Universidad
Rey Juan Carlos

DOCTORAL THESIS

Modeling 3D Hand Interactions for Accurate Contact and Manipulation

Author:

Suzanne Sorli

Supervisors:

Dan Casas Guix

Doctoral Program in Information and Communication Technologies

International Doctoral School

2024

Acknowledgements

This thesis marks the end of one of the longest and certainly most demanding adventures of my life. I would like to express my gratitude to all the people who have supported and accompanied me along the way, without whom this work would not have been possible.

First and foremost, I would like to express my deepest gratitude to my thesis supervisor, Dan Casas Guix, for his kind guidance and unwavering support throughout these years. Thank you for trusting me, for always encouraging me and for the enriching opportunities that have marked my career.

I would also like to thank Miguel Ángel Otaduy for welcoming me to the MSLab group as an intern. Without that first opportunity, I would not be here today. Your patience, invaluable advice and support have been crucial to my progress.

A huge thank you to Mickeal Verschoor and Marc Comino Trinidad for their guidance and teaching throughout each new project. Your presence has been indispensable and I am deeply grateful.

To my remarkable collaborators, Jiayi Wang and Ana Tajadura Jiménez, thank you for giving me the opportunity to work with you. These experiences have been extremely formative, and I am truly grateful for your confidence and invaluable insights.

I would also like to express my gratitude to the members of the thesis committee for their time and effort in evaluating my work, and to the reviewers for their valuable feedback on our publications.

Many thanks also to the institutions that financially supported this work, especially the Ministry of Economic Affairs and Digital Transformation (grant RTI2018-098694-B-I00 VizLearning) and the State Agency for Research (TED2021-132003B-I00 BLESIM), which made this work possible.

I would also like to thank all the people who have made my life abroad so memorable and enjoyable. Thank you to Antoine, Cristian, Igor, Rosa, Héctor, Dani, Lillo, Raquel, Melania, Pablo, and the whole MSLab group for your warm welcome and for making this adventure so special. You have been invaluable companions, both daily and during key moments.

A special thank you to Jordan, Leo, Valen, Marylis, Cristina, Alejandra, and Virginia. Every moment spent with you has been a true delight, and your friendship was a tremendous support throughout my time in Spain.

I must also thank those who have inspired and supported me throughout my academic journey. Thank you to my university friends, Yaniss, Clément, and Louis, for your precious friendship, and to my professors, David Vanderhaeghe and Loïc Barthe, who sparked my enthusiasm for computer graphics. A special thank you to

Pierre, for his exceptional support during the two challenging years of my Master's degree.

I am grateful to my siblings, Maud, Kevin, and Alexis, as well as Mathieu, Mathilde, and Aude, for their love and for providing a comforting refuge whenever needed. Thank you to my parents, Pascale and Jean-Marc, for giving me the opportunity to pursue my studies.

To my husband, Diego, thank you for your love, patience, and constant support, even during the most difficult times. Thank you for being my best friend, my partner in crime, and my home. Thank you for always believing in me. I love you more than anything.

I would like to thank my two beloved furry friends too, Navi and Midona, whose presence has made my remote working days so much more enjoyable over the past three years.

Finally, thank you to everyone who, in one way or another, has contributed to the success of this journey.

And lastly, thank you, reader, for taking the time to read my work. I hope you enjoy it.

Abstract

Over the past few years Augmented Reality (AR) and Virtual Reality (VR) has gained popularity and continues to grow in terms of interest and use, particularly focusing on user interaction. These technologies transform how people engage with digital content and immersive environments, generating considerable attention across sectors like entertainment, education, training and healthcare. The need for natural interactions in VR and AR has emerged to enhance immersion, accessibility and realism. The goal is to redefine human-computer interaction by seamlessly blending virtual and physical worlds, offering varied engagement levels from subtle virtual manipulation to full-body interactions within simulated environments.

AR and VR rely on two major technology components typically addressed independently: hand tracking and interactions involving both hand-object and hand-to-hand scenarios. Existing methods often simplify these challenges, limiting their real-world impact. For hand-object interaction, the most general approach involves using physics simulation to enable hands and objects to interact according to the laws of contact mechanics. However, differences in size and skeletal morphology between hand representations in simulators and tracking devices complicate this process. The first contribution of this thesis is a personalized soft-hand model paired with a pose retargeting strategy, formulated as an optimization problem, to connect tracked and simulated hand representations. This method integrates off-the-shelf hand tracking solutions with physics-based hand simulation without requiring a common hand representation, yet allows the hand model parametrization.

Hand-object interaction requires tracking the hand in the real world to map our gestures to a virtual scenario. This hand tracking problem is a growing research field with the potential to provide a natural interface for interacting with virtual environments. Common solutions use computer vision methods, often coupled with learning-based tracking algorithms, which can be depth-based or RGB-based. These methods output the skeletal morphology and configuration of a hand that best matches the user's actual hand, with some also estimating the hand shape. Given the ubiquity of RGB cameras, research has shifted towards RGB-based methods. Despite recent advances, the 3D tracking of two interacting hands from RGB images remains challenging due to issues like inter-hand occlusion, depth ambiguity, handness segmentation and collisions. Additionally, machine learning-based approaches face difficulties in training due to the challenge of obtaining sufficient, high-quality training data, further complicating the development of robust hand tracking systems.

To address these challenges, we propose the first system that simulates physically correct two-hand interactions with personalized hand shape and diverse appearances which generates precise synthetic data. This framework is a major component of a state-of-the-art algorithm for tracking two interacting hands from RGB

images. Furthermore, we tackle depth errors that prevent accurate hand-to-hand contact detection while tracking two-interacting hands by developing an image-based data-driven approach formulated as an image-to-image translation problem. To train our method, we introduce a new pipeline for automatically annotating dense surface contacts in hand interaction sequences. Consequently, our method estimates camera-space contacts during interactions, which can be plugged into any two-hand tracking framework.

Table of contents

List of Figures

List of Tables

1 Introduction

Augmented Reality (AR) and Virtual Reality (VR) are expanding areas of research and technological advancement that garnered increasing interest for decades, where user interaction takes center stage. They reshape how individuals engage with digital content and immersive environments. Their mission is to redefine human-computer interaction by creating immersive and interactive experiences that seamlessly blend the virtual and physical worlds while offering users a spectrum of engagement that ranges from subtle manipulations of virtual elements to full-body interactions within simulated worlds.

On one side, AR harmoniously integrates digital enhancements into the physical world and enables interactive experiences that combine the virtual with the real. Users can manipulate virtual objects, access contextual information or engage in collaborative activities using intuitive gestures, voice commands or touch interactions. AR empowers users to engage with digital content in a manner that is both immersive and contextually relevant. On the other side, VR transports users to immersive virtual environments where interaction is intrinsic to the experience. The users can manipulate virtual objects, explore simulated landscapes or collaborate/communicate with others. VR offers users a strong sense of control and immersion in the virtual realm.

However, despite the considerable advancements in headset technology over the last few years which have significantly enhanced performance and usability, there remains a critical issue - the use of unnatural and non-intuitive input devices such as controllers can prevent a full immersion. Given their dexterity, hands stand out as the most effective versatile interaction tool available to humans, making them ideal candidates to replace conventional input devices. Using our own hands would enable the users to manipulate and interact with virtual objects and environments closely emulating real-world interactions, but also communicate with natural gesture and interact with interfaces using a varied range of movements. This facilitates a seamless and intuitive user experience, eliminating the need for complex input devices or cumbersome controllers while making more accessible and engaging the experience for users of all skill levels; instead, users can rely on their innate motor skills and spatial awareness.

The evolution of AR and VR interaction methods has been propelled by ad-

vances in user interface design, interaction techniques and sensory feedback mechanisms. Researchers persistently explore innovations such as hand tracking, haptic feedback, spatial audio and gaze-based interactions to refine the realism and intuitiveness of user experiences within immersive environments. However, these advancements also present several challenges that will be presented in the following section.

1.1 Challenges

To date, VR has reached a high degree of visual realism, allowing the creation of truly immersive virtual experiences (Kaplan et al. 2021, Krichenbauer et al. 2018, Sun et al. 2018). As virtual objects appear more realistic, the next natural step is to interact with them (Chessa et al. 2019). Humans instinctively use both of their hands for interaction with real and virtual surroundings, and for gesturing and communication. Consequently, many applications require simultaneous hand pose estimation for both hands while they are in close interaction but also when they are interacting with virtual objects (Figure 1.1). However, this seemingly straightforward action entails additional challenges in VR: tracking both hands simultaneously and simulating hand-object interactions, which are typically addressed independently.

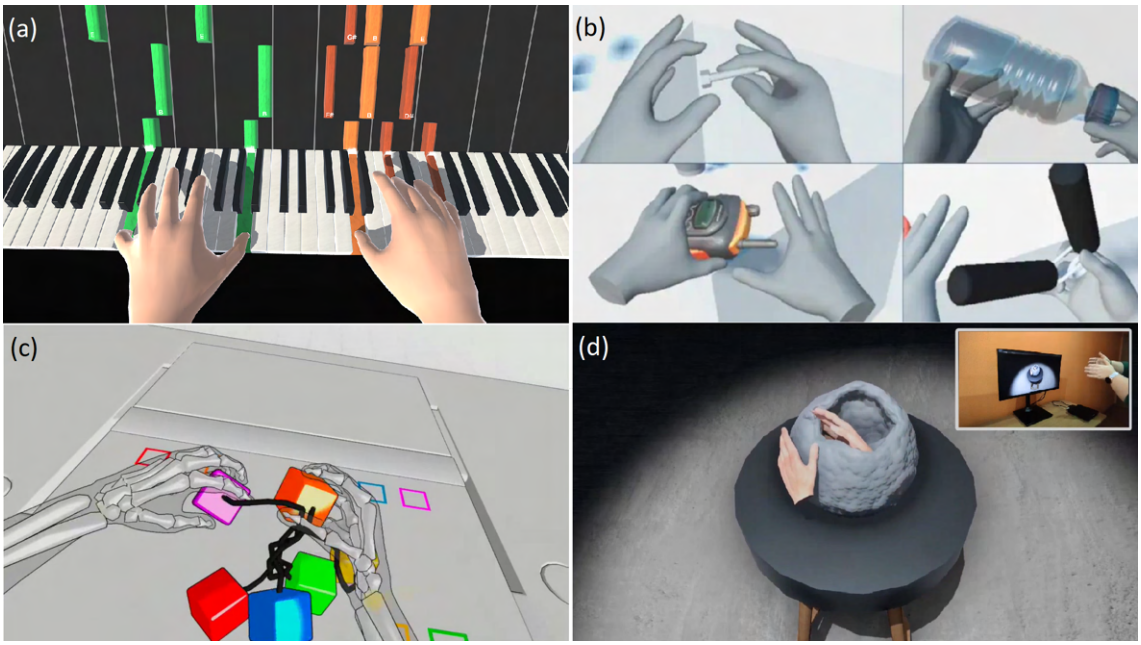


Figure 1.1: *Hand tracking and physics-based simulation enable a variety of practical applications in virtual reality. From playing instruments to manipulating tools and engaging in creative tasks like sculpting, these technologies open new possibilities for immersive learning, training, and exploration.* ***Source:*** *(a) VRtuos (2020), (b) VirtualGrasp (2022), (c) Hand Physics Lab (2021), (d) Barreiro et al. (2021).*

For hand-object interaction, the most general approach is the use of physics simulation to set up a system where hand and objects interact according to the law of contact mechanics. Several modern solutions facilitate interaction with virtual

objects by integrating rudimentary physics-based model on top of the tracked hand structure. However these models only support basic actions such as snapping and grabbing, in which the pose of the virtual hand is kept fixed once a grasp is detected. In parallel, a dedicated branch of research focuses on physics-based hand simulation, aiming to compute hand configurations that satisfies force equilibrium. This involves accounting for various forces, primarily driven by contacts and joint constraints, with the potential inclusion of soft-tissue deformation. The different approaches consider articulated hand representations (Borst & Indugula 2005, Ott et al. 2010), geometric flesh skinning (Duriez et al. 2008), local skin deformation at fingers (Jacobs & Froehlich 2011, Talvas et al. 2015), or full flesh deformation (Garre et al. 2011, Hirota & Tagawa 2016). The method of Verschoor et al. (2018) formulates the problem as an optimization. Yet, a critical observation arises: current approaches that incorporate a physics-based simulation step to model hand-object interaction (Hirota & Tagawa 2016, Kim & Park 2015, Verschoor et al. 2018) are missing an essential component that hinders their deployment in everyday VR applications, hand shape personalization.

Existing techniques use a fixed-size hand template model that does not adjust to the user's hand size, thereby hindering the effectiveness of state-of-the-art hand tracking solutions. Such discrepancy between real-world and virtual dimensions results in simulated hand deformations that significantly differ from those of the actual hand. Differences in hand size and skeletal morphology, particularly in the palm position, mean that the hand pose computed by hand tracking cannot be directly applied to hand-object simulations. If applied naively, it leads to inaccurate finger configurations, which complicate dexterous manipulation of virtual objects. Some finger configurations are even impossible to achieve when the pose of the tracked hand is applied directly to the simulated hand. This limitation is specially relevant in scenarios where the user wants to interact with virtual objects, resulting in unrealistic interactions and limiting natural engagement. Common failures include difficulties in pinching or squeezing virtual objects, where the desired virtual hand pose is difficult to achieve due to the shape mismatch. For instance, a real-world pinch might produce a virtual hand pose where the fingertips are far apart or interpenetrate in an unrealistic way. In Chapter 3 we propose a method to retarget hand poses between hands with different size and skeletal morphology by connecting any hand tracking solution with physics-based hand simulation.

Hand-object interaction requires tracking the hand in the real world to map our gestures to a virtual scenario. Such hand tracking problem is an expanding research field that can potentially provide a natural interface to interact with virtual environments. Typically, computer vision methods are employed to obtain as output the skeletal morphology as well as its pose, and more recently the configuration of a hand that best matches the user's actual hand (Mueller et al. 2019). Some modern commercial hand-tracking solutions, such as Oculus Quest (Facebook 2019) or HTC Vive (HTC 2016), circumvent these challenges by severely simplifying user interaction by means of marker-based hand controllers. This enables robust input

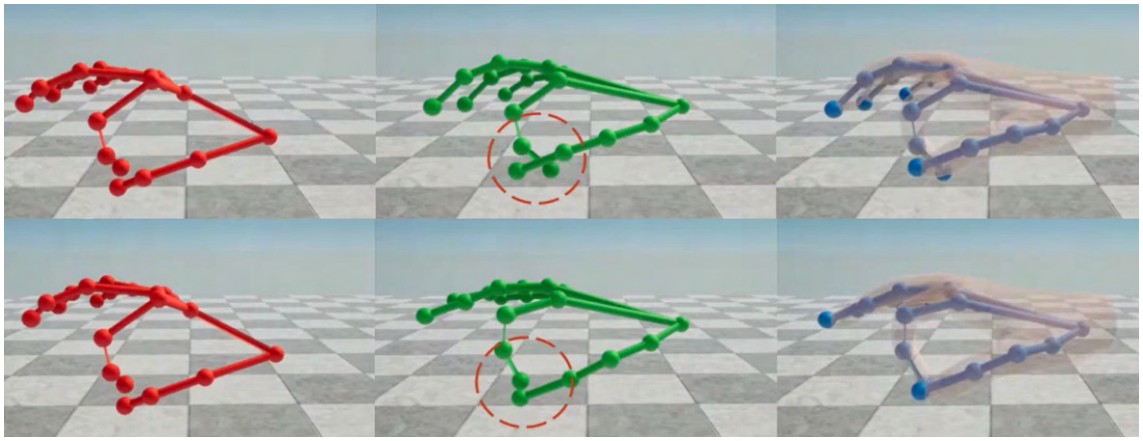
albeit at the price of reduced degrees of freedom; thereby, it reduces the realism of the immersive experience.

In contrast, recent hand tracking methods are able to estimate both hand shape and pose (Romero et al. 2017, Tkach et al. 2017), and some even enforce physically plausible hand poses which potentially enables a much richer input to the virtual world, directly from the user's bare hands: Tzionas et al. (2016) combined a generative model with physics-based simulation to track hands interacting with objects, however restrict the object to a set of predefined model class. Hasson et al. (2019) improved the quality of joint hand and object tracking by incorporating loss functions that favor plausible hand-object interactions, but the method reconstructs only the hand pose. Depending on the application, certain properties are essential for hand tracking methods, such as markerless capture, real-time performance, the ability to track two interacting hands or hands interacting with objects, automatic adaptation to the users' hand shape, or the use of a single RGB camera. However, achieving all these properties is challenging due to issues like frequent occlusion, depth-scale ambiguity, and the self-similarity of hand parts. To overcome these challenges, many methods have been proposed, often simplifying the problem by focusing on single-hand pose estimation, which has been successfully addressed using multi-camera setups (Ballan et al. 2012, Sridhar et al. 2013) or depth cameras (Malik et al. 2020, Sridhar et al. 2015). Rhodin et al. (2016) analyze the effects of camera positioning and the complexity of the acquisition system. Recently, the research focus has shifted towards methods that use a single RGB camera since these sensors are ubiquitous and have also shown very accurate and robust results at tracking single hands (Cai et al. 2018, Mueller et al. 2018, Zimmermann & Brox 2017, Zimmermann et al. 2019). Nevertheless, we argue that using hands as a natural and immersive HCI interface requires methods capable of tracking the *two hands in interaction* from a single RGB camera. Unfortunately, very few works exist that tackle the two-hand tracking scenario. Hybrid and discriminative approaches have shown great success in other scenarios, but they are heavily dependant on the quality and diversity of training data. This reliance presents a considerable challenge, as acquiring comprehensive training data is far from straightforward. On one hand, tasks such as segmentation and depth estimation can be extremely difficult to label manually, on the other hand, synthetic data is hard to obtain because its generation requires accurately simulating real-world conditions and interactions. In this thesis, we propose a new system simulating physically correct two-hand interactions with personalized hand shape and diverse appearances. This contribution led to the development of the first method to reconstruct two interacting hands from only monocular RGB video by using a hybrid method, detailed in Chapter 4. Other research efforts have been conducted to address this issue. Moon et al. (2020) used a discriminative method, trained on the first real dataset with large-scale 3D annotations. Li et al. (2022) approached the problem by formulating the output space as mesh vertices while Zhang et al. (2021) as a parametric model. Despite the promising outcomes, all these approaches share a common obstacle: residual errors in depth, shape, or hand pose estimation, which prevent accurate detection of hand-

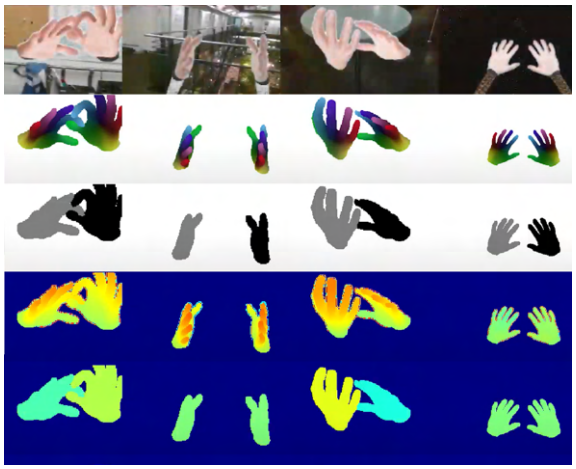
to-hand *contacts* and remains an open problem. In order to address this challenge, we present in Chapter 5, a method to explicitly learn to detect dense hand contact from RGB images of two interacting hands.

The focus of my thesis is to address two main challenges in hand tracking and hand-object interaction: resolving the discrepancies between real-world and virtual dimensions that lead to unrealistic hand deformations and accurately tracking both hands simultaneously with its resulting contacts. The contributions of my thesis are detailed in the following section, along with the resulting publications.

1.2 Summary of Contributions



(a) Chapter 3: Fine virtual manipulation with hands of different sizes.



(b) Chapter 4: Generating Annotated Data of Physically Accurate Two-Hand Interactions.

(c) Chapter 5: Hand-to-hand Contact from RGB images.

Figure 1.2: *Overview of the different methods implemented in this thesis to address the challenges mentioned in the previous section.*

The main contributions of this thesis can be summarized as follows:

- A pose retargeting strategy to connect the tracked hand and the simulated hand. Our approach works with any type of tracking or simulation method, as it stands at the interface between both tasks. We use an intermediate hand representation that shares the size and morphology of the simulated hand,

but which tries to match the configuration of the tracked hand. The retargeting strategy formulates an optimization of the pose of this intermediate hand, based on features that represent the pose of the tracked hand. (Chapter 3)

- An evaluation of the practical impact of the hand mismatch on the manipulation of virtual objects, comparing our pose retargeting strategy vs. naïve copy of the hand pose. To this end, we have carried out a user study which parallels task performance of virtual object manipulation. The study confirms that the mismatch of the hand representation is not critical for gross manipulation (i.e., large objects), but critical for fine manipulation (i.e., small objects). (Chapter 3)

- A new physically-correct synthetic data generation framework, which is able to account for interacting hands with varying hand identities, both in terms of shape and appearance. This work led to the development of the first monocular-RGB-based method for 3D motion capture of two strongly interacting hands, which simultaneously estimates hand pose and shape, while running in real time. (Chapter 4)

- To the best of our knowledge, the first image-based method to estimate hand-to-hand contacts from a single RGB image. Our method builds on top of existing two-hand tracking solutions, enriching them with a camera-space probability map of hand contact that provides many advantages: i) enables the detection of contact even when 3D tracking is inaccurate, ii) makes our solution compatible with any existing two-hand tracking framework (either depth-based and RGB-based methods), iii) can potentially be used as a new term in optimization-based methods for two-hand tracking. (Chapter 5)

- A new pipeline to automatically detect and annotate dense per-vertex surface contacts in real-world hand interaction sequences. (Chapter 5)

1.3 Publications

The following chapters are based on these publications:

- Sorli et al. (2021) **Fine Virtual Manipulation with Hands of Different Sizes.** - Suzanne Sorli, Dan Casas, Mickeal Verschoor, Ana Tajadura-Jiménez, Miguel A. Otaduy - In: Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR), 2021, CORE: A*.

- Wang et al. (2020) **RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video.** - Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, Christian Theobalt - In: ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia), 2020, JCR Q1.

  The main author of this work is Jiayi Wang, who designed the generative model fitting formulation as well as the multi-task Convolutional Neural Network,

while I was the main contributor to the physically-correct synthetic data generation framework used to train the machine learning predictor. The work is included in the thesis for completeness of the methods for tracking two interacting hands.

- **Hand-to-hand Contact from RGB Images.** - Suzanne Sorli, Marc Comino-Trinidad, Dan Casas - In: Under review, 2024.

1.4 Outline

This thesis is organized as follows:

- **Chapter 1 - Introduction.** This chapter presents and motivates the research in hand tracking and hand-object simulation and provides an overview of the challenges addressed in this thesis along with its contributions. It also details its outline and highlights the resulting publications.
- **Chapter 2 - Background.**
  This chapter summarizes the previous works that have inspired it, relating their advantages and weaknesses. We have divided this chapter in four main sections:
  - From Real to Virtual Hand that reviews the fundamental knowledge in hand modeling from its 3D representation to its articulation, required to understand this thesis,
  - Hand Simulation, that reviews the relevant works in physics-based hand simulation and presents the simulation model used in the methods described in Chapters 3, 4 and 5.
  - 3D Hand Tracking that describes the literature of vision-based and markerless 3D hand tracking methods, reviewing the different approaches to achieve it and the different scenarios possible.
  - Embodiment, discusses research on the perception of embodiment in VR users, highlighting how users can accept significant differences between their real and virtual hands.
- **Chapter 3 - Fine Virtual Manipulation with Hands of Different Sizes.**
  In this chapter, we introduce our pose retargeting strategy to connect the tracked hand and the simulated hand. In the first place, we describe the formulation and solution to the optimization problem, as well as a brief summary of the hand-object simulation using the CLAP library (Verschoor et al. 2018). In the second place, we compare task performance of virtual object manipulation using our pose retargeting strategy vs. naïve copy of the hand pose through a user study.
- **Chapter 4 - RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video.**
  This chapter describes a new framework for generating photorealistic and

physically accurate synthethic annotated data depicting sequences of interacting hand motions, which is a major component for a state-of-the-art method for tracking two hand in interactions. Finally, we provide the evaluation of an ablation experiment to demonstrate the effectiveness of our framework.

- **Chapter 5 - Hand-to-hand Contact from RGB Images.**
  This chapter presents our image-based data-driven method to estimate the contact in hand-to-hand interactions. Firstly, we formulate our method as an image-to-image translation problem. Then, we propose a new pipeline to automatically annotate dense surface contacts in hand interaction sequences. Finally, through an exhaustive evaluation, we qualitatively and quantitatively compare our method with the state-of-the-art and a straightforward RGB-only baseline.
- **Chapter 6 - Conclusions.**
  This chapter gives a final overview of the proposed methods and discusses the limitations as well as the different opportunities for future work

2 Background

2.1 From Real to Virtual Hand

Our hands are our main way of interacting with our surroundings. Palms and fingers enable us to perform more or less complex tasks that may require dexterity and/or strength, such as grabbing, pushing, pinching, lifting, or moving them in order to communicate. All these possibilities of activity justify the many research efforts to synthesize the most natural hands possible for integration into ever more immersive and realistic virtual worlds. The many methods used to model, animate and reconstruct hands in 3D are based on a simplified model of the biological structure of the hand, which we will introduce for greater clarity.

2.1.1 Biological Structure of the Hand

The real human hand is a complex structure made up partly of bone and cartilage held together by ligaments, and partly of muscles and tendons, all of which interact closely to orchestrate the complex movements of the fingers and palm together.

The skeleton of a hand comprises 27 bones divided into three groups: the carpus, the metacarpus and the finger skeleton, shown in Figure 2.1.

- The carpus, or wrist, is made up of eight short bones all lying in the same frontal plane,
- The metacarpus comprises five long bones in the palm of the hand, known as the metacarpals, which lie in a frontal plane and are generally oriented in the axis of the limb,
- The finger skeleton comprises fourteen long bones, called phalanges. The five fingers are numbered from most lateral (outer) to most medial (inner), and are called, in order, the thumb, index finger, middle finger, ring finger and pinky finger. Each finger has three phalanges, proximal, intermediate and distal, with the exception of the thumb, which has only two, proximal and distal. The phalanges are located in the axis of each finger.

These bones are linked by articulations (joints) that enable the hand's functionality and mobility. As each finger has three phalanges, it can articulate around the three corresponding joints (a joint shared with the preceding phalanx): around the

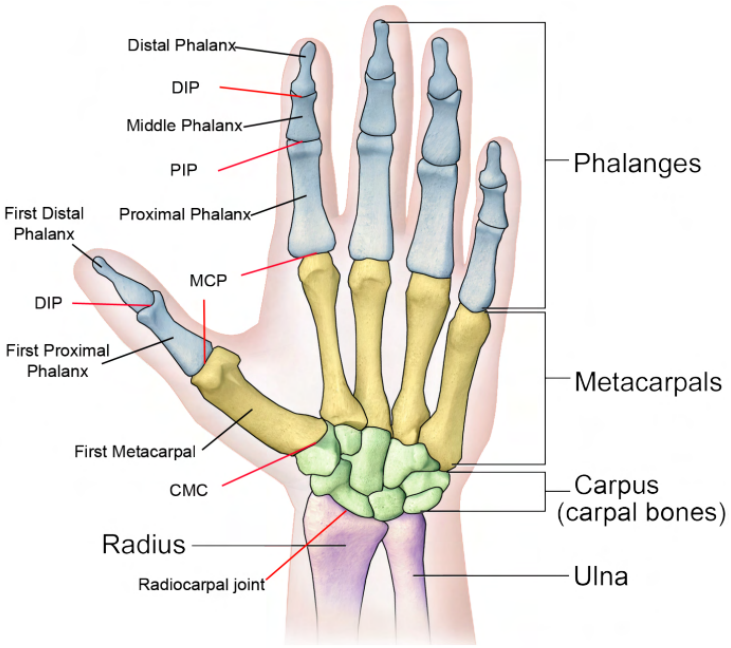


Figure 2.1: *Anatomical system of the hand showing the joints (red lines) and the bones (black lines). The finger skeleton is shown in blue, the metacarpal bones in yellow and the carpal bones in green. (Image from Wheatland et al. (2015)).*

metacarpophalangeal joints (MCP), which are located between the metacarpals and the proximal phalanges; around the *proximal interphalangeal joints* (PIP), around the proximal and middle phalanges; and finally around the *distal interphalangeal joints* (DIP), between the middle and distal phalanges. As the thumb does not have middle phalanx, it can only articulate around its jointed MCP and DIP. Movements are rotational and are as follows: *abduction*, where the fingers move away from each other, away from the middle of the hand; *adduction*, the opposite of abduction, where the fingers move towards each other, towards the middle of the hand; *flexion*, where the fingers bend towards the palm of the hand; *extension*, the opposite of flexion, where the fingers straighten towards the back of the hand. Figure 2.2 depicts these main movements. However, only MCP joints can perform these four types of movement, as DIP and PIP joints behave like hinges, and can only flex and/or extend. Finally, the wrist has a wide range of motion, being capable of flexion/extension, abduction/adduction and twisting.

Hands are among the most beautiful and complex pieces of engineering in the human body. It is evident that building a complete model would be a very difficult task, but also too computationally demanding. A compromise is therefore made between the features that are desirable and the simplicity/practicality of the 3D representation of the model.

2.1.2 3D Hand Representation

Commonly, geometric hand models have a skeleton composed of 20 bones and 16 joints. Each finger consists of 3 joints (DIP, PIP and MCP, with the exception of

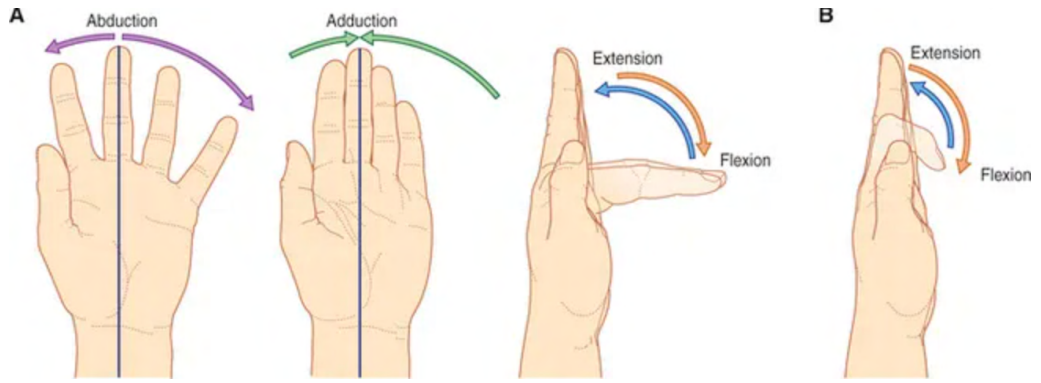


Figure 2.2: *The main types of movement possible for the hand when viewed from the side (**B**) and from the front (**A**): flexion (orange arrow), extension (blue arrow), abduction (purple arrow) and adduction (green arrow).(Image from Richard L. Drake (2023)).*

the thumb, which consists of CMC, MCP and DIP). The finger joints closest to the palm are connected to the Carpus (wrist), simplifying the skeleton considerably, with the last joint being the wrist. However, there is no consensus on the exact position of each joint, especially as not all hands and hand models have the same size and shape. Their Degrees of Freedom (DOF), which refer to the distinct ways in which each joint can move (such as flexing and extending), are depicted in Figure 2.3. They are explained by the natural movements of the different type of joints presented in the human hand previously. The skeleton resulting from this modeling (represented by the gray cones) is then used to animate the hand and is defined as a hierarchic kinematic chain of joints, a method originally proposed by Magnenat et al. (1988).

