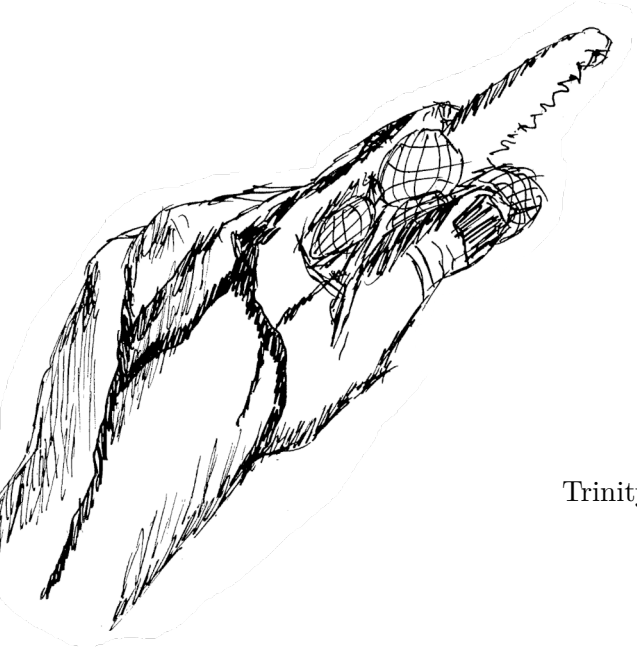
# Designing Hand-Object Interaction Representations For Better Grasp Priors

by

Théo Morales





**Dissertation**

Presented to the

Trinity College Dublin, the University of Dublin

in fulfillment

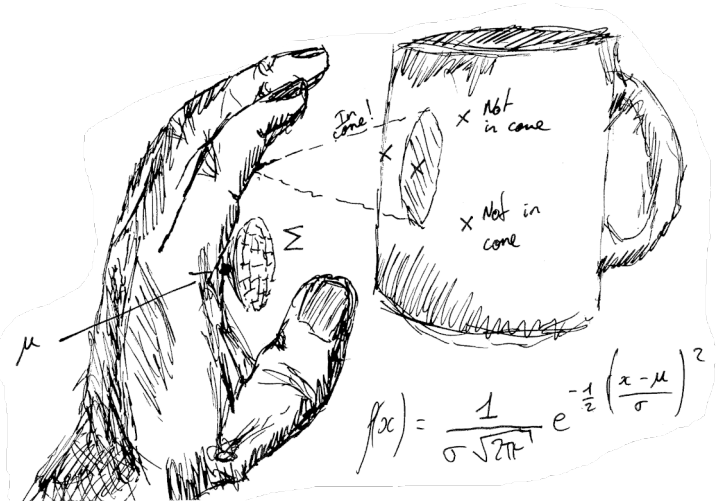
of the requirements

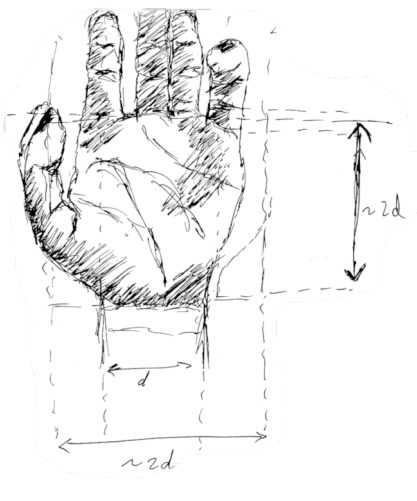
for the Degree of

**Doctor of Philosophy in Computer Science**

**Trinity College Dublin, the University of Dublin**

April 2025



# Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I, the undersigned, agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).



Théo Morales

March 20, 2025

## Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.



Théo Morales

March 20, 2025

I dedicate this thesis to my 10 years old self.

# Acknowledgments

I must start by thanking my advisor Dr. Gerry Lacey, who made me more pragmatic, thoughtful, careful and overall shaped me into an independent researcher. I am very grateful for his guidance throughout my PhD, for his advice and wisdom. Gerry made me a writer, an orator and a scientist. I am glad to have pursued my PhD with him. I also thank the d-real programme for giving me this opportunity, and in particular Dr. Stephen Carroll, Programme Manager, for being so nice to all of us and for being a thoughtful person. Stephen was a guide during these four years, his help is precious and I am grateful. I now thank my friends and colleagues, Dr. D, B, S and X in particular, who have brought me joy and comfort at all times, especially in the hard times. Thank you to all my friends in Stack B, and to all my friends outside who gave me their unconditional support. Thank you Omid and Sid for your guidance and friendship, it was a pleasure working with you. I finally thank my family and my parents in particular for giving me the resources that allowed me to pursue my dreams, and I thank my brother and sister for their support and encouragements. The final thanks go to my uncle who always pushed me to aim for greatness. He was a role model who truly inspired me from a young age, and I am here largely thanks to him. When he first put a computer in my hands, I discovered my true passion before the age of 10.

Théo Morales

*Trinity College Dublin, the University of Dublin*
*April 2025*

# Abstract

Accurate modelling of hand-object interactions (HOI) is critical for applications in areas such as computer vision, augmented/mixed/virtual reality, video games production and robotics. In computer vision, HOI modelling encompasses tasks like hand-object pose estimation, grasp synthesis, object manipulation and mesh reconstruction. These challenges share many characteristics but early HOI research focused on end-to-end methods tailored to specific tasks. These approaches suffered from limited transferability and generalization due to reliance on spurious correlations in the data. Recent research favours explicit representations that capture key features of hand grasps and object affordances, offering better generalization and transferability. Through the use of local coordinate frames, dense hand-object correspondences and contact mapping, explicit representations have shown promise in improving the accuracy of HOI tasks. However, designing representations that are both expressive and lightweight remains challenging, balancing the need for real-time applicability with model generalizability.

This thesis investigates the following research questions: (1) Can we design new shape, pose and contact representations in order to overcome the limitations of existing representations while maintaining their expressiveness? and (2) Can we design new models to demonstrate the expressiveness and versatility of these representations for various applications? These questions are explored within the scope of static and dynamic grasp denoising and synthesis, from the lens of deep learning for 3D vision. The focus of this research is on the expressiveness, versatility and efficiency of the proposed solutions when combined with state-of-the-art (SOTA) learning methods and test-time adaptation, as motivated by effective recent trends in the field.

This thesis answers the research questions via three main contributions. Firstly, we investigate

whether generalizable internal representations can be learned via meta-learning in a new paradigm for hand-object pose prediction. Secondly, we design a novel lightweight and fully-differentiable hand-object interaction field that captures shape and pose in an object-centric frame. We show how this representation can be used to denoise grasps with 8% higher accuracy than the state-of-the-art in static grasps. Thirdly, we extend the representation to model hand surface contacts with a parametric and continuous probabilistic representation of contacts. We demonstrate how this improved representation can be leveraged for synthesis and denoising of static grasps, outperforming specialized state-of-the-art methods in contact, pose and plausibility metrics, while achieving both tasks. This combined representation improves the SOTA, as it is continuous, fully-differentiable and approximately 70× faster to compute.

# Contents

**Acknowledgments** **v**

**Abstract** **vi**

**List of Tables** **xii**

**List of Figures** **xiii**

**List of Publications** **xv**

**Chapter 1 Introduction** **1**

1.1 Problem statement . . . . . 2

1.1.1 Challenges in hand-object interaction modelling . . . . . 2

1.1.2 The need for better hand-object interaction representations . . . . . 6

1.1.3 Limitations of recent representations . . . . . 6

1.2 Research questions . . . . . 7

1.3 Contributions . . . . . 8

**Chapter 2 Background** **9**

2.1 Mathematical notation . . . . . 9

2.2 Information Theory and Entropy . . . . . 9

2.3 The Kullback-Leibler Divergence . . . . . 11

2.4 Latent Variable Models . . . . . 13

2.4.1 Variational Inference . . . . . 16

2.4.2 Auto-encoding Variational Bayes . . . 17
2.4.3 Denoising Diffusion Probabilistic Models . . . 18
2.5 3D representations . . . 21
2.5.1 Explicit representations . . . 22
2.5.2 Implicit representations . . . 23
2.6 Overview of the literature in hand-object interaction . . . 26
2.6.1 Benchmarks and datasets . . . 26
2.6.2 Hand and object pose prediction . . . 28
2.6.3 Hand and object reconstruction . . . 31
2.6.4 Generating hand-object interactions . . . 32
2.6.5 Interaction representation . . . 35
2.6.6 Test-time adaptation . . . 37
2.6.7 Closing remarks . . . 38
2.7 Guidelines for hand-object interaction modelling . . . 38
2.8 Evaluation metrics . . . 40

**Chapter 3 Meta-learning grasp predictors** **42**
3.1 Background . . . 43
3.1.1 Optimization-based meta-learning . . . 44
3.2 Benchmarking grasp prediction through the lens of group distribution shifts . . . 46
3.2.1 Evaluating the effectiveness of meta-learning . . . 47
3.2.2 Empirical analysis . . . 55
3.2.3 Discussions . . . 58

**Chapter 4 CHOIR: Representing shape and pose** **60**
4.1 Motivation . . . 61
4.1.1 Related works and their limitations . . . 61
4.1.2 A tightly coupled, continuous and fully-differentiable representation in canonical frame . . . 66
4.2 Designing and learning CHOIR . . . 67

4.2.1 Explicit shape-pose representation as unsigned distances . . . 67
4.2.2 Learning a CHOIR distribution for grasp denoising . . . 70
4.2.3 From a representation to a hand . . . 73
4.3 Evaluation . . . 74
4.3.1 Quantitative & qualitative results . . . 75
4.4 Discussions and limitations . . . 78

**Chapter 5 $p$-CHOIR: Representing contacts 81**
5.1 Motivation . . . 82
5.1.1 Shape and pose are not enough . . . 82
5.1.2 Limitations of related works . . . 83
5.1.3 A hand-centric, localized, continuous and probabilistic contact representation . 87
5.2 $p$-CHOIR: Adding probabilistic contacts to CHOIR . . . 87
5.2.1 Parameterized representation of explicit contacts . . . 88
5.2.2 Learning a $p$-CHOIR distribution for grasp denoising and synthesis . . . 90
5.2.3 Aligning a hand mesh to the representation . . . 92
5.3 Evaluation . . . 94
5.3.1 Qualitative & quantitative results . . . 96
5.3.2 Multimodal model: grasp refinement & synthesis . . . 106
5.3.3 Ablation study . . . 107
5.3.4 Performance costs . . . 109
5.3.5 Limitations . . . 109

**Chapter 6 Reflections, limitations and conclusions 111**
6.1 Reflections on $p$-CHOIR . . . 112
6.1.1 Representing the shape and pose is not enough . . . 112
6.1.2 Contact representation alone is not enough . . . 113
6.1.3 A holistic representation is the way forward . . . 113
6.2 Limitations of $p$-CHOIR . . . 115
6.3 Future work . . . 117

6.4 Closing statement . . . . . 119

**Bibliography** **120**

# List of Tables


2.1 Review of static grasp datasets . . . . . 27
2.2 Review of dynamic manipulation datasets . . . . . 28

3.1 Image samples per split as the number of unseen objects grows in the test split . . . . 48
3.2 Distance heat map of mean hand shapes . . . . . 50
3.3 Average gradient norm per adaptation step: exp. 1 vs exp. 2 . . . . . 57

4.1 Anchor assignment comparison for CHOIR . . . . . 69
4.2 Quantitative evaluation of AggVED on the Perturbed ContactPose benchmark . . . . 76
4.3 Quantitative evaluation of AggVED on the Perturbed GRAB benchmark . . . . . 78

5.1 Qualitative evaluation of *JointDiffusion* on the Perturbed ContactPose benchmark . . 96
5.2 Quantitative evaluation of our approach on object splits . . . . . 96
5.3 Quantitative evaluation of *JointDiffusion* in grasp synthesis . . . . . 98
5.4 Evaluation of our multimodal model on static grasp generation . . . . . 106
5.5 Evaluation of our multimodal approach on static grasp refinement . . . . . 106
5.6 Ablation study of *p*-CHOIR . . . . . 109

# List of Figures


1.1 OpenPose hand model . . . . . 3
1.2 Examples of physically plausible and implausible grasps. . . . . 4
1.3 Data manifold illustration . . . . . 5

2.1 The forward process of the DDPM framework . . . . . 19
2.2 Mesh representation. . . . . 22
2.3 Voxel representation. . . . . 22
2.4 Point cloud representation. . . . . 23
2.5 Signed Distance Field representation . . . . . 24
2.6 Examples of recent HOI representations . . . . . 35

3.1 Illustration of training versus testing data distributions in object instances. . . . . 42
3.2 Example of tasks built from the DexYCB dataset . . . . . 49
3.3 An overview of the meta-learning system . . . . . 51
3.4 Relative Mean Per-Joint Pose Error (RMPJPE) & Mean Per-Corner Pose Error (RMPCPE) as functions of the imbalance level . . . . . 52
3.5 Relative Mean Per-Joint Pose Error (RMPJPE) as function of the test split for experiment 1 with 3 training set sizes . . . . . 53
3.6 Visualisation of weights per layer . . . . . 59

4.1 Decomposition of our Coarse Hand-Object Interaction Representation . . . . . 60
4.2 Limitations of purely object-centric representations . . . . . 64

4.3 Architecture of AggVED . . . . . 70
4.4 Qualitative comparison of AggVED with SOTA methods . . . . . 75
4.5 Quantitative evaluation of AggVED in pose error . . . . . 76
4.6 Evaluation of the plateau point of AggVED (dynamic *vs.* static models) . . . . . 77
4.7 Qualitative comparison of $\mu$AggVED and ContactOpt [1] . . . . . 80

5.1 Decomposition of our Probabilistic Coarse Hand-Object Interaction Representation . . 81
5.2 GraspTTA samples . . . . . 85
5.3 Illustration of the contact cone of tolerance . . . . . 88
5.4 Visualization of our probabilistic contact maps . . . . . 89
5.5 Architecture of *JointDiffusion* . . . . . 91
5.6 Qualitative evaluation of *JointDiffusion* on the Perturbed ContactPose benchmark . . 97
5.7 Qualitative evaluation of *JointDiffusion* on grasp synthesis . . . . . 100
5.8 Visualization of generated samples from *JointDiffusion* . . . . . 101
5.9 Failure cases on a comparison of *JointDiffusion* and ContactOpt for the Perturbed ContactPose benchmark . . . . . 102
5.10 Qualitative comparison of *JointDiffusion vs.* ContactOpt on the Perturbed ContactPose benchmark (1) . . . . . 103
5.11 Qualitative comparison of *JointDiffusion vs.* ContactOpt on the Perturbed ContactPose benchmark (2) . . . . . 104
5.12 Qualitative comparison of *JointDiffusion vs.* ContactOpt on the Perturbed ContactPose benchmark (3) . . . . . 105
5.13 Qualitative evaluation of the multimodal *JointDiffusion* on grasn synthesis . . . . . 107
5.14 Qualitative evaluation of the multimodal *JointDiffusion* on grasp refinement . . . . . 108

6.1 Limitations of the BPS representation . . . . . 116

# List of Publications

The following peer-reviewed publications are directly related to the research questions:

- T. Morales and G. Lacey, "**A new benchmark for group distribution shifts in hand grasp regression for object manipulation. Can meta-learning raise the bar?**", *Workshop on Distribution Shifts, NeurIPS 2022.*
- T. Morales, O. Taheri, G. Lacey, "**A Versatile and Differentiable Hand-Object Interaction Representation**", *Algorithm track, WACV 2025.*

Indirectly related to the research questions, the following peer-reviewed manuscripts were published during this research project:

- S. Ravi, P. Climent-Perez, T. Morales, C. Huesca-Spairani, K. Hashemifard, and F. Florez-Revuelta "**ODIN: An OmniDirectional INdoor dataset capturing Activities of Daily Living from multiple synchronized modalities**", *Workshop on Omnidirectional Computer Vision, CVPR 2022*

# Chapter 1

# Introduction

We interact with 140 objects per day on average [2]. There are many factors that make object manipulation a difficult task. While we take for granted the ability to grasp objects in our daily lives, this is not an innate ability. Although hands are highly dexterous and versatile, toddlers still struggle to hold objects properly. As we grow, we learn to manipulate objects through experience and observation. This experience allows us to manipulate previously unseen objects that resemble others or have the same utility. That is because we can reason about where to place our fingers on the object, depending on its shape and use. This *hand placement specific to an object* results in a taxonomy of hand grasps that we call *grasp affordances*. Object grasping is the exercise of carefully articulating the fingers in a way that is compatible with the object, and that makes sense from a functional, as well as physical, perspective. All in all, object grasping is a complex task which involves mainly two aspects. The first is an understanding of objects and their affordances, such that certain grasps can be easily applied to other shapes with the same use intent. The second is self-awareness of the hand's capabilities and limitations. From a computer algorithm's perspective, modelling hand-object interactions is a challenging task that requires careful consideration of both aspects.

Building algorithms to accurately model these hand-object interactions would open the door to a plethora of graphics and vision tasks: augmented/mixed/virtual reality, video games production, robotics manipulation, human-computer interaction, etc. Modelling the core mechanical, visual and functional characteristics of HOI allows all these challenges to be approached with the same foundations. Solidifying these foundations is thus a sensible approach to unifying the field of HOI and, as a

result, boosting performance across all tasks. Before solving the problem, we must define it by eliciting the underlying challenges of HOI and the limitations of current methods (Section 1.1). This paves the way for the research questions (Section 1.2) and the contributions of this thesis (Section 1.3).

## 1.1 Problem statement

Learning an accurate model of 3D hand-object interaction is an unsolved problem. While SOTA methods achieve impressive results across various tasks (*e.g.* pose prediction, 3D reconstruction, grasp generation, etc.), they still present failure cases for certain objects that require dexterous manipulation: *i.e.* a cup is easier to grasp than scissors. The field's fragmentation is due to the lack of a common representation for hand-object interaction. Designing or learning a versatile, rich, and unified representation would allow the community to improve the model's performance and generalization. There are several aspects characterizing this representation problem, and in this section, we investigate it from two perspectives.

### 1.1.1 Challenges in hand-object interaction modelling

Understanding and modelling hand-object interaction is challenging due to several factors. Firstly, hands are the sum of small parts, dexterous, and with many degrees of freedom which often creates self-occlusions. Their uniform texture and symmetry adds uncertainty to the observer and make a challenging case for texture-based computer vision algorithms: these are core challenges for image-based 3D tracking or reconstructing methods. Additionally, hand-object interactions naturally come with object occlusions, making it harder to estimate accurate hand pose and finger placement. Our natural intelligence allows us to tackle these challenges effortlessly. One explanation for this is that we build a mental model of grasping through experience acquired by watching our peers grasp objects, and by doing so ourselves. It allows us to predict and validate the plausibility of a partially observed grasp, thus reducing uncertainty in our predictions by a large margin compared to taking a random guess. This is the essence of the problem: what is the best way to mimic this in computer vision algorithms? To answer this question, we first need to introduce the paradigm of end-to-end deep learning methods in HOI tasks.

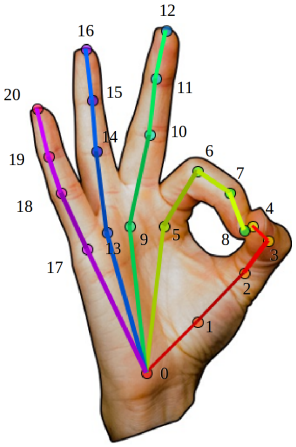
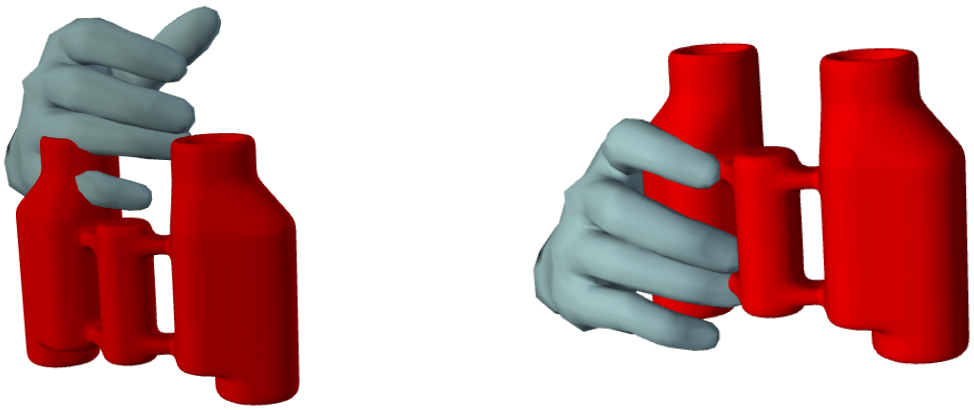



Figure 1.1: The 21 hand joints used in the OpenPose hand model. Source: OpenPose repository [3].

In the context of HOI modelling with deep learning, we refer to *end-to-end* as the paradigm of making predictions from raw inputs in a direct manner. For instance, where we would want to predict the 3D pose of a hand from an image, an end-to-end method makes a 3D prediction from the image only. This paradigm boils down to learning a direct mapping from raw inputs to outputs, without using intermediate representations and fitting. Originally, modelling hand-object grasping is treated as learning a mapping from one space to the other, such as images to 3D hand-object poses [4, 5, 6], videos to motions [7], or object meshes to hand poses [8, 1, 9]. From a computer vision perspective, the challenges are further exacerbated by the transition from 2D to 3D. This transition is particularly challenging when dealing with occlusions. The crux of the problem lies in mapping the space of possible hand grasps and finding the one that best fits the observation. In the grasp synthesis scenario, we are searching for a hand grasp that best fits the object, according to the training data distribution. Deep learning has shown promise in learning complex high-dimensional mappings from images to 3D space [10, 4, 11, 3, 12], but requires large amounts of data to fit such mappings. While deep learning excels in many high-dimensional tasks [13, 14], it is effectively always extrapolating either in data space or in embedding space [15]. All in all, the consensus on generalization in deep learning is that it is not truly attainable. The only effective way to generalize at test time is to cover the data distribution at training time. This is why large datasets of varied examples are crucial for deep learning models to perform well.



(a) Physically implausible grasp. (b) Physically plausible grasp.

Figure 1.2: Examples of physically plausible and implausible grasps. The hand in a is penetrating the object and the grasp is not stable. In b, the hand is stably grasping the object with no penetrations and sufficient contacts.

Due to the many degrees of freedom of the hand, the space of possible hand poses is vast. If we model the hand as the position of its joints in 3D space (see Fig. 1.1), the hand pose space is

$$\mathcal{P}_H = (\mathbb{R}^3)^{21} = \mathbb{R}^{63}, \tag{1.1}$$

thus one singular pose can be represented as a 63-dimensional vector. When considering object manipulation, we typically represent the object as a 3D point cloud of $N$ points, although other methods can be used (*i.e.* neural embeddings or signed distance functions). As such, the hand-object pair lives in a pose space defined as the Cartesian space of both hand space $\mathcal{P}_H$ and object space $\mathcal{P}_O$:

$$\begin{aligned} \mathcal{P}_O &= (\mathbb{R}^3)^N, \\ \mathcal{P}_{HO} &= \mathcal{P}_H \times \mathcal{P}_O = \mathbb{R}^{63+3N}, \end{aligned} \tag{1.2}$$

such that the dimensionality grows linearly with the number of object points $N$.

The main challenge of HOI modelling is revealed when considering the problem as a search in the pose space $\mathcal{P}_{HO}$ for the best hand-object pose that satisfies the laws of physics: no hand-object penetrations, stable grasp with sufficient contact, hand joints that comply with the biomechanical

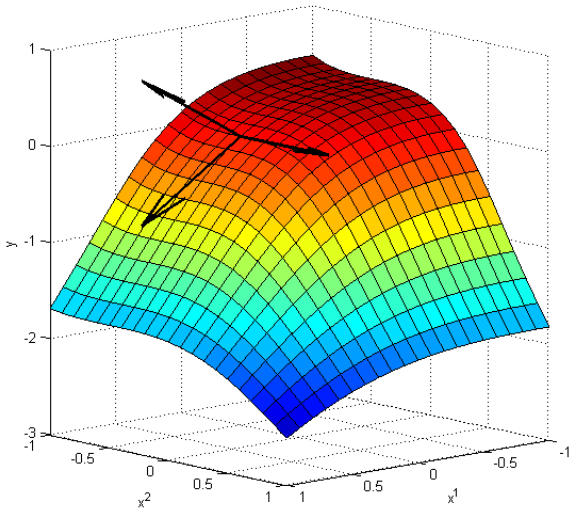



Figure 1.3: Illustration of a 2D manifold embedded in 3D space (source: [16]).

constraints of the hand, etc (see Fig. 1.2). All these factors constrain the pose space into a narrow "natural grasps" manifold which lives in a much larger "valid hand pose" manifold. This is the direct equivalent to natural images which are on a manifold embedded in the high-dimensional space of all possible images, as illustrated in Fig. 1.3. Hand poses that fall slightly outside the natural data manifold result in inaccuracies like subtle hand-object penetrations or slightly off-positioned fingers, which can significantly affect the realism and physical stability of the grasps. Effectively, the pose space is heavily tied to the object geometry and pose, resulting in a much higher dimensional space $\mathcal{P}_{HO}$ than the hand pose space $\mathcal{P}_H$ alone. This makes the problem of HOI even more challenging than hand pose prediction alone, hence the training data must be diverse in terms of hand poses, object shapes, and viewpoints.

Previous research in HOI focused on end-to-end methods tailored to specific downstream tasks, such as joint hand-object pose prediction or grasp synthesis. The goal of these methods is to perform the search in the pose space as a single step. They accomplish this by learning the high-dimensional mapping from training data as black-box functions using deep learning. While they have gradually improved the state-of-the-art on benchmarks, they are limited in several aspects. This thesis advocates for an alternative paradigm where the input and output of the learning method expose the essence of

hand-object interaction.

### 1.1.2 The need for better hand-object interaction representations

End-to-end methods, while effective on benchmarks, typically struggle to generalize to unknown objects [17]. This may be explained by the fact that the data distribution can be categorized in groups corresponding to object categories due to their tight relationship with the hand pose space. As a result, HOI models must deal with group distribution shifts that arise in the light of unseen object categories at test time. This hints at a deeper problem in deep learning: the reliance on spurious correlations in the training data – misleading cues in input-label pairs – to achieve good performance on the average case [18, 19, 20]. There is some evidence that feature selection or learning can help reduce the reliance on spurious correlations to improve generalization [21, 22]. In extensive experiments on contact modelling [23], it is shown that carefully selected input features – such as encoding the delta vectors from object points to the nearest hand mesh vertices – can improve the accuracy of the model. This observation is found across other domains in machine learning, such as speech recognition with Mel spectrograms [24] or action understanding with human pose keypoints [25], where feature selection is a crucial step in the model design [26].

We recall the question: what is the best way to mimic human grasp reasoning in computer vision algorithms? While an answer to this question may not be found, ongoing efforts in computer vision for HOI have shown that the end-to-end paradigm is limited [23, 17]. As more methods veer towards engineered hand-object representations that explicitly encode surface contacts and relative pose, the transferability of deep learning models to other HOI tasks is increased [27, 28, 29, 30]. Effectively, these richer representations constrain the problem with domain knowledge, reducing the search space by providing the model with more informative inputs. As a result, a lot of noise in the input space is removed, and the model can learn a more robust mapping from input to output. However, these representations are not without shortcomings, as we will see in the next section.

### 1.1.3 Limitations of recent representations

The current trends in HOI representations have shown that modelling contacts can greatly improve the physical plausibility and realism of the grasps. A common approach is to train a model to reason in 3D

space and predict coarse hand-object poses from images, and then refine them with a model trained on hand-object contacts [1, 27, 31, 32, 33]. This coarse-to-fine approach generates an estimate of how an unknown object is being grasped, while dense hand-object interactions (learned or simulated) serve as a test-time optimization (TTO) objective to refine the estimate.

Recently, there has been progress in dense contact map prediction, either directly from images or meshes [34, 1, 31, 35, 36]. However, they still have some limitations such as: compute-intensity (typically involving point cloud processing), the need for feature engineering [1], and being uninformative in cases where the hand is approaching but has not yet touched the object. To address these issues, recent methods have proposed to define the hand pose and shape in an object-centric space using distance fields [28], ray casting [27], or spring systems [37]. While these hand-object representations directly improve the refinement capabilities of TTO, they still require engineering, are compute-intensive, or are not fully differentiable.

