ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CVL Computer Vision Lab

3D Human-Object Interaction in Video

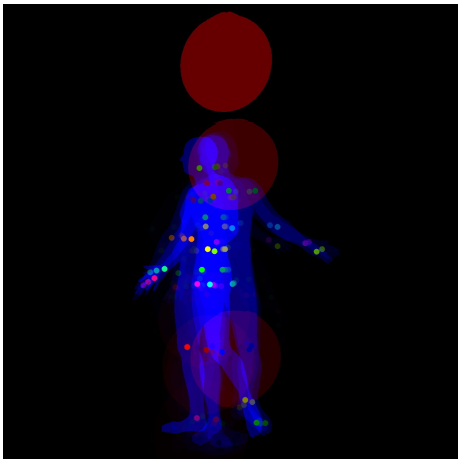
A New Approach to Object Tracking via Cross-Modal Attention

Master's Thesis

Lorenzo Germano

M.Sc. student in Mathematics in Data Science
TUM School of Computation, Information and Technology
Technische Universität München (TUM)

Mobility student at the Group of Image Communication and Understanding
Computer Vision Laboratory
Department of Electrical Engineering and Information Technology
Eidgenössische Technische Hochschule Zürich (ETH Zurich)



Advisors: Dr. Xi Wang, Dr. Gurkirt Singh
Supervisor at ETH: Prof. Dr. Luc van Gool
Supervisor at TUM: Prof. Dr. Daniel Cremers

February 14, 2024

Abstract

A novel framework for 6-DoF (Six Degrees of Freedom) object tracking in RGB video is introduced, named H2O-CA (Human to Object – Cross Attention). This framework adopts a sequence-to-sequence approach: it utilizes a method for the regression of avatars to parametrically model the human body, then groups offsets in a sliding-window fashion, and employs a cross-modal attention mechanism to attend human pose to object pose. The study commences by comparing datasets and regression methods for avatars in 5D (TRACE/ROMP/BEV/4DH) and scrutinizing various coordinate systems, including absolute, relative, and trilateration techniques, with the BEHAVE dataset being employed throughout. The significance of human pose in tracking tasks is explored by juxtaposing it with a baseline encoder model that relies solely on object pose. Various training configurations, differentiated by their loss functions, are investigated for the tracking task. Additionally, the framework is compared with other object-tracking methodologies (DROID-SLAM/BundleTrack/KinectFusion/NICE-SLAM/SDF-2-SDF/BundleSDF). The approach is particularly effective in scenarios influenced by human actions, such as lifting or pushing, which direct object movement, and in instances of partial or full object obstructions. Qualitative results are illustrated [here]. Although the fully recursive tracking approach does not achieve state-of-the-art performance, the potential of next-frame prediction and next-4 frames prediction is acknowledged. The primary application envisioned is in augmented reality (AR). The project's code will be made available [here].

Acknowledgements

I would like to express my deepest gratitude to the following individuals and organizations who have contributed to the completion of this thesis:

- My thesis advisors, Dr. Xi Wang, and Dr. Gurkirt Singh, for their guidance, and expertise throughout this research journey.
- My family, for their love, encouragement, and understanding during the challenging times of my academic pursuits.
- My friends and colleagues, for their camaraderie, valuable discussions, and motivation that kept me going.
- The faculty and staff of ETH Zurich, for their dedication to education and research, which enriched my academic experience.
- The faculty and staff of TUM, for coordinating the exchange.
- This work was supported in part by assistance from OpenAI's language model, ChatGPT, which provided information retrieval and data processing services.

I am profoundly grateful for the support and encouragement that I have received from all of you. This thesis would not have been possible without your contributions and belief in my abilities.

Thank you for being part of this incredible journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Focus of this Work

The research problem investigated is 6-DoF (Six Degrees of Freedom) object tracking, which consists of monitoring an object's position and orientation within a three-dimensional space from video. The term "degrees of freedom" pertains to the six independent parameters that delineate an object's state: three translational movements (forward/backward along the X-axis, left/right along the Y-axis, and up/down along the Z-axis) and three rotational movements (roll around the X-axis, pitch around the Y-axis, and yaw around the Z-axis). While position refers to the object's location in 3D space, pose encompasses both position and orientation, offering a comprehensive description of the object's placement and facing direction.

Tracking, in this context, signifies the continuous monitoring of these parameters to detect the object's dynamic movements through space. The accuracy of 6-DoF tracking can be affected by occlusion, a phenomenon where the object or parts of it are obscured from the tracking system's sensors, potentially interrupting the continuous tracking of the object's pose. Overcoming occlusion is crucial in maintaining the integrity of 6-DoF tracking, especially in applications like robotics, virtual/augmented reality, and autonomous navigation, where understanding the precise pose of objects or users is paramount for interaction and navigation.

The advent of deep learning has ushered in a new era in this field of computer vision, significantly advancing trackers. Despite these advancements, the accurate tracking of objects in the presence of complex human-object interactions remains a formidable challenge. Recent studies such as those by Petrov et al. [22], Xie et al. [39], and Wen et al. [37] have highlighted the importance of considering these interactions to enhance the robustness and accuracy of 6-DoF tracking systems. This study aims to build on top of these state-of-the-art methods to explore further methodologies that jointly consider the visibility of both humans and objects during interaction tracking.

The primary focus is on scenarios where the input is an RGB video, captured either by external cameras, body-mounted sensors such as wearable devices (e.g., smartphones or fitness watches), or derived from any shape-from-X approach. The scenarios of interest involve human-object interactions with non-deformable, non-transparent objects commonly found in everyday settings, such as boxes, or chairs. The interactions learned encompass a wide range of activities including lifting, carrying, sitting, pushing, and pulling with hands and feet, as well as more free-form interactions. It is, however, essential that the interactions do not involve multiple objects simultaneously or extend over prolonged periods. The expected outcome of this research is a robust methodology capable of providing a detailed 6-DoF representation of the object's trajectory, on par with the most recent literature.

Figure 1.1: Example of human-object interaction in a frame from Date01_Sub01_backpack_back, camera 1. Here, the background registration is highlighted in green, and the posed human and object meshes, respectively, in blue and red.

1.2 Thesis Organization

The structure of this thesis unfolds as follows:

- **Chapter 1: Introduction** – This chapter introduces the research, focusing on the scope, objectives, and significance of the work. It provides an overview of the thesis structure and lays the groundwork for the subsequent chapters.
  - **Section 1.1: Focus of this Work** – Discusses the main aims and objectives of the research.
  - **Section 1.2: Thesis Organization** – Outlines the structure and content of the thesis.
- **Chapter 2: Related Work** – Reviews existing literature and research relevant to the thesis topic. It contextualizes the research within the broader field, highlighting key theories, models, and methodologies.
  - **Section 2.1: Human Body Model** – Examines the human body models, with a specific focus on the SMPL model.
  - **Section 2.2: Human-Centered Object Tracking** – Discusses the challenges and methodologies related to human-centered object tracking.
  - **Section 2.3: Sequence Learning** – Explores the principles and approaches in sequence learning applicable to the research.
- **Chapter 3: Materials and Methods** – Details the datasets, tools, and methodologies employed in the research. This chapter ensures the reproducibility of the research and provides a foundation for the experiments.
  - **Section 3.1: Comparison of Datasets for Human-Object Interaction** – Compares and contrasts different datasets used in the study.

  - **Section 3.2: Comparison of Methods for Regression of Avatars (5D)** – Analyzes different regression methods for avatars.
  - **Section 3.3: Choice of Coordinate System** – Discusses the various coordinate systems used and their implications on the research.
  - **Section 3.4: Architectures** – Introduces and elaborates on the H2O-CA architecture and its significance in the research.

- **Chapter 4: Experiments and Results** – Presents a detailed account of the experiments conducted, the methodologies applied, and the results obtained.
  - **Section 4.1: Training** – Describes the training protocols and setups.
  - **Section 4.2: Evaluation Metrics** – Details the metrics used to evaluate the performance of the models.
  - **Section 4.3: Ablation Studies** – Analyzes the contribution of individual components of the system.
  - **Section 4.4: Quantitative Comparison** – Provides a quantitative comparison of the H2O-CA framework with other baselines and methodologies.

- **Chapter 5: Discussion** – Interprets the results, discussing the implications and significance of the findings. This chapter also addresses the broader impact of the research, its applications, and future work.
  - **Section 5.1: Independence of Location and Orientation** – Discusses how the model's performance is affected by location and orientation factors.
  - **Section 5.2: Use of Offsets** – Examines the role and impact of using offsets in the model.
  - **Section 5.3: Use of Masks** – Analyzes the use of masks and their significance in the model.
  - **Section 5.4: Choice of Modality for Human Representation** – Discusses the choice of human representation modality and its impact on the model's performance.
  - **Section 5.5: Limitations and Future Work** – Addresses the limitations of the current research and outlines potential future directions.
  - **Section 5.6: Applications** – Explores practical applications of the research and discusses ethical considerations in object tracking.

- **Chapter 6: Conclusion** – Summarizes the key findings of the research, their implications, and suggests avenues for future research.

- **Appendices** – Provides supplementary material to support the thesis, including tables for comparison of methods for regression of avatars (5D) and methods for 6DoF object tracking.

Chapter 2

Related Work

This section offers a comprehensive overview of the advancements in the various domains that will be discussed in the subsequent chapters.

2.1 Human body model

The evolution of techniques in the domain of human mesh recovery (HMR) has been considerable (see [33, 39] for a literature survey and references). Starting with the influential SMPLify method and its derivatives, there has been an emphasis on estimating 3D human pose and shape via iterative optimization. Using just a single image input, HMR has advanced this field by directly regressing SMPL parameters with CNNs. This method has seen several enhancements, with approaches delving into pseudo-ground truth generation leveraging temporal data, multiple views, and iterative optimization. There's also been the advent of distinct HMR architectural designs like PyMAF, PARE, HKMR, and HoloPose. Notably, HMR2.0 stands out by surpassing all its predecessors. Some strategies, such as GraphCMR, METRO, FastMETRO, and Mesh Graphormer, have even ventured into non-parametric predictions. The dynamics change when video is considered. Solutions have explored temporal encoding designs ranging from convolutional encoders like HMMR to more advanced mechanisms. In 2015, Loper et al. in "SMPL: A Skinned Multi-Person Linear Model" [15] presented a learned model of human body shape and pose-dependent shape variation. SMPL is a skinned vertex-based model that represents a wide variety of body shapes in natural human poses. The parameters of the model are learned from data including the rest pose template, blend weights, pose-dependent blend shapes, identity-dependent blend shapes, and a regressor from vertices to joint locations. Unlike previous models, the pose-dependent blend shapes are a linear function of the elements of the pose rotation matrices. Later in 2017, Romero et al. in "Embodied Hands: Modeling and Capturing Hands and Bodies Together" [26] noticed that when scanning or capturing the full body in 3D, hands are small and often partially occluded, making their shape and pose hard to recover. To cope with low resolution, occlusion, and noise, they developed a new model called MANO (hand Model with Articulated and Non-rigid defOrmations). MANO is learned from around 1000 high-resolution 3D scans of the hands of 31 subjects in a wide variety of hand poses. The model is realistic, low-dimensional, captures non-rigid shape changes with pose, is compatible with standard graphics packages, and can fit any human hand. They attach MANO to a standard parameterized 3D body shape model (SMPL), resulting in a fully articulated body and hand model (SMPL+H). In 2019, Pavlakos et al. in "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image" [21], to facilitate the analysis of human actions, interactions and emotions, compute a 3D model of human body pose, hand pose, and facial expression from a single monocular image. To achieve this, they use thousands of 3D scans to train a new, unified, 3D model of the human body, SMPL-X, that extends SMPL with fully articulated hands and an expressive face.

2.1.1 The SMPL model

The SMPL model [15] primarily relies on several parameters: $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and the global translation. Here, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ represents the body shape, capturing individual body characteristics from a principal component analysis. The pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{3K}$, where $K = 24$ for the number of joints, dictates the rotation of each body part in the kinematic tree, using the axis-angle representation. Each joint rotation is expressed in three dimensions, leading to $\boldsymbol{\theta}$ having 72 dimensions. The global translation vector, typically in $\mathbb{R}^3$, sets the model's overall position in 3D space. The axis-angle representation is defined by two components: a unit vector $\mathbf{u} = (u_x, u_y, u_z)$, representing the axis of rotation, and an angle θ, representing the magnitude of the rotation around this axis.

Given a rotation axis $\mathbf{u}$ and a rotation angle θ, any vector $\mathbf{v}$ in 3D space can be rotated using the Rodrigues' rotation formula [25]:

$$\mathbf{v}_{rot} = \mathbf{v}\cos\theta + (\mathbf{u} \times \mathbf{v})\sin\theta + \mathbf{u}(\mathbf{u} \cdot \mathbf{v})(1 - \cos\theta)$$

where $\mathbf{v}_{rot}$ is the rotated vector, $\times$ denotes the cross product, and $\cdot$ denotes the dot product.

Example

Consider a vector $\mathbf{v} = (1, 0, 0)$ that you wish to rotate around the axis $\mathbf{u} = (0, 1, 0)$ (the y-axis) by an angle of $\theta = 90°$ (or $\frac{\pi}{2}$ radians). Applying the Rodrigues' rotation formula, we get:

$$\mathbf{v}_{rot} = \begin{pmatrix}1\\0\\0\end{pmatrix}\cos\frac{\pi}{2} + \left(\begin{pmatrix}0\\1\\0\end{pmatrix} \times \begin{pmatrix}1\\0\\0\end{pmatrix}\right)\sin\frac{\pi}{2} + \begin{pmatrix}0\\1\\0\end{pmatrix}\left(\begin{pmatrix}0\\1\\0\end{pmatrix} \cdot \begin{pmatrix}1\\0\\0\end{pmatrix}\right)(1 - \cos\frac{\pi}{2})$$

After simplifying, it is found that $\mathbf{v}_{rot} = (0, 0, -1)$. This means the vector $\mathbf{v}$ has been rotated by 90 degrees around the y-axis, resulting in a new position along the negative z-axis.

The hierarchical structure is shown in Table 2.1.

This hierarchical joint structure ensures that parent joint movements have appropriate influences on their child's joints, facilitating a coherent and anatomically plausible representation of human motion and posture.

2.2 Human-centered object tracking

In the domain of human and object reconstruction, initial work primarily centered on the standalone modeling of appearance, largely overlooking interactional aspects. Human reconstruction efforts have explored a range from understanding basic forms using singular or multiple viewpoints to intricate details like hand gestures, facial expressions, self-contacts, and even clothing nuances. The utilization of parametric body models enhanced the accuracy of these reconstructions, leading to advancements where even sparse or single-camera feeds could generate reliable human models. However, the inherent ambiguity between depth and scale in capturing 3D humans from RGB data necessitated the shift towards using RGBD or volumetric data for a more dependable capture ([39]).

The field of object tracking from a human-centered perspective is relatively new, yet advancements have been noteworthy. A significant contribution to this domain is the work by Petrov et al. in their 2023 study, "Object pop-up" [22]. Their methodology involves training end-to-end a PointNet++ architecture to predict a coarse object center initialization from a single human point cloud, in turn, fed to a PointNet++ architecture that predicts point-wise offsets of a downsampled object mesh from the previous human encoding and a hot-encoded object class template. This approach underscores the importance of considering not only the body

Joint Index	Joint Name
0	Pelvis
1	Left Hip
2	Right Hip
3	Spine1
4	Left Knee
5	Right Knee
6	Spine2
7	Left Ankle
8	Right Ankle
9	Spine3
10	Left Foot
11	Right Foot
12	Neck
13	Left Collar
14	Right Collar
15	Head
16	Left Shoulder
17	Right Shoulder
18	Left Elbow
19	Right Elbow
20	Left Wrist
21	Right Wrist
22	Left Hand
23	Right Hand

Table 2.1: Hierarchy of joints in the SMPL model

parts in direct contact with the object but also those not immediately touching it. The prediction of point-wise offsets may produce a distortion of the mesh so the use of Procrustes alignment with the original template as the final processing step is a necessary aspect of their method.

Another pivotal work is the "BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects" framework by Wen et al., presented in 2023 [37]. This is a near real-time (10Hz) method for 6-DoF tracking of an unknown object from a monocular RGBD video sequence, while simultaneously performing neural 3D reconstruction of the object. This framework commences with matching features between consecutive segmented images to establish a coarse pose estimate. A subset of these posed frames is then stored in a memory pool for later refinement. The creation of a dynamic pose graph from the memory pool enables online optimization, which refines all the poses in the graph alongside the current pose. These updated poses are then re-stored in the memory pool. Crucially, all posed frames in the memory pool contribute to the learning of a neural object field. This field models both the geometry and visual texture of the object, while also adjusting their previously estimated poses.

In the same year, Xie et al. introduced a novel approach in their study, "Visibility Aware Human-Object Interaction Tracking from Single RGB Camera" [39], presenting a method that can jointly track full-body human interaction with a movable object from a monocular RGB camera. Their method begins with the reconstruction and tracking of 3D human and object models, as well as their interactions, from an RGB sequence and corresponding human-object masks. The core of this approach lies in the SMPL-T conditioned interaction field network (SIF-Net) that predicts neural fields (in this context, refer to continuous functions parameterized by neural networks), based on estimated SMPL meshes in camera space. This conditioning provides temporally consistent relative translation, essential for coherent 4D tracking. Furthermore, they introduce the HVOP-Net unit to predict object pose under occlusions by leveraging human motion and object visibility information. The joint optimization of human and object models ensures adherence to image observations and contact constraints.

Figure 2.1: SMPL joints regressed (regressor from [15]). The human mesh uses neutral shape parameters.

Among these, the framework that performs best on the BEHAVE dataset is the "BundleSDF" method. Notably, this framework has achieved a mean Chamfer distance (see [37]) of 1.16 cm. It must be pointed out that the point of view with the least object occlusion is selected for each scene. This performance surpasses that of the approach documented in [39], which records a mean Chamfer distance of 7.89 cm. For the sake of completeness, it is important to highlight that the measurements in [39] were obtained on keyframes only (a subset consisting of 3.9k frames) primarily for computational efficiency, and alignment with Procrustes analysis over a sliding window of size 1 was performed.

Challenges

The advancements in object-tracking by Wen et al. in "BundleSDF" (2023), Xie et al. in "Visibility-Aware Object Tracking in Human-Object Interaction" (2023), have notably progressed the field, yet they also leave open several challenges that highlight the complexity of object tracking in real-world scenarios.

In the "BundleSDF" framework by Wen et al., several challenges are identified. Firstly, the issue of long-term complete occlusions arises when an object is carried and faces away from the camera, obscuring it from view. Secondly, they note the difficulty in handling severe motion blur and abrupt displacement, especially when the object is swung by a human. Additionally, the distance from the camera during video capture poses a challenge, particularly for depth sensing, which is crucial for this method. Furthermore, the method faces limitations in tracking under conditions of severe occlusion, segmentation error, lack of texture

and geometric cues, and the re-appearance of the object affected by symmetric geometry. The requirement for depth modality limits its application to non-transparent objects and assumes that the object being tracked is rigid.

In contrast, Xie et al.'s method, while robust under heavy occlusions, assumes known object templates for tracking, which could be a limitation in scenarios where object templates are not available or are highly variable. Additionally, Xie et al. identify three typical failure cases in their supplementary notes. The first is related to heavy occlusion, especially when there are significant changes in object pose and contact locations between visible frames. This makes it challenging to accurately track pose and contact changes. The second issue arises from the inherent difficulty of pose prediction, even when the object is fully visible but adopts an uncommon pose. The third challenge is the tracking of symmetric objects, where the network struggles due to symmetry, despite minimizing 2D mask loss and contact constraints.

2.3 Sequence Learning

In the domain of visual object tracking, significant advancements have been made, particularly with the introduction of novel architectures like SeqTrack, proposed by Chen et al. in 2023 [3]. SeqTrack's foundation is an encoder-decoder transformer model, where the former extracts visual features from the input video frames, and the latter is designed to autoregressively generate a sequence of bounding box tokens. These tokens represent the position and dimensions of the object being tracked in each frame of the video. The generation of these tokens is based on the visual features extracted by the encoder, ensuring that the tracking is grounded in the actual visual content of the video. This interaction is facilitated by a masked multi-head attention mechanism. The masked nature of this attention mechanism ensures that the prediction for each frame is based only on the preceding frames, adhering to the causal nature of the problem.

