6	7	8	9	10
Position	Frontal	Time (s)	18.44 ± 4.08	19.00 ± 4.48	19.06 ± 3.50	18.60 ± 3.76	19.46 ± 3.30
		Success Rate	95%	80%	90%	80%	85%
		Number of Attempts	1.58	1.44	1.61	1.94	1.53
	Lateral	Time (s)	20.29 ± 5.33	20.25 ± 2.43	19.74 ± 3.71	19.48 ± 2.55	20.26 ± 2.50
		Success Rate	90%	90%	90%	95%	100%
		Number of Attempts	1.72	1.56	1.94	1.47	1.55

	Object ID		11	12	13	14	15
Position	Frontal	Time (s)	18.74 ± 4.22	19.19 ± 1.99	19.29 ± 6.67	19.52 ± 4.98	20.54 ± 7.68
		Success Rate	95%	90%	80%	85%	75%
		Number of Attempts	1.26	1.33	1.69	1.41	1.53
	Lateral	Time (s)	21.15 ± 5.82	19.83 ± 3.90	20.75 ± 3.32	19.83 ± 2.98	20.18 ± 3.55
		Success Rate	80%	80%	80%	65%	70%
		Number of Attempts	1.56	1.38	1.56	1.54	1.86

Fig. 12: 15 novel objects for system evaluation.

success rate are computed for each object with two different handover positions. Specifically, we follow the previous work [3] to record the number of times the robot tries to approach and grasp the objects as the number of attempts.

In summary, our handover system has 83% and 84% success rate for frontal and lateral handover position of human giver respectively. As illustrated in Tab. II, the slight difference between the two success rates demonstrates our system is able to adapt to not only the arbitrary poses and occlusions during handover but diverse handover positions. It is also reasonable that the time of lateral position is greater than frontal position because objects delivered from the lateral position are normally farther than the frontal position in our robot experiment platform.

Our handover system qualitative results of delivering 15 novel objects are shown in Fig. 13. Our system is capable of not only predicting grasps on several locations of the delivered objects, but also using numerous grasp taxonomies with different poses. The results demonstrate our handover system has an extensive variety of miscellaneous objects.

When conducting the simulation experiments on our proposed AHG-Net, any slight collision during grasp will change object pose, and even make the object directly fall on the floor, which causes a failure grasp case. However, the object is held by the human giver during handover, the human giver cannot guarantee to keep the same posture without any move, while the slight collision will not lead to an obvious pose change of the object either, which result in our handover system success rate is better than experiments on AHG-Net.

TABLE III: The average time cost for each module of handover system. Input: input module. Points: coarse handover point cloud generation module. Filter: hand points filtering module. Net: AHG-Net module. Execute: robot execution module.

	Input	Points	Filter	Net	Execute
Time (s)	2.17	0.96	0.35	0.12	16.28

We also record and compute the average time cost for each module of our system (listed in Tab. III), the input module takes 2.17 seconds for image capture, the next module spends 0.96 seconds generating coarse handover points, then we get high-precision object points after 0.35 seconds hand filtering, by waiting for 0.12 seconds to predict grasps proposals, execution module takes 16.28 seconds to finish handover process.

F. System Ablation Study on Robot

To examine the reasonability of our system setup, we carry out our system ablation study from three aspects: input, prediction, and planning. Specifically, to validate the necessity of using two cameras to accomplish the handover process, we modify the entire pipeline by only utilizing Azure Kinect RGB-D camera as a perceptual device. We then train two new grasp networks: AHG-Net (3-tax) by eliminating 2 grasp taxonomies Palmar Pinch and Large Wrap which behave similarly to Pen Pinch and Precision Sphere in some cases and AHG-Net (w. noise) by adding noise to the dataset to improve its robustness. We replace the original AHG-Net in the system with AHG-Net (3-tax) and AHG-Net (w. noise) to test the performance. RRT* [51] is another excellent path plan

