

# Multi-Modal Multi-View Perception Feature Tracking for Handover Human Robot Interaction Applications

Chaitanya Bandi<sup>a</sup> and Ulrike Thomas

Robotics and Human Machine Interaction Lab, Technical University of Chemnitz, Chemnitz, Germany  
{chaitanya.bandi, ulrike.thomas}@etit.tu-chemnitz.de

Keywords: Hand Pose, Hand-Object Pose, Body Pose, Handover, Human-Robot Interaction.

Abstract: Object handover is a fundamental task in human-robot interaction (HRI) that relies on robust perception features such as hand pose estimation, object pose estimation, and human pose estimation. While human pose estimation has been extensively researched, this work focuses on creating a comprehensive architecture to track and analyze hand and object poses, thereby enabling effective handover state determination. We propose an end-to-end architecture that integrates unified hand-object pose estimation with hand pose tracking, leveraging an early and efficient fusion of RGB and depth modalities. Our method incorporates existing state-of-the-art techniques for human pose estimation and introduces novel advancements for hand-object pose estimation. The architecture is evaluated on three large-scale open-source datasets, demonstrating state-of-the-art performance in unified hand-object pose estimation. Finally, we implement our approach in a human-robot interaction scenario to determine the handover state by extracting and tracking the necessary perception features. This integration highlights the potential of the proposed system for enhancing collaboration in HRI applications.


## 1 INTRODUCTION

Bi-directional handovers in human-robot interaction (HRI) involve the mutual transfer of objects between humans and robots, encompassing both robot-to-human and human-to-robot interactions. This dynamic exchange requires the robot to not only execute precise physical actions but also understand contextual cues to coordinate effortlessly with the human partner. In both directions, the process depends on accurate perception, intention recognition, and synchronized motion planning. For instance, in a robot-to-human handover, the robot must identify when the human is ready to receive the object by analyzing body posture, hand position, and gaze direction. Conversely, in a human-to-robot handover, the robot must detect when the human intends to release the object by monitoring cues like grip loosening or object trajectory. For human-to-robot handovers, the robot's role involves anticipating the human's intent, adjusting its gripper orientation to align with the object's pose, and ensuring a firm grasp at the right moment. This direction of communication also needs to consider safety of human subject and avoid collisions.

In this work, we design a model and test specifically for human-to-robot complex handover scenario nevertheless the model is applicable for handover application. We can achieve the seamless interaction model by fusing vision-based 3D hand pose tracking, unified hand-object pose tracking, and body pose tracking. The core idea is to leverage the relationships between the tracked features (hand pose, body pose, and object pose) to recognize the interaction state (handover). The fusion of data from multiple modalities ensures robustness and reduces ambiguities in complex or cluttered environments.

Human body pose estimation is a well-researched area, with recent advancements achieving robust performance even under conditions of partial body occlusion. Given this progress, our focus is not on contributing to this domain but rather on leveraging existing state-of-the-art methods capable of real-time 3D human pose estimation.

The primary contribution of this work lies in achieving unified hand-object pose estimation. Leveraging Intel RealSense D415 cameras, which provide both RGB and depth data, we utilize multimodal input to enhance the accuracy of hand-object pose estimation. To surpass state-of-the-art performance, our approach integrates feature fusion from multimodal

<sup>a</sup>  <https://orcid.org/0000-0001-7339-8425>

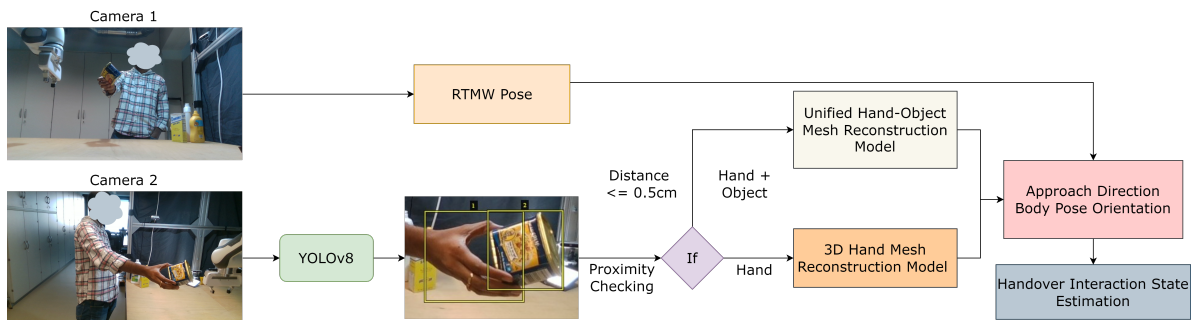


Figure 1: The proposed architecture provides an overview of a multi-camera setup designed for object handover interactions in a human-robot interaction environment. Among the three available camera views, only two are utilized, as the third view is deemed unnecessary for the handover application and is therefore excluded from the architecture.

data at early stages, along with cross-attention and self-attention mechanisms within the network. The complete process for estimating the handover interaction states is depicted in Figure 1. In the collaborative interaction scenario, two cameras are strategically positioned within the workspace. The first camera is placed to ensure a clear view of the subject’s upper body and face, capturing essential cues for interaction. The second camera is mounted on the left and right sides, respectively as illustrated in Figure 1. The RGB image from the first camera view is processed using RTMW3D (Jiang et al., 2024) to obtain 3D human pose estimation. Images from the second camera are fed into the YOLOv8 (Ultralytics, 2023) architecture to detect bounding boxes of the hand and identify the object regions, facilitating unified hand-object pose estimation.

After extracting the necessary information from YOLOv8 (Ultralytics, 2023), we proceed with two distinct tasks: 3D hand pose estimation and unified hand-object pose tracking. To avoid overload of loading all the models every time, we introduce proximity and geometric cues in addition to the bounding box intersection from object detection. For independent 3D hand mesh reconstruction, we adopt a process similar to the Vision Transformer (ViT) (Dosovitskiy et al., 2020) architecture. The input images are divided into patches and passed through a transformer encoder, which regresses the pose parameters of the MANO hand model.