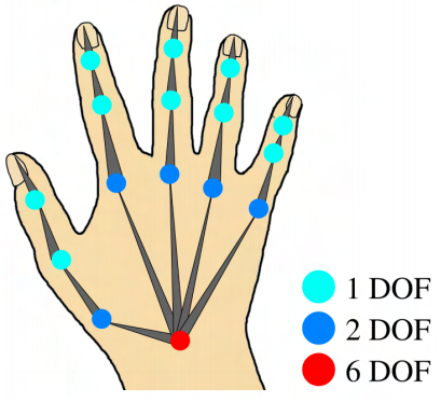


Figure 2.3: *Kinematic skeleton of a hand with its DOF. (Image from Mueller et al. (2017)).*

In other words, a kinematic skeleton is represented on the one hand as a tree of n joint nodes (nodes representing joints), with a root being by consensus the first node, defining the global transformation of the root joint (often chosen to be the wrist when dealing with the hand and spine for a human body); and secondly, by a set of rigid transforms $T \in SE(3)$ which gives the local frame of each joint noted $\{T_i\}_{i=0...n-1}$. Note that if you have a chain of connected bones, only the root can be moved, the rest can only be rotated. In fact, translating a child would mean

changing the scale of the bone, and this would not make the animation realistic (unless it is a special effect). In this model, an edge exists between two nodes if the corresponding joints are connected by a bone. Each ith node within the tree stores the transformation T_i which defines its local space relative to its parent. By convention, the joints are positioned at the origin of their own local coordinate. Therefore, we can retrieve the location of the ith joint in its parent's local space J_i^{parent} as

$$J_i^{parent} = T_i \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{2.1}$$

We can thus determine the global coordinate space of a node, that is to say, a coordinate system relative to the world, by multiplying the local transformations down the tree.

$$J_i^{global} = \left(\prod_{k \in \alpha_i} T_k^{local}\right) \cdot T_i^{local} \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{2.2}$$

formulating mathematically the fact that when a bone is moved, all the subsequent bones in the chain defined by α_n follow the movement.

2.1.3 Hand Personalization

Formulating a model of capable of reproducing hand shape variability and surface deformations of articulated hands with high detail would be a key to address the problem of inaccuracy due to the difference between the hand model and user's hands and lack of realism.

There are methods that can be used to build detailed surface parametric model such as Potamias et al. (2023) or Romero et al. (2017). Both of them built a model of shape variation from extensive high-resolution scans and deformations are then represented as a linear combination of blend shape basis. However, in the MANO model (Romero et al. 2017), they don't restrict the pose-dependent deformations to be modeled by linear blend skinning only but by learning pose-dependent corrective blend shapes instead from multiple poses, yielding more realistic posed meshes. They have used 1,000 scans of 30 different subjects in a variety of poses. Tkach et al. (2017) map an approximate parametric model of the hand to personalized video sequences. In our work, we leverage the MANO model of Romero et al. (2017).

A 3D mesh representation consists of a collection, commonly stored as a graph, of polygons defined by vertices, edges, and faces that approximates the shape of an object in 3D. The MANO model, which is made up of 778 vertices, 1538 faces and 16 joints, first has a basic mesh, a template, named $\bar{T}$, which is enriched with a

Figure 2.4: *Several instances of the MANO hand model with parameterized shapes in the mean pose.*

shape blendshape function to vary its shape/appearance as depicted in Figure 2.4. The following equation yields new hand instances with a parametrized shape:

$$T_S(\vec{\beta}) = \bar{T} + \sum_{n=1}^{n_c} \beta_n S_n. \tag{2.3}$$

With $\vec{\beta} = [\beta_0, \beta_1, .., \beta_{n_c}] \in \mathbb{R}^{n_c}$ the shape parameters as linear coefficients, and the vectors S_n, the first n_c principal components in a low-dimensional shape basis that they learnt from the scans. They also introduced a pose blend shape function that produces blend shapes that correct well-known artifacts from traditional LBS models resulting in more natural-looking finger bending, used to compute the final template which is defined as follows:

$$T_P(\vec{\beta}, \vec{\theta}) = T_S(\vec{\beta}) + \sum_{n=1}^{9K} (R_n(\vec{\theta}) - R_n(\vec{\theta*}))P_n \tag{2.4}$$

In this equation, $\vec{\theta} \in \mathbb{R}^{45}$ encodes the pose parameters in the form of a rotation of the joints, K counts the parts in the hand model and P_n are the pose blend shapes. $(R_n(\vec{\theta}) - R_n(\vec{\theta*}))$ will compute pose-dependent weights, with R_n indexing the n^{th} element from $\vec{\theta}$ and last, $\vec{\theta*}$ is the zero pose (see Romero et al. (2017) for further details). Finally, the general formula to produce the articulated mesh is then defined as:

$$M(\vec{\beta}, \vec{\theta}) = LBS(T_P(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W}) \tag{2.5}$$

with LBS, the traditional skinning function, applied to the final template T_P obtained in Equation 2.4. $J(\vec{\beta})$ is the function that computes 3D joint locations as a function of shape parameters in rest pose (on the template $\bar{T}$), and $\mathcal{W}$ is the LBS skinning/rigging weight matrix.

The computation of skeletal configurations of hand models is at the core of both hand tracking and hand simulation. These two lines of research differ in the input data and the formulation of the computational problem, but both solve the pose of the hand (i.e., joint or bone transformations).

2.2 Hand Simulation

2.2.1 Hand Simulation in VR

The natural integration of hands into VR applications has been tackled from two major perspectives. One is to track the user's hands and detect grasping of virtual objects, the other is to provide (force) feedback of the interaction with the virtual objects. Most hand models focus on the skeletal structure of the hand, and early models even approximated the hand as an articulated structure of rigid bodies (Borst & Indugula 2005, Jacobs et al. 2012, Ott et al. 2010).

To increase the realism of interactive hand simulation, the articulated skeleton of the hand can be used for driving the deformation of a skin surface, using geometric skinning techniques (Kry et al. 2002, Kurihara & Miyata 2004). Data-driven hand animation can also incorporate information about contact interactions with various objects (Kry & Pai 2006, Li et al. 2007, Pollard & Zordan 2005), but the lack of physics-based response complicates natural interaction with virtual objects.

The realism of hand simulation can be further increased by including a model of skin deformation that reacts based on the contact configuration with virtual objects. This skin deformation not only provides indentation and bulging effects akin to those in the real world, but it also provides a key property for smooth and stable grasping of objects. Skin is very soft and compliant under gentle forces, hence it quickly creates a large contact area when we touch and grasp objects, and this contact area provides a large and stable surface for frictional torque. It is, for example, the reason why we can grasp objects using just two fingers. To limit the computational complexity of simulating a full deformable hand, some works have explored the use of deformable finger tips (Barbagli et al. 2004, Ciocarlie et al. 2007, Frisoli et al. 2006). Some other methods connect a portion of the deformable skin to the skeleton (Duriez et al. 2008, Jacobs & Froehlich 2011, Rivers & James 2007). Some works with more detailed approaches also simulate parts of the deformation of the skin like Pouliquen et al. (2005). Considering deformable skin also complicates contact handling, as multiple contact points must be considered per finger tip. Talvas et al. (2015) simplified this problem by aggregating all the contacts happening on each finger tip. Kim & Park (2015) deformed the skin using a global skeletal blending method, and then added a local deformation model based on contact interactions. Holl et al. (2018) simulated a local frictional contact model to produce highly natural grasping poses.

When a real hand interacts with objects, the contact forces affect the deformation of the skin but also, indirectly, the motion of the skeleton. Capturing this effect correctly requires two-way coupling of skin deformation and skeletal dynamics. However, this coupled simulation increases considerably the computational complexity. Hirota & Tagawa (2016), for example, simulated the full deformation of the skin connected to the articulated skeleton, but applied the configuration of the tracked hand directly to the skeleton, hence eliminating the two-way coupling effect. This results in large stress and deformations in the skin, which appear unnatu-

ral or even turn unstable. Garre et al. (2011) approximated the full two-way-coupled dynamics solve in two steps. Verschoor et al. (2018) simulated the full coupled problem efficiently thanks to an energy minimization formulation.

Some works consider more detailed hand simulation, including for example muscle and tendon deformation (Sueda et al. 2008), or the high nonlinearity of skin (Perez et al. 2013). These complex models require specialized solvers for fast simulation, but it is difficult to guarantee real-time performance at all times (Perez et al. 2016).

2.2.2 Soft-Hand Simulation Model

In the following chapters, we use the simulation model described in Verschoor et al. (2018) to which contacts with deformable objects, including a second hand, have been added. This model simulates the behavior of the soft-hand featuring frictional contact with deformable objects. To achieve interactive frame rates, it avoids the use of constrained integration and formulate coupling and contact constraints as penalty forces (soft constraints). However, soft constraints might lead to unstable simulations due to the high force stiffness required to keep constraints satisfied. To guarantee its stability, the simulation method formulates the full dynamics as a unified energy minimization problem. This section introduces the dynamics formulation based on energy minimization and details the mechanics of various simulation elements in terms of energy components: the deformable objects, the soft articulated hand, and frictional contact.

Unified Energy Formulation

Verschoor et al. (2018) formulate the full dynamics problem as a unified energy minimization problem depending on a vector of generalized coordinates $\boldsymbol{q} \in \mathbb{R}^N$. The generalized coordinates comprise the nodes of the soft-tissue, $\boldsymbol{x}_\mathrm{H} \in \mathbb{R}^{3N_\mathrm{H}}$, the position and angular coordinates defining skeleton kinematics, $\{\boldsymbol{c}_\mathrm{S}, \boldsymbol{\phi}_\mathrm{S}\} \in \mathbb{R}^{N_\mathrm{S}}$, as well as an arbitrary number of deformable objects, $\boldsymbol{x}^i_\mathrm{O}$ to interact with in the virtual environment.

In the variational formulation, see *e.g.* Gast et al. (2015), the dynamics problem results from discretizing the Euler-Lagrange equations for some potential U and kinetic K energies. In the generalized coordinates system, the kinetic energy can be computed as $K = 1/2\dot{\boldsymbol{q}}^T \boldsymbol{M} \dot{\boldsymbol{q}}$. Here, $\boldsymbol{M}$ is the mass matrix of the system which is built assembling the mass matrices and inertia tensors of all deformable and rigid-body objects defined above. Verschoor et al. (2018) discretize Euler-Lagrange equations in time using forward difference approximations: $\ddot{\boldsymbol{q}} = (\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}_0) \cdot \Delta t^{-1}$ and $\dot{\boldsymbol{q}} = (\boldsymbol{q} - \boldsymbol{q}_0) \cdot \Delta t^{-1}$, for some time step Δt and initial positions and velocities $\boldsymbol{q}_0$ and $\dot{\boldsymbol{q}}_0$, respectively. Assuming the mass matrix of the system $\boldsymbol{M}$ is constant within each time step, we finally obtain the following backward-Euler discretization of the equations of motion:

$$\boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{M}\left(\frac{\boldsymbol{q} - \boldsymbol{q}_0 - \Delta t \dot{\boldsymbol{q}}}{\Delta t^2}\right) + \frac{\partial U}{\partial \boldsymbol{q}} = \boldsymbol{0}. \tag{2.6}$$

The expression for the potential energy U can be arbitrarily complex depending on the characteristics of the simulated scene. The solution to the resulting nonlinear system of equations $\boldsymbol{g}(\boldsymbol{q}) = 0$ provides a new value for the position of all generalized coordinates $\boldsymbol{q}$, and correponding velocities as $\dot{\boldsymbol{q}} = (\boldsymbol{q} - \boldsymbol{q}_0) \cdot \Delta t^{-1}$. Note that solving Equation 2.6 is formally equivalent to the optimality condition of an energy function:

$$E(\boldsymbol{q}) = \frac{\Delta t^2}{2} \ddot{\boldsymbol{q}}^T \boldsymbol{M} \ddot{\boldsymbol{q}} + U, \tag{2.7}$$

for which $\nabla_{\boldsymbol{q}} E = \boldsymbol{g}(\boldsymbol{q})$. This alternative formulation is convenient because the minimization problem can then be solved using any standard nonlinear optimization algorithm. To keep the simulation performance at interactive levels, Verschoor et al. (2018) only perform one Newton iteration per time step, with Preconditioned Conjugate Gradient for the linear solve. This simulation scheme allows us to add an arbitrary number of energy terms to the potential and solve the dynamics equation robustly in a unified way. This formulation can be exploited by modeling all interal forces, contact interactions and coupling mechanisms through energy terms that are summed up to the total:

$$U = U_{\mathrm{H}} + \sum U_{\mathrm{J}} + \sum U_{\mathrm{C}} + \sum U_{\mathrm{D}} + \sum U_{\mathrm{I}} + U_{\mathrm{T}} \tag{2.8}$$

Here, U_{H} models the deformable soft-tissue of the hand, U_{J} represents constraints at the skeletal joints, and U_{C} accounts for the coupling between the hand articulated skeleton and the skin. Alltogether these form our personalized soft-hand model. Additionaly, U_{D} models the behavior of all interactible deformable objects and U_{I} accounts for the interactions through frictional contact constraints. Finally, U_{T} is responsinble for the elastic tracking. In the following, we will briefly explain each of these energy terms.

Deformable Objects

Continuum deformation mechanics can be described as an integral over the object domain of an energy density function $\Psi(\boldsymbol{F})$. This energy function is defined in terms of the deformation gradient, $\boldsymbol{F}$, which is a metric of local deformation. The volumetric deformations are solved by numerically approximating such integral through the Finite Elements Method (FEM). The FEM model discretizes the volume using a tetrahedral mesh $\mathcal{V}$ and use linear interpolation functions to express the deformation within each element in terms of the deformation of the element nodes. This makes possible to approximate the total domain integral as a summation over the M discretization elements. Thus, for each deformable object:

$$U_{\mathrm{D}} = \sum_{M} \Psi(\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{X}) V_i, \tag{2.9}$$

where $\boldsymbol{x}$ and $\boldsymbol{X}$ are, respectively, the deformed and undeformed positions of the element nodes, and V_i is the rest volume of the tetrahedron. For linear interpolation functions, the deformation gradient $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{X})$ has a simple closed form and the elastic behavior of the object is completely determined by the chosen energy density

function Ψ. One desirable property of such a function is the invariance to rotation, *i.e.*, rotational components of the element deformation should not contribute to the increase of the elastic potential. However, this property requires considering high order expressions of Ψ which might be costly to simulate. Instead, Verschoor et al. (2018) use a linear co-rotational model (Müller et al. 2002). Every iteration, the rotation of each deformed tetrahedral element is estimated w.r.t. the rest configuration $\boldsymbol{R}_e$, by computing the polar decomposition of the deformation gradient $\boldsymbol{F}$. Then, a quadratic energy density is formulated depending on the unrotated deformation gradient $\boldsymbol{F}' = \boldsymbol{R}_e^T \boldsymbol{F}$. This way it is possible to keep the simulation at interactive rates and ignore deformation artifacts due to rotations.

Articulated Soft-Hand

The personalized soft-hand model is modeled through a set of rigid-bodies with joint constraints, which are mechanically coupled to a soft-tissue layer. The nonlinear continuum mechanics of the skin are modeled in the same way as any of the other deformable objects in the virtual environment. Thus, U_{H} in Equation 2.8 is formally equivalent to Equation 2.9. However it is known that the nonlinear response is an essential property for an accurate depiction of skin behavior. In order to keep the solution as simple as possible while capturing the nonlinear nature of skin dynamics, we adopt the extension to the linear co-rotational material in Verschoor et al. (2018), where the elastic potential Ψ is augmented with a quickly growing energy term when its value exceeds some specific threshold.

For the articulated rigid-body representing the hand, joint constraints are modeled as a penalty forces. This soft constraint energy penalizes the squared distance between the joint position expressed in terms of both rigid-body coordiantes. For two arbitrary rigid-bodies A and B, this results in an energy

$$U_{\mathrm{J}} = \frac{1}{2} k_{\mathrm{J}} \| \boldsymbol{c}_A + \boldsymbol{R}_A r_A - \boldsymbol{c}_B + \boldsymbol{R}_B r_B \|, \tag{2.10}$$

where $\boldsymbol{c}_A$, $\boldsymbol{c}_B$, $\boldsymbol{R}_A$, and $\boldsymbol{R}_B$ are the positions and rotations of the rigid-bodies, and $\boldsymbol{r}_A$ and $\boldsymbol{r}_B$ are the vectors from the center of mass ot the location of the joint, expressed in the local reference system of each body.

Finally, to ensure the global motion of the soft-tissue is driven by the articulated skeleton, Verschoor et al. (2018) introduce an additional soft constraint that penalizes the distance between the two. The penalty energy penalizes the distance between a few coupling points expressed in terms of both mesh elements nodes and rigid-body coordinates. First, a containing capsule representing each bone is defined and its intersection is computed with the tetrahedral mesh. For each element e intersected by the capsule, its intersection surface is uniformly sampled and the average positions of all the points is computed to obtain a representative coupling point $\boldsymbol{x}_e$. This point can be kinematically expressed in terms of both the capsule rigid-body coordinates and the tetrahedral element nodes $\{\boldsymbol{x}_i\}$ through barycentric coordinates, $\{w_i\}$, $\boldsymbol{x}_e = \sum_i \boldsymbol{x}_i w_i$. This results in an energy term per intersected

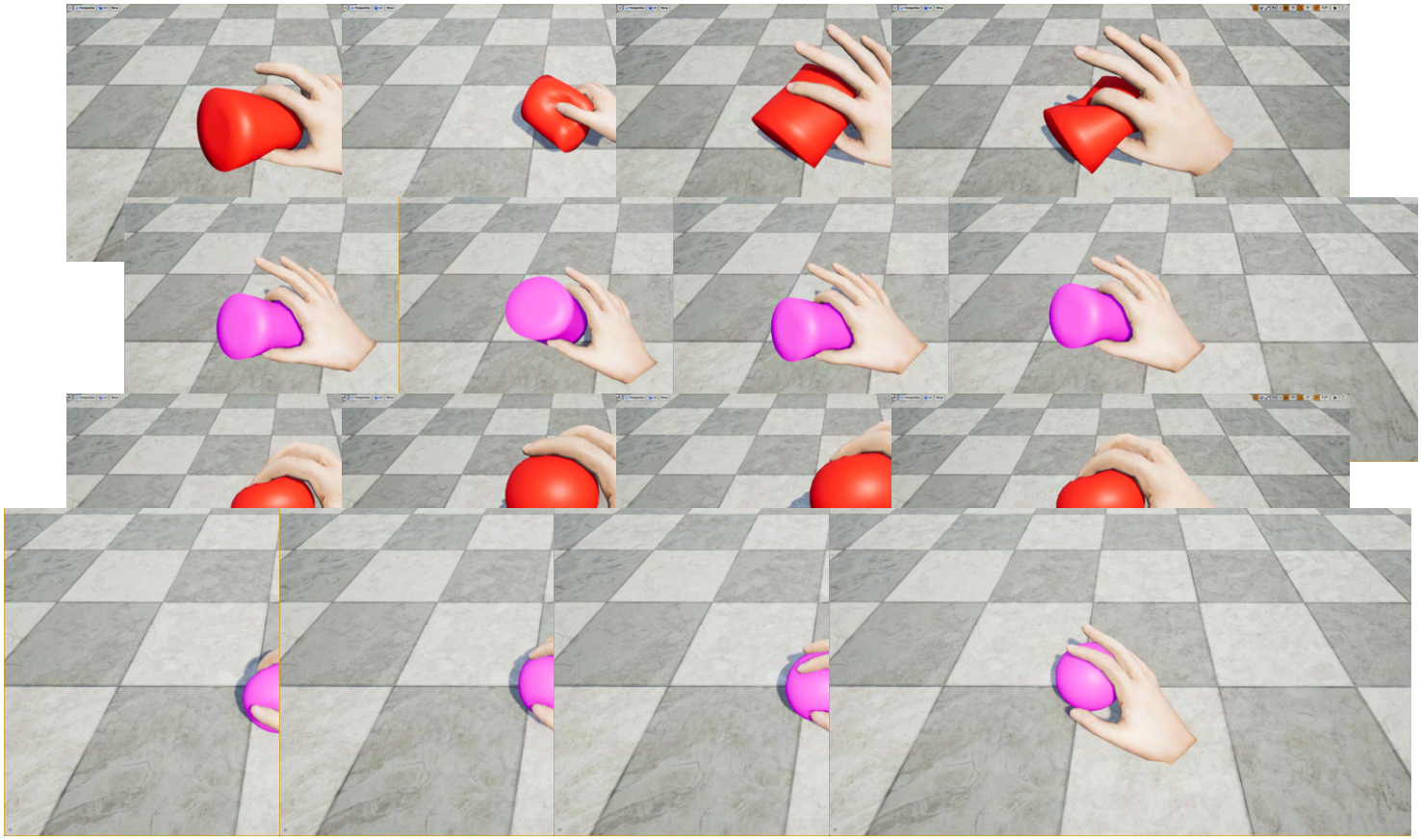


Figure 2.5: *Example of interaction between the simulated hand and various deformable objects using the CLAP library (Verschoor et al. 2018).*

element

$$U_{\mathrm{C}} = \frac{1}{2} k_{\mathrm{C}} \frac{S}{S_T} \| \boldsymbol{c}_c + \boldsymbol{R}_c \boldsymbol{r}_e - \sum_i \boldsymbol{x}_i w_i \|, \tag{2.11}$$

where $\boldsymbol{c}_c$ and $\boldsymbol{R}_c$ are the position and rotation of the capsule and $\boldsymbol{r}_e$ is the position of the coupling point in the local reference system of the capsule. To ensure the coupling force is proportional to the intersected area, each energy is weighted considering the total number of capsule samples, S_{T}, and the number of samples in the area intersected by the element, S.

Frictional Contact

Frictional contact is essential for a correct modeling of the interactions between the hand and the objects of the environment, especially for solving complex tasks involving grasping. The soft-hand simulation considers non-penetration and friction soft constraints between the surface nodes and faces of any two arbitrary objects in the scene. In practical terms, this implies finding, for each node, the closest surface face in any deformable object and query for potential collisions. To make this complex problem tractable, Verschoor et al. (2018) compute a signed distance field (SDF) for each object, which has to be periodically updated to account for possible deformations. At the beginning of the dynamic step, before solving the optimization problem in Equation 2.6, the SDFs are updated and a query is executed

for collisions of surface nodes. When a surface node $\boldsymbol{x}_n$ collides with a face, *i.e.*, resulting in a negative distance query, an anchor point $\boldsymbol{x}_a$ is initialized on the surface of the contacted object. This defines two additional energy terms to account for contact and friction forces as soft constraints, $U_\mathrm{I} = U_\mathrm{N} + U_\mathrm{F}$, where

$$U_N = \frac{1}{2} w k_N \|\boldsymbol{n}\boldsymbol{n}^T(\boldsymbol{x}_a - \boldsymbol{x}_n)\|^2, \tag{2.12}$$

$$U_F = \frac{1}{2} w k_F \|(\boldsymbol{I} - \boldsymbol{n}\boldsymbol{n}^T)(\boldsymbol{x}_a - \boldsymbol{x}_n)\|^2. \tag{2.13}$$

Here, w is a weighing value that is proportional to an area associated with the colliding node and $\boldsymbol{n}$ is the normal of the surface at the anchor. Note that such normal kinematically depends on the the nodes of the colliding face, which are simulated degrees-of-freedom. However, to simplify collision handling, such normals are considered constant within each time step. To account for both static and dynamic friction regimes, the position of the anchor point is updated allowing it to slide. In particular, Verschoor et al. (2018) use the Coulomb friction model to determine the friction regime. If the friction force $\boldsymbol{f}_\mathrm{F} = -\delta U_\mathrm{F}/\delta \boldsymbol{x}_n$ and the normal force, $\boldsymbol{f}_\mathrm{N} = -\delta U_\mathrm{N}/\delta \boldsymbol{x}_n$ satisfy the Coulomb condition $\|\boldsymbol{f}_\mathrm{F} \leq \mu \boldsymbol{f}_\mathrm{N}\|$, for some friction coefficient μ, then friction is in static regime and the anchor is kept fixed. Otherwise, it is in dynamic regime and the anchor slides tangential to the contact surface. In practice, the position of the anchor is recomputed after each step to ensure the Coulomb condition holds.

Elastic Tracking

To control the simulated hand by tracking the motion of the user's real hand, an additional penalty energy U_t is added that tries to minimize the positions and rotations of the articulated rigid body representing the hand and the tracked data. This way, it strongly reduces the impact of discontinuities introduced by the tracking device and large deviations when the simulated hand is constrained by virtual contact. Additionally, to prevent problems due to different simulation and tracking frequencies, Verschoor et al. (2018) use an intermediate skeletal representation that will be leveraged in the work described in Chapter 3, that allows to connect off-the-shelf solutions for hand tracking and physics-based hand simulation, avoiding the need to share a common hand representation.

2.3 3D Hand Tracking

Most hand tracking methods rely on either vision-based or wearable sensors. On one hand, wearable sensors are commonly designed as gloves, making them practical and easy to use. They feature electromagnetic or mechanical sensors that accurately track the flexion and abduction angles of each finger joint and the palm. However, they suffer from limited finger dexterity and are not suitable to all hand sizes. On the other hand, vision-based sensors, or cameras, have become very popular in our daily lives. Ubiquitous and affordable, cameras can capture a broad spectrum of light from visible to infrared. Vision-based sensors offer unrestricted hand

movement and are valuable for applications that includes intricate objects manipulation. Nevertheless, to track hands properly, they depend on consistent hand visibility and can be affected by cluttered backgrounds.

2.3.1 Vision-Based Hand Tracking

Within the methods used in vision-based tracking, we can namely find marker-based and markerless methods. Marker-based hand tracking is a system that relies on the placement of physical markers or tags attached to the hand to track its movements. These markers are usually distinct, easily identifiable objects or patterns that are tracked by cameras or other sensors to determine the hand's position and orientation in space. These methods enable precise tracking of the hand even in complex environments, but require the user to wear or hold markers, which can sometimes be cumbersome. In contrast, markerless 3D hand trackers can accurately detect and track the movements of a person's hand without the need for any physical markers or attachments, by typically using computer vision techniques. It has been an actively researched problem for decades, which can be explained by the fact that it enables many important applications, e. g. in human–computer interaction, activity recognition, or robotics. The problem of tracking a *single* hand is nowadays a mature topic in the Computer Vision community. Pioneering works use multi-camera (Ballan et al. 2012, Oikonomidis et al. 2011*b*, Sridhar et al. 2013, Wang et al. 2011) or marker-based setups such as color gloves (Wang & Popović 2009) to facilitate the detection of hand joints. In the rest of this thesis, we will focus on markerless methods.

Vision-based hand tracking takes as input images of hands or key feature points, and computes the skeletal configuration of the hand that best reproduces the input data. Modern methods can be classified into three large sets.

Generative Methods

In principle, generative approaches work by finding a hand configuration that minimizes an objective function, more precisely, an energy function that defines the proximity between the current hand model configuration and the pixel-wise or heuristic image features. Initialization involves defining a vector of parameters that will then be iteratively updated in order to better fit the input image. By using optimization algorithms to minimize this energy, generative methods attempt to recover the correct model parameters (Karunratanakul et al. 2023, Melax et al. 2013, Oikonomidis et al. 2011*a*,*b*, Qian et al. 2014, Sridhar et al. 2013, Tagliasacchi et al. 2015, Tkach et al. 2016). Generative approaches optimization is sensitive to model initialization, however, they do not require training data and annotations that can be tedious to obtain and can easily include prior information by incorporating losses and constraints during optimisation.

Discriminative Methods

As opposed to generative approaches, discriminative approaches work by inferring the correct hand configuration parameters from the input image. These

approaches thus require a training step and learning-based algorithms to interpret the image features. These algorithms are trained over a large scale image dataset that includes ground truth annotations, whether synthetically or manually obtained.

Numerous investigations have been conducted varying the method and the data: regressing joint locations from depth data (Baek et al. 2018, Chen et al. 2019, Ge et al. 2018, 2016, 2017, Li & Lee 2019, Oberweger et al. 2015, Sinha et al. 2016, Tompson et al. 2014, Wan et al. 2017), from infrared images less fragile to motion blur than depth images (Park et al. 2020), for occluded hand in cluttered scenes (Mueller et al. 2017, Rogez et al. 2014) and based on RGB-based methods to estimate 3D pose (Cai et al. 2018, Chen et al. 2022, Ge et al. 2019, Iqbal et al. 2018, Mueller et al. 2018, Panteleris et al. 2018, Pavlakos et al. 2024, Spurr et al. 2018, Yang et al. 2019, Zimmermann & Brox 2017). Some methods also introduce feature injection and fusion (Park et al. 2022, Xu et al. 2023) to extract as much information as possible from the visible parts to improve the reconstruction and estimation quality. However, obtaining accurate 3D ground truth annotations is arduous and requires specific conditions such as multi-view environments. Moreover, data scarcity makes discriminative methods prone to over-fitting to dataset biases, and without explicit knowledge of hand geometry, reconstruction failures can distort hands. In some cases, manually label data may be a solution, but for tasks such as segmentation or depth estimation it is hardly possible.

Hybrid Methods

To reap the benefits of both approaches, hybrid approaches are investigated to join the complementary advantages of generative and discriminative methods. One way to combine these approaches consists of using a discriminative approach to extract a visible features from an image, such as joint positions or finger tips (Baek et al. 2019, Hampali et al. 2020, Pavlakos et al. 2019, Taylor et al. 2016, Xiang et al. 2019, Ye et al. 2016), segmentation (Sridhar et al. 2015), or dense correspondence (Mueller et al. 2019). These features are then used to define an energy in an optimization problem. In reverse, it is also possible to use the hand model parameters such as shape and pose, as a constrained output space as prior information, and use them in the optimization energy as losses to train the image-to-pose mapping function (Boukhayma et al. 2019, Zhou et al. 2016, Zimmermann et al. 2019). Some studies propose self-supervised methods in order to minimize the need for labeled data (Chen et al. 2021, Jiang et al. 2023, Tu et al. 2023, Wan et al. 2019).

2.3.2 Hand-Object Tracking

Also related to this thesis are the recent tracking methods that are able to estimate the hand pose while manipulating rigid objects (Sridhar et al. 2016, Taheri et al. 2020). However, physically-correct interactions cannot be enforced since forces are not modeled.

One of the important ongoing challenges in hand tracking is the reconstruction of two-hand motion and/or the interaction of hands with other objects. In

this regard, hand simulation models provide a convenient prior to hand tracking algorithms. Tzionas et al. (2016) combined a generative model with physics-based simulation to track hands interacting with each other and/or with objects. Hasson et al. (2019), Mueller et al. (2018), Rogez et al. (2015) improved the quality of joint hand and object tracking by incorporating loss functions that favor plausible hand-object interactions by reconstructing the hand pose only, while Doosti et al. (2020), Grady et al. (2021), Hampali et al. (2020), Hasson et al. (2020), Karunratanakul et al. (2020), Kyriazis & Argyros (2014), Liu et al. (2021), Park et al. (2022), Tekin et al. (2019), Tzionas et al. (2016), Xu et al. (2023) restrict the object class. Garcia-Hernando et al. (2018) produced annotated benchmarks of hand-object interaction, which can serve to further improve tracking algorithms. Mueller et al. (2019) presented the first method that can track in real time two hands in intricate contact. They generated synthetic two-hand simulation data using a hand simulation engine, and then used these annotated data to train a learning-based tracking algorithm. Their method also fits the shape of a parametric hand model to the captured images. We use a similar pipeline in the methods described in Chapters 4 and 5.

