2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# Overcoming Hand and Arm Occlusion in Human-to-Robot Handovers: Predicting Safe Poses with a Multimodal DNN Regression Model

Catherine Lollett[†1], Advaith Sriram[†2], Mitsuhiro Kamezaki[3], and Shigeki Sugano[4]

***Abstract*— Handovers play a key role in human-robot interactions. However, current research focuses on visible-hand handovers, thereby heavily relying on hand detection. Large objects in human-robot interactions present a unique challenge: they inherently block the person's hands and arms from the robot's view. This occlusion raises the robot's risk of unintended physical contact with the person, leading to discomfort and safety concerns. This study aims to develop a model that can determine a pose for the robot that ensures a handover that avoids physical contact with the person, especially in scenarios when hands and arms are occluded. Toward this goal, a three-branch multimodal Deep Neural Network (DNN) regression model was implemented. First, a robust human-pose keypoints detection to calculate shoulder-elbow angles is applied. Secondly, we extract the refined object's segmented mask. Thirdly, we compute two intrinsic object properties. The concatenated outputs from these branches pass through extra dense layers, resulting in the prediction of the robot's 14 arms-joint angles. Compared to an only keypoint data processed-based model, our multimodal approach made a 17.7% accuracy improvement. The experiments highlight each pipeline step's significance, showing important results even when hands and arms were heavily occluded, adjusting to different variations.**

## I. INTRODUCTION

Handovers are in our daily lives, occurring in countless forms: from a caregiver providing a patient a glass of water to a mechanic receiving a tool from an assistant; handovers are a cornerstone of cooperation and efficiency [1]. With the improvements in robotic intelligence, robots will join us in these key acts of collaboration. The challenge lies in ensuring they can do it safely, integrating them into our lives [2].

However, there exists evidence that direct physical contact between the robot and the person may lead to discomfort and safety issues [3] - [5]. Real-world scenarios often involve hands and arms hidden fully or partially from the robot's view, defined as occlusions, as is the case when handling large objects. Such occlusions may cause the robot to misclassify the person's hand position, potentially resulting in unwanted contact during a handover. Given significant pose keypoints occlusions, inaccuracies in keypoint detection frequently arise. A simplistic geometric or keypoints-only approach will often lead to misclassifications.

The combination of the lack of research in scenarios with hand occlusions during large object handovers and the strong reliance on hand detection underscores a need for further exploration in this field. Moreover, the existing dependence on external sensors adds complexity and cost to these systems, further emphasizing the necessity for more integrated and cost-effective solutions. Particularly, scenarios involving handovers from a human giver to a robot receiver are a field that has not been sufficiently explored [6].

In light of these challenges, we aim to implement a model capable of determining an optimal position for the robot that ensures a handover pose while avoiding any direct physical contact with the person, emphasizing situations when the hands and arms are occluded.

Our study involves four key steps as a preliminary approach to predict a safe pose handover system, even in the presence of hand and arm occlusions:

*1. Shoulder-Elbow Angles Calculation:* Based on a robust human-pose detection system, we calculate the persons' shoulders and elbows angles with an arm correction function in case of noisy arm's keypoints data.

*2. Mask Refinement:* We select the object's mask using segmentation, with a refinement step. Masks with a person's form are discarded, and the remaining ones are examined for proximity to a point on the person's torso, calculated from one arm's vector intersecting the opposing elbow's vector.

*3. Object's Intrinsic Attributes Calculation:* By analyzing the object's mask and person's pose, we calculate two objects' intrinsic attributes: the Fourier Descriptor Vector and its related Pose Vector.

*4. Multimodal DNN Regression Model:* The shoulder-elbow angles, object mask segmentation, and object's intrinsic attributes branches are inputs for the multimodal DNN regression model that predicts the robot's 14 arms-joint angles for a robot's optimal position for a safe handover.

Our main contribution is a preliminary multimodal DNN regression model that determines the robot's arm position for a handover avoiding physical contact with the person that includes:

1. Robustness to hand and arm occlusions, providing a reliable handover pose prediction even under these occlusions.

2. Non-intrusiveness, as the model is designed to execute a pose for the robot that avoids any physical contact with the person during the handover.

3. Generalization, ensuring that the model can predict an optimal position regardless of the individual involved.

4. Portability and extensivity, as the model requires only the robot's integrated camera and a processing computer.

5. Near real-time processing, allowing for quick response and adaptation during the handover process.

[†]C. Lollett and A. Sriram are equal contributors to this work and are designated as co-first authors.

[1]C. Lollett is with Future Robotics Organization, Waseda University, Tokyo, Japan (e-mail: clollett@akane.waseda.jp).

[2]A. Sriram and [4]S. Sugano are with the Department of Modern Mechanical Engineering, Waseda University, Tokyo, Japan.

[3]M. Kamezaki is with the Department of Electrical Engineering and Information Systems, The University of Tokyo, Tokyo, Japan.