Building on this foundation, our proposed contribution focuses on estimating the unified hand-object pose for real-time tracking in handover interaction scenarios. This unified approach enables precise and efficient tracking of both the hand and the object, enhancing the system’s reliability during dynamic interactions. For unified hand-object pose estimation, we rely on multi-modal data from intelrealsense camera. To reduce computational complexity in later stages, we first fuse the RGB and depth information using an

attention mechanism. The fused data is then passed through a unified backbone network based on a MobileNetV2 (Sandler et al., 2018) feature pyramid network (Lin et al., 2017) (FPN). ROI aligned information from hand and object are forwarded to separate hand and object encoders using attention mechanism which are later decoded using cross-attention to obtain outputs. From the hand decoder, MANO pose and shape parameters are obtained, which are then processed through the MANO model to reconstruct the 3D hand mesh. Meanwhile, the object decoder regresses 2D keypoint correspondences, which are matched with 3D keypoints to compute the object’s 6D pose using the Perspective-n-Point (PnP) (Lepetit et al., 2009) algorithm.

## 2 RELATED WORK

This work focuses on three perception features: 3D body pose estimation, 3D hand mesh recovery, and 6D object pose estimation. We then perform unified hand-object pose estimation. For 3D body pose estimation we rely on existing state-of-the-art works. In this section, we discuss recent work related to hand mesh reconstruction and unified hand-object pose estimation.

### 2.1 3D Hand Mesh Reconstruction

The work introduced in (Zimmermann and Brox, 2017) one of the first deep learning frameworks for 3D hand pose estimation from RGB images. It employed a keypoint-based regression method to predict the 3D pose and introduced a dataset to facilitate this task. The model demonstrated robustness in single-view hand pose estimation but lacked the ability to model the hand’s detailed shape. Hand PointNet (Ge et al., 2018) utilized point clouds to estimate hand poses directly, avoiding reliance on inter-

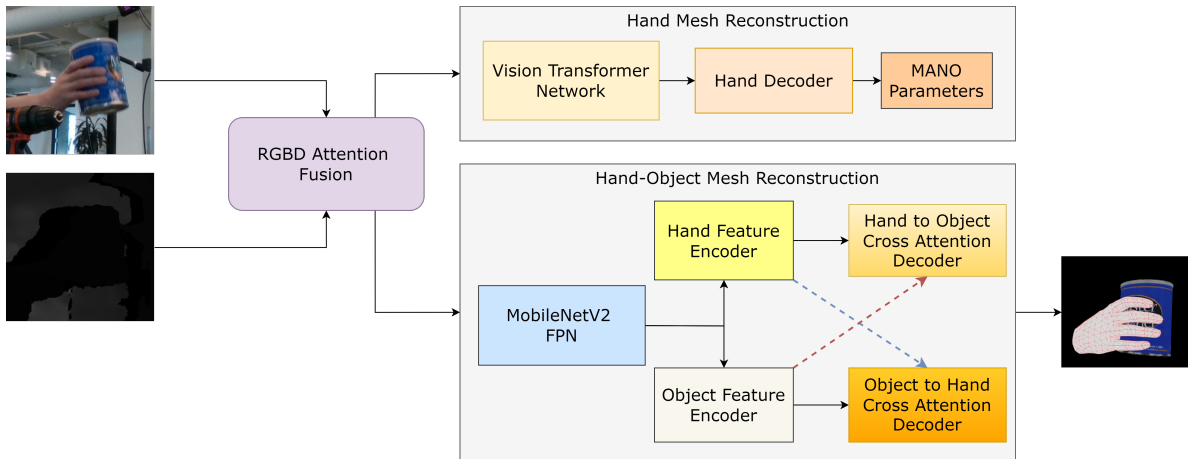


Figure 2: The architecture of the proposed 3D hand pose estimation and unified hand-object pose estimation.

mediate 2D representations. By operating on point sets, this method was robust to occlusions and noise. The approach effectively captured geometric features but required depth input, limiting its applicability in RGB-only scenarios. The work introduced in (Baek et al., 2019) contains a neural rendering framework that iteratively refines hand pose estimations by comparing the rendered hand image with the observed input. This iterative approach improved pose accuracy and made the network more resilient to occlusions and ambiguous poses.

In contrast, to achieve accurate hand pose estimation, many works adopt a model-based method utilizing the differentiable MANO model introduced in (Romero et al., 2017). This approach enables the simultaneous estimation of 3D hand pose and shape, represented as a detailed mesh. The authors in (Ge et al., 2019) propose a method for estimating both hand shape and pose by predicting the parameters of the MANO hand model. By leveraging the differentiable nature of MANO (Romero et al., 2017), the method reconstructed realistic hand meshes while maintaining computational efficiency. Later many research works such as (Cai et al., 2019), (Moon et al., 2020), (Park et al., 2022a), (Pavlakos et al., 2024) were developed on MANO based hand model with different backbones.

## 2.2 Unified Hand-Object Pose Estimation

HOPE-Net (Wang et al., 2022) integrates hand and object pose estimation into a unified framework using a shared latent space. The network employs a disentangled representation for joint and independent pose estimations of hands and objects. The use of multi-task learning allows simultaneous hand and ob-

ject pose prediction, resulting in efficient processing. A key advantage of this approach is its ability to handle occlusions effectively due to the shared feature space between hands and objects, enabling robust estimation under challenging conditions.