Two-Hand 3D Tracking.

Very few methods exist that tackle the challenging *two-hand* tracking problem. Oikonomidis et al. (2012) introduced one of the first attempts, which uses a multi-camera setup to circumvent the challenges caused by the unavoidable strong self-collisions. A few follow-up works also use multi-view setups (Han et al. 2020, 2018, Simon et al. 2017) and markers to ease the two-hand tracking problem, while others investigated the use of single depth sensors (Mueller et al. 2019, Taylor et al. 2017) which simplifies the setup but also provides sufficient cues to resolve depth ambiguities.

Closer to the work presented in Chapters 4 and 5, some recent methods consider a monocular RGB image as input. In Chapter 4, we introduce one of the first methods, which combined learned pixel-to-surface predictions with model-fitting to estimate the pose and shape of the two hands. Concurrently, Moon et al. (2020) proposed a method that directly predicts 3D hand joint positions from RGB images. Follow-up works attempt to model the inherent depth ambiguity by explicitly modeling a visibility term (Kim et al. 2021), or by using probabilistic models for hand part segmentation (Fan et al. 2021) or 3D pose (Wang et al. 2022).

Finally, the state-of-the-art method of Li et al. (2022) uses a graph representation combined with attention modules to infer vertex positions of two interacting hands, used as a baseline in Chapter 5. Yu et al. (2023) learn independent features for each hand and exploit attention-conditioned cross-hand prior to mitigate interdependencies. Despite tremendous advances in 3D hand tracking of two hands from RGB, these methods suffer from residual errors in depth estimation that prevent the computation of accurate hand contacts.

Contact from RGB

This thesis is also significantly related to the methods that, instead of just estimating hand poses, focus on estimating hand-object contact (Chen et al. 2023, He et al. 2021, Narasimhaswamy et al. 2020, Xie et al. 2022). This problem has been addressed by data-driven pipelines that leverage large annotated datasets of hands manipulating rigid objects (Liu et al. 2021, Shan et al. 2020, Taheri et al. 2020). However, most of these methods estimate hand contact given 3D information of the scene or the target object (Cao et al. 2021, Grady et al. 2021, Jiang et al. 2021, Taheri et al. 2020). Instead in Chapter 5, we attempt to estimate camera-space contact labels for two hands without using any additional cue or annotation. Closest to us are the few methods that estimate body-body contacts directly from RGB images. For example, Fieraru et al. (2020) learn to estimate coarse subject-to-subject labels using a manually annotated dataset of human interactions. Huang et al. (2022) are able to predict dense per-vertex contact labels by exploiting image information, without reconstructing 3D poses or 3D bodies.

2.4 Embodiment

There is a wide line of research that studies the effect of the representation of the hand on the embodiment of the VR user (Argelaguet et al. 2016). To ground this discussion, we draw on the widely accepted definition of embodiment proposed by Kilteni et al. (2012), which refers to the sensation of being situated within a virtual body, having control over it, and perceiving it as one's own, particularly in virtual reality applications. This concept consists of three core subcomponents: self-location, agency, and body ownership. Agency refers to the perception of controlling the virtual body with intentional actions (Braun et al. 2018). Body ownership reflects the experience of perceiving the virtual body as the source of sensations (Braun et al. 2018). Finally, self-location captures the feeling of being physically present in a virtual environment, even while being aware of the true location of one's body.

Building on this understanding of embodiment, most works focus on investigating how various aspects of VR visualization and interaction influence this phenomenon. For example, studies often analyze the virtual hand illusion (Lin et al. 2019). There is a general consensus that VR users can still experience a sense of embodiment even when the virtual hand differs significantly from their real hand in shape, size, motion, or appearance (Sanchez-Vives et al. 2010). Due to the inherent imprecision of proprioception, which is the subconscious ability to perceive the position, movement, and orientation of one's body parts without relying on visual input (Tuthill & Azim 2018), VR users can tolerate significant discrepancies between the position and pose of virtual hands relative to their own hands. Such discrepancies can even be leveraged in the VR simulation to model contacts with virtual objects in a more realistic way (Cosco et al. 2013).

This line of research is orthogonal to this thesis. Its conclusions suggest that achieving embodiment does not necessarily require a perfect match between the

user's real and virtual hand. Instead, it is possible even when notable differences exist, such as those in the shape and size of the virtual hand. This highlights the need for methods to reconcile the tracked and simulated hand representations, as described in Chapter 3. Some authors have addressed the problem of motion retargeting across characters of very diverse morphology (Hecker et al. 2008), or even within video-to-video (Chan et al. 2019).

3 Fine Virtual Manipulation with Hands of Different Sizes

3.1 Introduction

When virtual objects appear real, the next natural step is to reach out and start interacting with them (Chessa et al. 2019). But this apparently simple action entails additional tasks in VR, introduced in Section 1.1: hand tracking and hand-object interaction, which are typically solved independently. For hand tracking, the common solution is to use computer vision methods, which output the skeletal morphology and configuration of a hand that best matches the user's actual hand (Mueller et al. 2019), see Section 2.3 for a detailed review of the existing methods and discussion. For hand-object interaction, the most general approach is to find the configuration of a simulated hand that takes the tracked hand as goal, but is subject to a model of hand biomechanics and the laws of contact mechanics (Verschoor et al. 2018). Some modern commercial hand-tracking solutions, such as Oculus Quest, provide some limited hand interactions by building an ad-hoc physics-based model on top of the tracked hand morphology. In this Chapter, we address challenges arising in the connection of hand tracking and hand-object simulation, and as a result, we aim for VR animations of fine object manipulation, commanded by interactive hand tracking of the user's hands.

When connecting hand tracking and hand-object simulation, we find that the hand models used in these two tasks may differ in size and skeletal morphology. These differences may be due to at least two major reasons: First, it is non-trivial to produce a simulation model that fits exactly the size and morphology of the user's hand. Even though embodying the user in an avatar with different hand size is perfectly viable from a perceptual point of view (Argelaguet et al. 2016), it is not free of technical difficulties. Second, to leverage existing work in hand tracking and hand simulation, it is convenient to integrate off-the-shelf solutions, but it is unlikely that these solutions use hand representations with the same skeletal morphology. For instance, there is no consensus on the placement of joints across different hand representations, particularly at the palm.

Due to these differences in hand size and skeletal morphology, the hand pose computed by hand tracking cannot be directly input to hand-object simulation. If the pose is applied naïvely, it results in inaccurate finger configurations, which complicate dexterous manipulation of virtual objects. Thanks to visual feedback

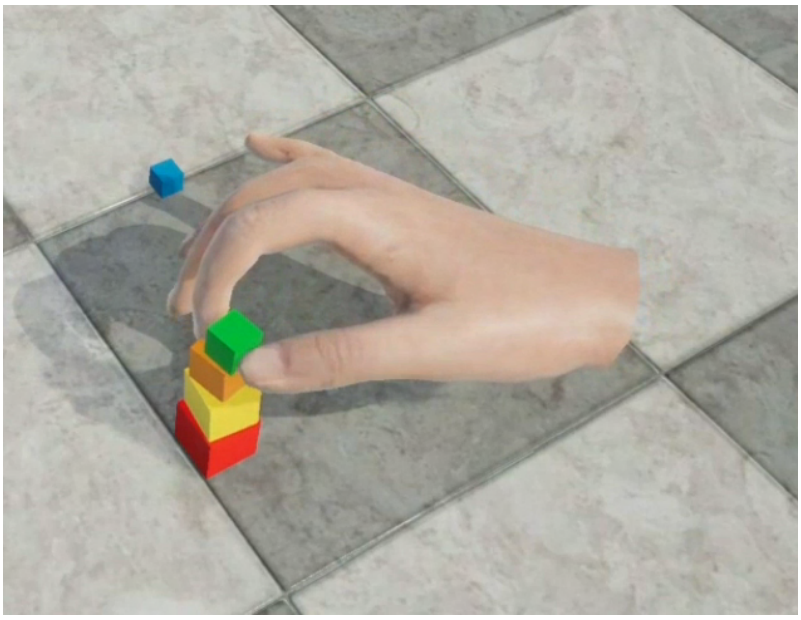


Figure 3.1: *Natural physics-based interaction with small objects. This VR scene was implemented by connecting off-the-shelf hand tracking and hand simulation solutions, which use hand models with different skeletal morphology.*

of the simulated hand, the user may correct the real hand pose and try to work around the mismatch. We have found that this is sufficient for gross manipulation of virtual objects. However, some finger configurations are impossible to reach when the pose of the tracked hand is applied naïvely to the simulated hand, which altogether prevents dexterous fine manipulation of virtual objects.

In this chapter, we introduce a pose retargeting strategy to connect the tracked hand and the simulated hand. Our approach works with any type of tracking or simulation method, as it stands at the interface between both tasks. We use an intermediate hand representation that shares the size and morphology of the simulated hand, but which tries to match the configuration of the tracked hand. The retargeting strategy formulates an optimization of the pose of this intermediate hand, based on features that represent the pose of the tracked hand. We have used finger tip positions as features, as they represent key information for fine manipulation. We describe the formulation and solution to the optimization problem in Section 3.2, together with a brief summary of the hand-object simulation using the CLAP library (Verschoor et al. 2018).

We have evaluated the practical impact of the hand mismatch on the manipulation of virtual objects, and we have compared our pose retargeting strategy vs. naïve copy of the hand pose. To this end, we have carried out a user study, discussed in Section 3.3, comparing task performance of virtual object manipulation. We have confirmed that the mismatch of the hand representation is not critical for gross manipulation (i.e., large objects), but it is critical for fine manipulation (i.e., small objects). With our pose retargeting approach, the performance of pick and drop actions for small objects is significantly faster than the performance of

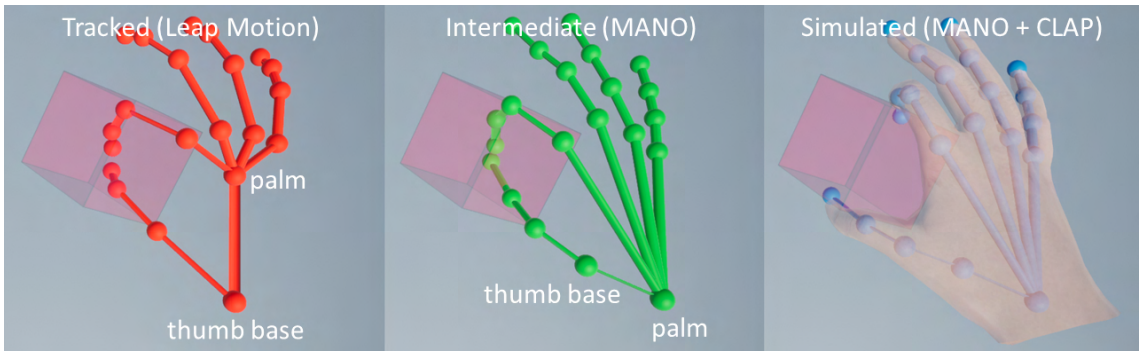


Figure 3.2: *Left: Tracked hand (in red), obtained using a Leap Motion tracker. Right: VR simulated hand (skeleton in blue, flesh semitransparent) implemented using CLAP (Verschoor et al. 2018) with MANO hand representation (Romero et al. 2017). The skeletal morphology differs, in particular at the palm and thumb joints. Moreover, the simulated hand is constrained by contact with VR objects, while the tracked hand is not. Middle: We connect both hands using an intermediate representation (in green), which shares the morphology of the simulated hand but matches the pose of the tracked hand.*

naïve strategies, and users also report increased precision, naturalness and ease of manipulation.

3.2 Tracking-Based Hand Animation

As discussed in the introduction of this chapter, we wish to drive a VR simulated hand model using as input interactive hand tracking data. However, the representations of the simulated hand and the tracked hand may differ in size and skeletal morphology. Furthermore, the simulated hand is constrained by contact with objects in the VR scene, while the tracked hand is not.

We use an intermediate hand representation to connect the user's tracked hand and the VR simulated hand. This intermediate hand shares some properties with the tracked hand (i.e., it is not constrained by other VR objects), and other properties with the simulated hand (i.e., its size and skeletal morphology). We characterize all three hand instances by their skeletal pose θ. Then, formally we denote the three following hand poses: θ^{t} for the user's tracked hand, θ^{i} for the intermediate hand representation, and θ^{s} for the VR simulated hand. Let us emphasize that, even though we use the same symbol θ to conceptually represent pose for all three hand instances, the joint angles of the tracked hand may have a different geometric interpretation from those of the intermediate and simulated hands, due to the differences in skeletal morphology. Recall that our method is general and works for any hand representation, hand tracker, and hand simulation method. In our implementation, we use Leap Motion as tracker, with its corresponding hand representation, and the MANO representation (Romero et al. 2017) for the simulated hand. Figure 3.2 shows schematically the interconnection of all three hand instances.

The intermediate hand representation serves as target configuration for the VR

hand simulation. By matching the skeletal morphology of the simulated hand, it is easy to formulate input forces and torques for each bone in the simulated VR hand. These forces and torques are combined with contact forces and elastic deformation forces to produce the overall smooth simulation of the VR hand.

We start this section by describing a pose retargeting strategy to compute the intermediate hand θ^{i}. We motivate the formulation of the strategy to optimize fine manipulation tasks, and we describe an efficient solution algorithm. We conclude the section with a summary of the physics-based hand simulation method.

3.2.1 Hand Pose Retargeting

Given a pose of the tracked hand θ^{t}, we wish to compute a pose of the intermediate hand θ^{i}, such that it retains the most relevant characteristics, despite skeletal differences. We do this by defining a set of features $\vec{f}$, and solving an optimization problem. Prior to this, we apply a uniform scale to the tracked hand, such that it matches the overall size of the simulated hand. We do this by fitting a bounding box to an open palm pose.

The pose features should describe important characteristics of the pose, but with no assumptions about the skeletal morphology or size. In this chapter, we focus on the ability to manipulate small objects with high dexterity; therefore, we define the feature vector $\vec{f}$ by concatenating the positions of finger tips. In Section 3.3 we demonstrate that our pose retargeting approach is effective at producing dexterous manipulations of fine objects.

Based on this feature vector, we formulate the computation of the intermediate hand pose as the solution to the following constrained optimization problem:

$$\theta^{\mathrm{i}} = \arg\min_{\theta^{\mathrm{i}}} \frac{1}{2} \left(\vec{f}(\theta^{\mathrm{i}}) - \vec{f}(\theta^{\mathrm{t}})\right)^{\mathrm{T}} W \left(\vec{f}(\theta^{\mathrm{i}}) - \vec{f}(\theta^{\mathrm{t}})\right) + R(\theta^{\mathrm{i}}), \qquad (3.1)$$
$$\text{s.t. } \vec{c}(\theta^{\mathrm{i}}) \geq 0.$$

In a nutshell, the optimization finds the pose of the intermediate hand that produces features (i.e., finger tip positions) as close as possible to those of the tracked hand. W represents a (diagonal) weight matrix for the different features, which allows us to put more emphasis on the motion of the thumb and the index, for very accurate pinching. $R(\theta^{\mathrm{i}})$ is a regularization term; we use a small spatial and temporal regularization, to smooth interphalangeal rotations and avoid temporal discontinuities. $\vec{c}(\theta^{\mathrm{i}})$ represents constraints, to handle joint limits in the optimization.

We solve the optimization problem in Equation 3.1 iteratively using the Gauss-Newton method (Nocedal & Wright 2006). On each iteration, given a current estimate θ^{i}_0 of the pose of the intermediate hand, we linearize the feature vector as $\vec{f}(\theta^{\mathrm{i}}_0) + \vec{f}\theta^{\mathrm{i}}\,\Delta\theta^{\mathrm{i}}$, and the active constraints as $\vec{c}(\theta^{\mathrm{i}}_0) + \vec{c}\theta^{\mathrm{i}}\,\Delta\theta^{\mathrm{i}}$. Then, with Lagrange multipliers λ to enforce the active constraints, each iteration of Gauss-Newton

amounts to solving the following linear system:

$$\begin{pmatrix} \vec{f}\theta^{\mathrm{i}\,\mathrm{T}}\, W\, \vec{f}\theta^{\mathrm{i}} + R\theta^{\mathrm{i}} & \vec{c}\theta^{\mathrm{i}\,\mathrm{T}} \\ \vec{c}\theta^{\mathrm{i}} & 0 \end{pmatrix} \begin{pmatrix} \Delta\theta^{\mathrm{i}} \\ \lambda \end{pmatrix} = \begin{pmatrix} \vec{f}\theta^{\mathrm{i}\,\mathrm{T}}\, W\, \left(\vec{f}(\theta^{\mathrm{t}}) - \vec{f}(\theta_0^{\mathrm{i}})\right) - R\theta^{\mathrm{i}\,\mathrm{T}} \\ -\vec{c}(\theta_0^{\mathrm{i}}) \end{pmatrix}. \tag{3.2}$$

To solve the linear system, we compute a Cholesky factorization of the matrix $\vec{f}\theta^{\mathrm{i}\,\mathrm{T}}\, W\, \vec{f}\theta^{\mathrm{i}} + R\theta^{\mathrm{i}}$, then we use a Schur-complement approach to compute the Lagrange multipliers λ, and we conclude by computing the pose update $\Delta\theta^{\mathrm{i}}$.

Let us pay some attention to the computation of gradients $\vec{f}\theta^{\mathrm{i}}$. Without loss of generality, finger tip positions $\vec{f}_j$ can be computed from the hand pose as a concatenation of relative rotations:

$$\vec{f}_j = \vec{x}_{\mathrm{wrist}} + R_{\mathrm{palm}} \left(\vec{x}_{\mathrm{palm}} + R_1 \left(\vec{x}_1 + R_2 \left(\vec{x}_2 + R_3\, \vec{x}_3\right)\right)\right). \tag{3.3}$$

$\vec{x}_k$ and R_k denote, respectively, phalanx bone vectors and joint rotations. For gradient computations, we use a tangent-space representation of rotations (Taylor & Kriegman 1994). In a nutshell, given the current joint rotation R_0, a joint axis $\vec{k}$, and a differential rotation angle ϕ, the joint rotation can be expressed as $R = \left(I + \phi\, \mathrm{skew}(\vec{k})\right) R_0$. The gradient with respect to the rotation angle is then simply $R\phi = \mathrm{skew}(\vec{k})\, R_0$. We use this expression to compute the gradient of finger tip positions with respect to each component of the hand pose in Equation 3.3, and thus we assemble the full gradient $\vec{f}\theta^{\mathrm{i}}$ to be used in Equation 3.2. In some joints, rotations have two degrees of freedom, and are expressed as the concatenation of two one-degree-of-freedom rotations.

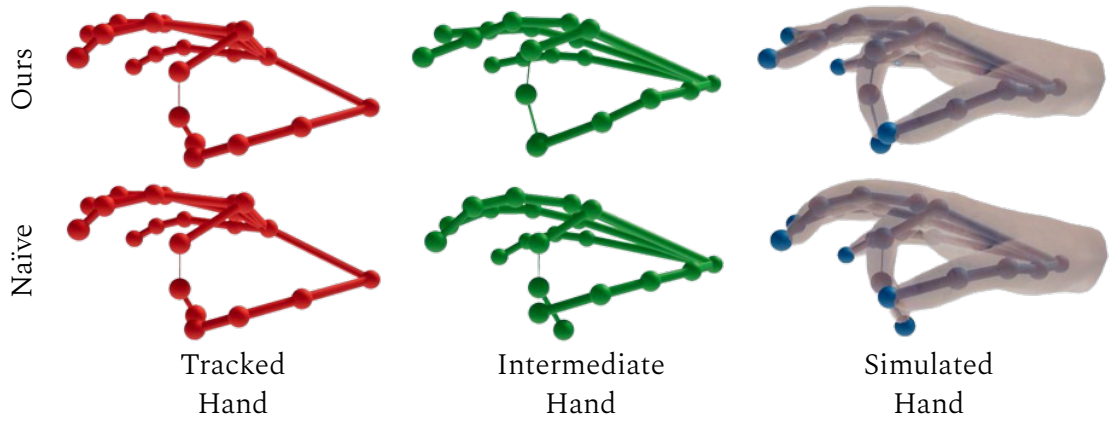


Figure 3.3: *With our pose retargeting (top), thumb-index pinch motions of the tracked hand are accurately reproduced on the simulated hand, despite skeletal differences. Naïve retargeting (bottom) does not reach comparable accuracy, which complicates fine manipulation.*

Figure 3.3 depicts the accuracy of our pose retargeting strategy on a thumb-index pinch pose. We also compare the accuracy of a naïve retargeting strategy. For

this, we align the intermediate and tracked hands using the base positions of the four fingers (which are very similar on the Leap Motion and MANO skeletons), we place the palm and thumb base joints based on the dimensions of the MANO skeleton, and then we copy the relative Leap Motion pose to the intermediate MANO hand. As shown in the figure, the naïve strategy fails to reproduce pinch poses correctly, which complicates fine manipulation.

The constrained Gauss-Newton solver requires on average 16 iterations per frame. This amounts to a total cost of 5 ms. per frame in our implementation on a commodity processor.

3.2.2 Hand Simulation

On every simulation frame, once the pose of the intermediate hand is computed following the retargeting approach described above, we use it to command the simulation of the VR hand. As mentioned in the introduction of this chapter, we use the freely available CLAP simulation library to this end (Verschoor et al. 2018). The decomposition of the hand animation into two subproblems, retargeting and simulation, simplifies the overall approach, and allows us to leverage off-the-shelf simulation libraries. As a result, we can create with no effort VR scenes with physics-based contact and natural hand interaction, as the one shown in Figure 3.1.

The hand simulation used in our method is detailed in Section 2.2.2. In essence, the simulation degrees of freedom are: the positions and orientations of bones (gathered as the pose of the simulated hand, θ^s), the nodes of a tetrahedral decomposition of the skin, and the position and orientation of a grasped object (if it exists). These degrees of freedom are computed jointly on each simulation frame, by solving an optimization-like formulation of implicit integration of the deformation dynamics (Gast et al. 2015, Martin et al. 2011). The formulation of dynamics gathers multiple potential energy terms, which model the physical behavior of the hand, its interaction with the grasped object, and also the command provided by hand tracking. A comprehensive description of the energy terms is provided in Section 2.2.2. In practice, to set up the hand simulation, we first generate the rest-shape geometry of the hand surface and the skeleton based on particular values of the shape parameters of the MANO model (Romero et al. 2017). Then, we tetrahedralize the volume of the hand, connect internal nodes of the tetrahedral mesh to the bones, and pass this simulation model to the CLAP library (Verschoor et al. 2018). The simulation model consists of 16 bones and 2,291 tetrahedra. The cost of the physics-based simulation is 51 ms. on average per frame. This adds some latency, but it did not seem to affect the quality of interaction.

3.3 User Experiment

To evaluate our hand pose retargeting strategy and compare it to naïve retargeting, we have designed a user experiment. In this experiment, users must execute a manipulation task, and the results confirm that our pose retargeting strategy enables more effective manipulation of virtual objects when fine dexterity is needed.

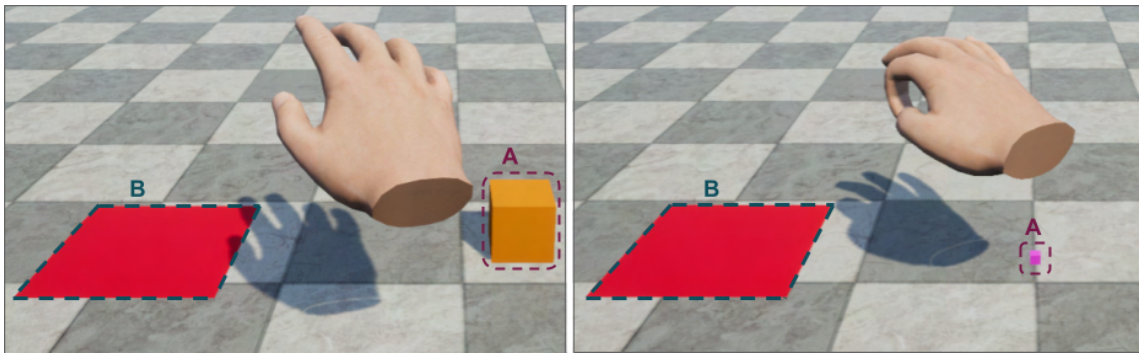


Figure 3.4: *We have studied a cube manipulation task, where users were asked to move a cube from A to B. In a user experiment, we have compared Our pose retargeting Strategy vs. a Naïve Strategy on two Manipulation scenarios: Gross Manipulation (i.e., large cube, left) and Fine Manipulation (i.e., small cube, right).*

In this section, we describe the experiment and discuss the results.

3.3.1 Methods

Participants. A total of 20 right-handed participants (age: range = 18-34, Mean (M) = 26.2, Standard Deviation (σ) = 4.04; 15 male and 5 female) took part in the user study. They received no compensation. In addition to age, we documented their hand size (range = 15.9-20.5 cm, M = 18.80, σ = 1.23) and prior VR experience (10 participants had experience, 10 had none), in order to test the influence of these variables. All participants confirmed correct vision with the HMD. The study was conducted in accordance with the 1964 Declaration of Helsinki and was granted ethical approval by the local ethics committee at Universidad Rey Juan Carlos. All participants provided informed written consent beforehand.

Materials and experimental design. We studied a manipulation task where users were asked to pick a cube with their thumb and index finger from position A, move it to position B, and drop it there, as shown in Figure 3.4. The interaction between the user's hand and the cube, as well as the interaction of the cube with the floor, were simulated with full physics-based contact, as described in Section 3.2.2. However, we disabled contact between the simulated hand and the floor. In the absence of haptic feedback, we found that the response of floor-hand contact could be unintuitive and distort the experiment. The scenarios were developed in Unreal Engine, and were displayed on an Oculus Rift HMD, with head tracking, to optimize vision-motion correlation and make the manipulation task very natural. The hand of the user was tracked using a Leap Motion device, mounted frontally on the HMD for optimal tracking accuracy of grasping and pinching poses. The HMD was sanitized after each use, and soft components were covered with disposable hygienic pads.

The simulated VR hand was the same throughout the experiment. It was a large hand, 21.4 cm long, generated by adjusting the value of the main component of the

statistical MANO model (Romero et al. 2017). Therefore, the skeletal morphology of both the simulated and intermediate hands corresponded to the MANO model. The skeletal morphology of the tracked hand corresponded instead to the morphology of the Leap Motion model. Moreover, the size of the tracked hand was adapted to each user, thanks to the built-in functionality of the Leap Motion.

Two different strategies were compared in the study: the pose retargeting strategy described in Section 3.2.1, referred to as Ours, and a naïve retargeting strategy carried out by aligning the palms of the tracked and intermediate hands and then directly copying the joint angles of the tracked hand to the intermediate hand, referred to as Naïve.

In addition to the retargeting strategy, two different manipulation scenarios were studied: manipulation of a large cube 6 cm wide (i.e. Gross Manipulation) and a small cube 1 cm wide (i.e. Fine Manipulation), as shown in Figure 3.4.

Hypothesis. The initial hypothesis of the study is that the retargeting strategy may have an effect on the dexterity of manipulation; therefore, it may affect task performance on the Fine Manipulation scenario in which the small cube is manipulated, and to a lesser or no extent on the Gross Manipulation scenario in which the large cube is manipulated.

Experimental procedure. In each experimental trial, users were asked to execute the cube manipulation task. Each participant tested both types of Manipulation (i.e. Gross vs. Fine) under both Strategy conditions (i.e. Naïve vs. Ours), a total of five trials per Manipulation and Strategy. Having five repetitions of the condition allowed us to evaluate the effect of task learning. Participants completed five experimental blocks, each with four trials, one per experimental condition; on each block, the order of the four combinations of Manipulation and Strategy was randomized, to avoid potential bias due to task learning. Each experimental block lasted on average six minutes, and the full procedure lasted 30 minutes.

Measures and questionnaire. The time needed to complete the manipulation task in each experimental condition was measured. In addition, participants were asked to complete a questionnaire during the last experimental block, after each trial, i.e., once per condition. The questionnaire contained three items (5-point Likert-type), and was used to assess participants' subjective feelings of dexterity of manipulation in each experimental condition, in terms of Precision (defined to participants as "the movements of the virtual hand respond precisely to the movements made in reality", and ranging from "Not precise at all" to "Very precise"), Ease (defined to participants as "repetitions needed to achieve the objective", and ranging from "A lot of effort" to "Very little effort"), and Naturalness (defined to participants as "the way of grasping virtual objects corresponds to the way of grasping objects in the real world" and ranging from "Not at all natural" to "Very natural").

Data analysis. Data were statistically analyzed using R software. Time data was analyzed with repeated measures analyses of variance (ANOVA) with 2x2x5 within-subject factors Manipulation, Strategy and Repetition. In case of significant inter-

actions between factors, these were followed by t-tests comparing all conditions against each other to understand if there were differences between them, with the p-value adjusted with the recommended Tukey method for comparing a family of estimates (Wobbrock et al. 2011).

For questionnaire data, we conducted non-parametric Friedman tests to assess significant differences between the four conditions (i.e., Gross and Fine Manipulation under both Strategy conditions, i.e., Naïve vs. Ours). Significant results were followed by pairwise comparisons using Wilcoxon signed-rank tests comparing all four conditions against each other, with the p-value adjusted using the Bonferroni multiple testing correction method.

3.3.2 Results and Analysis

Time to complete the task. As shown in Figure 3.5, the Manipulation condition influenced the time to complete the task, but this influence was different depending on the Strategy. The ANOVA on the time data showed a significant difference between Manipulation conditions (F-Statistics (F)(1,19)=15.21, P-Value (p)<0.001), due to longer times needed to complete the task for the Fine than for the Gross Manipulation. Critically, there was a significant interaction between Manipulation and Strategy ($F(1,19)=8.78$, $p=0.008$). T-tests comparing the four conditions against each other showed the expected significant difference between the Gross and Fine Manipulation for the Naïve Strategy ($t(19)=4.881$, $p<0.001$), but this difference did not reach significance for Our Strategy ($p=0.34$). Importantly, t-tests also revealed that, while there were no significant differences between Strategies for the Gross Manipulation ($p=0.85$), the time to complete the task for the Fine Manipulation was significantly smaller for Our Strategy than for the Naïve Strategy condition ($t(19)=3.52$, $p=0.006$). This result indicates that the Fine Manipulation was easier to perform with Our Strategy. This is further evidenced by the results showing a significant difference between the Fine Manipulation with Naïve Strategy vs. Gross Manipulation with Our Strategy ($t(19)=4.36$, $p=0.001$), but not between the Gross Manipulation with Naïve Strategy vs. Fine Manipulation with Our Strategy ($p=0.11$).