The SeqTrack builds upon the foundational work in transformers, such as that by Vaswani et al. in 2017 [34], and integrates concepts from more recent developments in the field, like the MixFormer architecture proposed by Cui et al. [4].

The use of transformers in domains other than natural language processing (NLP) is very promising in different fields. "Generating Interacting Protein Sequences using Domain-to-Domain Translation" [18], for example, is a method for generating protein domain sequences intended to interact with another protein domain. Using data from natural multi-domain proteins, they cast the problem as a translation problem from a given interactor domain to the new domain to be generated, i.e. they generate artificial partner sequences conditional on an input sequence. Likewise, it is possible to cast the object tracking task into a sequence-to-sequence problem, with different domains (human and object). This motivates the choice of exploring a transformer-based tracker with cross-attention.

Furthermore, *Pose2Room* [20] uses a spatiotemporal pose encoder that operates on two levels: spatially, it uses graph convolution across skeleton bones to extract per-frame pose features from the relationships between intra-skeleton joints. Temporally, it employs a 1D convolution over a window frame to analyze the movement of each joint, understanding inter-frame relations. This dual focus on both spatial and temporal dimensions allows for a more nuanced understanding of human poses and movements over time. This mechanism enables the model to pay selective attention to different parts of the pose data, which is crucial for accurately inferring the 3D structure of the scene and the objects within it based on human movements. Graph convolution relates to the self-attention mechanism used by both encoder and decoders, as it allows for the incorporation of graph structure into the learning process, enabling the model to capture and leverage the relational information inherent in graph data such as the human skeleton.

In literature, there are also examples of transformers-based trackers (see [35] for a literature review), like "Transforming Model Prediction for Tracking", used to estimate the weights of a discriminative correla-

tion filter used for the tracker, or "SeqTrack" [3], a method that generates a sequence in an autoregressive manner. They perform a study where they compare this autoregressive method with another bidirectional method that predicts all coordinate values simultaneously. In the bidirectional method, the input sequence consists of four special tokens similar to the start token. The decoder receives the sequence and predicts the four coordinates in a batch. The causal attention mask is removed, allowing tokens to attend to each other. The study highlights that the bidirectional method performs much inferior to the autoregressive one, demonstrating that the causal relation between tokens is important for sequence modeling in tracking.

Chapter 3

Materials and Methods

3.1 Comparison of datasets for human-object interaction

Among the datasets available, three stand out for their comprehensive labeling of human-object interactions:

- *InterCap* [10]: comes with pseudo ground truth SMPL and object registrations at 30fps, in approximately 420 GB. Azure Kinect sensors are set up with a simple multi-view RGB-D capture system that minimizes the effect of occlusion while providing reasonable inter-camera synchronization. It contains 10 subjects (5 males and 5 females) interacting with 10 objects of various sizes and affordances, including contact with the hands or feet. In total, InterCap has 223 RGB-D videos, resulting in 67,357 multi-view frames, each containing 6 RGB-D images.

- "HuMoR" [24]: adopts the generalized notion of "whole-body grasps". GRAB (GRasping Actions with Bodies), contains full 3D shape and pose sequences of 10 subjects interacting with 51 everyday objects of varying shape and size. Given MoCap markers, Rempe et al. fit the full 3D body shape and pose, including the articulated face and hands, as well as the 3D object pose. This gives 3D meshes over time, from which they compute contact between the body and the object.

- "BEHAVE" [1]: a full body human-object interaction dataset with multi-view RGBD frames and corresponding 3D SMPL and object fits along with the annotated contacts between them. Bhatnagar et al. record circa 15k frames at 5 locations with 8 subjects performing a wide range of interactions with 20 common objects. They use this data to learn a model that can jointly track humans and objects in natural environments with an easy-to-use portable multi-camera setup.

  Throughout the rest of the research, "BEHAVE" dataset has been used, specifically utilizing the official test split comprising all 1fps sequences with subject 3 and selected sequences with subjects 4 and 5. Initial statistical analysis, as depicted in Figures 3.1 and 3.2, reveals a uniform distribution of data across various components and camera views.

  In Figure 3.1, the histograms for the sign statistics of the first (x-axis) and second component (y-axis) demonstrate a similar distribution, with the majority of data points falling into the [-1,1] m interval across the four cameras. For the third component (z-axis), a predominant frequency at [1.5, 3.5] m is observed, indicating a possible skewness or an imbalance in the distribution of this component across the four cameras. By working with offsets, this potential issue is overcome.

  In Figure 3.2, the histograms for the sign statistics of the components (rows) across the four cameras (columns) demonstrate the first two components being balanced, while the third is found only positive.

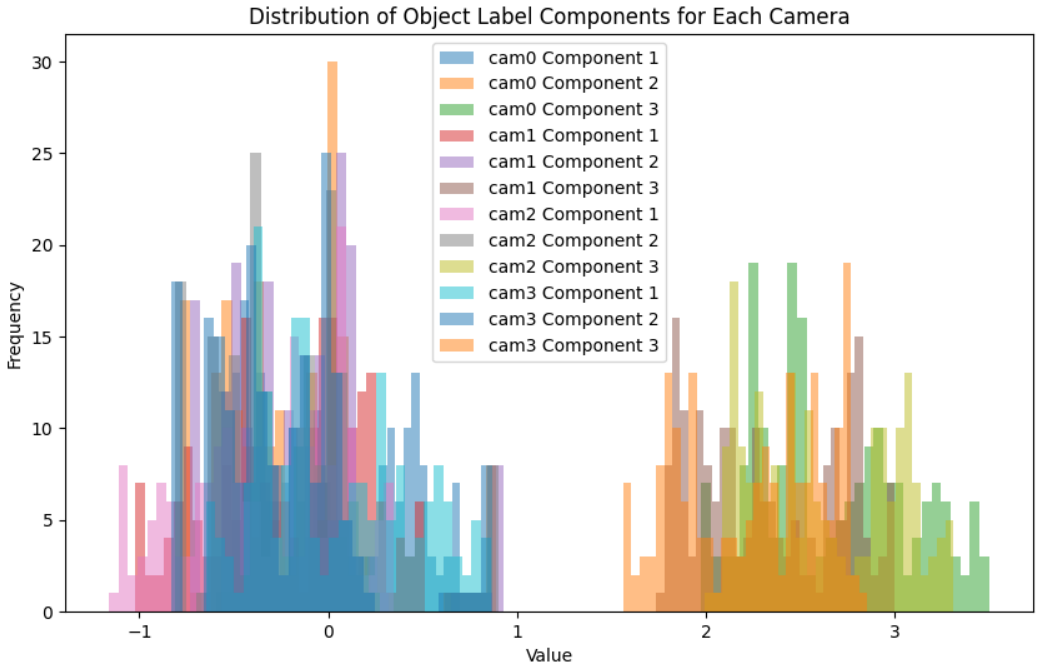


Figure 3.1: Distribution of the components of the objects' position for each camera on the test set.

3.2 Comparison of methods for regression of avatars (5D)

In 2023, Goel et al. present "Humans in 4D" [7], an approach to reconstruct humans and track them over time. At the core of their approach is a transformer-based version of a network for human mesh recovery (HMR 2.0), that uses the Hungarian algorithm for tracking. In 2021, Sun et al. present "Monocular, One-stage, Regression of Multiple 3D People" [29]. This paper focuses on the regression of multiple 3D people from a single RGB image. Existing approaches predominantly followed a multi-stage pipeline that first detects people in bounding boxes and then independently regresses their 3D body meshes. In contrast, the authors propose to Regress all meshes in a One-stage fashion for Multiple 3D People (termed ROMP). Their method simultaneously predicts a body center heatmap and a mesh parameter map, which can jointly describe the 3D body mesh on the pixel level. Through a sampling process, the body mesh parameters of all people in the image are extracted from the mesh parameter map. This one-stage framework is free of the complex multi-stage process and more robust to occlusion. By the same author and successive is "Putting People in their Place" (2022) [31], where the focus is put on multi-body images and their relative depth. Inferring the depth of a person in an image, however, is fundamentally ambiguous without knowing their height. This is particularly problematic when the scene contains people of very different sizes, e.g. from infants to adults. While previous work reasons in the image plane, in BEV, a Bird's-Eye-View is assumed to explicitly reason about depth. BEV reasons simultaneously about body centers in the image and in-depth and, by combining these, estimates 3D body position. Finally, in the most recent work from Sun et al., titled "TRACE" [30], the additional problem of tracking in global coordinates, which is critical for many applications, is addressed. This is particularly challenging when the camera is also moving, entangling human and camera motion. To address these issues, the authors adopt a novel 5D representation (space, time, and identity) that enables end-to-end reasoning about people in scenes. This method, called TRACE, introduces several novel architectural components. Most importantly, it uses two new "maps" to reason about

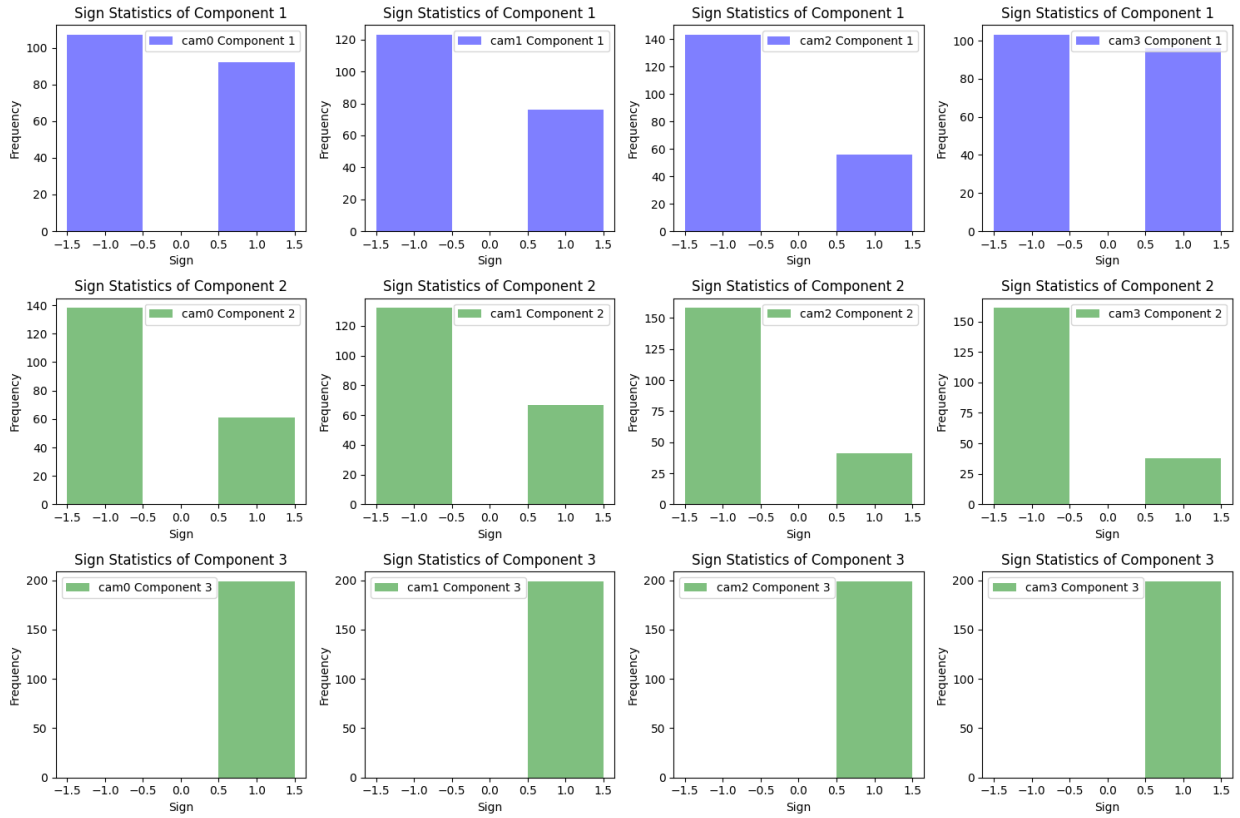


Figure 3.2: Distribution of the sign of components of objects' position for each camera on the test set.

the 3D trajectory of people over time in camera, and world, coordinates. An additional memory unit enables persistent tracking of people even during long occlusions. TRACE is the first one-stage method to jointly recover and track 3D humans in global coordinates from dynamic cameras. The same task is addressed by Ye et al. In 2023, "Decoupling Human and Camera Motion from Videos in the Wild" [40] is presented as a method to reconstruct global human trajectories from videos in the wild. This method decouples the camera and human motion, placing people in the same world coordinate frame. Most existing methods did not model the camera motion; methods that relied on the background pixels to infer 3D human motion usually require a full scene reconstruction, which is often not possible for in-the-wild videos. However, following the intuition that, even when existing SLAM systems cannot recover accurate scene reconstructions, the background pixel motion still provides enough signal to constrain the camera motion, they show that relative camera estimates along with data-driven human motion priors can resolve the scene scale ambiguity and recover global human trajectories.

A quantitative evaluation of various methodologies has been organized in Table A.1 found in Appendix A.1. The assessment is grounded on the Chamfer distance metric, further detailed in Section 4.2, and utilizes the test split of the *BEHAVE* dataset for consistency across comparisons. The analysis revealed that the TRACE framework exhibits superior performance in minimizing the Chamfer distance. As evidenced in the table, TRACE achieves a mean Chamfer distance of 3.67 cm, which is notably lower than its counterparts; ROMP and 4DH both register a mean of 4.60 cm, while BEV stands slightly better at 4.44 cm. The results of the benchmarking highlight the TRACE methodology as the best backbone.

Moving forward, this research will concentrate on leveraging ground truth human poses. This choice is

aimed at understanding the impact of human input on the efficacy of object tracking neglecting the contribution of HMR methods (a potentially confounding variable).

3.3 Choice of coordinate system

In the realm of human and object localization within the literature, researchers predominantly utilize three coordinate systems to represent spatial information: absolute coordinates, a hybrid system involving absolute coordinates for the initial frame followed by relative frame-to-frame offsets, and a hybrid system with cross-offset involving absolute coordinates for the initial frame followed by relative frame-to-frame offsets for the human and frame-to-frame offsets relative to the human offsets for the object. See 3.3. Analogous considerations can be made for the orientation.

For the sake of clarity, let H_t and O_t denote the positions of a human and an object at time t, respectively.

3.3.1 Absolute Position Coordinates

In this system, the position of a human or an object at any given frame is defined in absolute terms, usually in a fixed coordinate space. Mathematically, this can be expressed as:

$$H_t = (x_t^H, y_t^H, z_t^H)$$
$$O_t = (x_t^O, y_t^O, z_t^O)$$

3.3.2 Relative Frame-to-Frame Offsets

Here, the position for any frame (beyond the first) is given as an offset from the previous frame. The position in the first frame remains absolute. This can be represented as:

For the first frame (t = 1):

$$H_1 = (x_1^H, y_1^H, z_1^H)$$
$$O_1 = (x_1^O, y_1^O, z_1^O)$$

For subsequent frames (t > 1):

$$H_t = H_{t-1} + \Delta H_t$$
$$O_t = O_{t-1} + \Delta O_t$$

where ΔH_t and ΔO_t are the offsets (changes) in human and object positions between consecutive frames, respectively.

3.3.3 Relative Frame-to-Frame Cross-Offsets

In this system, the position of the human is updated using the relative frame-to-frame offsets approach, while the object's position is determined by a second-level offset, factoring in the relative motion with respect to the human. This is formulated as:

For the first frame (t = 1):

$$H_1 = (x_1^H, y_1^H, z_1^H)$$
$$O_1 = (x_1^O, y_1^O, z_1^O)$$

For subsequent frames (t > 1):

$$H_t = H_{t-1} + \Delta H_t$$

$$\Delta O_t = \Delta O_{t-1} + \Delta\Delta O_t$$

Here, $\Delta\Delta O_t$ represents the second-level offset for the object. It is calculated based on the difference between the human's motion and the object's motion relative to the previous frame:

$$\Delta\Delta O_t = (O_t - H_t) - (O_{t-1} - H_{t-1}) = \Delta O_t - \Delta H_t$$

This nuanced representation captures the dynamic interplay between the human and the object, considering not just their individual motions but also their relative movements. This relative perspective significantly reduces the dependency on a global coordinate system, allowing the system to focus more on local movements and changes. For the remaining part, this work uses the second coordinate system.

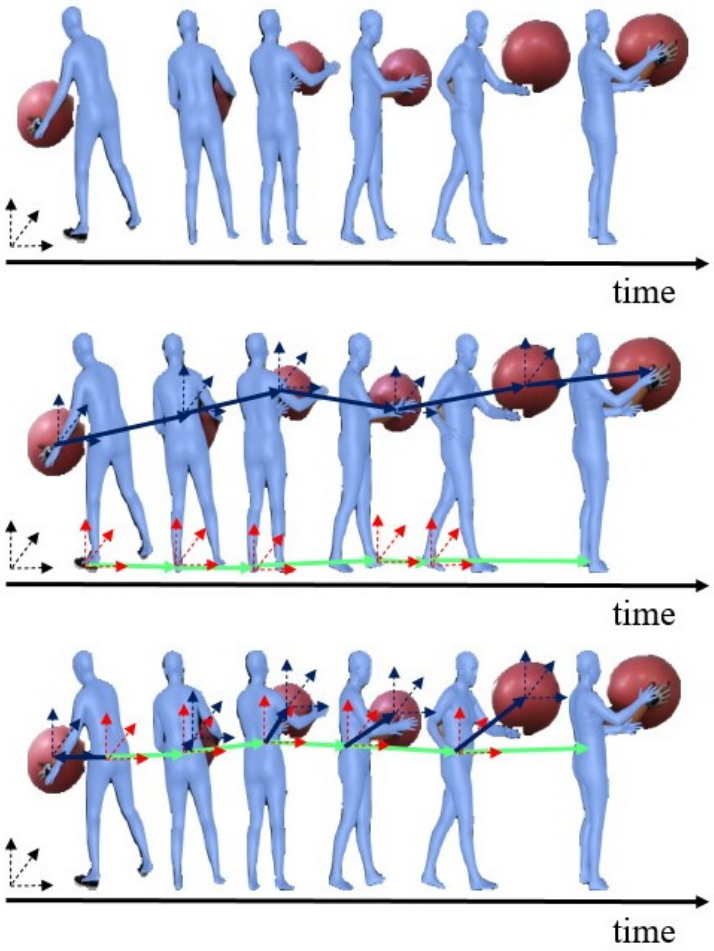


Figure 3.3: Comparison of coordinate systems to parametrize human-object interaction. In the first row absolute coordinates are used for every frame (human and object); in the second row offsets in location and orientation are shown: dark blue for the object, green (location), and red (orientation) for the human – only the left foot joint is shown; in the third row, relative offsets of the human body are shown from the pelvis (root joint) in green (location) and red (orientation), whereas offsets of the object with respect to the human offset are shown in dark blue.

3.3.4 Trilateration

In this section, the use of trilateration for localizing an object is discussed. Let there be n anchor points with known positions $A_i = (x_i, y_i, z_i)$ for $i = 1, 2, \ldots, n$, and known distances d_i from these anchors to the object. The position of the object $P = (x, y, z)$ is to be determined. The trilateration problem can be formulated as finding P that satisfies the following set of equations:

$$\begin{aligned}(x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 &= d_1^2,\\ (x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 &= d_2^2,\\ &\vdots\\ (x - x_n)^2 + (y - y_n)^2 + (z - z_n)^2 &= d_n^2.\end{aligned}$$

This set of equations represents the geometric condition that the distance between the object and each anchor is equal to the measured distance d_i. This is a fundamental technique used in numerous applications to determine the position of a point in space from anchors. This approach is exemplified in applications such as GPS, where satellites serve as anchors, indoor positioning systems utilizing signals like Wi-Fi or Bluetooth, and navigation systems for aircraft and ships. In 3D space, distances from at least four non-coplanar anchors are needed, where each distance forms a sphere around the anchor, and the point in question is located at the intersection of these spheres. Localizing objects relative to the human body has the potential to establish a view-independent localization system, where object coordinates are consistently defined in relation to the absolute positions of body joints. This approach could offer an intuitive and stable method for tracking objects in relation to the human body. Empirical evidence, however, shows that this approach is not advisable. This representation might not be suitable for a neural architecture as these type of networks inherently assume that the underlying regression task (tracking) can be approximated through a series of linear transformations and non-linear activations, trained to a local minimum. The robustness and distribution of local minima in space are explored in the following experiments, assuming all body joints of the SMPL model (24) as anchors.