HOISDF (Xu et al., 2022) employs global signed distance fields (SDFs) for simultaneous learning of hand and object shapes. It leverages SDFs to encode mutual constraints between hands and objects, focusing on global plausibility rather than fine-grained details. The approach includes a U-Net-based encoder-decoder for hierarchical feature extraction and SDF decoders for estimating distances to hand and object surfaces. This method excels in handling occlusions and capturing robust global information. Later many works improved and extended based of SDFs (Chen et al., 2022b), (Chen et al., 2023).

The framework (Qu et al., 2023) combines neural rendering and model-based fitting for joint hand-object pose estimation. The method uses offline learning to build generative implicit models for hand and object geometry. During online inference, rendering-based model fitting refines poses under geometric constraints. A key advantage is the ability to generate smooth and stable pose sequences for videos, reducing jitter and improving temporal consistency.

Dense Mutual Attention (Zhao et al., 2023) introduces a novel approach for estimating 3D hand-object poses by explicitly modeling fine-grained interactions using a dense mutual attention mechanism. This method aims to improve the physical plausibility and quality of pose estimations while maintaining real-time inference speed. The approach constructs hand and object graphs based on their mesh structures. Each node in the hand graph aggregates features from all nodes in the object graph through learned attention weights, and vice versa. This dense

interaction captures detailed dependencies between the hand and object, enhancing interaction modeling.

HFL-Net (Wang et al., 2023) presents a framework that integrates hand and object pose estimation into a unified process by focusing on capturing mutual constraints and interactions. The core contribution lies in a harmonious feature learning strategy, which emphasizes extracting joint features that represent both the hand and the object while maintaining their distinct identities. The approach leverages advanced neural architectures to encode fine-grained hand-object relationships and applies attention mechanisms to dynamically prioritize critical interaction regions. Experimental results show that this method achieves superior accuracy and robustness, particularly in scenarios involving occlusions or complex hand-object interactions, making it well-suited for real-world applications in human-robot collaboration and augmented reality.

The work (Hoang et al., 2024) proposes a novel approach to hand-object pose estimation that combines multiple modalities, such as RGB and depth images, to enhance the accuracy and robustness of the estimation. The method employs adaptive fusion techniques to intelligently combine information from different sensory inputs, optimizing the model’s ability to handle varying input conditions. The core innovation of this work lies in the introduction of interaction learning, which models the dynamic interactions between the hand and object to improve pose predictions, especially in challenging scenarios involving complex hand-object interactions.

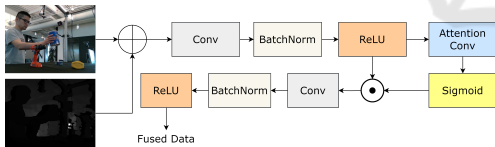


Figure 3: Early efficient RGBD fusion. Attention-based fusion of RGB and depth image.

### 3 METHODOLOGY

In this work, we aim to develop a comprehensive model for object handover with a strong focus on safety. To achieve this, we utilize perception feature extraction networks capable of real-time operation. These include 3D human body pose estimation, 3D hand pose estimation, and unified hand-object pose estimation. Rather than designing all components from scratch, we leverage existing state-of-the-art methods. Specifically, for 3D human body pose estimation, we adopt the recently introduced RTMW

model (Jiang et al., 2024), which offers high accuracy and real-time performance, making it suitable for multi-person whole-body pose estimation scenarios. The model processes input images to detect multiple people and their detailed poses simultaneously, even in crowded or dynamic scenarios. By balancing speed and precision, RTMW demonstrates robust performance in real-time applications such as sports analytics, augmented reality, and human-robot interaction. Its real-world usability is enhanced by its ability to handle occlusions and variations in body configurations.

To optimize system performance, we chose to track features continuously, except for 3D human body pose estimation, to avoid unnecessary computational overhead. To minimize redundant processing, we implemented a hand-object proximity detection method to bypass 6D object pose estimation when it is not required. The proximity detection relies on two simple but effective approaches. The first approach involves monitoring the intersection of bounding boxes over time, as detected using YOLOv8. However, due to the cluttered arrangement of objects, multiple items may overlap within certain durations. To address this, we incorporated additional criteria, including depth proximity and geometric cues. Specifically, we check if the depth of both the hand and object is within close range (less than 0.5 cm) and overlaps persist for a set number of frames. When these conditions are met, we assume the object is in the human hand and trigger unified hand-object pose estimation. Otherwise, we only compute 3D hand pose reconstruction, reducing unnecessary computational load. The complete architecture is illustrated in Figure 2

The process begins by passing the RGB image through the YOLOv8 (Ultralytics, 2023) object detection model, which has been retrained for this work to detect YCB (Calli et al., 2017) objects and human hands. The model outputs bounding boxes for all YCB (Calli et al., 2017) objects and the hand, if present. Using this bounding box information, proximity is assessed based on depth and geometric cues. The RGB and depth images are then cropped to focus on either the hand pose or the unified hand-object pose, guided by the bounding box and proximity data. To ensure consistency, the cropped regions maintain their original aspect ratios and are resized to dimensions of  $224 \times 224 \times 3$  for the RGB image and  $224 \times 224 \times 1$  for the depth image.



### 3.1 RGBD Attention-Based Fusion

The initial step involves performing efficient early-stage RGB-D attention fusion. Direct fusion at this stage often results in information loss, so we employ an attention mechanism with learnable parameters to selectively integrate critical depth information into the model. This approach eliminates the need for additional networks, such as PointNet++ (Qi et al., 2017) or CNNs, which can introduce latency and hinder real-time inference. By integrating depth information efficiently, the system maintains high performance without compromising real-time processing capabilities. The process of Efficient RGBD fusion with attention mechanism is illustrated in Figure 3.

### 3.2 Hand Mesh Reconstruction Network

The backbone of the Hand Mesh Reconstruction (HMR) network is the vision transformer (ViT) (Dosovitskiy et al., 2020). we follow the similar process as the work (Pavlakos et al., 2024) to encode the hand features using vision transformer. The encoded features are then decoded to obtain the mano parameters. The MANO parameters are then forwarded to the MANO model to obtain 3D hand mesh and 3D hand joint locations.