## 1.2 Research questions

In the direction of solving the main challenges of hand-object interaction representations, this thesis investigates the following research questions:

1. Can we design new shape, pose and contact representations in order to overcome the limitations of existing representations while maintaining their expressiveness?

2. Can we design new models to demonstrate the expressiveness and versatility of these representations for various applications?

These questions are explored within the scope of static and dynamic grasp denoising and synthesis, from the lens of deep learning for 3D vision. The focus of this research is on the expressiveness, versatility and efficiency of the proposed solutions when combined with state-of-the-art (SOTA) learning methods and test-time adaptation, as motivated by effective recent trends in the field.

## 1.3 Contributions

The main contribution of this dissertation is $p$-CHOIR: a lightweight, continuous and fully-differentiable Coarse Hand-Object Interaction Representation with Probabilistic contact representation. $p$-CHOIR is the culmination of experiments and findings, which we present in three core chapters, led by a background chapter. These core chapters are structured as the following.

**Chapter 3** We first investigate the challenges of HOI modelling through the lenses of group distribution shifts and representation learning. We reframe the problem of hand-object pose prediction under the meta-learning paradigm and analyse the learning of specialized internal representations. This chapter is based on the article: T. Morales and G. Lacey, "A new benchmark for group distribution shifts in hand grasp regression for object manipulation. Can meta-learning raise the bar?", in *Workshop on Distribution Shifts, Neural Information Processing Systems (NeurIPS 2022).*

**Chapter 4** We introduce CHOIR as a novel shape and pose representation for joint modelling of the hand-object pair applied to grasp denoising. This representation is 100× faster to compute than the SOTA, while achieving SOTA results in grasp denoising accuracy with an 8% improvement on the Perturbed ContactPose benchmark.

**Chapter 5** We enhance CHOIR with a novel probabilistic dense contact representation, resulting in $p$-CHOIR. We evaluate this improved solution on two applications: static grasp refinement and synthesis. To do so, we propose a novel diffusion model architecture with joint feature learning for the two components of $p$-CHOIR: shape & pose and contacts. It achieves state-of-the-art results on the Perturbed ContactPose benchmark for grasp denoising and on the ContactPose benchmark for grasp synthesis, and is fully differentiable unlike other SOTA methods. This chapter is based on the article: T. Morales, O. Taheri and G. Lacey, "A Versatile and Differentiable Hand-Object Interaction Representation", in *Algorithm track, Winter Applications in Computer Vision (WACV 2025).*

Prior to these core chapters, we present the background on the relevant deep learning methods and a literature review on HOI modelling.

# Chapter 2

# Background

This chapter is dedicated to providing the necessary theoretical background as well as a broad overview of the literature on hand-object interaction modelling and relevant fields.

## 2.1 Mathematical notation

In this thesis, we remain consistent with the mathematical notation and try to employ a generally accepted form. In certain sections, we may adopt a temporary notation which only applies to that section, in which case we shall define the notation expressly.

We use $p(\cdot|\cdot;\rho)$ to denote that the distribution $p$ depends on the parameters $\rho$. We use $f_\pi(\cdot)$ to imply that the function $f$ is parameterized by $\pi$. For dataset samples in general, we employ $X$. To represent vectors and vector-valued functions we use a small bold letter, such as $\boldsymbol{x}$. Scalar values are represented as non-bold small letters, such as $x$. In the broad sense, the set of latent variables is referred as $Z$ and the set of observed variables as $X$.

## 2.2 Information Theory and Entropy

The primary goal of information theory is to quantify how much information is in data. Claude Shannon, the father of Information Theory, introduced the concept based on a measure called Entropy. Claude Shannon claimed that the right way to define *the information content of an outcome* $\boldsymbol{x} = a_i$

for an ensemble $X$ is

$$\begin{aligned} h(x = a_i) &= \log_2 \frac{1}{P(x = a_i)} \\ &= -\log_2 P(x = a_i). \end{aligned} \tag{2.1}$$

An ensemble is defined as $X = \{x, A_x, P_x\}$ where:

- $x$ is a random variable
- $A_x = \{a_1, a_2, a_3, \ldots, a_i\}$ is the set of possible outcomes
- $P_x = \{p_1, p_2, p_3, \ldots, p_i\}$ is the set of probabilities such that $p(x = a_i) = p_i$ and $\sum_i p_i = 1, p_i \geq 0$.

If we consider a joint distribution of two independent variables $P(X, Y)$, the information content is additive. For an ensemble $XY$ such that an event is a pair of independent random variables $x, y$ with probability $P(x, y) = P(x)P(y)$, the entropy of one outcome is defined as:

$$\begin{aligned} h(x = a, y = b) &= \log_2 \frac{1}{P(x = a, y = b)} \\ &= \log_2 \frac{1}{P(x = a)} + \log_2 \frac{1}{P(y = b)} \end{aligned} \tag{2.2}$$

This means that the information content of the joint outcome of two independent events is the sum of the information content of each independent event. For example, the toss of a coin and the weather outside are independent: the information content of both outcomes is the sum of each.

In Information Theory, the notion of entropy is used all over the place. This quantity defines the average information content (or element of surprise) in the transmission of information. It is defined for a probability distribution as

$$H(X) = \sum_{x_i} P(X = x_i) \log_2 \frac{1}{P(X = x_i)} \tag{2.3}$$

where $P(X = x_i)$ is the probability of event $x_i$ to occur, given the PDF/PMF $p(X)$. A distribution with relatively spread out probability mass (information) has an exponentially higher entropy than one with all its mass concentrated around a single point.

If a system $S$ has $N$ possible states, then its state $s$ can be represented in $\log_2 N$ bits. In fact, if all

its states are uniformly distributed, we can compute the system's entropy using $\log_2$ as:

$$\begin{aligned} H(S) &= -\sum_{i=1}^{N} p(s_i) \log_2 p(s_i) \\ &= -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} \\ &= -\frac{1}{N} \log_2 (\prod_{i=1}^{N} \frac{1}{N}) \\ &= -\frac{1}{N} \log_2 (N^{-N}) \\ &= -\frac{1}{N} (-N) \log_2 N \\ &= \log_2 N \end{aligned} \tag{2.4}$$

Therefore, we can use $\log_2$ to measure the entropy of a distribution of $N$ possible states. Then, we can interpret the result as the minimum number of bits needed to encode a state sampled from that distribution. Using $\log_2$ is well suited to computer science and information encoding, but it generalizes to any base. Hence, Shannon's entropy is a measure of information content.

In Information Theory, we want to find the information in a signal $x \in X$ that is relevant to the information of interest in a signal $y \in Y$. In other words, we want to extract the information in $x$ such that we can predict $y$. The Information Bottleneck is a way to extract a short code out of $X$ that preserves the maximal information about $Y$. We "squeeze" the information that $X$ provides about $Y$ into a limited set of code words $\tilde{X}$ through this bottleneck. The Information Bottleneck is leveraged in the auto-encoder framework, as well as its variational inference version, the variational auto-encoder. Another fundamental concept at the heart of variational inference is the Kullback-Leibler divergence.

## 2.3 The Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence measure, heavily based on the measure of Entropy, allows us to calculate the minimum amount of information required to transmit a message, hence enabling us to compress data. However, it is just a measure, and we still need to find the encoding scheme that leads to this compression. Entropy tells us the theoretical lower bound on the number of bits needed

to encode the message, thus we can measure the information content of our data. When working with complex and high dimensional problems, we often replace observed data or a complex and intractable distribution with a simpler approximate one. Being able to quantify the information content in the true data distribution, we want to quantify how much information is lost when we substitute our observed distribution for a parametrized approximation. The Kullback-Leibler (KL) divergence quantifies how much information is lost when we approximate the true distribution with a simpler one.

Let us consider the case of a true data distribution $p(X)$ and its approximate distribution $q(X)$. In such scenario, the KL divergence $D_{KL}(p||q)$ is a modification of the formula for Entropy which accounts for $q$:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot (\log p(x_i) - \log q(x_i)) \quad = \mathbb{E}_{p(x)} \left[\log p(x) - \log q(x)\right] \tag{2.5}$$

The KL divergence can be rewritten, for discrete distributions, as

$$D_{KL}(p||q) = \sum_{i=1}^{N} \underbrace{p(x_i)}_{A} \cdot \underbrace{\log \frac{p(x_i)}{q(x_i)}}_{B}. \tag{2.6}$$

There are two main terms in the common definition of the KL divergence: $A$ and $B$. The first is the probability of $x_i$ according to the base distribution. It acts as a weight for the divergence measure: events that are more likely to occur have a stronger effect on the divergence of the distribution. The second defines the divergence between the two distributions as the logarithm of a fraction of probabilities. By looking at this fraction, we can observe three behaviours of this log term:

1. If $q(x_i) > p(x_i)$ the fraction is between 0 and 1 and the logarithm decreases exponentially with respect to the difference between $q(x_i)$ and $p(x_i)$. A larger difference implies an exponentially lower negative divergence.

2. If $q(x_i) < p(x_i)$ the fraction is higher than 1 and the logarithm increases exponentially with respect to the difference between $q(x_i)$ and $p(x_i)$. A larger difference implies an exponentially higher positive divergence.

3. If $q(x_i) = p(x_i)$ the logarithm is 0 and there is no divergence.

A direct observation from this breakdown is that the divergence is not a distance metric, since it comprises a direction and is therefore asymmetric. This is a direct consequence of the fact that we are measuring the loss of information in an approximation of the original distribution: if the metric was symmetric, there would be no way of knowing if the approximating distribution is adding information or removing it.

The KL divergence is a fundamental concept in Variational Inference, and is a cornerstone of Latent Variable Models such as Variational Auto-Encoders, diffusion models, and other generative models.

## 2.4 Latent Variable Models

In the context of data modelling and generative models, the goal is to learn the probability distribution of some data, $p(X)$, such that we can generate new examples through sampling. This can be achieved using numerical methods (Markov Chain Monte Carlo, Gibbs sampling, etc.) or analytical methods (maximum likelihood estimation, etc.); these are often infeasible for high-dimensional data or very complex data distributions One way to simplify the problem is by using latent variables that explain the observed ones. This is a common approach in machine learning, as it allows redefining the complex joint distribution $p(X, Z)$ in terms of simpler components. In this section, we introduce the concept of Latent Variable Models (LVMs) and the two core frameworks that derive from them and that we use in our main contributions: Variational Auto-Encoders (VAEs) and Denoising Diffusion Probabilistic Models (DDPMs).

Latent Variable Models (LVMs) aim to model the joint probability distribution of data with their explanatory latent variables. A variable is said latent when it is not observed in the data, but is inferred by the model. Latent variables can help generate coherent samples from the true data distribution because they explain implicit decisions or contain structural information on the generative process. In the LVM framework, we assume the data $X$ to be generated by some random process involving $Z$, where there is at least one $Z$ which causes the process to generate something very similar to $X$. In essence, LVMs are a family of methods that define a framework to:

1. Define a probability distribution $p(Z|\theta)$ for the latent variables.

2. Infer those latent variables by extending such distribution as a joint distribution $p(X, Z|\theta)$ over the observed data $X$ and the latent variables $Z$.

3. Use the Bayes theorem to compute the posterior distribution $p(Z|X, \theta)$ over the latent variables.

The main challenge now is to find the parameters $\theta$ that govern the model. The remainder of this section is dedicated to explaining the algorithms, and their limitations, that can be used to find these algorithms.

To better understand the challenges in estimating the model parameters, let us consider two scenarios with respect to the availability of data:

- Access to the complete data, where we assume each $Z$ to be given with each $X$: $\{X, Z\}$.
- Access to the incomplete data, where only $X$ is observed: $\{X\}$.

With the complete data, it is assumed that maximizing the log likelihood is straightforward and trivial – at least for the exponential family of distributions – while direct optimization of $p(X|\theta)$ is difficult. In practice, we only have access to the incomplete data, such that the log-likelihood for continuous latent variables is given by

$$\log p(X|\theta) = \log \left\{ \int p(X, Z|\theta) dZ \right\} \tag{2.7}$$

where $\theta$ are the parameters of the model. Due to the integral being inside the logarithm, the resulting expression for the maximum likelihood estimator (MLE) is complicated, and the marginal does not belong to the exponential family even if the joint distribution does. Indeed, computing the integral is computationally expensive and often intractable. Obtaining a maximum likelihood estimator for $p(X|\theta)$ via analytical methods is very hard due to $Z$ being unobserved. Using gradient-based methods is not feasible either, as the problem is *intractable* due to the integral over $Z$, since $Z$ is often high dimensional for complex and high-dimensional data (images and the like). For high-dimensional continuous or discrete $Z$, numerical integration may not be feasible either due to the curse of dimensionality. However, having access to the full data would make the computation of the MLE straightforward.

Assuming we can infer $Z$ from $X$ (*i.e.* we know $Z_1 = a, X_1 = b, \ldots$), we can then define the posterior conditional distribution $p(Z|X, \theta)$. This would allow the model to learn the joint distribution $p(X, Z|\theta)$ through its expected value under this posterior; this is the idea of the Expectation Maximization

(EM) algorithm. EM is a general method for finding the MLE of a latent variable probabilistic model under the assumption that solving for the complete-data likelihood function $p(X, Z|\theta)$ is significantly easier than for the incomplete-data likelihood. To achieve this, the general EM algorithm decomposes the incomplete data log-likelihood function into two components in a more manageable form. Such decomposition is defined as

$$\log p(X|\theta) = \mathcal{L}(q, \theta) + \mathrm{KL}(q\|p) \tag{2.8}$$

where

$$\begin{aligned} \mathcal{L}(q, \theta) &= \int_Z q(Z) \log \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\} dZ \\ \mathrm{KL}(q\|p) &= -\int_Z q(Z) \log \left\{ \frac{p(Z|X, \theta)}{q(Z)} \right\} dZ \end{aligned} \tag{2.9}$$

and with $q(Z)$ a distribution over latent variables $Z$, typically an approximate distribution to the true posterior $p(Z|X)$ to make the optimization easier. Effectively, this gives rise to a relatively simple maximization problem for the lower bound of the incomplete data likelihood function. In Eq. (2.8), the Kullback-Leibler divergence term $\mathrm{KL}(q\|p)$, by definition, satisfies $\mathrm{KL}(q\|p) \geq 0$ if and only if $p(Z|X, \theta) = q(Z)$; *i.e.* if the model learns a perfect posterior. Therefore, since $\mathcal{L}(q, \theta) \leq \log p(X|\theta)$, it follows that $\mathcal{L}(q, \theta)$ is a lower bound on the incomplete data likelihood $\log p(X|\theta)$.

The EM algorithm is executed in two steps, where we start with model parameters $\theta^{\mathrm{old}}$ and iteratively update these parameters until convergence. After substituting $q(Z) = p(Z|X, \theta^{\mathrm{old}})$ into Eq. (2.9), the lower bound takes the form

$$\begin{aligned} \mathcal{L}(q, \theta) &= \int_Z p(Z|X, \theta^{\mathrm{old}}) \log p(X, Z|\theta) dZ - \int_Z p(Z|X, \theta^{\mathrm{old}}) \log p(Z|X, \theta^{\mathrm{old}}) dZ \\ \mathcal{L}(q, \theta) &= \mathcal{Q}(\theta, \theta^{\mathrm{old}}) + \mathrm{const} \end{aligned} \tag{2.10}$$

where $\mathcal{Q}(\theta, \theta^{\mathrm{old}})$ is referred to as the $\mathcal{Q}$ function and the constant is the entropy of the distribution $q(Z)$ and does not depend on $\theta$. The $\mathcal{Q}$ function corresponds to the expectation over the complete data log likelihood, $\log p(X, Z|\theta)$ for a fixed value of $\theta$, under the posterior $p(Z|X, \theta^{\mathrm{old}})$. With the logarithm now *inside* the integral, and with $\mathcal{Q}$ being an expectation under a posterior distribution we can sample from, the objective is much simpler and tractable. In the E(xpectation) step of EM, we maximize the lower bound $\mathcal{L}(q, \theta^{\mathrm{old}})$ with respect to $q(Z)$ while holding $\theta^{\mathrm{old}}$ fixed. Since $Z$ is integrated out,

$\log p(X|\theta^{\text{old}})$ does not depend on $q(Z)$, and thus we can maximize the lower bound by minimizing the KL divergence term $\text{KL}(q\|p)$. Minimizing this term is equivalent to finding the auxiliary distribution $q(Z)$ that is closest to the true posterior $p(Z|X,\theta)$. In the M(aximization) step, we maximize the lower bound with respect to $\theta$ while holding $q(Z)$ fixed. Essentially, the expectation of the complete-data log likelihood is maximized. Because the parameters $\theta$ are inside the logarithm, the M step is much simpler than maximizing $p(X|\theta)$ directly.

By training the model to produce training set samples – or very similar to – with high probability, we also train it to produce dissimilar samples with very low probability. This problem is a central task in the application of probabilistic models: we need to learn the conditional posterior $p(Z|X,\theta)$ of the latent variables $Z$ for the observed variables $X$. Overall, the EM algorithm remains a mathematical framework that leaves some choices to be made by a domain expert: what distribution to use for $Z$? What information should $Z$ contain about $X$? EM is also limited when it comes to dealing with high-dimensional data and complex, non-linear relationships between $X$ and $Z$. For this family of problems, such as modelling natural images or hand-object interaction pairs, EM is not powerful enough to learn the model parameters. A direct solution to this problem is to use Variational Inference (VI) methods, which provide a way to *approximate* the intractable posterior $p(Z|X,\theta)$ with a simpler, tractable distribution, such that the optimization problem can be solved more easily.

### 2.4.1 Variational Inference

The basic idea of Variational Inference is to analytically *approximate* the posterior distribution $p(Z|X,\theta)$ by assuming a specific form, often chosen to be the Gaussian distribution. Restricting the range of functions over which to optimize has a tremendous effect on computational efficiency. Compared to sampling methods (Markov Chain Monte Carlo, Gibbs sampling, etc.), which can generate exact results by learn an accurate approximation of the true distribution, Variational Inference methods trade accuracy for computational efficiency. The word *variational* has origins in the *Calculus of Variations* [38]. Unlike traditional calculus where the goal is to find the derivatives of functions, here the goal is to find the derivatives of *functionals*. A *functional* is a mapping from function to output. A widely used functional in Machine Learning is Entropy, which computes a real value for a given probability density function.

Essentially, Variational Inference lets us approximate the posterior distribution $p(Z|X,\theta)$ by optimizing a functional that measures the difference in information between our approximation and the true posterior. This functional is often the Kullback-Leibler divergence, as is used in EM. Variational Inference extends the EM algorithm by redefining the parameters of the model as stochastic variables absorbed in $Z$, such that Eq. (2.8) becomes

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{2.11}$$

and more specifically

$$\mathcal{L}(q) = \int_Z q(Z) \log \left\{ \frac{p(X,Z)}{q(Z)} \right\} dZ \tag{2.12}$$

With the parameters of the model now absorbed into $Z$, we still need a rich family of approximating distributions. The restriction over this family is achieved using a parametric distribution $q(Z|\omega)$, such as the Gaussian distribution. The goal is now to find the optimal parameter $\omega^*$ that minimizes the KL divergence between $q(Z|\omega)$ and the true posterior $p(Z|X)$.

Variational Inference introduces the idea of restricting the family of distributions to approximate the true posterior. Such restriction is achieved using parametric distributions, to reduce the search space and take advantage of known properties. It then proposes to maximize a lower bound by minimizing the KL divergence to find a candidate member in this family. However, VI does not provide a complete framework for efficient inference and learning in probabilistic models, especially when dealing with continuous latent variables and intractable posterior distributions. This task becomes even more challenging when dealing with high-dimensional data and complex problems.

### 2.4.2 Auto-encoding Variational Bayes

While EM is a general approach for solving intractable problems with LVMs, it remains a theoretical framework that leaves several practical issues open. For instance, the EM algorithm assumes a specific form of a parametric distribution for the latents, and does not focus on learning good latent variables from the data. This is especially limiting for generative modelling where the goal is to learn a good compressed representation of the data in the latent space. The EM algorithm also assumes that the likelihood and the prior are tractable, which is not always the case. In the context of large datasets or

high-dimensional data, the computation of the marginal likelihood can be infeasible. As the quantity of data grows, the convergence of the EM algorithms suffers, and batch optimization becomes too costly. However, we are interested in data modelling problems that typically involve these challenges.

To address the limitations of the EM algorithm, the Auto-encoding Variational Bayes framework [39] introduces a new perspective on LVMs using Variational Inference. The main objective of this framework is to offer a flexible way of learning latent representations from the data itself. Effectively, the authors propose to apply the Variational Inference framework to an autoencoder, such that the generative model's parameters are learned jointly with the parameters $\omega$ of the true posterior approximation $q(Z|\omega)$ called the "recognition model". This recognition model is a neural network that takes an input $X$ and outputs the parameters of the approximate posterior distribution $q(Z|X;\omega)$. The generative model is then a neural network that takes a sample $z$ from this approximate posterior distribution and generates a sample $\hat{X}$ from the data space. This method, called the Variational Auto-encoder (VAE), is a generative model that jointly learns the recognition model and the generative model. It is trained by maximizing the Evidence Lower Bound (ELBO) loss, which is derived from Eq. (2.11) and is defined as follows:

$$\mathcal{L}(\theta, \omega; X) = \mathbb{E}_{q(Z|X;\omega)}[\log p(X|Z;\theta)] - \mathrm{KL}(q(Z|X;\omega)\|p(Z)) \tag{2.13}$$

This loss is a trade-off between the reconstruction error and the KL divergence between the approximate posterior and the prior. The ELBO loss is optimized using stochastic gradient descent, allowing for batch optimization and efficient learning on large datasets.

The VAE framework is a cornerstone in the field of deep learning and generative modelling, with many subsequent works building from it [40, 41, 42]. Recently, the VAE framework has been extended to more complex data distributions with the Normalizing Flows [43] and the more recent diffusion models.

### 2.4.3 Denoising Diffusion Probabilistic Models

In recent years, a new family of generative models arose in the class of Latent Variable Models, called *diffusion models*. Their key idea is to exploit random walks with drift to model highly complex data

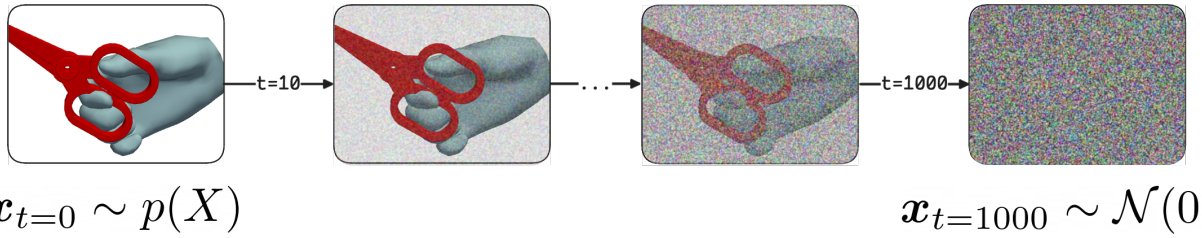



Figure 2.1: At each step $t$ of the diffusion process, the sample at step $t-1$ is corrupted with a noise sample drawn from a Gaussian distribution. After enough diffusion steps (*i.e.* 1000), the sample $\boldsymbol{x}_t$ is indistinguishable from white noise $\epsilon \sim \mathcal{N}(0, (I))$.

distributions. These methods are originally based on non-equilibrium thermodynamics [44], but a more recent branch uses Variational Inference (VI) for more efficient training on large datasets: Denoising Diffusion Probabilistic Models [45]. These are the focus of this section.

Denoising Diffusion Probabilistic Models (DDPM) have recently made their prowess in distribution learning – aka generative modelling – for high dimensional problems, from modelling natural images [13, 46] to videos [47, 48]. The prowess of DDPMs is all the more impressive in scenarios where we aim to generate a sample conditioned on a context variable, such as in image inpainting or text-to-image generation. The key idea is to model the data distribution as a Markov chain, where each step of the chain is a latent variable that is corrupted by noise.

The unreasonable effectiveness of modern DDPMs is due to two factors. The first is their powerful distribution learning framework, based on non-equilibrium thermodynamics [44] and improved with Variational Inference. Training a DDPM with VI allows to efficiently learn the backbone neural network parameters in the abundance of training data. The second is the harnessing of a powerful neural network backbone with strong inductive priors for image feature extraction. The backbone used for image modelling is the U-Net architecture [49], with modern deep learning mechanisms such as Multi-Head Cross Attention (MHCA) and Multi-Head Self Attention (MHSA). While MHSA improves the learning of deep features, MHCA enables context conditioning via relevant feature extraction and fusion. These mechanisms allow modelling complex relationships between context and target information.

### The forward process

In its original formulation [45], the DDPM gradually corrupts a data sample $\boldsymbol{x} \sim p(\boldsymbol{x})$ through additive noise sampled from a standard normal distribution at each step of the corruption process (see Fig. 2.1); this is referred to as the *forward* or *diffusion process*. This diffusion process defines a Markov chain where each step $t$ defines a latent variable $\boldsymbol{z}_t$. This latent is sampled from a conditional distribution depending only on the previous step's latent $\boldsymbol{z}_{t-1}$, such that

$$\begin{aligned} q(\boldsymbol{z}_1|\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{z}_1; \sqrt{1-\beta_1}\boldsymbol{x}, \beta_1 \boldsymbol{I}) \\ q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) &= \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t \boldsymbol{I}) \end{aligned} \tag{2.14}$$

with diffusion parameters $\beta_1 \dots \beta_T$. However, with some mathematical gymnastics [45, 50], the distribution of $\boldsymbol{z}_t$ can be reformulated such that it only depends on the data sample $\boldsymbol{x}$:

$$q(\boldsymbol{z}_t|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{\alpha_t}\boldsymbol{x}, (1-\alpha_t)\boldsymbol{I}), \tag{2.15}$$

where $\alpha_t = \prod_{s=1}^{t} 1 - \beta_s$. With enough diffusion steps (typically $T = 1000$) and small values of $\beta_t$, the final distribution $q(\boldsymbol{z}_T|\boldsymbol{z}_{T-1})$ resembles a standard normal Gaussian.

The goal is to learn the *reverse process*, where we sample a latent variable $\boldsymbol{z}_t$ from a tractable distribution and traverse back along the Markov chain. This allows to obtain a sample from the original, intractable, data distribution.

### The reverse process

The reverse process consists in learning a conditional distribution $p(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ for each step $t$. Since these distributions depend on the data distribution $p(\boldsymbol{x})$, they can be complex and intractable. We approximate them as normal distributions parameterized by neural networks $\boldsymbol{f}_{\theta_1} \dots \boldsymbol{f}_{\theta_T}$:

$$\begin{aligned} q(\boldsymbol{z}_T) &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t; \theta_t) &= \mathcal{N}(\boldsymbol{f}_{\theta_t}(\boldsymbol{z}_t), \sigma^2 \boldsymbol{I}) \\ q(\boldsymbol{x}|\boldsymbol{z}_1; \theta_1) &= \mathcal{N}(\boldsymbol{f}_{\theta_1}(\boldsymbol{z}_1), \sigma^2 \boldsymbol{I}) \end{aligned} \tag{2.16}$$

where the output $\boldsymbol{f}_{\theta_t}(\boldsymbol{z}_t)$ is the mean and the hyperparameter $\sigma^2$ defines the covariance. The model is trained by maximizing the log-likelihood of the data given the parameters $\theta$, which is achieved by maximizing the evidence lower bound (ELBO) [51].