Localization error as a function of anchors misplacement in trilateration

In this experiment, the impact of the localization error of anchors on the localization error of an object in a 3D space using multilateration is investigated. This setup reproduces a faulty HMR method as the backbone of an architecture that does object tracking. The process involves 24 randomly generated anchors within a cube of side length of 50 units. The true position of the object is set at the coordinates (25, 25, 25). Multilateration is employed to estimate the position of the object by minimizing the sum of squared differences between the measured distances and the Euclidean distances from the estimated position to each anchor. The estimated position (P_{est}) is calculated by minimizing the sum of squared differences between measured distances and Euclidean distances from P_{est} to each anchor.

$$\text{Minimize } S = \sum_{i=1}^{24} (d_{measured,i} - ||P_{est} - A_i||)^2$$

Where A_i is the position of the i-th anchor, $d_{measured,i}$ is the measured distance from the i-th anchor to the object, and $||\cdot||$ represents the Euclidean norm. Measurement errors (100) are simulated by adding normally distributed noise to the true distances, with the standard deviation of the noise varying from 0 to 5 units (at most 1/10 of the size of the box is close to the experienced error on the regression of joints). This procedure is repeated 100 times, and the localization error, defined as the Euclidean distance between the true and estimated positions, is recorded for each level of measurement error. Finally, the average

localization error is plotted against the measurement error magnitude to visualize the relationship between measurement precision and localization accuracy. See figure 3.4. Trilateration is a problem where the average localization error grows almost linearly with the measurement error in the radii (axis x indicates the std of a normal, from which the error is sampled).

Distribution of local minima as a function of anchors misplacement in trilateration

In this experiment, the objective is to study the top k intersection points (in terms of local minima) of multiple spheres in a 3-dimensional space when solving a trilateration problem. The process involves 10 randomly generated anchors within a cube of side length 50 units. The true position of the object is set at the coordinates (25, 25, 25). Measurement errors are simulated by adding normally distributed noise to the true distances, with the standard deviation of the noise varying from 0 to 5 units. The Newton-Raphson method is used to iteratively refine an initial guess of the intersection point. During each iteration, a function computes the Jacobian matrix and a residual vector based on the current estimate of the intersection points. The function seeks to find the intersection points of n spheres, given by their centers C_i and radii r_i.

$$\text{Find } P \text{ such that } ||P - C_i|| = r_i \text{ for each } i$$

The Newton-Raphson method updates the estimate P using the Jacobian matrix J and the residual vector R.

$$P_{new} = P_{old} - J^{-1}R$$

This procedure is repeated for 100 iterations. After a certain number of iterations or upon convergence, the list of potential solutions is sorted based on their errors, and the top k solutions with the minimum errors are selected. See figures 3.5 and 3.6. From the graphs, it emerges that non-global minima are found easily outside a threshold interval of 0.01 units. The set threshold is much larger than the state-of-the-art object localization error.

In conclusion, the local minima vary largely in space, resulting in a coordinate system not suitable for deep learning-trained architectures.

3.4 Architectures

In this section, H2O-CA, an original architecture for 6-DoF tracking is presented. The rationale behind the specific design choices is thoroughly discussed in Section 5.

3.4.1 H2O-CA

Characterization

An overview of the pipeline is presented in Figure 3.7. Orientation and translation are processed independently, trained as in Figure 4.1, and used at inference as in Figure 3.7. Details can be found in Tables 3.1, **??**, and 3.2. The computation of joints follows a methodology referenced in [28]. The positional encoding function used is as outlined in [34]. The model consists of 3.7 M trainable parameters and 0 non-trainable parameters. The total estimated model parameter size is 14.849 MB.

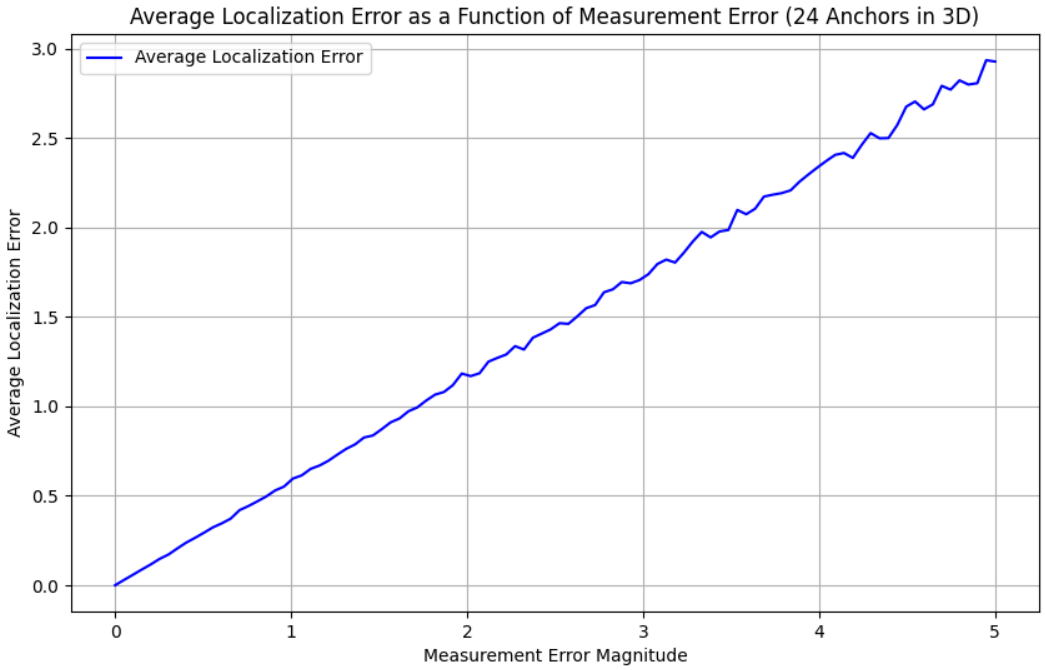


Figure 3.4: Average localization error as a function of measurement error (24 anchors in 3D).

Parameter	**Description**	**Value**
frames_subclip	Number of frames in a subclip	12
masked_frames	Number of frames to be masked at the end of the sequence	4
num_heads	Number of heads in the multi-head attention mechanisms	4
d_model	Dimensionality of the model's feature space	128

Table 3.1: H2O-CA, Summary of parameters.

Loss design

In "On the Continuity of Rotation Representations in Neural Networks" [44] Zhou et al. demonstrate that for 3D rotations, all representations are discontinuous in the real Euclidean spaces of four or fewer dimensions. Thus, widely used representations such as quaternions and Euler angles are discontinuous and difficult for neural networks to learn. They show that the 3D rotations have continuous representations in 5D and 6D, which are more suitable for learning. Hence the loss function calculates the loss between two rotations provided in axis-angle format by first converting them to matrices. The conversion process involves normalizing the axis part of the axis-angle vector and constructing a skew-symmetric cross-product matrix, skew. The rotation matrix R is derived using Rodrigues' rotation formula [25]:

$$R = I + \sin(\theta) \cdot \text{skew} + (1 - \cos(\theta)) \cdot \text{skew}^2 \tag{3.1}$$

where I signifies the identity matrix. The procedure is as follows:

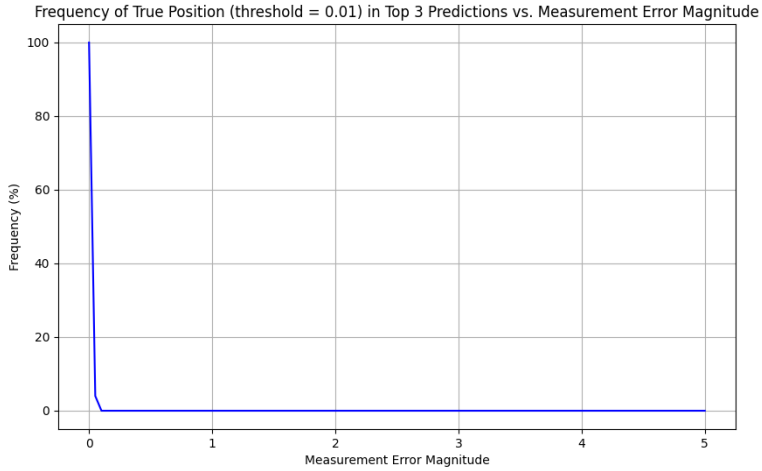


Figure 3.5: Frequency of true position (threshold = 0.01) in top 3 predictions vs. measurement error.

Name	Type	Params	Description
mlp_output_pose	MLP	387	MLP for pose output, mapping from d_model to 3
mlp_output_trans	MLP	387	MLP for translation output, mapping from d_model to 3
mlp_smpl_pose	MLP	9.3 K	MLP for SMPL pose, mapping from 72 to d_model
mlp_smpl_joints	MLP	9.3 K	MLP for SMPL joints, mapping from 72 to d_model
mlp_obj_pose	MLP	512	MLP for object pose, mapping from 3 to d_model
mlp_obj_trans	MLP	512	MLP for object translation, mapping from 3 to d_model
transformer_model_trans	Transformer	1.8 M	Transformer model for translation prediction
transformer_model_pose	Transformer	1.8 M	Transformer model for orientation prediction

Table 3.2: H2O-CA, Summary of number of parameters. Here, pose refers to orientation.

Let $(\mathbf{a}, \theta)$ be the axis-angle representation for rotation,
and let $\mathbf{t}_1$ and $\mathbf{t}_2$ be the translation vectors.

Normalize the rotation axis: $$\mathbf{a} = \frac{\mathbf{a}}{\|\mathbf{a}\|},$$

Construct the skew-symmetric matrix: $$\text{skew} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix},$$

Compute the rotation matrix (Rodrigues' formula): $$R = I + \sin(\theta) \cdot \text{skew} + (1 - \cos(\theta)) \cdot \text{skew}^2,$$

Form the 6D representation (first two columns of R) : $$R_{6D} = [R_{:,1}, R_{:,2}],$$

Compute MSE for the rotation part: $$\text{MSE}_{\text{rotation}} = \frac{1}{6} \sum_{i=1}^{6} (R_{6D_{1i}} - R_{6D_{2i}})^2,$$

Compute MSE for the translation part: $$\text{MSE}_{\text{translation}} = \frac{1}{3} \sum_{i=1}^{3} (t_{1i} - t_{2i})^2,$$

Calculate the total Loss (sum of the above errors): $$\text{Loss} = \text{MSE}_{\text{rotation}} + \text{MSE}_{\text{translation}}. \tag{3.2}$$

Two types of losses are considered, next-frame and next-4 frames. Next-n loss, in general, adds up the losses

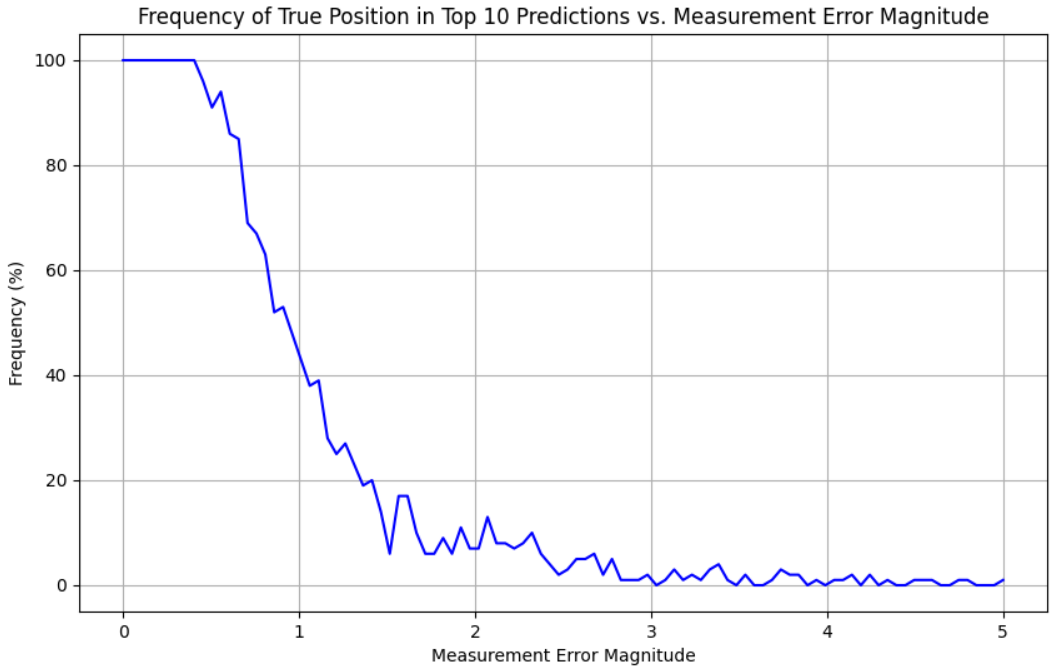


Figure 3.6: Frequency of true position (threshold = 0.01) in top 10 predictions vs. measurement error.

of the first n masked frames.

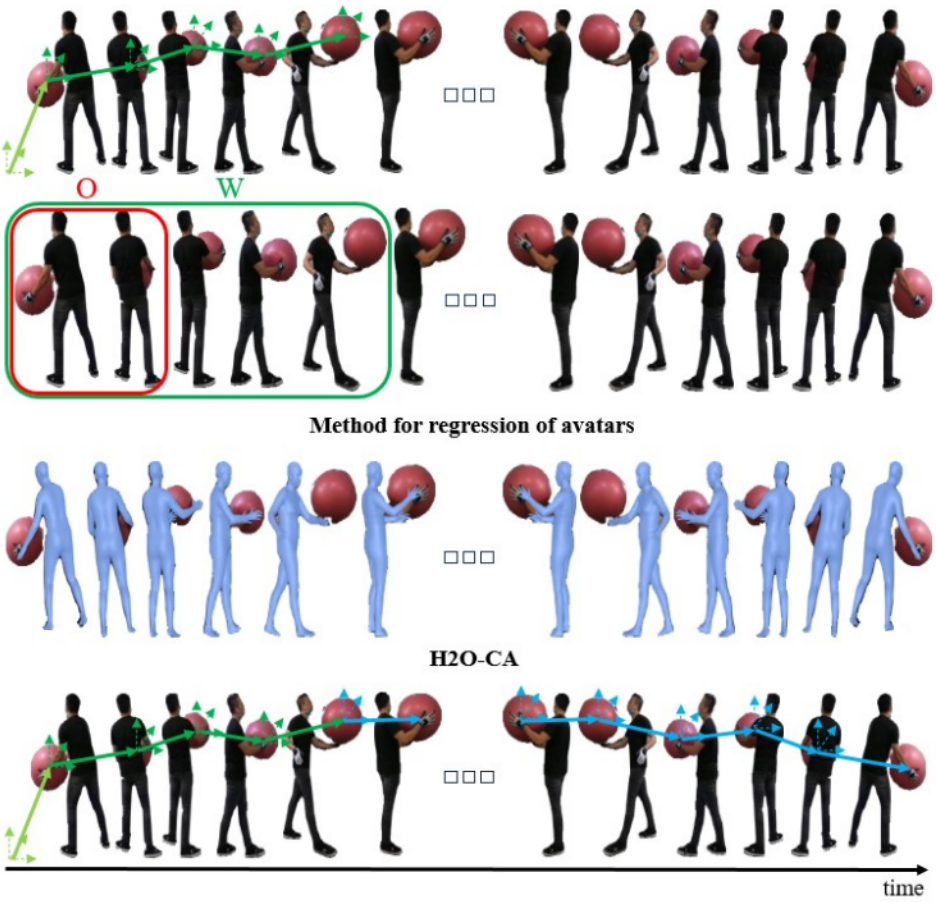


Figure 3.7: H2O-CA pipeline. In step 1, in a fully recursive approach, the first 8 frames of the video are equipped with an arbitrary reference frame, and successive relative offsets of the position and orientation of the object are computed, as in 3.3. In step 2, it is portrayed the sliding window W in input (width 12, offset 1), and the sliding window O of offsets (width 2, offset 1). In step 3, a method for regression of avatars has been applied ("TRACE"). In step 4, the regressive unit H2O-CA (see 3.8) yields, after hot initialization (green), fully recursive predictions (light blue).

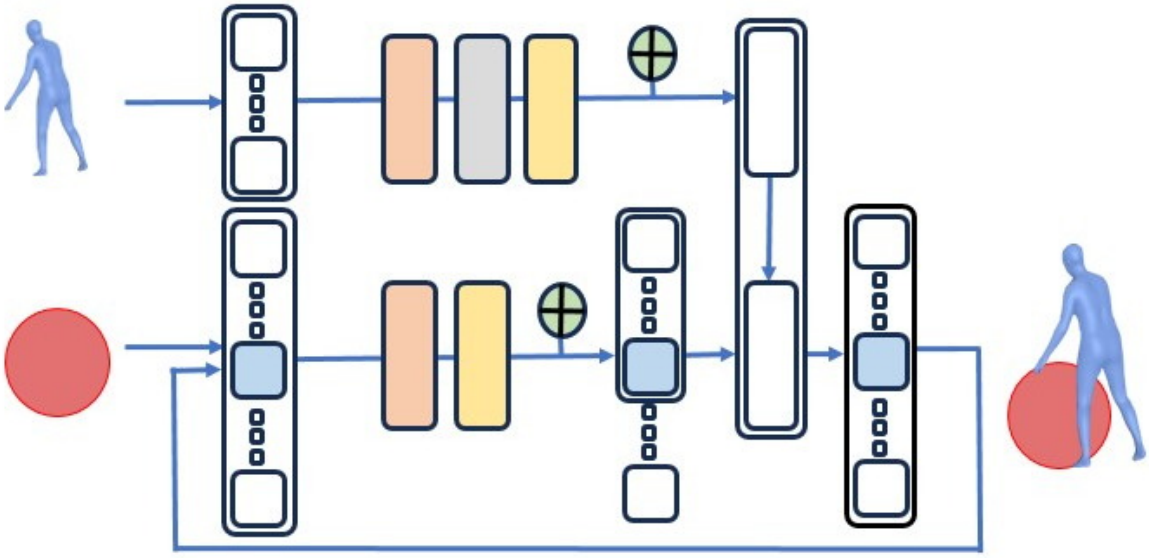


Figure 3.8: H2O-CA Inference setup: the video is processed in a sliding window fashion. Input tensors are flattened (in orange), normalized (gray), and linearly encoded to the hidden dimension of the transformer (128, in yellow). Positional encoding is added (green). Input to the decoder is masked (last 4 frames). The human pose is fed to the encoder, and the object's masked online predictions are fed to the decoder. Output to the transformer is linearly decoded and the 9th frame is fed as input to the next iteration (in a fully recursive approach).

Chapter 4

Experiments and Results

4.1 Training

A 2-stage strategy is adopted:

1. **Initial Training with Window Recursive Approach**: The model is first exposed to a subset of the dataset in a windowed fashion (one forward pass for each window). This strategy ensures that the model captures generalizable features without overly adapting to the particularities of any single scene-camera pair.

2. **Fine-Tuning with Full Sequence Recursive Setting**: Subsequently, the model is fine-tuned in a comprehensive sequence recursive framework (forward pass consists of a chain of 12 units), enhancing its ability to refine predictions based on the entire sequence context.

A detailed exposition of the rationale behind design choices is articulated in Section 5.