#### 3.2.1 MANO Parametric Model

The MANO (Model-based Articulated hand tracking using a Nonlinear representation) hand parametric model is a statistical 3D model that represents human hand shapes and poses in a compact and efficient form. It is an adaptation of the SMPL (Skinned Multi-Person Linear) model, customized for hand pose and shape estimation. MANO parameterizes a 3D hand mesh using two components: pose parameters  $\theta \in \mathbb{R}^{K \times 3}$ , which control the rotation of  $K = 16$  joints in axis-angle format, and shape parameters  $\beta \in \mathbb{R}^N$ , which define individual hand shape variations based on  $N = 10$  principal components derived from a dataset of scanned hand shapes.

The MANO model outputs a triangulated 3D mesh with  $V = 778$  vertices connected by faces to form the hand's surface. The pose and shape parameters ( $16 \times 3 + 10 = 58$ ) are combined with a linear blend skinning algorithm to deform the mesh according to the desired articulation and morphology. This allows for realistic and anatomically plausible hand representations. A key feature of MANO is its ability to directly regress joint locations, making it suitable for both hand pose estimation and applications requiring high-quality hand-object interaction modeling.

### 3.3 Architecture

The input to the proposed architecture is a fused image of size  $224 \times 224 \times 3$ . This image is divided into 16 non-overlapping patches, which are then forwarded to the Vision Transformer (ViT) (Dosovitskiy et al., 2020) architecture to encode hand-specific features. The ViT-H backbone outputs a sequence of tokens that encapsulate the encoded hand information. To decode these features, a transformer decoder is employed. It processes the output tokens from the ViT and regresses the MANO parameters similar to the work in (Pavlakos et al., 2024). These parameters are subsequently passed to the MANO model, which generates 3D hand joint locations and 3D mesh vertices.

### 3.4 Hand-Object Mesh Reconstruction Network

Once the proximity is triggered, the system performs unified hand-object pose estimation. The fused image  $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{224 \times 224 \times 3}$  is forwarded as input to the Hand-Object Mesh Reconstruction (HOMR) network. For feature extraction from  $\mathbf{F}_{\text{fused}}$ , the MobileNetV2 FPN (Lin et al., 2017) architecture is utilized, which ensures computational efficiency while capturing rich feature representations.

MobileNetV2 (Sandler et al., 2018) is a lightweight and efficient convolutional neural network architecture designed for mobile and embedded vision applications. The core innovation in MobileNetV2 is the use of inverted residual blocks and linear bottlenecks. MobileNetV2 FPN (Lin et al., 2017) (Feature Pyramid Network) combines the efficient MobileNetV2 backbone with the multi-scale feature processing capabilities of FPN for improved object detection and segmentation tasks. In the FPN architecture, features from different stages of the network are combined to form a feature pyramid, allowing the model to leverage both high-resolution and high-level semantic information. Once the hand-object features from MobileNetV2 FPN are extracted, the region of interest (ROI) aligned information of each of the features are extracted. The ROI aligned features are then forwarded to the deformable transformer (Zhu et al., 2021) (DETR) encoder for both hand and object.

The input to the deformable multi-headed transformer attention is a feature map of size  $7 \times 7 \times 256$ , which corresponds to a spatial resolution of  $7 \times 7$  with 256 feature channels. This input is first flattened into a sequence of size  $49 \times 256$ , where 49 is the total number of spatial tokens ( $7 \times 7$ ). A learn-

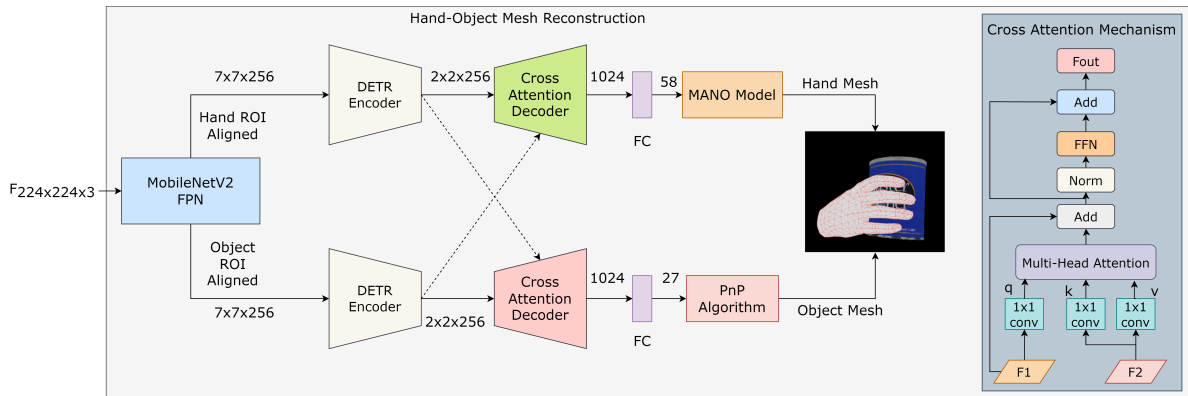


Figure 4: The architecture of the Hand-Object Mesh Reconstruction (HOMR) network. This network employs a MobileNet-FPN backbone, deformable transformers, and a cross-attention mechanism to achieve unified hand-object pose estimation.

able positional embedding of size  $49 \times 256$  is added to the input sequence to incorporate spatial information. The input is then projected into query, key, and value tensors, each of size  $49 \times 256$ . These tensors are reshaped for multi-head attention into the dimensions  $B \times \text{num\_heads} \times 49 \times \text{head\_dim}$ , where  $\text{head\_dim} = \frac{\text{embed\_dim}}{\text{num\_heads}}$ . Offsets for deformable sampling are predicted through a linear layer, producing a tensor of size  $49 \times \text{num\_heads} \times 2$  for each spatial token. These offsets dynamically determine the sampling locations within the feature map and  $B$  is the batch size, number of heads is 8, and head dimension is 128.