When training a modern diffusion model [50], the goal of learning $p(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ is in fact simplified to learning to predict the noise sample $\boldsymbol{\epsilon}_t$ that was added to the data sample $\boldsymbol{x}$ to produce the latent variable $\boldsymbol{z}_t$. This is seen by reparameterizing Eq. (2.15) as:

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2.17}$$

We may now replace the model of Eq. (2.16) $\hat{\boldsymbol{z}}_{t-1} = \boldsymbol{f}_{\theta_t}(\boldsymbol{z}_t)$ with $\hat{\boldsymbol{\epsilon}} = \boldsymbol{g}_{\psi_t}(\boldsymbol{z}_t)$. With the help of stochastic gradient descent, the training objective is simply formulated as the Mean Squared Error (MSE) between the predicted and ground-truth noise samples:

$$\begin{aligned} \mathcal{L}(\psi_{1\ldots T}) &= \sum_{i=1}^{N}\sum_{t=1}^{T} \|\boldsymbol{g}_{\psi_t}(\boldsymbol{z}_{it}) - \boldsymbol{\epsilon}_{it}\|^2 \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T} \|\boldsymbol{g}_{\psi_t}(\sqrt{\alpha_t}\boldsymbol{x}_i + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{it}) - \boldsymbol{\epsilon}_{it}\|^2 \end{aligned} \tag{2.18}$$

where $N$ is the number of training samples and $T$ is the number of diffusion steps. In practice, the time step is sampled from a uniform distribution for each batch element $\boldsymbol{x}_i$. As expected, the modelling capacity of diffusion models is bound by the complexity of the backbone neural network $\boldsymbol{g}$. In practice, the choice of neural network depends on the task and the necessary inductive priors, such as the U-Net [49] for images or a Transformer for animation [52]. The choice of inductive priors is tightly linked to the modality of the data and its representation. Since this thesis is about modelling interactions between entities in a 3D world, the following section delves into the various forms of 3D representations.

## 2.5 3D representations

In Computer Graphics and Vision, we process objects, shapes and scenes in various ways. There are many ways to represent these in 3D, either for input or for output, and the choice of representation

often depends on the task and problem at hand. 3D representations are essential to HOI modelling, as the input, output or even intermediate representation, often fall directly into the 3D domain. In this section, we discuss the most common 3D representations and categorize them in two main types: explicit and implicit. For simplicity, we shall employ the term "scene" to refer to any 3D entity to be represented.

### 2.5.1 Explicit representations

The most straightforward way to represent a scene is to *explicitly* define it in a structured way, such that it can be written down and understood using a specification. In this section, we give a non-exhaustive list of the most common explicit representations and their properties.

**Meshes / polygons** The most commonly used shape and surface representation, they are a collection of vertices (*i.e.* points in space), edges and faces that define the scene. This structured representation simply defines connectivity between vertices, such that a rendering algorithm can draw faces. They have the expressivity of point clouds with additional connectivity and neighbourhood information, enabling inductive priors in deep learning [53]. However, processing them or generating them is rather challenging due to their intricate structure that requires strong inductive priors(*i.e.* inter-penetration issues). Hence, most generative methods learn to deform a mesh template, which makes it hard to learn an arbitrary topology [54]. Nevertheless, when leveraged with a powerful learning framework, this representation can provide highly-detailed results with the advantage of being accurately texturable [55].

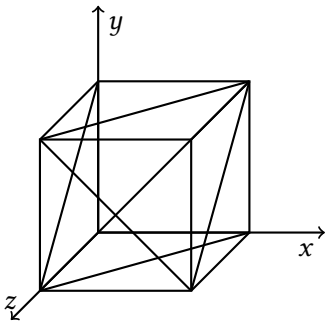


Figure 2.2: Mesh representation.

**Voxel grids** Similarly to image pixels, they represent scenes as three-dimensional pixels, called voxels for *vo*lume *el*ements. Voxels represent volumes in a regular grid, and as such they are not an efficient format for arbitrary shape definition. They

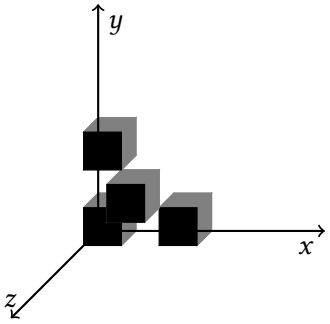


Figure 2.3: Voxel representation.

do not scale well as they grow cubically in memory, and are inherently highly redundant. However, the grid structure allows processing voxels with 3D convolutions, which enables parameter sharing and inductive priors in a variety of tasks [56, 57, 58].

**Point clouds** They simply represent points in space via a set of absolute coordinates, making them an adequate choice for unstructured data captured from a scene (*i.e.* with a 3D scanner). As they do not waste memory, they are more efficient than voxel grids, but they cannot represent volume. Point clouds are arbitrarily scattered in space, hence they lack connectivity and neighbourhood information. Due to this, point cloud processing in deep learning often relies on pooling [59, 60, 61], set-based [62], graph [63] or distance-based [64] methods. In addition, obtaining a shape from a point cloud (*i.e.* a mesh representation or voxel grid) necessitates using a heuristics-based algorithm such as Ball-Pivoting [65].

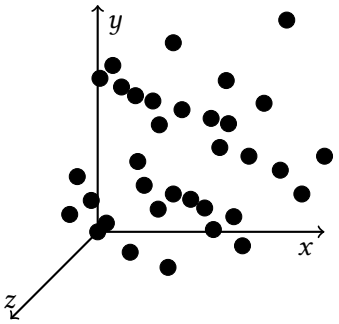


Figure 2.4: Point cloud representation.

### 2.5.2 Implicit representations

Unlike explicit representations, implicit representations define a scene as a continuous function. The input to the function is a query point in space, and the output is a descriptor $y \in \mathbb{R}^N$ of the scene with respect to that point, *i.e.* for a 3D scene:

$$\boldsymbol{x} \mapsto \mathbb{R}^N, \quad \boldsymbol{x} \in \mathbb{R}^3 \tag{2.19}$$

Implicit representations allow representing arbitrary topologies without the constraints of discrete representations, resulting in smooth surfaces. The properties of implicit representations make them well suited for deep learning and generative models, as they can be learned end-to-end. However, they are not without limitations, as they tend to produce overly smooth surfaces that do not catch intricate surface details and require many function evaluations to recover a dense surface. In addition, the nature of implicit representations makes it hard to use surface-based losses or regularization terms

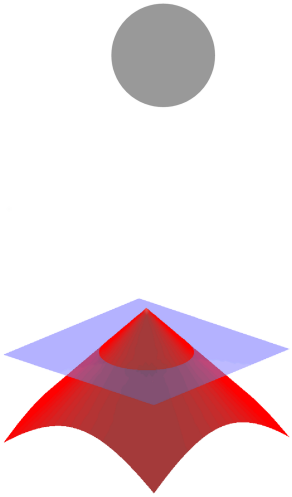


Figure 2.5: Illustration of the Signed Distance Field representation [68]: the red graph shows the positive and negative distance values to the shape above, from each point on the blue plane.

during training.

A typical implicit scene representation is the occupancy of the surface. For a query point, the value $y$ of the function is a binary value that indicates whether the point is inside or outside the surface, akin to a continuous voxel grid. Occupancy Networks [66] is a pioneering work that demonstrates the application of an occupancy-based implicit scene representation for 3D shape reconstruction. This concept is taken a step further with image feature extraction for high definition 3D human reconstruction [67]. However, since Occupancy Networks rely on a probabilistic representation of the surface boundary, it can result in ambiguities in complex shapes. Moreover, the convergence of Occupancy Networks is not optimal and requires large amounts of training samples to cover the space near the decision boundary. A more efficient and accurate alternative is the family of Neural Distance Fields, a neural implicit representation of distance fields.

### Distance Fields

A very commonly used implicit representation is the distance field. It represents the scene as the distance of any point in space to the surface. There exists two types of distance fields: signed and

unsigned.

**Signed Distance Fields** A signed distance field (SDF) maps each point in space to the signed distance to the closest point on the surface (see Fig. 2.5). Where the distance is null, *i.e.* $f(\boldsymbol{x}) = 0$, is called the zero-level set or the boundary of the surface. On either side of the surface boundary, the distance is signed or unsigned, where the convention is negative on the outside and positive on the inside. Conveniently, the SDF's gradient is a unit vector pointing from any $\boldsymbol{x} \in \mathbb{R}^3$ to the surface, which is useful for ray marching and rendering. It is common to truncate the SDF, as this helps to constrain the function to a finite range, which is beneficial for training deep learning models.

**Unsigned Distance Fields** An unsigned distance field (UDF) is a special case of the SDF where the distance is unsigned. It is always non-negative and also has a zero-level set. UDFs recently became a viable and more efficient alternative to SDFs for representing scenes with deep learning, especially in the case of non-watertight (open) surfaces where SDFs are not a viable representation. Neural Unsigned Distance Fields provide an easier learning objective by removing the sign, which can be ambiguous in intricate areas of a scene [69]. Overall, UDFs do not require any assumptions on the properties of a scene and are hence easier to work with. This idea is exploited in the Basis Point Set representation [64], a special case of the UDF which uses a fixed set of measurement points to produce a discrete UDF vector. This in fact results in an explicit shape representation.

From a deep learning perspective, learning distance fields – also known as Neural Distance Fields (NDFs) – is more efficient than their occupancy-based equivalent neural representation. Mainly, NDFs allow recovering an explicit representation via their zero-level set, eliminating the ambiguity of Occupancy Networks [66, 70]. NDFs also require less training samples thanks to their non-probabilistic nature and their more informative representation. This makes NDFs a popular choice of representations for scene reconstruction [69, 71, 72] or generation [73, 74, 75] with deep learning. In the next section, the connection between 3D representations and HOI modelling becomes apparent, as we discuss concrete examples of how these representations are used in the context of HOI.

## 2.6 Overview of the literature in hand-object interaction

Despite advances and impressive results across a wide variety of HOI tasks, learning an accurate HOI prior model remains a challenging and unsolved task problem. While this thesis focuses on the representation aspect of an HOI prior, we here provide a broad overview of the literature in hand-object interaction. To navigate the vast literature, we propose to subdivide HOI modelling works into the following categories:

- Benchmarks & datasets
- Pose prediction
- 3D reconstruction
- Interaction generation
- Representations

We choose this specific ordering to reflect the hierarchy and chronology of HOI tasks. In addition, we close this section with a brief review of the literature on test-time adaptation (TTA) methods, as this paradigm has become a cornerstone of HOI modelling in the recent literature.

### 2.6.1 Benchmarks and datasets

Hand-object interaction is studied from the perspective of various tasks, such as joint pose prediction from images or manipulation synthesis from object and human body trajectories. As such, there exists a large amount of datasets and benchmarks to evaluate these tasks. Due to the high overlap between some HOI tasks, it is common for a dataset to contain enough annotation modalities such that it can be used for several tasks. Therefore, in this section we attempt to provide a clear view of the recently published datasets through a broad categorization of HOI tasks.: *static grasps* and *dynamic manipulation*.

In Table 2.1, we give a non-fully-exhaustive list of published datasets that cover *static grasps* in HOI. This category covers all HOI interactions where the hand pose remains relatively static, without involving dexterous manipulation. This often involves grasping and holding objects, or handover. We

| Datasets | # subjects | # objects | # hand poses | Nature | Markerless | Object pose | Hand mesh | Contacts | Bimanual |
|---|---|---|---|---|---|---|---|---|---|
| | Criteria | | | | | | | | |
| TUB Dataset [76] | 17 | 25 | – | Real | ✗ | ✗ | ✗ | ✗ | ✗ |
| SynthHands / EgoDextr [77] | 4 | 7 | – | Hybrid | ✓ | ✗ | ✗ | ✗ | ✗ |
| ObMan [78] | – | 8 | 21K | Synthetic | ✓ | ✗ | ✗ | ✗ | ✗ |
| ContactDB [79] | 50 | 50 | – | MoCap | ✓ | – | – | ✓ | ✗ |
| ContactPose [80] | 50 | 25 | 2.3K | Hybrid | ✓ | ✓ | ✓ | ✓ | ✓ |
| HO3Dv3 [81, 82] | 10 | 10 | – | Real | ✓ | ✓ | ✓ | ✗ | ✗ |
| OakInk-Image [83] | 12 | 20 | 1K | Real | ✗ | ✓ | ✓ | ✓ | ✓ |
| OakInk-Shape [83] | 12 | 20 | 50K | Synth-MoCap | – | ✓ | ✓ | ✓ | ✓ |
| DexYCB [84] | 10 | 20 | 1K | Real | ✓ | ✓ | ✓ | ✗ | ✗ |
| AffordPose [85] | – | 641 | 26.7K | Synthetic | ✓ | ✓ | ✓ | ✗ | ✗ |

Table 2.1: Static grasp datasets in HOI in chronological order, from 2016 to 2023. Most datasets are of real images, a few are a mixture of real and synthetic images, and some are purely motion capture (MoCap) or synthesized hand grasps (Synth-MoCap). Missing or irrelevant criteria are marked '–'.

choose to ignore older datasets that we deem outdated and superseded by more recent ones due to their limited relevancy. Table 2.1 shows relatively few datasets providing contact maps or considering bimanual grasps (*e.g.* holding a camera with two hands). Then, in Table 2.2 we propose a curated list of datasets focusing on dynamic manipulation. There is a recently growing interest in this area of HOI, as it remains a neglected domain which we can now investigate thanks to the progress on static grasp modelling. These datasets aim to capture and annotate natural hand-object manipulation, and more recently bimanual manipulation of deformable or articulated objects.

From Table 2.1 and Table 2.2, we can observe a recent trend towards intent-based and affordance-based grasping. This is notable with the recently published OakInk2 [93], TACO [94] and AffordPose [85] which include richer annotations such as object affordances, intent labels or functional taxonomies. These works allow for a new exploration of the problem space where additional information is considered for modelling grasps. Nevertheless, static grasp datasets are still relevant and an essential part of learning accurate HOI priors. The works mentioned in the following sections are all evaluated on benchmarks discussed in this section.

| Datasets | # subjects | # objects | # frames | Nature | Markerless | Object pose | Hand mesh | Contacts | Bimanual | Articulated objects |
|---|---|---|---|---|---|---|---|---|---|---|
| | Criteria | | | | | | | | | |
| FPHAD / FHAB [86] | 6 | 4 | 100K | Real | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| FreiHAND [87] | 32 | 27+ | 37K | Real | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| GRAB [9] | 10 | 51 | 1.6M | MoCap | – | ✓ | ✓ | ✓ | ✓ | ✗ |
| H2O [88] | 4 | 8 | 572K | Real | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ |
| HOI4D [89] | 9 | 800 | 2.4M | Real | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| ARCTIC [90] | 10 | 11 | 2.1M | Real | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ |
| HMDO [91] | – | 12 | 21.6K | Real | ✓ | ✓ | ✓ | ✓ | ✓ | ✓* |
| AssemblyHands [92] | 34 | – | 3.03M | Real | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| OakInk2 [93] | 9 | 75 | 4.01M | Real | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| TACO [94] | 14 | 196 | 5.2M | Real | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| ContactArt [95] | – | 80 | 550K | Hybrid | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |

Table 2.2: Dynamic manipulation datasets in HOI in chronological order, from 2018 to 2024. All datasets are of real videos, except for GRAB [9] which is motion capture (MoCap) and ContactArt [95] which used RGB-D to capture and rendered textured animations. Missing or irrelevant criteria are marked '–'. * HMDO [91] is currently the only dataset considering and annotating deformable object manipulation.

### 2.6.2 Hand and object pose prediction

Hand or joint hand-object pose prediction from images form the basis of all existing works in HOI due to two main reasons. The first being that accurately predicting the hand pose under object occlusions requires a model of hand-object interaction. The second reason is that predicting the pose accurately is useful for a wide range of real-world applications, namely in mixed reality and monitoring. Pose prediction is a first step into modelling 3D hand-object interactions, and is arguably one of the simplest forms of the modelling problem. In itself, pose prediction can help with other problems, such as action understanding. In this section, we dive into the literature on hand and joint hand-object pose prediction, skipping the earliest works [96, 97, 98] and considering only deep-learning-based methods.

Initially, the widely used approach was to regress joint keypoints as 3D vectors, rotation vectors,

or both. To achieve this, the predominant paradigm for pose prediction methods is the ”end-to-end” approach, where a neural network is trained to predict the pose in one forward pass. Several works adopt a 2D to 3D approach, which typically requires two stages. Of them, earlier works generated synthetic data to train a neural network [99, 100]. [99] was one of the earliest works to propose a 3D pose prior. The authors trained a model to lift 2D keypoints into plausible 3D coordinates using synthetic data, obtaining results competitive with methods that employ depth images. A later work followed the same idea and contributed with a public synthetic dataset for two-hand global 3D pose estimation from egocentric view [101]. [102] introduced novel biomechanical constraints as a simple mechanism to improve 3D skeleton prediction from 2D supervision. This results in a weakly supervised method that can be used as an add-on to any keypoint prediction method. Subsequent works seem to mark a switch to direct 3D modelling, such as neural network architectures based on graph convolutions [6, 103, 104], with [105] blending graph convolutions with Transformers for SOTA results with reduce parameter count. [106] proposed a weakly supervised 3D CNN trained on voxelized depth maps, and [107] augmented hand point clouds with object occlusions for sparsely supervised hand pose and shape prediction in object interaction sequences. Regarding real time applications, most works follow a 1-stage approach to favour efficiency, with tailored architectures [108, 109, 110, 111] or a hybrid neural network with a gating mechanism for an adaptive compute budget [112]. On a different note, several works have focused on extracting better features via joint-aware feature fusion [113, 114], pixel-wise local feature extraction from depth images [115, 116, 117], hand-object feature fusion from two backbones [118] or feature aggregation of both hands [119].

Setting apart from the joint regression approach is a pioneering work that set the standard approach for hand representation: the MANO parametric hand mesh [120]. The authors proposed a parametric hand mesh with shape and pose parameters. The model was fitted to a large collection of accurate 3D hand scans from dozens of participants, which allowed to compute the statistics on shape and pose variations of hands. Consequently, MANO proposes a pose parametrization in PCA space, in addition to the axis-angle format. MANO is a powerful hand model that uses classical graphics techniques (*i.e.* linear blend skinning) for fast rendering on GPUs, while being fully differentiable. Hence, it allows training neural network pose predictors efficiently with a statistical prior, which made MANO the de facto hand model for training neural networks in the field. Nevertheless, there has been

attempts at improving upon MANO, such as [121] which proposes a neural occupancy representation for 3D-keypoint-driven mesh reconstruction. While most works that built on the MANO model were mostly aimed at mesh reconstruction with accurate pose, which will be covered in the next section, several works doing purely pose prediction have emerged [119, 118, 122, 123]. One such notable work for joint hand-object pose prediction is HOISDF [123], which leverages Signed Distance Fields as an intermediate representation for a powerful 3D shape and pose prior. The authors use this rich interaction prior to guide the regression of the hand shape and pose as MANO parameters and joint offsets, alongside the object 6D pose. HOISDF uses point-wise cross-attention to enhance the SDF features, resulting in SOTA hand-object pose accuracy on the challenging DexYCB dataset [84].

Recent methods have leveraged the recent advances in deep learning architectures. For instance, some works profited from the long-range modelling capabilities of Transformers for either joint hand-object pose prediction [124, 125], pose prediction in bi-manual manipulation [126, 33], or for a foundational hand reconstruction model [127] trained from large scale data. More recent works have exploited the power of diffusion models [128, 129] for hand pose estimation.

[130] proposed a challenge to measure the generalization of HOPE models on unseen poses, shapes and objects for different input modalities. The main finding is that all methods fail to accurately predict grasps on object shapes that are out of distribution, *i.e.* when the object category was not seen during training. This seems to indicate that most methods boil down to retrieval and memorization of training set grasps. Regarding this generalization to unknown objects, methods with a novelty on the training procedure have been proposed, such as [131] with a pose dictionary module to learn with limited data in a semi-supervised fashion, which yields more robust predictions than SOTA deep learning architectures trained on fully-labelled data, although it is not evaluated on unknown objects. Akin to this method, [132] propose a novel online data enrichment strategy for articulated hand-object pose prediction which allows a baseline to outperform the state of the art. It explores the pose space via optimized sample synthesis to enrich the data, thus making it model-agnostic, however, this only takes effect at training time, hence not allowing adaptation to unknown objects. Finally, [133] proposed to improve the accuracy of an off-the-shelf method by learning to denoise its output in a physics simulation with reinforcement and imitation learning. The method produces refined hand and object pose predictions for virtual reality and in-the-wild applications.

### 2.6.3 Hand and object reconstruction

Compared to predicting the hand pose from images, reconstructing hands in 3D requires little engineering effort [134, 135, 5], unless high-fidelity reconstruction is desired [136]. Because of advances in deep learning, pose prediction and modern hardware, modern neural network architectures with larger capacity are doing the heavy lifting [127, 5]. However, the field of 3D reconstruction also includes high-fidelity reconstruction of the in-hand object [137, 106, 138], which is an entirely different task from hand pose prediction. The field of Computer Graphics saw recent paradigm shifts with the advent of light fields [139] and neural rendering in general [140]. With these recent developments came more powerful techniques to achieve 3D reconstruction, hence the more active research area of HOI reconstruction in recent years. In this section, we cover the literature on hand-object 3D reconstruction, emphasizing modern techniques and high fidelity. While there is some overlap with the previous section, we here focus on methods that reconstruct both the hand and the hand-held object.

A large body of works in hand-object 3D reconstruction uses implicit functions to model the object shape and MANO [120] for the hand. Given the growing availability of rich annotated HOI datasets, these methods are typically single-image-based and trained from 3D data. One of the first works to introduce the Signed Distance Field (SDF) for HOI modelling is Grasping Field [28]. The authors propose an implicit neural HOI prior that can generate grasps or infer them from images. However, the method assumes the object mesh to be given, which thus restricts its use to hand mesh reconstruction only. The concept of SDF-based modelling was then taken further by [141] who found an important yet obvious finding: the hand shape and pose is a useful constraint to recover the object shape. The authors propose to use an off-the-shelf method to recover an initial hand pose and condition the object SDF on it, yielding impressive results. [142] took this idea one step further and conditioned the object SDF on an estimate of the object pose as well, with additional temporal consistency from monocular videos. Around the same time, [143] introduced a directed distance field to alleviate the poor fidelity of SDFs for local hand-object cues, using ray casting and binary intersection labels. Lastly, [144] proposed to constrain the object shape even further by explicitly modelling hand contacts. The authors used a Graph Convolutional Network (GCN) to condition an SDF decoder on features obtained from 3D convolutions of contact points and an initial hand mesh. Explicitly modelling the hand-object interaction seems to be a global trend across the entirety of HOI applications.

Before the NeRF method [139] took the field by storm, there has been an impactful attempt at reconstructing hand-object pairs from videos without relying on 3D annotations [145]. The authors proposed a method that leverages photometric consistency across time in an object manipulation video with sparse 3D annotations. The method achieved impressive results, even with 1% of video frames annotated. However, it relies on the full knowledge of the object mesh, and effectively relaxes the problem to parametric shape and pose prediction. In this direction, [146] used a 4D NeRF to reconstruct a hand-held object from a monocular video, akin to the multi-view paradigm often seen with NeRF-based methods. However, the method assumes the hand pose to remain mostly fixed. In the same vein, other methods have used NeRF-based neural rendering to learn implicit hand-object shapes, without relying on MANO [147, 148]. In particular, [148] achieves high-fidelity 4D hand-object reconstruction, handling both articulated hand and object, via two neural implicit models. Each model fits a canonical scene (for either the hand or the object) which represents distances and colours. The authors demonstrate how distances can be converted to point densities such that the implicit models are trained to render images by compositional neural rendering. While the results are impressive, the method converges in about 24 hours.

To summarize, implicit neural representations seem to dominate the field of HOI reconstruction, with the data availability and compute budget driving the choice of method. However, other works have shown that predicting vertices is still a viable option [149, 150], and some works have leveraged the fame of diffusion models [151, 30]. In particular, [150] introduced an important finding for HOI modelling: unifying hand and object features in a hyperbolic space yields better results than typical Euclidean-space-based methods, mostly due to metric learning. While the methods discuss in this section take on different approaches, it is not obvious whether this particular task can be well achieved with a general prior interaction model. In the next section, we look at HOI generation, a closer task to direct HOI modelling than 3D reconstruction.

### 2.6.4 Generating hand-object interactions

Interaction generation, split into static and dynamic domains, has received much attention recently. In the static domain, the goal is to synthesize realistic grasps, *i.e.* the hand pose, for a given object. The dynamic domain is more challenging, as the goal is to generate an entire object interaction sequence,

which typically involves an approaching hand grasp and some dexterous object manipulation. In this section, we briefly introduce the most recent works in static and dynamic manipulation generation.

### Static grasp synthesis

In the static domain, many classic methods use physical constraints to satisfy realistic grasps [152, 153, 154, 155, 156]. Newer methods take a learning-based approach and use big datasets of hand-object interactions to learn grasps [9, 157, 79, 80, 8, 158, 159, 160]. These often generate the pose parameters of a model directly [79, 80], or estimate an implicit representation for the grasp [159]. Grasping Field [159] is one of the first works to introduce implicit neural representations for HOI modelling, where the hand and object shape and pose are represented in 3D via the Signed Distance Field (SDF). The authors propose two variants of the model, where one is conditioned on images and given object mesh, and the other is conditioned on the object mesh only, for reconstruction and generation tasks.

Most static grasp synthesis methods follow a two-stage approach, which was first made popular by GRABNet [9]. In the first stage, a coarse grasp is sampled from a learn distribution of hand poses, typically as MANO [120] parameters. In the second stage, the grasp is refined, either by a feedforward network or by stochastic gradient descent (SGD) with some error signal. For the first stage, GRABNet learns a direct mapping from object geometry (encoded in the BPS representation [64]) to a grasp in MANO parameters. Then, the authors propose a refinement network trained to denoise grasps, such that the coarse hand parameters are refined after three forward passes. Later methods started shifting towards explicit contact modelling. Thus, the denoising approach of GRABNet was followed by ContactOpt and S2Contact [1, 161], where both methods predict the hand and object contact maps from noisy grasps, before refining the hand pose via SGD with a contact alignment loss. However, these methods are both applied to the task of grasp denoising and do not allow to sample coarse grasps in a first stage. A compromise was proposed by GraspTTA [8], where the authors proposed a model to learn a coarse grasp distribution as in GRABNet, and another model to regress the object contact map. Finally, a test-time adaptation stage refines the initial grasp with a contact alignment objective. More recent works have followed the trend but incorporated more information into the contact maps. For instance, ContactGen [162] models per-finger contacts on the object surface. The authors introduced an object-centric contact representation which encodes contact locations, hand

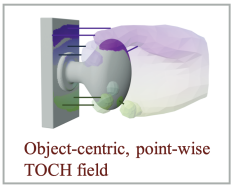
parts and directions. Those are learned as a cascade of 3 decoders, and the methods skips stage 1 to entirely sample the three stages of the contact map. Finally, Contact2Grasp [163] took this idea further and skips stage 2 by learning the mapping from the contact map to the hand pose. The method becomes a single-pass network, where the contact map is but an intermediate representation. In addition, the authors proposed a clever TTO algorithm that is penetration-aware and only affects the penetrating fingers.

**Dynamic manipulation synthesis**

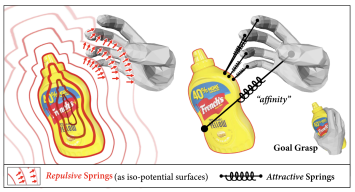
Similarly to static grasps, dynamic manipulation synthesis methods define contact constraints and use optimization to satisfy them [164, 165, 166, 167, 168, 169]. For better motion realism, recent methods use reinforcement learning (RL) for hand grasp generation [170, 171, 172, 173, 174, 175, 176]. However, as seen for static grasp synthesis, the generated manipulation sequences are mostly inaccurate and require further refinement.

To further refine grasps there exist optimization-based methods [27, 29, 177] or learning-based ones [178, 179]. For the former, ManipNet [29] proposed ambient interaction sensors to model hand-object interactions. These sensors represent local hand-object cues with a voxel grid and signed distances from the hand to the object surface. ManipNet is trained to output the refined grasp and new signed distances for the current frame, given a sequence of coarse hand trajectories with an object shape. The predicted signed distances are then used to refine the predicted pose with test-time optimization (TTO). Instead of directly operating on the poses, TOCH [27] proposed to refine a novel explicit interaction field from noisy hand motions, and then use it in an optimization process to refine the hand poses. However, TOCH was applied to manipulation denoising, as a dynamic equivalent to ContactOpt and S2Contact [1, 161]. More recent methods took inspiration from these works, such as GEARS [180] and GRIP [178], but opt for a learning-based refinement approach instead of optimization-based. GEARS introduced spatio-temporal attention to predict joint displacements from a representation similar to ManipNet's. Similarly, GRIP employed ManipNet's representation with improvements in expressivity using the BPS representation [64] as well as temporal smoothing in latent space.