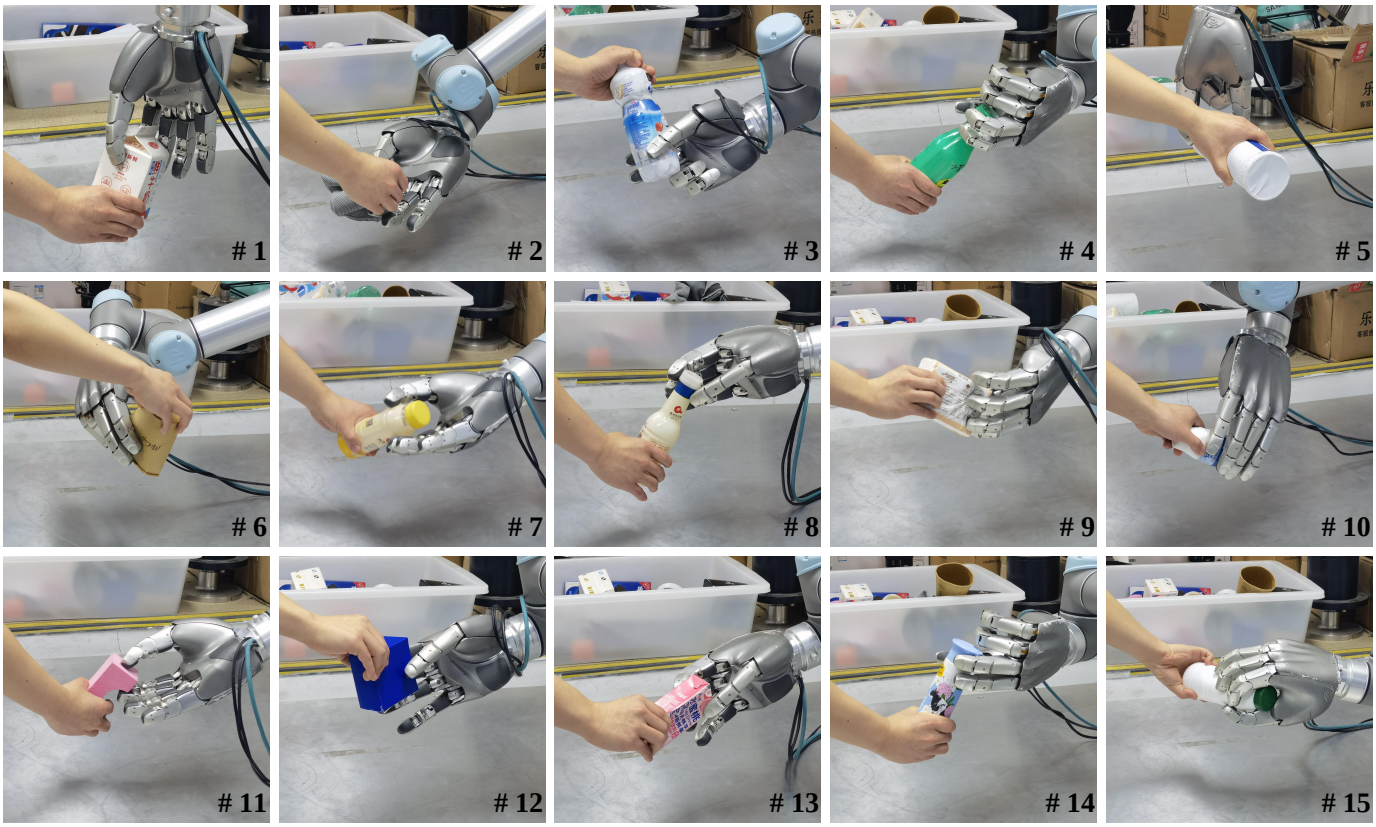


Fig. 13: Qualitative experimental results for handover by our system with 15 novel objects with diverse shapes and sizes.

algorithm, we change RRT-Connect to RRT-star in the robot execution module to evaluate how the path planning algorithm affects final handover performance. Human giver delivers 15 objects same as selected in the above experiments 6 times only from the frontal position. Path planning is allowed to take 1 second to complete at most. We keep the success determination criteria in line with those described in Sec. VII-C. The ablation experiment results are illustrated in Tab. IV.

Camera. The input module uses both Azure Kinect RGB-D camera and Ensenso N35 camera exceeds the performance of one that only uses Azure Kinect RGB-D camera by a large margin. Unlike the object grasp work utilize a parallel-jaw gripper as the end effector [3], [4], [25], [52], which performs well only taking point clouds captured by a low-precision camera, anthropomorphic hand dexterous grasp configuration prediction requires more precise perception device. Azure Kinect is indispensable for our system setup, the captured RGB image is utilized for hand and object detection for further mask extraction and coarse points generation. Although some literature proposes methods directly segment hand on depth images [53], [54], their performances are not guaranteed on arbitrary poses and occlusions of human hand during handover, which will affect further hand points filtering and collision check. We use the respective features of two cameras to quickly extract high-precision object point clouds, which provides a guarantee for the execution of downstream modules. The failure cases of only using Azure Kinect camera is shown in Fig. 14.