## II. RELATED WORKS

Safety is crucial in HRI (Human-Robot Interaction). Robots must ensure they grasp the object without touching the human partner. This is achieved through vision, being cautious due to perception noise [7]. However, few studies address datasets where the hand and arms are occluded due to their challenging nature. Moreover, handovers from humans to robots (H2R) remain under-explored [6].

As our research focuses on H2R handovers, this will be our main point of focus. Our comparison is bifurcated: handovers involving small objects and those concerning large objects.

*Small Objects Handovers:* Several works use Kinect and keypoint detection. For instance, [8] employs Kinect's data to detect keypoints on a visible subject's body during robot-human handovers. Similarly, [9] and [10] employ Kinect RGB-D. Another group, like [11], uses RGB for gesture recognition, using a multimodal classification for handover-initiation detection. [12], relies on the YOLO detector and an RGB-D camera, integrating body and hand segmentation with object detection, limiting it to YOLO's object categories. Within the scope of wearable sensors and motion capture, [13] developed their approach using a wearable sensory system, while [14] used the PhaseSpace Motion Capture system for wrist pose sense. Shifting to depth image, point clouds, and multisensory methods, [15] merges RGB, depth image, and point clouds, using sensors like Kinect RGB-D and Ensenso N35. Additionally, [16] used a set of 12 OptiTrack Flex 13 motion capture cameras. Elsewhere, [17] focuses on visual object tracking, and [18] uses QR codes and sensors, which might face real-world constraints.

These methods may face challenges with larger objects due to their heavy dependence on full hand and body pose detection, that can be limited to specific object categories, involve complex systems, or require invasive wearables. In contrast, our approach addresses these challenges with robust data models, and non-invasive methods to handle occlusions.

*Large Object Handover:* For larger objects, [19] uses QR code detection, showing also concern about most of the handover systems only focusing on one-handed small objects' handover. They explain, that big objects require us to finish the handover task with two hands. [20] proposes a multi-sensor fusion approach that needs calibration. [21] uses two Kinect V2 sensors, and prioritizes trajectory prediction. [22], [23] use large distances between robot and human during large object handovers that may overlook hand collisions.

Large object handover methods have limitations. Dependence on markers like QR codes, calibration, and assumptions reduce usability. Our solution offers flexibility, avoiding external markers and presumptions, ensuring a wide range of real-world unpredictability. We also consider hand occlusions, ensuring safer human-robot interactions.

In summary, our novel approach addresses the challenges with hand and arm occlusions in handover scenarios that have not been thoroughly studied previously. Prioritizing versatility, broader object detection, and safer interactions, our goal is to create a safer H2R handover scenario.

## III. APPROACH

Given the complexities in human-robot interactions and the highly noisy data from significant occlusion of a person's pose keypoints, a simplistic geometric or keypoints-only approach will lead to misclassification. Based on successful multimodal models [24], we design a three-branch multi-modal DNN regression model.

The based data for the branches of our model is calculated from a frame captured by the robot's fisheye camera, which undergoes a de-fisheye distortion correction [25], from now on, referred to as "de-fisheye" camera frame. The choice of using fisheye camera data over RGB or depth, along with the distortion correction's fine-tuning, is aimed at searching for a balance between maintaining the contextual richness, processing speed, and precision to the human pose detection.

The first branch focuses on human-pose detection, where we calculate shoulder-elbow angles. The second branch extracts the segmented mask of the handed-over object. The third branch computes two intrinsic object properties. Then, each branch's concatenated outputs are further processed through dense layers. This leads to the final prediction of the robot's 14 arms-joint angles, defining an optimal pose for the robot to accomplish a handover avoiding physical contact with the person. Fig. 1 shows the model's full pipeline. Next, we will detail each branch's input calculation.

### A. Branch 1: Shoulder-Elbow Angles Calculation

The input data is calculated through a two-step process.

1. Human Body-Pose Keypoints Extraction and Arms Keypoints Correction: Over the de-fisheye camera frame, we detect the person's keypoints using Alphapose [26], an occlusion-robust, high-speed library [27] - [30] that has proven effective with an earlier version in our previous works [31]. This keypoint data feeds our `correct arm` function, part of the pose correction process. The heavy occlusions leads to misclassified keypoints. This function was tailored to adjust the positioning of the elbow keypoint within the arm's structure when there is a keypoints's misclassification. In a typical human arm, the shoulder-to-elbow and elbow-to-wrist distances should be roughly equal. Large discrepancies can indicate incorrect arm keypoint detection. To detect such discrepancies, the system applies a function to each arm to compute the Euclidean distances between these points:

$$d_{\text{elbow-shoulder}} = \sqrt{(x_{\text{elbow}} - x_{\text{shoulder}})^2 + (y_{\text{elbow}} - y_{\text{shoulder}})^2} \tag{1}$$

$$d_{\text{elbow-wrist}} = \sqrt{(x_{\text{elbow}} - x_{\text{wrist}})^2 + (y_{\text{elbow}} - y_{\text{wrist}})^2}, \tag{2}$$

and corrects the elbow's position if the absolute difference between these distances exceeds a predefined tolerance,

$$|d_{\text{elbow-shoulder}} - d_{\text{elbow-wrist}}| > \text{tolerance}, \tag{3}$$

to the midpoint between the shoulder and wrist,

$$(x_{\text{elbow\_c}} = \frac{x_{\text{shoulder}} + x_{\text{wrist}}}{2}, y_{\text{elbow\_c}} = \frac{y_{\text{shoulder}} + y_{\text{wrist}}}{2}). \tag{4}$$