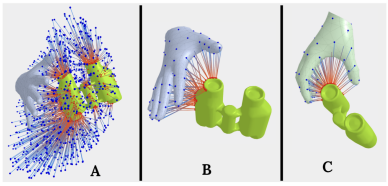
Finally, recent methods have proposed manipulation priors learned with diffusion models for both




(a) TOCH field.




(b) Contact Potential Field.




(c) GRIP representation.

Figure 2.6: Examples of recent HOI representations (images taken from [27] for TOCH, [37] for CPF, and [178] for GRIP).

static [30] and dynamic generation [181]. In particular, G-HOP [30] introduced an interaction grid where the object and hand are represented as an SDF and fused together in a grid. A diffusion model is then trained to learn the distribution of such interaction grid, conditioned on text to generate according hand-object grasps. The obtained prior can be used for static grasp synthesis and 4D reconstruction of interaction clips. The field of HOI modelling through generation is particularly relevant to this thesis, as it involves contacts and object affordances modelling for the learning of a rich interaction prior, as G-HOP so demonstrates. In the next section, we narrow down the focus to the HOI representation itself, which has recently come as the foundation for HOI generation tasks.

### 2.6.5 Interaction representation

In the early days of HOI modelling, the literature showed a growing trend to model contacts as a way to improve pose accuracy, generation realism, and physical plausibility in general. Thus, a large body of works focused on modelling contacts either as a classification problem with per-vertex contact probabilities [1, 161, 35, 8, 161], or more recently as a distribution learning problem with entire contact map sampling [162, 163]. However, these approaches are limited in how much they can express about hand-object interaction. In practice, these methods cannot represent approaching hands, full hand poses where not all fingers are in contact, or dynamic manipulation. To alleviate these issues, richer representations arose in the field.

As was seen with grasp and manipulation synthesis, explicit & expressive HOI representations are increasingly gaining traction in the field. One of the earliest instances of continuous and implicit HOI representations is Grasping Field [28], which introduced a semantic SDF with hand-parts labels. However, it is implicit and still relies on whole hand and object point clouds as inputs, leading to

a high-dimensional model. With the experiments on contact modelling of the ContactPose dataset [23], the authors showed the importance of feature selection and engineering for the performance of neural network models in HOI. In this direction, a number of subsequent works followed. Yang et al. [37] proposed Contact Potential Field (CPF) with coarse anchors and an innovative physics-based spring-mass system on the object surface (see Fig. 2.6b). Despite its promising direction, it involves minimizing relatively intricate energy functions, proving time-consuming to optimize at test-time. In addition, the representation is limited to a short range near the object surface, and thus cannot represent full hand poses. To solve this issue, TOCH [27] proposed to represent the hand pose from the object's surface, similarly to CPF, via ray casting and intersection detection. TOCH defines a field which encodes the hand shape and pose as distances and coordinates in the MANO canonical frame (see Fig. 2.6a). The result is a rich representation able to transfer grasps by transposition to another object, with the ability to encode approaching hands. However, the use of ray-casting from the object's surface limits the expressivity of TOCH for the occluded parts of the hand. Additionally, TOCH fields are not fully differentiable and expensive to compute. A similar work followed, named CAMS [177], which proposed to encode every finger for each contact point on the object. CAMS produces finger embeddings with distances as well, in addition to normal vectors and joint-to-joint angles. This results in a richer representation allowing to generate realistic grasps via fully-continuous TTO. A recent collection of works took a recent route to model dynamic manipulation using static features [29, 180, 178]. ManipNet [29] originally formulated local hand-object cues with a voxel grid and signed distances from the hand to the object. Following this idea, GRIP [178] improved it via a global hand representation in canonical space with the Basis Point Set representation [64] and a targeted local object representation with distances from the hand to the nearest object points (see Fig. 2.6c). GEARS [180] sampled object points using spheres defined around the hand joints, thus relying on PointNet-based architectures, an albeit less efficient representation than distance-based ones.

The literature survey in HOI representations sheds light on new literature gaps. We identify three main limitations in the literature, and aim to address them in this thesis. The first one is the lack of continuity or full differentiability. This makes the test-time optimization algorithm rely on random restarts to avoid local minima, and is not a desirable property of the data for neural network training.

In addition, non-continuous representations tend to be more costly to compute than distance-based metrics, in particular those relying on ray casting. The second limitation is the lack of a single semantic space for the hand and object representations, which hinders feature extraction and abstract modelling. Finally, the third limitation relates to the interaction range between the hand and the object, and to how much of the hand is represented. With object-centric representations, certain fingers may not be represented due to self-occlusions of the hand.

To summarize, the field of HOI representation is rapidly evolving towards richer and distance-based hand-object correspondences. Distances from basis points attract the most attention, as they are efficient to compute and simplify the learning for neural networks. However, Signed Distance Fields are still a popular choice in many areas of HOI modelling, especially for 3D reconstruction. The recent G-HOP [30] method introduced an interaction grid to represent the hand and object shapes as a unified SDF. This is a similar approach to Grasping Field [28], to the difference that G-HOP defines it explicitly, allowing to use it as input and output in a diffusion model. This is a promising approach that very closely follows our main contribution in this thesis. While there is a growing interest in this area of HOI modelling, the recent representations suffer from common limitations. In this thesis, we address these limitations and make significant contributions to the field.

### 2.6.6 Test-time adaptation

Broadly speaking, this area relates to methods that enhance a model's performance on a specific case at test time. This was explored using meta-learning in [182] for human motion prediction, with a specialisation of MAML significantly outperforming the original on this problem. [183] derived a meta-learning algorithm for online object tracking, with state-of-the-art accuracy. Similarly, [184] propose an inner-loop-based meta-leaner for video object segmentation which greatly outperforms a baseline in accuracy and fine-tuning time. [185] directly applied MAML on gaze estimation to train person-specific models with few examples and outperform existing methods. Other methods [186, 187] use a regular training stage but adopt an online adaptation strategy for new test-cases, which implies defining appropriate self-supervision objectives. Test-time adaptation has been applied to single-view mesh reconstruction for in-the-wild videos in [188] using self-supervision and temporal consistency. Recently, in single-view mesh reconstruction, [189, 190, 191] investigate the use of learnt object priors

and propose different methods for inferring new priors on unknown objects.

### 2.6.7 Closing remarks

The broad literature of HOI modelling shows a recent push towards the coarse-to-fine paradigm for hand-object interaction, where a coarse hand pose is first generated or reconstructed, and then is refined via optimization with pseudo-ground-truth or using learning-based methods. In addition, there is a clear trend in explicitly representing manipulation and grasping as hand-object correspondences. Even for dynamic modelling tasks, static features seem to be dominant in the field. The field of interaction generation is particularly crowded with general priors that model contacts or interaction explicitly, which seems to indicate that it is important for HOI modelling as a whole. Effectively, HOI generative models act as general priors that can be conditioned on a number of modalities (images, text, object meshes, etc.) to achieve any other task as a posterior model. In this thesis, we employ this approach to demonstrate the superiority of our representation.

## 2.7 Guidelines for hand-object interaction modelling

The literature survey reveals key insights into the challenges of HOI modelling from a deep learning perspective. There is a growing need for adaptability in modeling methods for the challenges of this domain. Indeed, it is reasonable to assume that adaptability is needed for generalization and compliance to constraints, without which we would only be able to perform curve fitting and distribution learning. In reality, grasps are heavily tied to not only the object geometry, but also the functionality, grasp affordances, and intent. The hard truth is that the true joint distribution of object and grasp is very complex and difficult to model end-to-end. Therefore, we need to investigate what is needed for adaptability in HOI tasks, either in terms of input representations, learning objectives or algorithms. These insights lead to the following guidelines – or desired characteristics – for the development of HOI interaction modelling:

- ☞ **Specialized inference**: To improve the model's performance on a single instance at test time, or to adapt to distribution shifts, the inference phase should specialize the model for each interaction. This can be achieved using specialized model parameters or by using an intermediate

representation to serve as a proxy objective to refine the prediction. This falls under the umbrella of *Test-Time Adaptation* (TTA) and has been partly demonstrated by our meta-learning experiments. Numerous works have shown that specialized inference can improve the model's performance across different tasks, from grasp synthesis [8, 1, 9, 163, 162] to hand-object mesh reconstruction [7, 10, 192, 151, 144].

☞ **Explicit HOI formulation**: To better guide the learning process, domain constraints should be explicitly encoded, via the learning objective or an augmented input (*i.e.* rich HOI representation). There are multiple ways of implementing biomechanical or contact-based constraints. For instance, weak biomechanical constraints can be added through the use of a parametric hand model [120], and stronger ones via skeletal kinematics [158]. For contact-based constraints, regularizers and loss functions are better suited and complementary to the former [8, 1, 192]. Incorporating these explicit constraints is akin to designing better inductive biases for the model, which can be done using rich explicit representations and suitable neural network architectures. Augmenting the input hand-object pair via rich representations results in more plausible and realistic predictions [27, 177, 180, 178, 163].

☞ **Canonical frame of reference**: To generalize to unknown object poses and geometries, and to reduce the learning task to modelling relative hand-object relationships, a canonical frame of reference should be used. Transitioning from a world coordinate system to a canonical frame eliminates the complexity of learning the spatial position of both the hand and the object with respect to the world or the camera, and instead focuses on the relative hand-object pose information. Typical canonical frames include the object-centric frame [37, 27, 180] or the hand-centric frame [136, 143, 141], or even both for independent modelling of contacts and distances [1, 8]. The use of canonical frames is now the standard in HOI modelling, as seen in SOTA methods that generalize well to novel object categories [27, 180, 177, 193, 178, 9].

☞ **More training data**: To achieve generalization over a broad range of object categories and grasp types, large datasets comprising dozens of objects are preferred. Our meta-learning experiments show that at least 6 objects are needed to leverage the benefits of meta-learning, in the context of this HOI task and with this model parameterization. This is in line with the recent

trend of using large-scale datasets for training, such as OakInk [83], AffordPose [85], DexYCB [84], ARCTIC [90], and GRAB [9] to name a few. Methods that have been trained or larger and more diverse datasets have shown better generalization across previous, smaller datasets [83].

**A case study on specialized inference** In [194], we show that TTA and specialized inference can be effective against distribution shifts. In this work, we annotate a dataset of human action recognition and full-body pose in wide-angle images and cluttered environments. Due to the nature of wide-angle images, off-the-shelf pose predictors yield unsatisfying results with unnatural poses. These poor results are emphasized in highly occluded sequences, such as when the legs of a person are behind a bed, due to the distribution shift problem. Indeed, most pose prediction methods being trained on clean datasets fail in such scenarios. To tackle this challenge, we annotate the 3D poses with the SOTA body pose predictor HuMoR[25]. This method uses specialized inference via test-time optimization to predict body pose sequences in camera space, where the camera parameters are given via calibration. The specialized inference comes in the form of model parameter optimization over a video sequence weakly labelled with 2D keyoints. To deal with the archaic nature of human actions in cluttered environments, we split each video sequences into sub-sequences of unoccluded human motion. This is achieved with a novel heuristic-based detection algorithm and with a fine-tuned human body segmentation model.

In conclusion, the proposed guidelines provide recommendations for the development of HOI models that are more adaptable, generalizable, and robust to distribution shifts. By following those, we can expect to see a general improvement over the current state-of-the-art methods in various tasks, as has been demonstrated by recent works in the field. In the rest of this dissertation, we follow these guidelines, with an emphasis on *specialized inference*, to develop a novel representation that is amenable to various HOI tasks.

## 2.8 Evaluation metrics

In our quantitative experiments, we use the following metrics to evaluate the final hand-object pair obtained from our method.

- **Mean Per-Joint Pose Error (MPJPE)/(R-MPJPE) (mm)**: L2 norm between ground-truth (GT) and predicted hand joints. We compute both absolute (MPJPE) and root-aligned (R-MPJPE) metrics. The former tells us about the position of the hand around the object, and the latter tells us about the hand grasp error regardless of the spatial pose.
- **Mean Per-Vertex Pose Error (MPVPE) (mm)**: L2 norm between ground-truth (GT) and predicted hand joints. We compute both absolute (MPJPE) and root-aligned (R-MPJPE) metrics. The former tells us about the position of the hand around the object, and the latter tells us about the hand grasp error regardless of the spatial pose.
- **Intersection Volume (IV) ($cm^3$)**: A measure of hand-object mesh penetration. It is computed by voxelizing the hand and object meshes (1mm voxels) and computing the volume of the intersecting voxels.
- **Contact coverage (%)**: The percentage of hand points that are within 2mm of the object surface. To compute it, we sample 3000 points on the hand surface, and equally on the object surface. This metric is an indication of how much contact the hand has with the object.
- **Hand contact F1/precision/recall (%)**: The precision and recall scores are measured on binary hand contact maps obtained by upsampling the MANO mesh and computing the Chamfer distance to the object point cloud. Hand vertices within 2mm of their nearest object point are considered in contact, to emulate soft tissue deformation as in [1]. A high precision means few false positives, while a high recall means many false negatives. The F1 score is the harmonic mean of both and is a measure of predictive performance.
- **Simulation Displacement (SD) (cm)**: The distance of displacement of the object in world space when applying inward forces to the hand grasp in a physics simulation. This tells how stable the grasp is, since more hand-object contact patches result in higher friction and therefore lower displacement.

# Chapter 3

# Meta-learning grasp predictors

Based on "**A new benchmark for group distribution shifts in hand grasp regression for object manipulation. Can meta-learning raise the bar?**"

Article published at the *Workshop on Distribution Shifts, NeurIPS 2022*

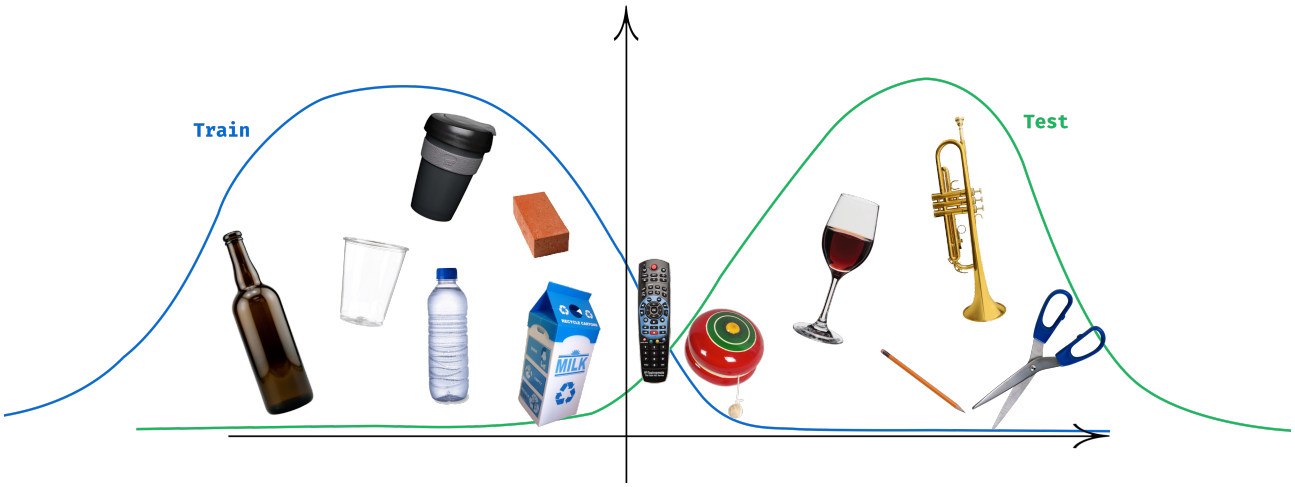



Figure 3.1: Illustration of training versus testing data distributions in object instances.

We begin with an investigation of the challenges of learning hand-object interaction representations. For this, we look at the task of joint hand-object pose prediction through the lens of group

distribution shifts. These are a special case of distribution shifts, a phenomenon where the test-time input distribution differs from the training-time distribution. When the distribution shift is consistent across groups (*e.g.* object categories, subjects, or environments), it is referred to as a group distribution shift (see Fig. 3.1). We investigate this issue with a novel benchmark for hand-object pose prediction, on which we measure the efficacy of meta-learning, a test-time adaptation framework for learning to learn.

In the following sections, we first introduce meta-learning, which is used to investigate the effect and requirement for adaptability in neural networks for HOI tasks. We then present our benchmark for group distribution shifts in hand-object pose prediction.

## 3.1 Background

Distribution shifts are one cause of the performance degradation of a model during inference. They arise when the training and testing data come from different distributions. There are several types of distribution shifts, such as covariate shift [195], label shift [195], or group shift [196]. This phenomenon is typically encountered in the context of few-shot or zero-shot learning. In these scenarios, the model is expected to accurately solve a task on unseen data, either with few examples or none. This "unseen data" is implied to be drawn from a different distribution than the training data (distribution shift), or using different classification labels (label shift), or both. In the adversity of distribution shifts, such as these scenarios, a model trained in traditional supervised learning will under-perform and its predictive accuracy will degrade. Test-Time Adaptation (TTA) is a way to alleviate this degradation, typically by updating the model's parameters at test time [197].

One approach to TTA is meta-learning, for *learning to learn*, where a machine learning model is trained to learn new tasks with few data points and/or few optimization steps. Meta-learning is typically applied in the few-shot classification scenario, although some methods are applied to regression problems. For the latter, three main categories: optimization-based methods, memory-based methods, and lastly generative models for functions. Optimization-based methods [198] were first designed with a Long-Short Term Memory (LSTM) architecture to learn the update rule of a learner network. Later, inner-loop-based methods were introduced by MAML [199] which removes

extra parameters and learns a good initialization for the learner to quickly adapt to novel tasks via second-order gradient descent. Improvements were proposed by [200] regarding the instability and generalization. [201] analysed the extent to which MAML only reuses learnt features and proposed to apply inner-loop optimization on the head of a CNN only. Generalization was improved by [202] via latent embedding optimization, at a high computational cost, while [203] tackled the first-order version of MAML to improve its computational cost and generalization. A different paradigm is memory-based methods [204, 205], which adopt a strategy closer to human-like learning by storing representations of concepts in modifiable memory cells. Neural Processes [206, 42, 207], introduced as a blend between Gaussian Processes and Neural Networks, model uncertainty. They are especially suited to regression problems by meta-learning functions from sets of observations, similarly to [208, 209]. However, by using latent space aggregations they tend to underfit the context for high-dimensional tasks. This is underlined by [73] proposing to meta-learn an SDF for implicit shape representation using MAML. For an extensive review of the literature on meta-learning, we refer the reader to [210] and [211]. In this chapter, we focus on optimization-based meta-learning, which is the most versatile and widely used approach.

### 3.1.1 Optimization-based meta-learning

In meta-learning, "the goal of the trained model is to quickly learn a new task from a small amount of new data" [199]. Hence, the idea of a task is central to meta-learning. A task is defined as $\mathcal{T}_i = \{p_i(\mathbf{x}), p_i(\mathbf{y}|\mathbf{x}), \mathcal{L}_i\}$ and comprises a distribution of samples $p_i(\mathbf{x})$, a distribution of ground-truth labels $p_i(\mathbf{y}|\mathbf{x})$ and a loss function $\mathcal{L}_i$. By targeting similar tasks, a model can make use of the shared structure in the data across tasks for a meta-learning approach. This allows the model to learn a generic – or across-task – prior, while having the plasticity required to adjust its parameters for task-specific knowledge.

Meta-learning is often presented in the context of few-shot learning (also known as $K$-shot learning), where the goal is to optimize a new objective from a few examples, referred to as the context set. During training, the meta-learner optimizes the learner on various objectives from their context set (of $K \leq 10$ samples typically) and evaluates its performance on a validation set of $Q$ unseen samples called the query set. The model's performance on the tasks aggregate of the latter is a measure of its ability to

learn quickly.

In a single-task supervised learning context, we consider a single dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})_K\}$ of pairs of images and labels. The optimization problem is thus to minimize the loss over the entire dataset to find the optimal model parameters $\theta$ as:

$$\min_{\theta} \mathcal{L}(\theta, \mathcal{D}), \tag{3.1}$$

where $\mathcal{L}$ is the *Mean Squared Error* for regression tasks. It is minimized such that the update rule at each optimization step is given by

$$\theta_{t+1} = \theta_t - \alpha_t \nabla f_{\theta_t} \tag{3.2}$$

with a static or adaptive learning rate $\alpha$ for each step $t$ and the optimized function $f$ parameterized by $\theta_t$ at a step $t$. The most characteristic aspect of the meta-learning approach is that the update rule is instead learnt end-to-end, such that the parameter update can be reformulated, as defined by [198], to

$$\theta_{t+1} = \theta_t + g_t(\nabla f_{\theta_t}, \phi) \tag{3.3}$$

where $g$ is the learnt optimizer parameterized by $\phi$ and $f$ is the optimizee.

Typically, an optimization-based meta-learning system is composed of an outer block called the optimizer, which learns the update rule of the base learner, and of an inner block called the optimizee, which is the base learner that directly learns the task. The optimizee is thus optimized by the optimizer block to rapidly learn novel tasks. A general definition of the optimization problem for meta-learning is given as

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_i(\theta, \mathcal{D}_i) \tag{3.4}$$

for $N$ tasks where each task is a small dataset $\mathcal{D}_i$. For MAML and other optimization-based meta-learning algorithms, this meta-objective becomes

$$\min_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})}^{N} \mathcal{L}_{\mathcal{T}_i}(f_\psi) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})}^{N} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}) \tag{3.5}$$

for $N$ tasks sampled from $p(\mathcal{T})$, with post-adaptation parameters $\psi$. As for the update rule, it is

defined as such:

$$\theta_{t+1} = \theta_t - \beta \nabla_{\theta_t} \sum_{\mathcal{T}_i \sim p(\mathcal{T})}^{N} f_{\theta_t - \alpha \nabla_{\theta_t} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_t})}. \tag{3.6}$$

In most cases, the learnt optimizer is only effective during training. However, some algorithms propose to learn per-layer and/or per-step learning rates for SGD as the learnable optimizer, and use them at test-time during the adaptation phase [200]. For the remainder of this chapter, the term *meta-learning* will refer to optimization-based meta-learning and the derivations of MAML in particular.

## 3.2 Benchmarking grasp prediction through the lens of group distribution shifts

Joint hand-object pose regression methods, and hand pose regression in object grasping, are not commonly benchmarked for in-the-wild data or a wide variety of object grasps [212]. They aim to generalize to unseen poses on the same objects by learning from a large collection of diverse poses and grasps [82, 84, 86]. This may be due to the cost of annotating 3D pose and the limited availability of 3D scanned objects, which encourages researchers to reuse objects' meshes. In addition, synthetic datasets of realistic object grasps are hard to produce and lead to domain adaptation challenges. However, such models have limited use if they are only accurate for a limited set of objects.

To accurately predict the pose of a hand occluded by an object, a deep learning model must learn prior knowledge of various grasps on diverse objects. Neural networks are usually trained to keep a balance between generalization and specialization as their knowledge is frozen when deployed. However, temporarily learning at test time would allow them to specialize their parameters for a specific object grasp while remaining generalizable by forgetting this specialized knowledge. TTA methods demonstrate improvements in regression accuracy for distribution shifts and in-the-wild data in other domains, such as human pose estimation [187], gaze estimation [185], mesh reconstruction [188] and video object segmentation [184]. We propose to achieve this on the grasp prediction problem with a meta-learning algorithm, where the goal is to quickly learn new tasks from a few examples at test time [213, 214, 198]. We then evaluate this method on a novel benchmark for group distribution

shifts in hand-object pose regression for object grasping. How does the performance of a CNN pose predictor evolve as the test set grasps diverge from the training set? We answer this question and look at the advantages and limitations of meta-learning for this application via experiments and empirical analysis.

In this section, we:

1. Reformulate grasp prediction in the context of multitask learning such that meta-learning can be applied, as in [185] with person-specific gaze estimation.

2. Propose a new benchmark for group distribution shifts associated with object categories in hand grasp regression based on the DexYCB dataset [84].

3. Prove that meta-learning is effective at tackling these group distribution shifts for hand grasp regression.

In the evaluation of our method, we find an increase in accuracy for unknown objects from 6 training objects upwards. We compare the relative error of the meta-learning with a baseline on our benchmark, and provide a comprehensive analysis of the limitations of the method.

### 3.2.1 Evaluating the effectiveness of meta-learning

In this section, we explain the task set creation process necessary to apply meta-learning to the grasp prediction problem.

#### Task set creation

To cast pose prediction as a multitask problem and apply meta-learning, we must create a dataset of tasks from a dataset of samples. For such a task set, we require that:

- Each task is composed of a support set of $K$ randomly sampled images corresponding to one manipulation sequence of an object by one subject, and a query set of $Q$ distinct random samples from the sequence.

- A series of $\Omega$-objects-left-out splits is used, where $\Omega$ is the number of objects absent in the training split and placed in the test split. Thus, all images associated with $\Omega + \min(5, \frac{\Omega}{2} + (\Omega$

Table 3.1: **Image samples per split as the number of unseen objects grows in the test split** – As the test set size increases for each added object, the training set size decreases proportionally. We cap the validation objects so as to maximise the number of dataset versions.

| # objects in test split | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|
| Training | 242, 704 | 221, 688 | 183, 104 | 163, 608 | 117, 824 | 98, 704 | 80, 072 | 58, 016 | 42, 288 |
| Validation | 60, 248 | 62, 552 | 79, 232 | 82, 240 | 105, 040 | 101, 744 | 99, 952 | 100, 896 | 99, 072 |
| Test | 100, 960 | 119, 672 | 141, 576 | 158, 064 | 181, 048 | 203, 464 | 223, 888 | 245, 000 | 262, 552 |

mod 2)) objects are removed from the training split: the samples associated with 5 or fewer objects are used for validation while the ones for other $\Omega$ objects are used for testing. This ensures that there is no overlap between the training, validation and test splits.

- For all values of $\Omega$, the objects are randomly sampled with a fixed random seed for reproducibility across experiments.

To assess the generalization of a grasp prediction model to unseen objects, we need a dataset of 2D images of the largest amount of manipulated objects with full 3D hand joint annotations, as well as 6D object pose. The review of [212] provides an extensive list of available datasets, but their limited amount of objects was the main factor in choosing the unlisted and recently released DexYCB [84]. We began the experiment with the HO-3Dv3 dataset [82, 81] which contains 10 different objects. This limitation hindered the experimental results, as more objects needed to have enough training, validation, and testing object categories since they could not overlap in each split. We further used the DexYCB after realizing that the results were not conclusive. It contains 582, 000 annotated frames with 10 subjects and 21 objects, of which the *large clamp* is not annotated and therefore removed. We discard all frames where the hand is not in contact with the object, such that at least two fingertips are within the object bounding box. That oriented cuboid is computed using Trimesh [215] on the associated object mesh, before applying transforms. Table 3.1 shows the final size of each split for every variant of the dataset, and Fig. 3.2 shows examples of tasks built from DexYCB.

We use the DexYCB dataset to create our task set and run our experiments, although the procedure applies to any object manipulation dataset (see the survey of [212] for an overview of hand-object pose datasets). In the following paragraph, we provide a similarity study of object grasps from this dataset

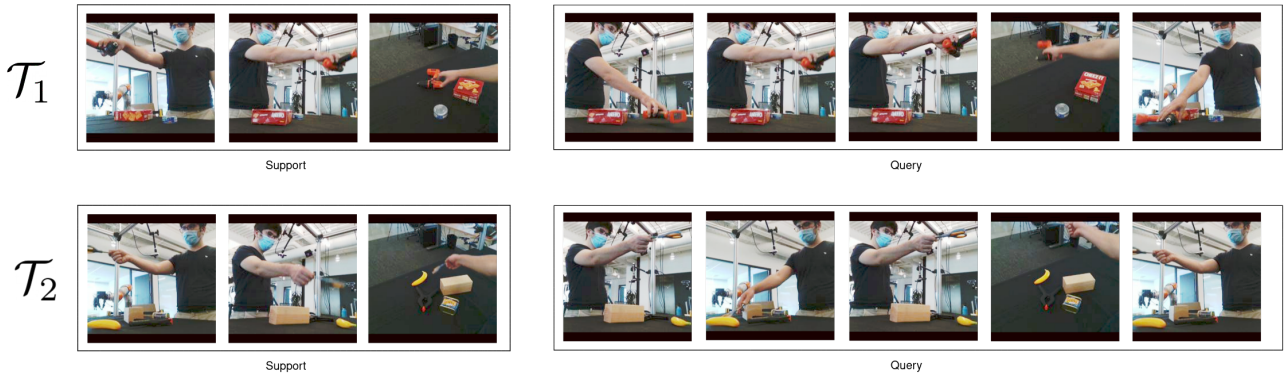



Figure 3.2: **Example of tasks built from the DexYCB dataset** – $\mathcal{T}_1$ and $\mathcal{T}_2$ are two tasks composed of a support set of 3 training images, and a query set of 5 test images for 3-shot learning. $\mathcal{T}_1$ is the manipulation of a drill, while $\mathcal{T}_2$ is the grasping of scissors. The tasks contain several viewpoints but only one subject and intent of use in their support and query sets.