Network. The performance gap between AHG-Net and AHG-Net (3-tax) illustrates our selected 5 grasp taxonomies

TABLE IV: Handover system ablation study

Camera		Network			Path Planning		Success Rate ↑	Time (s) ↓
Azure Kinect	Ensenso N35	AHG-Net (3-tax)	AHG-Net	AHG-Net (w. noise)	RRT*	RRT-Connect		
✓		✓			✓		48.89%	19.56
✓			✓		✓		56.67%	19.77
✓				✓	✓		68.89%	19.53
✓		✓				✓	48.89%	20.50
✓			✓			✓	64.44%	19.82
✓				✓		✓	76.67%	19.79
✓	✓	✓			✓		67.78%	19.88
✓	✓		✓		✓		70.00%	19.43
✓	✓			✓	✓		67.78%	19.71
✓	✓	✓				✓	72.22%	19.22
✓	✓		✓			✓	83.33%	19.42
✓	✓			✓		✓	82.22%	19.66

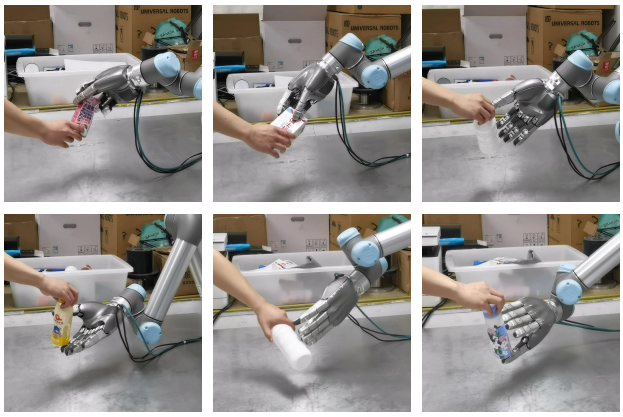


Fig. 14: The failure cases of only using Azure Kinect camera. The performance drops obviously by taking low-quality point clouds captured by low-precision camera.

are beneficial for increasing the handover success rate. Compared with Pen Pinch, Palmar Pinch is more inclined to grasp the cylinder with a larger radius or a flatter surface. Pen Pinch and Palmar Pinch grasp examples are shown in Fig. 13-4 and Fig. 13-3 respectively. Similarly, Large Wrap normally intends to wrap the objects tightly other than grasping objects by fingertips adopted by Precision Sphere. Fig. 13-7 and Fig. 13-15 show the difference between Precision Sphere and Large Wrap. Because of the random poses and occlusions, captured object point clouds are unpredictable, eliminating some taxonomies will cause the robot executes an unsuitable grasp on delivered objects.

AHG-Net (w. noise) outperforms original AHG-Net by a large margin when only Azure Kinect camera is used. However, the performance of AHG-Net (w. noise) drops a little bit when both cameras are used. The reason for this performance reduction is mainly because the training of AHG-Net (w. noise) is carried out on the point cloud with added noise, but for a high-precision camera such as the Ensenso N35, the extracted object point clouds usually have negligible noise that makes the point clouds fed into the network have a certain difference in data distribution with the point cloud with noise added during training, which leads to its performance is not as good as the original AHG-Net. For the results that AHG-Net (w. noise) does not gain better performance on one-camera system compared with two-camera system, Azure Kinect interpolates the depth values at surface edges (depth discontinuities) in many cases, just the degrees of distortion of the object point clouds are different. Due to this effect, even with the addition of noise to train the network, it is difficult to fully adapt to the point cloud distortion from the camera. For the parallel-jaw gripper, the network usually only needs to predict the 6D pose of the grasp and the gripper opening width, while for the anthropomorphic hand, its high degrees of freedom of the joints make the network often perform unsatisfactorily on low-quality point clouds.

Path Planning. In terms of time efficiency, there is no significant difference between the two planning algorithms. However, the improved success rate of RRT-Connect indicates it fits our handover system better than RRT*. The most considerable shortcoming of RRT* is its slow path searching time for our task. RRT* can always find a more optimal path than RRT-Connect once it succeeds within 1 second, which leads to no influence on time efficiency. Yet unable to find a solution more than 3 times also happens frequently by using RRT* causes many failure cases based on our criteria. The time cost is a substantial indicator of the handover system, long waiting time will make object holding posture unstable, and cause tiredness and impatience of human givers.