The attention mechanism computes attention scores of size  $B \times \text{num\_heads} \times 49 \times 49$  using scaled dot-product attention. These scores are used to compute a weighted sum of the value tensor, resulting in an attended output of size  $B \times \text{num\_heads} \times 49 \times \text{head\_dim}$ . The outputs from all heads are concatenated back into the shape  $B \times 49 \times 256$ . After applying a final projection layer, the output is reshaped back into the original spatial resolution of  $7 \times 7 \times 256$ . To meet the desired output size of  $2 \times 2 \times 256$ , bilinear interpolation is applied to downsample the spatial dimensions from  $7 \times 7$  to  $2 \times 2$ , while preserving the 256 feature channels. The final output is a tensor of size  $B \times 2 \times 2 \times 256$ .

The extracted features are forwarded to the cross-attention decoder layer, where the query for the hand decoder consists of the object-encoded information, while the query for the object decoder is derived from the hand decoder’s features. After performing the cross-attention mechanism, a fully connected layer is employed to generate the respective output features. Specifically, the hand decoder outputs 58 parameters representing the MANO (Romero et al., 2017) hand model, and the object decoder outputs 27 features corresponding to 9 keypoints, each with 3 dimensions.

For the object keypoints, the first two dimensions  $(x, y)$  represent the 2D location of the keypoint, while the third dimension represents the confidence score of the keypoint being accurately predicted.

The predicted MANO parameters are subsequently passed into the MANO model to compute the hand mesh vertices and the 3D keypoints of the hand. Similarly, using the predicted 2D keypoints and known 3D correspondences of the object, the 6D pose of the object is estimated by solving the Perspective-n-Point (Lepetit et al., 2009) (PnP) problem iteratively. This approach ensures accurate estimation of both the hand’s mesh structure and the object’s pose in a unified framework. The complete HOMR network is illustrated in Figure 4

### 3.4.1 Loss Function

To train the network, we define a composite loss function that minimizes the L2 distances between the predicted and ground truth values of  $\mathbf{H}$  (object keypoints),  $\theta$  (pose parameters),  $\beta$  (shape parameters),  $\mathbf{V}$  (3D vertices), and  $\mathbf{J}$  (3D joints). The total loss for hand pose estimation, denoted as  $\mathcal{L}_{\text{overall}}$ , is formulated as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{Obj}} + \mathcal{L}_{3\text{D}} + \mathcal{L}_{\text{MANO}}$$

The term  $\mathcal{L}_{\text{Obj}}$  corresponds to the L2 loss for 2D object keypoint location predictions, ensuring accurate localization of keypoints in the 2D space:

$$\mathcal{L}_{\text{Obj}} = \sum_{i=1}^K \left\| \mathbf{o}_i - \mathbf{o}_i^{\text{gt}} \right\|_2^2$$

Here,  $\mathbf{o}_i$  and  $\mathbf{o}_i^{\text{gt}}$  denote the predicted and ground truth keypoints for the  $i$ -th keypoint, respectively, and  $K$  is the total number of keypoints.

The term  $\mathcal{L}_{3\text{D}}$  accounts for the L2 loss between the predicted and ground truth 3D vertices ( $\mathbf{V}$ ) and joint

Table 1: Comparison with the state-of-the-art on the FreiHAND dataset.

Method	PA-MPJPE $\downarrow$	PA-MPVPE $\downarrow$	F@5 $\uparrow$	F@15 $\uparrow$
I2UV-HandNet (Chen et al., 2021)	6.7	6.9	0.707	0.977
METRO (Lin et al., 2021)	6.5	6.3	0.731	0.984
Tang et al. (Tang et al., 2021)	6.7	6.7	0.724	0.981
MobRecon (Chen et al., 2022a)	5.7	5.8	0.784	0.986
AMVUR (Jiang et al., 2023)	6.2	6.1	0.767	0.987
HaMeR (Pavlakos et al., 2024)	6.0	5.7	0.785	0.990
<b>Ours</b>	5.7	5.6	0.797	0.990

Table 2: Comparison with the state-of-the-art on the HO-3D dataset.

Method	PA-MPJPE $\downarrow$	PA-MPVPE $\downarrow$	F@5 $\uparrow$	F@15 $\uparrow$
Liu et al. (Liu et al., 2021)	9.9	9.5	0.528	0.956
HandOccNet (Park et al., 2022b)	9.1	8.8	0.564	0.963
I2UV-HandNet (Chen et al., 2021)	9.9	10.1	0.500	0.943
Hampali et al. (Hampali et al., 2020)	10.7	10.6	0.506	0.942
Hasson et al. (Hasson et al., 2019)	11.0	11.2	0.464	0.939
METRO (Lin et al., 2021)	10.4	11.1	0.484	0.946
MobRecon (Chen et al., 2022a)	9.2	9.4	0.538	0.957
AMVUR (Jiang et al., 2023)	8.3	8.2	0.608	0.965
HaMeR (Pavlakos et al., 2024)	7.7	7.9	0.635	0.980
<b>Ours</b>	7.7	7.8	0.635	0.978

coordinates ( $\mathbf{J}$ ), promoting accurate 3D mesh reconstruction and joint localization:

$$\mathcal{L}_{3D} = \|\mathbf{V} - \mathbf{V}^{\text{gt}}\|_2^2 + \|\mathbf{J} - \mathbf{J}^{\text{gt}}\|_2^2$$

The term  $\mathcal{L}_{\text{MANO}}$  imposes L2 losses on the MANO shape parameters ( $\beta$ ) and pose parameters ( $\theta$ ), ensuring accurate estimation of the hand’s pose and shape:

$$\mathcal{L}_{\text{MANO}} = \|\beta - \beta^{\text{gt}}\|_2^2 + \|\theta - \theta^{\text{gt}}\|_2^2$$

Here,  $\mathbf{V}^{\text{gt}}$ ,  $\mathbf{J}^{\text{gt}}$ ,  $\beta^{\text{gt}}$ , and  $\theta^{\text{gt}}$  represent the ground truth 3D vertices, joint coordinates, shape parameters, and pose parameters, respectively. The combined loss  $\mathcal{L}_{\text{overall}}$  ensures robust hand pose, object pose and mesh estimation by optimizing both spatial accuracy and parametric consistency.