Scheduler

Schedulers are used to adjust the learning rate during training. Two types of schedulers are tried, opting for the second one:

- ReduceLROnPlateau Scheduler: this scheduler reduces the learning rate when a metric has stopped improving. The parameters like patience (5), factor (0.01), and threshold (0.75) define the conditions for the reduction. [23]

- CyclicLR Scheduler: this scheduler varies the learning rate between a base ($1e-7$) and a maximum value ($1e-4$) cyclically. [14]

Optimizer

Several optimizers are explored: the L-BFGS optimizer, as described in the study by [13], with a learning rate set by the scheduler, a maximum of 20 iterations per optimization step, and the 'strong_wolfe' line search function. Adaptive learning rate methods like Adagrad [6], with a learning rate set by the scheduler and weight decay of $1e-4$; RMSprop [8], with a learning rate set by the scheduler, decay rate α of 0.99, ε of $1e-08$, weight decay of $1e-4$, and momentum of 0.9; and Adadelta [41], with a learning rate set by the scheduler, decay rate ρ of 0.9, ε of $1e-06$, and weight decay of $1e-4$ are tried. Furthermore, the Adam optimizer, particularly its variant introduced by Loshchilov and Hutter [16], optimizes the model by

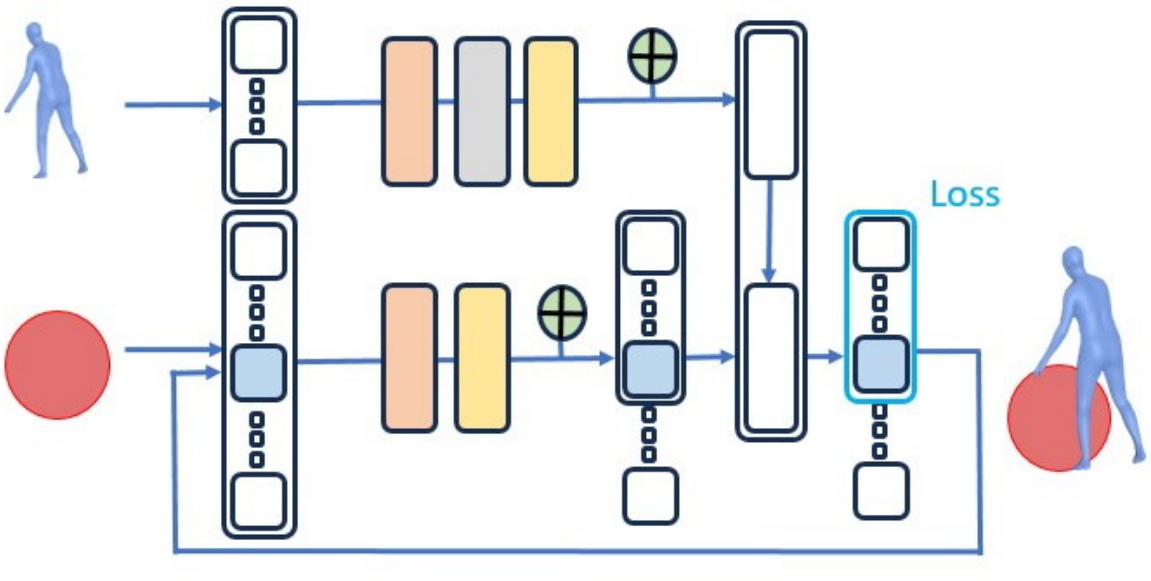


Figure 4.1: H2O-CA training setup: the loss function considers the last 4 predictions.

computing adaptive learning rates for each parameter, with a learning rate set by the scheduler, β parameters of (0.9, 0.999), and weight decay of $1e-4$. If no specific optimizer is matched, the default optimizer used is Adam with the same parameters as AdamW. AdamW is chosen.

System

The GPU resource mainly used is NVIDIA GeForce GTX TITAN X. The computational framework is PyTorch, with its extension PyTorch Lightning. The model's performance and training process are monitored and logged using Weights&Biases (wandb).

4.2 Evaluation metrics

Meshes are simplified via [11] to 1500 vertices for computational reasons.

Area Under the Curve (AUC)

The Area Under the Curve (AUC) is computed as the integral of the function $\mathrm{f}(\theta)$, representing the recall values over a range of threshold values. The AUC provides an aggregate measure of performance across different levels of strictness. Mathematically, it is expressed as:

$$\mathrm{AUC} = \int_{\theta_{\min}}^{\theta_{\max}} \mathrm{f}(\theta)\, d\theta \tag{4.1}$$

In this context:

- $\theta_{\min} = 0$ represents the minimum threshold, indicating the least permissive condition.
- $\theta_{\max} = 0.1$ denotes the maximum threshold, reflecting the most permissive condition (same as [37]).

- $\mathrm{f}(\theta) = \text{recall}(\text{ADD values of the scene-camera})$ is the function representing the recall values for Average Direct Distance (ADD) of the scene-camera pair, evaluated at different thresholds θ.

Chamfer Distance

$$\text{Chamfer Distance} = \frac{1}{N}\sum_{p\in P}\min_{\hat{p}\in\hat{P}}\|p-\hat{p}\|_2 + \frac{1}{M}\sum_{\hat{p}\in\hat{P}}\min_{p\in P}\|\hat{p}-p\|_2 \tag{4.2}$$

Chamfer Distance measures the average closest point distance between two sets of points P and $\hat{P}$, where P is the set of ground truth points and $\hat{P}$ is the set of predicted points. It is symmetric and measures the bidirectional closest point distances.

Average Direct Distance (ADD)

$$\text{ADD} = \frac{1}{N}\sum_{i=1}^{N}\|(Rp_i+T)-(\hat{R}\hat{p}_i+\hat{T})\|_2 \tag{4.3}$$

ADD computes the average Euclidean distance between corresponding points p_i on the model object and $\hat{p}_i$ on the predicted object, where N is the number of points considered. R and T represent the rotation matrix and translation vector of the ground truth pose, respectively, while $\hat{R}$ and $\hat{T}$ represent the rotation matrix and translation vector of the predicted pose.

Vertex-to-Vertex and Center Error

$$E_{v2v} = \|T-\hat{T}\|_F, \quad E_c = \|c(T)-c(\hat{T})\|_2 \tag{4.4}$$

Vertex-to-Vertex Error (E_{v2v}) measures the Frobenius norm difference between the predicted vertices set T and the ground truth vertices set $\hat{T}$. It evaluates the point-to-point discrepancy across the entire structure of the objects. On the other hand, Center Error (E_c) calculates the Euclidean distance between the centers of the predicted and the ground truth objects, denoted by $c(T)$ and $c(\hat{T})$ respectively. While E_{v2v} assesses the overall shape accuracy, E_c focuses on the positional accuracy of the object as a whole.

Mean Squared Error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \tag{4.5}$$

Mean Squared Error (MSE) quantifies the average of the squared differences between the predicted values $\hat{Y}_i$ and the actual observed values Y_i. It serves as a criterion for evaluating the performance of regression models, emphasizing the penalization of larger errors due to the squaring of each term. This characteristic ensures that larger discrepancies between predicted and observed values have a disproportionately high impact on the overall error metric, encouraging models to avoid significant deviations from the actual data points.

Procrustes Alignment Error

$$E_{proc} = \min_{R,t,s}\|sRT-\hat{T}+t\|_F \tag{4.6}$$

Procrustes Alignment Error (E_{proc}) quantifies the discrepancy between a predicted shape T and a ground truth shape $\hat{T}$ after optimal alignment. This alignment involves finding the best uniform scaling factor s,

rotation matrix R, and translation vector t that minimally distort T to match $\hat{T}$. The Frobenius norm in the equation calculates the point-to-point distance between the optimally aligned predicted shape and the ground truth, offering a measure of shape similarity that is invariant to scale, translation, and rotation.

4.3 Ablation Studies

Table 4.1 presents various ablation experiments. The first experiment aims to understand the performance degradation of the tracker as it makes inferences in increasingly larger batch sizes, n. Two batch sizes, n=1 and n=4, are examined. The training losses are, respectively, next-frame and next-4 frames, with an inference setup in which both the absolute pose of the object and the online predictions are reset to ground truth at the start of each batch. When used as a next-frame predictor, the model shows improvements of +22.2% in ADD and +1.05 cm in CD over the next-4 frame predictors (refer to rows 1 and 4). This result is expected: a next-4 frame setup reduces the number of ground truth (GT) parameters required by a factor of four, at the expense of accuracy. Performance further deteriorates (see row 2) when the training duration set in the loss, is shorter than the inference period (1 vs. 2).

The second experiment investigates whether using a loss that accounts for predictions beyond n (i.e., next-4 loss for n=1 at inference time) can be beneficial. The results (comparing rows 1 and 3) indicate significantly better performance with the next-frame loss trained H2O-CA, showing an increase of +5.51% in ADD and 0.72 cm in CD over its next-4 frames counterpart. This improvement can be attributed to the next-frame loss being optimized precisely for the inference setting.

Further, evaluation in a fully recursive setting (mimicking a proper tracker) was found to yield poor performance. Visual inspection revealed that the model is highly sensitive to its input, leading to rapid error accumulation. Consequently, a training setup where a chain of 12 blocks of H2O-CA forms a single computational graph was investigated (see row 5). The model, trained with a next-4 frames loss and evaluated in a next-4 frames setup, demonstrated a performance boost (an increase of +10.8% in ADD and +0.62 cm in CD, compare rows 4 and 5) when the forward (and backward) passes encompass multiple chained H2O-CA blocks.

Lastly, the significance of human pose was examined by removing its contribution and training a transformer encoder (from [34]) solely on object data. The results underscore the advantage (+1.79% in ADD and +0.54 cm in CD, compare rows 6 and 3) of integrating human cues into the model. Due to computational constraints, only camera 3 was considered in this analysis. The evaluations were conducted using ground truth templates of the objects.

Table 4.1: Summary of results of ablation.

Setup	Next-n Prediction	ADD (%) ↑	CD (cm) ↓
H2O-CA w/ next-frame loss	1	81.29	0.95
	4	53.80	2.20
H2O-CA	1	75.78	1.67
	4	59.09	2.00
H2O-CA chain	4	69.89	1.38
H2O-Encoder	1	73.99	1.13

4.4 Quantitative comparison

A comparison of different object trackers to H2O-CA is presented in Table A.2 in Appendix A.2. The model is trained with a next-4 frames loss setup and evaluated in a next-frame setup. As [37, 22, 39, 32, 32, 45, 45, 19, 2, 27] perform 3D reconstruction as well, H2O-CA uses ground truth templates of the objects. At inference time, the model performs next-frame prediction of the object offsets, in a non-recursive manner.

4.4.1 Baselines

Table 10 by Wen et al. is fully integrated. They compare against "DROID-SLAM" [32], "NICE-SLAM" [45], "KinectFusion: Real-Time Dense Surface Mapping and Tracking" [19], "BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models" [36], and "SDF-2-SDF Registration for Real-Time 3D Reconstruction from RGB-D Data" [27] using the respective open-source implementations with the best-tuned parameters. They additionally include the baseline results from their leaderboard. The inputs to these evaluated methods are the RGBD video and the first frame's mask indicating the object of interest. Wen et al. augment the comparison methods with the same video segmentation masks used in their framework for a fair comparison, to focus on 6-DoF object pose tracking and 3D reconstruction performance. In the case of tracking failure, no re-initialization is performed to test long-term tracking robustness. Since "DROID-SLAM" and "BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models" cannot reconstruct an object mesh, they augment these methods with TSDF Fusion ([5, 43]) for shape reconstruction evaluation. For "NICE-SLAM" and "BundleSDF", Wen et al. initialize the neural volume's bound using only the first frame's point cloud.

Results presented in "Object pop-up" [22] (Table 1) are also integrated for comparison. "Object pop-up" [22] proposes a simple yet informative baseline. Given the input point cloud, they recover the most similar in the training dataset in an L2 sense. Then, they recover the object handled by that subject and pose it in space in the same way. Here, the center-to-center error (see 4.2) is compared to the mean squared error between centers. For more details please see [22].

Table 4.2: Comparison of NN, Object pop-up, and H2O-CA across different datasets and metrics.

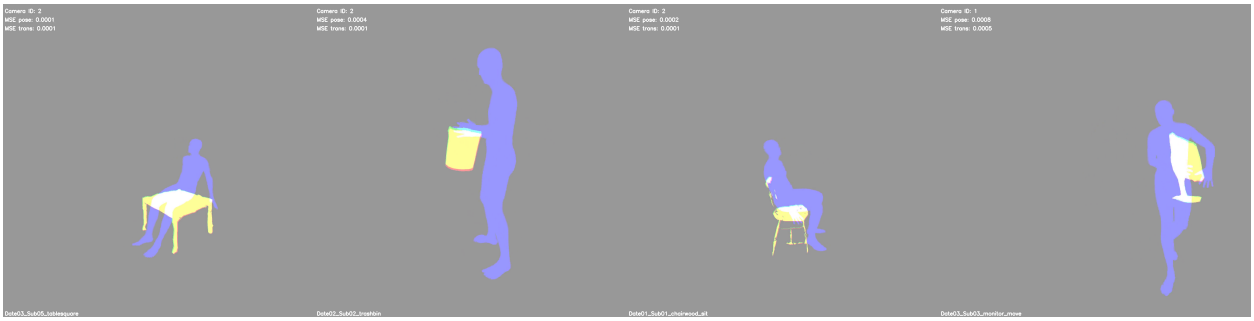
Methods	**GRAB**		**BEHAVE**		**BEHAVE-Raw**	
	$\mathbf{E_c} \downarrow$	$\mathbf{E_{v2v}} \downarrow$	$\mathbf{E_c} \downarrow$	$\mathbf{E_{v2v}} \downarrow$	$\mathbf{E_c} \downarrow$	$\mathbf{E_{v2v}} \downarrow$
NN Baseline [22]	0.0362	0.1445	0.0802	0.3445	-	-
"Object pop-up" [22]	0.0237	0.0943	0.0663	0.2900	0.0806	0.3143
H20-CA	-	-	**0.01965**	-	-	-

Another baseline by "Visibility Aware Human-Object Interaction Tracking from Single RGB Camera" [39] is reported (Table 1). Xie et al. evaluate the performance of human and object reconstruction using Chamfer distance (see 4.2) between predicted SMPL and object meshes, and their relative ground truths. "CHORE: Contact, Human and Object REconstruction from a single RGB image" [38] uses Procrustes alignment (see 4.6) on combined SMPL and object meshes for each frame before computing errors. As this does not reflect the real accuracy in terms of the relative translation between nearby frames in a video, joint Procrustes alignment in a sliding window is performed by the authors. More specifically, they combine all SMPL and object vertices within a sliding window and compute a single optimal Procrustes alignment to the ground truth vertices. This alignment is then applied to all SMPL and object vertices within this window. Table4.3 reports both the errors using per-frame alignment (w=1) and alignment with a sliding window of 10s (w=10).

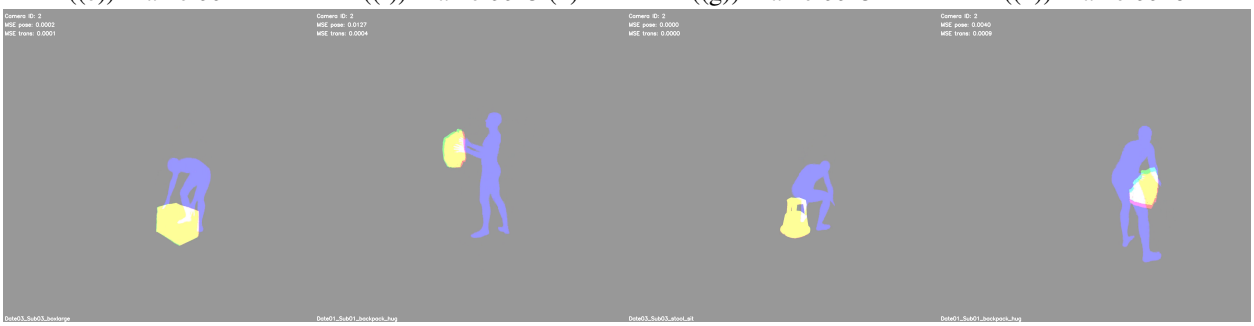
Table 4.3: Human and object tracking results on BEHAVE dataset (unit: cm). w is the temporal window size used for Procrustes alignment where $w = 1$ means per-frame Procrustes and $w = 10$ means alignment over a sliding window of 10s.

Dataset	**Methods**	**Align w=1**		**Align w=10**	
		SMPL ↓	**Obj. ↓**	**SMPL ↓**	**Obj. ↓**
BEHAVE	CHORE [38]	5.55	10.02	18.33	20.32
	VAHOIT [39]	5.25	8.04	7.81	8.49
	H2O-CA	-	**0.02**	-	-

((a)) Frame 0006 ((b)) Frame 0008 ((c)) Frame 0009 ((d)) Frame 0001



((e)) Frame 0014 ((f)) Frame 0015 (1) ((g)) Frame 0015 ((h)) Frame 0016



((i)) Frame 0022 ((j)) Frame 0023 ((k)) Frame 0001 ((l)) Frame 0018

Figure 4.2: Collection of frames. Qualitative results of H2O-CA. Simplified GT mesh in green, simplified prediction mesh in red. GT human mesh in blue.

Chapter 5

Discussion

In this section, an analysis of the performance and an explanation of the most important design choices of H2O-CA made throughout this work is provided.

5.0.1 Performance analysis of H2O-CA

The H2O-CA method has demonstrated notable performance across various metrics compared to other well-established methods. However, the assumptions underpinning this method are significant: it requires the ground truth object template and the ground truth human and object parameters of all frames to perform inference. This requirement explains the strong results. Focusing on the ADD metric, which is crucial for accurate object tracking, H2O-CA consistently shows competitive performance. For instance, in Date03_Sub03_chairblack_lift.1, it achieves an ADD score of 79.50%, while the next best method, Bundle, scores 11.43%. This superiority is explained by considering the complete object shape (the object has thin wooden slats on the back, which are difficult to reconstruct in 3D). There are other videos in which the object template is regular, yet other methods perform poorly. Some of these cases are illustrated [here]. For example, Date03_Sub05_boxlarge is tracked with 20.02 % ADD as second-best. The video presents several occlusions caused by the subject. For almost every listed sequence, H2O-CA maintains a CD score of around 0.02 cm, showcasing its precision in maintaining the structural integrity of the object during tracking. Similar considerations hold for the other comparisons. However, even with these impressive results, a relaxation of the next-frame prediction setting, up to a fully recursive approach, has not yet reached state-of-the-art performance.

5.1 Independence of location and orientation

In the context of human-object interaction, the 6-DoF representation of an object with respect to the human body varies: in the scenario of a static posture (illustrated in Figure 5.1, left image), the ideal regressor should behave like the identity function for both position/orientation when focusing on any joint. In the second scenario (right image), a dynamic interaction unfolds as the individual lifts an object. In this case, the offset in orientation in the ankle joint results in an upward movement of the object, exemplified here by a briefcase. Its offset in orientation is null, while its offset in position varies (approximately) linearly with the ankle's offset in orientation.

While rotation and translation are independent quantities in classical mechanics, they are linearly related in a hierarchical parametric model of the human body (such as SMPL), or of a robotic arm: for the root joint J_1, its child J_2, and the child's child J_3, a rotation of the root joint J_1 affects the global positions and orientations of J_2 and J_3. Letting R_1, R_2, and R_3 be the rotation matrices and $\mathbf{t_1}, \mathbf{t_2}$, and $\mathbf{t_3}$ be the

Figure 5.1: Two interactions.

translation vectors for joints J_1, J_2, and J_3 respectively, the child joint J_3 has global transformation:

$$T_3 = \begin{bmatrix} R_1 & \mathbf{t_1} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} R_2 & \mathbf{t_2} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} R_3 & \mathbf{t_3} \\ 0 & 1 \end{bmatrix}$$

The design choice to separate the prediction of orientation and position in two branches is due to the different domains of the types of offsets, and the fact that the mechanism for which a rotation influences a translation, and vice versa, is observed for all the joints lower in the hierarchy (the hands, in the example above have similar offsets in position/orientation to the briefcase). As the offsets of the elements at the base of the hierarchies (hands, feet, and head) can be used as a proxy of those of the object, the design choice to limit this type of redundant information has been made.