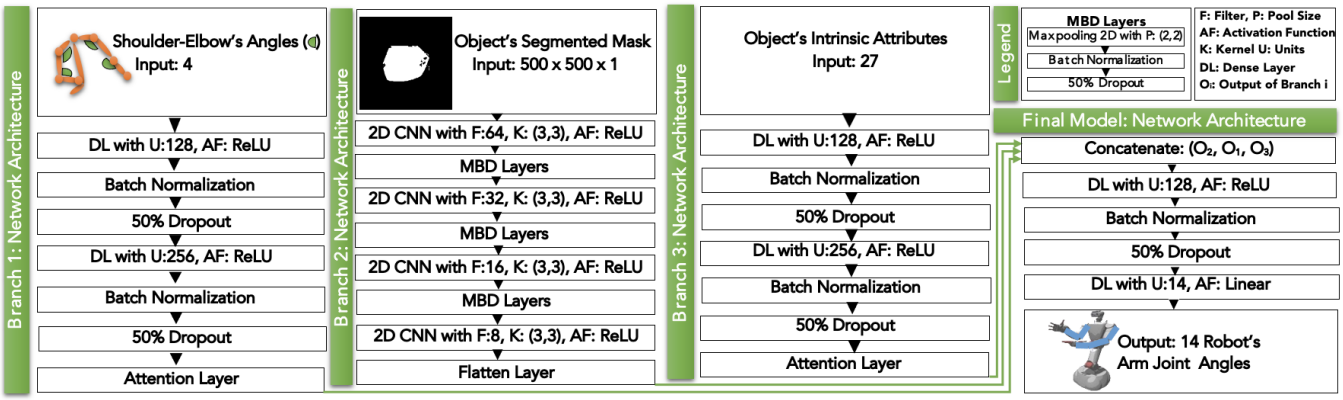


Fig. 1. Model Architecture Overview. From left to right – Branch 1: Shoulder-Elbow Angles, Branch 2: Refined Object's Segmented Mask, Branch 3: Object's Intrinsic Attributes, and Final Multimodal DNN Regression Model.

This correction is applied only when the shoulder-to-elbow and elbow-to-wrist distances are significantly unequal, suggesting possible incorrect keypoint detection. The outcome is a corrected arm in a relaxed pose. While this does not accurately represent the noisy arm's actual position, it allows the model to concentrate on the other arm which potentially holds accurate data, thereby reducing noise.

2. Shoulder-Elbow Angles Calculation: We calculate four specific angles that will be the input of this branch: the angles at the elbows and the shoulders for both the left and right arms. The angle calculation is performed as follows:

Given the pose keypoints $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$:

$$\mathbf{BA} = \mathbf{A} - \mathbf{B}, \tag{5}$$

$$\mathbf{BC} = \mathbf{C} - \mathbf{B}, \tag{6}$$

$$\text{norm}_{\mathbf{BA}} = \|\mathbf{BA}\| + \epsilon, \tag{7}$$

$$\text{norm}_{\mathbf{BC}} = \|\mathbf{BC}\| + \epsilon, \tag{8}$$

$$\cos(\theta) = \text{clip}\left(\frac{\mathbf{BA} \cdot \mathbf{BC}}{\text{norm}_{\mathbf{BA}} \cdot \text{norm}_{\mathbf{BC}}}, -1, 1\right), \tag{9}$$

$$\angle ABC = \arccos(\cos(\theta)) \cdot \frac{180}{\pi}. \tag{10}$$

Vectors $\mathbf{BA}$ and $\mathbf{BC}$ denote paths from point B to A and C, corresponding to elbow-to-shoulder and elbow-to-wrist, or shoulder-to-neck and shoulder-to-elbow. Their norms include a constant $\epsilon$ to avoid zero division. The cosine of angle $\theta$ is clipped between [-1, 1] to handle null or infinite values. Lastly, $\theta$ is changed to degrees.

### B. Branch 2: Refined Extraction of Object's Segmented Mask

We prioritize precise object mask extraction, helping to isolate the object and providing insights about it. Given the person holds the object, its position is typically near the extended shoulder midpoint vector's intersection. However, overlap with the person's mask might occur, especially for shorter objects. Making an effort to get noiseless data, the process is divided into several stages:

*1. Upper Body Mask Creation:*

To discern whether a segmented region represents a person or the held object, given the pose keypoints, we first focus on creating a mask for the upper body. The process is:

**Keypoint Selection:** A subset of detected keypoints representing the person's upper body, each one represented as a 2D point $(x, y)$.