Overall, our system setup is reasonable according to our ablation study. We use two cameras with different precision to extract object points efficiently and effectively. Selected 5 grasp taxonomies are robust to unpredictable single-view points input and promote the robot grasp delivered objects dexterously. RRT-Connect is adopted to quickly search for a path to reach the object. Experiment results demonstrate our system has the best performance over 8 system setups.

G. Comparison on Robot

To validate how our proposed method benefits handover system, we replace AHG-Net with HGC-Net [21] and DVGG [17] from our prior work. Same 15 novel objects are delivered 6 times for each with arbitrary poses. We keep other modules consistent and adopt identical criteria for success number count.

TABLE V: Comparison results of different grasp prediction methods.

Method	Success Rate ↑	Time (s) ↓
HGC-Net [21]	70.00%	19.86
DVGG [17]	28.89%	19.62
Ours	**83.33%**	**19.42**

As the comparison results are shown in Tab. V, our proposed method outperforms other methods by a large margin. HGC-Net is designed for grasping objects placed on the desk, while DVGG adopts Conditional Variational Auto-Encoder to generate the grasp on objects with object points completion. The former one performs well when the delivered objects are not too far from the desk. However, when captured object points are too few, or objects are held too high, HGC-Net will have a higher probability to cause inference failure. DVGG passes single-view object points to the point cloud completion module to recover the entire geometric information of the object. The point cloud completion module restricts that the object must be captured from a specific direction and angle in order to obtain better results for the grasp sampler. Therefore, our proposed method solves two problems compared with them: (1) our method can handle the variety of height changes of delivered objects; (2) our method can handle the variety of postures of delivered objects.

H. User Study

We conduct a user study experiment with 6 participants to illustrate the robustness and adaption of our handover system.

Objects Selection. We prepare two object sets for the participants. Specifically, *Set*-I contains 15 objects used in the above experiments, *Set*-II contains 7 novel objects with more complex geometric shapes (shown in Fig. 15).

Fig. 15: 7 novel objects for user study.

Experiment Setup. We keep the same system pipeline and success criteria as used in previous experiments. Each participant is asked to deliver total of 15 objects over 2 object sets. Participants are told to randomly select 8 object from *Set*-I and use all objects from *Set*-II. All objects are passed from the frontal position. The camera start to capture once they stably hold the object and keep the same posture.

The experiment results in Tab. VI demonstrate our handover system can not only adapt to different users but generalize to diverse objects with miscellaneous geometric shapes and sizes. The success rate and time cost are not affected by adding novel and more complicated objects. Our system is able to finish the majority of object handovers within 1 attempt, which further evaluates the reasonability and reliability of the system.

TABLE VI: The experiment results of user study

	Success Rate	Time (s)
User 1	80.00%	19.34
User 2	73.33%	19.60
User 3	73.33%	19.26
User 4	80.00%	19.23
User 5	80.00%	19.41
User 6	86.67%	19.56
Average	78.89%	19.40

VIII. Conclusion

We present a human-to-robot handover framework by utilizing an anthropomorphic robot hand, which dexterously receives and grasps the objects delivered by the human giver. We implement the system from the framework that takes images captured by two cameras to accomplish hand-object detection and segmentation, object points extraction, and precise anthropomorphic grasp proposals prediction. We also build a large-scale dataset for grasp prediction on single-view object points and propose a lightweight network associated with the dataset. Experiments demonstrate the handover system not only adapts to different users but generalizes to diverse novel objects with miscellaneous shapes and sizes.

Future work spans three directions: (1) we can improve the current system as close-loop to react to the motion of the human giver so that the grasp planner can update the grasp pose online and in real-time; (2) to further explore the capabilities of anthropomorphic robot hand, the grasp planner should generate reasonable grasps on the affordance of the objects for the purpose of subsequent manipulation; (3) promote the reproducibility of the system by using a single camera without degrading the performance of the system.

References

[1] A. Edsinger and C. C. Kemp, "Human-robot interaction for cooperative manipulation: Handing objects to one another," in *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2007, pp. 1167–1172.

[2] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, "Object handovers: a review for robotics," *IEEE Transactions on Robotics*, 2021.