Regarding the effect of Repetition, there was a significant main effect ($F(4,76)=10.01$, $p<0.001$), showing an effect of learning with more repetitions. Further, a significant interaction of Repetition with the factor Manipulation ($F(4,76)=4.71$, $p=0.002$) was found; as it can be seen in Figure 3.6-left, the effect of learning with more repetitions was larger for the Fine than for the Gross Manipulation, because of the task being easier for the Gross Manipulation, which was expected. Importantly, there was also a significant interaction of Repetition with the factor Strategy ($F(4,76)=3.69$, $p=0.008$), showing that the effect of learning with more repetitions was larger for the Naïve Strategy than for Our Strategy. As it can be seen in Figure 3.6-right, for the very first trial it took less time to perform the task with Our Strategy than with the Naïve strategy ($t(19)=3.82$, $p=0.01$). This suggests that Our Strategy makes the task easier and more natural than the Naïve Strategy. There was not a significant

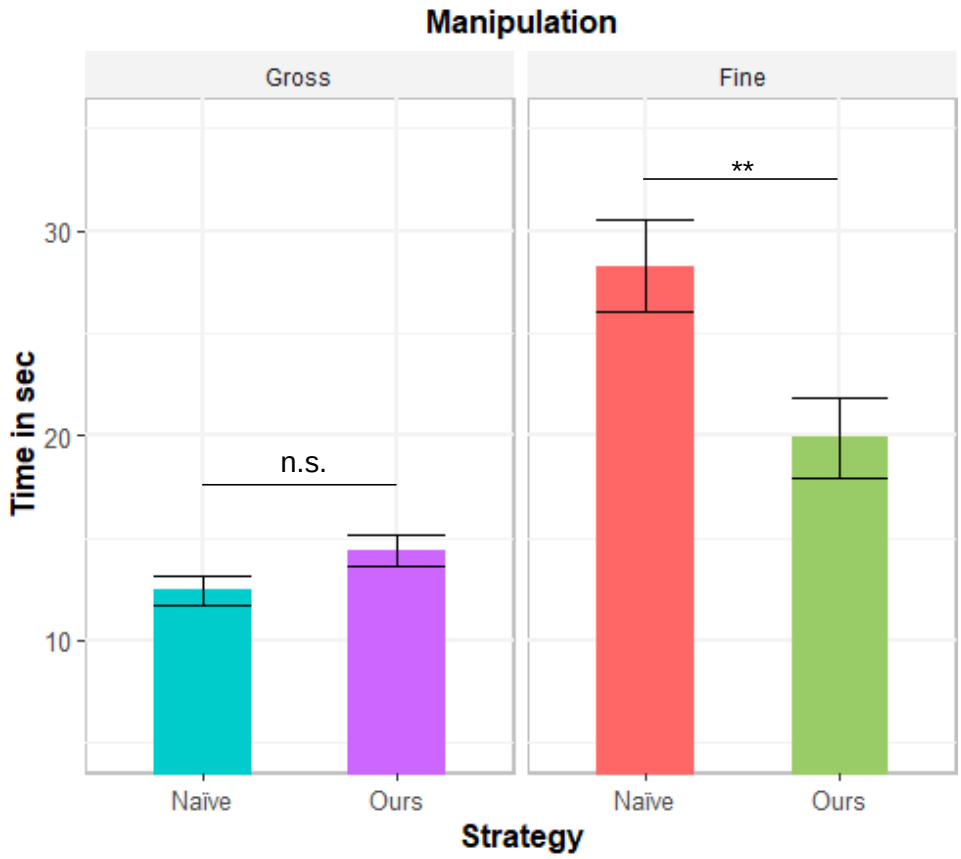


Figure 3.5: *Mean (±Standard Error (SE)) time to complete the task for the four experimental conditions. Asterisks denote significant differences between means (** denotes p<0.01). n.s. denotes no significant differences between means. Note that only the relevant comparisons are displayed in the figure; results from all comparisons are described in the main text.*

triple interaction between the factors Repetition, Manipulation and Strategy, accounting for different effects of Repetition for the two types of Manipulation with the two different Strategies.

To test the potential influence of the individual factors age, hand size and prior VR experience on the observed effects, we ran additional ANOVAs with factors Manipulation and Strategy in which we added age and hand size as covariates, and prior VR experience as a between-subjects factor. We observed a significant interaction of the factor Manipulation only for the factor hand size ($F(1,18)=8.38$, $p=0.01$). As shown in Figure 3.7, while overall it took less time for participants with smaller hand sizes to complete the task, this facilitation effect for smaller hand sizes was more evident for the Fine Manipulation condition. Importantly, the factors hand size, VR experience and age did not interact significantly with the factor Strategy (all $ps>0.12$), which indicates that the results related to Strategy were not significantly affected by these factors.

Self-report measures. Figure 3.8 shows the participants' subjective feelings of

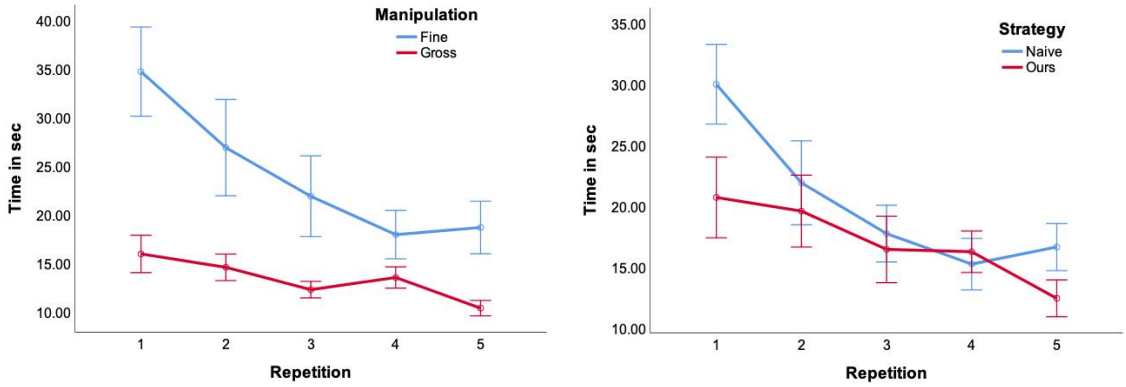


Figure 3.6: *Mean (±SE) time to complete the task across repetitions for the two Manipulation conditions (left) and for the two Strategy conditions (right).*

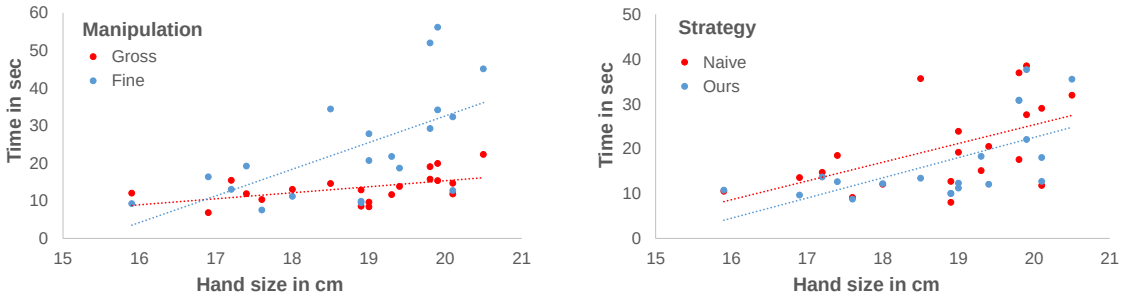


Figure 3.7: *Mean (±SE) time to complete the task for the two Manipulation conditions (left) and for the two Strategy conditions (right) according to participant's hand size.*

dexterity of manipulation, in terms of Precision, Ease and Naturalness, for each of the four experimental conditions. The precision score was statistically significantly different across conditions ($X^2(3)=14.43$, p=0.002). This was also the case for the effort score ($X^2(3)=15.36$, p=0.001) and the naturalness score ($X^2(3)=12.34$, p=0.006).

In terms of precision, pairwise Wilcoxon signed rank tests comparing the four conditions against each other showed that participants felt more precise in the Fine Manipulation with Our Strategy than with the Naïve Strategy (p=0.017), as shown in Figure 3.8-left. Further, for the Naïve Strategy, the Fine Manipulation felt significantly less precise than the Gross Manipulation (p=0.021), while this was not the case for Our Strategy, for which there were not significant differences between Manipulation conditions. Other comparisons between conditions were not significant either.

In terms of effort, pairwise Wilcoxon signed rank tests comparing the four conditions against each other showed that participants felt they had applied significantly less effort in the Fine Manipulation with Our Strategy than with the Naïve Strategy (p=0.013), as shown in Figure 3.8-middle. The Fine Manipulation with the Naïve Strategy required also more effort than the Gross Manipulation both with

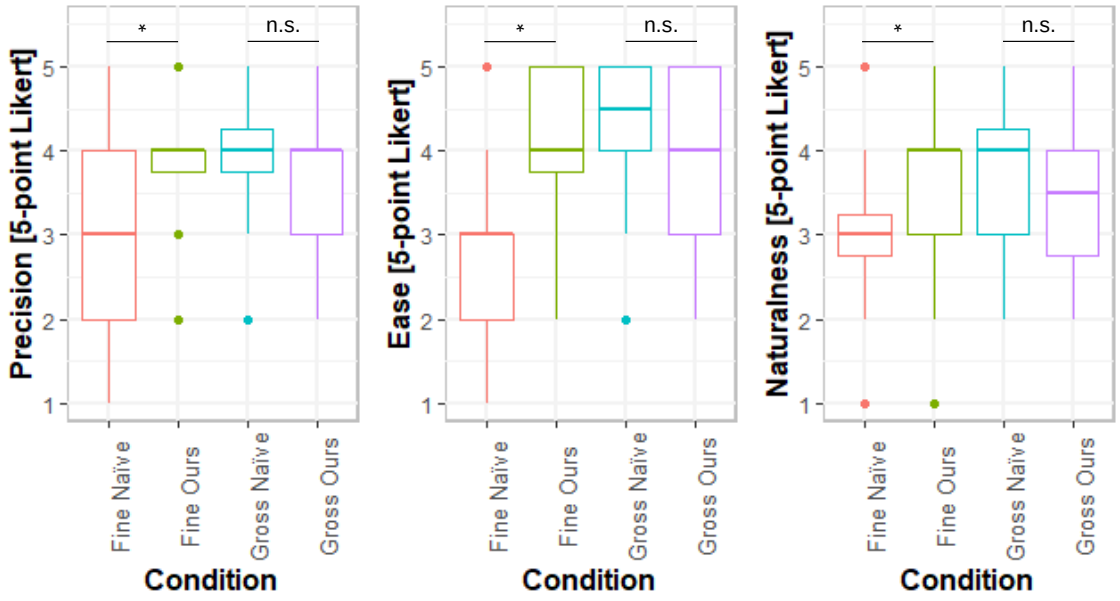


Figure 3.8: *Median (±Range) self-reported scores for Precision, Ease and Naturalness. Asterisks denote significant differences between means (* denotes $p < 0.05$); n.s. denotes no significant differences between means. Note that only the relevant comparisons are displayed in the figure; results from all comparisons are described in the main text.*

the Naïve Strategy (p=0.008) and with Our Strategy (p=0.016). Other comparisons between conditions were not significant.

In terms of naturalness, pairwise Wilcoxon signed rank tests comparing the four conditions against each other showed that participants reported higher naturalness of the Fine Manipulation with Our Strategy than with the Naïve Strategy (p=0.047), as shown in Figure 3.8-right. Further, for the Naïve Strategy, the Fine Manipulation felt significantly less natural than the Gross Manipulation (p=0.019), while this was not the case for Our Strategy, for which there were not significant differences between Manipulation conditions. Other comparisons between conditions were not significant either.

3.3.3 Discussion

The analysis of task performance and self-reporting of the user experiment suggests benefits of the proposed pose retargeting strategy. Moreover, these benefits appear independent of hand size, VR experience or age. First and foremost, Our Strategy exhibits significantly better performance than the Naïve Strategy for Fine Manipulation. In addition, the Naïve Strategy performs significantly worse on Fine Manipulation vs. Gross Manipulation, while Our Strategy does not exhibit a significant performance difference on these two Manipulation conditions.

Hand size has an effect on performance for Fine Manipulation regardless of the retargeting Strategy, but Our Strategy performs better than the Naïve Strategy

consistently across hand sizes.

The questionnaires suggest that Our pose retargeting Strategy feels significantly more precise, easier, and more natural for Fine Manipulation. For Gross Manipulation, Our Strategy scores slightly lower than the Naïve Strategy, but the difference is not significant. This is likely due to the inherent easiness of the Gross Manipulation scenario, which is confirmed when analyzing task performance across repetitions: Fine Manipulation benefits from learning more significantly than Gross Manipulation. Similarly, the analysis of task performance across repetitions indicates that the performance gain of Our Strategy is even larger initially, which again suggests that it is more natural, i.e., it requires less training.

3.4 Limitations and Future Work

In this chapter, we have proposed a method to retarget hand poses between hands with different size and skeletal morphology. The method serves for connecting off-the-shelf solutions for hand tracking and physics-based hand simulation, avoiding the need to share a common hand representation. The results of the user study indicate that our method is effective for fine manipulation, achieving performance and naturalness comparable to gross manipulation. From an applied point-of-view, our hand retargeting approach could accelerate the development of VR training applications requiring high dexterity and fine manipulation.

The key technical insight of the method is to formulate pose retargeting as the optimization of finger tip positions. This approach is motivated by maximizing the accuracy of pinch poses, which are key for fine manipulation. One interesting avenue of future work would be to explore more diverse feature vectors and/or objective functions, including other points in the hand, pose likelihood, etc. Similarly, the user study could be extended by covering more diverse interaction tasks.

Our hand retargeting method assumes that the shape of the simulated VR hand is given. This is the case when the VR application uses a hand of a fixed size, but one interesting extension would be to allow the simulation of personalized hands, which would require changing the shape of the simulated hand. Our current approach approximates this step by estimating a uniform scale, which could be extended to the estimation of, e.g., statistical shape parameters (Romero et al. 2017).

Currently, the retargeting method is applicable only to hands with similar skeletal topology, e.g, with five fingers. However, the approach could be extended to connect hands with very diverse skeletons, e.g., with a different number of fingers. The technical challenge is to define relevant feature metrics for such diverse hands.

Finally, in this chapter, we developed a method to improve the interaction between a hand and objects of various sizes. However, the method is limited to interactions with one hand at a time due to the challenges of simultaneously tracking two hands in close interactions, with or without objects. The upcoming chapter introduces a novel framework for generating synthetic data for scenarios involving two interacting hands, which has been crucial for developing a pioneering method

for tracking and reconstructing these interactions using only a single RGB camera.

4 Generating Annotated Data of Physically Accurate Two-Hand Interactions

4.1 Introduction

With the abundance of smart and mobile devices, interaction paradigms with computers are changing rapidly and moving farther away from the traditional desktop setting. With the recent progress on virtual and augmented reality (VR/AR), hand pose estimation has gained further attention as direct, natural, and immersive way to interact. Yet, tracking two interacting hands in close interactions while automatically adapting to the users' hand shape is still an ongoing challenge, and the task is further complicated when using a single RGB camera due to frequent occlusion, depth-scale ambiguity, and the self-similarity of hand parts.

To ease the problem, many previous works on 3D hand pose estimation use special depth cameras providing partial 3D information and they commonly rely on a learning component. Nevertheless, many of them focused on tracking a single isolated hand (Yuan et al. 2018), with only a few exceptions that are able to handle object interactions (Panteleris et al. 2015, Sridhar et al. 2016, Tzionas & Gall 2015) or interactions with a second hand (Mueller et al. 2019, Taylor et al. 2016, 2017) which we review in Chapter 2. In recent years, the research focus has shifted towards methods that use a single RGB camera since these sensors are ubiquitous (Cai et al. 2018, Mueller et al. 2018, Zimmermann et al. 2019). Despite significant progress, there are still very few methods explicitly designed to reconstruct close two-hand interactions from single RGB input. While hybrid and discriminative approaches have succeeded in other scenarios, they rely heavily on high-quality, diverse training data, which is difficult to acquire, especially for two interacting hands scenarios. Tasks such as segmentation and depth estimation are extremely difficult to label manually, and generating synthetic data is challenging because it requires accurate simulation of real-world conditions and interactions. As mentioned in Section 1.1, simulation systems often lack a parametrizable hand model in terms of shape and/or appearance, which restricts the diversity of the data, hence, limiting its effectiveness in training robust machine learning models. In this tesis, we address these limitations by proposing a system that simulates and generates physically accurate two-hand interactions with personalized hand shapes and diverse appreances.

This advancement has led to the development of the first method capable of reconstructing two interacting hands from monocular RGB images, using a hybrid

approach (Wang et al. 2020). Our framework leverages the output of a commercial depth-based hand tracker (*LeapMotion* 2016) to drive the two simulated hands, allowing us to generate a comprehensive set of essential data for training learning components. This data includes dense per-pixel surface matchings, hand segmentation, photorealist renderings, intra-hand and inter-hand distance maps, and 2D keypoints for both hands, while accounting for variations in hand identities, including different shapes and textures, and handling interactions between the two hands.

4.2 Method

Our main objective in the scenario of developing a tracker for two interacting hands, is to provide precise annotations, and photorealistic data that is practically impossible to obtain in real-life conditions for highly-challenging poses, in order to train robust models. To this end, and similar in spirit to Zhao et al. (2013) and Mueller et al. (2019), we present a motion capture-driven physics-based simulation to generate physically-correct hand sequences (*e.g.*, without self-collisions, with accurate inter-hand contact, and with a soft-skin layer) where two hands realistically interact in a large variety of poses. To increase the realism and diversity of simulated hand sequences, and in contrast to existing approaches that use a hand template of fixed shape and appearance in the simulation framework, as mentioned in Section 1.1, we extend the surface-based parametric model of MANO to a volumetric representation that is subsequently fed into the simulation (Verschoor et al. 2018) described in Section 2.2.2. This allows us to synthesize complex hand motions driven by a motion capture sequence, including 2D keypoint positions and heatmaps $\mathcal{H}^{\mathrm{GT}}$, dense correspondence images $\mathcal{C}^{\mathrm{GT}}$, segmentation masks $\mathcal{S}^{GT}$, relative depth maps $\mathcal{D}^{\mathrm{GT}}_{\mathrm{intra}}$, and relative inter-hand distance maps $\mathcal{D}^{\mathrm{GT}}_{\mathrm{inter}}$, with varying subject identities. We can therefore generate data with varying hand shapes.

Additionally, we further extended the MANO model with photorealistic appearances by a standard texture mapping approach. Hand textures were generated by reprojecting multi-camera imagery into a still hand image to which the MANO model was fitted (Qian et al. 2020). The ability to render physically plausible two-hand interactions for various hand shapes and appearances enables models trained with our data to generalize better to real world scene diversity.

The different ground truth annotations types generated by our framework are defined as follows:

4.2.1 Dense Matching

Each pixel $\gamma = (u, v) \in \Gamma$ from $\mathcal{C}^{GT}$ contains the 4-channel color value $\mathcal{M}(\gamma)$ that uniquely determines the surface point of the 3D hand model which is visible at this pixel. We call the mapping from the color vector to the 3D model surface *dense matching encoding*, depicted in Figure 4.2. Note that the dense matching encoding is the same for the left and right hand, where we make use of the segmentation mask $\mathcal{S}$ for disambiguation, shown in Figure 4.1 (second row). Hence, the first 3 channels

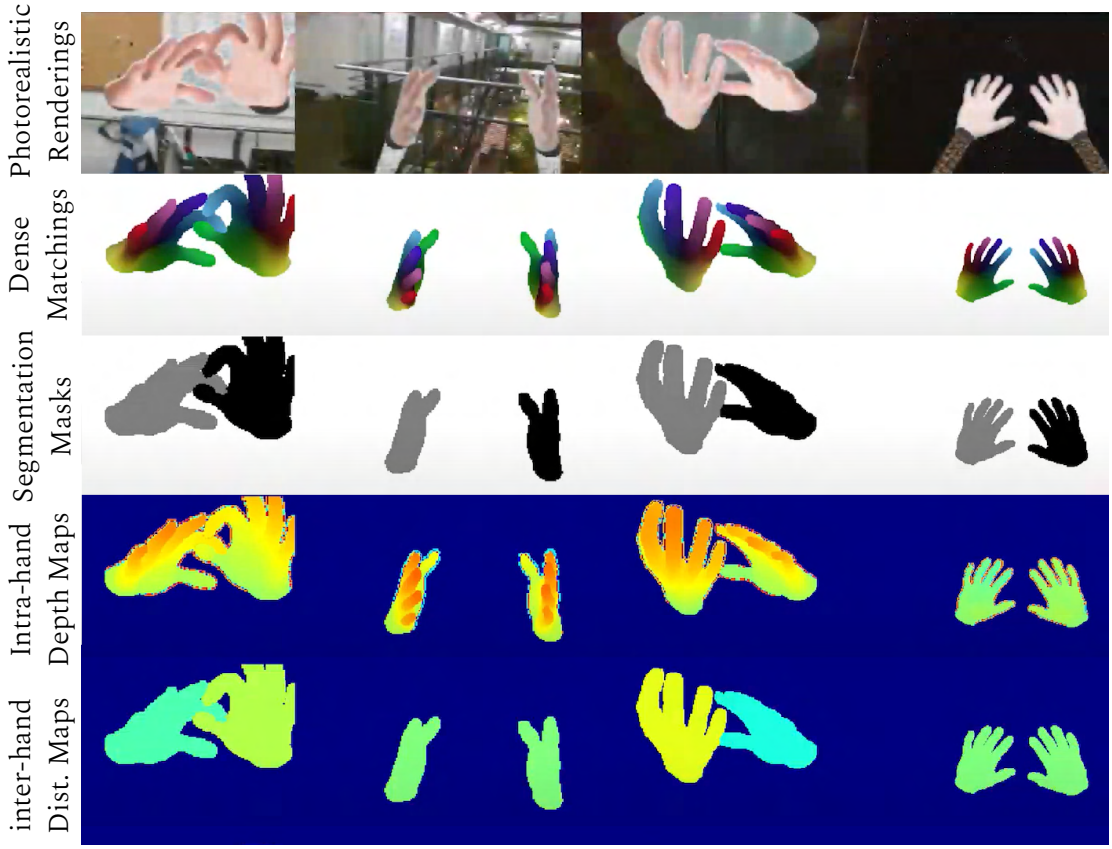


Figure 4.1: *Examples of images generated by our framework. From top to bottom: photorealistic renderings, dense matching maps, segmentation masks, intra-hand depth maps, inter-hand distance maps.*

correspond to the dense matching on the hand surface, and the last channel encodes the segmentation label, set as $\sigma(\text{LEFT}) = 0.0$ and $\sigma(\text{RIGHT}) = 0.5$.

We use the same encoding as Mueller et al. (2019) to embed the hand surface to a 3D feature space for our dense matching map. This is done using the method of Bronstein et al. (2006) to approximately preserve geodesic distances in the feature space. We then map the feature space to an HSV color space cylinder which results in each finger being assigned a different hue.

4.2.2 Segmentation

A specific value is assigned to each vertices from the hand model to label the pixels corresponding to the left or right hand, generating a segmentation mask $\mathcal{S}^{GT} \in \{\text{LEFT}, \text{RIGHT}, \text{BG}\}^{h \times w}$ that separates the hands from its surroundings/background, see Figure 4.1 (third row).

4.2.3 Intra-Hand Relative Depth

The framework further generates an intra-hand relative depth map $\mathcal{D}^{GT}_{\text{intra}} \in \mathbb{R}^{h \times w}$. For each hand pixel, it contains the estimated depth difference of this hand point

Figure 4.2: *Dense matching encoding of MANO model, front and back.*

to the root of the respective hand, shown in Figure 4.1 (fourth row). Note that $\mathcal{D}^{GT}_{\text{intra}}$ is scale-normalized due to the inherent ambiguity in RGB images.

4.2.4 Inter-Hand Distance

Our framework also provides the distance in depth between the two hands as a RGB image $\mathcal{D}^{GT}_{\text{inter}} \in \mathbb{R}^{h \times w}$. Every pixel in $\mathcal{D}^{GT}_{\text{inter}}$ that belongs to a hand contains the distance of its root joint from the other hand's root (in the case for only a single hand being visible, we assign a constant value to all pixels). Note that each pixel in the output can thus be seen as member of an ensemble, see Figure 4.1 (fifth row). Analogous to the intra-hand relative depth, we also normalize the inter-hand distance with the size of the hand. We summarize the ensemble with one relative distance value d_h per hand by calculating the median over all pixels that belong to the respective hand based on the segmentation mask $\mathcal{S}$, i.e.

$$d_h = \underset{\gamma \in \Gamma, \mathcal{S}(\gamma)=h}{\text{median}} \mathcal{D}^{GT}_{\text{inter}}(\gamma)\,. \tag{4.1}$$

We set the robust relative distance $d_{\text{inter}} = \text{mean}(d_{\text{left}}, -d_{\text{right}})$. When the two hands are close, d_{left} and d_{right} can be degenerate and have the same sign. In this case, d_{inter} is set to 0.

4.2.5 2D Keypoints

Let $\mathcal{J}_{\text{total}}$ be the set of all 12 keypoints, namely the fingertips and wrist of each of the two hands. Our framework yields the 2D keypoints position that are then translated into heatmaps $\mathcal{H}^{GT} \in \mathbb{R}^{h \times w \times |J_{\text{total}}|}$, a one-channel image for each of the keypoints. Each ground-truth heatmap contains a Gaussian with radius $0.07 \cdot r_c$, where r_c is the edge length of the larger edge of a tight hand crop, scaled to have maximum value 1, centered at the 2D keypoint position. Note that the ground truth is also provided for occluded keypoints, enabling their use even under strong occlusions, which are common in two-hand interactions.

The framework presented in this chapter was an essential component in developing a state-of-the-art method, which we will briefly summarize in the following sections for the completeness of this thesis. Additionally, we also show results on real-world captured data in Figure 4.6.

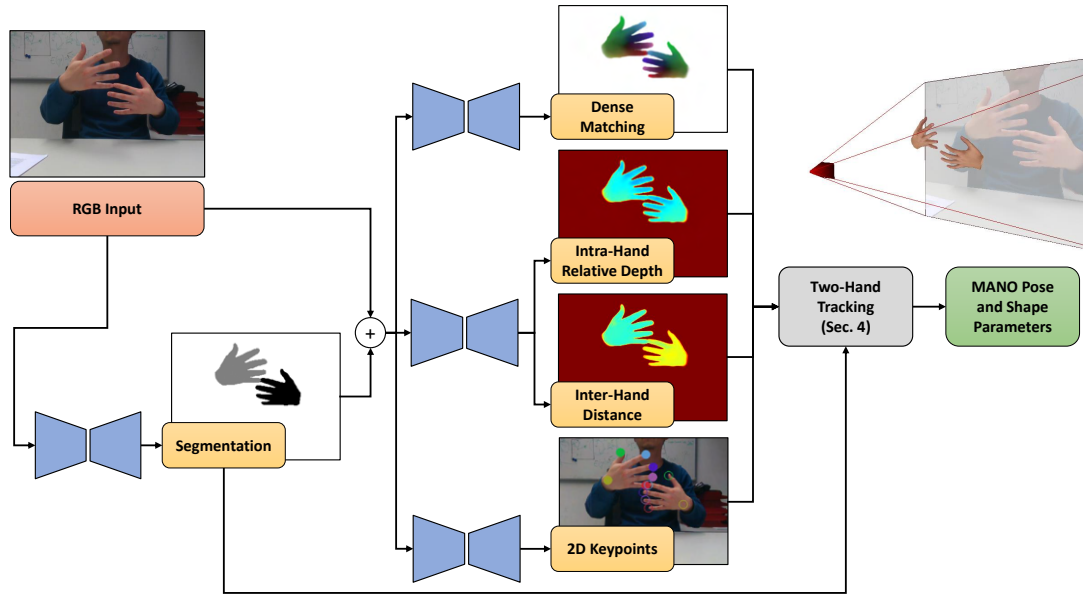


Figure 4.3: *Illustration of the RGB2Hands approach. The RGB input image is processed by neural predictors that estimate segmentation, dense matching, intra-hand relative depth, inter-hand distances, as well as 2D keypoints. This is then used within the two-hand tracking energy minimization framework. The output are pose and shape parameters of the 3D MANO model (Romero et al. 2017) of both hands, which directly give rise to a bimanual 3D reconstruction.*

4.3 Two-Hand Tracking Framework

4.3.1 Overview

Given a monocular RGB image that depicts a two-hand interaction scenario, the goal is to recover the global 3D pose and 3D surface geometry by fitting a parametric hand model to both hands in the input image. Such a model-fitting task requires information extracted from the input image to be used as a fitting target, which however represents a major challenge when using monocular RGB data only. Previous methods that rely on depth data (Mueller et al. 2019, Taylor et al. 2017) are implicitly provided with a much richer input (*i.e.*, global depth), which is the fundamental ingredient for an accurate 3D pose and shape fit. Per-pixel estimation of correct 3D hand depth from a single RGB image is very challenging.

Note that, in particular in the two-hand case, inconsistent depth estimates per hand would lead to incorrectly captured interactions in 3D. Thus, the method and the scene representation need to be able to handle these ambiguities well. Therefore, Wang et al. (2020) design an alternative representation of dense 3D geometry information, tailored for a two-hand scenario, which is amenable to be directly extracted from RGB images based on a machine learning pipeline summarized in Section 4.3.2. To this end, inter-hand distance and intra-hand depth maps are regressed, in combination with robust 2D keypoints. This design choice explicitly

provides sufficient information to resolve depth ambiguities in the model-fitting step. Furthermore, dense per-pixel surface matchings are also regressed to the parametric hand model directly from input images. This step is designed to be robust against the significant skin tone and illumination variability in RGB images.

The hand representation builds on the parametric surface hand model MANO proposed by Romero et al. (2017), which we summarize below and in Section 2.1.2. Subsequently, the model-based fitting framework will be outlined.

Parametric Pose and Shape Model

MANO was built from more than 1,000 scans of 30 subjects performing a large variety of poses, and consequently the model is capable of reproducing hand shape variability and surface deformations of articulated hands with high detail. Specifically, for a single hand, MANO outputs a set of 3D vertex positions $\mathcal{X}$ of an articulated 3D hand mesh, *i.e.*

$$\mathcal{X}(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \mathbf{W}), \tag{4.2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and $\boldsymbol{\theta} \in \mathbb{R}^{51}$ are the shape and pose parameters with the latter consisting of 45 articulation parameters and 6 global rotation and translation parameters. $T(\cdot)$ is a parametric hand template in rest pose with pose-dependent corrections to reduce skinning artifacts, $J(\cdot)$ computes the 3D position of the hand joints, and $\mathbf{W}$ is a matrix of rigging weights used by the skinning function W (based on linear blend skinning). See Romero et al. (2017) and Section 2.1.2 for further details.

As the method from Wang et al. (2020) targets a two-hand scenario, two sets of shape and pose parameters $(\boldsymbol{\beta}_h, \boldsymbol{\theta}_h), h \in \{\text{left}, \text{right}\}$ are used, for the left and right hand respectively. To simplify the notation, the parameters of both hands as $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\text{left}}, \boldsymbol{\beta}_{\text{right}}) \in \mathbb{R}^{20}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{left}}, \boldsymbol{\theta}_{\text{right}}) \in \mathbb{R}^{102}$ are stacked, and a unique set of vertices $\mathcal{X} = (\mathcal{X}_{\text{left}}, \mathcal{X}_{\text{right}})$ is defined, where the dependence of $\mathcal{X}$ on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ for brevity may be omitted.