**Polygon Formation:** Using these keypoints, we construct a polygon $P$ with vertices arranged to roughly outline the person's body.

**Binary Mask Generation:** After forming polygon $P$, a binary mask matching the input image's size is created. Pixels inside $P$ are set to 1, and those outside to 0:

$$M(x, y) = \begin{cases} 1 & \text{if } (x, y) \text{ is inside } P \\ 0 & \text{otherwise} \end{cases}$$

where $M$ is the binary mask, and $(x, y)$ the pixel coordinates.

*2. Lower Chest Point Calculation:*

To identify the object's position, determining the lower chest point is crucial. We assume that the object lies in the midpoint between one shoulder joint and the opposite arm's elbow joint, near the lower chest:

$$L = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right) \tag{11}$$

where $(x_1, y_1)$ and $(x_2, y_2)$ are the coordinates of one shoulder joint and the opposing arm's elbow joint.

*3. Object Mask Extraction using Semantic Segmentation:*

After detailed research, we opted for FastSAM [32] for semantic segmentation, aiming for a balance between precision and speed. Each segmented object's overlap with the person's mask is processed. Objects that overlap significantly with the person's mask are discarded, as they likely form part of the person. The nearest object to the earlier-determined lower chest point is identified as the object of interest.

*4. Final Mask Processing:* Our strategy emphasizes the object's intrinsic shape by isolating and standardizing it. In the final mask processing, we expand the mask with a set border size, ensuring no object parts are accidentally truncated. We then detect the bounding box around the non-zero mask regions and crop the mask. This cropped mask is resized for consistency in next stages. This cropping step ensures the model is not biased by the position of the

segmented mask in the image. The result is a mask of the held object.

### C. Branch 3: Intrinsic Object Attributes

Data related to the object being held is key information to enhance accuracy. Fourier descriptor offers a nuanced insight into the object's shape, encapsulating its periodic contour components. Concurrently, by associating the person's pose with these descriptors we can provide the object's spatial orientation. Merging these vectors, we have a descriptor that integrates the object's inherent shape and its pose.

*1. Fourier Descriptor Vector:* The Fourier Descriptor translates the object's final mask contour $S$, into the frequency domain, representing the object's shape. The descriptor $F$ is:

$$F(k) = \sum_{n=0}^{N-1} S(n) \times e^{-j(2\pi kn/N)} \tag{12}$$

where $k$ is the frequency component, $N$ is the number of points in the contour, and $j$ is the imaginary unit. We transformed the imaginary components into their real-valued equivalents for model compatibility.

*2. Pose Vector:* Derived from arm-related keypoints, the pose vector captures these points' distances from their centroid. For keypoint $K_i$ and centroid $Q$, distance $d_i$ is:

$$d_i = \sqrt{(K_{i,x} - Q_x)^2 + (K_{i,y} - Q_y)^2} \tag{13}$$

The pose vector is then the collection of these distances.

*3. Combined Descriptor:* We normalize the Fourier Descriptor Vector and concatenate it with the pose vector. This normalization ensures uniform magnitude, making the descriptor robust to object size variations. The final input for Branch 3 is this combined descriptor.

### D. Multimodal DNN Regression Model Final Concatenation

We concatenate results from the three branches, and after processing through dense layers, we generate 14 joint arm angles for the robot, 7 for each arm. This ensures a handover that avoids any physical contact with the person, even when hands or arms are occluded.

The model, trained over 100 epochs with a 0.001 learning rate, employs the Adam optimizer and a 16-batch size. Mean Squared Error (MSE) measures the prediction accuracy. Fig. 1 shows the inputs, network topography, and output of the full model and its branches.

## IV. EXPERIMENTAL EVALUATION AND RESULTS

In H2R handovers, the challenges occlusion presents, especially when dealing with occluded hands and arms, cannot be understated. Accurately predicting a safe pose for handovers in such cases is highly challenging. With this perspective, our experimental framework was designed to simulate varied scenarios where occlusion is prevalent.


Fig. 2. Experiment setting overview. From left to right – a. Robot's base poses. Poses vary in x and y coordinates based on the handed-over object, height, and depth variations. b. Robot's poses' heights and depths c. Experiments' objects and corresponding sizes (width × depth × height).

### A. Evaluation Framework

We used the humanoid robot Dry-AIREC [33], which will be referred to as AIREC from now on in this paper.

*Handover Pose Variations*

We set a baseline by testing four distinctive AIREC's handover poses, shown in Fig. 2.a. For each pose, the person's pose mirrors the robot's pose:

- One side: Right hand to the right, left hand down. See Fig. 2.a.1.
- One side mirrored: Left hand to the left, right hand down. See Fig. 2.a.2.
- Sides: Both hands sideways. See Fig. 2.a.3.
- Down: Both hands downwards. See Fig. 2.a.4.