[3] W. Yang, C. Paxton, A. Mousavian, Y.-W. Chao, M. Cakmak, and D. Fox, "Reactive human-to-robot handovers of arbitrary objects," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 3118–3124.

[4] P. Rosenberger, A. Cosgun, R. Newbury, J. Kwan, V. Ortenzi, P. Corke, and M. Grafinger, "Object-independent human-to-robot handovers using real time robotic vision," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 17–23, 2020.

[5] W. Yang, B. Sundaralingam, C. Paxton, I. Akinola, Y.-W. Chao, M. Cakmak, and D. Fox, "Model predictive control for fluid human-to-robot handovers," *arXiv preprint arXiv:2204.00134*, 2022.

[6] P.-K. Chang, J.-T. Huang, Y.-Y. Huang, and H.-C. Wang, "Learning end-to-end 6dof grasp choice of human-to-robot handover using affordance prediction and deep reinforcement learning," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

[7] K. Yamane, M. Revfi, and T. Asfour, "Synthesizing object receiving motions of humanoid robots with human motion database," in *2013 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2013, pp. 1629–1636.

[8] A. Koene, S. Endo, A. Remazeilles, M. Prada, and A. M. Wing, "Experimental testing of the coglaboration prototype system for fluent human-robot object handover interactions," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 249–254.

[9] A. Koene, A. Remazeilles, M. Prada, A. Garzo, M. Puerto, S. Endo, and A. M. Wing, "Relative importance of spatial and temporal precision for user satisfaction in human-robot object handover interactions," in *Third International Symposium on New Frontiers in Human-Robot Interaction*, 2014.

[10] M. Prada, A. Remazeilles, A. Koene, and S. Endo, "Implementation and experimental validation of dynamic movement primitives for object handover," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2014, pp. 2146–2153.

[11] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: learning attractor models for motor behaviors," *Neural Computation*, vol. 25, no. 2, pp. 328–373, 2013.

[12] G. J. Maeda, G. Neumann, M. Ewerton, R. Lioutikov, O. Kroemer, and J. Peters, "Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks," *Autonomous Robots*, vol. 41, no. 3, pp. 593–612, 2017.

[13] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[14] M. K. Pan, E. Knoop, M. Bächer, and G. Niemeyer, "Fast handovers with a robot character: Small sensorimotor delays improve perceived qualities," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6735–6741.

[15] V. Micelli, K. Strabala, and S. Srinivasa, "Perception and control challenges for effective human-robot handoffs," 2011.

[16] J. Konstantinova, S. Krivic, A. Stilli, J. Piater, and K. Althoefer, "Autonomous object handover using wrist tactile information," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2017, pp. 450–463.

[17] W. Wei, D. Li, P. Wang, Y. Li, W. Li, Y. Luo, and J. Zhong, "Dvgg: Deep variational grasp generation for dextrous manipulation," *IEEE Robotics and Automation Letters*, 2022.

[18] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 107–11 116.

[19] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8709–8719.

[20] W. Wei, Y. Luo, F. Li, G. Xu, J. Zhong, W. Li, and P. Wang, "Gpr: Grasp pose refinement network for cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4295–4302.

[21] Y. Li, W. Wei, D. Li, P. Wang, W. Li, and J. Zhong, "Hgc-net: Deep anthropomorphic hand grasping in clutter," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 714–720.

[22] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[23] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dexterous grasping via eigengrasps: A low-dimensional approach to a high-complexity problem," in *Robotics: Science and Systems manipulation workshop-sensing and adapting to the real world*, 2007.

[24] C. Borst, M. Fischer, and G. Hirzinger, "A fast and robust grasp planner for arbitrary 3d objects," in *Proceedings 1999 IEEE International Conference on Robotics and Automation (ICRA)*, vol. 3. IEEE, 1999, pp. 1890–1896.

[25] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3571–3578.

[26] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3629–3635.

[27] G. Xu, Y. Tao, B. Jiang, P. Wang, Y. Luo, and J. Zhong, "Pois: Policy-oriented instance segmentation for ambidextrous robot picking," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 743–749.

[28] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "Contactpose: A dataset of grasps with object contact and hand pose," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 361–378.

[29] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2386–2393.

[30] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1471–1481.

[31] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5031–5041.

[32] X. Yu, B. Li, W. He, Y. Feng, L. Cheng, and C. Silvestre, "Adaptive-constrained impedance control for human-robot co-transportation," *IEEE transactions on cybernetics*, 2021.