5.2 Use of offsets

Most recent work uses offsets: *Pose2Room* [20], published in 2022, leverages human motion and interactions as key indicators, classifying objects and predicting their 3D bounding boxes using observed human trajectories. They use a relative position encoder in two aspects – a relative position encoding of the joints to the hip (root joint) and a pose feature that is independent of position. This approach is rooted in the understanding that interactions between humans and objects are typically confined to local areas within a scene.

5.3 Use of masks

The choice of masking the input is inspired by the unit HVOP-Net [39]. The authors obtain more robust rotation under heavy occlusions with respect to infilling the object pose of invisible frames using spherical linear interpolation between two visible frames (SLERP), SmoothNet [42], a fully connected neural network trained to smooth the object motion, or CMIB [12], a SoTA method for human motion infilling. They use masking only on frames where the tracked object is occluded beyond a set threshold, whereas H2O-CA indiscriminately masks the last 4 frames of a window.

In addition, as self-attention is permutation-invariant to the tokens, and a window is an ordered set of frames, positional encoding is added to the input.

5.4 Choice of modality for human representation

Literature covers several modalities for human representation (point cloud, RGB, segmentation mask, depth map, IMUs). The choice of neglecting RGB input serves the setting of an AR/VR experience where offsets are registered by a body suit with mounted sensors and instances like showing a skill remotely while moving physical objects, changing environments, and navigating a building, should be replicated as full interactions

including the person and the scene changes (although possibly not be visible) on the AR/VR device. A tracking solution based solely on wearable sensors, however heavily engineered and class-dependent, comes from *Interaction Replica* [9]: the process involves automatically detecting contacts from the human pose alone, correcting the human trajectory to satisfy the physical contact, and inferring the 3D object pose trajectory coherent with the human motion. Degrees of freedom of an object are assumed to be known, and the method differs between use cases (object that slides, object hinged). H20-CA in its fully recursive version requires offsets of the joints (and initial object pose), not necessarily obtained from a method for avatar regression from video, thus being quite light in requirements. Other methods use less demanding settings like in [37] the object is assumed to be segmented in the first frame only, and unknown and the video is monocular RGBD, or in "Object pop-up"[22] an input point cloud and an object class are assumed.

5.4.1 Human Mesh Choice

Although the utilization of the SMPL model [15] imposes certain limitations on the quality of human mesh reconstruction (HMR), in principle, a more sophisticated model like SMPL-X [21]—which offers increased articulation, enhanced hand and foot models, improved expressivity, and greater body shape variation—should yield higher performance. However, "Object pop-up" [22] presents an ablation study that excludes the points from the hands (setting them in a rest pose) to focus on the remaining points of the subject. This method allows for an analysis of how much the network relies on finger pose information. Point saliency analysis evaluates how changes in each part's position influence the model's output by considering the magnitude of the backpropagated gradients. The findings indicate that the head and feet are more significant than the hands' joints for object pose regression. Given this insight, and considering the trade-off between model complexity and performance, a simpler model (SMPL with 85 parameters) is preferred over a more complex one (SMPL-X with 179 parameters). This decision is based on the point saliency results, which suggest that the additional expressivity offered by SMPL-X does not significantly contribute to the performance in the context of this specific task.

5.5 Limitations and future work

The trained methods assume contact between the human and the object. As the offset prediction is based on human movement, whenever the human-object interaction is not physical, the problem is ill-posed and this method fails. One possible solution is presented in [9], where contact intervals (active time windows for tracking) are predicted based on hand movement. Test time adaptation could be necessary for dealing with interactions not covered by the training data (lifting, carrying, sitting, pushing, and pulling with hands and feet). A more robust training procedure is essential for achieving a fully recursive approach.

5.6 Applications

Object tracking technology, with its wide-ranging applications across various sectors such as sports analysis, virtual realities (VR, MR, AR, and the Metaverse), healthcare, robotics, animation, security, workplace safety, driver assistance, human-computer interaction, and forensics, holds the potential to revolutionize efficiency, safety, and user experience. From enhancing sports performance analysis and immersive gaming experiences to aiding patient rehabilitation, ensuring workplace safety, and improving public security, the implications of accurate and reliable object tracking are profound. However, the implementation of this technology is not without its challenges and ethical considerations. Concerns about privacy, data security, potential biases, autonomy, and the risk of dehumanization necessitate a cautious and well-regulated approach. Ensuring the protection of individual privacy, securing sensitive data, minimizing biases, obtaining

informed consent, and maintaining a balance between technological and human care are crucial. Moreover, adherence to local and international regulatory standards is paramount. While object tracking technology offers numerous benefits, a responsible and ethical implementation, which respects individual rights and aligns with societal norms, is essential for harnessing its full potential.

Chapter 6

Conclusion

In conclusion, the H2O-CA (Human to Object – Cross Attention) framework presents a novel approach to 6-DoF object tracking in RGB video, utilizing a sequence-to-sequence methodology that significantly leverages human pose information. Despite the fully recursive tracking approach not achieving state-of-the-art results, the potential unearthed by next-frame prediction methods is noteworthy. Central to this method is the recognition that human actions frequently determine the trajectory of objects in real-world situations. This hints at the human figure as a proxy for the object's pose, especially under conditions of occlusion. This intuition is verified in the ablation study, showing +1.79% in ADD and +0.54 cm in CD on BEHAVE. The comparative analysis with established object-tracking methodologies accentuates H2O-CA's robustness and adaptability, with its promising results, marked +9.82% ADD and +4.64 cm CD on the mean.

Looking forward, it is hoped that this work will foster further research in this direction. A more robust training procedure is essential for achieving a fully recursive approach. However, considerable progress has been made by exploring various loss functions, training configurations, and design choices. These insights mark substantial steps toward realizing a comprehensive and reliable tracking framework.

As a commitment to advancing the field and encouraging collaborative efforts, the project's codebase will be made available [here].

Appendix A

Tables of comparison

A.1 Comparison of methods for regression of avatars (5D)

Video	TRACE	ROMP	BEV	4DH
Overall				
Mean	**3.67**	4.60	4.44	4.60
Per scene				
Date03_Sub04_backpack_back	**2.07**	4.47	4.39	3.83
Date03_Sub05_chairblack	5.39	4.28	4.08	**3.91**
Date03_Sub03_boxmedium	**3.60**	4.70	4.59	4.51
Date03_Sub04_keyboard_move	**4.34**	4.58	4.41	5.01
Date03_Sub04_tablesmall_hand	**3.55**	4.79	4.63	3.98
Date03_Sub04_stool_move	**3.68**	4.80	4.67	4.37
Date03_Sub05_stool	**3.54**	4.71	4.52	3.88
Date03_Sub03_toolbox	**3.24**	4.65	4.54	4.18
Date03_Sub03_basketball	3.94	4.24	4.11	**3.55**
Date03_Sub04_chairblack_liftreal	4.67	4.63	4.50	**3.87**
Date03_Sub04_yogamat	**3.82**	4.57	4.45	11.65
Date03_Sub03_boxtiny	**3.49**	4.27	4.18	7.62
Date03_Sub04_keyboard_typing	**3.84**	5.08	4.82	4.04
Date03_Sub04_yogaball_sit	**3.61**	4.74	4.53	4.89
Date03_Sub04_chairblack_sit	**4.38**	4.87	4.64	5.04
Date03_Sub03_stool_sit	5.23	5.25	4.99	**4.63**
Date03_Sub03_chairblack_sitstand	**4.99**	5.18	5.00	5.80
Date03_Sub04_chairwood_lift	**2.73**	4.73	4.59	3.99
Date03_Sub04_boxmedium	**2.75**	4.81	4.68	4.03
Date03_Sub04_boxlarge	**3.72**	4.57	4.43	3.90
Date03_Sub03_suitcase_move	**3.41**	4.46	4.31	3.71
Date03_Sub04_backpack_hand	**3.90**	4.73	4.64	4.51
Date03_Sub03_trashbin	4.00	4.42	4.28	**3.60**
Date03_Sub05_boxmedium	**2.71**	4.50	4.35	3.63
Date03_Sub03_monitor_move	**1.40**	3.43	3.20	3.00
Date03_Sub05_chairwood	**3.20**	4.43	4.27	3.67
Date03_Sub05_monitor	**1.70**	4.13	3.92	3.26
Date03_Sub05_keyboard	**3.68**	4.90	4.73	3.90
Date03_Sub04_chairwood_hand	**3.86**	4.52	4.37	4.57
Date03_Sub05_basketball	4.86	4.33	4.18	**3.31**

Date03_Sub04_trashbin	**3.29**	4.54	4.40	4.01
Date03_Sub03_tablesmall_move	**2.88**	4.31	4.12	3.76
Date03_Sub03_chairwood_lift	3.63	4.13	3.96	**3.43**
Date03_Sub04_tablesquare_sit	**3.10**	4.88	4.63	4.29
Date03_Sub03_chairwood_hand	**3.60**	4.04	3.86	5.67
Date03_Sub04_stool_sit	4.34	4.92	4.65	**4.26**
Date03_Sub03_chairblack_sit	**4.06**	6.13	5.84	6.61
Date03_Sub05_yogamat	**3.29**	4.66	4.54	3.73
Date03_Sub05_suitcase	4.53	4.65	4.53	**3.83**
Date03_Sub04_tablesquare_hand	**4.03**	5.11	4.94	4.42
Date03_Sub03_yogaball_play	**2.48**	4.45	4.27	6.22
Date03_Sub04_boxsmall	**3.17**	4.60	4.51	3.93
Date03_Sub04_tablesmall_lift	**3.66**	4.81	4.70	3.93
Date03_Sub04_chairwood_sit	**2.76**	4.51	4.25	3.91
Date03_Sub05_boxtiny	**2.80**	4.38	4.24	3.48
Date03_Sub05_tablesquare	**2.22**	4.79	4.60	4.02
Date03_Sub04_toolbox	**3.04**	4.59	4.45	3.88
Date03_Sub05_boxsmall	4.35	4.43	4.32	**3.71**
Date03_Sub03_backpack_back	**3.22**	4.25	4.14	3.42
Date03_Sub03_keyboard_move	**3.05**	4.32	4.11	3.82
Date03_Sub04_monitor_hand	4.98	5.11	4.92	**4.16**
Date03_Sub04_tablesquare_lift	**3.90**	4.58	4.45	3.96
Date03_Sub04_basketball	4.32	4.74	4.65	**3.86**
Date03_Sub03_backpack_hand	**3.65**	4.35	4.21	4.34
Date03_Sub04_yogaball_play	**2.57**	4.03	3.86	5.76
Date03_Sub03_boxlarge	**2.81**	3.99	3.85	13.87
Date03_Sub03_yogamat	3.75	4.57	4.46	**3.73**
Date03_Sub04_suitcase_lift	**3.28**	4.62	4.51	3.92
Date03_Sub03_chairblack_hand	4.68	4.47	4.27	**3.82**
Date03_Sub03_tablesmall_lean	4.32	4.96	4.82	**4.24**
Date03_Sub04_suitcase_ground	**2.88**	4.53	4.43	3.88
Date03_Sub05_backpack	4.24	4.51	4.37	**3.62**
Date03_Sub05_toolbox	4.11	4.82	4.68	**3.93**
Date03_Sub03_stool_lift	**3.65**	4.76	4.64	9.29
Date03_Sub04_boxlong	9.71	5.02	**4.87**	4.97
Date03_Sub03_keyboard_typing	**1.67**	4.94	4.67	4.11
Date03_Sub04_monitor_move	4.12	5.02	4.88	**3.91**
Date03_Sub05_plasticcontainer	4.11	4.54	4.42	**3.64**
Date03_Sub03_plasticcontainer	**3.01**	4.72	4.63	3.68
Date03_Sub03_tablesquare_lift	**2.48**	4.81	4.71	3.92
Date03_Sub03_chairblack_lift	**3.26**	3.85	3.70	5.23
Date03_Sub05_yogaball	**2.99**	4.13	3.94	3.45
Date03_Sub04_chairblack_hand	**2.96**	4.66	4.50	4.14
Date03_Sub05_boxlong	4.01	4.48	4.36	**3.65**
Date03_Sub03_suitcase_lift	**3.66**	4.62	4.51	10.29
Date03_Sub05_trashbin	**2.65**	4.19	4.05	3.44
Date03_Sub04_plasticcontainer_lift	4.48	4.56	4.45	**3.90**
Date03_Sub03_boxsmall	**3.13**	4.12	4.00	6.74
Date03_Sub03_backpack_hug	**3.67**	4.50	4.37	4.27
Date03_Sub05_boxlarge	4.82	4.67	4.56	**3.91**
Date03_Sub04_boxtiny	**3.80**	4.74	4.65	3.94

Date03_Sub04_tablesmall_lean	**3.75**	4.50	4.34	3.78
Date03_Sub03_tablesmall_lift	**3.04**	4.44	4.27	7.53
Date03_Sub03_tablesquare_sit	4.75	5.12	4.84	**4.51**
Date03_Sub03_yogaball_sit	**4.43**	4.82	4.59	11.09
Date03_Sub05_tablesmall	3.99	4.88	4.75	**3.98**
Date03_Sub03_tablesquare_move	**2.29**	3.99	3.84	3.69
Date03_Sub03_chairwood_sit	4.80	5.48	5.29	**4.69**
Date03_Sub03_boxlong	**4.03**	4.30	4.18	10.25
Date03_Sub04_backpack_hug	**3.56**	4.99	4.85	4.25
Per scene-camera				
Date03_Sub04_backpack_back.2	**2.06**	4.40	4.31	3.64
Date03_Sub04_backpack_back.0	**2.18**	4.61	4.52	3.59
Date03_Sub04_backpack_back.3	**2.07**	4.67	4.55	3.93
Date03_Sub04_backpack_back.1	**2.12**	4.28	4.25	4.18
Date03_Sub05_chairblack.1	5.36	4.25	**4.07**	4.27
Date03_Sub05_chairblack.3	5.39	4.39	4.19	**3.59**
Date03_Sub05_chairblack.2	5.50	4.13	3.93	**3.73**
Date03_Sub05_chairblack.0	5.69	4.41	4.21	**4.00**
Date03_Sub03_boxmedium.0	**3.63**	4.39	4.33	3.71
Date03_Sub03_boxmedium.2	**3.55**	5.01	4.87	3.76
Date03_Sub03_boxmedium.3	**3.60**	5.06	4.92	3.73
Date03_Sub03_boxmedium.1	**3.55**	4.39	4.34	6.69
Date03_Sub04_keyboard_move.2	**4.29**	4.60	4.41	6.23
Date03_Sub04_keyboard_move.0	4.35	4.56	4.45	**3.60**
Date03_Sub04_keyboard_move.1	4.32	4.35	**4.20**	5.92
Date03_Sub04_keyboard_move.3	4.34	4.84	4.67	**3.95**
Date03_Sub04_tablesmall_hand.1	**3.59**	4.89	4.73	4.25
Date03_Sub04_tablesmall_hand.3	**3.55**	4.72	4.60	3.92
Date03_Sub04_tablesmall_hand.2	**3.53**	4.59	4.37	3.92
Date03_Sub04_tablesmall_hand.0	**3.68**	4.95	4.84	3.78
Date03_Sub04_stool_move.3	**3.68**	4.74	4.60	3.78
Date03_Sub04_stool_move.1	**3.62**	4.79	4.67	5.79
Date03_Sub04_stool_move.2	**3.74**	4.63	4.48	3.86
Date03_Sub04_stool_move.0	**3.65**	5.04	4.96	3.85
Date03_Sub05_stool.0	**3.46**	4.64	4.45	3.71
Date03_Sub05_stool.1	**3.56**	4.65	4.44	3.83
Date03_Sub05_stool.3	**3.54**	4.81	4.64	3.87
Date03_Sub05_stool.2	**3.54**	4.76	4.55	4.09
Date03_Sub03_toolbox.1	**3.20**	4.46	4.39	5.92
Date03_Sub03_toolbox.3	**3.24**	4.81	4.70	3.55
Date03_Sub03_toolbox.0	**3.18**	4.68	4.59	3.58
Date03_Sub03_toolbox.2	**3.23**	4.65	4.54	3.59
Date03_Sub03_basketball.0	3.86	3.96	3.85	**3.02**
Date03_Sub03_basketball.1	3.88	3.81	**3.72**	4.41
Date03_Sub03_basketball.3	3.94	4.68	4.52	**3.37**
Date03_Sub03_basketball.2	3.88	4.61	4.46	**3.37**
Date03_Sub04_chairblack_liftreal.3	4.67	4.85	4.70	**3.78**
Date03_Sub04_chairblack_liftreal.1	4.65	4.46	4.34	**4.33**
Date03_Sub04_chairblack_liftreal.2	4.62	4.44	4.30	**3.71**
Date03_Sub04_chairblack_liftreal.0	4.63	4.84	4.73	**3.62**
Date03_Sub04_yogamat.1	**3.81**	4.43	4.30	32.57

Date03_Sub04_yogamat.3	3.82	4.68	4.51	**3.71**
Date03_Sub04_yogamat.0	3.79	4.69	4.61	**3.47**
Date03_Sub04_yogamat.2	3.85	4.53	4.45	**3.49**
Date03_Sub03_boxtiny.0	3.50	4.16	4.11	**3.18**
Date03_Sub03_boxtiny.1	**3.44**	4.16	4.06	10.76
Date03_Sub03_boxtiny.2	**3.48**	4.35	4.28	12.16
Date03_Sub03_boxtiny.3	3.49	4.44	4.29	**3.43**
Date03_Sub04_keyboard_typing.1	**3.82**	4.42	4.06	4.15
Date03_Sub04_keyboard_typing.3	**3.84**	5.73	5.55	4.00
Date03_Sub04_keyboard_typing.0	3.83	4.71	4.44	**3.75**
Date03_Sub04_keyboard_typing.2	**4.01**	5.55	5.35	4.22
Date03_Sub04_yogaball_sit.3	**3.61**	4.81	4.65	4.35
Date03_Sub04_yogaball_sit.0	**3.57**	5.11	4.95	4.03
Date03_Sub04_yogaball_sit.2	**3.65**	4.56	4.30	4.26
Date03_Sub04_yogaball_sit.1	**3.54**	4.52	4.32	6.73
Date03_Sub04_chairblack_sit.1	**4.33**	4.66	4.46	6.85
Date03_Sub04_chairblack_sit.0	4.43	4.79	4.59	**3.84**
Date03_Sub04_chairblack_sit.3	4.38	5.16	4.93	**4.30**
Date03_Sub04_chairblack_sit.2	**4.36**	4.93	4.65	5.02
Date03_Sub03_stool_sit.3	5.23	5.04	4.78	**4.36**
Date03_Sub03_stool_sit.1	5.23	5.46	**5.17**	5.35
Date03_Sub03_stool_sit.2	5.18	5.13	4.86	**4.46**
Date03_Sub03_stool_sit.0	5.17	5.33	5.09	**4.30**
Date03_Sub03_chairblack_sitstand.1	**4.94**	5.40	5.19	10.88
Date03_Sub03_chairblack_sitstand.3	4.99	4.92	4.64	**4.30**
Date03_Sub03_chairblack_sitstand.2	4.92	5.05	4.96	**4.19**
Date03_Sub03_chairblack_sitstand.0	4.93	5.33	5.11	**3.73**
Date03_Sub04_chairwood_lift.2	**2.84**	4.88	4.75	3.81
Date03_Sub04_chairwood_lift.3	**2.73**	5.17	4.98	4.01
Date03_Sub04_chairwood_lift.0	**2.69**	4.56	4.47	3.67
Date03_Sub04_chairwood_lift.1	**2.71**	4.40	4.25	4.43
Date03_Sub04_boxmedium.0	**2.69**	4.87	4.75	3.67
Date03_Sub04_boxmedium.1	**2.71**	4.77	4.64	4.54
Date03_Sub04_boxmedium.3	**2.75**	4.88	4.76	4.00
Date03_Sub04_boxmedium.2	**2.73**	4.73	4.61	3.84
Date03_Sub04_boxlarge.1	**3.77**	4.30	4.18	4.39
Date03_Sub04_boxlarge.2	**3.71**	4.59	4.45	3.78
Date03_Sub04_boxlarge.3	**3.72**	4.80	4.67	3.89
Date03_Sub04_boxlarge.0	3.74	4.64	4.51	**3.53**
Date03_Sub03_suitcase_move.2	**3.46**	4.65	4.48	3.56
Date03_Sub03_suitcase_move.1	**3.36**	4.20	4.06	4.08
Date03_Sub03_suitcase_move.0	**3.36**	4.34	4.19	3.45
Date03_Sub03_suitcase_move.3	**3.41**	4.71	4.58	3.75
Date03_Sub04_backpack_hand.3	**3.90**	4.66	4.56	5.34
Date03_Sub04_backpack_hand.1	**3.88**	4.73	4.67	5.13
Date03_Sub04_backpack_hand.2	**3.86**	4.74	4.65	4.02
Date03_Sub04_backpack_hand.0	3.83	4.79	4.68	**3.65**
Date03_Sub03_trashbin.1	**3.95**	4.25	4.15	4.07
Date03_Sub03_trashbin.2	3.98	4.37	4.21	**3.41**
Date03_Sub03_trashbin.0	3.97	4.53	4.38	**3.43**
Date03_Sub03_trashbin.3	4.00	4.56	4.43	**3.48**