### B. Object, Height and Depth Variations

To introduce variations to AIREC's poses, each was matched with five differently sized objects, shown in Fig. 2.c. Also, our evaluation with AIREC included height and depth variations: Low, Fig. 2.b.1, High, Fig. 2.b.2, Near, Fig. 2.b.3, and Far, Fig. 2.b.4. Overall, our test framework covered 80 pose variations per participant: 4 poses × 5 objects × 2 heights × 2 depths.

### C. Participants

10 participants for training data; 5 participants for testing.

### D. Experiments

*Experiment 1: Objective Testing*

This experiment evaluates each branch's impact on our model by measuring error. While other studies use just segmentation or keypoints, we aim to assess individual and combined branch performances. The sub-experiments are:

(1.1) Only the pose branch with final layers.

(1.2) Only the segmentation branch with final layers.

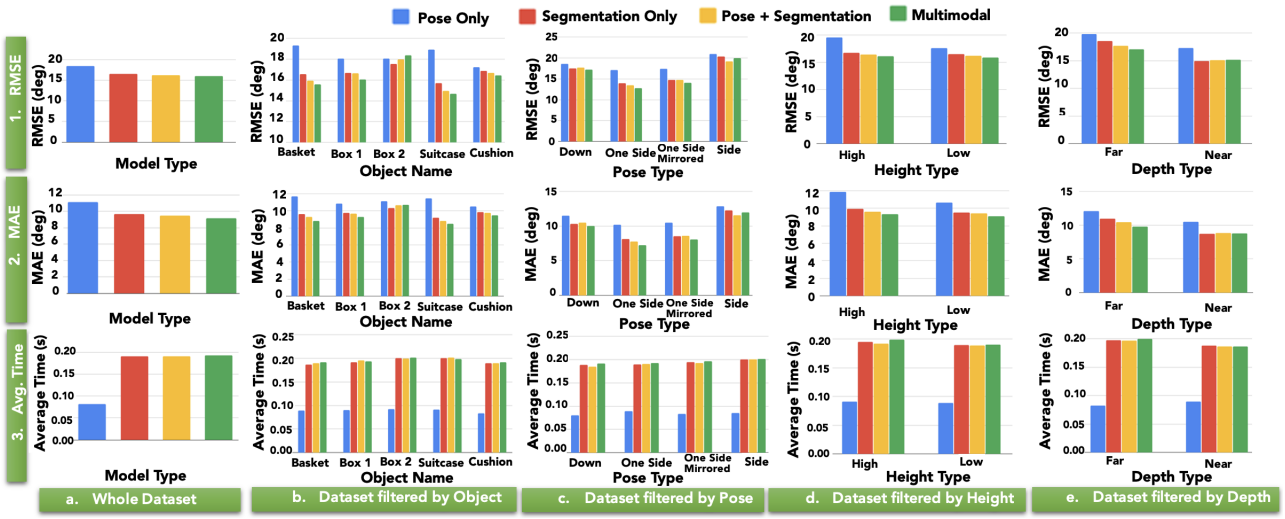


Fig. 3. Experiment 1: Objective Testing Results Graphs. From top to bottom: 1. RMSE Comparison, 2. MAE Comparison, 3. Average Time Comparison; From left to right: (a) Overall Comparison, (b) Comparison by Object, (c) Comparison by Pose, (d) Comparison by Height, (e) Comparison by Depth.

TABLE I

OBJECTIVE TESTING RESULTS OF MAE (DEGREES), RMSE (DEGREES), AND RT (SECONDS) ON THE WHOLE DATASET PER MODEL.

| Model | Objective Testing Results | | |
|---|---|---|---|
| | MAE | RMSE | RT |
| Pose Only | 11.16 | 18.42 | 0.08 |
| Segmentation Only | 9.67 | 16.57 | 0.19 |
| Pose + Segmentation | 9.51 | 16.26 | 0.19 |
| Multimodal (Final Proposal) | 9.19 | 16.00 | 0.19 |

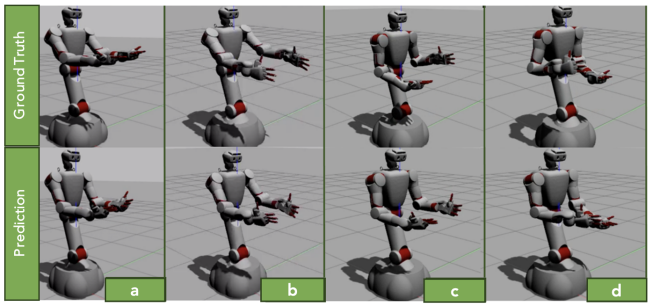


Fig. 4. Pose Predictions. Top to bottom: Ground Truth Pose. Model's Predicted Pose. (a) Pose: Down, Object: Suitcase, Height: High, Depth: Far. (b) Pose: Sides, Object: Basket, Height: High, Depth: Far. (c) Pose: One Side Mirrored, Object: Box 2, Height: Low, Depth: Near. (d) Pose: One Side, Object: Cushion, Height: Low, Depth: Near.