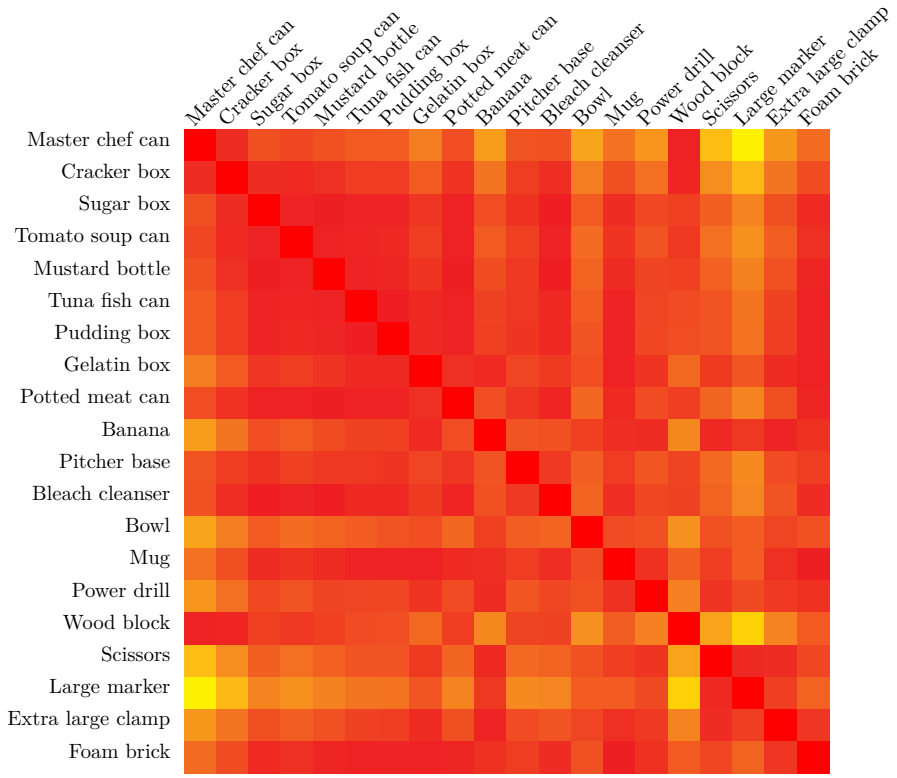
and more implementation details.

**The DexYCB dataset** In the DexYCB dataset, there are approximately 20K samples per object with 100 manipulation sequences each, providing the across-task diversity required for the training of neural networks and to minimize overfitting. Meta-learning shows its power over standard gradient descent optimization in the presence of across-task diversity, such that each task is an objective requiring specialization. We analyse the diversity of object-specific grasps using the Generalized Procrustes Analysis and the Procrustes distance to compare mean hand shape similarity. It is defined as the sum of the squared vertex distances:

$$d = \sum_{i}^{N} [(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2 + (z_{i1} - z_{i2})^2] \tag{3.7}$$

for two 3D shapes with $N$ matching vertices. The heat map of hand poses similarity in Table 3.2 shows relatively few light cells, indicating that most grasps are similar to each other. This is partly because subjects may grasp the same object in various and unusual ways, thus resulting in an uninformative mean hand shape. Due to this, we construct tasks from individual manipulation sequences such that there is only one grasp per task (see Section 3.2.1).

Table 3.2: **Distance heat map of mean hand shapes** – Heat map of Procrustes distances of mean hand shapes obtained via Generalised Procrustes Analysis; similarity ranges from yellow to red (best seen in colour). The *Large marker* induces the most different grasps in the dataset, while all box-shaped objects induce similar grasps (as seen by the more uniform upper left square). Note that all objects have several grasps associated with them, which vary depending on the intent of use of the subject.




**On the non-mutual exclusiveness of our tasks** Based on information theory, the sampled tasks must be *mutually exclusive* such that no single model can solve all tasks. In some cases where the task is identifiable (*i.e.* the classification labels are consistently the same across batches of tasks), the meta-learner may simply discard the context set and memorize the predictions to minimize the error. In other cases, such as predicting the pose of an object from an image, if the task can be solved from the target inputs alone, then the model may minimize the loss while discarding the context set by memorizing a function that solves all the tasks. As a result, we get a zero-shot model that makes average decisions for all tasks without adaptation, and thus poor performance on novel tasks. This is referred to as memorization overfitting [216]. This is a concrete risk in our experiments, as nothing is

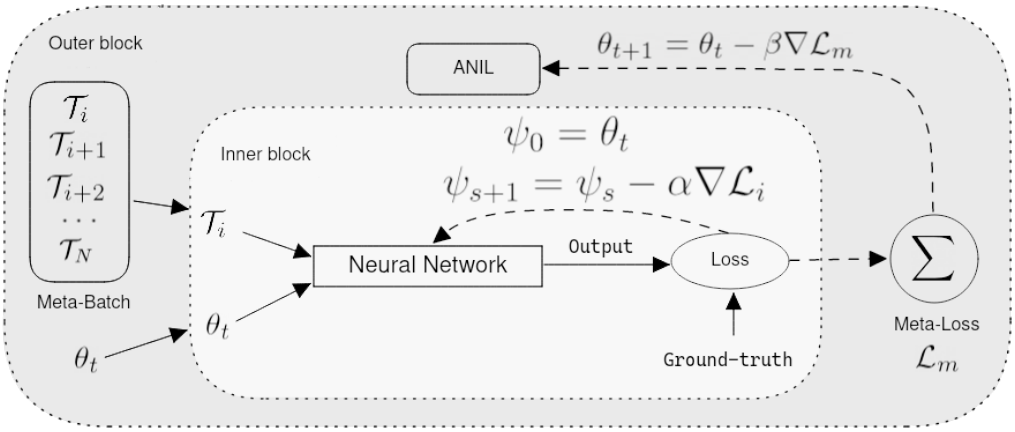


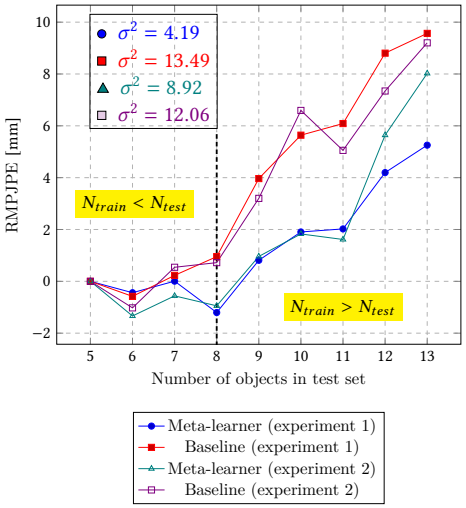

Figure 3.3: **An overview of the meta-learning system** – The outer block optimizes the inner block's parameters by back-propagating the second-order gradients of the meta-loss. This cost function is the cumulative loss of the inner block on all $N$ tasks, computed with the adapted parameters $\psi_S$ obtained after $S$ optimization steps and with the same initial parameters $\theta_t$.

preventing the optimizee from learning a function that predicts the hand and object poses from the target inputs alone. One solution is to reduce the information flow between the input $x$ and the target $y$ in the query set $\mathcal{Q}$, such that the meta-learner is forced to use the information in $\mathcal{Q}$ [217].

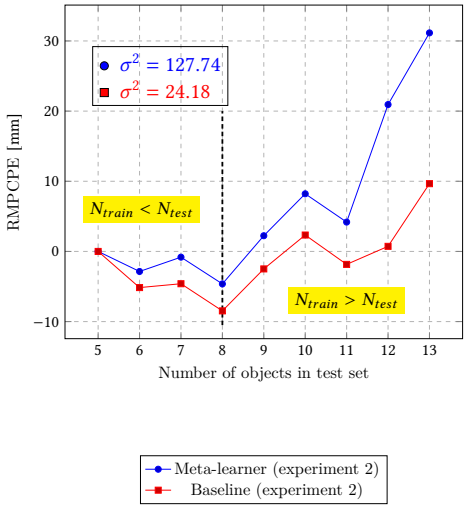
### Experimental framework

To evaluate the effectiveness of meta-learning at tackling object group shifts and dealing with highly imbalanced dataset splits, we apply ANIL [201] with the improvements of MAML++ [200] to a ResNet18 [218] baseline. As seen in Fig. 3.3, our framework consists of an outer block (ANIL) and an inner block (ResNet18).

**Outer block** The outer block trains the inner block to learn good parameters that are amenable to fast adaptation with few gradient descent steps and examples. It is based on MAML and is a re-implementation of ANIL [201] which brings performance improvements to the former by only adapting the head of the network and learning a common feature extractor for all tasks. We further implement some of the improvements proposed by [200], such as the Multi-Step Loss to improve the gradient stability and Derivative-Order Annealing to reduce the second-order derivative cost in early epochs. Furthermore, to deal with the meta-overfitting phenomenon that arises from the non-mutual-




(a) **Hand only** – For both experiments, the error increases less steeply and has lower variance $\sigma^2$ as the number of test objects grows for the meta-learner. The latter is better at adapting to dataset imbalance.



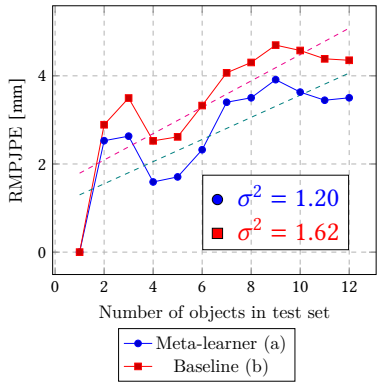

(b) **Object only** – Both methods exhibit a high variance but the meta-learner is worse by an order of magnitude. The latter isn't able to reduce the object pose error.

Figure 3.4: **Relative Mean Per-Joint Pose Error (RMPJPE) & Mean Per-Corner Pose Error (RMPCPE) as functions of the imbalance level** – All curves are aligned to the origin to show the relative changes and compare the generalisability of both methods. The more objects are in the test set, the more deprived the training set (see Table 3.1 in Section 3.2); the dashed line shows the equilibrium between the two. Both experiments are plotted in Fig. 3.4a since they both regress the hand pose, while only experiment 2 predicts the object pose and is plotted in Fig. 3.4b on its own. The meta-learner is overall better at handling object class imbalance, and thus generalisability to unknown grasps. However, this behaviour is not seen for the prediction of the object corners coordinates.

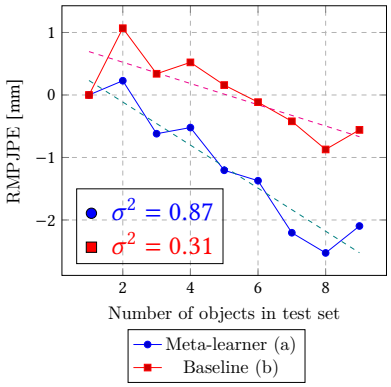
exclusiveness of our regression tasks, we implement the regularization method of [217]. We use the *learn2learn* [219] library to facilitate the implementation and minimize mistakes.

**Inner block** For the inner block, the baseline pose regression neural network, we use the same backbone for both experiments: ResNet18 [218] pre-trained on ImageNet [220]. The choice of a shallow ResNet architecture is motivated by the speed of training and low memory footprint for meta-learning. It allows a fair comparison without being impacted by the limitations of meta-learning methods regarding more complex architectures. Combining these more complex and efficient architectures with meta-learning should be the focus of future work. For a fair comparison, the baseline is trained with weight decay regularization to minimize overfitting.

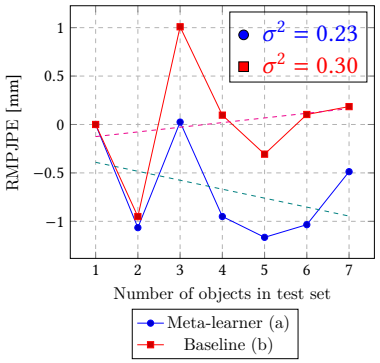
The baseline is trained with a batch size of 64 and a learning rate $\alpha = 10^{-3}$ for 100 epochs. For




(a) **3 training objects** – Both methods have a similar variance and error curves, with similar slopes ($p = 0.61$). There is insufficient prior knowledge for the meta-learner to be effective.




(b) **6 training objects** – The error decreases more steeply for the meta-learner ($p = 0.02$), although its variance is higher. There is sufficient prior knowledge for the meta-learner to be effective.




(c) **9 training objects** – The meta-learner has a lower variance but both error slopes are similar ($p = 0.38$). Both models generalise well with a non-significant improvement from the meta-leaner.

Figure 3.5: **Relative Mean Per-Joint Pose Error (RMPJPE) as function of the test split for experiment 1 with 3 training set sizes** – The dashed lines are the linear regression models.

the meta-learner, we use a meta-batch size of 8 and found learning rates $\alpha = 10^{-5}$ and $\beta = 10^{-2}$ via hyperparameter search using *Weights & Biases* [221], with 300 epochs. Both methods are trained with normalized inputs according to ImageNet's statistics since we use a pre-trained ResNet backbone, and align the ground-truth poses to the root of the hand. Using wrist-aligned pose labels greatly improves the convergence of both methods, and the absolute position in 3D space is not in the scope of this chapter. For the meta-learner, we use $K = 10$ for the support set, $Q = 50$ for the query set, and 15 adaptation steps. This implies that during evaluation, there are $K \times (\frac{T}{K+Q})$ samples – with $T$ being the size of the test split – virtually removed from the test split, which is a slight disadvantage for the baseline. It is acceptable because the goal is to measure error changes relative to various test splits. In fact, the pose prediction error of the meta-learner is roughly 1.5× that of the baseline for both experiments. When evaluating the meta-learner, we average all metrics over 5 runs to account for the run-time stochasticity coming from the support and query sets sampling. For all experiments, we normalize the error against each model's performance on the easiest setting (5 test objects) to compare the relative error changes.

With this framework, we design two experiments: (1) we evaluate the relative error of the 3D

hand pose regression for manipulation samples of unknown objects, (2) we reiterate on the joint hand-object pose regression by incorporating the object bounding cuboid coordinates in the targets, as in [6, 132, 124].

**Experiment 1: hand pose only** Fig. 3.4a shows the meta-learner's ability to deal with object group shifts on a macro scale. The size of the training set decreases inversely and in proportion to the size of the test set, thus the error is expected to go up for the baseline as the test set grows. The error of both models rises steeply from 8 objects in the test set, where the gap starts to widen, but less steeply for the meta-learner. It corresponds to 8 training objects as well (with 4 for validation); a more or less balanced dataset. By fitting a linear regression model on both curves, we prove that the difference in the two slopes is statistically significant with a $p$-value of 0.0031.

We look at the micro scale by freezing the training split for 3 levels of prior knowledge, and progressively adding objects to the test split; the results are shown in Fig. 3.5. Both methods are trained and evaluated on 3 train/test splits; the average curve is shown for the 3 sizes. The meta-learner and the baseline behave identically in Fig. 3.5a since the training set is too small for the meta-learner to collect sufficient prior knowledge for adaptation: 3 training objects are not sufficient prior knowledge. As the training set size and variability grow for Fig. 3.5b, the meta-learner reduces the error significantly better than the baseline: 6 objects are enough prior knowledge. However, with 9 training objects in Fig. 3.5c, both curves have a similar horizontal slope and a variance roughly a third of that of Fig. 3.5a. In short, the baseline and the meta-learner generalize equally well. For this case, the improvement of the meta-learner may nonetheless be revealed with more test objects if available. The findings are different from in Fig. 3.4a since the models are here compared in terms of accuracy for a given training set size, and not their ability to deal with dataset imbalance.

**Experiment 2: joint hand-object pose** Analogously to the previous experiment, Fig. 3.4a shows a mild improvement in the hand pose regression for the meta-learner when trained for the task of joint hand-object pose regression. This improvement however is lesser than for the first experiment, as reflected by the smaller gap between the two curves and a $p$-value of 0.3186. The null hypothesis cannot be rejected: the meta-learner shows no significant improvement in generalization. Fig. 3.4b confirms this phenomenon and shows even worse results than the baseline. We tried increasing the

network's capacity but obtained similar results with underfitting. Therefore, we can hypothesize that this joint regression causes interference in the adaptation to both tasks, and this problem would have to be processed separately in future work. In Section 3.2.2, we run more experiments to assess the hypotheses made on the meta-learner in an attempt to explain the results.

### 3.2.2 Empirical analysis

From these experiments, we can conclude three things: (1) for hand grasp prediction, the meta-learner is better able to deal with dataset imbalance than the traditional SGD baseline, as shown in Fig. 3.4a this is especially true when using less than 8 training objects (or more than 8 test objects), (2) the meta-learner improves the accuracy with enough training data (i.e. 6 objects) on unknown objects, and (3) the meta-learner is ineffective for joint hand-object pose regression, and is even worse than the traditional baseline for object pose prediction (see Fig. 3.4b). This weakness is investigated in the following section.

#### Hypothesis 1: the model learns specialized object-specific parameters

ANIL and similar methods learn prior knowledge of related tasks during training so that they can adapt to specialized parameters for a novel task with few optimization steps [211]. But does it translate to learning object- or grasp-specific parameters in the proposed framework? To verify this, we analyze the significance of differences between object groups in the network's head parameters, post adaptation on the context set. We first reduce the dimensionality of the weight matrices of each layer via Principal Component Analysis. Then, we compute the pairwise distances using the cosine similarity metric. We also pre-process the embedding vectors at each layer using the Euclidean distance metric. To verify the statistical significance of the differences between groups, we employ Permutation Analysis Of Variance (PERMANOVA) [223]. It is a non-parametric multivariate variant of Analysis Of Variance (ANOVA) [224] which employs random permutations of the vectors between groups to test the null hypothesis that the group centroids and dispersions are equivalent. This null hypothesis would suggest that all parameters of the meta-learner *do not specialize according to object groups*. The results of PERMANOVA show that the null hypothesis is not invalidated in any significant way, with a $p$-value $> 0.3$, for the PCA-reduced weights of both layers. However, the analysis shows that the null

hypothesis is significantly invalidated for the embeddings of both layers, with $p$-values of $0.002$ for the first layer, and $0.001$ for the second layer. This indicates that, while the meta-learner weights do not seem to form clusters and specialize by objects significantly, there is some for of clustering observed in the embedding. However, since the raw data can be clustered according to grasp archetypes as seen in Table 3.2, it is to be expected that this structure is preserved during embedding. In addition, we compute the Silhouette scores [225] of the PCA-reduced weight vectors to obtain a quantitative measure of the quality of clustering. The Silhouette score measures cohesion within clusters and separation between clusters, with a score ranging from $-1$ for worst clustering quality to $1$ for best clustering quality. The first layer's weight vector scores $-0.03$ while the second one scores $-0.02$, confirming the previous analysis. For visual inspection, we show UMAP [222] embeddings for the high-dimensional and normalized weight vectors in Fig. 3.6. The plots show no consistent clustering between object groups and illustrate the validity of the null hypothesis.

The absence of clusters in the layer weights reveals that no object-specific information is encoded in the parameters during adaptation. It reveals that the parameters do not specialize to a specific grasp or object after adaptation, thus refuting this hypothesis. This may well be due to the *non-mutual exclusiveness* of the tasks incurring memorization overfitting (see Section 3.2.1), for which the regularizer of [217] did not help enough. Other possible reasons are the lack of diversity in the grasps of the DexYCB dataset, or the lack of strong constraints in the inner objective. To encourage specialization, the problem should be properly formulated in the meta-objective.

**Hypothesis 2: Using Oriented Bounding Box (OBB) coordinates in the training signal constrains the hand-object pose**

In experiment 2, we hypothesize that using Oriented Bounding Box (OBB) coordinates in the training signal is enough to jointly constrain the hand-object pose and that OBBs differ enough to lead away from the initialization in the loss landscape. In truth, Fig. 3.4b shows that it limits the adaptation for the hand pose and leads to worse generalization than the baseline for the object pose.

[226] define the shape as *"all the geometrical information that remains when location, scale and rotational effects are filtered out from an object"*. Therefore, we can consider that most OBBs have almost identical shapes, except for unusually thin or wide objects such as *large marker* or *pitcher base*.

Table 3.3: **Average gradient norm per adaptation step: exp. 1 vs exp. 2** – Entries follow an *exp1/exp2* format, with larger values in bold. The average gradient norm for experiment 2 is consistently larger than for experiment 1, hinting that jointly regressing the hand and object poses requires deviating further away from the initialization during adaptation.

| Step | Tomato soup can | Banana | Scissors | Foam brick | Mug |
|---|---|---|---|---|---|
| 1 | 204/**394** | 231/**638** | 270/**677** | 195/**380** | 236/**505** |
| 2 | 193/**360** | 219/**576** | 255/**613** | 185/**338** | 225/**458** |
| 3 | 182/**330** | 207/**522** | 242/**557** | 176/**302** | 214/**416** |
| 4 | 173/**303** | 197/**475** | 229/**509** | 167/**273** | 205/**381** |
| 5 | 164/**280** | 187/**435** | 217/**467** | 159/**248** | 195/**349** |
| 6 | 155/**259** | 178/**401** | 207/**430** | 152/**227** | 187/**322** |
| 7 | 148/**241** | 169/**371** | 197/**398** | 145/**209** | 179/**298** |
| 8 | 141/**225** | 161/**344** | 188/**369** | 139/**194** | 172/**277** |
| 9 | 134/**210** | 153/**321** | 180/**344** | 133/**181** | 165/**259** |
| 10 | 128/**198** | 146/**301** | 172/**322** | 127/**169** | 159/**243** |

We postulate that due to this, the adaptation phase should produce small gradients for the object keypoints and thus the hand keypoint regression should not be severely impacted. To verify this, we plot the norm of the gradients during the adaptation phase in Table 3.3. For both experiments, on the same randomly sampled test objects, the gradient norms of the first 10 adaptation steps are averaged over batches of the same task (with re-sampling of the same object manipulation query and support set). Both models were trained on the same six objects. In general, in Table 3.3 we observed a $2-3\times$ increase in gradient norms from experiments 1 to 2. Experiment 2 has consistently much larger gradients, meaning that jointly regressing the hand and object poses requires further deviating from the initialization during adaptation. This could reflect poorly learned meta-parameters. We hypothesize that this is due to the complex hand-object relationship, which is not expressed in the meta-objective, causing interference in the meta-optimization landscape. To address this issue, the constraints should be explicitly defined in the objective, or the two problems should be decoupled.

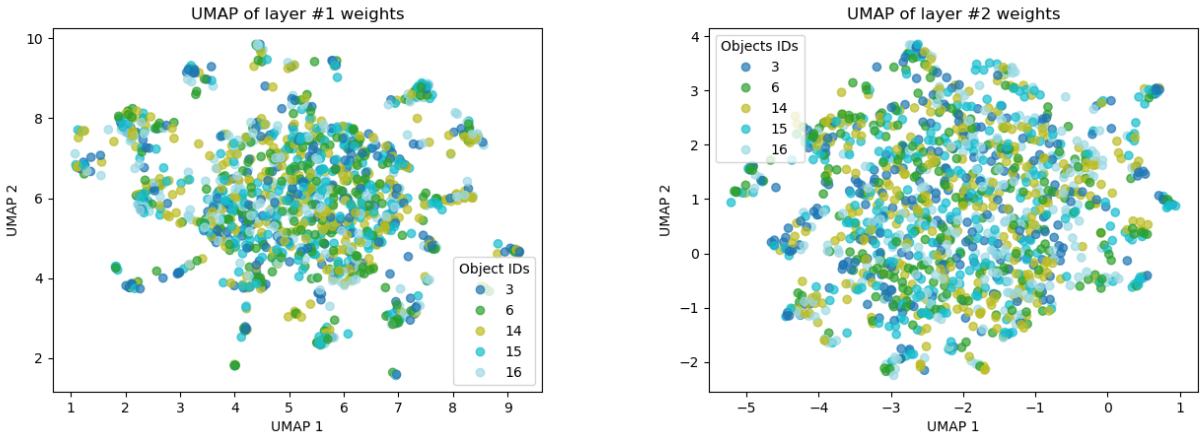
Another plausible reason for the large disparity in gradient norms between the two experiments is the use of absolute coordinates in the training signal. This may lead to a higher variance in the gradients, as the model has to learn to predict the absolute coordinates of the keypoints instead of their relative positions. To alleviate this issue, we suggest using a canonical frame of reference for the keypoints, to reduce the variance in the gradients and improve the generalization of the model.

### 3.2.3 Discussions

In this chapter, we hypothesize that applying meta-learning to a 3D keypoint regression baseline allows it to generalize to unseen objects and outperform the traditional baseline trained with SGD. To validate this hypothesis, we devised a benchmark where we gradually reduced the diversity of grasps in the training set and, conversely, increased the diversity in the test set. The meta-learner was expected to outperform the SGD baseline, where the training data are sparse. Our results show measurable improvements over the baseline, where it reduced the rate at which the error rises with fewer than 8 training objects. On the other hand, it generalized better than the baseline with 6 training objects.

For the task of joint hand-object pose regression, the meta-learner's ability to deal with object group shifts is dampened and is worse than the baseline for the object pose. In our empirical analysis shows that this may be due to interference in the optimization. Furthermore, we find that the meta-learner does not learn grasp-specific parameters during adaptation, which may impede its performance. We hypothesize that explicitly defining the problem in the meta-objective would alleviate this issue.

The findings of this chapter all converge towards the same conclusion: for effective learning of hand-object interaction models, the optimization must be constrained on several grounds. We demonstrated that learning implicit models of hand grasps specialized to object groups is challenging in the absence of explicit HOI constraints. These results indicate that following the HOI modelling guidelines of the previous chapter, *i.e.* using explicit representations instead, is a more viable way to address the first research question: *can we design new shape, pose and contact representations in order to overcome the limitations of existing representations while maintaining their expressiveness?* Moreover, with an implicit representation approach there is no obvious way to address the second research question: *can we design new models to demonstrate the expressiveness and versatility of these representations for various applications?*. In the case of the meta-learner, the internal representations are strongly tied to the learning objective, hence requiring more complex learning methods for transferability to other tasks. In the next chapter, we design an intermediate hand-object representation as a way to address these challenges, following the guidelines on HOI modelling derived from the literature Section 2.7.




(a) **Weights of first layer** – No clustering visible between objects (IDs in the legend).

(b) **Weights of second layer** – No clustering visible between objects (IDs in the legend).

Figure 3.6: **Visualisation of weights per layer** – The post-adaptation parameters of the meta-learner's head are embedded with UMAP [222]. The embeddings of the adapted weights of each layer, for experiment 1, show no signs of separate clusters for the 5 different objects in the test set (best seen in colour). This implies that they mostly encode the same information regarding the hand or joint hand-object pose for all objects.