Overview of Model-Based Fitting Formulation

In order to track two interacting hands in an image sequence the parametric MANO model is used within an energy minimization framework. To this end the fitting energy $f(\boldsymbol{\beta}, \boldsymbol{\theta})$ is introduced as

$$f(\boldsymbol{\beta}, \boldsymbol{\theta}) = \Phi(\boldsymbol{\beta}, \boldsymbol{\theta}) + \Omega(\boldsymbol{\beta}, \boldsymbol{\theta}), \tag{4.3}$$

where $\Phi(\cdot)$ is the image fitting term that accounts for fitting the model to the observed RGB image, and $\Omega(\cdot)$ is the regularizer that has the purpose of obtaining a plausible and well-behaved tracking result, for details, please refer to Wang et al. (2020). By minimizing the fitting energy f the pose and shape parameters $\boldsymbol{\theta} \in \mathbb{R}^{102}, \boldsymbol{\beta} \in \mathbb{R}^{20}$ (of both hands) are jointly estimated for each frame of the image sequence.

Image-fitting Term

Due to the 2D nature of RGB images and the so-resulting depth ambiguities, as well as the additional level of difficulty caused by interactions between the left and right hand, the novel image-fitting term Φ is designed carefully in order to allow for a reliable fit of the parametric hand model. In particular it uses specific information that the multi-task CNN (see Sec. 4.3.2) extracts from 2D images that enables the estimation of correct and coherent 3D pose of both hands in interaction, and minimizes the risk of implausible interaction capture due to ambiguous 3D pose estimates of each individual hand. Wang et al. (2020) propose to combine five components, where the following terms are used

1. the dense 2D fitting term Φ_{dense},
2. the silhouette term Φ_{sil},
3. the 2D keypoint term Φ_{key},
4. the intra-hand relative depth term Φ_{intra}, and
5. the inter-hand distance term Φ_{inter}.

An emphasis should be placed on the fact that existing methods that are capable of tracking *two hands in interaction* avoid 3D pose ambiguities by heavily relying on depth-based input data that is used in their image-fitting term, which, however severely simplifies the problem. In contrast, the proposed energy terms Φ_{dense}, Φ_{intra}, Φ_{inter} have the purpose of compensating the lack of available depth information and enable 3D consistent two-hand reconstructions by using a strong neural prior that extracts suitable information from RGB images only.

With that, the complete image fitting term that accounts for the model-to-image fitting reads

$$\Phi(\boldsymbol{\beta}, \boldsymbol{\theta}) = \Phi_{\text{dense}} + \Phi_{\text{sil}} + \Phi_{\text{key}} + \Phi_{\text{intra}} + \Phi_{\text{inter}} \,, \tag{4.4}$$

where the explicit dependence on $(\boldsymbol{\beta}, \boldsymbol{\theta})$ of the individual terms have been omitted for the sake of readability.

The camera intrinsics are assumed to be known and $\Pi : \mathbb{R}^3 \to \Gamma$ defines the projection from camera space onto the image plane. When this is not available, plausible intrinsics can be provided to obtain results accurate up to a scale.

One crucial part for defining the image fitting term is the *dense matching map* $\psi : \mathcal{X} \to \Gamma$, which predicts for each vertex $\boldsymbol{x} \in \mathcal{X}$ the corresponding pixel position $(u, v) \in \Gamma$ in the input image. The term ψ is assumed to be known, see Wang et al. (2020) for further details. In the following, when summing over vertices in the set $\mathcal{X}$, only those vertices that are visible are considered, where a vertex $\boldsymbol{x}$ is considered to be visible whenever $\psi(\boldsymbol{x}) \neq \varnothing$.

4.3.2 Dense Matching and Depth Regression

In order to obtain the predictions that were described in the previous section, including predictions for segmentation, dense matching, intra-hand depth, inter-

hand distance and 2D keypoints, the RGB input image are fed to a fully-convolutional neural network. This enables the method to work on entire images without requiring a potentially error-prone bounding box estimation for each hand. Since the network is trained using a large training corpus, it successfully learns priors to handle the inherent ambiguities in monocular RGB data. In the following, the network outputs are briefly described, see Wang et al. (2020) for more details on losses and architecture.

Network Outputs

The network architecture comprises two stages. In the first stage the network performs per-pixel segmentation into left hand, right hand, and background pixels. Then, the architecture branch into multiple subnetworks to regress dense matching, 2D keypoints, intra-hand relative depth, and inter-hand distance (the latter two using a shared multi-task subnetwork). The input for the second stage are both the original RGB input image, as well as the segmentation masks predicted in the first stage. Figure 4.4 shows all outputs predicted from test images.

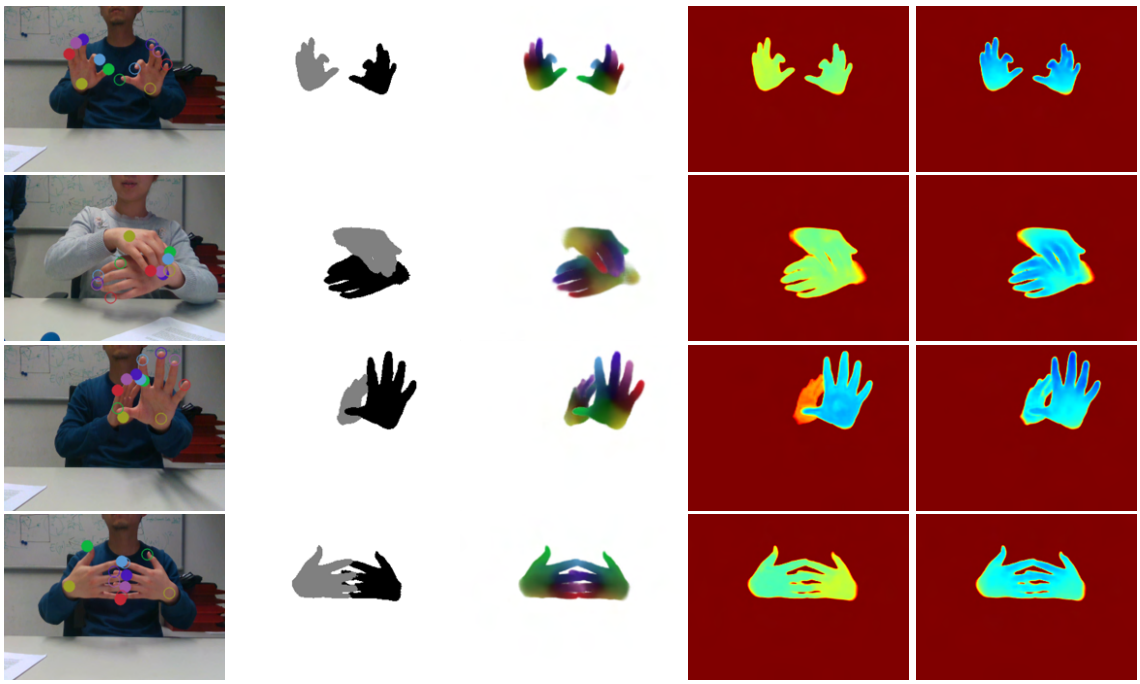


Figure 4.4: *Visualization of network outputs. From left to right: 2D keypoints, segmentation, dense matching map, inter-hand distance, intra-hand relative depth.*

4.3.3 Training Data

For training the regressor in a supervised manner, for a given RGB image containing two potentially interacting hands, a ground-truth relative depth map $\mathcal{D}^{\text{GT}}_{\text{intra}}$, the relative inter-hand distance map $\mathcal{D}^{\text{GT}}_{\text{inter}}$, a dense matching image $\mathcal{M}^{\text{GT}}$, and 2D joint position heatmaps $\mathcal{H}^{\text{GT}}$ are ideally required. Existing datasets like the *Rendered Hands Dataset (RHD)* Zimmermann & Brox (2017) or *Panoptic* Joo et al. (2017)

only provide a subset of the required annotations (see Table 4.1) and, in particular, do not have dense matching annotations. The former does also not show realistic and physically plausible close two-hand interactions, an important requirement for the intended setting. The recent FreiHand dataset (Zimmermann et al. 2019) provides crops of single hands with annotated MANO fits, sometimes even with objects, but no two-hand frames. Generating synthetic interacting hands images from these would require compositing and would lead to unrealistic interaction. Therefore, since manual annotation of the required labels is impossible, a new set of strategies is proposed to obtain annotations for both real and synthetic images. The existing datasets *RHD* and *Panoptic* have been added to the proposed real and synthetic datasets to increase data diversity and hence improve generalization. Table 4.1 presents a summary of the different datasets used for training, and gives details about the ground-truth annotations available in each of them. The procedure for creating the synthetic dataset has been presented in the Section 4.2 of this same chapter, being a contribution of this thesis. See Wang et al. (2020) for further details on the strategy for the real-world dataset. Furthermore, in Section 4.3.4, an ablation study, which systematically evaluates the impact of specific components by removing or modifying them, is provided to demonstrate how the real data (with slightly noisy annotations) helps bridge the real-synthetic domain gap, and how the perfectly annotated synthetic data from our framework mitigates the influence of noise.

	Segmentation	Dense Corrs.	Intra-Hand	Inter-Hand	2D Keypoints
Synthetic Data	✓	✓	✓	✓	✓
Real Data	✓	✓	✓	✓	✓
RHD	✓	✗	✓	✓	✓
Panoptic	✗	✗	✗	✗	✓

Table 4.1: *Available annotations in existing hand tracking datasets and the proposed datasets from Wang et al. (2020).*

4.3.4 Experiments

The proposed RGB two-hand tracking approach is experimentally evaluated in order to demonstrate its merits. First, the dataset and metrics used in the evaluation is introduced. Subsequently, an ablation study is conducted that evidences the importance of the individual components. Afterwards, the proposed method is compared quantitatively and qualitatively to other related works. Moreover, additional qualitative two-hand tracking results are presented. In this chapter we will only review the ablation experiment that evaluates the effect of the importance of

using the real and the synthetic dataset to demonstrate the effectiveness of our data generation framework presented in the previous section, while the rest of the evaluation results and the details can be found in Wang et al. (2020).

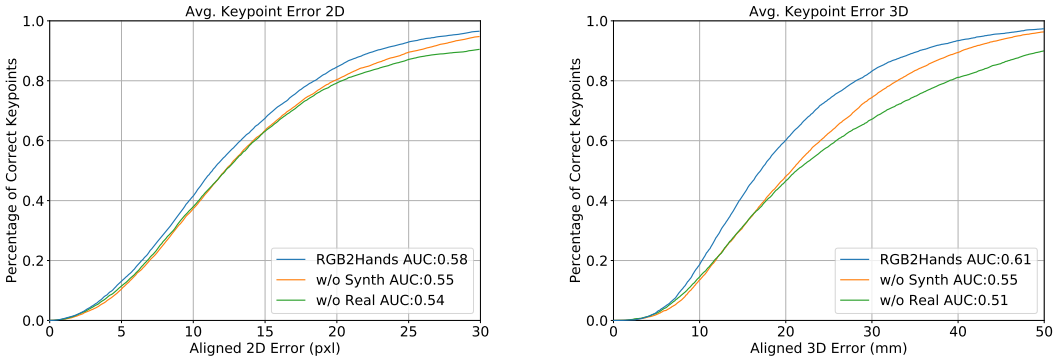


Figure 4.5: *Training data ablation study on the RGB2HANDS dataset.*

Ablation Study

The behavior of the hand tracker is analyzed when training the prediction networks either without the real dataset, or without the synthetic dataset provided by our framework, respectively, see Figure 4.5. When not using the real dataset, or when omitting the synthetic dataset, the PCK curves drop substantially (see green and orange lines in Figure 4.5), compared to using both datasets (blue line).

4.4 Discussion & Future Work

Overall we have presented a framework that simulates and generates data for physically accurate two-hand interactions. This framework has led to the development of a compelling 3D two-hand tracking and reconstruction method, that produces promising results, even on challenging sequences of two interacting hands. Despite the overall good performance of the proposed method, particularly for close hand-hand interaction settings, there are also some downsides that need to be addressed in the future. Currently, the method may not always be able to correctly track very fast hand motions, since in this case motion blur may lead to unreliable predictions of the neural network. One way to address this is to generate additional training data with motion blur using our framework, so that the neural network is able to handle such challenging cases more effectively.

Moreover, it is difficult to find a good trade-off between the MANO pose prior and the other energy terms, so that one has to sacrifice either pose variability or pose plausibility. This is most noticeable for thumb articulations (Figure 4.7, left). This could for example be addressed by equipping the MANO model with a kinematic skeleton, and then enforcing explicit joint limit constraints while still using the pose space to capture correlations in joint articulations. Due to inherent depth ambiguity, our method may also have difficulties reconstructing interactions where high precision in relative hand positioning is required; e.g. slotting a ring onto a

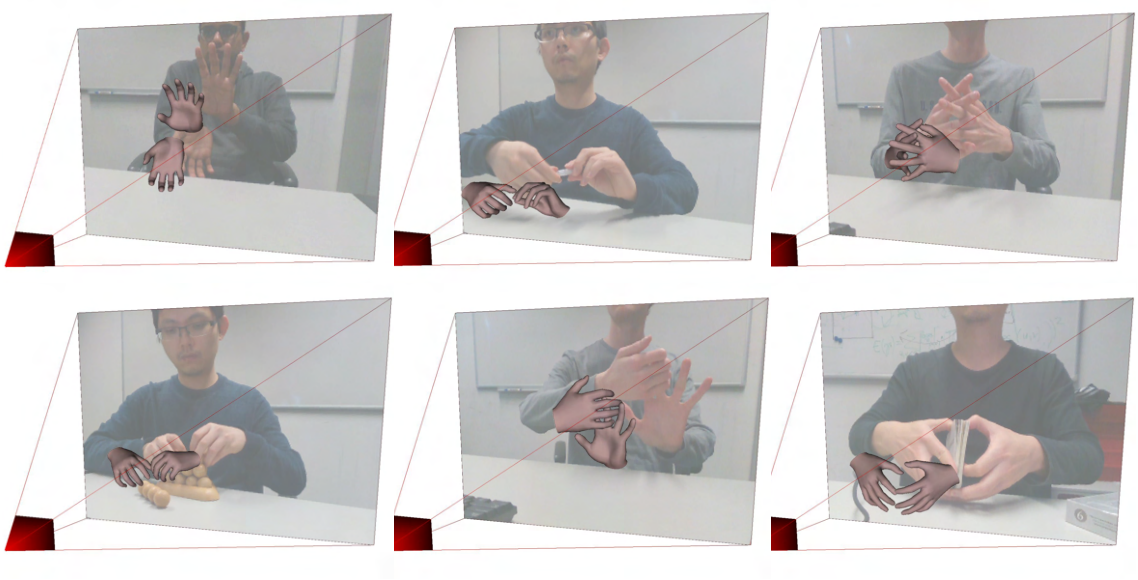


Figure 4.6: *Results of our RGB2Hands method.*

finger (see Figure 4.7, right). For such tasks introducing a new energy term or additional cues from a depth sensor or a stereo camera might be required. It would be interesting as well to explore the explicit use of the temporal dimension, so that for example hand shape information can be integrated over time, in a similar spirit to bundle adjustment in multi-view reconstruction. Moreover, temporal neural network architectures can be used to obtain temporally smoother predictions and thus further improve temporal tracking consistency.

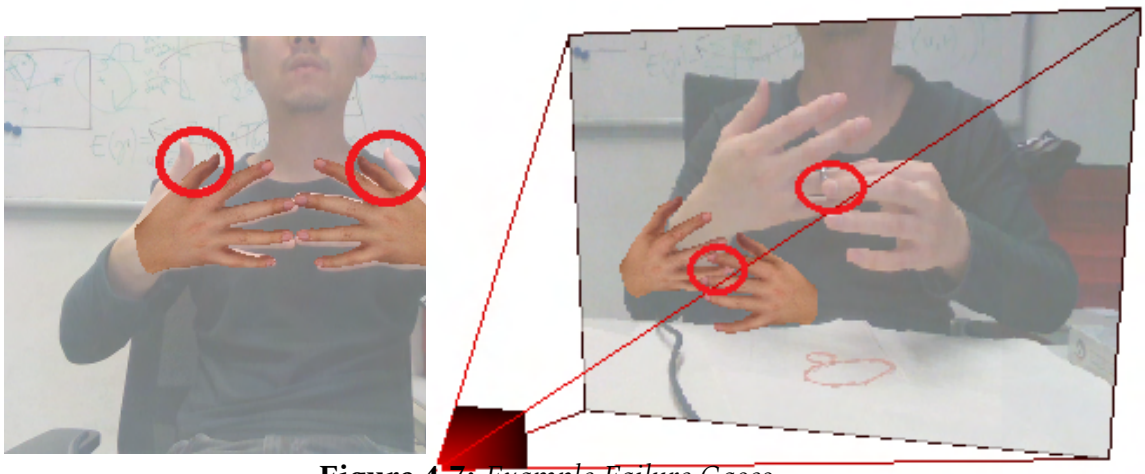


Figure 4.7: *Example Failure Cases*

4.5 Conclusion

In this chapter we have presented a framework to generate physically accurate two-hand interactions data that takes hand variability in terms of shape and appearance into account. This contribution enables the creation of a new synthetic dataset, which is combined with a new real dataset for which annotations were ob-

tained based on RGB-D frames. The real data, despite having slightly noisy annotations, helps bridge the real-synthetic domain gap, while the perfectly annotated synthetic data, obtained from our framework, mitigates the influence of the noise. This combined dataset was used to train a multi-task neural network predictor, one of the two key components of the first approach specifically tailored for tracking and reconstructing two hands in interaction in global 3D using only RGB images (Wang et al. 2020). The second crucial component is a parametric 3D hand model. Together, these two strong priors tackles the major challenge of depth ambiguities inherent in this context.

The proposed approach is shown to outperform previous RGB-only methods in complex two-hand interaction scenarios, both quantitatively and qualitatively, and even achieves comparable performance to a state-of-the-art depth-based real-time approach. However, despite the encouraging results achieved in two-hand tracking scenarios, residual errors in depth, shape, or hand pose estimation remain. These inaccuracies prevent the precise detection of hand-to-hand contacts, which in turn limits the ability to fully and intuitively interact with the surrounding. In the following chapter, we will address this persistent challenge by introducing an image-based approach designed to estimate hand-to-hand contacts from single RGB images, that could potentially be used as a new energy term. This method is based on a similar pipeline that generates physically accurate contact data, which will be used to train our learning-based method.

5 Hand-to-hand Contact from RGB Images

5.1 Introduction

One of the main motivations for estimating the 3D pose and shape of hands captured from the real world is to provide a natural interface to any human-computer interaction (HCI) problem, where a user wants to use his or her hands to interact with a virtual environment. However, hand tracking is a highly challenging task due to the natural self-occlusions, the similar appearance between the two hands, and the many degrees of freedom that we need to estimate in order to recover the hand motion.

To tackle the grand challenge of tracking hands, many methods have been proposed (summarized and discussed in Section 2.3). To simplify the problem, most of them focus on single-hand pose estimation, which has been successfully addressed using multi-camera setups (Ballan et al. 2012, Sridhar et al. 2013) or depth cameras (Malik et al. 2020, Sridhar et al. 2015) to further simplify the task, although recent methods from RGB-only have also shown very accurate and robust results at tracking single hands (Mueller et al. 2018, Zimmermann & Brox 2017, Zimmermann et al. 2019). However, we argue that the use of hands as a natural HCI interface requires methods that are capable of tracking the *two hands in interaction* from a single RGB camera. Unfortunately, very few works exist that tackle the two-hand tracking scenario (Li et al. 2022, Moon et al. 2020), one of them being described in Chapter 4 (Wang et al. 2020) and, despite the promising results, they all share a common lim-

Figure 5.1: *Given a single RGB image of two hands in interaction, our model infers camera space 2D contact maps (in orange) that encode the pixels where the two hands are in contact. We demonstrate that our contact maps can be plugged into 3D hand tracking frameworks to improve accuracy.*

itation, referred in the previous conclusion in Section 4.4: residual errors in depth, shape, or hand pose estimation prevent the accurate detection of hand-to-hand *contacts*.

To circumvent these shortcomings in this chapter, we propose an image-based method to estimate hand-to-hand contacts from a single RGB image. Our method builds on top of existing two-hand tracking solutions, enriching them with a camera-space probability map of hand contact. We argue that framing the hand contact problem on top of the 3D hand tracking problem provides many advantages: first, it enables the detection of contact even when 3D tracking is inaccurate, for example, due to errors in global pose estimation which produce 3D hands that do not touch in hand interaction motions; second, it makes our solution compatible with any existing two-hand tracking solution, for example, it can be plugged into either depth-based (Mueller et al. 2019) and RGB-based (Wang et al. 2020) methods; and third, it can potentially be used as a new term in optimization-based methods for two-hand tracking, similar to other object-human contact terms that enforce soft constraints in the tracker (Sridhar et al. 2016).

We formulate our method as an image-to-image translation problem. Assuming a single RGB image as input, we first estimate pixel-to-surface correspondences (Alp Güler et al. 2018) through a state-of-the-art 3D hand tracking method (Li et al. 2022). We then use a UNet architecture to translate the pixel-to-surface image into a probability contact map image. Since our network takes as input a pixel-to-surface encoding, it is agnostic to scene lighting conditions, shadows, and camera sensor modalities, which typically hinder the generalization of RGB methods.

To train our method, we propose a new pipeline to automatically annotate dense surface contacts in hand interaction sequences. To this end, we first use a commercial depth-based hand tracker to capture a collection of hand interaction sequences. We then fit a parametric surface hand model (Romero et al. 2017), and feed the resulting hand mesh sequences into a physics-based simulator (Verschoor et al. 2018), described in Section 2.2.2. This allows us to automatically detect and annotate per-vertex collisions, which we use to render ground truth contact maps for a variety of viewpoints. Additionally, for each frame, we also render ground truth pixel-to-surface correspondences automatically provided through the UV parameterization of the hand mesh. Finally, using a large dataset of pairs of pixel-to-surface images and contact maps, we train our UNet network.

Through an exhaustive evaluation, we qualitatively and quantitatively demonstrate that our method is able to predict hand-to-hand contact labels with an accuracy higher than existing methods. In summary, our main contribution is twofold:

- To the best of our knowledge, the first method to explicitly learn to detect dense hand contact from RGB images, which can be plugged into any two-hand tracking framework.
- A pipeline to annotate dense per-vertex surface contacts in real-world sequences of two-hand interactions.

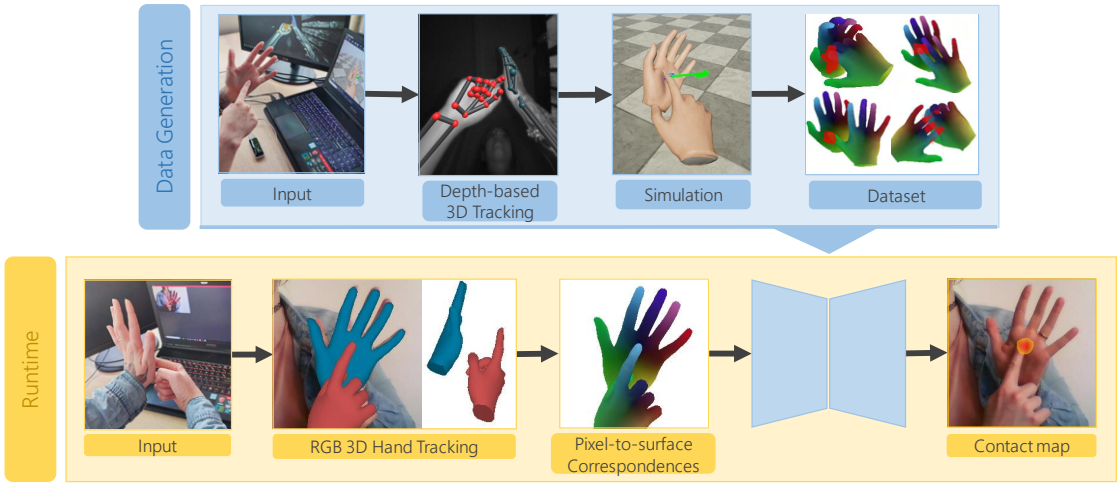


Figure 5.2: *Overview of our method to detect hand-to-hand interaction from RGB images. We first propose a pipeline to automatically annotate hand-to-hand contacts in sequences captured with a depth camera, which we use to create a dataset of pairs of surface-to-pixel correspondences and their contact map (top). At run time, we predict the 3D pose of each hand using a state-of-the-art method (Li et al. 2022), but this often fails in capturing a global pose, which prevents hand contact estimation. We render the pixel-to-surface correspondences of the tracked hands and use our network to infer the accurate contact map (bottom).*

5.2 Method

Our primary goal is to predict dense contact labels to augment the information inferred by a 3D hand pose estimation method when reconstructing two interacting hands from in-the-wild RGB images. To this end, we propose to use a strategy based on a popular image-to-image translation network (Ronneberger et al. 2015). Importantly, instead of working directly on RGB images, we first convert them into pixel-to-surface correspondences. This helps to circumvent many challenges related to in-the-wild imagery (e.g., lighting, shadows, background, etc.). Therefore, our system learns to convert estimated pixel-to-surface correspondences into contact probability maps.

Figure 5.2 presents the different steps of our approach, which we describe in detail in the rest of this section. We first construct a dataset (Section 5.2.1) by simulating a wide variety of interaction sequences using a state-of-the-art hand simulator (Verschoor et al. 2018) in a similar way to the work presented in the preceding chapter, and subsequently train our model (Section 5.2.2). At inference time (Section 5.2.3), we use a state-of-the-art method (Li et al. 2022) to reconstruct the pose and shape of both hands but, as will be shown later, it sometimes fails at providing 3D hand poses that are in contact due to the depth ambiguity. We therefore render the hand poses using their color-coded dense correspondences and feed them into the network to infer the corresponding contact probability maps. Finally, we demonstrate that our contact maps enable us to refine the estimated hand poses by

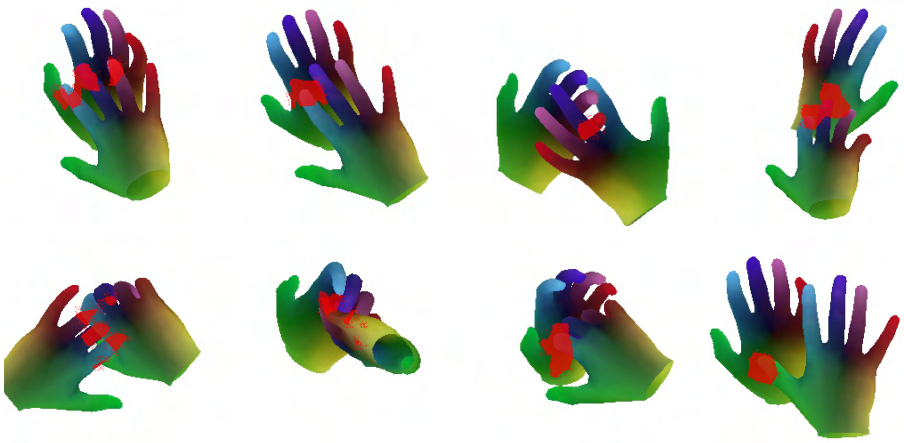


Figure 5.3: *Sample frames from the simulated sequences of our train set, with overlay ground truth contact map in red. Our dataset includes a wide variety of hand poses and interactions.*

state-of-the-art hand tracking (Li et al. 2022) to improve the 3D contact accuracy (Section 5.2.4).

5.2.1 Dataset Generation

We require a dataset containing several sequences with two interacting hands and their corresponding contact labels. To this end, we first capture hand motion data using a commercial depth sensor (*LeapMotion* 2016) and then fit the MANO (Romero et al. 2017) parametric surface model to the captured motions to generate sequences of hand meshes in interaction. Following, we feed these meshes into the physics-based hand simulator CLAP (Verschoor et al. 2018), which was extended by Mueller et al. (2019) to handle pairs of hands. This finally allows us to compute per-vertex contact labels.

We captured 26 sequences to which we fitted the 3D hand models and performed physics-based simulations. Subsequently, each sequence was rendered at 30 frames per second (FPS), resulting in between 1800 and 3600 frames per sequence. This amounted to a total of 3.8×10^5 frames, all rendered at a resolution of 256 × 256 pixels. Additionally, at train time, we augment the data by generating 10 random rotations ranging between -100 and 100 degrees.

For each frame, we generated two kinds of information: a pixel-to-surface map, which will be used as input to the network; and a binary mask with ones at the pixels where we have detected contact, which will be used as the prediction target. We use the pixel-to-surface map $\mathcal{N} : \Omega \rightarrow [0, 1]^4$ from Mueller et al. (2019) also used in Chapter 4, that assigns each pixel in the image domain Ω a color $[0, 1]^4$ where the first 3 channels encode the dense correspondence to the hand surface and the last channel serves as a segmentation mask where 0 indicates the right hand, 0.5 the left hand and 1 indicates that the pixel does not correspond to any hand. Figure 5.3 depicts some samples of our training set, overlaying the contact map labels on

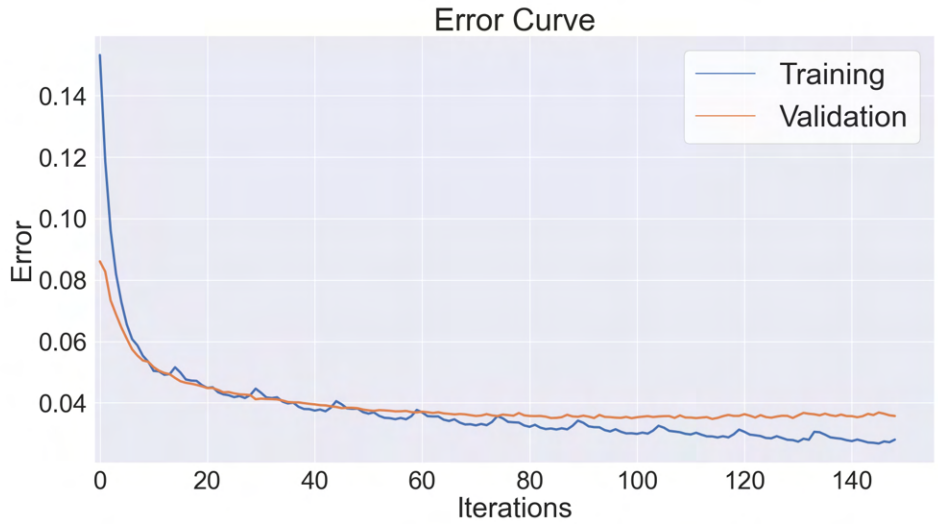


Figure 5.4: *Error values for our train and validation sets during training. An iteration in the horizontal axis corresponds to 250 batches of 8 images.*

top of the pixel-to-surface image.

5.2.2 Network Details

We learn to generate a binary mask that highlights regions with inter-hand contact in image space. For this task, we use an adapted version of the image-to-image translation network UNet (Ronneberger et al. 2015), an architecture consisting of an encode-decoder scheme with skip connections. We take the Pytorch implementation by Buda et al. (2019) as a reference. This uses 32 initial feature channels, which are doubled at each of the four levels on the multi-resolution pyramid. Each convolutional block consists of two 3×3 convolutions+activation followed by a max-pooling (left side of the pyramid) or upsampling (right side of the pyramid) units. We also use LeakyReLUs with a negative slope set of 0.1 and upsampling deconvolutions with bilinear upsampling units. Our network takes a 4-channel image as input, consisting of the encoded hand dense-correspondences (*i.e.*, the pixel-to-surface encoding) and the hand segmentation mask, and predicts a contact probability map. To train our network, we use the standard binary cross-entropy loss to compare against the ground-truth contact maps together with an Adam optimizer, with a learning rate 10^{-4} and a batch size of 8. Figure 5.4 depicts the loss curve while training, for train and validation sets.