(1.3) Pose and segmentation branches with final layers.
(1.4) Final proposed pipeline combining pose, segmentation, and object intrinsic attributes.

Using the real AIREC robot, we captured ground truth data from its fisheye camera and its joints. For Experiment 1, predictions were simulated to mimic real AIREC behavior. Each participant executed all 80 poses, using 8 frames each, resulting in 6400 training and 3200 testing instances. Models were further analyzed by the different attributes related to the variations: object, pose, height, and depth.

*Experiment 2: Subjective Testing*

Here we evaluate the quality of the robot's predicted real-world pose attributes such as height, depth, hand width relative to object size, and pose with our proposed pipeline. Unlike Experiment 1, real AIREC was used for predictions. Each participant tested all 80 poses, totaling 400 instances.

### E. Evaluation Metrics

*Experiment 1: Objective Testing*

For an objective evaluation, we used the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as they are standard metrics for evaluating regression models:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

where $y_i$ represents the actual value, $\hat{y}_i$ denotes the predicted value, and $n$ is the total number of observations.

We selected RMSE for its sensitivity to larger errors and MAE for consistent penalties. A large RMSE-MAE difference suggests sporadic larger errors, while similar values show uniform errors. Also running time (RT) was measured.

*Experiment 2: Subjective Testing*

We subjectively scored the robot's final pose using four primary attributes: arm height (*height*), depth (*depth*), overall pose (*pose*), and arm width relative to the object (*object*). Each attribute correctly executed earned one point. Minor deviations, such as an arm height discrepancy of 2-5 cm or a sizeable hand tilt, deducted 0.25 points from that attribute. Entirely wrong attributes scored zero.

In both experiments, instances where AlphaPose failed to detect a great portion of the keypoints, or which AIREC cannot physically perform, were filtered out from the evaluation.

### F. Results and analysis

**Experiment 1:** All models' results on the complete dataset are in Table I and Fig. 3. a. Performance on datasets

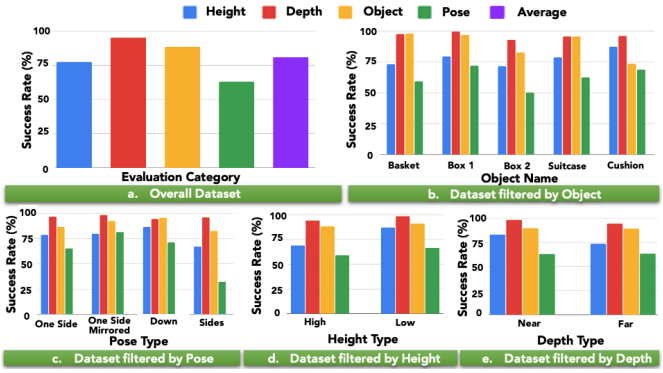


Fig. 5. Experiment 2: Subjective Testing Results Graphs. (a) Overall Comparison, (b) Comparison by Object, (c) Comparison by Pose, (d) Comparison by Height, (e) Comparison by Depth.

TABLE II

SUBJECTIVE TESTING SUCCESS RATE RESULTS IN % OF MULTIMODAL MODEL FOR HEIGHT, DEPTH, OBJECT, POSE AND AVERAGE OVERALL

| Exp | Subjective Testing | | | | |
|---|---|---|---|---|---|
| | Height | Depth | Object | Pose | Average |
| Multimodal | 77.41 | 95.25 | 88.36 | 62.74 | 80.94 |

filtered by the different attributes are in Fig. 3. b, c, d, e. Experiment's instances are shown in Fig. 4.

*Experiment 1.1: Pose Only Model – RT: 0.081 s*

The model, while fastest, shows the most error among all models. This model is challenged in estimating *High* over *Low* and *Far* over *Near*. The *Cushion* has the best MAE, due to its lower height since it is handover flat during the experiments, leading to fewer occluded keypoints.

*Experiment 1.2: Segmentation Only Model – RT: 0.19 s*

With a MAE of 9.67°, this model predicts better at a slightly increased run time. The degree by which RMSE exceeds MAE indicates occasional larger errors. It struggles with *High* and *Far* poses, likely from increased keypoints occlusions in the upper-chest during high hand positions, resulting in noisy data. The *Suitcase* performs best, while *Box 2*, which occludes most keypoints, has the highest MAE.

*Experiment 1.3: Pose + Segmentation Model – RT: 0.19 s*

Combining pose and segmentation shows improvements with same time. MAE gap for *High* and *Low* narrows, but *Far* still challenging in this model as well. *Box 2* keeps the highest MAE.

*Experiment 1.4: Multimodal Model – RT: 0.19 s*

Our full model excels with 9.19° of MAE, at a slightly longer time than the *Pose Only Model*. In general lines, it estimates *Low* better than *High* and struggles more with *Far* than *Near*. *Box 2* persistently challenges.

*Overall Discussion*

*Performance Metrics:* The Multimodal Model outperforms others in MAE and RSME.