Date03_Sub05_boxmedium.2	**2.72**	4.60	4.45	3.64
Date03_Sub05_boxmedium.0	**2.69**	4.42	4.25	3.50
Date03_Sub05_boxmedium.3	**2.71**	4.67	4.49	3.72
Date03_Sub05_boxmedium.1	**2.69**	4.35	4.24	3.66
Date03_Sub03_monitor_move.3	**1.40**	3.45	3.32	3.04
Date03_Sub03_monitor_move.2	**1.38**	3.43	3.21	2.75
Date03_Sub03_monitor_move.0	**1.34**	3.40	3.13	2.81
Date03_Sub03_monitor_move.1	**1.46**	3.44	3.19	3.42
Date03_Sub05_chairwood.1	**3.27**	4.32	4.16	3.62
Date03_Sub05_chairwood.3	**3.20**	4.53	4.35	3.90
Date03_Sub05_chairwood.0	**3.16**	4.46	4.30	3.45
Date03_Sub05_chairwood.2	**3.18**	4.43	4.28	3.73
Date03_Sub05_monitor.1	**1.72**	3.68	3.48	3.10
Date03_Sub05_monitor.2	**1.80**	4.20	3.95	3.55
Date03_Sub05_monitor.3	**1.70**	4.59	4.40	3.33
Date03_Sub05_monitor.0	**1.71**	4.12	3.98	3.07
Date03_Sub05_keyboard.0	**3.67**	4.60	4.47	3.91
Date03_Sub05_keyboard.3	**3.68**	5.21	5.04	3.88
Date03_Sub05_keyboard.2	**3.63**	5.29	5.04	3.92
Date03_Sub05_keyboard.1	**3.75**	4.55	4.41	3.90
Date03_Sub04_chairwood_hand.3	**3.86**	4.56	4.44	3.97
Date03_Sub04_chairwood_hand.1	**3.92**	4.49	4.28	6.47
Date03_Sub04_chairwood_hand.2	**3.83**	4.44	4.31	3.89
Date03_Sub04_chairwood_hand.0	4.38	4.61	4.48	**3.74**
Date03_Sub05_basketball.3	4.86	4.40	4.27	**3.19**
Date03_Sub05_basketball.1	4.84	4.08	3.95	**3.27**
Date03_Sub05_basketball.2	4.83	4.58	4.36	**3.52**
Date03_Sub05_basketball.0	4.90	4.26	4.17	**3.23**
Date03_Sub04_trashbin.3	**3.29**	4.75	4.59	3.97
Date03_Sub04_trashbin.2	**3.25**	4.40	4.30	3.77
Date03_Sub04_trashbin.0	**3.24**	4.67	4.53	3.69
Date03_Sub04_trashbin.1	**3.27**	4.38	4.28	4.61
Date03_Sub03_tablesmall_move.1	**2.86**	4.09	3.92	4.55
Date03_Sub03_tablesmall_move.0	**2.83**	4.39	4.21	3.31
Date03_Sub03_tablesmall_move.2	**2.99**	4.26	4.04	3.47
Date03_Sub03_tablesmall_move.3	**2.88**	4.54	4.38	3.60
Date03_Sub03_chairwood_lift.3	3.63	4.42	4.22	**3.48**
Date03_Sub03_chairwood_lift.0	3.64	3.86	3.76	**3.10**
Date03_Sub03_chairwood_lift.1	**3.58**	3.89	3.71	3.71
Date03_Sub03_chairwood_lift.2	3.60	4.41	4.21	**3.43**
Date03_Sub04_tablesquare_sit.0	**3.08**	4.88	4.71	3.87
Date03_Sub04_tablesquare_sit.1	**3.04**	4.91	4.63	4.80
Date03_Sub04_tablesquare_sit.3	**3.10**	4.94	4.65	4.34
Date03_Sub04_tablesquare_sit.2	**3.09**	4.79	4.54	4.13
Date03_Sub03_chairwood_hand.0	3.52	3.99	3.85	**3.05**
Date03_Sub03_chairwood_hand.2	**3.62**	4.16	4.00	8.56
Date03_Sub03_chairwood_hand.3	3.60	4.46	4.23	**3.42**
Date03_Sub03_chairwood_hand.1	3.53	3.64	**3.45**	7.15
Date03_Sub04_stool_sit.1	**4.25**	4.65	4.40	4.76
Date03_Sub04_stool_sit.0	4.20	4.95	4.67	**3.84**
Date03_Sub04_stool_sit.2	4.23	5.00	4.72	**4.20**

Date03_Sub04_stool_sit.3	4.34	5.13	4.89	**4.17**
Date03_Sub03_chairblack_sit.3	**4.06**	5.67	5.26	5.19
Date03_Sub03_chairblack_sit.0	**4.03**	6.23	5.95	4.62
Date03_Sub03_chairblack_sit.2	**4.03**	6.03	5.76	5.14
Date03_Sub03_chairblack_sit.1	**4.03**	6.51	6.24	11.34
Date03_Sub05_yogamat.0	**3.34**	4.83	4.72	3.61
Date03_Sub05_yogamat.1	**3.41**	4.49	4.43	3.59
Date03_Sub05_yogamat.3	**3.29**	4.86	4.69	3.88
Date03_Sub05_yogamat.2	**3.26**	4.52	4.40	3.85
Date03_Sub05_suitcase.0	4.46	4.49	4.33	**3.61**
Date03_Sub05_suitcase.2	4.49	4.77	4.66	**3.91**
Date03_Sub05_suitcase.1	4.51	4.61	4.49	**3.84**
Date03_Sub05_suitcase.3	4.53	4.73	4.62	**3.97**
Date03_Sub04_tablesquare_hand.0	4.05	5.21	5.12	**3.99**
Date03_Sub04_tablesquare_hand.2	**4.03**	4.98	4.78	4.31
Date03_Sub04_tablesquare_hand.1	**4.05**	5.29	5.03	4.96
Date03_Sub04_tablesquare_hand.3	**4.03**	4.92	4.82	4.35
Date03_Sub03_yogaball_play.3	**2.48**	4.98	4.77	3.57
Date03_Sub03_yogaball_play.1	**2.45**	3.99	3.79	5.21
Date03_Sub03_yogaball_play.0	**2.46**	4.24	4.14	3.30
Date03_Sub03_yogaball_play.2	**2.52**	4.67	4.51	12.34
Date03_Sub04_boxsmall.0	**3.14**	4.70	4.62	3.61
Date03_Sub04_boxsmall.3	**3.17**	4.75	4.62	3.74
Date03_Sub04_boxsmall.1	**3.18**	4.38	4.36	4.50
Date03_Sub04_boxsmall.2	**3.11**	4.59	4.50	3.81
Date03_Sub04_tablesmall_lift.0	3.71	5.04	4.89	**3.64**
Date03_Sub04_tablesmall_lift.2	**3.57**	4.62	4.56	3.97
Date03_Sub04_tablesmall_lift.3	**3.66**	4.68	4.55	3.92
Date03_Sub04_tablesmall_lift.1	**3.58**	4.88	4.76	4.13
Date03_Sub04_chairwood_sit.3	**2.76**	4.57	4.33	3.82
Date03_Sub04_chairwood_sit.1	**2.73**	4.46	4.20	4.39
Date03_Sub04_chairwood_sit.2	**2.72**	4.61	4.35	3.85
Date03_Sub04_chairwood_sit.0	**2.74**	4.41	4.13	3.54
Date03_Sub05_boxtiny.1	**2.78**	4.18	4.10	3.38
Date03_Sub05_boxtiny.3	**2.80**	4.63	4.43	3.43
Date03_Sub05_boxtiny.0	**2.78**	4.50	4.34	3.47
Date03_Sub05_boxtiny.2	**2.82**	4.31	4.18	3.62
Date03_Sub05_tablesquare.1	**2.26**	4.79	4.62	4.03
Date03_Sub05_tablesquare.2	**2.24**	4.64	4.44	4.14
Date03_Sub05_tablesquare.3	**2.22**	4.82	4.63	4.01
Date03_Sub05_tablesquare.0	**2.21**	4.93	4.76	3.87
Date03_Sub04_toolbox.0	**3.01**	4.64	4.53	3.68
Date03_Sub04_toolbox.3	**3.04**	4.63	4.50	3.85
Date03_Sub04_toolbox.1	**3.00**	4.51	4.37	4.08
Date03_Sub04_toolbox.2	**3.03**	4.59	4.43	3.86
Date03_Sub05_boxsmall.3	4.35	4.55	4.42	**3.85**
Date03_Sub05_boxsmall.0	4.32	4.38	4.27	**3.52**
Date03_Sub05_boxsmall.1	4.36	4.28	4.17	**3.79**
Date03_Sub05_boxsmall.2	4.34	4.52	4.42	**3.69**
Date03_Sub03_backpack_back.1	**3.16**	3.89	3.88	3.49
Date03_Sub03_backpack_back.0	3.22	4.13	4.02	**3.18**

Date03_Sub03_backpack_back.2	**3.29**	4.43	4.29	3.44
Date03_Sub03_backpack_back.3	**3.22**	4.61	4.42	3.60
Date03_Sub03_keyboard_move.3	**3.05**	3.97	3.70	3.32
Date03_Sub03_keyboard_move.2	**3.18**	4.32	4.05	4.71
Date03_Sub03_keyboard_move.0	**3.03**	4.36	4.17	3.39
Date03_Sub03_keyboard_move.1	**2.99**	4.57	4.41	3.75
Date03_Sub04_monitor_hand.2	4.84	4.87	4.73	**3.85**
Date03_Sub04_monitor_hand.3	4.98	5.52	5.28	**3.97**
Date03_Sub04_monitor_hand.0	4.85	5.45	5.32	**4.02**
Date03_Sub04_monitor_hand.1	6.31	4.73	**4.53**	4.74
Date03_Sub04_tablesquare_lift.0	3.86	4.75	4.63	**3.60**
Date03_Sub04_tablesquare_lift.1	**3.90**	4.25	4.19	4.42
Date03_Sub04_tablesquare_lift.2	3.87	4.52	4.36	**3.64**
Date03_Sub04_tablesquare_lift.3	**3.90**	4.87	4.73	4.22
Date03_Sub04_basketball.0	4.24	4.77	4.70	**3.51**
Date03_Sub04_basketball.2	4.28	4.71	4.65	**3.76**
Date03_Sub04_basketball.3	4.32	4.85	4.70	**3.96**
Date03_Sub04_basketball.1	4.23	4.66	4.56	**4.17**
Date03_Sub03_backpack_hand.3	**3.65**	4.65	4.50	3.74
Date03_Sub03_backpack_hand.1	**3.59**	4.07	4.00	6.78
Date03_Sub03_backpack_hand.0	3.66	4.24	4.12	**3.32**
Date03_Sub03_backpack_hand.2	3.62	4.49	4.31	**3.42**
Date03_Sub04_yogaball_play.3	**2.57**	4.32	4.13	3.28
Date03_Sub04_yogaball_play.1	**2.51**	3.70	3.61	13.57
Date03_Sub04_yogaball_play.2	**2.50**	4.05	3.87	3.18
Date03_Sub04_yogaball_play.0	**2.62**	4.04	3.90	2.99
Date03_Sub03_boxlarge.3	**2.81**	4.40	4.20	3.17
Date03_Sub03_boxlarge.0	**2.79**	3.71	3.59	3.06
Date03_Sub03_boxlarge.1	**2.79**	3.65	3.55	21.16
Date03_Sub03_boxlarge.2	**2.77**	4.30	4.13	25.80
Date03_Sub03_yogamat.0	3.70	4.72	4.62	**3.32**
Date03_Sub03_yogamat.3	3.75	4.49	4.38	**3.73**
Date03_Sub03_yogamat.1	**3.88**	4.62	4.51	4.33
Date03_Sub03_yogamat.2	3.70	4.44	4.31	**3.56**
Date03_Sub04_suitcase_lift.3	**3.28**	4.68	4.58	3.77
Date03_Sub04_suitcase_lift.0	**3.34**	4.62	4.49	3.56
Date03_Sub04_suitcase_lift.1	**3.24**	4.51	4.43	4.51
Date03_Sub04_suitcase_lift.2	**3.26**	4.68	4.55	3.77
Date03_Sub03_chairblack_hand.2	4.63	4.72	4.51	**3.67**
Date03_Sub03_chairblack_hand.1	4.62	4.25	**4.06**	4.36
Date03_Sub03_chairblack_hand.0	4.61	4.21	4.04	**3.41**
Date03_Sub03_chairblack_hand.3	4.68	4.75	4.53	**3.83**
Date03_Sub03_tablesmall_lean.0	4.24	4.93	4.78	**3.98**
Date03_Sub03_tablesmall_lean.3	4.32	4.69	4.62	**4.01**
Date03_Sub03_tablesmall_lean.2	4.28	5.03	4.88	**4.05**
Date03_Sub03_tablesmall_lean.1	**4.28**	5.12	4.96	4.87
Date03_Sub04_suitcase_ground.3	**2.88**	4.75	4.61	3.70
Date03_Sub04_suitcase_ground.0	**2.84**	4.63	4.54	3.60
Date03_Sub04_suitcase_ground.1	**2.85**	4.43	4.32	4.40
Date03_Sub04_suitcase_ground.2	**2.95**	4.39	4.29	3.73
Date03_Sub05_backpack.3	4.24	4.49	4.38	**3.79**

Date03_Sub05_backpack.2	4.20	4.53	4.38	**3.69**
Date03_Sub05_backpack.0	4.19	4.46	4.36	**3.42**
Date03_Sub05_backpack.1	4.20	4.53	4.35	**3.61**
Date03_Sub05_toolbox.0	4.16	4.90	4.78	**3.83**
Date03_Sub05_toolbox.2	**4.04**	4.71	4.60	4.08
Date03_Sub05_toolbox.3	4.11	4.83	4.72	**3.92**
Date03_Sub05_toolbox.1	4.13	4.85	4.66	**3.88**
Date03_Sub03_stool_lift.2	**3.63**	5.04	4.87	14.76
Date03_Sub03_stool_lift.3	**3.65**	4.75	4.67	3.70
Date03_Sub03_stool_lift.1	**3.74**	4.75	4.67	13.89
Date03_Sub03_stool_lift.0	3.63	4.48	4.36	**3.55**
Date03_Sub04_boxlong.2	9.65	5.13	4.92	**4.88**
Date03_Sub04_boxlong.0	9.66	4.99	4.89	**3.88**
Date03_Sub04_boxlong.1	9.64	4.98	**4.83**	6.69
Date03_Sub04_boxlong.3	9.71	4.96	4.85	**4.07**
Date03_Sub03_keyboard_typing.2	**1.66**	4.75	4.47	3.84
Date03_Sub03_keyboard_typing.0	**1.66**	5.02	4.83	3.73
Date03_Sub03_keyboard_typing.3	**1.67**	4.56	4.23	4.11
Date03_Sub03_keyboard_typing.1	**1.68**	5.34	5.06	4.76
Date03_Sub04_monitor_move.2	4.08	5.13	4.92	**3.85**
Date03_Sub04_monitor_move.3	4.12	5.08	4.94	**4.06**
Date03_Sub04_monitor_move.0	4.09	4.94	4.85	**3.87**
Date03_Sub04_monitor_move.1	4.08	4.95	4.81	**3.88**
Date03_Sub05_plasticcontainer.2	4.10	4.66	4.54	**3.67**
Date03_Sub05_plasticcontainer.1	4.12	4.54	4.45	**3.73**
Date03_Sub05_plasticcontainer.3	4.11	4.56	4.36	**3.69**
Date03_Sub05_plasticcontainer.0	4.06	4.41	4.29	**3.47**
Date03_Sub03_plasticcontainer.2	**2.99**	4.58	4.56	3.62
Date03_Sub03_plasticcontainer.3	**3.01**	4.79	4.67	3.86
Date03_Sub03_plasticcontainer.0	**3.02**	4.80	4.70	3.59
Date03_Sub03_plasticcontainer.1	**2.96**	4.72	4.61	3.66
Date03_Sub03_tablesquare_lift.2	**2.59**	4.92	4.83	3.75
Date03_Sub03_tablesquare_lift.0	**2.47**	4.74	4.64	3.58
Date03_Sub03_tablesquare_lift.1	**2.49**	4.96	4.91	4.57
Date03_Sub03_tablesquare_lift.3	**2.48**	4.60	4.39	3.77
Date03_Sub03_chairblack_lift.1	**3.23**	3.41	3.31	3.79
Date03_Sub03_chairblack_lift.2	3.24	4.00	3.84	**3.10**
Date03_Sub03_chairblack_lift.3	**3.26**	4.35	4.14	5.84
Date03_Sub03_chairblack_lift.0	**3.20**	3.76	3.64	8.25
Date03_Sub05_yogaball.0	**3.01**	4.06	3.90	3.21
Date03_Sub05_yogaball.2	**2.93**	4.29	4.08	3.75
Date03_Sub05_yogaball.1	**3.00**	3.81	3.63	3.31
Date03_Sub05_yogaball.3	**2.99**	4.42	4.21	3.53
Date03_Sub04_chairblack_hand.0	**2.95**	4.88	4.73	4.04
Date03_Sub04_chairblack_hand.3	**2.96**	4.64	4.49	3.75
Date03_Sub04_chairblack_hand.2	**2.91**	4.47	4.33	3.96
Date03_Sub04_chairblack_hand.1	**2.94**	4.70	4.51	4.71
Date03_Sub05_boxlong.3	4.01	4.58	4.46	**3.68**
Date03_Sub05_boxlong.0	3.98	4.38	4.30	**3.57**
Date03_Sub05_boxlong.2	4.01	4.60	4.45	**3.63**
Date03_Sub05_boxlong.1	4.07	4.39	4.26	**3.70**