5.2.3 Inference

At inference time, we can use any commercial RGB camera to capture sequences of an individual performing interactions with his or her hands. To provide an initial estimation of the hand tracking, we can use any state-of-the-art method for simultaneous two-hand tracking on the frames from the input sequences (e.g. Li et al. (2022)), and then convert the tracked sequences to MANO (Romero et al. 2017) representation. We render the corresponding hand meshes, from the same cam-

era viewpoint, using a texture that encodes the surface-to-pixel color-coded correspondences. The rendered images are fed into the trained network which yields the desired contact maps.

5.2.4 Application: Global 3D Hand Pose Optimization

Our inferred contact maps open the door to many downstream tasks in the area of hand tracking and human-computer interaction, where detecting accurate contact is crucial. To showcase a potential approach, we have used inferred maps to improve the accuracy of state-of-the-art 3D hand tracking methods with respect to 3D contact detection. To this end, starting from the 3D hand poses estimated with Li et al. (2022), we optimize the global 3D position (*e.g.*, 3 DOFs) of the hand closest to the camera such that the vertices within the areas where contact is detected (in camera space) are in contact in 3D space. This effectively refines the hand positions estimated with state-of-the-art methods, estimating 3D hand poses that are in actual contact.

Results of this downstream task are showcased Figure 5.8 (right), where we demonstrate the effectiveness of our method for optimizing the global 3D pose of hands tracked in real-world sequences.

5.3 Evaluation and Results

We qualitatively and quantitatively evaluate our method in both synthetic scenes and real scenes. For each scene, we compare the detected two-hands contact using our method and the state-of-the-art method of Li et al. (2022). We also implement a straightforward RGB-only baseline consisting of a UNet network that directly predicts camera space contacts from RGB (*i.e.*, without using pixel-to-surface estimation). As a contact metric, we use True Positive Rate (TPR), True Negative Rate (TNR), and ROC curve. A pixel is labeled as a contact if the predicted probability is higher than a threshold τ_m. We consider a true positive a pixel that is labeled as contact and is predicted as contact; a true negative a pixel that is predicted as no contact and is labeled as no contact; a false positive a pixel that is predicted as contact and is labeled as non-contact; and a false negative a pixel that is predicted as non-contact but it is labeled as a contact.

5.3.1 Evaluation on Synthetic Data.

To evaluate our method on synthetic data, we need a collection of hand interaction sequences with ground truth contact annotations and photorealistic 3D renders. To this end, we first capture data using Leap Motion (*LeapMotion* 2016) hand tracker and then convert the tracked skeletons into MANO (Romero et al. 2017) mesh representation. We then feed hand meshes into a physics-based simulator (Verschoor et al. 2018) that automatically computes ground truth contact maps. Finally, we also render the hand meshes using a photorealistic texture and background images. Figure 5.5 (left) presents a few frames of a synthetic test sequence, with ground truth contact labels overlayed on top. To test our method with these

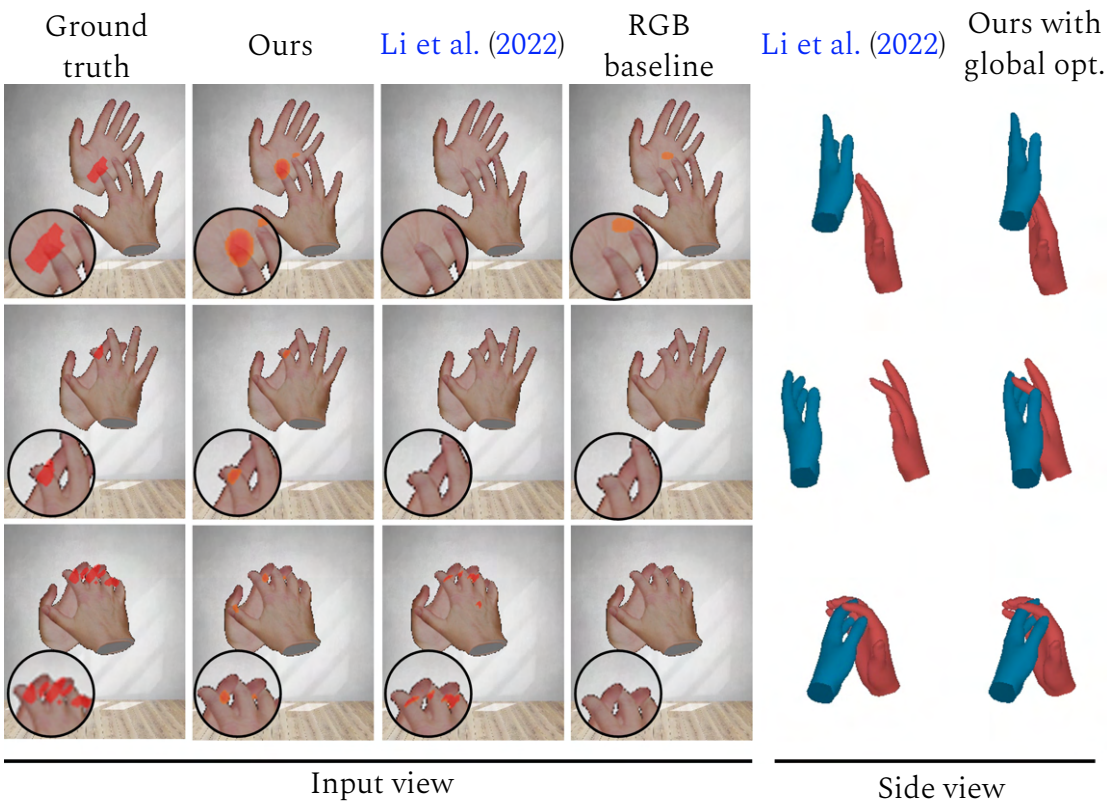


Figure 5.5: *Comparison to Li et al. (2022) and the RGB baseline on synthetic data, each frame with a zoom-in inset to contact area. Global 3D errors in Li et al. (2022) (5th column, notice colliding or too separated hands) prevent the detection of contacts using the method of Li et al. (2022) (3rd column). In contrast, our method (2nd column) closely matches the ground truth contacts (1st column, automatically annotated using simulation (Verschoor et al. 2018)), which can be used to optimize the global 3D position (6th column). Notice all colormaps were generated for values larger than 0.5 (for the RGB Baseline most values were under this threshold)*

synthetic sequences, we first estimate the 3D hand pose of the two hands (Li et al. 2022) and render the reconstructed meshes using a pixel-to-surface texture map. We then feed the renders into our network to estimate contacts. Figure 5.5 shows representative frames of our results and compares them with ground truth contacts, the RGB-only baseline, and contacts computed using the output of the state-of-the-art 3D hand tracking method by Li et al. (2022). As it is obvious from the side view renders, the method of Li *et al.* can fail at providing a reliable 3D global pose, which precludes the computation of hand contact based on the regressed 3D information. Also, as can be seen in the figure, the RGB-only baseline mostly predicts noisy or weak contact signals. In contrast, our image-based method is capable of estimating dense contact maps that match the input image. Finally, the right column shows the result of optimizing the global hand pose using the inferred contact maps, which effectively computes 3D hand poses that are in actual contact.

Figure 5.6: *To evaluate our approach, we show qualitative results on our test set. For this experiment, we use the ground truth pixel-to-surface image directly as input to our method (i.e., no hand tracking). This allows us to disentangle residual errors due to tracking issues. Results demonstrate that our predicted contact maps (bootom) closely match the ground truth (top).*

We also conducted a quantitative evaluation of our method assuming ground truth 3D hand poses. This is an important test that allows us to disentangle the errors of our approach from the errors of the 3D hand-tracking method that we build upon. To this end, we render ground truth pixel-to-surface sequences of simulated hand interacting sequences, and the corresponding segmentation map, and pass it to our network. We then compute the F1 score of the output contact maps and compare it to the ground truth labels. Figure 5.7 depicts this evaluation over 600 frames of the test sequence, demonstrating consistently high F1 values that highlight the precision and robustness of our method.

Finally, in Figure 5.6 we show representative *test* frames of our dataset. This demonstrates that our approach generalizes to unseen hand poses at test time.

5.3.2 Evaluation on Real Data.

To showcase that our approach generalizes to unseen poses and natural images with uncontrolled lighting captured with various sensors, we quantitatively evaluate our results on real-world data captured from a desktop webcam and a mobile phone camera. To this end, we record hand sequences of interacting hands and manually annotate ground-truth contact maps using an interactive tool where users paint contact on top of each frame. In total, we annotate 6,540 frames captured with two different RGB sensors. Importantly, notice that none of the manual annotations were used to train our model, they are only used for evaluation purposes, since our method is solely trained on synthetic contact maps that are computed automatically using our physics-based hand simulator driven by motion captured data.

Figure 5.8 presents a qualitative evaluation of representative frames of our method

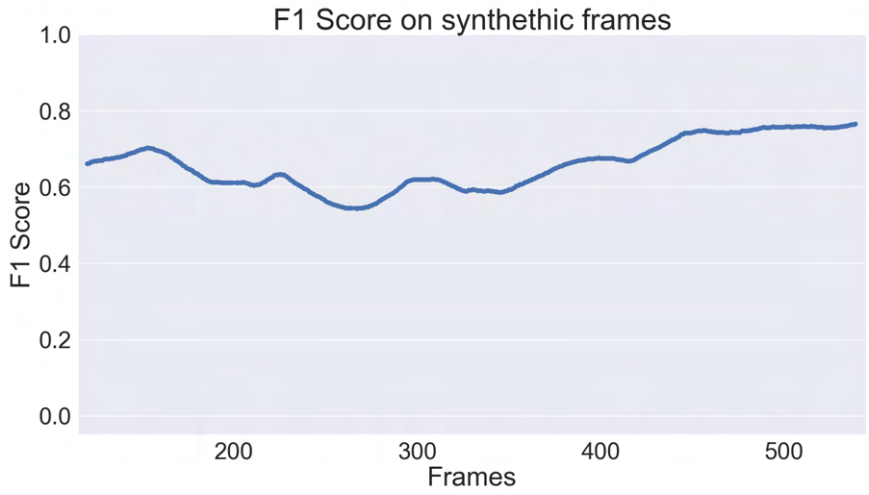


Figure 5.7: *Quantitative analysis of the predicted contact maps of our method on synthetic data.*

	TPR ↑	TNR ↑	AUC ROC ↑	RMSE ↓
RGB baseline	1 %	99 %	0.84	0.028
Ours	**56** %	**99** %	**0.96**	**0.026**
Ours w/o seg.	38 %	98 %	0.92	0.036
Li et al. (2022)	8 %	99 %	0.55	0.033

Table 5.1: *Quantitative evaluation of the sequence shown in Figure 5.10 (bottom).*

in a real-world test sequence. Results demonstrate that our predictions closely match the ground truth annotations, while the state-of-the-art method of Li et al. (2022) cannot infer contact due to errors in global pose estimation. To highlight such errors, we show side views of the Li et al. (2022) predictions, where it is very noticeable that hands are not in contact. In contrast, using our predicted camera-space contact maps enables the optimization of the global pose of the hand, yielding a correct 3D global pose estimation with accurate contact (right column). Additionally, for completeness, we show contact maps inferred with an RGB-only baseline (i.e., not using dense pose estimations as input to the network), which fails to predict accurate labels in most of the cases.

To evaluate our approach under a different lighting condition, we captured and annotated two additional real-world sequences, shown in Figure 5.10, and report quantitative evaluations in Table 5.1 and Figure 5.9. Our approach consistently overcomes the competing methods, yielding a True Positive Rate (TPR) 56% for our method, 8% for the method of Li et al. (2022), and 1% with the RGB-only baseline, using a threshold of 0.40. Additionally, the ROC score for our method is 0.96, 0.55 for the method of Li et al. (2022) and 0.84 for the RGB-only baseline, which clearly highlights the superiority of our approach. Note that, for fairness in these quantitative analyses, we use thresholds different from 0.5, as otherwise, there were

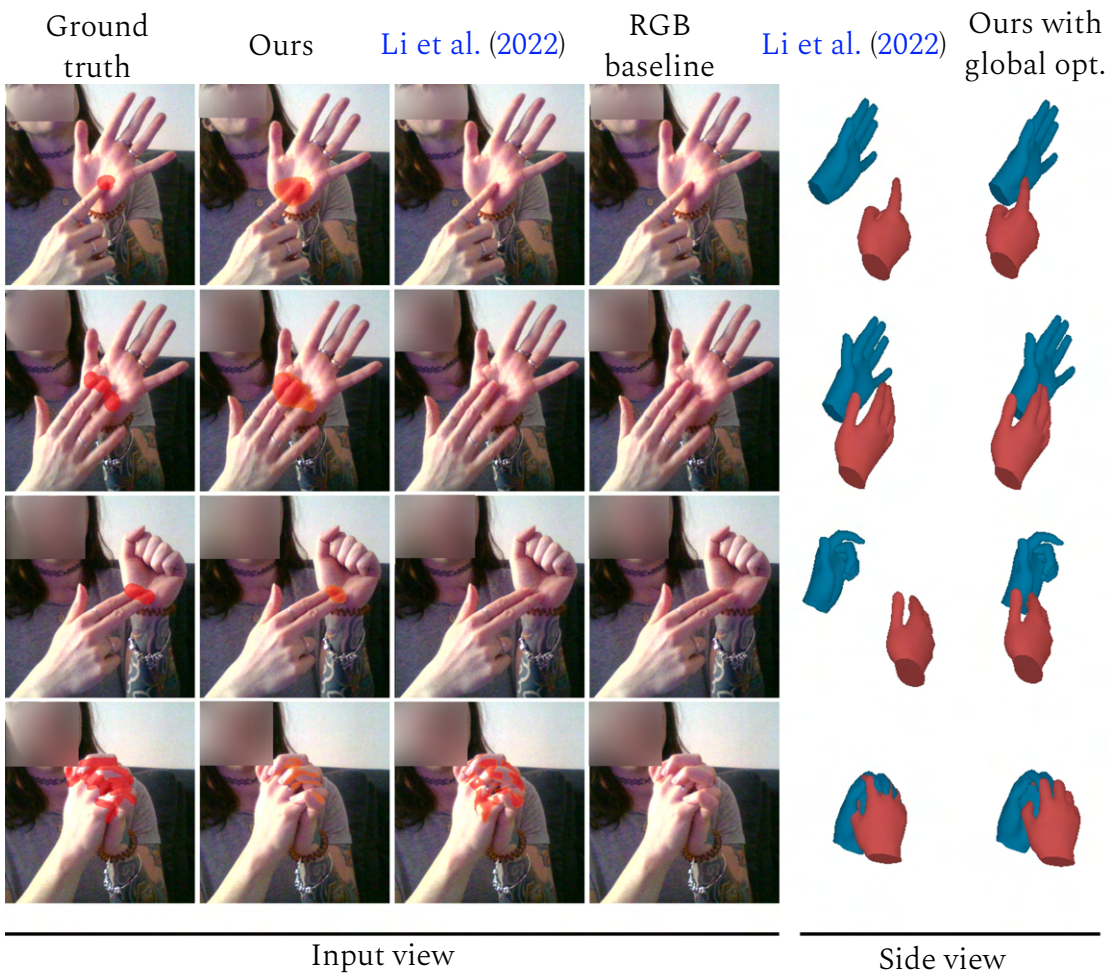


Figure 5.8: *Comparison to Li et al. (2022) on real-world data captured from a webcam. Errors in global 3D hand pose estimation (5th column, notice colliding or too separated hands) prevent the detection of contacts using the method of Li et al. (2022) (3rd column). In contrast, our method (2nd column) closely matches the ground truth contacts (1st column, manually annotated), enabling the accurate estimation of 3D contacts by optimizing the global pose of the hand (6th column).*

no predicted positive contacts for the RGB baseline. Thus, for each experiment, we selected the threshold closest to 0.5 that would produce some contact using the RGB baseline method. These results demonstrate that, regardless of the scene conditions, the contact accuracy of our method consistently outperforms the contacts detected using the 3D hand tracking method of Li et al. (2022) and the RGB-only baseline.

5.3.3 Qualitative Results.

For completeness, in Figures 5.1 and 5.10 we also show qualitative results of our method for a variety of subjects and sensors.

Overall, our results demonstrate the success of our method in the challenging

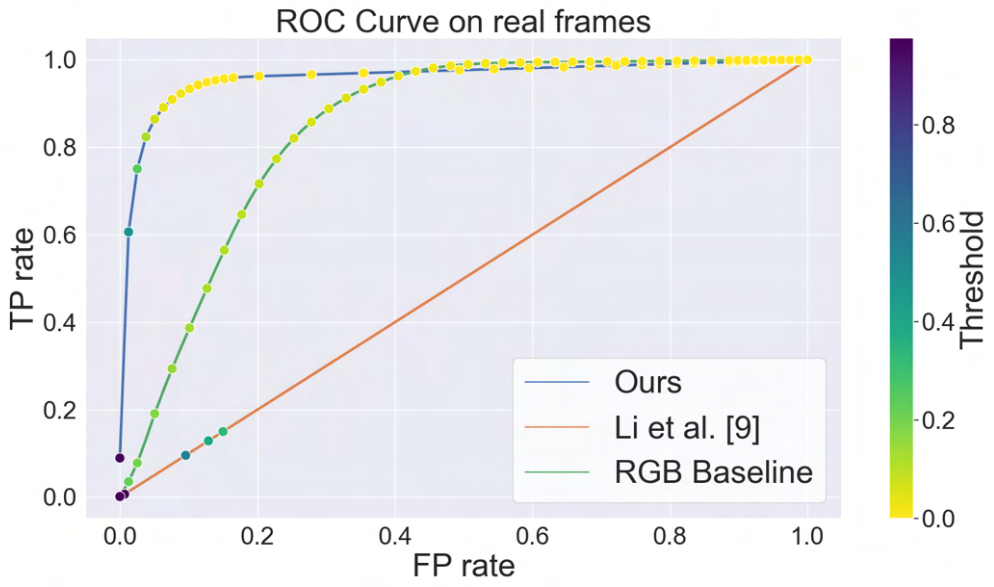


Figure 5.9: *Quantitative evaluation of the real sequence shown in Figure 5.10 (center).*

problem of estimating contact between two hands. To the best of our knowledge, our approach is the first to tackle this problem directly from RGB images.

5.4 Conclusions, Limitations, and Future Work

We have presented a method for camera-space contact estimation of two interacting hands. Our approach builds on top of existing state-of-the-art two-hand tracking systems (e.g. Li et al. (2022)), which often suffer from residual errors that prevent the computation of hand contacts. We first render the estimated hand meshes using a pixel-to-surface texture and feed the renders to an image-to-image translation network that yields dense contact labels. Our estimated contacts consistently achieve higher F1 scores than the raw tracking systems on both synthetic and real scenes. Therefore, we can estimate accurate contact even when the underlying tracking fails. A key benefit of our method is that it is designed as a standalone component, not limited to a specific input sensor or modality. Since our network takes as input rendered pixel-to-surface correspondences, our method circumvents many challenges common in real-world images such as illumination, sensor noise, or shadows.

By accurately estimating hand-to-hand contacts, our method could improve significantly the effectiveness of HCI systems, providing a more natural and immersive user experience. Contact detection is a fundamental component for realistic hand-hand and hand-object interactions and gesture recognition. Thus, our approach not only addresses a significant gap in existing hand tracking technologies but also paves the way for more intuitive and effective virtual interaction interfaces.

Limitations. As with most other learning-based methods, our system strongly relies on the quality of the training corpus. We believe we have produced a fairly rich set of

Figure 5.10: *Qualitative results of on real-world test sequences of different subjects and different lighting conditions. Our method predicts contact maps for a wide variety of hand-hand interactions directly from single RGB input.*

training sequences and shown the network's generalization capabilities. However, there are still some residual cases where we fail to predict correct contact labels. For instance, Figure 5.11 shows some of our failure cases. One intuition that can explain these failures is that our training set mostly contains cases with positive contact labels, i.e., we showcase just a few scenarios where there is no contact at all.

Figure 5.11: *Failure cases of our approach. Our method sometimes gives false positives if only one hand is partially visible (left) or in very ambiguous hand interactions (center). Finally, our method sometimes struggle with with heavily occluded and articulated hands (right).*

Another limitation of our method is that it relies on the quality of the underlying tracker, up to some extent. While we aim at solving the situations where the tracker cannot infer contact due to the depth ambiguities for the hands, our method will not work if the tracking is very unstable or estimates 3D poses that significantly differ from the observations.

Future work. We are currently considering multiple lines for future work. First, we want to investigate whether it is also possible to infer the forces involved in the interaction. Since we leverage a physics-based simulator to generate data, we can also generate labels for external contact forces (*e.g.*, magnitude, direction) and we could try to learn them. Second, at the moment our pipeline runs at interactive rates (3-5 fps) but does not achieve real-time performance. Improving the performance of each of the steps of our system is required to speed up the system. Finally, our method heavily depends on the initial 3D poses estimated by the baseline state-of-the-art hand tracking method. Future research should look into relaxing this dependency.

6 Conclusions

6.1 General Conclusions

Overall, this thesis aims to enhance hand-object and hand-hand interactions to offer immersive and intuitive experiences and interfaces. We achieve this objective by proposing tailored methods for each scenario. First, we developed a method to retarget hand poses between hands of different sizes and skeletal morphologies, improving hand-object interactions. Second, we presented a new framework to generate physically accurate two-hand interaction data, creating a training dataset to develop the first real-time tracking system capable of reconstructing two interacting hands from monocular RGB video. Additionally, we introduced an image-based approach to estimate hand-to-hand contacts, enhancing the overall quality of hand-hand interactions beyond what current RGB-based state-of-the-art of two hand tracking methods can achieve due to errors in depth, shape or pose estimations.

Chapter 3 is dedicated to hand-object interactions and addresses the gap between tracked hand data and simulated hand models through a pose retargeting strategy that avoids the necessity of sharing a common hand representation. Our approach is versatile, functioning effectively with any tracking or simulation method by acting as an intermediary between the two. Additionally, we assess the practical implications of hand representation discrepancies on manipulating virtual objects by comparing our pose retargeting strategy with a naïve hand pose copying approach revealing that while the mismatch in hand representation has minimal impact on manipulating larger objects, it markedly affects the precision required for manipulating smaller objects. The proposed method culminated with a presentation under the title *Fine Virtual Manipulation with Hands of Different Sizes* (Sorli et al. 2021) at the **2021 IEEE International Symposium on Mixed and Augmented Reality** (**ISMAR**) Conference, which is ranked A* in the CORE conference ranking system.

Chapters 4 and 5, instead, focused on scenarios involving two interacting hands. RGB cameras are ubiquitous but do not provide any depth cues, which poses a significant obstacle in such scenarios. In Chapter 4, we developed a framework capable of generating precise data on sequences involving two hands in close interaction, including intra-hand depth and inter-hand distance, for training. This was

a key component in designing a pioneering real-time method for capturing both the skeletal pose and 3D surface geometry of hands from a single RGB camera, explicitly considering close interactions. This method has been shown to outperform previous RGB-only methods in complex two-hand interaction scenarios. However, in some cases, it still suffers from residual errors in depth, shape, or hand pose estimation, which prevent accurate detection of hand-to-hand contacts despite the promising outcomes and continues to present an ongoing problem. In Chapter 5, we introduced an image-based data-driven method to estimate the contact in hand-to-hand interactions. At inference time, we estimate pixel-to-surface correspondences using state-of-the-art hand tracking and then use our network to predict accurate hand-to-hand contact. We validate our approach extensively with both qualitative and quantitative analyses on real-world data, demonstrating its superior accuracy over current state-of-the-art hand-tracking methods.

This research culminated with a conference presentation and a journal publication titled *RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video*, in the **ACM Transactions on Graphics** (**TOG**) journal (JCR Q1) as fourth author (Wang et al. 2020), and an article currently under review titled *Hand-to-hand Contact from RGB Images*.

6.2 Discussion and Future Work

The upcoming sections will delve into the technical contributions, emphasizing the strengths of the methods, as well as their limitations and avenues for future research.

6.2.1 Fine Virtual Manipulation with Hands of Different Sizes

The proposed method seamlessly connects existing hand tracking solutions with physics-based hand simulation, eliminating the necessity for a common hand representation. User study results show that this approach is effective for fine manipulation, achieving performance and naturalness comparable to gross manipulation. From an application perspective, our hand retargeting approach could advance VR training systems that demand high levels of dexterity and precise manipulation.

The key technical insight of the method lies in optimizing fingertip positions to enhance accuracy, particularly for pinch poses critical to fine manipulation. Future research could explore more diverse feature vectors and objective functions, incorportaing other hand points and pose likelihoods. Additionally, the user study could be extended by covering a broader range of interaction tasks that could further validate our approach.

The retargeting is based on the assumption that the shape of the simulated VR hand is predefined. This assumption holds true when the VR application uses a hand of a fixed size. A potential extension would be to enable the simulation of personalized hands, which would require modifying the shape of the simulated hand. Currently, our method approximates this step by estimating a uniform scale, but

future work could involve estimating more detailed parameters, such as statistical shape descriptors, as proposed by Romero et al. (2017).

Currently, the proposed retargeting method is limited to hands with similar skeletal topology, e.g, with five fingers. However, the approach could be extended to connect hands with very diverse skeletons, e.g., with a different number of fingers or phalanxes. The main technical challenge is defining appropriate feature metrics for these varied hand structures. Addressing this would broaden the versatility and inclusiveness of our approach across a broader range of VR scenarios.

6.2.2 Generating Annontated Data of Physically Accurate Two-Hand Interactions

The presented framework enables the simulation and generation of data for realistic two-hand interactions. This framework was a crucial component in the development of an effective 3D two-hand tracking and reconstruction method, that produces promising results, demonstrating compelling performance even in complex and demanding scenarios.

While the proposed method generally performs well, especially in scenarios involving close hand-hand interactions, there are limitations that need to be addressed in the future. Currently, the method faces difficulties in tracking very fast hand movements, as motion blur can lead to unreliable neural network predictions. To mitigate this issue, generating additional training data with motion blur using our framework could enhance the neural network's ability to handle such challenging cases more effectively.

Moreover, achieving an optimal trade-off between the MANO pose prior and other energy terms is challenging, and often requires a compromise between pose variability and pose plausibility. One possible solution is to integrate a kinematic skeleton into the MANO model, which would enforce explicit joint limit constraints, while still using the pose space to capture correlations in joint movements.

Due to inherent depth ambiguity, our method may also face difficulties with interactions that require high precision in the relative positioning of hands. For such tasks it may be necessary to incorporate a new energy term or additional information from a depth sensor or stereo camera.

Exploring the explicit use of the temporal dimension could also provide further insights and using temporal neural network architectures could contribute to smoother predictions and thus further improve temporal tracking consistency.

6.2.3 Hand-to-hand Contact from RGB Images

We introduced a method for estimating camera-space contact points between two interacting hands. Our approach enriches existing state-of-the-art two-hand tracking systems (Li et al. 2022), which often suffer from residual errors that hinder accurate hand contact computation. Our method consistently achieves higher F1 scores for contact estimation than the raw tracking systems in both synthetic and real-world scenarios, enabling precise contact estimation even when the underlying

tracking fails. A significant advantage of our method is that it is designed as a standalone component, not limited to a specific input sensor or modality. Since our network takes as input rendered pixel-to-surface correspondences, our method circumvents many challenges common in real-world images such as illumination, sensor noise, or shadows.

Like most learning-based methods, the performance of our system heavily relies on the quality of the training data. We believe we have produced a fairly rich set of training sequences and demonstrated the generalization capabilities of the proposed network. However, there are still instances where our method fails to predict correct contact labels. One possible explanation for these failures is that our training set mostly consists of cases with positive contact labels.

The proposed method also relies partially on the quality of the underlying tracker. While our approach addresses mitigates issues where the tracker fails to infer contact due to depth ambiguities between hands, it struggles when the tracking is highly unstable or generates 3D poses that significantly deviate from actual observations.

For future work, several directions are considered. Firstly, we want to investigate the possibility of inferring the forces involved in interactions. By leveraging the proposed physics-based simulator framework to generate data, we can generate labels for external contact forces (*e.g.*, magnitude, direction) and attempt to learn them. Secondly, although our pipeline currently runs at interactive rates (3-5 fps), it does not achieve real-time performance. Improving the performance of each system component is necessary to speed up the overall process. Lastly, our method heavily depends on the initial 3D poses estimated by the baseline state-of-the-art hand tracking method. Future research should look into relaxing this dependency.

6.3 Final Remarks

In this thesis, we have introduced several innovative methods and approaches that made significant advancements in the field of hand-object and hand-hand interactions in VR and AR. Our contributions have been demonstrated through rigorous experimentation and practical applications, enhancing the realism and immersion of VR experiences. The proposed solutions address critical challenges in hand-object and two-hand scenarios, providing a more natural and intuitive interaction model that will benefit both commercial applications and research. As VR and AR technologies continue to evolve and gain popularity, the demand for sophisticated interaction models will only grow, and we are confident that our solutions will be essential in addressing this need.

We are optimistic that the advancements in AR and VR will soon make these technologies more prevalent and integrated into everyday life. The progress achieved in this thesis contributes to its broader adoption by improving the usability and accessibility of these technologies. By developing more practical and intuitive interfaces, we aim to eliminate the need for cumbersome sensors and devices. Combined

with ongoing improvements in technology and machine learning algorithms, these developments will significantly enhance the proposed experiences. We hope that these advancements will attract a wider audience, unfamiliar with this virtual world, and inspire a passion for it similar to my own.

An important step forward would involve prioritizing hand diversity and accommodating individuals with disabilities. While current hand tracking and new interface models offer hopeful solutions, accessing this technology remains challenging for many.