*Prediction Challenges:* All models find the *Sides* pose challenging due to near-complete keypoints occlusion (Fig. 3. c) as well as happens with *Box 2* (Fig. 3. b) because of its height. Predictions for *High* and *Far* sometimes are challenging due to heavy occlusion (Fig. 3. d, e).

*Time Metrics:* Prediction times are mostly consistent across models, leading the *Pose Only* model.

*Insights:* A progression of 17.7% accuracy improvement from Pose Only to Multimodal model is clear. The latter balances time and accuracy.

**Experiment 2**: Success rates for the Multimodal Model on the complete dataset are in Table II and Fig. 5. a. Performance on datasets filtered by the different attributes are shown in Fig. 5. b, c, d, e.

*Depth Estimation:* The model excels in depth estimation, possibly due to clearer features or consistent data.

*Object Estimation:* The model adeptly distinguishes various object sizes, adjusting the robot's hands.

*Height Estimation:* Height success is notable but trails depth and object attributes, suggesting improvement areas.

*Pose Estimation:* Pose estimation is the most challenging for the model. The accuracy was notably affected by the *Side* pose (Fig. 5. c). Meanwhile, One Side, One Side Mirrored, and Down poses achieved generally good results.

*Conclusive Insights:* The model excels in depth and object estimations despite dataset challenges. While pose estimation can improve, it is noteworthy that even for humans, this dataset is very challenging. Incorporating a video-based system could enhance the detection capabilities. Also, given the numerous instances of highly tilted hands, integrating tactile data via touch sensors could significantly refine performance. This approach would mimic human adaptability in object handling, leading to a better interaction understanding.

*Limitations:* Our research operates under specific constraints. We assumed both of the person's hands were involved in the handover. Given AIREC's large hands, the robot mainly interacted with objects that fit its hand size. We did not account for unconventional holds, like holding an object from the side. AIREC's movements have inherent limitations, restricting our experiments' scope.

## V. CONCLUSION AND FUTURE WORKS

This study emphasizes the need for safe pose predictions in robot-human handovers when human hands and arms are largely occluded. Occlusions trigger unintended human contact, causing discomfort and safety concerns. While many models sidestep these circumstances, our DNN regression multimodal approach addresses them, searching for robot arm-joint angle predictions for contact-free handovers.

Our results reduce unintentional physical contact. Using shoulder-elbow angles, refined segmentation, and object attributes, our model guides the robot, preventing accidental touches even with heavy occlusions.

Future enhancements include tactile feedback in robot hands for tailored adjustments and a video-based system.

## ACKNOWLEDGMENT

This research was funded through JST Moonshot R&D.

We extend our whole gratitude to all those who participated in the experiments within this project.

## REFERENCES


[1] K. Strabala, M.K. Lee, A. Dragan, J. Forlizzi, S.S. Srinivasa, M. Cakmak, and V. Micelli, "Toward seamless human-robot handovers," in Journal of Human-Robot Interaction, vol. 2, no. 1, pp. 112–132, 2013.

[2] M. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer, "Human-robot interaction in handing-over tasks," in Proceedings of IEEE International Symposium on Robot and Human Interactive Communication, pp. 107–112, 2008.

[3] T.L. Chen, C.H. King, A.L. Thomaz, and C.C. Kemp, "Touched by a robot: An investigation of subjective responses to robot-initiated touch," in Proceedings of the International conference on Human-robot interaction, pp. 457–464, 2011.

[4] E.A. Sisbot, L.F. Marin-Urias, X. Broquere, D. Sidobre, and R. Alami, "Synthesizing robot motions adapted to human presence: A planning and control framework for safe and socially acceptable robot motions," in International Journal of Social Robotics, vol. 2, no. 3, pp. 329–343, 2010.

[5] P.A. Lasota, T. Fong, and J.A. Shah, "A survey of methods for safe human-robot interaction," in Foundations and Trends® in Robotics, vol. 5, no. 4, pp. 261–349, 2017.

[6] H. Nemlekar, D. Dutia, and Z. Li, "Object Transfer Point Estimation for Fluent Human-Robot Handovers," in International Conference on Robotics and Automation, pp. 2627–2633, 2019.

[7] V. Ortenzi, A. Cosgun, T. Pardi, W.P. Chan, E. Croft, and D. Kulić, "Object Handovers: A Review for Robotics," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1855–1873, 2021.

[8] J. Aleotti, V. Micelli, and S. Caselli, "Comfortable robot to human object hand-over," in IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 771–776, 2012.

[9] W. Yang, C. Paxton, M. Cakmak, and D. Fox, "Human Grasp Classification for Reactive Human-to-Robot Handovers," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11123–11130, 2020.

[10] W. Yang, C. Paxton, A. Mousavian, Y.-W. Chao, M. Cakmak, and D. Fox, "Reactive Human-to-Robot Handovers of Arbitrary Objects," in IEEE International Conference on Robotics and Automation (ICRA), pp. 3118–3124, 2021.