# Chapter 4

# CHOIR:

# Representing shape and pose

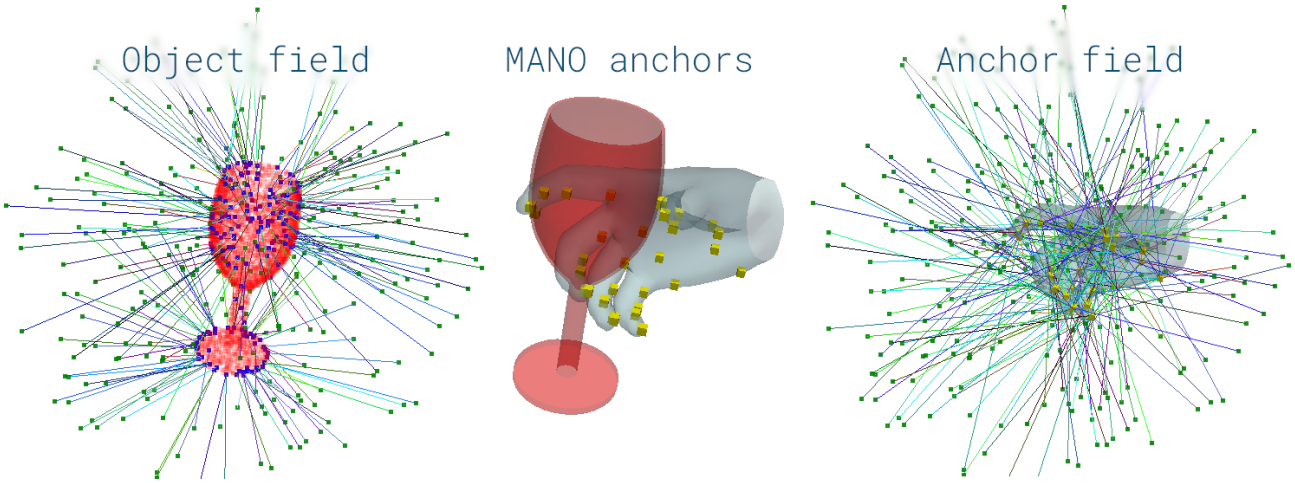



Figure 4.1: Decomposition of our Coarse Hand-Object Interaction Representation (CHOIR). (left) Each ray represents the assignment of the basis points (BPS) to the object points. The blue points on the object surface are the represented points and ray shades denote the pair-wise distance (best viewed in colour). (middle) The cubes on the hand mesh represent the MANO anchors[37]. (right) Hand field computed from BPS points to the MANO anchors. CHOIR is a concatenation of both object and hand fields in an object-centric frame.

Having established the limitations of HOI representations (Section 1.1.3), we now seek to address them with a novel representation for the shape and pose of the hand-object pair. In this chapter, we thus introduce CHOIR: a novel lightweight, continuous and fully-differentiable shape and pose repre-

sentation for hand-object interactions. We first motivate our choice of representation by discussing the limitations of current methods and using the guidelines we developed in Chapter 3. This is followed by a detailed explanation of our core novelty, and finally with experimental results on the application of static grasp refinement. This chapter closes with a discussion of the method's limitations, which leads to the next chapter.

## 4.1 Motivation

The literature in HOI modelling has been shifting towards designing representations that are more amenable to learning generalizable models with increased predictive performance. However, these representations are not without limitations. In this section, we take a close look at the recent literature in HOI representations, discuss their approach, limitations and weaknesses. We then define the scope of the novelty of this chapter.

### 4.1.1 Related works and their limitations

By design, end-to-end methods do not generalize well outside their training distribution. It was shown that their performance can be trivially improved by carefully selecting the right features [80]. There was a recent shift towards designing HOI representations which precisely encode rich grasp features and dense hand-object correspondences. These methods are a clear improvement over the previous paradigm. Notably, at the time this work was conducted, three novel HOI representations stood out: Grasping Field [28], Contact Potential Field [37], and TOCH [27].

#### Grasping Field (Karunratanakul et al. 2020)

Grasp synthesis from object geometry was first introduced by Grasping Field [28] using an implicit representation. The concept is straightforward: modelling the hand shape and pose as a semantic Signed Distance Function (SDF) of hand parts. This SDF is modelled by a variational encoder-decoder neural network trained on 3D data with hand part labels. This is a weak form of *explicit HOI formulation* (from our guidelines in Section 2.7). Grasping Field essentially results in an HOI prior which can be used to synthesize plausible grasps from an object mesh, using nothing but a neural

network. The concept of HOI prior is taken further with different neural network architectures that incorporate image representations. This allows learning SDFs for the hand and the object, which is applied to image-based hand-object reconstruction. Hand and object meshes are then recovered via traditional methods, such as marching cubes [227]. Grasping Field [28] thus demonstrates a simple yet versatile technique to learn an implicit HOI prior that allows for reconstruction of explicit mesh representations with SOTA plausibility metrics. However, while it learns plausible grasps, the method heavily relies on observation embeddings based on raw inputs. During training, the model embeds the hand-object pair point clouds with PointNet [59] or the image with ResNet [218]. This approach remains close to the paradigm of end-to-end methods, where no effort is placed on using richer features to constrain the learning. In effect, the model's performance is constrained by the learning of these embeddings, where the embedding network (*i.e.* PointNet [59]) is the bottleneck. While Grasping Field [28] is in itself an implicit representation, the main model acts as a decoder for the representation learned by PointNet, hence it can be viewed as an end-to-end method and may suffer from the weaknesses associated with them. As no ablation study is provided, it is difficult to gauge how much information is captured by the PointNet encoder and how much prior knowledge is stored in the variational encoder-decoder. This limitation can be removed by using richer features or an expressive explicit HOI representation, such that the embedding network may be omitted or unified with the encoder-decoder.

**Contact Potential Field (Yang et al. 2020)**

Contact Potential Field (CPF) [37] propose a mass-spring system for image-based grasp reconstruction, to reconstruct a hand mesh given an HOI image and an object mesh. The method defines a passive repulsive field and an active attractive field, both from the object's surface. The attractive field is activated at object vertices where the hand should be in contact, such that the attraction of each object-hand vertex pair is parameterized by a scalar. This translates to an effective energy minimization objective for gradient-descent-based TTO. CPF thus follows our guidelines proposed in Chapter 3, in particular *specialized inference* with the TTO stage, *explicit HOI formulation* and *canonical frame of reference* with the field formulation. [37] introduce the idea of coarse hand surface regions, defined as anchor points on the MANO [120] parametric hand mesh. These anchors simplify the field by

reducing each hand-object vertex pair to a pair of object vertex and MANO [120] anchor, using the closest anchor per hand vertex. To learn the field, the authors propose a network that predicts the attractive energy in multiple steps. It first predicts the contact probabilities of the object's vertices as a regression task. The network then regresses the affinity for all anchors, for each object vertex. Finally, the model predicts the anchor elasticity for each object vertex. This forms a cascade of information that culminates towards the prediction of the elasticity, or attraction scalar. CPF [37] achieved SOTA results in grasp reconstruction from images, but is not devoid of limitations. Primarily, it uses heuristics to compute the ground truth attraction scalar (or elasticity), and others to limit the effect of the repulsive field. The final result is obtained after 400 TTO iterations, which may be considered slow. Overall, CPF [37] proposes an elegant solution, but the extensive engineering and use of heuristics result in a learning method that is intricate, and in several hyperparameters that may require returning for new benchmarks. Furthermore, the neural network is designed to work with images and coarse mesh predictions for informed prediction. While this allows for image-based grasp reconstruction, [37] does not bring evidence that CPF can serve as a powerful grasp prior for other tasks, such as grasp denoising or synthesis. Finally, CPF [37] does not unify the representation of the hand-object interaction and the geometry of the object; the latter is encoded using PointNet [59]. This may hinder generalization to other object instances from the test set, however this is not evaluated by the authors. A solution to this limitation would be to design a representation that encodes the object geometry, preferably in a common space with the hand shape and pose.

#### TOCH (Zhou et al. 2022)

The idea of defining grasps as an object-based representation was taken further by TOCH [27]. The authors propose to represent grasps in HOI as an object-centric field, similarly to CPF [37]. TOCH [27] differs from the latter by using a simpler form for the grasp representation, based on signed distances (akin to Grasping Field [28]). Furthermore, TOCH not only encodes contacts but the inner hand surface from the object's perspective, enabling the modelling of approaching hands in a manipulation sequence. This is accomplished by casting rays from points sampled on the object's surface and along their normal vectors. The TOCH field formulation defines three components. The first is a binary interaction label for each object surface point, which indicates whether a corresponding hand point

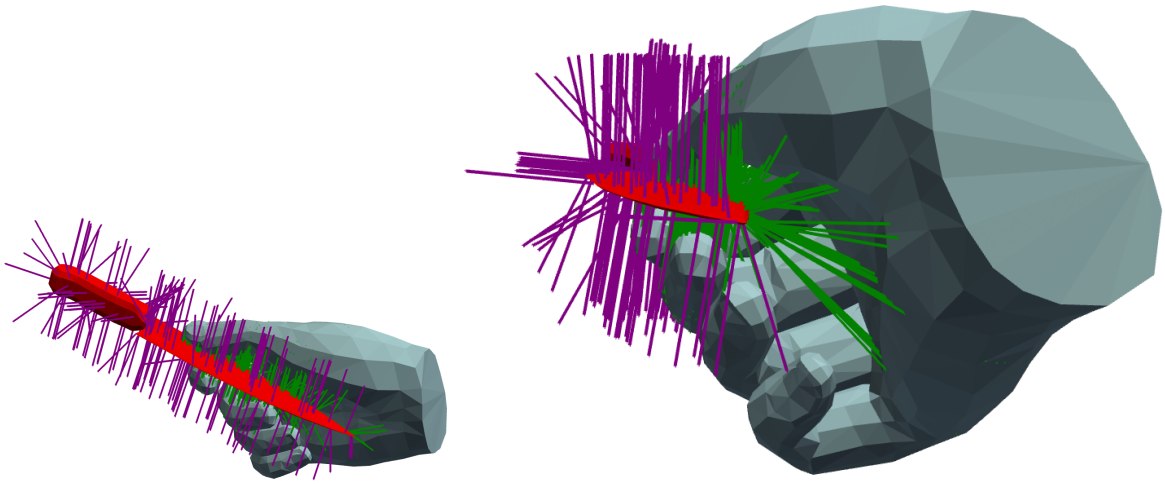


Figure 4.2: Ray-based methods like TOCH [27] and CPF [37] do not model the entire hand shape and pose. In this example, the green rays are hitting the hand surface while the purple rays are not. On the right, it is clearly visible that some fingers are much less represented than others. This shows how some object rays may never hit certain phalanges or occluded hand parts, leading to uncertainty about the hand pose: this is a flaw in entirely object-centric methods.

exists along its ray. The second is a signed distance to this corresponding hand point, and the third is the coordinates of such point on the MANO template mesh (in canonical pose). In a way, TOCH [27] is a simplification of CPF [37] which uses ray casting instead of contact heuristics. Consequently, it also follows our guidelines proposed in Chapter 3. While TOCH is a more accurate and generalizable formulation of grasping, it comes at a heavier computational cost. Another limitation of TOCH, and thus of CPF, is that it does not model the entire hand shape and pose, but only the hand portions directly facing the object's surface. This may be limited to the palm in some cases, or to small portions of the hand in certain grasps (see Section 4.1.1). For their applicability, TOCH fields are demonstrated on the problem of manipulation refinement, or grasp denoising, using a temporal denoising autoencoder. The authors also demonstrate the transferability of grasps to other object instances, thanks to the object-centric representation of TOCH. Nevertheless, TOCH fields remain slow to compute and are not fully differentiable nor continuous, which are desirable properties for curve fitting with neural networks. As a result, TOCH relies on random restarts to find a good solution during TTO, which would not be necessary with an SGD optimizer. Furthermore, the authors do not demonstrate the method's ability to serve as an interaction prior that can be used for other tasks,

such as grasp synthesis.

**Strengths and limitations** These three methods share common strengths when compared to traditional end-to-end methods. In particular, they all propose novel ways to express grasp affordances and hand-object correspondences that are richer than the previous paradigm. This novel type of hand-object features, in conjunction with test-time optimization, apply our guidelines from Chapter 3 with success. Additionally, Grasping Field [28] and TOCH [27] show promise in more general-purpose representations, while CPF [37] is in theory usable as a general prior but remains to be evaluated as such.

These approaches to expressive HOI representations are all limited in their own aspect. While they each try to improve upon previous approaches, they eventually introduce new limitations or disregard others. While Grasping Field [28] and CPF [37] propose continuous and fully-differentiable HOI representations, they are tied to the learning method. On the other hand, TOCH [27] is a prime example of a learning method-agnostic representation that is applicable to more than one task. In addition, TOCH defines a richer representation for hand grasps that are object-centric, allowing for a more expressive model with SOTA results in grasp refinement. However, TOCH is not continuous nor fully-differentiable, and is compute intensive. Those properties, while not required for a deep learning model, are desirable as they make optimization easier and improve convergence. There is extensive research on the continuity of functions to be fit by neural networks and the benefits of employing continuous representations as inputs [228]. In fact, the universal approximation theorem states that any continuous function that is continuously differentiable can be approximated with an arbitrary precision by a neural network with an upper bound in depth [229, 230], or in width if using ReLU activations [231]. Thus, an ideal representation would be continuous and fully differentiable, and would not be tied to a particular learning method.

Another major limitations of the presented methods is how they encode the hand-object pair. Indeed, they either do not encode the object geometry, or are relying on PointNet [59] to do so, which inherits from the limitations of end-to-end methods. By encoding both the hand and object geometries in a shared space, we can expect the model to learn a richer mapping from object to hand grasp due to the common semantics of the input features. By contrast, embedding the object

in a latent vector to condition the embedding of the hand grasp (as shown in Grasping Field [28]) introduces new challenges. We hypothesize that the object embedding may not be informative enough to condition the hand grasp due to the information bottleneck [232].

### 4.1.2 A tightly coupled, continuous and fully-differentiable representation in canonical frame

To design a representation that is easily learnable by neural networks, and that follows our previous guidelines, it should take the best of previous methods. Hence, it should be continuous and fully differentiable. Furthermore, it should encode the hand and the object in a shared canonical frame in a manner as to consistently encode the geometry of the hand-object pair. This can be achieved with a frame of reference that is independent of the object geometry and hand grasp. Moreover, we believe that HOI can be better modelled in a shared semantic space. Unsigned Distance Fields (UDFs) [69] are an efficient alternative to signed distances as used by TOCH. The Basis Point Set (BPS) representation [64], a discrete case of UDF, is a good candidate for a better representation. It is a method for representing point clouds with a fixed set of basis points, which yields higher mesh reconstruction accuracy with a simple MLP than when using point clouds with PointNet [64, 59]. The use of a fixed basis is desirable for our *canonical frame of reference* guideline. BPS encodes point clouds or meshes with unsigned distances, which were shown to yield more stable gradients during optimization than signed distances [69]. Since the BPS representation doesn't require sampling, unlike Unsigned Distance Fields (UDFs) [69], nor surface normals for sign computation, it encodes meshes in around 1ms for 1024 basis points on a modern GPU. A lightweight representation also allows us to perform test-time optimization closer to real-time, a desirable property for our *specialized inference* guideline.

In the next section, we introduce our tightly coupled HOI representation defined in canonical space that is continuous and fully differentiable, based on unsigned distances.

## 4.2 Designing and learning CHOIR

In this section we introduce CHOIR, a *Coarse Hand-Object Interaction Representation* that is computable for a whole sequence of several hundred frames in less than 2ms on a modern GPU. To demonstrate its effectiveness, we apply it to the problem of noisy grasp refinement as in TOCH [27] and ContactOpt [1]. We assume that predictions from an off-the-shelf HOI reconstruction method are noisy, and we wish to refine them in terms of pose accuracy. By demonstrating that this task can be achieved with a conditional generative model, it would support the claim that CHOIR is a general representation for HOI modelling. Additionally, to static grasp denoising, we demonstrate the modelling capabilities of CHOIR for dynamic grasps with a unified model. To do so, we hypothesize that there is redundant information across several frames of the hand-object interaction. If we consider the static grasping scenario, we can hypothesize that multiple predictions of the same input are highly correlated, but corrupted by i.i.d. noise due to the inevitable inaccuracy of the model. For dynamic interaction sequences, neighbouring frames contain redundancy since movement is a fluid sequence of poses in time. Furthermore, single-frame-based methods produce predictions that are corrupted with additive noise, which we assume to be i.i.d. across frames as a relaxation. Thus, the true pose parameters are easier to disambiguate when we can validate the prediction across more than one frame. To this end, we propose a denoising variational encoder-decoder neural network (AggVED) with a latent aggregation mechanism for an arbitrary amount of noisy choir observations to maximize the signal-to-noise ratio. Prior to introducing the architecture of AggVED, we introduce the formulation of CHOIR.

### 4.2.1 Explicit shape-pose representation as unsigned distances

**Notation** In the following, $x$ on its own denotes a generic data sample of any form.

In CHOIR, the object geometry and relative hand pose are represented with unsigned distances from a common set of points, following the Basis Point Set (BPS) representation [64]. The latter is a lightweight 3D point cloud representation with fixed dimensionality that enables the use of lightweight neural network architectures like the MLP, or the use of convolutions with a regular point grid. Define

a dataset $\mathcal{D} = \{X^1, \ldots, X^N\}$ consisting of $N$ point clouds where each point cloud $X^n$ is composed of points $\{\boldsymbol{x}_1^n, \ldots, \boldsymbol{x}_{K_n}^n\}$. In the original formulation, a 3D basis point set $\mathcal{B} = \{\boldsymbol{b}^1, \ldots, \boldsymbol{b}^M\}$ is sampled uniformly from a unit ball in $\mathbb{R}^3$ [64]. Alternatively, the basis points can be defined as a regular grid to enable the use of convolutions. However, experiments have shown that random sampling in the unit ball results in a BPS representation which yields slightly better performance in classification tasks [64]. Since we do not use convolutions in our method, we opt for the random sampling variant. Then, the dataset is normalized such that $X^n$ fits in the grid and its centroid is at the origin. Finally, the BPS representation of each point cloud $X^n$ is computed as

$$X_{\text{BPS}}^n = \begin{bmatrix} \min_k \|\boldsymbol{b}^1 - \boldsymbol{x}_k^n\|_2^2 \\ \min_k \|\boldsymbol{b}^2 - \boldsymbol{x}_k^n\|_2^2 \\ \cdots \\ \min_k \|\boldsymbol{b}^M - \boldsymbol{x}_k^n\|_2^2 \end{bmatrix}. \tag{4.1}$$

Following our motivations for a better representation, we choose to encode the hand-object pair in a common semantic space with a canonical frame. For the latter, we use the BPS representation which can be viewed as a high-dimensional strictly positive coordinate system. As for the common semantic space, we encode the object geometry is encoded with the traditional BPS representation, while the hand mesh uses fixed anchors instead of the nearest surface point. This is doable since the hand mesh is based on the parametric MANO template [120], and it simplifies the objective since the anchors are fixed. As such, CHOIR represents the hand-object interaction as a coarse hand pose in a canonical object frame using the MANO parametric hand mesh [120]. This coordinate system is obtained by shifting the hand-object pair at the object's centroid and applying the inverse of the object's affine transformation.

To encode the hand shape and pose with the object, CHOIR is defined as a concatenation of the BPS representation of the object mesh and of the distances from the BPS to the 32 pre-assigned MANO anchors proposed by [37]. In effect, a CHOIR specifies an object point cloud $X^n$ together with a hand mesh $H$. The anchor distances $\boldsymbol{d}_H = [d_H^1, \ldots, d_H^M]^T$ are given by

$$d_H^j = \|\boldsymbol{b}^j - \delta_H(j)\|_2^2 \tag{4.2}$$

Table 4.1: Average reconstruction error for MANO meshes fitted onto ground-truth CHOIRs with the *ordered* and *random* anchor assignment schemes. Mean Per-Joint Pose Error (MPJPE) and Mean Per-Vertex Pose Error (MPVPE) are averaged across the entire ContactPose [23] dataset.

| | *Ordered* | *Random* |
|---|---|---|
| MPJPE (mm) | 0.18 | 0.19 |
| MPVPE (mm) | 0.22 | 0.22 |

where the function $\delta_H(j)$ returns the anchor for point $\boldsymbol{b}^j$ and hand mesh $H$. Note that the same MANO anchor can be assigned to multiple basis points. We propose two assignment schemes $\delta_H$: (1) a repeating pattern of the 32 ordered indices and (2) a shuffled version of the latter. Table 4.1 shows that both the (1) *ordered* and (2) *random* anchor assignment schemes produce the same reconstruction error when fitting a ground-truth CHOIR from the ContactPose [23] dataset. The Mean Per-Joint Pose Error (MPJPE) and Mean Per-Vertex Pose Error (MPVPE) metrics were averaged across the entire dataset. Note that with ground-truth hand-object meshes, the obtained CHOIR allows fitting a MANO mesh with less than 1mm error. As we do not find any difference in accuracy between the two, we use the *ordered* assignment for all experiments.

To sum up, a CHOIR (see Fig. 4.1) is defined as

$$x_{\text{CHOIR}} = [X^n_{\text{BPS}}, \boldsymbol{d}_H] \in \mathbb{R}^{M\times 2}. \tag{4.3}$$

As opposed to TOCH [27] which uses an object-centric approach for MANO mesh coordinates at the intersection of object-cast rays, we encode coarse hand-object correspondences via the unsigned distances to the object surface and anchor points on the hand. As a result, our representation is continuous and fully differentiable, which facilitates the learning of a neural network. Nevertheless, our experiments show that we are able to fit hand meshes onto ground-truth CHOIRs with less than 1mm absolute mean per-joint pose error. Thus, denoising CHOIRs is the accuracy bottleneck; in the next subsection, we present our denoising variational encoder-decoder enhanced with aggregation capability.

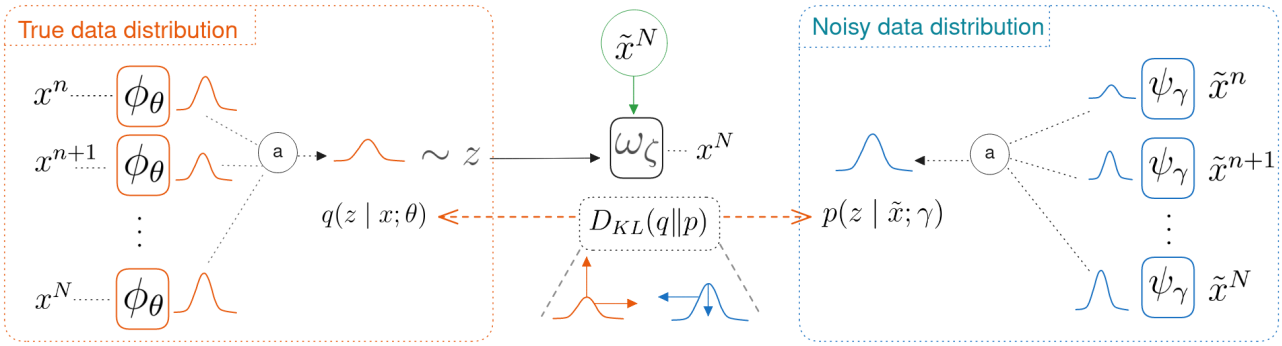



Figure 4.3: Architecture of AggVED: each observation frame in the context window is passed to $\phi_\theta$ and $\psi_\gamma$ as CHOIRs $x$ and noisy CHOIRs $\tilde{x}$ respectively. Individual distributions are predicted per frame, then aggregated into one from which the latent variable $z$ is sampled. The noisy encoder $\psi_\gamma$ is solely trained through the Kullback-Leibler divergence ($D_{KL}$) between its predicted distribution and $\phi_\theta$'s which approximates the true data distribution. This makes $\phi_\theta$ the teacher and $\psi_\gamma$ the student. The green path is for the dynamic model: it helps to condition the decoder on the target frame since the latent is now less informative due to the variations in all frames.

### 4.2.2 Learning a CHOIR distribution for grasp denoising

**Notation** In the following, we assume $x$ to be a CHOIR, and we use this generic notation for simplicity.

We propose to use a Variational Encoder-Decoder architecture for denoising noisy CHOIR fields. We hypothesize that there is more useful information in a set of $N$ noisy observations than in a single one for the purpose of predicting the underlying grasp. To this end, we extend a student-teacher variational encoder-decoder with an aggregation mechanism in the predicted normal distribution parameters. The architecture, named AggVED, is inspired by the student-teacher model of HuMoR [25] and the Latent Neural Process (LNP) [42]. Our proposed method is flexible enough to be applied to static grasps, where we have multiple noisy observations for the same static hand pose, and dynamic grasps, where we have multiple frames of a motion and the hand pose varies slightly from frame to frame. Despite working on more frames, our method is competitive with the SOTA methods with just a single observation.

Given a set of noisy CHOIRs for the same static grasp – or of consecutive frames in a hand-object interaction sequence – AggVED is trained to predict a denoised CHOIR for the underlying grasp or the

last frame of the motion. It is composed of two encoders and a decoder as shown in Fig. 4.3. The first encoder $\phi_\theta(x)$ learns the approximate posterior distribution $q(z|x;\theta)$ of the latent variable $z$ given the ground-truth CHOIR $x$. The second encoder $\psi_\gamma(\tilde{x})$ learns the conditional prior distribution $p(z|\tilde{x};\gamma)$ of $z$ given the noisy CHOIR $\tilde{x}$. Finally, the decoder $\omega_\zeta(z)$ learns the conditional data distribution $p(x|z;\zeta)$ of the true CHOIR $x$ given the latent variable $z$. As in the typical variational autoencoder framework [39], integrating over the true data distribution $p(x|z)$ is intractable:

$$\int p(x|z)p(z)dz. \tag{4.4}$$

Therefore we train the decoder jointly with the first encoder $\phi_\theta(x)$ through maximization of the ELBO objective (see Section 2.4.2) defined as

$$L(\zeta,\theta,\gamma;x) = \ln p(x;\zeta) - D_{KL}(q(z|x;\theta)||p(z|\tilde{x};\gamma)). \tag{4.5}$$

Note that unlike in the variational autoencoder (VAE), we do not use a normal Gaussian $\mathcal{N}(0,I)$ as the prior distribution $p(z)$. Instead, we learn a conditional prior $p(z|\tilde{x};\gamma)$ through the encoder $\psi_\gamma(\tilde{x})$ as in [25, 42]. The Kullback-Leibler divergence ($D_{KL}$) term of Eq. (4.5) is the only error signal for $\psi_\gamma(\tilde{x})$, thus making this architecture a student-teacher model. To achieve convergence, one must carefully balance the reconstruction term of Eq. (4.5) and the KL-divergence term to allow learning the true data distribution before learning the mapping from the noisy data distribution to the latent space. To this end, we define our final training objective as

$$\mathcal{L} = \ln p(x;\zeta) - w \cdot D_{KL}(q(z|x;\theta)||p(z|\tilde{x};\gamma)) \tag{4.6}$$

where we grow $w$ exponentially during training. For our static grasp experiments, we use MLPs with 5 layers of width 1024 for the encoders and 3 layers of width 1024 for the decoder, with a latent dimension of 128. For the dynamic sequences, we use 3 layers of width 2048 for the decoder because it is conditioned on the input CHOIR (see Fig. 4.3. For all experiments, we make use of skip-connections for the decoder as proposed by Dieng et al. [233] and of residual connections for the encoders.

Let $\mathcal{G} = \{y \mid y \coloneqq x + \epsilon_x \text{ for } x \in \mathcal{X}, \epsilon_x \sim \mathcal{N}(0,\sigma^2)\}$ be a set of noisy CHOIRs where $\mathcal{X} = \{x,\ldots,x\}$ for

a static grasp or $\mathcal{X} = \{x^1, x^2, \ldots, x^N\}$ for a dynamic grasp sequence of $N$ observations. We hypothesize that aggregating the parameters $\eta = \phi_Y(\tilde{x})$ of the latent distributions for all observations helps to disambiguate the true grasp of the noise for two reasons. First, the same latent $z$ explains all elements of $\mathcal{X}$, and hence $\mathcal{G}$, for static grasps. This is also true for dynamic sequences because all grasps in a sequence are similar for an appropriate choice of $N$ (we explore this in Section 4.3). Second, the noise is additive and i.i.d. Since we assume observations to be order-independent the aggregation operator needs to be permutation invariant. Based on the Neural Processes [42, 207], we explore two operators with that property, *mean* for the static grasp setting and *cross-attention* for the dynamic and temporal setting. We define the aggregated parameters $\bar{\eta}$ through cross-attention as

$$\begin{aligned} \bar{\eta} &= \text{MultiHead}(Q, K, V), \text{ where} \\ Q &= [\tilde{x}^N], \\ K &= [\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^N], \\ V &= [\eta_1, \eta_2, \ldots, \eta_N]. \end{aligned} \tag{4.7}$$

In practice, to reduce the model's footprint, $\tilde{x}$ is set to the anchor distances instead of the full CHOIR. This is valid since the object representation is known and only useful for the encoders. We refer the reader to the original material [234] for a formulation of multi-head attention. Note that the cross-attention aggregator necessitates a query, hence it is suited to the dynamic grasp setting where the last frame $x^N$ of the context window may be used as the query. As opposed to mean aggregation, this operator applies weights to the latent distributions such that the information extracted is most relevant to the last frame in a sequence. In further references, we name the mean aggregation variant $\mu AggVED$ and the cross-attention one $\times AggVED$.