Date03_Sub03_suitcase_lift.0	3.60	4.78	4.71	**3.54**
Date03_Sub03_suitcase_lift.3	**3.66**	4.70	4.58	3.79
Date03_Sub03_suitcase_lift.2	**3.61**	4.44	4.30	13.21
Date03_Sub03_suitcase_lift.1	**3.60**	4.57	4.47	19.57
Date03_Sub05_trashbin.1	**2.74**	3.87	3.82	3.34
Date03_Sub05_trashbin.3	**2.65**	4.47	4.26	3.52
Date03_Sub05_trashbin.0	**2.75**	4.03	3.89	3.38
Date03_Sub05_trashbin.2	**2.80**	4.38	4.25	3.54
Date03_Sub04_plasticcontainer_lift.2	4.44	4.49	4.37	**3.64**
Date03_Sub04_plasticcontainer_lift.1	4.45	4.53	**4.40**	4.66
Date03_Sub04_plasticcontainer_lift.3	4.48	4.62	4.55	**3.73**
Date03_Sub04_plasticcontainer_lift.0	4.45	4.62	4.51	**3.56**
Date03_Sub03_boxsmall.3	**3.13**	4.34	4.19	3.36
Date03_Sub03_boxsmall.0	**3.07**	3.83	3.74	3.15
Date03_Sub03_boxsmall.1	**3.07**	3.94	3.83	10.83
Date03_Sub03_boxsmall.2	**3.10**	4.39	4.30	9.02
Date03_Sub03_backpack_hug.0	3.58	4.37	4.24	**3.28**
Date03_Sub03_backpack_hug.2	3.62	4.67	4.52	**3.50**
Date03_Sub03_backpack_hug.3	3.67	4.78	4.63	**3.52**
Date03_Sub03_backpack_hug.1	**3.62**	4.25	4.17	6.64
Date03_Sub05_boxlarge.0	4.77	4.65	4.50	**3.61**
Date03_Sub05_boxlarge.3	4.82	4.79	4.67	**4.19**
Date03_Sub05_boxlarge.1	4.85	4.57	4.49	**3.87**
Date03_Sub05_boxlarge.2	4.78	4.71	4.60	**3.99**
Date03_Sub04_boxtiny.3	3.80	4.80	4.68	**3.67**
Date03_Sub04_boxtiny.0	**3.61**	4.78	4.69	3.77
Date03_Sub04_boxtiny.1	**3.64**	4.70	4.62	4.50
Date03_Sub04_boxtiny.2	**3.62**	4.70	4.62	3.73
Date03_Sub04_tablesmall_lean.3	**3.75**	4.71	4.53	3.79
Date03_Sub04_tablesmall_lean.0	3.84	4.37	4.25	**3.47**
Date03_Sub04_tablesmall_lean.2	3.71	4.79	4.59	**3.68**
Date03_Sub04_tablesmall_lean.1	**3.72**	4.19	4.02	4.14
Date03_Sub03_tablesmall_lift.3	**3.04**	4.83	4.63	3.74
Date03_Sub03_tablesmall_lift.0	**3.01**	4.43	4.31	18.55
Date03_Sub03_tablesmall_lift.1	**2.98**	4.15	3.98	3.48
Date03_Sub03_tablesmall_lift.2	**3.02**	4.44	4.25	3.67
Date03_Sub03_tablesquare_sit.2	4.71	5.16	4.91	**4.26**
Date03_Sub03_tablesquare_sit.0	4.71	5.08	4.73	**4.21**
Date03_Sub03_tablesquare_sit.1	**4.75**	5.10	4.83	5.20
Date03_Sub03_tablesquare_sit.3	4.75	5.14	4.92	**4.36**
Date03_Sub03_yogaball_sit.3	4.43	5.01	4.85	**3.97**
Date03_Sub03_yogaball_sit.1	4.42	4.53	**4.34**	14.91
Date03_Sub03_yogaball_sit.0	4.38	5.08	4.81	**3.90**
Date03_Sub03_yogaball_sit.2	**4.43**	4.69	**4.43**	20.31
Date03_Sub05_tablesmall.3	**3.99**	4.89	4.75	4.06
Date03_Sub05_tablesmall.0	3.91	4.99	4.85	**3.87**
Date03_Sub05_tablesmall.1	3.95	4.93	4.77	**3.93**
Date03_Sub05_tablesmall.2	**3.89**	4.75	4.64	4.07
Date03_Sub03_tablesquare_move.0	**2.31**	3.72	3.65	3.30
Date03_Sub03_tablesquare_move.1	**2.24**	3.66	3.51	4.32
Date03_Sub03_tablesquare_move.2	**2.32**	4.30	4.14	3.49

Date03_Sub03_tablesquare_move.3	**2.29**	4.30	4.12	3.66
Date03_Sub03_chairwood_sit.3	4.80	5.19	4.93	**4.62**
Date03_Sub03_chairwood_sit.0	4.79	5.68	5.50	**4.50**
Date03_Sub03_chairwood_sit.2	4.79	5.26	5.08	**4.60**
Date03_Sub03_chairwood_sit.1	**4.73**	5.73	5.54	5.02
Date03_Sub03_boxlong.2	**4.02**	4.37	4.23	11.60
Date03_Sub03_boxlong.0	3.98	4.25	4.12	**3.28**
Date03_Sub03_boxlong.1	**3.94**	4.27	4.16	21.25
Date03_Sub03_boxlong.3	4.03	4.33	4.22	**3.28**
Date03_Sub04_backpack_hug.3	**3.56**	5.13	5.00	4.22
Date03_Sub04_backpack_hug.2	**3.42**	4.95	4.81	3.84
Date03_Sub04_backpack_hug.0	**3.62**	4.99	4.88	3.86
Date03_Sub04_backpack_hug.1	**3.40**	4.92	4.76	5.03

Table A.1: Comparison of methods for regression of avatars on BEHAVE [1], using Chamfer distance (cm).

A.2 Comparison of methods for 6DoF object tracking

Video	Metric	DROID-SLAM [32]	BundleTrack [2]	KinectFusion [19]	NICE-SLAM [45]	SDF-2-SDF [27]	Bundle [37]	H20-CA
Date03_Sub03_boxlarge.2	ADD-S (%) ↑	72.59	52.88	21.09	7.05	24.78	92.63	-
	ADD(%) ↑	21.04	13.00	11.00	3.02	7.97	**86.72**	69.50
	CD(cm) ↓	8.61	11.61	8.80	24.79	41.97	1.46	**0.02**
Date03_Sub03_boxlong.3	ADD-S (%) ↑	44.05	27.77	5.59	10.21	54.87	77.0	-
	ADD(%) ↑	14.06	20.31	1.83	1.58	13.75	32.58	**71.64**
	CD(cm) ↓	4.88	1.61	11.55	49.75	26.47	3.05	**0.02**
Date03_Sub03_boxmedium.2	ADD-S (%) ↑	75.98	86.25	11.84	12.60	5.86	92.57	-
	ADD(%) ↑	39.16	50.04	4.26	3.11	3.10	**85.24**	72.78
	CD(cm) ↓	14.49	3.28	3.23	49.73	44.36	1.25	**0.02**
Date03_Sub03_boxsmall.3	ADD-S (%) ↑	8.50	36.32	5.60	4.64	0.84	70.83	-
	ADD(%) ↑	5.40	20.93	4.09	2.84	0.78	51.64	**70.39**
	CD(cm) ↓	3.92	11.73	10.93	36.46	42.29	2.92	**0.02**
Date03_Sub03_boxtiny.3	ADD-S (%) ↑	41.70	52.40	9.94	6.80	19.97	88.01	-
	ADD(%) ↑	22.44	39.34	7.64	4.35	6.49	**74.24**	67.81
	CD(cm) ↓	3.27	13.31	14.64	46.00	26.47	2.08	**0.02**
Date03_Sub03_chairblack_hand.3	ADD-S (%) ↑	67.82	86.19	45.88	26.62	31.81	95.52	-
	ADD(%) ↑	10.89	70.83	6.86	3.26	0.66	**90.28**	82.72
	CD(cm) ↓	15.82	11.35	12.08	12.89	32.25	2.4	**0.01**
Date03_Sub03_chairblack_lift.1	ADD-S (%) ↑	32.98	27.85	21.15	59.55	14.29	28.03	-
	ADD(%) ↑	11.96	15.61	9.90	10.70	4.68	11.43	**79.50**
	CD(cm) ↓	51.95	20.86	8.15	31.12	34.36	6.46	**0.02**
Date03_Sub03_chairblack_sit.3	ADD-S (%) ↑	97.25	98.87	95.32	32.52	43.08	98.81	-
	ADD(%) ↑	94.49	**98.06**	91.39	13.08	16.75	97.95	84.07
	CD(cm) ↓	4.71	4.57	3.48	32.12	26.66	4.63	**0.01**
Date03_Sub03_chairblack_sitstand.3	ADD-S (%) ↑	92.75	98.64	97.26	78.19	34.57	98.56	-
	ADD(%) ↑	86.31	**97.73**	95.24	55.40	14.49	97.64	83.94
	CD(cm) ↓	4.25	2.49	2.1	27.56	37.39	4.75	**0.01**
Date03_Sub03_chairwood_hand.3	ADD-S (%) ↑	72.52	98.24	86.28	39.39	36.43	97.80	-
	ADD(%) ↑	34.60	**96.03**	56.04	9.64	10.83	94.75	82.97
	CD(cm) ↓	10.65	7.97	8.22	30.33	47.06	0.92	**0.01**
Date03_Sub03_chairwood_lift.3	ADD-S (%) ↑	60.58	60.24	7.83	19.90	11.50	81.39	-
	ADD(%) ↑	19.99	35.33	4.61	2.38	2.73	48.19	**78.39**
	CD(cm) ↓	13.52	10.30	8.09	44.33	49.34	6.3	**0.02**
Date03_Sub03_chairwood_sit.2	ADD-S (%) ↑	74.28	99.27	93.88	84.39	9.81	99.31	-
	ADD(%) ↑	52.26	98.92	85.09	68.02	8.78	**99.01**	84.98
	CD(cm) ↓	17.37	4.47	5.12	49.36	40.59	3.87	**0.01**
Date03_Sub03_monitor_move.1	ADD-S (%) ↑	9.14	32.37	15.03	63.27	30.76	51.38	-
	ADD(%) ↑	8.43	13.24	9.25	32.85	7.59	24.54	**80.75**
	CD(cm) ↓	2.83	19.83	2.46	40.97	28.32	3.09	**0.01**
Date03_Sub03_plasticcontainer.2	ADD-S (%) ↑	55.25	61.63	16.48	23.24	4.57	84.65	-
	ADD(%) ↑	13.84	44.28	8.37	4.15	2.14	58.15	**62.28**

	CD(cm) ↓	12.81	8.70	26.06	27.79	41.60	5.62	**0.02**
Date03_Sub03_stool_lift.2	ADD-S (%) ↑	73.65	18.15	19.38	23.13	8.30	94.42	-
	ADD(%) ↑	37.80	15.43	9.64	6.05	2.64	**82.53**	71.18
	CD(cm) ↓	10.56	26.86	5.73	45.05	50.46	1.37	**0.02**
Date03_Sub03_stool_sit.2	ADD-S (%) ↑	85.03	98.68	97.88	26.56	5.44	98.67	-
	ADD(%) ↑	69.71	96.64	91.98	16.93	4.03	**96.65**	84.59
	CD(cm) ↓	5.66	3.13	1.54	36.35	46.52	1.68	**0.01**
Date03_Sub03_suitcase_lift.0	ADD-S (%) ↑	69.41	81.78	16.46	35.64	49.11	90.27	-
	ADD(%) ↑	24.89	52.97	9.47	8.74	14.43	**76.77**	75.01
	CD(cm) ↓	10.22	6.95	14.05	13.97	33.97	2.3	**0.02**
Date03_Sub03_suitcase_move.0	ADD-S (%) ↑	71.95	35.00	22.58	37.79	77.35	94.41	-
	ADD(%) ↑	41.66	17.16	9.59	9.76	41.32	79.25	**81.97**
	CD(cm) ↓	9.34	17.34	3.35	27.54	26.47	1.32	**0.01**
Date03_Sub03_tablesmall_lean.3	ADD-S (%) ↑	52.26	98.72	93.40	50.70	32.52	98.55	-
	ADD(%) ↑	44.13	**96.52**	80.17	18.00	18.45	95.37	84.66
	CD(cm) ↓	6.01	8.13	8.50	37.76	32.43	14.38	**0.01**
Date03_Sub03_tablesmall_lift.2	ADD-S (%) ↑	46.86	48.88	12.70	45.02	15.03	70.67	-
	ADD(%) ↑	23.23	26.54	10.25	15.97	7.92	44.03	**79.38**
	CD(cm) ↓	11.10	44.79	10.56	40.56	46.03	7.03	**0.01**
Date03_Sub03_tablesmall_move.3	ADD-S (%) ↑	48.65	94.12	93.57	37.25	1.66	98.31	-
	ADD(%) ↑	28.78	84.67	75.58	12.00	1.64	**95.16**	84.10
	CD(cm) ↓	11.34	22.70	8.37	29.06	26.47	5.22	**0.01**
Date03_Sub03_tablesquare_lift.1	ADD-S (%) ↑	85.52	96.58	10.33	5.05	3.30	97.02	-
	ADD(%) ↑	50.60	91.95	4.79	1.52	2.25	**92.9**	78.06
	CD(cm) ↓	7.14	2.15	30.80	44.14	36.26	0.68	**0.02**
Date03_Sub03_tablesquare_move.2	ADD-S (%) ↑	97.09	99.36	99.21	15.44	41.38	99.35	-
	ADD(%) ↑	92.17	**98.98**	98.60	10.71	22.26	98.96	83.22
	CD(cm) ↓	4.22	2.86	2.26	43.09	50.02	2.31	**0.01**
Date03_Sub03_tablesquare_sit.3	ADD-S (%) ↑	81.23	99.09	98.97	64.13	57.54	99.1	-
	ADD(%) ↑	78.30	98.65	98.26	33.85	35.25	**98.71**	82.54
	CD(cm) ↓	3.04	1.49	1.13	37.66	36.43	2.22	**0.01**
Date03_Sub03_toolbox.3	ADD-S (%) ↑	0.08	26.69	2.50	5.96	9.01	92.39	-
	ADD(%) ↑	0.08	20.25	1.44	3.53	1.52	68.97	**76.91**
	CD(cm) ↓	1.42	34.63	22.42	44.52	26.47	1.70	**0.01**
Date03_Sub03_trashbin.1	ADD-S (%) ↑	72.44	30.27	52.37	24.45	5.90	91.31	-
	ADD(%) ↑	48.50	21.79	30.18	11.60	2.07	73.23	**74.83**
	CD(cm) ↓	8.67	15.10	14.71	47.01	42.50	4.62	**0.02**
Date03_Sub03_yogamat.2	ADD-S (%) ↑	45.99	17.04	17.27	14.54	69.35	95.8	-
	ADD(%) ↑	21.05	12.27	4.61	3.16	21.24	**73.06**	72.30
	CD(cm) ↓	9.66	15.32	11.58	57.95	26.47	0.92	**0.02**
Date03_Sub04_boxlarge.0	ADD-S (%) ↑	78.77	50.00	11.32	17.14	22.68	90.81	-
	ADD(%) ↑	39.96	44.56	8.91	7.66	6.57	59.99	**70.84**

	CD(cm) ↓	9.15	94.26	4.76	25.77	41.14	2.55	**0.02**
Date03_Sub04_boxlong.2	ADD-S (%) ↑	30.54	24.48	6.40	5.92	7.04	13.53	-
	ADD(%) ↑	8.48	13.05	4.60	2.60	2.49	5.37	**68.88**
	CD(cm) ↓	8.74	76.45	8.43	37.69	26.47	24.72	**0.02**
Date03_Sub04_boxmedium.0	ADD-S (%) ↑	5.05	29.29	5.40	14.67	6.06	92.65	-
	ADD(%) ↑	2.50	8.91	2.99	2.69	2.24	30.34	**76.41**
	CD(cm) ↓	4.12	69.32	5.83	26.99	26.47	1.27	**0.02**
Date03_Sub04_boxsmall.0	ADD-S (%) ↑	0.07	38.07	19.26	18.48	5.40	88.35	-
	ADD(%) ↑	0.07	23.81	11.46	10.55	2.98	64.11	**70.26**
	CD(cm) ↓	3.07	48.46	6.40	22.40	48.37	2.78	**0.02**
Date03_Sub04_boxtiny.0	ADD-S (%) ↑	1.36	12.90	2.92	5.57	11.97	42.99	-
	ADD(%) ↑	0.81	7.40	2.19	1.76	3.44	28.52	**59.40**
	CD(cm) ↓	34.18	68.38	2.07	29.79	26.47	3.54	**0.02**
Date03_Sub04_chairblack_hand.1	ADD-S (%) ↑	74.11	93.52	40.70	45.71	19.26	96.61	-
	ADD(%) ↑	20.40	86.55	15.73	10.10	2.03	**93.0**	82.35
	CD(cm) ↓	8.91	3.79	15.32	28.98	38.09	1.35	**0.01**
Date03_Sub04_chairblack_liftreal.1	ADD-S (%) ↑	47.82	64.32	11.18	6.90	1.37	40.10	-
	ADD(%) ↑	10.85	20.65	4.57	1.66	0.36	10.04	**77.48**
	CD(cm) ↓	81.37	17.57	5.37	25.04	26.47	7.95	**0.02**
Date03_Sub04_chairblack_sit.1	ADD-S (%) ↑	80.91	90.64	73.12	24.95	38.99	97.69	-
	ADD(%) ↑	56.35	83.45	46.21	11.76	23.92	**95.25**	83.14
	CD(cm) ↓	7.04	4.86	9.53	24.96	38.25	3.61	**0.01**
Date03_Sub04_chairwood_hand.0	ADD-S (%) ↑	61.54	68.00	4.54	30.96	37.45	94.38	-
	ADD(%) ↑	17.25	33.62	3.33	6.18	1.80	**86.84**	82.02
	CD(cm) ↓	12.81	11.76	31.75	27.24	26.47	1.32	**0.01**
Date03_Sub04_chairwood_lift.3	ADD-S (%) ↑	64.87	29.10	16.22	32.87	16.12	54.47	-
	ADD(%) ↑	36.92	10.57	7.70	9.45	5.79	12.13	**75.85**
	CD(cm) ↓	12.69	11.21	6.22	42.00	35.90	19.81	**0.02**
Date03_Sub04_chairwood_sit.1	ADD-S (%) ↑	76.25	98.15	71.86	56.97	31.97	98.14	-
	ADD(%) ↑	32.16	**95.67**	45.56	35.31	9.82	94.83	84.73
	CD(cm) ↓	10.16	6.93	13.44	30.31	34.57	1.04	**0.01**
Date03_Sub04_monitor_hand.3	ADD-S (%) ↑	98.21	99.41	98.81	60.24	12.56	99.38	-
	ADD(%) ↑	96.86	**99.24**	95.69	23.50	5.32	99.21	85.32
	CD(cm) ↓	4.13	4.35	3.04	14.61	38.55	3.30	**0.01**
Date03_Sub04_monitor_move.3	ADD-S (%) ↑	6.31	16.72	15.52	4.93	4.07	10.83	-
	ADD(%) ↑	4.62	8.44	6.47	4.10	2.31	5.52	**76.66**
	CD(cm) ↓	7.67	16.76	2.16	34.00	25.43	4.12	**0.02**
Date03_Sub04_plasticcontainer_lift.2	ADD-S (%) ↑	45.35	40.99	12.05	7.59	12.95	73.63	-
	ADD(%) ↑	12.91	23.34	8.37	2.86	6.86	36.16	**65.38**
	CD(cm) ↓	7.08	71.91	6.20	34.26	41.69	5.76	**0.02**
Date03_Sub04_stool_move.0	ADD-S (%) ↑	74.77	46.72	30.19	18.13	76.73	55.24	-
	ADD(%) ↑	48.47	27.65	21.74	7.14	44.05	31.78	**77.61**