## 4 EXPERIMENTATION

This section presents a comprehensive evaluation of the proposed approach on three widely used RGB-D datasets: FreiHand (Only hand interactions) (Zimmermann et al., 2020), HO-3D (Zhang et al., 2020) and DexYCB (Mishra et al., 2020) (these contain hand-object interactions). These datasets are designed to reflect realistic hand pose scenarios, offering a robust benchmark for assessing the performance of hand pose estimation techniques in practical settings. Our analysis includes a detailed comparison with leading RGB-based and depth-based methods, allowing us to effectively validate the robustness and accuracy of our approach against state-of-the-art alternatives.

### 4.1 Implementation Details

For hand and YCB (Calli et al., 2017) object detection, we utilize bounding box annotations from all three datasets. While YOLOv8 (Ultralytics, 2023) is employed for detection tasks, we do not conduct an extensive evaluation of its transfer learning performance, as this aspect has already been thoroughly explored in prior studies.

The HMR network was trained for 70 epochs using the Adam optimizer. To improve generalization, a weight decay of  $5 \times 10^{-4}$  was applied, which was scheduled to update every 10 epochs. During training, the aspect ratios of all input images were preserved to ensure realistic representations of hand poses. The images were resized to a resolution of  $224 \times 224$  pixels while maintaining their original proportions.

The HOMR architecture was trained under a setup similar to the hand mesh reconstruction network, with a few adjustments. The HOMR model was trained for 100 epochs using the Adam optimizer, with a weight decay of  $5 \times 10^{-4}$  applied every 10 epochs. The training process also preserved the aspect ratios of all input images, which were resized to a fixed resolution of  $224 \times 224$  pixels to align with the network’s input requirements while retaining critical spatial information.

### 4.2 Datasets and Evaluation Metrics

The **HO-3D (Hand-Object 3D) dataset** (Zhang et al., 2020) is a publicly available resource designed for research in hand pose estimation and hand-object interaction analysis. It provides a comprehensive collection of RGB-D images capturing real-world interactions between hands and various objects. The dataset emphasizes scenarios involving natural hand poses while manipulating objects, making it highly suitable for studying complex hand-object interactions.

The **DexYCB dataset** (Mishra et al., 2020) is a comprehensive resource designed for studying hand-object interactions, particularly focusing on 6D object pose estimation and 3D hand pose estimation. It features a diverse set of RGB-D sequences capturing real-world interactions with objects from the YCB object set, a widely used benchmark for robotic manipulation research.

The **FreiHand dataset** (Zimmermann et al., 2020) is a high-quality resource for advancing research in 3D hand pose estimation and shape reconstruction. It is specifically designed to provide challenging and realistic scenarios, featuring diverse hand poses captured from real-world settings. The dataset



Figure 5: The qualitative samples of the DexYCB dataset obtained from the HOMR network.

includes 134,000 samples collected from 32 unique subjects, ensuring significant variation in hand shape, size, and pose.

For FreiHand dataset and HO-3D dataset, we report the F-scores, the mean joint error (PAMPJPE), and the mean mesh error (PAMPVPE) in millimeters after performing Procrustes alignment. For DexYCB dataset, we report non procrustes aligned MPJPE. For 6D object pose estimation, we compute ADD-S (Average Distance of Model Points with Symmetry). The Average Distance of Model Points (ADD) is a widely used metric to evaluate the accuracy of 6D object pose estimation. It calculates the mean distance between corresponding 3D points of the ground truth object model and the estimated object model under a predicted pose. In particular, for symmetric objects, the ADD-s variant is employed to handle symmetry. ADD-s is defined as:

$$\text{ADD-s} = \frac{1}{|M|} \sum_{\mathbf{x} \in M} \min_{\mathbf{y} \in M} \|(\mathbf{R}\mathbf{x} + \mathbf{t}) - (\mathbf{R}_{\text{gt}}\mathbf{y} + \mathbf{t}_{\text{gt}})\|, \quad (1)$$

where  $M$  is the set of 3D model points,  $\mathbf{R}$  and  $\mathbf{t}$  are the predicted rotation and translation of the object, and  $\mathbf{R}_{\text{gt}}$  and  $\mathbf{t}_{\text{gt}}$  are the ground truth rotation and translation. The term  $\min_{\mathbf{y} \in M}$  accounts for symmetry by finding the closest point  $\mathbf{y}$  in the model set  $M$  for each transformed point  $\mathbf{x}$ .

ADD-S measures the average alignment error between the predicted and ground truth poses. Lower ADD-s values indicate more accurate pose predictions, making it a key metric for evaluating object pose estimation in scenarios involving symmetrical objects.

Table 3: Performance comparison with state-of-the-art methods on hand pose estimation on the HO3D dataset.

Method	PA-MPJPE↓	PA-MPVPE↓	F@5†	F@15†
Hasson et al. (Hasson et al., 2020)	11.4	11.4	42.8	93.2
Hasson et al. (Hasson et al., 2019)	11.0	11.2	46.4	93.9
Hampali et al. (Hampali et al., 2020)	10.7	10.6	50.6	94.2
Liu et al. (Liu et al., 2021)	10.1	9.7	53.2	95.2
HFL-Net (Wang et al., 2023)	8.9	8.7	57.5	96.5
Ours	8.87	8.79	58.5	96.9

Table 4: Performance comparison on the object pose estimation task for Cleanser, Bottle, and Can categories.