Bibliography

Alp Güler, R., Neverova, N. & Kokkinos, I. (2018), Densepose: Dense human pose estimation in the wild, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 52

Argelaguet, F., Hoyet, L., Trico, M. & Lecuyer, A. (2016), The role of interaction in virtual embodiment: Effects of the virtual hand representation, *in* '2016 IEEE Virtual Reality (VR)', pp. 3–10. ↑ 23, 25

Baek, S., In Kim, K. & Kim, T.-K. (2018), Augmented skeleton space transfer for depth-based hand pose estimation, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Baek, S., Kim, K. I. & Kim, T.-K. (2019), Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21, 87

Ballan, L., Taneja, A., Gall, J., Gool, L. V. & Pollefeys, M. (2012), Motion Capture of Hands in Action using Discriminative Salient Points, *in* 'European Conference on Computer Vision (ECCV)'. ↑ 4, 20, 51

Barbagli, F., Frisoli, A., Salisbury, K. & Bergamasco, M. (2004), Simulating human fingers: a soft finger proxy model and algorithm, *in* 'HAPTICS '04. Proc. 12th International Symposium on'. ↑ 14

Barreiro, H., Torres, J. & Otaduy, M. A. (2021), Natural tactile interaction with virtual clay, *in* 'Proc. of World Haptics Conference'.
URL: *http://gmrv.es/Publications/2021/BTO21* ↑ vii, 2, 87

Borst, C. W. & Indugula, A. P. (2005), Realistic virtual grasping, *in* 'Proc. of IEEE Virtual Reality Conference'. ↑ 3, 14, 86

Boukhayma, A., Bem, R. d. & Torr, P. H. (2019), 3d hand shape and pose from images in the wild, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H. & Philipsen, A. (2018), 'The senses of agency and ownership: a review', *Frontiers in psychology* **9**, 535. ↑ 23

Bronstein, M. M., Bronstein, A. M., Kimmel, R. & Yavneh, I. (2006), 'Multigrid multidimensional scaling', *Numerical linear algebra with applications* **13**(2-3), 149–171. ↑ 41

Buda, M., Saha, A. & Mazurowski, M. A. (2019), 'Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm', *Computers in Biology and Medicine* **109**. ↑ 55

Cai, Y., Ge, L., Cai, J. & Yuan, J. (2018), Weakly-supervised 3d hand pose estimation from monocular rgb images, *in* 'European Conference on Computer Vision', Springer, Cham, pp. 1–17. ↑ 4, 21, 39

Cao, Z., Radosavovic, I., Kanazawa, A. & Malik, J. (2021), Reconstructing hand-object interactions in the wild, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 12417–12426. ↑ 23, 88

Chan, C., Ginosar, S., Zhou, T. & Efros, A. A. (2019), Everybody dance now, *in* 'The IEEE International Conference on Computer Vision (ICCV)'. ↑ 24

Chen, X., Liu, Y., Yajiao, D., Zhang, X., Ma, C., Xiong, Y., Zhang, Y. & Guo, X. (2022), Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21, 87

Chen, Y., Dwivedi, S. K., Black, M. J. & Tzionas, D. (2023), Detecting Human-Object Contact in Images, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.
URL: *https://hot.is.tue.mpg.de* ↑ 23, 88

Chen, Y., Tu, Z., Ge, L., Zhang, D., Chen, R. & Yuan, J. (2019), So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning, *in* 'The IEEE International Conference on Computer Vision (ICCV)'. ↑ 21

Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R. & Yuan, J. (2021), Model-based 3d hand reconstruction via self-supervised learning, *in* 'Conference on Computer Vision and Pattern Recognition'. ↑ 21

Chessa, M., Maiello, G., Klein, L. K., Paulun, V. C. & Solari, F. (2019), Grasping objects in immersive virtual reality, *in* '2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)', pp. 1749–1754. ↑ 2, 25, 86

Ciocarlie, M., Lackner, C. & Allen, P. (2007), Soft finger model with adaptive contact geometry for grasping and manipulation tasks, *in* 'World Haptics Conference'. ↑ 14

Cosco, F., Garre, C., Bruno, F., Muzzupappa, M. & Otaduy, M. A. (2013), 'Visuo-haptic mixed reality with unobstructed tool-hand integration', *IEEE Transactions on Visualization and Computer Graphics* **19**(1), 159–172. ↑ 23

Doosti, B., Naha, S., Mirbagheri, M. & Crandall, D. (2020), Hope-net: A graph-based model for hand-object pose estimation, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 22

Duriez, C., Courtecuisse, H., Plata Alcalde, J.-P. d. & Bensoussan, P.-J. (2008), 'Contact skinning', *Eurographics conference (short paper)* . ↑ 3, 14, 86

Facebook (2019), 'Oculus Quest', `https://www.oculus.com/quest`. Accessed: 2019-11-22. ↑ 3

Fan, Z., Spurr, A., Kocabas, M., Tang, S., Black, M. & Hilliges, O. (2021), Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-pixel Part Segmentation, *in* '3DV'. ↑ 22, 88

Fieraru, M., Zanfir, M., Oneata, E., Popa, A.-I., Olaru, V. & Sminchisescu, C. (2020), Three-dimensional reconstruction of human interactions, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 23

Frisoli, A., Barbagli, F., Ruffaldi, E., Salisbury, K. & Bergamasco, M. (2006), A limit-curve

based soft finger god-object algorithm, *in* 'Haptic Interfaces for Virtual Environment and Teleoperator Systems, 14th Symposium on'. ↑ 14

Garcia-Hernando, G., Yuan, S., Baek, S. & Kim, T.-K. (2018), First-person hand action benchmark with RGB-D videos and 3D hand pose annotations, *in* 'Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 409–419. ↑ 22

Garre, C., Hernandez, F., Gracia, A. & Otaduy, M. A. (2011), Interactive simulation of a deformable hand for haptic rendering, *in* 'Proc. of World Haptics Conference'. ↑ 3, 15, 86

Gast, T. F., Schroeder, C., Stomakhin, A., Jiang, C. & Teran, J. M. (2015), 'Optimization integrator for large time steps', *IEEE Transactions on Visualization and Computer Graphics* **21**(10), 1103–1115. ↑ 15, 30

Ge, L., Cai, Y., Weng, J. & Yuan, J. (2018), Hand pointnet: 3d hand pose estimation using point sets, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Ge, L., Liang, H., Yuan, J. & Thalmann, D. (2016), Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Ge, L., Liang, H., Yuan, J. & Thalmann, D. (2017), 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1991–2000. ↑ 21

Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J. & Yuan, J. (2019), 3d hand shape and pose estimation from a single rgb image, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 10833–10842. ↑ 21, 87

Grady, P., Tang, C., Twigg, C. D., Vo, M., Brahmbhatt, S. & Kemp, C. C. (2021), ContactOpt: Optimizing contact to improve grasps, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 22, 23, 88

Hampali, S., Rad, M., Oberweger, M. & Lepetit, V. (2020), 'Honnotate: A method for 3d annotation of hand and object poses'.
URL: *https://arxiv.org/abs/1907.01481* ↑ 21, 22, 87

Han, S., Liu, B., Cabezas, R., Twigg, C. D., Zhang, P., Petkau, J., Yu, T.-H., Tai, C.-J., Akbay, M., Wang, Z., Nitzan, A., Dong, G., Ye, Y., Tao, L., Wan, C. & Wang, R. (2020), 'Megatrack: monochrome egocentric articulated hand-tracking for virtual reality', *ACM Trans. Graph.* **39**(4).
URL: *https://doi.org/10.1145/3386569.3392452* ↑ 22, 88

Han, S., Liu, B., Wang, R., Ye, Y., Twigg, C. D. & Kin, K. (2018), 'Online optical marker-based hand tracking with deep labels', *AMC TOG* **37**(4), 166. ↑ 22, 88

Hand Physics Lab (2021), https://www.holonautic.com/hand-physics-lab. ↑ vii, 2, 87

Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M. & Schmid, C. (2020), Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction, *in* 'CVPR'. ↑ 22

Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M. J., Laptev, I. & Schmid, C. (2019), Learning joint reconstruction of hands and manipulated objects, *in* 'Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 4, 22

He, T., Gao, L., Song, J. & Li, Y.-F. (2021), Exploiting Scene Graphs for Human-Object Interaction Detection, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 15984–15993. ↑ 23, 88

Hecker, C., Raabe, B., Enslow, R. W., DeWeese, J., Maynard, J. & van Prooijen, K. (2008), 'Real-time motion retargeting to highly varied user-created morphologies', *ACM Trans. Graph.* **27**(3), 1–11. ↑ 24

Hirota, K. & Tagawa, K. (2016), Interaction with virtual object using deformable hand, *in* 'IEEE Virtual Reality (VR)', IEEE, pp. 49–56. ↑ 3, 14, 86

Holl, M., Oberweger, M., Arth, C. & Lepetit, V. (2018), Efficient physics-based implementation for realistic hand-object interaction in virtual reality, *in* '2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)', pp. 175–182. ↑ 14

HTC (2016), 'HTC Vive', `https://www.vive.com`. Accessed: 2019-11-22. ↑ 3

Huang, C.-H. P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D. & Black, M. J. (2022), Capturing and inferring dense full-body human-scene contact, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 13274–13285. ↑ 23

Iqbal, U., Molchanov, P., Breuel Juergen Gall, T. & Kautz, J. (2018), Hand pose estimation via latent 2.5 d heatmap regression, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 118–134. ↑ 21, 87

Jacobs, J. & Froehlich, B. (2011), A soft hand model for physically-based manipulation of virtual objects, *in* '2011 IEEE Virtual Reality Conference (VR)', pp. 11–18. ↑ 3, 14, 86

Jacobs, J., Stengel, M. & Froehlich, B. (2012), A generalized god-object method for plausible finger-based interactions in virtual environments, *in* '2012 IEEE Symposium on 3D User Interfaces (3DUI)', pp. 43–51. ↑ 14

Jiang, H., Liu, S., Wang, J. & Wang, X. (2021), Hand-object contact consistency reasoning for human grasps generation, *in* 'Proceedings of the International Conference on Computer Vision'. ↑ 23, 88

Jiang, Z., Rahmani, H., Black, S. & Williams, B. M. (2023), A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 758–767. ↑ 21

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T. S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S. & Sheikh, Y. (2017), 'Panoptic studio: A massively multiview system for social interaction capture', *IEEE Transactions on Pattern Analysis and Machine Intelligence* . ↑ 46

Kaplan, A., Cruit, J., Endsley, M., Beers, S., Sawyer, B. & Hancock, P. (2021), 'The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis', *Human Factors* **63**(4), 706–726. ↑ 2, 86

Karunratanakul, K., Prokudin, S., Hilliges, O. & Tang, S. (2023), Harp: Personalized hand reconstruction from a monocular rgb video, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 20

Karunratanakul, K., Yang, J., Zhang, Y., Black, M. J., Muandet, K. & Tang, S. (2020), Grasping field: Learning implicit representations for human grasps, *in* '2020 International Conference on 3D Vision (3DV)', IEEE Computer Society, Los Alamitos, CA, USA, pp. 333–344. **URL:** *https://doi.ieeecomputersociety.org/10.1109/3DV50981.2020.00043* ↑ 22

Kilteni, K., Groten, R. & Slater, M. (2012), 'The sense of embodiment in virtual reality', *Presence: Teleoperators and Virtual Environments* **21**(4), 373–387. ↑ 23

Kim, D. U., Kim, K. I. & Baek, S. (2021), End-to-end detection and pose estimation of two interacting hands, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 11189–11198. ↑ 22, 88

Kim, J.-S. & Park, J.-M. (2015), Physics-based hand interaction with virtual objects, *in* 'IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 3814–3819. ↑ 3, 14, 86

Krichenbauer, M., Yamamoto, G., Taketom, T., Sandor, C. & Kato, H. (2018), 'Augmented reality versus virtual reality for 3d object manipulation', *IEEE Transactions on Visualization and Computer Graphics* **24**(2), 1038–1048. ↑ 2, 86

Kry, P. G., James, D. L. & Pai, D. K. (2002), Eigenskin: Real time large deformation character skinning in hardware, *in* 'ACM SIGGRAPH Symp. on Computer Animation', pp. 153–160. ↑ 14

Kry, P. G. & Pai, D. K. (2006), 'Interaction capture and synthesis', *ACM Transactions on Graphics (TOG)* **25**(3), 872–880. ↑ 14

Kurihara, T. & Miyata, N. (2004), Modeling deformable human hands from medical images, *in* '2004 ACM SIGGRAPH / Eurographics Symp. on Computer Animation', pp. 355–363. ↑ 14

Kyriazis, N. & Argyros, A. (2014), Scalable 3d tracking of multiple interacting objects, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 3430–3437. ↑ 22

LeapMotion (2016), https://developer.leapmotion.com/orion. ↑ 40, 54, 56

Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T. & Liu, Y. (2022), Interacting attention graph for single image two-hand reconstruction, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 2761–2770. ↑ ix, x, 4, 22, 51, 52, 53, 54, 55, 56, 57, 59, 60, 61, 67, 88

Li, S. & Lee, D. (2019), Point-to-pose voting based hand pose estimation using residual permutation equivariant layer, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Li, Y., Fu, J. L. & Pollard, N. S. (2007), 'Data-driven grasp synthesis using shape matching and task-based pruning', *IEEE Transactions on Visualization and Computer Graphics* **13**(4), 732–747. ↑ 14

Lin, L., Normoyle, A., Adkins, A., Sun, Y., Robb, A., Ye, Y., Di Luca, M. & Jörg, S. (2019), The effect of hand size and interaction modality on the virtual hand illusion, *in* '2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)', pp. 510–518. ↑ 23

Liu, S., Jiang, H., Xu, J., Liu, S. & Wang, X. (2021), Semi-supervised 3d hand-object poses

estimation with interactions in time, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 14687–14697. ↑ 22, 23, 88

Magnenat, T., Laperriere, R. & Thalmann, D. (1988), Joint-dependent local deformations for hand animation and object grasping, Canadian Inf. Process. Soc, p. 26–33. Montreal Univ., Que., Canada.
URL: *http://infoscience.epfl.ch/record/98750* ↑ 11

Malik, J., Abdelaziz, I., Elhayek, A., Shimada, S., Ali, S. A., Golyanik, V., Theobalt, C. & Stricker, D. (2020), Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 4, 51

Martin, S., Thomaszewski, B., Grinspun, E. & Gross, M. (2011), 'Example-based elastic materials', *ACM Trans. Graph.* **30**(4). ↑ 30

Melax, S., Keselman, L. & Orsten, S. (2013), Dynamics based 3d skeletal hand tracking, *in* 'Proceedings of Graphics Interface 2013', GI '13, Canadian Information Processing Society, CAN, p. 63–70. ↑ 20

Moon, G., Yu, S.-I., Wen, H., Shiratori, T. & Lee, K. M. (2020), InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image, *in* 'European Conference on Computer Vision (ECCV)'. ↑ 4, 22, 51, 88

Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D. & Theobalt, C. (2018), GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB, *in* 'Proceedings of Computer Vision and Pattern Recognition (CVPR)'.
URL: *http://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/* ↑ 4, 21, 22, 39, 51, 87

Mueller, F., Davis, M., Bernard, F., Sotnychenko, O., Verschoor, M., Otaduy, M. A., Casas, D. & Theobalt, C. (2019), 'Real-time pose and shape reconstruction of two interacting hands with a single depth camera', *ACM TOG* **38**(4), 49. ↑ 3, 21, 22, 25, 39, 40, 41, 43, 52, 54, 87, 88

Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D. & Theobalt, C. (2017), Real-time hand tracking under occlusion from an egocentric rgb-d sensor, *in* 'International Conference on Computer Vision (ICCV)'. ↑ vii, 11, 21

Müller, M., Dorsey, J., McMillan, L., Jagnow, R. & Cutler, B. (2002), Stable real-time deformations, *in* 'Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation', SCA '02, ACM, New York, NY, USA, pp. 49–54.
URL: *http://doi.acm.org/10.1145/545261.545269* ↑ 17

Narasimhaswamy, S., Nguyen, T. & Hoai, M. (2020), Detecting hands and recognizing physical contact in the wild, *in* 'Advances in Neural Information Processing Systems'. ↑ 23, 88

Nocedal, J. & Wright, S. J. (2006), *Numerical Optimization*, second edn, Springer, New York, NY, USA. ↑ 28

Oberweger, M., Wohlhart, P. & Lepetit, V. (2015), Training a feedback loop for hand pose estimation, *in* 'IEEE International Conference on Computer Vision (ICCV)', pp. 3316–3324. ↑ 21

Oikonomidis, I., Kyriazis, N. & Argyros, A. A. (2011*a*), Efficient model-based 3d tracking of hand articulations using kinect., *in* 'BMVC', Vol. 1, p. 3. ↑ 20

Oikonomidis, I., Kyriazis, N. & Argyros, A. A. (2011*b*), Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 2088–2095. ↑ 20

Oikonomidis, I., Kyriazis, N. & Argyros, A. A. (2012), Tracking the articulated motion of two strongly interacting hands, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 1862–1869. ↑ 22, 88

Ott, R., Vexo, F. & Thalmann, D. (2010), 'Two-handed haptic manipulation for CAD and VR applications', *Computer Aided Design & Applications* **7**(1). ↑ 3, 14, 86

Panteleris, P., Kyriazis, N. & Argyros, A. A. (2015), 3d tracking of human hands in interaction with unknown objects., *in* 'BMVC', pp. 123–1. ↑ 39

Panteleris, P., Oikonomidis, I. & Argyros, A. (2018), Using a single rgb frame for real time 3d hand pose estimation in the wild, *in* '2018 IEEE Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 436–445. ↑ 21

Park, G., Kim, T.-K. & Woo, W. (2020), 3d hand pose estimation with a single infrared camera via domain transfer learning, *in* '2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)', pp. 588–599. ↑ 21

Park, J., Oh, Y., Moon, G., Choi, H. & Lee, K. M. (2022), Handoccnet: Occlusion-robust 3d hand mesh estimation network, *in* '2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 1486–1495. ↑ 21, 22, 87

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D. & Black, M. J. (2019), Expressive Body Capture: 3D Hands, Face, and Body from a Single Image, *in* 'Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21, 87

Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D. & Malik, J. (2024), Reconstructing hands in 3D with transformers, *in* 'CVPR'. ↑ 21, 87

Perez, A. G., Cirio, G., Hernandez, F., Garre, C. & Otaduy, M. A. (2013), Strain limiting for soft finger contact simulation, *in* 'World Haptics Conference (WHC)', IEEE, pp. 79–84. ↑ 15

Perez, A. G., Cirio, G., Lobo, D., Chinello, F., Prattichizzo, D. & Otaduy, M. A. (2016), Efficient nonlinear skin simulation for multi-finger tactile rendering, *in* 'Proc. of Haptics Symposium', IEEE.
URL: *http://www.gmrv.es/Publications/2016/PCLCPO16* ↑ 15

Pollard, N. S. & Zordan, V. B. (2005), Physically based grasping control from example, *in* 'Proc. of the ACM SIGGRAPH/Eurographics symposium on Computer animation', ACM, pp. 311–318. ↑ 14

Potamias, R. A., Ploumpis, S., Moschoglou, S., Triantafyllou, V. & Zafeiriou, S. (2023), Handy: Towards a high fidelity 3d hand shape and appearance model, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 4670–4680. ↑ 12

Pouliquen, M., Duriez, C., Andriot, C., Bernard, A., Chodorge, L. & Gosselin, F. (2005), Real-time finite element finger pinch grasp simulation, *in* 'Eurohaptics Conference, 2005

and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2005. World Haptics 2005. First Joint', pp. 323–328. ↑ 14

Qian, C., Sun, X., Wei, Y., Tang, X. & Sun, J. (2014), Realtime and robust hand tracking from depth, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1106–1113. ↑ 20, 87

Qian, N., Wang, J., Mueller, F., Bernard, F., Golyanik, V. & Theobalt, C. (2020), Html: A parametric hand texture model for 3d hand reconstruction and personalization, *in* 'European Conference on Computer Vision'.
URL: *https://api.semanticscholar.org/CorpusID:220702354* ↑ 40

Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.-P., Schiele, B. & Theobalt, C. (2016), 'EgoCap: Egocentric marker-less motion capture with two fisheye cameras', *ACM Transactions on Graphics (TOG)* **35**(6), 162. ↑ 4

Richard L. Drake, A. Wayne Vogl, A. W. M. M. (2023), *Gray's Anatomy for Students, 5th Edition*, ELSEVIER. ↑ vii, 11

Rivers, A. R. & James, D. L. (2007), 'Fastlsm: fast lattice shape matching for robust real-time deformation', *ACM Trans. Graph.* **26**(3), 82–es.
URL: *https://doi.org/10.1145/1276377.1276480* ↑ 14

Rogez, G., Khademi, M., Supančič III, J., Montiel, J. M. M. & Ramanan, D. (2014), 3D hand pose detection in egocentric RGB-D images, *in* 'Workshop at the European Conference on Computer Vision', Springer, pp. 356–371. ↑ 21

Rogez, G., Supančič, J. S. & Ramanan, D. (2015), First-person pose recognition using egocentric workspaces, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 4325–4333. ↑ 22

Romero, J., Tzionas, D. & Black, M. J. (2017), 'Embodied hands: Modeling and capturing hands and bodies together', *ACM Trans. Graph.* **36**(6), 245:1–245:17.
URL: *http://doi.acm.org/10.1145/3130800.3130883* ↑ vii, viii, 4, 12, 13, 27, 30, 32, 37, 43, 44, 52, 54, 55, 56, 67, 90

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241. ↑ 53, 55

Sanchez-Vives, M. V., Spanlang, B., Frisoli, A., Bergamasco, M. & Slater, M. (2010), 'Virtual hand illusion induced by visuomotor correlations', *PLoS ONE* **5**(4). ↑ 23

Shan, D., Geng, J., Shu, M. & Fouhey, D. (2020), Understanding human hands in contact at internet scale, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 23, 88

Simon, T., Joo, H., Matthews, I. & Sheikh, Y. (2017), Hand keypoint detection in single images using multiview bootstrapping, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 22, 88

Sinha, A., Choi, C. & Ramani, K. (2016), Deephand: Robust hand pose estimation by completing a matrix imputed with deep features, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 4150–4158. ↑ 21

Sorli, S., Casas, D., Verschoor, M., Tajadura-Jiménez, A. & Otaduy, M. A. (2021), Fine virtual manipulation with hands of different sizes, *in* 'Proc. of the International Symposium on Mixed and Augmented Reality (ISMAR)'.
URL: *http://gmrv.es/Publications/2021/SCVTO21* ↑ 6, 65, 92, 93

Spurr, A., Song, J., Park, S. & Hilliges, O. (2018), Cross-modal deep variational hand pose estimation, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Sridhar, S., Mueller, F., Oulasvirta, A. & Theobalt, C. (2015), Fast and Robust Hand Tracking Using Detection-Guided Optimization, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 4, 21, 51

Sridhar, S., Mueller, F., Zollhöefer, M., Casas, D., Oulasvirta, A. & Theobalt, C. (2016), Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input, *in* 'European Conference on Computer Vision (ECCV)'. ↑ 21, 39, 52

Sridhar, S., Oulasvirta, A. & Theobalt, C. (2013), Interactive Markerless Articulated Hand Motion Tracking using RGB and Depth data, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 2456–2463. ↑ 4, 20, 51, 87

Sueda, S., Kaufman, A. & Pai, D. K. (2008), 'Musculotendon simulation for hand animation', *ACM Trans. Graph.* **27**(3). ↑ 15

Sun, Q., Patney, A., Wei, L.-Y., Shapira, O., Lu, J., Asente, P., Zhu, S., Mcguire, M., Luebke, D. & Kaufman, A. (2018), 'Towards virtual reality infinite walking: Dynamic saccadic redirection', *ACM Trans. Graph.* **37**(4). ↑ 2, 86

Tagliasacchi, A., Schroeder, M., Tkach, A., Bouaziz, S., Botsch, M. & Pauly, M. (2015), 'Robust Articulated-ICP for Real-Time Hand Tracking', *Computer Graphics Forum (Symposium on Geometry Processing)* **34**(5), 101–114. ↑ 20, 87

Taheri, O., Ghorbani, N., Black, M. J. & Tzionas, D. (2020), GRAB: A Dataset of Whole-Body Human Grasping of Objects, *in* 'European Conference on Computer Vision (ECCV)'. ↑ 21, 23, 88

Talvas, A., Marchal, M., Duriez, C. & Otaduy, M. A. (2015), 'Aggregate constraints for virtual manipulation with soft fingers', *IEEE Transactions on Visualization and Computer Graphics* **21**(4), 452–461. ↑ 3, 14, 86

Taylor, C. J. & Kriegman, D. J. (1994), Minimization on the Lie Group SO(3) and related manifolds, Technical report, Yale University. ↑ 29

Taylor, J., Bordeaux, L., Cashman, T., Corish, B., Keskin, C., Sharp, T., Soto, E., Sweeney, D., Valentin, J., Luff, B. et al. (2016), 'Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences', *ACM Transactions on Graphics (TOG)* **35**(4), 143. ↑ 21, 39

Taylor, J., Tankovich, V., Tang, D., Keskin, C., Kim, D., Davidson, P., Kowdle, A. & Izadi, S. (2017), 'Articulated Distance Fields for Ultra-fast Tracking of Hands Interacting', *ACM Trans. Graph.* . ↑ 22, 39, 43, 88

Tekin, B., Bogo, F. & Pollefeys, M. (2019), H+o: Unified egocentric recognition of 3d hand-object poses and interactions, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 22

Tkach, A., Pauly, M. & Tagliasacchi, A. (2016), 'Sphere-meshes for real-time hand modeling and tracking', *ACM Transactions on Graphics (TOG)* **35**(6), 222. ↑ 20, 87

Tkach, A., Tagliasacchi, A., Remelli, E., Pauly, M. & Fitzgibbon, A. (2017), 'Online generative model personalization for hand tracking', *ACM Transactions on Graphics (TOG)* **36**(6), 243. ↑ 4, 12

Tompson, J., Stein, M., Lecun, Y. & Perlin, K. (2014), 'Real-time continuous pose recovery of human hands using convolutional networks', *ACM Transactions on Graphics (ToG)* **33**(5), 1–10. ↑ 21

Tu, Z., Huang, Z., Chen, Y., Kang, D., Bao, L., Yang, B. & Yuan, J. (2023), 'Consistent 3d hand reconstruction in video via self-supervised learning', *IEEE Transactions on Pattern Analysis amp; Machine Intelligence* **45**(08), 9469–9485. ↑ 21

Tuthill, J. C. & Azim, E. (2018), 'Proprioception', *Current Biology* **28**, R194–R203.
URL: *https://api.semanticscholar.org/CorpusID:235330764* ↑ 23

Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M. & Gall, J. (2016), 'Capturing hands in action using discriminative salient points and physics simulation', *International Journal of Computer Vision (IJCV)* **118**(2), 172–193.
URL: *http://files.is.tue.mpg.de/dtzionas/Hand-Object-Capture* ↑ 4, 22

Tzionas, D. & Gall, J. (2015), 3d object reconstruction from hand-object interactions, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 729–737. ↑ 39

Verschoor, M., Lobo, D. & Otaduy, M. A. (2018), Soft hand simulation for smooth and robust natural interaction, *in* '2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)', pp. 183–190. ↑ vii, ix, 3, 7, 15, 16, 17, 18, 19, 25, 26, 27, 30, 40, 52, 53, 54, 56, 57, 86

VirtualGrasp (2022), https://www.virtualgrasp.com/. ↑ vii, 2, 87

VRtuos (2020), https://vrtuos.eu/. ↑ vii, 2, 87

Wan, C., Probst, T., Gool, L. V. & Yao, A. (2019), Self-supervised 3d hand pose estimation through training by fitting, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 21

Wan, C., Probst, T., Van Gool, L. & Yao, A. (2017), Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 680–689. ↑ 21

Wang, J., Luvizon, D., Mueller, F., Bernard, F., Kortylewski, A., Casas, D. & Theobalt, C. (2022), HandFlow: Quantifying View-Dependent 3D Ambiguity in Two-Hand Reconstruction with Normalizing Flow, *in* 'International Symposium on Vision, Modeling, and Visualization (VMV)', pp. 99–106. ↑ 22, 88

Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M. A., Casas, D. & Theobalt, C. (2020), 'RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video', *ACM Transactions on Graphics (TOG)* **39**(6). ↑ xi, 6, 40, 43, 44, 45, 46, 47, 48, 50, 51, 52, 66, 92, 94

Wang, R., Paris, S. & Popović, J. (2011), 6D Hands: Markerless Hand-tracking for Computer Aided Design, *in* 'Proceedings of the 24th annual ACM symposium on User interface software and technology', pp. 549–558. ↑ 20

Wang, R. Y. & Popović, J. (2009), 'Real-Time Hand-Tracking with a Color Glove', *ACM Transactions on Graphics* **28**(3). ↑ 20

Wheatland, N., Wang, Y., Song, H., Neff, M., Zordan, V. & Jörg, S. (2015), 'State of the Art in Hand and Finger Modeling and Animation', *Computer Graphics Forum* . ↑ vii, 10

Wobbrock, J. O., Findlater, L., Gergle, D. & Higgins, J. J. (2011), The aligned rank transform for nonparametric factorial analyses using only anova procedures, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', CHI '11, Association for Computing Machinery, New York, NY, USA, p. 143–146. ↑ 33

Xiang, D., Joo, H. & Sheikh, Y. (2019), Monocular total capture: Posing face, body, and hands in the wild, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 10965–10974. ↑ 21, 87

Xie, X., Bhatnagar, B. L. & Pons-Moll, G. (2022), CHORE: Contact, Human and Object REconstruction from a single RGB image, *in* 'European Conference on Computer Vision (ECCV)', Springer. ↑ 23, 88

Xu, H., Wang, T., Tang, X. & Fu, C.-W. (2023), H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 17048–17058. ↑ 21, 22, 87

Yang, L., Li, S., Lee, D. & Yao, A. (2019), Aligning latent spaces for 3d hand pose estimation, *in* 'The IEEE International Conference on Computer Vision (ICCV)'. ↑ 21, 87

Ye, Q., Yuan, S. & Kim, T.-K. (2016), Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation, *in* 'European Conference on Computer Vision (ECCV)', Springer, pp. 346–361. ↑ 21

Yu, Z., Huang, S., Fang, C., Breckon, T. P. & Wang, J. (2023), 'Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction'.
URL: *https://arxiv.org/abs/2303.05938* ↑ 22, 88

Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Yong Chang, J., Mu Lee, K., Molchanov, P., Kautz, J., Honari, S., Ge, L., Yuan, J., Chen, X., Wang, G., Yang, F., Akiyama, K., Wu, Y., Wan, Q., Madadi, M., Escalera, S., Li, S., Lee, D., Oikonomidis, I., Argyros, A. & Kim, T.-K. (2018), Depth-based 3d hand pose estimation: From current achievements to future goals, *in* 'The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'. ↑ 39

Zhang, B., Wang, Y., Deng, X., Zhang, Y., Tan, P., Ma, C. & Wang, H. (2021), Interacting two-hand 3d pose and shape reconstruction from single color image, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 11354–11363. ↑ 4

Zhao, W., Zhang, J., Min, J. & Chai, J. (2013), 'Robust realtime physics-based motion control for human grasping', *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* **32**(6), 207:1–207:12.
URL: *http://doi.acm.org/10.1145/2508363.2508412* ↑ 40

Zhou, X., Wan, Q., Zhang, W., Xue, X. & Wei, Y. (2016), Model-based deep hand pose estimation, *in* 'IJCAI'. ↑ 21

Zimmermann, C. & Brox, T. (2017), Learning to estimate 3D hand pose from single RGB images, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 4903–4911. ↑ 4, 21, 46, 51

Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M. & Brox, T. (2019), Freihand: A dataset for markerless capture of hand pose and shape from single rgb images, *in* 'The IEEE International Conference on Computer Vision (ICCV)'. ↑ 4, 21, 39, 47, 51

Appendix

A Resumen

En los últimos años, la Realidad Aumentada (RA) y la Realidad Virtual (RV) han ganado popularidad y siguen creciendo en términos de interés y uso, centrándose especialmente en la interacción con el usuario. Estas tecnologías transforman la forma en que las personas se relacionan con los contenidos digitales y los entornos inmersivos, generando una atención considerable en sectores como el entretenimiento, la educación, la formación y la sanidad. La necesidad de interacciones naturales en RV y RA ha surgido para mejorar la inmersión, la accesibilidad y el realismo. El objetivo es redefinir la interacción persona-ordenador combinando a la perfección los mundos virtual y físico, ofreciendo distintos niveles de interacción, desde una sutil manipulación virtual hasta interacciones corporales completas en entornos simulados.

La RA y la RV se basan en dos importantes componentes tecnológicos que suelen abordarse de forma independiente: el seguimiento de la mano y las interacciones que implican escenarios mano-objeto y mano-mano. Los métodos existentes a menudo simplifican estos retos, lo que limita su impacto en el mundo real. Para la interacción mano-objeto, el enfoque más general consiste en utilizar la simulación física para que las manos y los objetos interactúen de acuerdo con las leyes de la mecánica de contacto. Sin embargo, las diferencias de tamaño y morfología esquelética entre las representaciones de las manos en los simuladores y los dispositivos de seguimiento complican este proceso. La primera aportación de esta tesis es un modelo personalizado de mano blanda combinado con una estrategia de retargeting de poses, formulada como un problema de optimización, para conectar las representaciones de manos simuladas y del seguimiento. Este método integra las soluciones de seguimiento de la mano disponibles en el mercado con la simulación de la mano basada en la física sin necesidad de una representación común de la mano, pero permite la parametrización del modelo de la mano.