Ultimately, we propose to maximize the information content of the latent variable by aggregating the predicted distributions from several observed noisy CHOIRs. This latent variable is then decoded into the denoised grasp for the last frame of the context window or the underlying static grasp. In the next subsection we show how this denoised CHOIR is used to obtain MANO parameters in the object coordinate system.

Algorithm 1: Minimal Python code for the TTO loss.

```python
anchor_dist = torch.cdist(bps, anchors) # Anchors predicted in TTO
distances = torch.gather(anchor_dist, 2, anchor_ids)
choir_loss = F.mse_loss(
    distances,
    choir[..., -1]
) # Agreement of anchors and CHOIR
```

### 4.2.3 From a representation to a hand

After predicting denoised CHOIRs from observations given by an off-the-shelf hand-object pose predictor, obtaining refined hand meshes and poses is done via TTO. While the state-of-the-art methods are either not fully differentiable, require several optimization stages, or rely on random restarts [27, 1], fitting a MANO mesh and pose parameters to CHOIR is straightforward and fast. Due to its distance-only based structure, the fitting loss is obtained in very few lines of Python code as seen in Algorithm 1.

Our TTO objective is composed of the reconstruction loss $\mathcal{L}_{\text{rec}}$ described in Algorithm 1, a shape regularizer $\mathcal{L}_{\text{shape}}$ and a pose regularizer $\mathcal{L}_{\text{pose}}$ such that the final objective is

$$\mathcal{L}_{\text{TTO}} = \lambda_1 \cdot \mathcal{L}_{\text{rec}} + \lambda_2 \cdot \mathcal{L}_{\text{shape}} + \lambda_3 \cdot \mathcal{L}_{\text{pose}}. \tag{4.8}$$

The shape regularizer prevents the hand mesh from over-deforming to satisfy the reconstruction loss. It is defined as

$$\mathcal{L}_{\text{shape}} = \|\boldsymbol{\beta}\|_2^2 \tag{4.9}$$

with $\boldsymbol{\beta}$ being the MANO shape parameters. As for the pose regularizer, it prevents strong deviation from the initial MANO pose estimate $\boldsymbol{\theta}$ and is defined as

$$\mathcal{L}_{\text{pose}} = \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{init}}\|_2^2. \tag{4.10}$$

With this objective, we optimize the MANO pose and shape parameters and rotation and translation of the wrist joint with the Adam optimizer [235] with a learning rate of $8 \times 10^{-2}$ for around 150 iterations. We set $\lambda_1$ to 1000 to bring the loss into the millimetre scale and found $\lambda_2 = 1 \times 10^{-4}$ and

$\lambda_3 = 1 \times 10^{-8}$ to work well in practice. All parameters are initialized with the estimates from which CHOIR observations are built, which leads to fast convergence on the denoised CHOIRs.

In the next section, we evaluate the combined CHOIR + AggVED + TTO solution on several benchmarks.

## 4.3 Evaluation

Our solution is constituted of (a) a representation (CHOIR), (b) a learning method (AggVED), and (c) a hand-mesh fitting algorithm (TTO). We evaluate this solution in two settings, using several key metrics as detailed in Section 2.8.

**Static grasps** We replicate the benchmark proposed by Grady et al. [1] which consists of a perturbed version of the ContactPose dataset [23]. ContactPose comprises highly accurate hand poses for static grasps of 25 objects by 50 participants. Grady et al. [1] define large perturbations on the hand poses as 3 additive and i.i.d. noise components: (1) translation noise $\epsilon_t \sim \mathcal{N}(0, 5)$ in cm, (2) pose noise $\epsilon_\theta \sim \mathcal{N}(0, 0.05)$ in PCA space, and (3) rotation noise $\epsilon_R \sim \mathcal{N}(0, 15)$ in radians. However, they propose to split the dataset in training and testing splits without a validation split. We instead split the dataset with 70% data for the training split, 10% for the validation split and the last 20% for the test split. The training split contains 16 perturbed version of each sample and 4 for the validation and test splits. We retrain ContactOpt on these new splits, and the recent SOTA method TOCH [27], specialized in denoising dynamic grasps (on the GRAB [9] benchmark), which has not been evaluated on static grasps with this amount of noise. In addition, we include the pose error of S2Contact [161] as reported by the authors. Furthermore, we propose object-based splits to explore the generalizability of our method compared to ContactOpt, with 2 objects held out for the validation split and 5 for the test split. We retrain ContactOpt and TOCH on these splits as well and compare it with our method.

**Dynamic grasps** Zhou et al. [27] propose to perturb the dynamic grasp sequences of the GRAB dataset [9] with additive and i.i.d. Gaussian noise. GRAB consists of motion sequences of 51 objects being manipulated by 10 participants. The large number of objects in the dataset allows it to be split into training, validation, and testing by objects. We replicate their experiment with translation noise

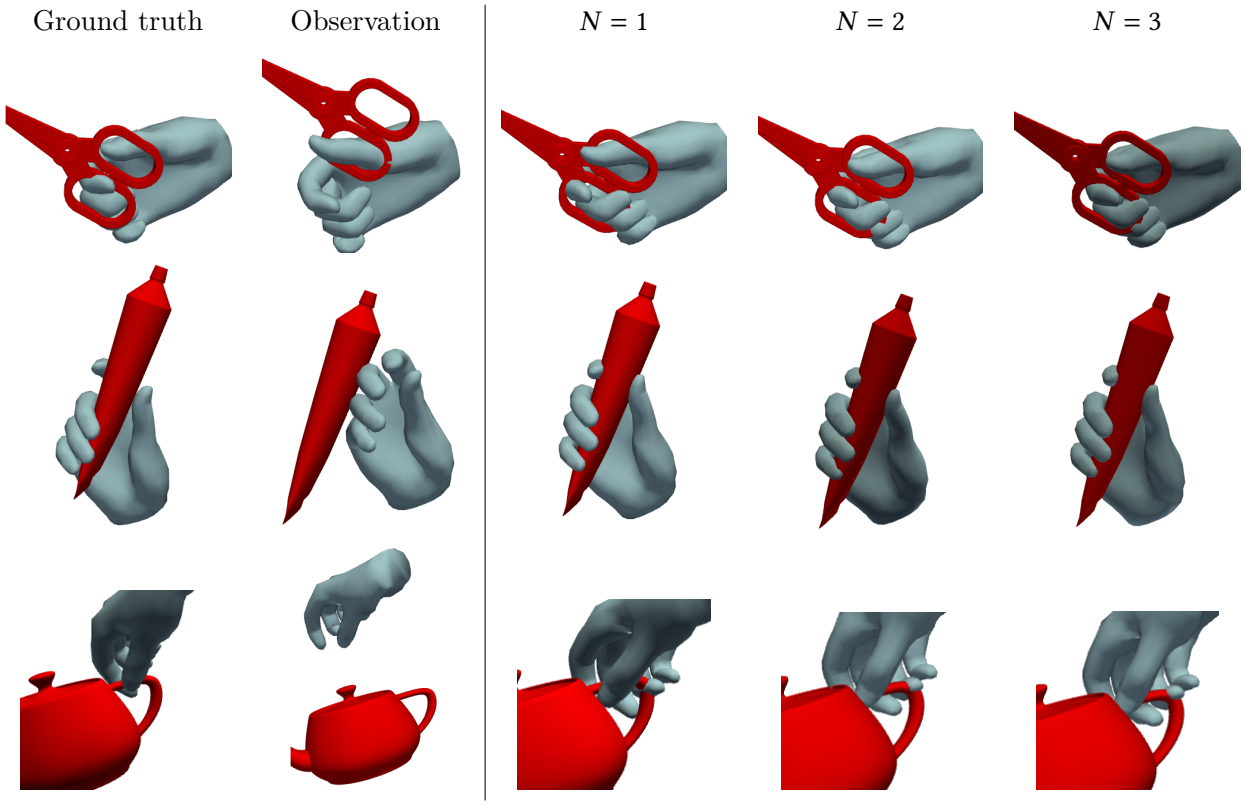



Figure 4.4: Qualtiative results on ContactPose for the grasp splits. We show a sample observation from one of the 4 perturbed versions of the grasp, and the hand meshes fitted on the predicted CHOIRs for $N = 1$, $N = 2$ and $N = 3$ observations. With 3 observations, the fitted hand meshes are closer to the ground truth even for perturbed samples with large translation noise.

$\epsilon_t \sim \mathcal{N}(0, 0.01)$ (in cm) and hand pose noise $\epsilon_\theta \sim \mathcal{N}(0, 0.3)$ (in PCA space) and report their figures for comparison.

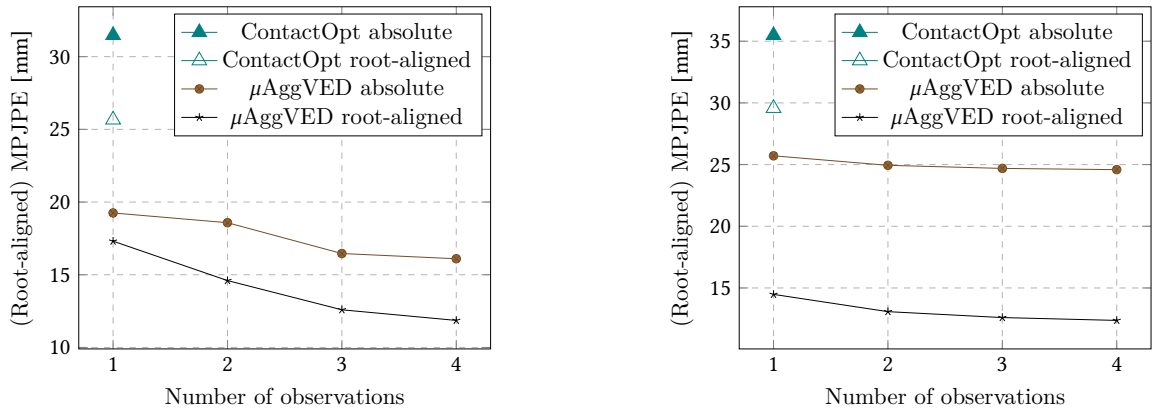
### 4.3.1 Quantitative & qualitative results

#### Refining static grasps

To evaluate static grasp denoising on the Perturbed ContactPose benchmark, we retrain all methods on grasp splits and object splits. The grasp splits are used to evaluate the generalization to novel grasps on objects seen during training, while the object splits evaluate the generalization to novel objects (absent from the training split).

| **Method** | **IV** ($cm^3$) ↓ | **MPJPE** (mm) ↓ | **MPVPE** (mm) ↓ | **Contact coverage** (%) ↑ |
|---|---|---|---|---|
| Perturbed Data | 8.46* | 81.76 | 81.47 | 2.36 |
| ContactOpt [1] | 12.83* | 31.47 | 31.62 | **19.70*** |
| S2Contact [161] | – | 33.65 (29.44*) | 30.64* | – |
| TOCH [27] | 10.14 | 26.96 | 27.26 | – |
| $\mu$AggVED (*ours*) | **5.17** | **19.25** | **19.51** | 2.41 |

Table 4.2: Evaluation of our approach with $N = 1$ observation against ContactOpt [1] on the Perturbed ContactPose benchmark, with grasp splits. * means reported figures, and – means missing figures. Our method improves over the SOTA methods on most metrics except for the contact coverage, where ContactOpt [1] directly predicts contacts on a given surface and thus maximizes contact coverage during TTO. The first row compares the perturbed samples to the ground-truth. Our method offers a higher fidelity to the ground-truth pose, thus better at modelling the noise distribution.

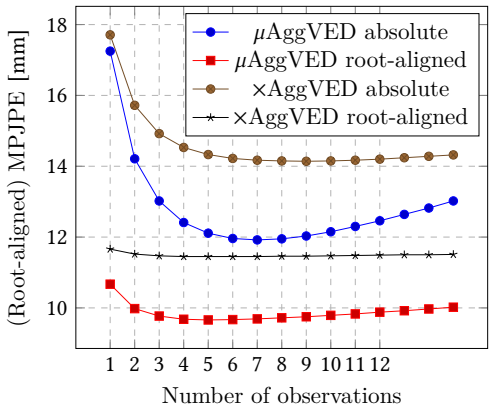



(a) Grasps splits.

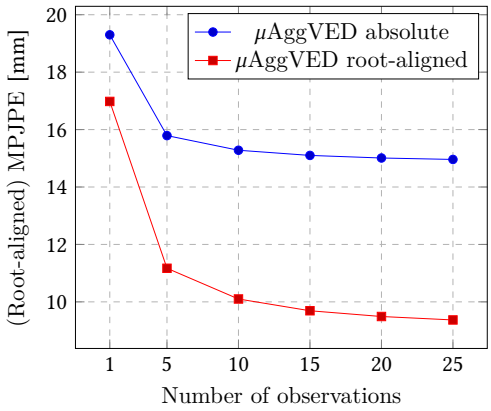
(b) Object splits.

Figure 4.5: Evaluation of our approach with $N$ observations ranging from 1 to 4 against ContactOpt [1] on the Perturbed ContactPose benchmark. $N = 1$ gives a fair comparison. Generalizing to unknown objects remains challenging.

**Grasp splits** For a fair comparison with other methods, we first evaluate our solution with only 1 observation. As shown on Table 4.2, our method improves over the SOTA method TOCH [27] by 8mm of kinematic error (MPJPE) in absolute space and of vertex error for hand mesh. We also demonstrate a 2× lower intersection volume (IV) than TOCH, which indicates that our method produces less hand-object interpenetration and more plausible grasps. While ContactOpt [1] yields a much higher contact coverage, this comes at the detriment of ground-truth fidelity and is caused by their predictions being restricted to the hand and object surfaces. Our method results in less contact coverage but more realistic grasps. We then evaluate our method with more than 1 observations against ContactOpt




(a) **Dynamic model.** μAggVED worsens past 7 observations since the latent dimension is restricted and more observations induce more grasp variations. ×AggVED mitigates this due to more expressiveness through cross-attention.




(b) **Static model.** The accuracy does not get worse as observations are added since they represent the same underlying grasp. A plateau is reached at around 15 observations.

Figure 4.6: Evaluation of the plateau point for the dynamic and static models. While the dynamic model worsens past 7 observations, the static model does not suffer as much from this effect and the error reduces up to 25 observations.

[1] in Section 4.3.1, where we can see a steady decrease in the MPJPE. This shows the advantage of our aggregation mechanism over methods that utilize only 1 observation to refine a static grasp. With $N = 4$ observations, our method yields a 15mm lower error than ContactOpt [1]. In Fig. 4.6, we show that our method gets better in predictive accuracy as we add observations, and seems to reach a plateau around 15 observations. This is because all observations are of the same static grasp and thus there is no information loss for the aggregate. We show qualitative results with 1 to 3 observations in Fig. 4.4, and a qualitative comparison with ContactOpt in Fig. 4.7. The latter, while showing how well our method recovers the original pose, exposes a limitation of CHOIR: the lack of contact representation and penetration regularization.

**Object splits** We re-iterate the previous experiment on the per-object splits, where both methods now suffer from the distribution shift. While our method maintains a 10mm gap as seen on Fig. 4.5b, it is not able to effectively reduce the error with more observations. Generalizing to unknown objects is an active field of research where new representation learning methods are needed [236].

| **Method** | **IV** ($cm^3$) ↓ | **MPJPE** (mm) ↓ | **MPVPE** (mm) ↓ |
|---|---|---|---|
| Perturbed Data | 2.20* | 19.1 | N/A |
| TOCH [27] | 1.78* | **10.3*** | 12.1* |
| $\mu$AggVED | 2.65 | 11.92 | **11.85** |
| ×AggVED | 3.08 | 14.14 | 14.04 |

Table 4.3: Evaluation of solution CHOIR+$\mu$AggVED+TTO with $N = 7$ observations against the ground-truth and TOCH [27] on the perturbed GRAB benchmark. We chose $N = 7$ observations for our method as per the results shown in Fig. 4.6a.* means reported figures.

**Refining dynamic grasps**

To showcase the flexibility of our solution, we evaluate it on a benchmark of perturbed dynamic grasp sequences from the GRAB dataset. We compare our method with the SOTA by using a sliding window of $N = 7$ frames as the observations and predicting the last frame's CHOIR. This way we simulate a temporal model without changing our architecture. We train the model with $N$ ranging from 2 to 15 since less than 2 observations may hinder model convergence because of ambiguity in dynamic grasps. Table 4.3 shows that our method is competitive with the SOTA in hand pose error, achieving only 2mm worse than TOCH, and better in terms of hand mesh fidelity as shown by a lower MPVPE. The choice of $N = 7$ is motivated by the curve of Fig. 4.6 which shows that $\mu$AggVED's predictions start to deteriorate past this number. On the other hand, Fig. 4.6 shows that ×AggVED alleviates this issue by using attention weights for the aggregate, hence minimizing information loss occurred by variations in large windows. While our temporal model is not as accurate as the SOTA, it is flexible in the compute budget and more suited to real-time inference with CHOIR being 1000× faster to compute for a GRAB sequence.

## 4.4 Discussions and limitations

We evaluated CHOIR with our learning method on static and dynamic grasp denoising using two benchmarks. Our solution clearly demonstrates that if more noisy observations of the same grasp are available, the denoising accuracy is improved. This approach seems to reach anplateau at around 15 observations for static grasps, and at 7 observations for dynamic manipulation. This hints at the lack of expressivity of our temporal aggregation scheme. Furthermore, while our method is more accurate

than the SOTA method for static grasp refinement, it displays less physical plausibility and realism in the generated grasps. This is especially seen in Fig. 4.7 where the pose is accurately recovered but the contacts are not. Indeed, CHOIR does not model contacts or dense hand-object correspondences, which limits its expressivity for realistic grasp modelling. In addition, this chapter only explores noisy grasps refinement for the application of CHOIR. This does not show sufficient proof that CHOIR is a suitable representation for better HOI priors, as applicability to other HOI tasks is a core criterion for a general grasping prior. In the following chapter, we explore the issues of contact modelling and applicability for CHOIR. However, improvements on dynamic manipulation modelling are left out of the scope of this thesis, as they can be tackled entirely by the learning method, as TOCH [27] demonstrates better than ×AggVED.

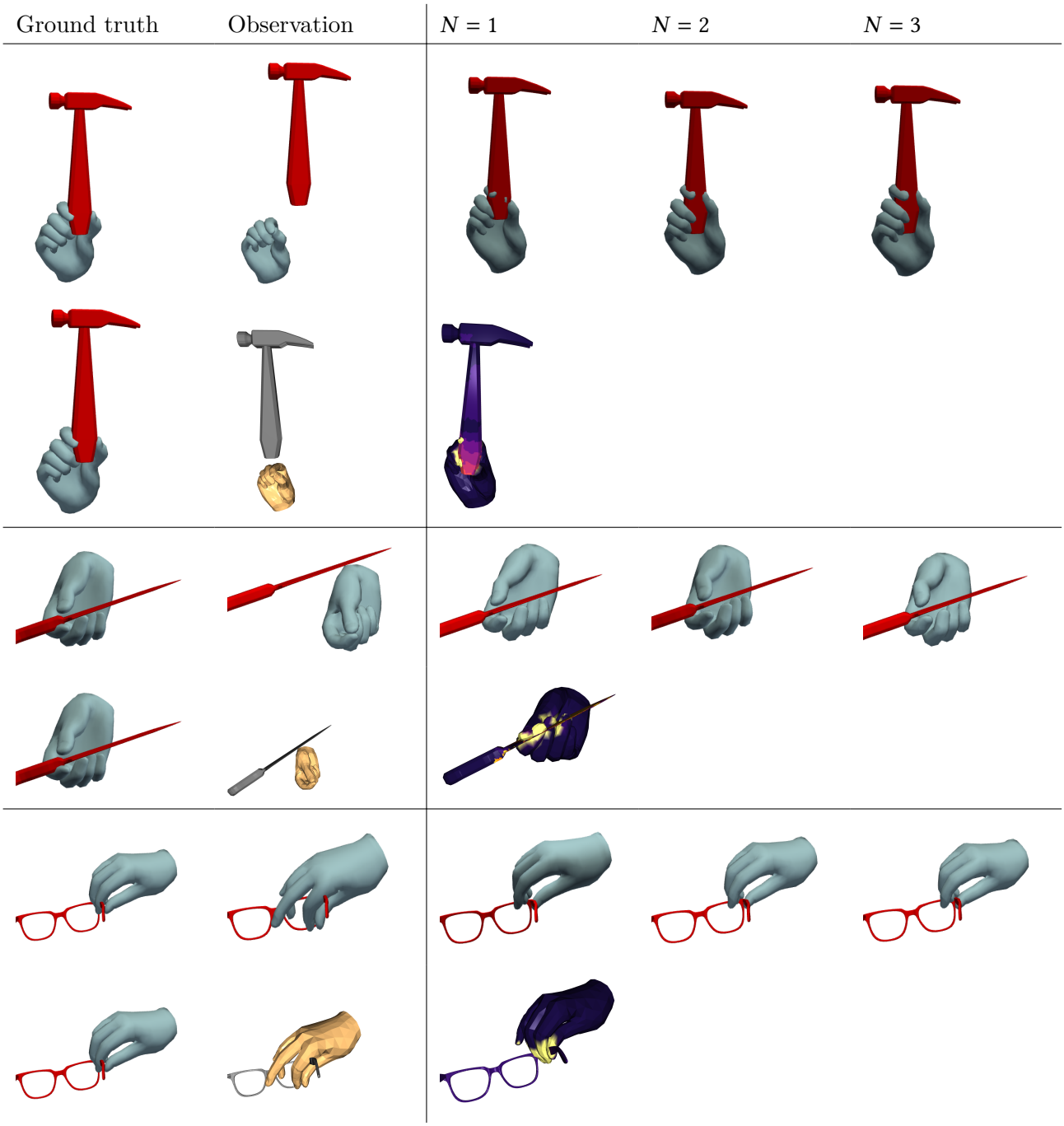



Figure 4.7: Qualitative comparison of $\mu$AggVED (top row of each section) and ContactOpt [1] for $N$ observations, where ContactOpt [1] only handles 1 observations. $\mu$AggVED exhibits grasps closer to the ground truth than ContactOpt [1], especially for $N = 3$ observations as clearly seen on the *knife* example. However, our method shows poor physical plausibility due to the lack of contact modelling and penetration regularization.

# Chapter 5

# *p*-CHOIR: Representing contacts

Based on "**A Versatile and Differentiable Hand-Object Interaction Representation**"

---

Article published at *WACV 2025*

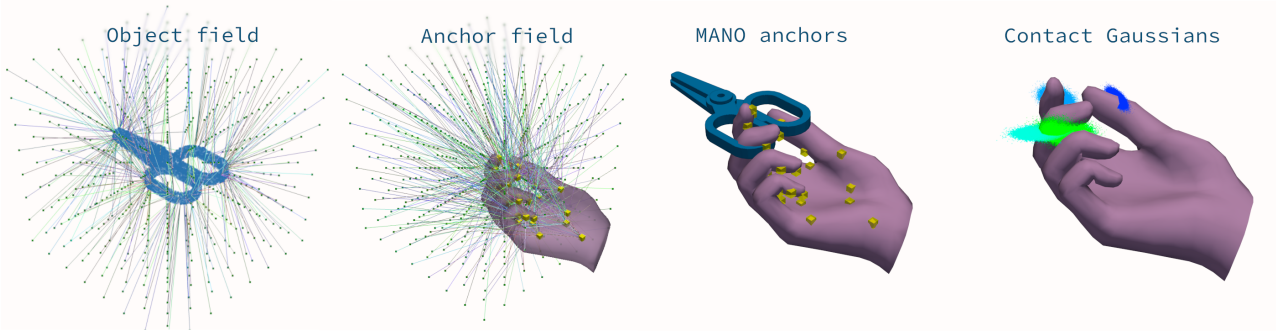



Figure 5.1: Decomposition of our Probabilistic Coarse Hand-Object Interaction Representation (*p*-CHOIR). From left to right, our representation encodes the object geometry with point-wise distances in a regular grid (coloured rays), the hand shape and pose as point-wise distances to 32 MANO anchors on the mesh surface (yellow cubes), and the hand contact points as probability densities form multivariate 3D Gaussian distributions (coloured point clouds). *p*-CHOIR is a fully-differentiable, lightweight representation of the hand-object pair in an object-centric coordinate system.

Up to this point, CHOIR represents the shape and pose of a hand-object pair in a lightweight, continuous and differentiable manner. This shows promise in real-time HOI denoising, since CHOIR yields SOTA hand pose accuracy. However, it is clear that in terms of contact metrics and qualitative samples, our proposed method is lacking. To address this, we focus this chapter on extending CHOIR with a dense hand contact map in a fully probabilistic manner (as a continuous function) for a compact parameterization, hence making it easier to achieve fit by a neural network. To clarify the distinction, we name this extended representation $p$-CHOIR, for *probabilistic* CHOIR.

We begin with a motivation for contact modelling in HOI, something often taken for granted in recent literature. By dissecting the recent methods in the literature, we address the pitfalls and limitations that go against our HOI modelling guidelines. These make clear the necessity for a continuous and dense contact map in the representation.

## 5.1 Motivation

### 5.1.1 Shape and pose are not enough

As the experimental results of grasp denoising show, CHOIR does not yet enable the generation of sufficiently realistic grasps. The generated grasps are plausible in terms of hand shape and pose, but they lack the physical plausibility of hand-object contacts. This is because our distance field representation learns an implicit relationship between the hand and the object, but no hard constraints are imposed or even exist by design. At inference, the TTO only aligns a hand mesh with the predicted field without regards to mesh penetration or contact with the object.

Hard contact constraints are difficult to integrate into the output hand and object representation, since it is typically achieved through physical simulation of the grasp[237, 152, 238]. However, CPF[37] modifies the MANO parametric hand mesh[120] to add anatomical constraints to the finger joints. This approach is close to integrating hard constraints but essentially boils down to energy minimization, where the hand fingers are strongly discouraged from bending in physically impossible ways. Another approach uses a physics-based deformable model to constrain the hand shape to be in a natural pose, as well as to not self-penetrate [239]. However, these methods do not apply to hand-object contact modelling. Unlike hand shape modelling, where a hand has fixed universal constraints, this harder

task requires a flexible model that can generalize to numerous object shapes. Therefore, the vast majority of recent works in contact modelling take a different approach, which can be categorized as soft constraints.

Soft contact constraints are typically modelled as contact maps [1, 34], correspondences [27] or distances[29, 37]. These predictions are then used to fit or refine an initial mesh with an agreement loss and regularization terms in a TTO stage. These methods can achieve high realism and physical plausibility (indicated by low physical simulation displacement), but they come with a trade-off for the hand-object penetration. Nevertheless, these methods are significantly faster to compute and more flexible than their simulation-based counterparts, which rely on meshes being watertight and with perfect mesh representations.

We take inspiration from recent works that adopt soft contact constraints to give CHOIR the ability to model contacts. We first dissect the most notable methods and discuss their limitations; then we propose a novel extension to CHOIR that enables the modelling of contacts, leading to $p$-CHOIR.

### 5.1.2 Limitations of related works

There are two main categories for hand-object contact representation in the literature. The first is a holistic representation which encodes the hand pose relative to the object or some other frame of reference. The previous chapter focuses on these methods, and up to this point CHOIR is a holistic representation that omits the contacts. The second category is a singular representation that focuses on the surface contacts, either on one surface only (object or hand) or on both. There is a large body of work that models contacts in this way, as it the task can be simplified to a 3D segmentation or a point-wise regression task on point clouds. To integrate hand-object contacts into CHOIR, we opt for a singular approach as a dense counterpart to our holistic approach for the shape and pose.

Before diving into the details of our representation method, we discuss the limitations of the most notable singular approaches to the problem: ContactOpt [1], GraspTTA [8] and $S^2$ Contact [161]. These limitations motivate the choice of our method.