[11] J. Kwan, C. Tan, and A. Cosgun, "Gesture recognition for initiating human-to-robot handovers," in arXiv preprint arXiv:2007.09945, 2020.

[12] P. Rosenberger et al., "Object-Independent Human-to-Robot Handovers Using Real Time Robotic Vision," in IEEE Robotics and Automation Letters, vol. 6, no. 1, pp. 17–23, 2020.

[13] W. Wang, R. Li, Z.M. Diekel, Y. Chen, Z. Zhang, and Y. Jia, "Controlling Object Hand-Over in Human–Robot Collaboration Via Natural Wearable Sensing," in IEEE Transactions on Human-Machine Systems, 2018.

[14] M. Bianchi et al., "Touch-Based Grasp Primitives for Soft Hands: Applications to Human-to-Robot Handover Tasks and Beyond," in IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, pp. 7794-7801, 2018.

[15] H. Duan, P. Wang, Y. Li, D. Li and W. Wei, "Learning Human-to-Robot Dexterous Handovers for Anthropomorphic Hand," in IEEE Transactions on Cognitive and Developmental Systems, 2022.

[16] M.K.X.J. Pan, E.A. Croft, and G. Niemeyer, "Evaluating Social Perception of Human-to-Robot Handovers Using the Robot Social Attributes Scale (RoSAS)," in ACM/IEEE International Conference on Human-Robot Interaction, 2018.

[17] N. Marturi, M. Kopicki, A. Rastegarpanah et al., "Dynamic grasp and trajectory planning for moving objects," in Autonomous Robots, vol. 43, pp. 1241–1256, 2019.

[18] J. Male, G.A. Al, A. Shabani, and U. Martinez-Hernandez, "Multi-modal sensor-based human-robot collaboration in assembly tasks," in IEEE International Conference on Systems, Man, and Cybernetics, pp. 1266-1271, 2022.

[19] W. He, J. Li, Z. Yan and F. Chen, "Bidirectional Human–Robot Bimanual Handover of Big Planar Object With Vertical Posture," in IEEE Transactions on Automation Science and Engineering, vol. 19, no. 2, pp. 1180-1191, 2021.

[20] S.E. Ovur and Y. Demiris, "Naturalistic Robot-to-Human Bimanual Handover in Complex Environments Through Multi-Sensor Fusion," in IEEE Transactions on Automation Science and Engineering, 2023.

[21] Z. Yan, W. He, Y. Wang, L. Sun and X. Yu, "Probabilistic Motion Prediction and Skill Learning for Human-to-Cobot Dual-Arm Handover Control," in IEEE Transactions on Neural Networks and Learning Systems, 2022.

[22] L. v. der Spaa, M. Gienger, T. Bates and J. Kober, "Predicting and Optimizing Ergonomics in Physical Human-Robot Cooperation Tasks," in IEEE International Conference on Robotics and Automation (ICRA), pp. 1799-1805, 2020.

[23] T. Asfour et al., "ARMAR-6: A Collaborative Humanoid Robot for Industrial Environments," in IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids), pp. 447-454, 2018.

[24] N. Saito, T. Ogata, S. Funabashi, H. Mori, and S. Sugano, "How to Select and Use Tools? Active Perception of Target Objects Using Multimodal Deep Learning," in IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 2517-2524, 2021.

[25] E. Pereira, "duducosmos/defisheye: v1.2.1," Zenodo, v1.2.1, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8116565

[26] H.S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[27] B. Li, J. Zou, L. Wang, X. Li, Y. Li, R. Lei, and S. Sun, "The overview of multi-person pose estimation method," in Signal and Information Processing, Networking and Computers: Proceedings of the 5th International Conference on Signal and Information Processing, Networking and Computers, pp. 600-607, 2019.

[28] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," in Journal of Visual Communication and Image Representation, 2021.

[29] M.M. Desai, and H.K. Mewada, "Review on human pose estimation and human body joints localization," in International Journal of Computing and Digital Systems, 2021.

[30] S. Juraev, A. Ghimire, J. Alikhanov, V. Kakani and H. Kim, "Exploring Human Pose Estimation and the Usage of Synthetic Data for Elderly Fall Detection in Real-World Surveillance," in IEEE Access, vol. 10, pp. 94249-94261, 2022.

[31] C. Lollett, M. Kamezaki, and S. Sugano, "Driver's Drowsiness Classifier using a Single-Camera Robust to Mask-wearing Situations using an Eyelid, Lower-Face Contour, and Chest Movement Feature Vector GRU-based Model," in IEEE Intelligent Vehicles Symposium (IV), pp. 519-526, 2022.

[32] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast Segment Anything," arXiv preprint arXiv:2306.12156, 2023.

[33] T. Miyake, Y. Wang, G. Yan and S. Sugano, "Skeleton recognition-based motion generation and user emotion evaluation with in-home rehabilitation assistive humanoid robot," in IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), Ginowan, Japan, pp. 616-621, 2022.