Method	Cleanser†	Bottle†	Can†	Average†
Liu et al. (Liu et al., 2021)	88.1	61.9	53.0	67.7
HFL-Net (Wang et al., 2023)	81.4	87.5	52.2	73.3
Ours	85.4	86.3	51.4	74.3

### 4.3 Comparison to the State-of-the-Art

In this study, we implement two distinct networks: HMR and HOMR. For the model trained with the HMR network, we evaluate and compare the 3D hand pose and 3D mesh errors against state-of-the-art methods using the HO-3D and FreiHand datasets. For the model trained with the HOMR network, which is a unified framework, we perform comparisons on both the HO-3D (Zhang et al., 2020) and DexYCB (Mishra et al., 2020) datasets, benchmarking the results against state-of-the-art techniques.

#### 4.3.1 HMR Network Comparisons

Initially, we trained the HMR network using the FreiHand dataset (Zimmermann et al., 2020). A detailed comparison with state-of-the-art methods on the FreiHand dataset is provided in Table 1. The evaluation follows the standard protocol, with metrics reported for assessing 3D joint and 3D mesh accuracy. The PA-MPVPE and PA-MPJPE metrics are presented in millimeters and low the error higher the 3D pose ac-



Table 5: Comparison of hand pose estimation results with state-of-the-art methods on the DexYCB dataset.

Method	MPJPE↓	PAMPJPE↓	RGB-D
Hasson (Hasson et al., 2019)	17.6	-	RGB
Hasson (Hasson et al., 2020)	18.8	-	RGB
Tze et al. (Tse et al., 2022)	15.3	-	RGB
Liu et al. (Liu et al., 2021)	15.27	6.58	RGB
DMA (Zhao et al., 2023)	12.7	-	RGB
HFL-Net (Wang et al., 2023)	12.56	5.47	RGB
Hoang et al. (Hoang et al., 2024)	12.15	4.54	RGBD
<b>Ours</b>	11.9	4.61	RGBD

Table 6: Performance comparison of the object pose estimation on DexYCB dataset.

Method	AUC↑	ADD-S < 2cm↑
Hasson et al. (Hasson et al., 2019)	0.69	0.65
Hasson et al. (Hasson et al., 2020)	0.75	0.71
Cao et al. (Cao et al., 2021)	0.70	0.72
Chen et al. (Chen et al., 2022b)	0.72	0.74
Chen et al. (Chen et al., 2023)	0.75	0.77
Hoang et al. (Hoang et al., 2024)	0.84	0.82
<b>Ours</b>	0.86	0.83

curacy.

To assess the performance of our model on hand-object datasets, we further evaluate the HMR network using the HO-3D (Zhang et al., 2020) dataset. Consistent with the evaluation on the FreiHand dataset, we report PA-MPVPE and PA-MPJPE metrics, both expressed in millimeters. A detailed comparison of the results is presented in Table 2. From these comparisons, it is evident that our model achieves error rates comparable to HaMeR (Pavlakos et al., 2024). The slight differences in error values can be attributed to our use of a fused RGB and Depth image approach, where the depth fusion introduces marginal variations in performance.

### 4.3.2 HOMR Network Comparisons

We evaluate the performance of the HOMR network on two datasets: HO-3D (Zhang et al., 2020) and DexYCB (Mishra et al., 2020). The evaluation includes both hand pose estimation errors and object pose estimation metrics. Our proposed HOMR network is compared against existing state-of-the-art methods for hand-object pose estimation on HO-3D dataset. The detailed results are presented in Table 3. From the comparison, it is evident that the F-scores and mesh error (PA-MPVPE) achieved by our method surpass those of the current state-of-the-art approaches. Additionally, the joint error (PA-MPJPE) is slightly lower than that of the most recent state-of-the-art methods.

Limited comparisons regarding object pose estimation on the HO-3D (Zhang et al., 2020) dataset have been presented in prior works. Two studies reported the ADD-0.1D error for four objects from the YCB dataset (Calli et al., 2017). For a fair evaluation,

we compare these specific objects, and the results are detailed in Table 4. From the comparison, it can be observed that the average object pose estimation error in our method is slightly higher than the state-of-the-art methods.

Similarly, limited works have reported hand-object pose estimation performance on the DexYCB (Mishra et al., 2020) dataset. Based on our research, we compare the results with the state-of-the-art methods. The reported values for hand pose estimation are presented in Table 5. For object pose evaluation, not all works use same metrics so we compute ADD-S because it was mostly mentioned by research works. The ADD-S and area under the curve (AUC) for object pose evaluation is mentioned in the Table 6. Few qualitative samples obtained from HOMR network on DexYCB dataset is illustrated in Figure 5. The primary limitation of this work arises when the hands are significantly occluded, leading to failures in accurately estimating hand joints.

## 5 CONCLUSIONS

In this work, we present a comprehensive architectural framework tailored for human-robot interaction applications, particularly focusing on tasks such as object handover. Our key contribution lies in unified hand-object pose estimation, achieved through an early-stage fusion of RGB and depth modalities. The fused data is processed by a MobileNetV2 FPN-based backbone to extract region-of-interest (ROI) aligned features for both the hand and the object. These features are subsequently encoded using a deformable transformer, with cross-attention-based decoding employed to estimate both hand and object parameters. From these parameters, we derive 3D hand mesh reconstructions and 6D object pose estimations. The proposed models are evaluated on large-scale open-source datasets, demonstrating competitive, state-of-the-art performance. Our future work will focus on thoroughly evaluating the proposed system within a human-robot interaction (HRI) workspace. While we have tested the inference speed in real-time and conducted preliminary tests on a limited number of samples to validate the system’s functionality in the HRI environment, further efforts will include creating a new dataset and testing the system in entirely unseen environments to assess its robustness and generalization capabilities.