	CD(cm) ↓	7.95	25.46	5.33	45.27	26.47	1.25	**0.01**
Date03_Sub04_stool_sit.0	ADD-S (%) ↑	0.51	98.15	97.56	41.58	9.88	98.14	-
	ADD(%) ↑	0.45	**95.57**	83.62	11.90	5.68	95.19	84.58
	CD(cm) ↓	4.30	3.67	2.87	28.76	33.65	2.79	**0.01**
Date03_Sub04_suitcase_ground.0	ADD-S (%) ↑	59.70	96.59	14.83	18.85	6.36	96.93	-
	ADD(%) ↑	20.56	92.75	12.21	8.12	5.23	**93.61**	80.18
	CD(cm) ↓	10.41	1.91	3.18	22.11	37.86	1.17	**0.01**
Date03_Sub04_suitcase_lift.2	ADD-S (%) ↑	34.95	31.68	25.40	29.32	11.21	71.91	-
	ADD(%) ↑	18.14	10.65	11.03	11.75	2.95	64.51	**76.50**
	CD(cm) ↓	5.53	58.81	8.84	49.20	47.01	1.91	**0.02**
Date03_Sub04_tablesmall_hand.0	ADD-S (%) ↑	61.21	29.93	16.53	39.31	21.32	92.94	-
	ADD(%) ↑	37.48	10.46	8.17	8.22	7.03	**85.62**	82.86
	CD(cm) ↓	9.09	9.89	24.59	35.72	42.35	8.45	**0.01**
Date03_Sub04_tablesmall_lean.0	ADD-S (%) ↑	78.16	98.44	96.66	17.29	33.80	98.49	-
	ADD(%) ↑	66.19	95.09	87.52	14.88	18.06	**95.34**	83.75
	CD(cm) ↓	13.51	8.27	7.89	46.25	40.48	9.36	**0.01**
Date03_Sub04_tablesmall_lift.3	ADD-S (%) ↑	43.3	18.38	10.62	18.74	37.41	30.33	-
	ADD(%) ↑	26.87	8.81	6.95	7.78	9.85	11.81	**75.26**
	CD(cm) ↓	6.87	12.59	7.99	19.63	38.33	5.53	**0.02**
Date03_Sub04_tablesquare_hand.0	ADD-S (%) ↑	93.83	98.95	91.30	63.69	33.70	98.82	-
	ADD(%) ↑	46.35	**97.41**	72.31	19.86	24.62	96.69	82.48
	CD(cm) ↓	6.56	3.80	3.81	39.31	38.35	1.63	**0.01**
Date03_Sub04_tablesquare_lift.3	ADD-S (%) ↑	75.82	48.09	12.99	49.94	5.41	96.13	-
	ADD(%) ↑	26.25	16.71	7.92	3.48	3.08	**90.62**	77.11
	CD(cm) ↓	11.98	8.40	10.71	24.59	43.44	0.7	**0.02**
Date03_Sub04_tablesquare_sit.2	ADD-S (%) ↑	93.00	99.18	98.94	63.02	15.54	99.25	-
	ADD(%) ↑	82.80	98.94	97.97	35.62	11.70	**99.07**	83.56
	CD(cm) ↓	4.10	2.27	3.42	40.13	53.83	2.99	**0.01**
Date03_Sub04_toolbox.3	ADD-S (%) ↑	30.35	15.10	7.02	4.66	54.25	80.91	-
	ADD(%) ↑	17.38	9.44	4.37	3.70	29.63	58.0	**64.20**
	CD(cm) ↓	2.47	45.67	13.61	30.08	26.47	3.99	**0.02**
Date03_Sub04_trashbin.0	ADD-S (%) ↑	78.62	66.63	34.18	16.89	4.10	95.62	-
	ADD(%) ↑	54.54	34.15	21.41	8.34	3.14	63.9	**76.73**
	CD(cm) ↓	6.16	5.33	18.05	50.63	47.54	1.05	**0.02**
Date03_Sub04_yogamat.3	ADD-S (%) ↑	25.56	33.14	11.67	15.06	51.85	85.55	-
	ADD(%) ↑	4.67	8.74	6.92	3.53	5.65	58.87	**72.46**
	CD(cm) ↓	16.85	18.22	3.58	42.54	26.47	2.4	**0.02**
Date03_Sub05_boxlarge.1	ADD-S (%) ↑	66.41	42.28	19.60	9.10	34.96	94.47	-
	ADD(%) ↑	15.43	6.25	2.49	1.86	5.68	20.02	**76.51**
	CD(cm) ↓	11.90	15.68	8.48	38.84	32.29	1.13	**0.02**
Date03_Sub05_boxlong.3	ADD-S (%) ↑	3.26	35.26	2.70	5.87	16.01	88.02	-
	ADD(%) ↑	0.56	3.40	1.63	0.96	3.29	59.52	**73.72**

	CD(cm) ↓	10.53	67.09	27.56	38.73	37.49	36.11	**0.02**
Date03_Sub05_boxmedium.2	ADD-S (%) ↑	27.94	20.85	28.36	32.95	7.65	84.52	-
	ADD(%) ↑	18.57	12.96	13.51	15.80	3.60	47.87	**78.04**
	CD(cm) ↓	10.12	12.18	5.82	44.70	43.36	2.51	**0.02**
Date03_Sub05_boxsmall.3	ADD-S (%) ↑	73.21	4.39	5.12	37.15	77.48	93.89	-
	ADD(%) ↑	38.61	3.95	2.58	10.21	37.97	**75.64**	73.38
	CD(cm) ↓	5.23	12.72	2.86	27.07	26.47	2.0	**0.02**
Date03_Sub05_boxtiny.3	ADD-S (%) ↑	23.63	9.58	14.49	5.77	1.27	54.23	-
	ADD(%) ↑	12.80	5.32	5.57	2.81	0.87	40.9	**66.88**
	CD(cm) ↓	37.93	8.70	2.3	49.22	26.47	3.49	**0.02**
Date03_Sub05_chairblack.1	ADD-S (%) ↑	56.78	45.68	32.90	58.88	4.77	69.13	-
	ADD(%) ↑	30.54	39.49	27.99	18.43	2.22	43.12	**81.96**
	CD(cm) ↓	18.52	27.39	25.19	23.76	39.51	8.36	**0.01**
Date03_Sub05_chairwood.1	ADD-S (%) ↑	69.21	92.03	46.51	21.12	13.16	90.43	-
	ADD(%) ↑	30.01	79.75	28.99	11.83	6.28	75.08	**83.86**
	CD(cm) ↓	16.58	5.14	13.52	51.69	43.14	7.78	**0.01**
Date03_Sub05_monitor.1	ADD-S (%) ↑	67.18	64.86	73.46	71.04	15.64	89.05	-
	ADD(%) ↑	55.08	46.39	48.30	52.76	8.71	75.96	**81.83**
	CD(cm) ↓	7.28	52.23	4.19	32.80	31.32	7.37	**0.01**
Date03_Sub05_plasticcontainer.3	ADD-S (%) ↑	51.10	41.66	24.21	23.13	71.54	76.33	-
	ADD(%) ↑	16.60	15.12	9.06	4.40	29.34	48.3	**76.78**
	CD(cm) ↓	23.18	24.71	7.18	32.27	26.47	6.34	**0.02**
Date03_Sub05_stool.2	ADD-S (%) ↑	80.38	96.42	94.69	40.17	33.24	98.27	-
	ADD(%) ↑	60.87	86.80	75.13	27.59	9.88	**94.41**	82.48
	CD(cm) ↓	9.21	6.66	4.55	43.11	45.26	4.13	**0.01**
Date03_Sub05_suitcase.2	ADD-S (%) ↑	71.70	81.13	63.07	25.34	30.69	97.39	-
	ADD(%) ↑	30.48	68.68	26.68	7.04	4.30	**94.31**	81.69
	CD(cm) ↓	6.27	2.31	6.47	29.06	43.58	0.96	**0.01**
Date03_Sub05_tablesmall.1	ADD-S (%) ↑	50.53	56.23	51.31	23.32	59.44	71.39	-
	ADD(%) ↑	34.29	39.52	36.06	9.47	6.41	55.86	**80.36**
	CD(cm) ↓	12.63	27.61	9.87	56.51	32.95	17.35	**0.01**
Date03_Sub05_tablesquare.2	ADD-S (%) ↑	35.60	96.16	66.70	7.55	35.69	97.93	-
	ADD(%) ↑	23.15	87.43	53.86	5.28	25.12	**94.5**	82.23
	CD(cm) ↓	8.73	3.63	29.20	52.47	35.46	1.27	**0.01**
Date03_Sub05_toolbox.1	ADD-S (%) ↑	55.27	24.17	23.30	13.54	52.24	89.64	-
	ADD(%) ↑	36.92	19.30	15.41	6.45	29.98	71.47	**73.39**
	CD(cm) ↓	7.80	17.49	4.94	38.37	26.47	2.94	**0.01**
Date03_Sub05_trashbin.3	ADD-S (%) ↑	78.78	56.89	24.44	32.29	40.09	92.2	-
	ADD(%) ↑	48.88	16.38	14.24	14.19	6.16	56.67	**77.26**
	CD(cm) ↓	7.46	8.89	5.58	36.88	26.47	2.28	**0.01**
Date03_Sub05_yogamat.3	ADD-S (%) ↑	62.56	66.92	8.02	25.46	17.43	96.6	-
	ADD(%) ↑	21.54	8.33	3.84	5.52	1.42	**78.41**	76.89

	CD(cm) ↓	8.92	12.68	2.99	40.50	26.47	1.04	**0.02**
Mean	ADD-S (%) ↑	56.14	59.06	38.37	28.80	25.71	83.63	-
	ADD(%) ↑	32.29	45.03	28.45	11.93	10.05	67.52	**77.34**
	CD(cm) ↓	11.24	19.27	9.36	36.03	35.99	4.66	**0.02**

Table A.2: Per-video comparison on BEHAVE. ADD is AUC (0 to 0.5 m) percentages for pose evaluation. CD is Chamfer distance (cm).

Bibliography

[1] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. "BEHAVE: Dataset and Method for Tracking Human Object Interactions". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2022, pp. 15914–15925. DOI: 10.1109/CVPR52688.2022.01547. URL: https://ieeexplore.ieee.org/document/9879007 (visited on 01/13/2024) (cit. on pp. 11, 44).

[2] *BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models | IEEE Conference Publication | IEEE Xplore*. URL: https://ieeexplore.ieee.org/abstract/document/9635991 (visited on 01/24/2024) (cit. on pp. 27, 45).

[3] X. Chen, H. Peng, D. Wang, H. Lu, and H. Hu. "SeqTrack: Sequence to Sequence Learning for Visual Object Tracking". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2023, pp. 14572–14581. DOI: 10.1109/CVPR52729.2023.01400. URL: https://ieeexplore.ieee.org/document/10203645 (visited on 01/14/2024) (cit. on pp. 9, 10).

[4] Y. Cui, C. Jiang, G. Wu, and L. Wang. "MixFormer: End-to-End Tracking With Iterative Mixed Attention". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–18. DOI: 10.1109/TPAMI.2024.3349519. URL: https://ieeexplore.ieee.org/document/10380715 (visited on 01/14/2024) (cit. on p. 9).

[5] B. Curless and M. Levoy. "A Volumetric Method for Building Complex Models from Range Images". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. URL: https://doi.org/10.1145/3596711.3596726 (cit. on p. 27).

[6] J. Duchi, E. Hazan, and Y. Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12.61 (2011), pp. 2121–2159. URL: http://jmlr.org/papers/v12/duchi11a.html (visited on 01/15/2024) (cit. on p. 23).

[7] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. "Humans in 4D: Reconstructing and Tracking Humans with Transformers". In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023, pp. 14783–14794. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Goel_Humans_in_4D_Reconstructing_and_Tracking_Humans_with_Transformers_ICCV_2023_paper.html (visited on 01/15/2024) (cit. on p. 12).

[8] A. Graves. *Generating Sequences With Recurrent Neural Networks*. version: 5. June 5, 2014. arXiv: 1308.0850[cs]. URL: http://arxiv.org/abs/1308.0850 (visited on 01/15/2024) (cit. on p. 23).

[9] V. Guzov, J. Chibane, R. Marin, Y. He, T. Sattler, and G. Pons-Moll. *Interaction Replica: Tracking human-object interaction and scene changes from human motion*. Mar. 31, 2023. DOI: 10.48550/arXiv.2205.02830. arXiv: 2205.02830[cs]. URL: http://arxiv.org/abs/2205.02830 (visited on 01/12/2024) (cit. on p. 31).

[10] Y. Huang, O. Tehari, M. J. Black, and D. Tzionas. *InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction*. version: 1. Sept. 25, 2022. arXiv: 2209.12354[cs]. URL: http://arxiv.org/abs/2209.12354 (visited on 06/09/2023) (cit. on p. 11).

[11] *jannessm/quadric-mesh-simplification: Fast python implementation of the quadric mesh simplification algorithm from http://mgarland.org/files/papers/quadrics.pdf*. URL: https://github.com/jannessm/quadric-mesh-simplification (visited on 01/15/2024) (cit. on p. 24).

[12] J. Kim, T. Byun, S. Shin, J. Won, and S. Choi. "Conditional motion in-betweening". In: *Pattern Recognition* 132 (2022), p. 108894. DOI: https://doi.org/10.1016/j.patcog.2022.108894. URL: https://www.sciencedirect.com/science/article/pii/S0031320322003752 (cit. on p. 30).

[13] *LBFGS — PyTorch 2.1 documentation*. URL: https://pytorch.org/docs/stable/generated/torch.optim.LBFGS.html#lbfgs (visited on 01/15/2024) (cit. on p. 23).

[14] J. Li and X. Yang. "A Cyclical Learning Rate Method in Deep Learning Training". In: *2020 International Conference on Computer, Information and Telecommunication Systems (CITS)*. 2020 International Conference on Computer, Information and Telecommunication Systems (CITS). Oct. 2020, pp. 1–5. DOI: 10.1109/CITS49457.2020.9232482. URL: https://ieeexplore.ieee.org/document/9232482 (visited on 01/15/2024) (cit. on p. 23).

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16 (cit. on pp. 5, 6, 8, 31).

[16] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. Jan. 4, 2019. DOI: 10.48550/arXiv.1711.05101. arXiv: 1711.05101[cs,math]. URL: http://arxiv.org/abs/1711.05101 (visited on 01/15/2024) (cit. on p. 23).

[17] C. Mayer, M. Danelljan, G. Bhat, M. Paul, D. P. Paudel, F. Yu, and L. Van Gool. "Transforming Model Prediction for Tracking". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2022, pp. 8721–8730. DOI: 10.1109/CVPR52688.2022.00853. URL: https://ieeexplore.ieee.org/document/9879113 (visited on 01/12/2024) (cit. on p. 9).

[18] B. Meynard-Piganeau, C. Fabbri, M. Weigt, A. Pagnani, C. Feinauer, and L. Cowen. "Generating Interacting Protein Sequences using Domain-to-Domain Translation". In: *Bioinformatics* 39 (July 3, 2023). DOI: 10.1093/bioinformatics/btad401 (cit. on p. 9).

[19] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. "KinectFusion: Real-Time Dense Surface Mapping and Tracking". In: Oct. 2011, pp. 127–136. DOI: 10.1109/ISMAR.2011.6092378 (cit. on pp. 27, 45).

[20] Y. Nie, A. Dai, X. Han, and M. Nießner. *Pose2Room: Understanding 3D Scenes from Human Activities*. July 14, 2022. DOI: 10.48550/arXiv.2112.03030. arXiv: 2112.03030[cs]. URL: http://arxiv.org/abs/2112.03030 (visited on 01/13/2024) (cit. on pp. 9, 30).

[21] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10975–10985 (cit. on pp. 5, 31).

[22] I. A. Petrov, R. Marin, J. Chibane, and G. Pons-Moll. "Object pop-up: Can we infer 3D objects and their poses from human interactions alone?" In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2023, pp. 4726–4736. DOI: 10.1109/CVPR52729.2023.00458. URL: https://ieeexplore.ieee.org/document/10204148 (visited on 01/12/2024) (cit. on pp. 1, 6, 27, 31).

[23] *ReduceLROnPlateau — PyTorch 2.1 documentation*. URL: https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html (visited on 01/29/2024) (cit. on p. 23).

[24] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas. "HuMoR: 3D Human Motion Model for Robust Pose Estimation". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, Oct. 2021, pp. 11468–11479. DOI: 10.1109/ICCV48922.2021.01129. URL: https://ieeexplore.ieee.org/document/9711220/ (visited on 10/10/2023) (cit. on p. 11).

[25] *Rodrigues' rotation formula*. In: *Wikipedia*. Page Version ID: 1199303067. Jan. 26, 2024. URL: https://en.wikipedia.org/w/index.php?title=Rodrigues%27_rotation_formula&oldid=1199303067 (visited on 01/29/2024) (cit. on pp. 6, 18).

[26] J. Romero, D. Tzionas, and M. J. Black. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6 (Nov. 2017) (cit. on p. 5).

[27] M. Slavcheva, W. Kehl, N. Navab, and S. Ilic. "SDF-2-SDF Registration for Real-Time 3D Reconstruction from RGB-D Data". English. In: 126.6 (June 2018). Publisher Copyright: © 2017, Springer Science+Business Media, LLC, part of Springer Nature., pp. 615–636. DOI: 10.1007/s11263-017-1057-z (cit. on pp. 27, 45).

[28] *SMPL layer for PyTorch*. URL: https://github.com/gulvarol/smplpytorch?tab=readme-ov-file (cit. on p. 17).

[29] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei. "Monocular, One-stage, Regression of Multiple 3D People". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). ISSN: 2380-7504. Oct. 2021, pp. 11159–11168. DOI: 10.1109/ICCV48922.2021.01099. URL: https://ieeexplore.ieee.org/document/9710639 (visited on 01/15/2024) (cit. on p. 12).

[30] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black. "TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2023, pp. 8856–8866. DOI: 10.1109/CVPR52729.2023.00855. URL: https://ieeexplore.ieee.org/abstract/document/10204771 (visited on 01/15/2024) (cit. on pp. 12, 21).

[31] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black. "Putting People in their Place: Monocular Regression of 3D People in Depth". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2022, pp. 13233–13242. DOI: 10.1109/CVPR52688.2022.01289. URL: https://ieeexplore.ieee.org/document/9879825 (visited on 01/15/2024) (cit. on p. 12).

[32] Z. Teed and J. Deng. "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 16558–16569. URL: https://proceedings.neurips.cc/paper/2021/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html (visited on 01/24/2024) (cit. on pp. 27, 45).

[33] Y. Tian, H. Zhang, Y. Liu, and L. Wang. "Recovering 3D Human Mesh from Monocular Images: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) (cit. on p. 5).

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 01/14/2024) (cit. on pp. 9, 17, 26).

[35] *visionml/pytracking*. original-date: 2019-04-03T10:17:53Z. Jan. 29, 2024. URL: https://github.com/visionml/pytracking (visited on 01/29/2024) (cit. on p. 9).

[36] B. Wen and K. Bekris. "BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2021, pp. 8067–8074. DOI: 10.1109/IROS51168.2021.9635991 (cit. on p. 27).

[37] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. "BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2023, pp. 606–617. DOI: 10.1109/CVPR52729.2023.00066. URL: https://ieeexplore.ieee.org/document/10203995 (visited on 01/13/2024) (cit. on pp. 1, 7, 8, 24, 27, 31, 45).

[38] X. Xie, B. L. Bhatnagar, and G. Pons-Moll. "CHORE: Contact, Human and Object REconstruction from a single RGB image". In: *European Conference on Computer Vision (ECCV)*. Springer. Oct. 2022 (cit. on pp. 27, 28).

[39] X. Xie, B. L. Bhatnagar, and G. Pons-Moll. "Visibility Aware Human-Object Interaction Tracking from Single RGB Camera". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2023, pp. 4757–4768. DOI: 10.1109/CVPR52729.2023.00461. URL: https://ieeexplore.ieee.org/document/10204516 (visited on 01/13/2024) (cit. on pp. 1, 5–8, 27, 28, 30).

[40] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa. "Decoupling Human and Camera Motion from Videos in the Wild". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075. June 2023, pp. 21222–21232. DOI: 10.1109/CVPR52729.2023.02033. URL: https://ieeexplore.ieee.org/document/10204131 (visited on 01/15/2024) (cit. on p. 13).

[41] M. D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. Dec. 22, 2012. DOI: 10.48550/arXiv.1212.5701. arXiv: 1212.5701[cs]. URL: http://arxiv.org/abs/1212.5701 (visited on 01/15/2024) (cit. on p. 23).

[42] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu. "SmoothNet: A Plug-and-Play Network for Refining Human Poses in Videos". In: *Computer Vision – ECCV 2022*. Ed. by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 625–642. DOI: 10.1007/978-3-031-20065-6_36 (cit. on p. 30).

[43] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. "3DMatch: Learning local geometric descriptors from RGB-D reconstructions". English. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Publisher Copyright: © 2017 IEEE. Nov. 2017, pp. 199–208. DOI: 10.1109/CVPR.2017.29 (cit. on p. 27).

[44] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. "On the Continuity of Rotation Representations in Neural Networks". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 5745–5753. URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html (visited on 01/19/2024) (cit. on p. 18).

[45] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys. "NICE-SLAM: Neural Implicit Scalable Encoding for SLAM". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, pp. 12786–12796. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Zhu_NICE-SLAM_Neural_Implicit_Scalable_Encoding_for_SLAM_CVPR_2022_paper.html (visited on 01/24/2024) (cit. on pp. 27, 45).