[33] Z. Li, X. Li, Q. Li, H. Su, Z. Kan, and W. He, "Human-in-the-loop control of soft exosuits using impedance learning on different terrains," *IEEE Transactions on Robotics*, 2022.

[34] W. Yang, C. Paxton, M. Cakmak, and D. Fox, "Human grasp classification for reactive human-to-robot handovers," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 11 123–11 130.

[35] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 581–600.

[36] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Generating grasp poses for a high-dof gripper using neural networks," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1518–1525.

[37] H. Duan, P. Wang, Y. Huang, G. Xu, W. Wei, and X. Shen, "Robotics dexterous grasping: The methods based on point cloud and deep learning," *Frontiers in Neurorobotics*, vol. 15, p. 73, 2021.

[38] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[39] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9869–9878.

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[41] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 264–277, 1999.

[42] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[43] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation (ICRA). Symposia Proceedings*, vol. 2. IEEE, 2000, pp. 995–1001.

[44] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2015.

[45] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," *arXiv preprint arXiv:1912.05604*, 2019.

[46] S. Katyara, F. Ficuciello, D. G. Caldwell, B. Siciliano, and F. Chen, "Leveraging kernelized synergies on shared subspace for precision grasping and dexterous manipulation," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[47] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2012, pp. 5026–5033.

[48] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, 2019.

[49] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[50] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous robots*, vol. 34, no. 3, pp. 189–206, 2013.

[51] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 846–894, 2011.

[52] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2901–2910.

[53] A. K. Bojja, F. Mueller, S. R. Malireddi, M. Oberweger, V. Lepetit, C. Theobalt, K. M. Yi, and A. Tagliasacchi, "Handseg: An automatically labeled dataset for hand segmentation from depth images," in *2019 16th Conference on Computer and Robot Vision (CRV)*. IEEE, 2019, pp. 151–158.

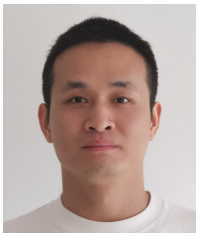
[54] D. H. Nguyen, T. N. Do, I.-S. Na, and S.-H. Kim, "Hand segmentation and fingertip tracking from depth camera images using deep convolutional neural network and multi-task segnet," *arXiv preprint arXiv:1901.03465*, 2019.

Haonan Duan received the B.E. degree in mechanical engineering from East China University of Science and Technology, Shanghai, China in 2017, and the M.S. degree in mechanical engineering from The University of Texas at Dallas, Richardson, Texas, USA in 2019, and the M.S. degree in information sciences from The University of Pittsburgh, Pittsburgh, Pennsylvania, USA in 2021. He is currently pursuing the Ph.D. degree in control theory and control engineering in Institute of Automation, Chinese Academy of Sciences, Beijing, China and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. His current research interests include computer vision, robot manipulation and human-robot interaction.

Peng Wang received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (CASIA) in control theory and control engineering, Beijing, China, in 2010. He is currently a full professor with Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, and a visiting professor with the CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China, and a visiting professor with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China. He has published more than 80 journal and conference papers in artificial intelligence and robotics. He was awarded the First Prize of Science and Technology of Beijing, First Prize of Technological Invention of Chinese Association of Automation, Second Prize of Science and Technology of Beijing, and "Outstanding Reviewers of 2020, 2021" of IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. His current research interests include Robotic Vision, Robotic Learning, Robotic Dexterous Manipulation and Grasping, Anthropomorphic Robotic Hand, Neurorobotics, Brain-like Robotics, and Intelligent Robot systems.

Yiming Li received the B.E. degree in mechanical engineering from the College of Mechanical Engineering, Tongji University, Shanghai, China in 2019 and the M.S. degree in Control Theory and Control Engineering from Institute of Automation, Chinese Academy of Sciences, Beijing, China in 2022. His research focuses on robot manipulation, dexterous hand and machine learning.

Daheng Li received the B.S. degree from the Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China in 2020. He is currently pursuing the M.S. Degree in the School of Artificial Intelligence University of Chinese Academy of Sciences, Beijing, China. His current research interests include reinforcement learning, computer vision, robot control.



Wei Wei received the B.S. degree in PLA Army Academy of Artillery and Air Defense, Hefei, Anhui, China in 2013. He is currently pursuing the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include robotic dexterous grasping and manipulation.