## ACKNOWLEDGEMENTS

Funded by the German Federal Ministry of Education and Research (BMBF) – Project-ID 01IS23047B – aiRobot.

## REFERENCES

- Baek, S., Kim, K. I., and Kim, T.-K. (2019). Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1076.
- Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., and Thalmann, N. M. (2019). Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281.
- Calli, B., Siu, A., Walsman, A., Matusik, W., and Allen, P. (2017). The ycb object and model set: Towards common benchmarks for manipulation research. *arXiv preprint arXiv:1709.06965*.
- Cao, Z., Radosavovic, I., Kanazawa, A., and Malik, J. (2021). Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12417–12426.
- Chen, P., Chen, Y., Yang, D., Wu, F., Li, Q., Xia, Q., and Tan, Y. B. (2021). I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12909–12918.
- Chen, X., Liu, Y., Dong, Y., Zhang, X., Ma, C., Xiong, Y., Zhang, Y., and Guo, X. (2022a). Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12912–12921.
- Chen, Z., Hampali, S., Schmid, C., and Laptev, I. (2023). Gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12890–12900.
- Chen, Z., Hasson, Y., Schmid, C., and Laptev, I. (2022b). Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 231–248, Cham, Switzerland. Springer.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.
- Ge, L., Cai, Y., Weng, J., and Yuan, J. (2018). Hand pointnet: 3d hand pose estimation using point sets. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8417–8426.
- Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., and Yuan, J. (2019). 3d hand shape and pose estimation from a single rgb image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10825–10834.
- Hampali, S., Rad, M., Oberweger, M., and Lepetit, V. (2020). Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3196–3206.
- Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., and Schmid, C. (2020). Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 571–580.
- Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M. J., Laptev, I., and Schmid, C. (2019). Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816.
- Hoang, D.-C., Tan, P. X., Nguyen, A.-N., Vu, D.-Q., Vu, V.-D., Nguyen, T.-U., Hoang, N.-A., Phan, K.-T., Tran, D.-T., Nguyen, V.-T., Duong, Q.-T., Ho, N.-T., Tran, C.-T., Duong, V.-H., and Ngo, P.-Q. (2024). Multi-modal hand-object pose estimation with adaptive fusion and interaction learning. *IEEE Access*, 12:54339–54351.
- Jiang, T., Xie, X., and Li, Y. (2024). Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*.
- Jiang, Z., Rahmani, H., Black, S., and Williams, B. M. (2023). A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6276–6286.
- Lepetit, V., Moreno-Noguer, F., and Fua, P. (2009). Epnp: An accurate o(n) solution to the pnp problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.
- Lin, K., Wang, L., and Liu, Z. (2021). End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10690–10699.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125.
- Liu, S., Jiang, H., Xu, J., Liu, S., and Wang, X. (2021). Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14696.
- Mishra, A., Fathi, A., Jain, M., and Handa, A. (2020). Dex-ycb: A benchmark for dexterous manipulation of ob-

- jects in cluttered environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3473–3480.
- Moon, G., Yu, S.-I., Wen, H., Shiratori, T., and Lee, K. M. (2020). Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*.
- Park, J., Oh, Y., Moon, G., Choi, H., and Lee, K. M. (2022a). Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Park, J., Oh, Y., Moon, G., Choi, H., and Lee, K. M. (2022b). Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., and Malik, J. (2024). Reconstructing hands in 3D with transformers. In *CVPR*.
- Qi, C. R., Liu, W., Wu, C., Su, H., and Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Qu, W., Cui, Z., Zhang, Y., Meng, C., Ma, C., Deng, X., and Wang, H. (2023). Novel-view synthesis and pose estimation for hand-object interaction from sparse views. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15054–15065.
- Romero, J., Masi, I., Ranjan, A., Zhu, Z., Liu, Y., Shih, Y., Joo, H., Niebles, J. C., and Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4514–4523.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520.
- Tang, X., Wang, T., and Fu, C.-W. (2021). Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13909–13918.
- Tse, T. H. E., Kim, K. I., Leonardis, A., and Chang, H. J. (2022). Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1674.
- Ultralytics (2023). Yolov8: State-of-the-art object detection and segmentation. <https://github.com/ultralytics/ultralytics>.
- Wang, H., Wang, C., Li, H., and Li, Y. (2022). Hope net: Hierarchical object pose estimation. *IEEE Robotics and Automation Letters*, 7(4):7519–7526.
- Wang, H., Wang, C., Li, H., and Li, Y. (2023). Harmonious features learning for hand-object pose estimation. *IEEE Robotics and Automation Letters*, 8(2):1683–1690.
- Xu, Y., Wang, H., Wang, C., Li, H., and Li, Y. (2022). Hoisd: A hierarchical object-interaction dataset with spatial and functional dependencies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):10379–10393.
- Zhang, W., Wu, X., Luo, Z., Zhou, Z., Li, C., and Bao, X. (2020). Ho-3d: A dataset for 3d hand object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5867–5876.
- Zhao, Y., Wang, H., Wang, C., Li, H., and Li, Y. (2023). Interacting hand-object pose estimation via dense mutual attention. *IEEE Robotics and Automation Letters*, 8(2):1675–1682.
- Zhu, X., Su, W., Lu, L., Xu, B., Li, X., and Wang, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Zimmermann, C. and Brox, T. (2017). Learning to estimate 3d hand pose from single rgb images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921.
- Zimmermann, C., Rother, C., Saito, J., Pock, T., Sumer, H., Loper, M., Deigel, M., and Geiger, A. (2020). Freihand: A dataset for hand mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4316–4325.