La interacción mano-objeto requiere el seguimiento de la mano en el mundo real para trasladar nuestros gestos a un escenario virtual. Este problema de realizar seguimiento de mano es un campo de investigación en auge con el potencial de proporcionar una interfaz natural para interactuar con entornos virtuales. Las soluciones habituales utilizan métodos de visión por ordenador, a menudo combinados con algoritmos de seguimiento basados en el aprendizaje, que pueden basarse en la profundidad o en imagenes RGB. Estos métodos proporcionan la morfología esquelética y la configuración de una mano que mejor se ajusta a la pose de la mano real del usuario, y algunos también estiman su silueta. Dada la ubicuidad de las cámaras RGB, la investigación se ha orientado hacia los métodos basados en imagenes RGB. A pesar de los avances recientes, el seguimiento 3D de dos manos que interactúan a partir de imágenes RGB sigue siendo un reto debido a problemas como la oclusión entre manos, la ambigüedad de profundidad, la segmentación de la mano y las

colisiones. Además, los métodos basados en el aprendizaje automático se enfrentan a dificultades de entrenamiento debido a la dificultad de obtener datos de entrenamien to suficientes y de alta calidad, lo que complica aún más el desarrollo de sistemas de seguimiento de manos robustos.

Para hacer frente a estos desafíos, proponemos el primer sistema que simula interacciones entre dos manos físicamente correctas con siluetas de manos personalizadas y apariencias diversas que genera datos sintéticos precisos. Este sistema es un componente principal de un algoritmo de última generación para el seguimiento de dos manos interactuando a partir de imágenes RGB. Además, abordamos los errores de profundidad que impiden la detección precisa del contacto mano a mano durante el seguimiento de dos manos interactuando mediante el desarrollo de un método basado en datos de imagen, formulado como un problema de traslación de imagen a imagen. Para entrenar nuestro método, introducimos un nuevo procedimiento para anotar automáticamente los contactos densos de superficie en secuencias de interacción entre dos manos. En consecuencia, nuestro método estima los contactos en el espacio de la cámara durante las interacciones, lo que puede integrarse en cualquier dispositivo de seguimiento de dos manos.

A.1 Antecedentes

Actualmente la RV ha alcanzado un alto grado de realismo visual, permitiendo la creación de experiencias virtuales verdaderamente inmersivas (Kaplan et al. 2021, Krichenbauer et al. 2018, Sun et al. 2018). A medida que los objetos virtuales parecen más realistas, el siguiente paso natural es interactuar con ellos (Chessa et al. 2019). Los seres humanos utilizan instintivamente ambas manos para interactuar con el entorno real y virtual, así como para gesticular y comunicarse. En consecuencia, muchas aplicaciones requieren la estimación simultánea de la pose de ambas manos mientras están en estrecha interacción, pero también cuando interactúan con objetos virtuales (Figura A.1). Sin embargo, esta acción aparentemente sencilla conlleva retos adicionales en la RV: el seguimiento simultáneo de ambas manos y la simulación de las interacciones mano-objeto, que suelen abordarse de forma independiente.

Interacción Mano-Objeto

La interacción mano-objeto suele basarse en simulaciones físicas para modelar interacciones basadas en la mecánica del contacto. Las soluciones modernas suelen integrar modelos básicos basados en la física sobre las estructuras de las manos seguidas para facilitar la interacción con objetos virtuales. Sin embargo, estos modelos sólo suelen admitir acciones básicas como pellizcar y agarrar, en las que la postura de la mano virtual permanece fija una vez detectado el agarre. Paralelamente, una rama específica de la investigación se centra en la simulación de manos basada en la física, con el objetivo de calcular configuraciones de la mano que satisfagan el equilibrio de fuerzas teniendo en cuenta aspectos como las representaciones de manos articuladas (Borst & Indugula 2005, Ott et al. 2010), el modelado geométrico de la piel (Duriez et al. 2008), la deformación local de la piel en los dedos (Jacobs & Froehlich 2011, Talvas et al. 2015), o la deformación completa de la piel (Garre et al. 2011, Hirota & Tagawa 2016). El método de Verschoor et al. (2018) formuló este problema como una tarea de optimización. Sin embargo, surge una observación crítica: los métodos actuales que incorporan un paso de simulación basado en la física para modelar la interacción mano-objeto (Hirota & Tagawa 2016, Kim & Park 2015, Verschoor et al. 2018) carecen de un componente esencial que dificulta su implementación en aplicaciones cotidianas de RV, la personalización de la forma de la mano.

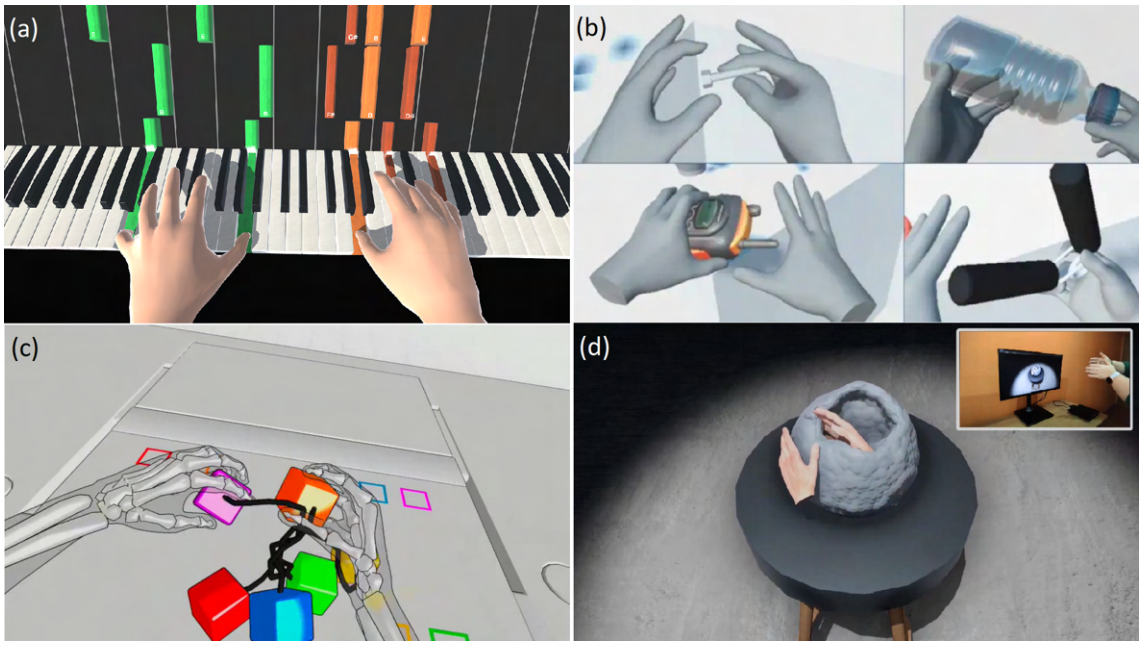


Figure A.1: *El seguimiento de las manos y la simulación basada en la física permiten una gran variedad de aplicaciones prácticas en realidad virtual. Desde tocar instrumentos hasta manipular herramientas y realizar tareas creativas como esculpir, estas tecnologías abren nuevas posibilidades de aprendizaje, formación y exploración inmersivos.* ***Fuente****: (a)* VRtuos *(2020), (b)* VirtualGrasp *(2022), (c)* Hand Physics Lab *(2021), (d) Barreiro et al. (2021).*

Seguimiento de Mano

La interacción mano-objeto requiere el seguimiento de la mano en el mundo real para trasladar nuestros gestos a un escenario virtual. Se trata de un campo de investigación en auge cuyo objetivo es proporcionar una interfaz natural para interactuar con entornos virtuales. Las soluciones habituales emplean métodos de visión por ordenador, a menudo combinados con algoritmos de seguimiento basados en el aprendizaje, que pueden basarse en la profundidad o en las imagenes RGB. Estos métodos generan la morfología esquelética y la configuración de una mano que mejor se ajusta a la mano real del usuario, y algunos también estiman la silueta de la mano. Los métodos modernos pueden clasificarse en tres grandes grupos. Los métodos generativos buscan una configuración de la mano que minimice una función de energía que defina la proximidad entre el modelo de mano actual y las características de la imagen (Qian et al. 2014, Sridhar et al. 2013, Tagliasacchi et al. 2015, Tkach et al. 2016). Los métodos discriminatorios infieren los parámetros de configuración de la mano a partir de imágenes de entrada utilizando algoritmos basados en el aprendizaje (Chen et al. 2022, Ge et al. 2019, Iqbal et al. 2018, Mueller et al. 2018, Park et al. 2022, Pavlakos et al. 2024, Xu et al. 2023, Yang et al. 2019). Los métodos híbridos combinan los puntos fuertes de los métodos generativos y discriminativos (Baek et al. 2019, Hampali et al. 2020, Mueller et al. 2019, Pavlakos et al. 2019, Xiang et al. 2019). Aunque recientemente se han logrado avances, el seguimiento 3D basado en RGB de dos manos que interactúan sigue siendo un reto debido a la oclusión entre manos, la ambigüedad de la profundidad, la segmentación de las manos y las colisiones. Además, los métodos basados en el aprendizaje automático se enfrentan a dificultades en el entrenamiento debido a la dificultad de

obtener datos de entrenamiento suficientes y de alta calidad, lo que complica el desarrollo de sistemas robustos.

Interacción Mano-Mano

Sostenemos que el uso de las manos como interfaz natural e inmersiva requiere métodos capaces de seguir las *dos manos en interacción* desde una única cámara RGB. Desafortunadamente, existen muy pocos trabajos que aborden el escenario de seguimiento de dos manos. Uno de los primeros intentos fue el de Oikonomidis et al. (2012), que utiliza una configuración multicámara para sortear las dificultades causadas por las inevitables y fuertes superposiciones. Algunos trabajos posteriores también utilizan nuevas configuraciones multivista (Han et al. 2020, 2018, Simon et al. 2017) y marcadores para facilitar el problema del seguimiento de dos manos, mientras que otros investigan el uso de sensores de profundidad individuales (Mueller et al. 2019, Taylor et al. 2017), lo que simplifica la configuración pero también proporciona pistas suficientes para resolver las ambigüedades de profundidad.

Algunos métodos recientes consideran una imagen RGB monocular como entrada. Moon et al. (2020) propuso un método que predice directamente las posiciones de las articulaciones de las manos en 3D a partir de imágenes RGB. Otros trabajos intentan modelar la ambigüedad de profundidad inherente modelando explícitamente un término de visibilidad (Kim et al. 2021), o utilizando modelos probabilísticos para la segmentación de la parte de la mano (Fan et al. 2021) o la pose 3D (Wang et al. 2022). Por último, el método de vanguardia de Li et al. (2022) utiliza una representación gráfica combinada con módulos de atención para inferir las posiciones de los vértices de dos manos que interactúan, y Yu et al. (2023) aprende características independientes para cada mano y explota la atención condicionada a la mano cruzada previa para mitigar las interdependencias. A pesar de los significativos avances en el seguimiento 3D de dos manos a partir de imágenes RGB, los errores residuales en la estimación de la profundidad siguen impidiendo el cálculo preciso de los contactos de las manos.

Detección de Contactos

Los métodos centrados en la estimación del contacto mano-objeto están estrechamente relacionados (Chen et al. 2023, He et al. 2021, Narasimhaswamy et al. 2020, Xie et al. 2022). Para abordar esta cuestión, se han desarrollado enfoques basados en datos que utilizan grandes conjuntos de datos anotados de manos que manipulan objetos rígidos (Liu et al. 2021, Shan et al. 2020, Taheri et al. 2020). Sin embargo, la mayoría de los métodos existentes estiman el contacto de la mano basándose en la información 3D de la escena o del objeto de interés (Cao et al. 2021, Grady et al. 2021, Jiang et al. 2021, Taheri et al. 2020).

A.2 Objetivos

El objetivo general de esta tesis es estudiar y diseñar soluciones eficaces para permitir una interacción inmersiva y natural en entornos virtuales. Para lograr esto, se abordan dos retos principales: resolver las discrepancias entre las dimensiones reales y virtuales que provocan deformaciones poco realistas, y diseñar procesos para generar datos físicamente correctos y precisos. Estos permitirán desarrollar dos métodos para permitir y mejorar las interacciones entre dos manos: uno para el seguimiento preciso de ambas manos simultáneamente en estrecha interacción y otro para la detección de los contactos resultantes.

Gestión de las Discrepancias entre Manos Reales y Virtuales

Para ofrecer una experiencia interactiva en el entorno que ofrece la tecnología de RV, uti-

lizamos por un lado, un simulador y por otro, un dispositivo de seguimiento para poder controlar la mano simulada. Sin embargo, esta tarea requiere la gestión de las discrepancias entre la mano real y la virtual, ya que estos dos elementos no comparten las mismas dimensiones ni las mismas morfologías de un modelo de mano a otro. Por lo tanto, esta gestión de las diferencias debe permitir la utilización de cualquier simulador y cualquier solución de seguimiento y, consecuentemente, ser independiente de los dos dispositivos y de las especificaciones de cada modelo de mano. Además, el método debe ser fiable, preciso y tener un rendimiento similar para poder ser utilizado en tiempo real.

Diseño de un Sistema de generación de Datos Físicamente Precisos

Para ofrecer una experiencia interactiva en entornos de realidad virtual utilizando métodos de seguimiento de dos manos basados en el aprendizaje, es fundamental contar con un dataset de entrenamiento de alta precisión y diverso. La capacidad de generalización de estos métodos depende en gran medida de la diversidad y calidad de los datos y anotaciones. Un conjunto variado de datos también ayuda a reducir sesgos, manejar casos aislados y considerar aspectos éticos importantes. La calidad de las anotaciones es crucial para garantizar la exactitud del seguimiento.

El proceso de recolección de datos y la anotación precisa son desafíos críticos, ya que cualquier error puede afectar significativamente el rendimiento del modelo. Por lo tanto, es necesario diseñar un método que no solo sea fiable y preciso, sino que también pueda ser escalable y adaptarse a diferentes condiciones y usuarios.

Diseño de un Método de Detección de los Contactos entre Manos

Para ofrecer una experiencia interactiva en entornos de realidad virtual, es esencial contar con un método robusto para estimar los contactos entre dos manos en interacción estrecha. Este método debe predecir con alta precisión los puntos de contacto, superando desafíos como la ambigüedad de la profundidad en imágenes RGB capturadas en entornos reales. En lugar de trabajar directamente con imágenes RGB, es beneficioso convertir estas imágenes en correspondencias de píxeles a superficie para evitar problemas relacionados con la iluminación, sombras y fondos complejos. Para desarrollar un método eficaz para la detección de contactos entre dos manos en interacción estrecha, es esencial contar con un conjunto de datos diverso y preciso, generado a partir de simuladores de manos avanzados.

A.3 Metodología

Para la realización de esta tesis, hemos seguido la siguiente metodología:

Revisión Bibliográfica

Para comprender el contexto de esta tesis, hemos realizado un exhaustivo análisis bibliográfico en el Capítulo 2, tratando el estado del arte por un lado del seguimiento de las manos y por otro de la simulación de las manos y los objetos. Abordamos los pilares fundamentales necesarios para la comprensión de esta tesis detallando la representación de la mano en el mundo virtual. Tras presentar los distintos modelos de simulación de manos y objetos, analizamos sus puntos fuertes y describimos el método que utilizamos para llevar a cabo nuestra investigación. A continuación, identificamos los distintos enfoques posibles para el seguimiento basado en la visión y centramos nuestros análisis en los métodos sin marcadores, en particular para el seguimiento simultáneo de ambas manos, que utilizan la visión por ordenador y ofrecen muchas ventajas para el propósito de esta tesis. Por último, también abordamos una línea ortogonal a nuestra investigación, la de la corporeidad, para tratar el contexto en su conjunto.

Estudio/Diseño y Desarrollo de un Modelo Intermediario para Restablecer las Discrepancias

En el capítulo 3, se detalla el diseño de una estrategia de retargeting de poses diseñada para cerrar la brecha entre los datos de la mano seguida y de los modelos de mano simulados. Nuestro enfoque es versátil y funciona eficazmente con cualquier método de seguimiento o simulación al actuar como un intermediario entre los dos. Utilizamos una representación intermedia de la mano que comparte el tamaño y la morfología de la mano simulada, mientras intenta igualar la configuración de la mano seguida. La estrategia de retargeting implica optimizar la pose de esta mano intermedia utilizando características que reflejan la pose de la mano seguida. Además, evaluamos el impacto práctico de la discrepancia entre las manos en la manipulación de objetos virtuales comparando nuestra estrategia de retargeting de poses con una copia simple de la pose de la mano. Para ello, realizamos un estudio con usuarios que evalúa el desempeño en tareas de manipulación de objetos virtuales. El estudio confirma que, aunque la discrepancia en la representación de la mano no es crítica para la manipulación de objetos grandes, es crucial para la manipulación fina de objetos pequeños.

Diseño de un Sistema para la Generación de Datos Anotados Físicamente precisos de Interacciones Mano a Mano

En el capítulo 4, se presenta un nuevo sistema para generar datos fotorrealistas y anotaciones físicamente precisas de interacciones entre dos manos, considerando la variabilidad en términos de silueta y apariencia. Estas condiciones son difíciles de replicar en situaciones reales para poses altamente complejas. Para lograr esto, extendemos el modelo paramétrico basado en la superficie de MANO (Romero et al. 2017) a una representación volumétrica, la cual se integra en el simulador descrito en el capítulo 2. Esto nos permite sintetizar secuencias de movimientos complejas de la mano basadas en grabaciones previas, generando así posiciones de puntos clave en 2D y heatmaps, imágenes de correspondencia densa, máscaras de segmentación, mapas de profundidad relativa intra-mano, y mapas de distancia relativa inter-mano, con identidades de sujetos variables.

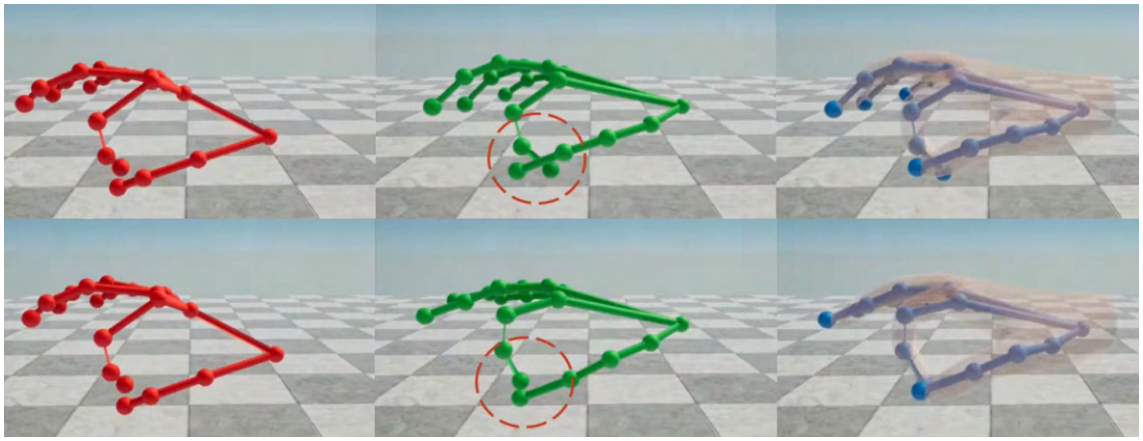
Este enfoque nos permite crear un conjunto de datos sintéticos que se combina con un conjunto de datos reales. Los datos sintéticos, perfectamente anotados gracias a nuestro sistema, mitigan la influencia del ruido presente en los datos reales. Este conjunto de datos combinado se utilizó para entrenar un predictor de red neuronal multitarea, uno de los componentes clave del primer método diseñado específicamente para seguir y reconstruir dos manos en interacción en 3D global utilizando únicamente imágenes RGB.

Estudio y Desarrollo de un Modelo Eficiente para la Deteccion de Contactos entre Ambas Manos

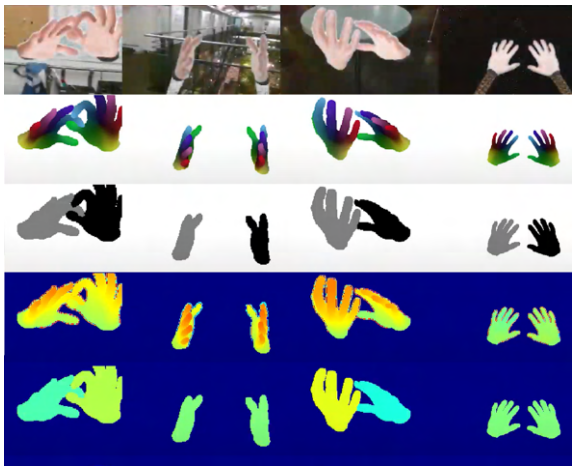
En el capítulo 5, proponemos un método basado en imágenes y datos para estimar el contacto en interacciones mano a mano. Nuestro método se basa en dispositivos de seguimiento de manos 3D que predicen la pose articulada de dos manos, enriquecidos con mapas de probabilidad de contacto en el espacio de la cámara. Para entrenar nuestro método, primero alimentamos datos de captura de movimiento de manos interactuando en un simulador de manos basado en física y calculamos puntos de contacto 3D densos. Luego, renderizamos estos mapas de contacto desde varios puntos de vista y creamos un conjunto de datos de pares de imágenes de píxel a superficie de las manos y sus etiquetas de contacto correspondientes. Finalmente, entrenamos una red de imagen a imagen que aprende a traducir correspondencias de píxel a superficie en mapas de contacto. En el tiempo de inferencia, estimamos las correspondencias de píxel a superficie utilizando un método de seguimiento

de manos de última generación y luego usamos nuestra red para predecir contactos precisos entre manos. Validamos cualitativamente y cuantitativamente nuestro método en datos del mundo real y demostramos que nuestras predicciones de contacto son más precisas que los métodos de seguimiento de manos más avanzados.

A.4 Resultados



(a) Capítulo 3: Manipulación Virtual Fina con Manos de Distintos Tamaños.



(b) Capítulo 4: Generación de Datos Anotados de Interacciones Mano a Mano Físicamente Precisas.

(c) Capítulo 5: Contacto Mano a Mano a partir de Imágenes RGB.

Figure A.2: *Visión general de los distintos métodos implementados en esta tesis para abordar los retos mencionados en la sección anterior.*

Las contribuciones principales de esta tesis pueden ser resumidas de la siguiente manera:

- Una estrategia de retargeting de poses para conectar la mano seguida y la mano simulada. Nuestro enfoque funciona con cualquier tipo de método de seguimiento o simulación, ya que se sitúa en la interfaz entre ambas tareas. Utilizamos una representación intermedia de la mano que comparte el tamaño y la morfología de la mano simulada, pero que intenta ajustarse a la configuración de la mano seguida. La estrategia de retargeting formula una optimización de la pose de esta mano intermedia, basada en características que representan la pose de la mano seguida. (Capítulo 3)

- Una evaluación del impacto práctico del desajuste de la mano en la manipulación de objetos virtuales, comparando nuestra estrategia de retargeting de la pose frente a la copia ingenua de la pose de la mano. Para ello, hemos llevado a cabo un estudio de usuarios que se asemeja a la realización de tareas de manipulación de objetos virtuales. El estudio confirma que el desajuste de la representación de la mano no es crítico para la manipulación grosera (objetos grandes), pero sí para la manipulación fina (objetos pequeños). (Capítulo 3)
- Un nuevo sistema de generación de datos sintéticos físicamente precisos, capaz de tener en cuenta las manos en interacción con diferentes identidades, tanto en términos de silueta como de aspecto. Este trabajo condujo al desarrollo del primer método monocular basado en imagenes RGB para la captura de movimiento 3D de dos manos que interactúan estrechamente, que estima simultáneamente la pose y la figura de la mano, mientras se ejecuta en tiempo real. (Capítulo 4)
- Hasta donde sabemos, se trata del primer método basado en imágenes para estimar los contactos entre manos a partir de una única imagen RGB. Nuestro método se basa en las soluciones de seguimiento a dos manos ya existentes, enriqueciéndolas con un mapa de probabilidad de contacto de manos en el espacio de la cámara que ofrece muchas ventajas: i) permite la detección del contacto incluso cuando el seguimiento 3D es impreciso, ii) hace que nuestra solución sea compatible con cualquier sistema de seguimiento a dos manos ya existente (tanto métodos basados en la profundidad como métodos basados en imagenes RGB), iii) puede utilizarse potencialmente como un nuevo término en los métodos basados en la optimización para el seguimiento a dos manos. (Capítulo 5)
- Un nuevo cojunto de procesos y sistemas para detectar y anotar automáticamente contactos de superficie densa por vértice en secuencias de interacción de manos en el mundo real. (Capítulo 5)

Además, los resultados de esta tesis se encuentran presentados en las siguientes publicaciones:

- Sorli et al. (2021) "**Fine Virtual Manipulation with Hands of Different Sizes.**" - Suzanne Sorli, Dan Casas, Mickeal Verschoor, Ana Tajadura-Jiménez, Miguel A. Otaduy - En: Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR), 2021, CORE: A*.
- Wang et al. (2020) "**RGB2Hands: Real-Time seguimiento of 3D Hand Interactions from Monocular RGB Video.**" - Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, Christian Theobalt - En: ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia), 2020, JCR Q1.
  El autor principal de este trabajo es Jiayi Wang, que diseñó la formulación de ajuste del modelo generativo, así como la red neuronal convolucional multitarea, mientras que yo fui el principal contribuyente del sistema de generación de datos sintéticos físicamente correctos utilizado para entrenar el predictor de aprendizaje automático. El trabajo se incluye en la tesis para complementar los métodos de seguimiento de ambas manos interactuando.
- "**Hand-to-hand Contact from RGB Images.**" - Suzanne Sorli, Marc Comino-Trinidad, Dan Casas - En: En proceso de revisión, 2024.

A.5 Conclusiones

Globalmente, esta tesis pretende mejorar las interacciones mano-objeto y mano-mano para ofrecer experiencias e interfaces inmersivas e intuitivas. Logramos este objetivo proponiendo métodos a medida para cada escenario. En primer lugar, desarrollamos un método para reorientar las poses de las manos entre manos de distintos tamaños y morfologías esqueléticas, mejorando las interacciones mano-objeto. En segundo lugar, presentamos un nuevo sistema para generar datos de interacción a dos manos físicamente precisos, creando un conjunto de datos de entrenamiento para desarrollar el primer método de seguimiento en tiempo real capaz de reconstruir dos manos interactuando a partir de vídeo monocular RGB. Además, introdujimos un enfoque basado en imágenes para estimar los contactos mano-mano, mejorando la calidad general de las interacciones mano-mano más allá de lo que los actuales métodos de seguimiento de dos manos basados en RGB pueden lograr debido a errores en las estimaciones de profundidad, forma o pose..

El capítulo 3 está dedicado a las interacciones mano-objeto y aborda la brecha existente entre los datos de la mano seguida y los del modelo de la mano simulada mediante una estrategia de retargeting de poses que evita la necesidad de compartir una representación común de la mano. Nuestro enfoque es versátil y funciona eficazmente con cualquier método de seguimiento o simulación actuando como intermediario entre ambos. Además, evaluamos las implicaciones prácticas de las discrepancias en la representación de la mano en la manipulación de objetos virtuales comparando nuestra estrategia de reorientación de la pose con un enfoque ingenuo de copia de la pose de la mano que revela que, mientras que la falta de coincidencia en la representación de la mano tiene un impacto mínimo en la manipulación de objetos más grandes, afecta notablemente a la precisión necesaria para manipular objetos más pequeños. El método propuesto culminó con una presentación bajo el título *Fine Virtual Manipulation with Hands of Different Sizes*. (Sorli et al. 2021) en la Conferencia **2021 IEEE International Symposium on Mixed and Augmented Reality** (**ISMAR**), clasificada A* en el sistema de clasificación de conferencias CORE.

Los capítulos 4 y 5, en cambio, se centran en escenarios en los que interactúan dos manos. Las cámaras RGB son omnipresentes, pero no proporcionan ninguna indicación de profundidad, lo que supone un obstáculo importante en este tipo de situaciones. En el capítulo 4, desarrollamos un sistema capaz de generar datos precisos sobre secuencias en las que intervienen dos manos en estrecha interacción, que incluyen la profundidad intra-mano y la distancia inter-mano, para el entrenamiento. Este fue un componente clave en el diseño de un método pionero en tiempo real para capturar tanto la pose esquelética como la geometría de la superficie 3D de las manos desde una única cámara RGB, considerando explícitamente las interacciones cercanas. Se ha demostrado que este método supera a otros anteriores basados únicamente en RGB en situaciones complejas de interacción a dos manos. Sin embargo, en algunos casos, sigue padeciendo errores residuales en la estimación de la profundidad, la silueta o la pose de la mano, lo que impide la detección precisa de los contactos mano a mano a pesar de los prometedores resultados y sigue representando un problema actual. En el capítulo 5, presentamos un método basado en datos de imagen para estimar el contacto en interacciones mano-mano. En el momento de la inferencia, estimamos las correspondencias píxel-superficie utilizando el seguimiento de manos más avanzado y, a continuación, utilizamos nuestra red para predecir con precisión el contacto mano-mano. Validamos ampliamente nuestro método con análisis cualitativos y cuantitativos de datos reales, demostrando su mayor precisión que los métodos actuales de seguimiento de manos.

Esta investigación culminó con una presentación en una conferencia y una publicación

en la revista **ACM Transactions on Graphics** (**TOG**) (JCR Q1) como cuarto autor, con el título *RGB2Hands: Real-Time seguimiento of 3D Hand Interactions from Monocular RGB Video* (Wang et al. 2020), y un artículo actualmente en revisión titulado *Hand-to-hand Contact from RGB Images*.

En esta tesis, hemos introducido varios métodos y enfoques innovadores que han supuesto avances significativos en el campo de las interacciones mano-objeto y mano-mano en RV y RA. Nuestras contribuciones se han demostrado mediante experimentos rigurosos y aplicaciones prácticas, mejorando el realismo y la inmersión de las experiencias de RV. Las soluciones propuestas abordan retos críticos en los escenarios mano-objeto y a dos manos, proporcionando un modelo de interacción más natural e intuitivo que beneficiará tanto a las aplicaciones comerciales como a la investigación. A medida que las tecnologías de RV y RA sigan evolucionando y ganando popularidad, la demanda de modelos de interacción sofisticados no dejará de crecer, y estamos seguros de que nuestras soluciones serán esenciales para satisfacer esta necesidad.

Somos optimistas en cuanto a que los avances en RA y RV pronto harán que estas tecnologías se generalicen e integren en la vida cotidiana. Las aportaciones realizadas en esta tesis contribuyen a su adopción más generalizada al mejorar la usabilidad y accesibilidad de estas tecnologías. Mediante el desarrollo de interfaces más prácticas e intuitivas, pretendemos eliminar la necesidad de sensores y dispositivos engorrosos. Combinados con las continuas mejoras de la tecnología y los algoritmos de aprendizaje automático, estos progresos potenciarán notablemente las experiencias propuestas. Esperamos que estos avances atraigan a un público más amplio, poco familiarizado con este mundo virtual, e inspiren una pasión por él similar a la mía.

Un importante paso hacia adelante consistiría en dar prioridad a la diversidad de las manos y acomodar a las personas con discapacidad. Aunque el seguimiento actual de las manos y los nuevos modelos de interfaz ofrecen soluciones esperanzadoras, el acceso a esta tecnología sigue siendo problemático para muchos.