### ContactOpt (Grady et al., 2021)

ContactOpt[1] is a foundational work in the field of static grasp refinement from noisy inputs. The authors propose a novel framework for denoising grasps by contact prediction and alignment. In a first stage, contacts are predicted on the hand and object surfaces via a classification task. 2048 points are sampled on the object surface, and the 778 MANO vertices are used for the hand. From there, PointNet++ [60] is trained to predict a contact value by classifying each vertex into 10 bins for a discretization of the range $[0, 1]$. In a second stage, the authors introduce a differentiable contact model that uses "capsule fields" placed along the normals of each vertex. These capsules approximately model soft tissue deformation and allow for a 2mm penetration of the mesh. This model yields a gradual contact drop-off which results in diffuse contact patches instead of binary contact points. This second stage is used to compute the ground-truth contact map to train the PointNet++ [60] classifier of the first stage (only for the hand, as the object heat maps are available in the training data). It is also used at inference for contact alignment by TTO, following our *specialized inference* and *explicit HOI formulation* guidelines (Section 2.7). A main strength of this method is how both contact maps are predicted from the joint hand-object pair, allowing the model to learn a mapping in a unified space. The authors train the first stage with noisy data to simulate an off-the-shelf pose predictor's inaccuracy. It is then evaluated by predicting accurate contacts, and a hand mesh (MANO) is fitted by minimizing the contact alignment loss. ContactOpt relies on random restarts to find a good solution, as it is prone to get stuck in bad local minima. This is due to the absence of any pose representation that would help better align the hand mesh with a richer grasping prior model, and is the main limitation of this method. It is unclear whether this method could be applied to other HOI tasks, but the absence of pose representation may result in a high failure rate, as already observed on challenging cases with ContactOpt's denoising capabilities.

### GraspTTA (Jiang et al., 2021)

Using the same concept as ContactOpt [1], GraspTTA [8] predicts an object contact map to use as error signal for the alignment of a hand mesh. Unlike ContactOpt, GraspTTA tackles the task of grasp synthesis, where the goal is to generate a plausible and realistic hand grasp given an object mesh. GraspTTA achieves this without a holistic contact-pose representation; they instead train a conditional

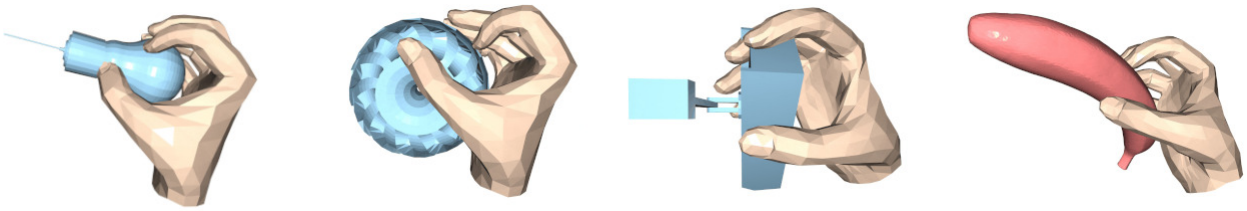


Figure 5.2: Grasp synthesis results from GraspTTA [8]. The generated grasps are not realistic and natural, with a mechanical look typical of claw crane machines.

variational auto-encoder (CVAE) to generate a grasp given an object. This acts as a strong guidance for the contact alignment and makes it less likely to be stuck in bad local minima. Similarly, the contact map is predicted with a PointNet-based method. But unlike ContactOpt, GraspTTA encodes the hand and object meshes separately and concatenates their embeddings into a latent representation. The latter is the input to a CNN-based contact regressor for the object contact map, here real valued in the interval $[0, 1]$. We hypothesize that embedding the hand and object separately may lead to sub-optimal learning due to the absence of local correspondences in the extracted features. Most importantly, the quality of the generated grasps is sub-par, as shown in Fig. 5.2: the hand poses are not realistic and natural, with a mechanical look typical of claw crane machines. This can be explained in two ways. Firstly, the authors use six hand prior contact regions that do not include the palm or the back of the hand, encouraging contact only on the inner phalanges: this results in "claw crane" grasps for intricate objects such as mugs and scissors. Secondly, since GraspTTA only predicts the object contact map, the model does not learn a rich prior of hand grasps, but only of object affordances. This is a major limitation of this method which results in unnatural grasps, especially for objects that require dexterity (as in contacts with the back and palm of the hand), such as scissors and mugs.

### $S^2$Contact (Tse et al., 2022)

In the same vein as ContactOpt[1], $S^2$ Contact [161] aims to refine noisy hand-object pairs by predicting accurate hand-object contact maps and realigning the grasp. The novel idea is to train a hand-object contact predictor from point clouds with semi-supervised learning. To do so, the authors exploit high-quality labelled 3D contact maps (from the ContactPose[80] dataset) to pre-train a student model. This model is then fine-tuned with pseudo-labels given by a teacher model trained on an Exponential Moving

Average (EMA) of the student's weights. The student-teacher model is trained with consistency losses to encourage the predictions to be physically realistic (geometric consistency) and visually consistent with 2D ground-truth segmentation masks (visual consistency). This training scheme allows ingesting more data than fully supervised methods, emphasizing our guideline on *more training data*. $S^2$ Contact [161] exposes the limitations of PointNet-based encoders in HOI, in particular, how they isolate the local features of points to allow permutation invariance and how they are inefficient. To address these limitations, the authors propose an efficient Graph Convolutional Network (GCN) architecture to capture the hand-object pair's global context. This allows the model to learn HOI embeddings that are more expressive and efficient than the PointNet-based methods. Without surprise, $S^2$ Contact [161] achieves state-of-the-art grasp prediction accuracy on several benchmarks, demonstrating the positive impact of more data and better feature extraction, as discussed in our guidelines.

**Strengths and limitations** All three methods approach contact representation as classification or regression problems on point cloud data with either a PointNet-based method or a more efficient alternative. A strength for these methods is how well they can model grasps without modelling the pose. The predicted contact maps act as a sufficient source of ground truth to align a hand mesh by optimization. However, it is less evident that this is also a downside of these methods compared to holistic representations: they fail to encode any prior knowledge for dexterous grasps where at least one finger is not in contact with the object. This limitation results in methods that produce consistent and compulsory finger contact, which does not reflect the complex reality of object manipulation. In addition, these prior models become useless when used for dynamic manipulation modelling, where free fingers still support the grasp (as seen on Section 4.1.1). Furthermore, we previously discussed how point-wise embedding methods are sub-optimal from a generalization standpoint, as these models must interpolate in high-dimensional data and mitigate high-frequency noise inherent in object geometry. There is a need to distil the representation to its essential elements. Methods such as TOCH [27] and CPF [37], which adopt a holistic approach, exemplify this well.

### 5.1.3 A hand-centric, localized, continuous and probabilistic contact representation

To overcome the limitations of previous methods, we propose to investigate a form of contact representation with different properties: dense as opposed to point-wise, continuous as opposed to discrete, and parametric as opposed to explicit. For this, we explore using parametric functions to model local contact patches on the surface of the hand. We hypothesize that fixed reference local frames are easier to learn due to a distillation of the information that is useful to the task. This hypothesis is supported by the aforementioned methods as well as the encouraging results of the BPS representation [64] and the previously mentioned holistic approaches. Indeed, these demonstrate the superiority of methods which treat geometries and spatial features as local neighbourhoods in fixed reference frames.

In summary, to allow for a richer prior model that handles dexterous grasps, we upgrade CHOIR to combine our canonical shape and pose representation with a novel singular contact representation based on parametric probability density functions in local hand frames.

## 5.2 *p*-CHOIR: Adding probabilistic contacts to CHOIR

The gaps in the literature on contact representations suggest that there is a need for a representation with certain properties, such as hand-centric, localized and continuous. To address these gaps, we explore a novel contact representation that is composed of parametric functions in local hand frames, for explicit contact predictions on the hand surface. In particular, we propose to leverage multivariate Gaussian distributions to represent dense contact patches in a compact form. This allows to infer contact probabilities for any point on the hand surface, while being an easier learning task than regressing all vertex probabilities due to the continuous and smooth form of the Gaussian distribution.

With this novel contact representation, we reformulate CHOIR as a holistic HOI representation with a probabilistic representation of hand contacts, leading to *p*-CHOIR. As such, it is lightweight, scalable, fully differentiable, and continuous. To evaluate our enhanced representation, we propose a unified neural network architecture to learn a distribution of *p*-CHOIR that can be applied to any downstream task, in principle. Concretely, we demonstrate the applicability of our representation in two settings: grasp refinement and grasp synthesis. We show that our model, named *JointDiffusion*,

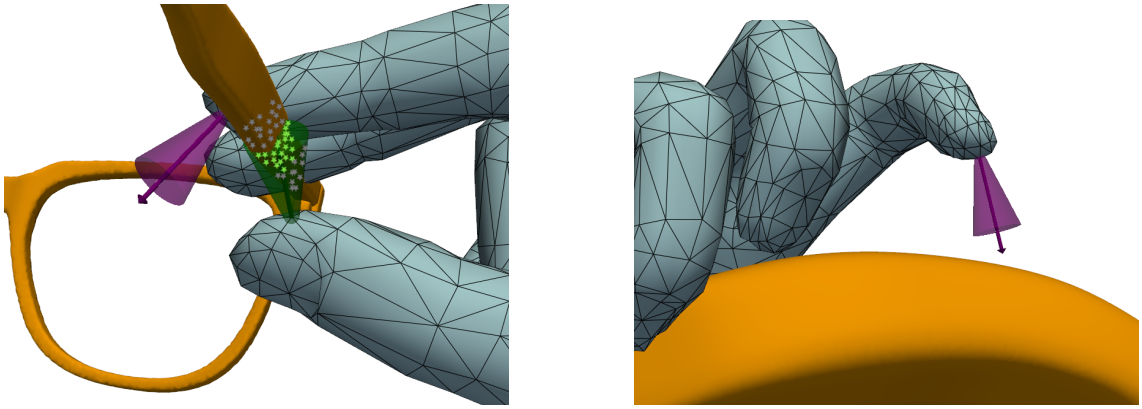


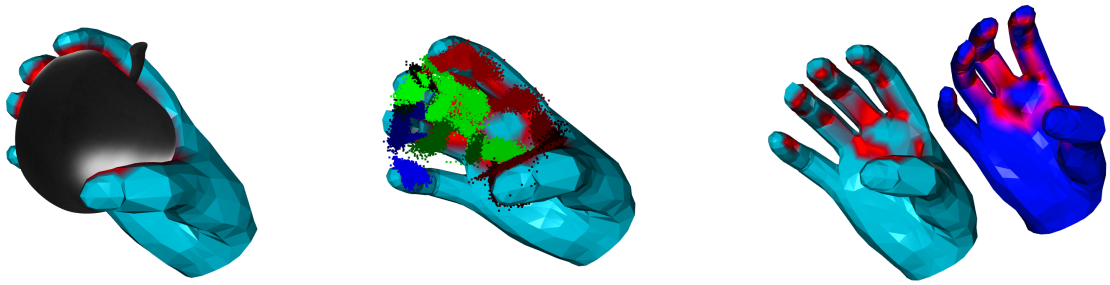
Figure 5.3: Illustration of the cone of tolerance used to determine the raw hand contact weights. For each hand vertex, the weights are the count of object points inside the vertex's cone defined along its normal vector. (Left) The green points on the object's surface are inside the cone, hence contributing to the hand vertex's weight while the grey points do not. (Left & Right) No object points are inside the purple cone: its vertex has a contact weight of 0.

can learn a prior $p$-CHOIR distribution that can be conditioned on either noisy partial $p$-CHOIRs or only objects shapes in BPS representation. This is achieved with a separate context encoder for each condition.

In this section, we detail the design of our novel contact representation and reformulate CHOIR into $p$-CHOIR. We then describe the learning method, *JointDiffusion*, that we use to learn the distribution of $p$-CHOIR. Finally, we present the new test-time optimization objectives that bring improvements in physical plausibility and grasp stability.

### 5.2.1 Parameterized representation of explicit contacts

Instead of representing hand contact points as a dense vector mapping each MANO vertex to a contact weight or binary class, we opt for a lightweight and coarse representation based on 3D multivariate Gaussian distributions. Given a hand-object pair $(H, X^n)$, we first compute the contact weights $w_i$ for each MANO vertex $v_i$. We define it as the count of all object surface points within a cone of tolerance defined at the root of $v_i$, in the direction of the vertex normal $n_i$ (see Fig. 5.3). This approach is inspired by ContactOpt's contact capsules [1], which includes vertices inside the mesh. With a cone, we exclude these vertices such that the contact map does not encode penetration patches in favour of a simpler optimization objective. Effectively, $w_i = |\mathcal{S}|$ where $\mathcal{S}$ is the set of points $x$ obeying the



(a) Raw hand contact weights in red.

(b) 3D Gaussian distributions as coloured pointclouds.

(c) Recovered contact weights (left) *vs.* raw contact weights (right).

Figure 5.4: Visualization of our probabilistic contact maps (best seen in colour). (a) The raw hand contact weights are computed with our cone of tolerance method. (b) 32 3D Gaussian distributions are fitted – one for each MANO anchor – on the raw weights to obtain contact probability densities. (c) Comparison of the recovered dense contact map from probability densities and of the raw contact weights. Our method leaves gaps in the contact map to allow for mesh penetration of up to 2*mm* and improve contact fitting.

following conditions (where we set $\lambda = 4\text{mm}$ and $\kappa = \frac{4}{3}\pi$):

$$\boldsymbol{x} \in X^n, \tag{5.1}$$

$$\|\boldsymbol{x} - \boldsymbol{v}_i\| \leq \lambda, \tag{5.2}$$

$$\arccos \frac{\boldsymbol{n}_i \cdot (\boldsymbol{x} - \boldsymbol{v}_i)^T}{\boldsymbol{n}_i \|\boldsymbol{x} - \boldsymbol{v}_i\|} \leq \kappa. \tag{5.3}$$

From this dense contact map, we fit one 3D multivariate Gaussian distribution per MANO anchor. This is done using the weighted mesh vertices such that the probability densities match the location of the vertices with the most contact weight. In effect, given the set of vertices $\mathcal{V}$ and the set of associated weights $\mathcal{W}$, we define the multiset $\mathcal{V}_w$ composed of each element $\boldsymbol{v}_i \in \mathcal{V}$ repeated $w_i \in \mathcal{W}$ times:

$$\mathcal{V}_w = \{\underbrace{\boldsymbol{v}_i, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_i}_{w_i \text{times}} \text{ for all } \boldsymbol{v}_i \in \mathcal{V}, w_i \in \mathcal{W}\}. \tag{5.4}$$

We then maximize the likelihood of the Gaussian parameters given the multiset $\mathcal{V}_w$. This allows us to encode a probabilistic dense contact map for the hand mesh as a set of 32 multivariate normal distributions (MVN). Each hand vertex then gets a contact probability by querying the probability density function of the nearest anchor's Gaussian (see Fig. 5.4). For anchor $j$, the MVN is parameter-

ized by a mean vector $\boldsymbol{\mu}^j \in \mathbb{R}^3$ and a covariance matrix $\Sigma^j \in \mathbb{R}^{3\times3}$. Since the latter must be positive semi-definite, it can be challenging to predict it with a neural network. One approach is to represent it as a lower triangular matrix $L^j$ obtained via Cholesky decomposition, such that $\Sigma^j = L^j L^{jT}$, and to enforce the diagonal entries to be positive.of our probabilistic contact representation for a given hand mesh $H$ is thus:

$$\boldsymbol{c}_H = \begin{bmatrix} \boldsymbol{\mu}^0 & \boldsymbol{l}^0 \\ \boldsymbol{\mu}^1 & \boldsymbol{l}^1 \\ \dots & \\ \boldsymbol{\mu}^{31} & \boldsymbol{l}^{31} \end{bmatrix} \in \mathbb{R}^{32\times9} \tag{5.5}$$

where $\boldsymbol{l}^j \in \mathbb{R}^6$ is the vector containing the elements of and below the diagonal of the matrix $L^j$. In summary, a $p$-CHOIR (see Fig. 5.1) is redefined as

$$x_{p\text{-CHOIR}} = [X^n_{\text{BPS}} \in \mathbb{R}^M, \boldsymbol{d}_H \in \mathbb{R}^M, \boldsymbol{c}_H \in \mathbb{R}^{32\times9}]. \tag{5.6}$$

$p$-CHOIR is now richer as it comprises more parameters with the contact representation. To fully exploit it, we employ a Denoising Diffusion Probabilistic Model (DDPM) for its SOTA distribution modelling capability. In addition, we show how to use the learned distribution in two tasks via separate context encoders, for grasp denoising and generation.

### 5.2.2 Learning a $p$-CHOIR distribution for grasp denoising and synthesis

Having defined a rich HOI representation that models shape, pose, and contacts, the goal is to learn a prior distribution for it. With the prior in hand, we can derive conditional posterior distributions for any downstream task via conditioning. To achieve this, we propose to leverage the modelling capability of the recently popularized Denoising Diffusion Probabilistic Models (DDPM). DDMPs excel at learning complex distributions [13] as Latent Variable Models. However, this learning framework rests on the modelling capacity of a backbone neural network which is typically tailored to the task with inductive biases. One common such backbone, for image and 3D data, is the U-Net [49].

The U-Net architecture is widely used as the backbone for DDPM in image generation literature [13]. This is because the architecture possesses several key features. Firstly, the autoencoder architec-

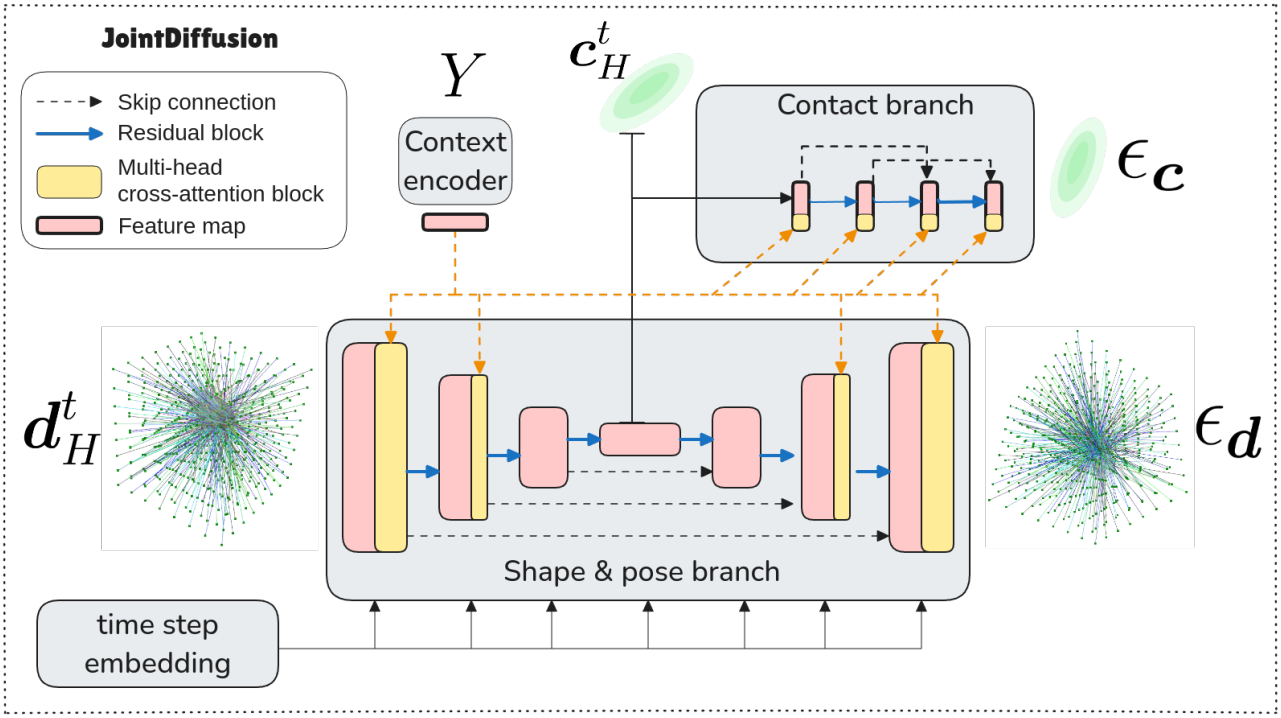



Figure 5.5: Architecture of *JointDiffusion*. The 3D U-Net predicts the noise sample $\epsilon_{\boldsymbol{d}}$ for the hand distance field $\boldsymbol{d}_H$. The contact prediction branch predicts the noise sample $\epsilon_{\boldsymbol{c}}$ for the contact Gaussian parameters $\boldsymbol{c}_H$ and receives the features of the U-Net's bottleneck as input. This joint learning encourages the U-Net to extract features relevant to both tasks, enhancing the accuracy of the learned $p$-CHOIR distribution.

ture extracts the most relevant features by using the information bottleneck technique [232]. Secondly, convolutions allow for translation equivariance in the learned representations, which improves feature extraction while still enabling invariance in the predictions. Thirdly, the U-Net is well-suited for modelling fine details in images, making it highly effective at predicting the fine details of the noise sample in the *reverse process*. These benefits led to our decision to use a regular grid for the BPS, as opposed to points uniformly sampled inside a unit sphere. This allows to use 3D convolutions and maximize parameter reuse in the model.

Our goal is to infer the posterior conditional distribution of hand poses $p(\boldsymbol{d}_H, \boldsymbol{c}_H | y)$ based on an observation $y$, where $\boldsymbol{d}_H$ and $\boldsymbol{c}_H$ are parts of $p$-CHOIR (see Eq. (5.6)). The context $y$ is either (1) a noisy hand-object pair, encoded as a partial $p$-CHOIR with missing contacts $\boldsymbol{c}_H$, or (2) an object point cloud encoded as $X^n_{\text{BPS}} \in \mathbb{R}^M$, *i.e.* a partial $p$-CHOIR with missing contacts and hand pose.

To learn this distribution we separately predict the noise samples for the distance field $\boldsymbol{d_h}$ and the contact Gaussians $\boldsymbol{c}_H$, denoted $\boldsymbol{\epsilon_d}$ and $\boldsymbol{\epsilon_c}$ respectively. This is motivated by the structure of $\boldsymbol{d}_H$ which allows the use of convolutions, while $\boldsymbol{c}_H$ is a vector in $\mathbb{R}^{32\times 9}$. Thus, our DDPM backbone, named *JointDiffusion*, (see Fig. 5.5) is composed of a 3D U-Net for the prediction of $\boldsymbol{\epsilon_d}$, and of a second decoder for the prediction of $\boldsymbol{\epsilon_c}$. This contact decoder is a fully connected residual network whose input is a concatenation of the latent variable $\boldsymbol{z}_t$ and of the latent features from the bottleneck layer of the U-Net. This encourages the model to learn pose and contact features in a shared space, such that the latent codes are relevant to both $\boldsymbol{d}_H$ and $\boldsymbol{c}_H$.

Our U-Net implementation uses MHSA to encourage relevant feature extraction and MHCA to condition the network on the context. The latter is embedded with an encoder identical to the U-Net encoder, with an additional spatial pooling mechanism via a shallow fully-connected network. We train one model per modality of context and include experiments on a multimodal model in supplementary material. As in the modern implementation [13], we use residual blocks and skip connections by concatenation to improve convergence and training stability. Similarly, we encode the diffusion step $t$ via the Transformer's positional encoding scheme [240] and embed this encoding with a shallow neural network. The feature maps are then conditioned on the diffusion step embedding by scale and shift, such that $\boldsymbol{g}_{\psi_t}$ is reparameterized as $\boldsymbol{g}_{\psi}(\cdot, t)$.

Running the inference on the DDPM, *i.e.* the reverse diffusion process, is akin to sampling from the posterior conditional distribution $p(\boldsymbol{d}_H, \boldsymbol{c}_H|y)$, given the observation (or condition) $y$. Therefore, the result is a $p$-CHOIR sample that represents a realistic hand grasp for a given object or noisy hand-object pair. In the next section, we cover the TTO algorithm that we design to recover a hand mesh from the $p$-CHOIR sample.

### 5.2.3 Aligning a hand mesh to the representation

While the state-of-the-art HOI fields are either not fully differentiable [27] or rely on random restarts [1, 27], fitting a hand mesh to $p$-CHOIR is done by gradient descent in two stages. Firstly, we fit a MANO mesh to the unsigned distance field of $p$-CHOIR, as in the previous chapter. Secondly, we adjust the hand contacts to the nearest object points using the contact Gaussians of $p$-CHOIR. These two stages rely on distance minimization, giving a smoother loss landscape than methods using

hand-object contact agreement objectives [1, 8].

### Coarse pose and shape fitting stage

As in the previous chapter, the first stage's objective $\mathcal{L}_{\text{PoseShape}}$ is composed of a reconstruction loss $\mathcal{L}_{\text{rec}}$, a shape regularizer $\mathcal{L}_{\text{shape}}$ and a pose regularizer $\mathcal{L}_{\text{pose}}$:

$$\begin{aligned} \mathcal{L}_{\text{PoseShape}} &= \lambda_1 \cdot \mathcal{L}_{\text{rec}} + \lambda_2 \cdot \mathcal{L}_{\text{shape}} + \lambda_3 \cdot \mathcal{L}_{\text{pose}} \\ \mathcal{L}_{\text{rec}} &= \|\boldsymbol{d}_H - \hat{\boldsymbol{d}_H}\|_2^2 \\ \mathcal{L}_{\text{shape}} &= \|\boldsymbol{\beta}_{\text{MANO}}\|_2 \\ \mathcal{L}_{\text{pose}} &= \|\boldsymbol{\theta}_{\text{MANO}} - \boldsymbol{\theta}_{\text{MANO}}^{\text{init}}\|_2 \end{aligned} \tag{5.7}$$

where $\boldsymbol{d}_H$ and $\hat{\boldsymbol{d}_H}$ are the respective ground-truth and predicted anchor distances (see Eq. (4.2)), $\boldsymbol{\beta}_{\text{MANO}}$ and $\boldsymbol{\theta}_{\text{MANO}}$ are the MANO shape and pose parameters, respectively. The shape regularizer prevents the hand mesh from over-deforming to satisfy the reconstruction loss, while the pose regularizer prevents strong deviation from the initial MANO pose estimate $\boldsymbol{\theta}_{\text{MANO}}^{\text{init}}$. Note that in the grasp synthesis case, we remove the pose regularizer.

We minimize $\mathcal{L}_{\text{PoseShape}}$ w.r.t. the MANO parameters alongside rotation and translation of the wrist joint with the Adam optimizer [235]. We set $\lambda_1$ to 1000 to bring the loss into the millimetre scale and found $\lambda_2 = 1 \times 10^{-4}$ and $\lambda_3 = 1 \times 10^{-8}$ to work well in practice. In the grasp refinement setting, all parameters are initialized with the noisy inputs from which $p$-CHOIR observations are built, leading to fast convergence ($\sim$ 150 iterations). Since the pose and shape encodings are unsigned distances, the fitting loss is obtained in very few lines of Python code (see supplementary material). This TTO stage fits a hand mesh to the predicted distance field, but does not take contacts and penetration into account. For this, we introduce stage two.

### Dense contact fitting stage

In the second stage, we refine the obtained hand grasp by minimizing the weighted distances from the MANO vertices to their nearest object points under some constraints. The weights of vertices $\boldsymbol{v}$ are obtained from the probability distribution function (PDF) $\Phi_j(\boldsymbol{v}_i)$ of the nearest anchor's contact

Gaussian (see Eq. (5.5)). For each MANO vertex $\boldsymbol{v}_i$, nearest anchor $j$ and nearest set of $K$ (such as 5) object points $\mathcal{K}^n$, the reconstruction loss is:

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N} \sum_{k=1}^{K} \Phi_j(\boldsymbol{v}_i) \cdot \|\boldsymbol{v}_i - \boldsymbol{p}_k\|_2^2, \quad \boldsymbol{p} \in \mathcal{K}^n. \tag{5.8}$$

This objective is minimized in conjunction with a penetration regularizer following [1], the shape regularizer defined in Eq. (5.7), and a pose regularizer to avoid deviating from the initial solution, defined as

$$\mathcal{L}_{\text{pose}} = \eta_1 \cdot \|R_{\text{MANO}} - R_{\text{MANO}}^{\text{stage1}}\|_2 + \eta_2 \cdot \|\boldsymbol{t}_{\text{MANO}} - \boldsymbol{t}_{\text{MANO}}^{\text{stage1}}\|_2 \tag{5.9}$$

where $R_{\text{MANO}}$ and $\boldsymbol{t}_{\text{MANO}}$ are the respective rotation matrix and translation vector of the MANO mesh, $\eta_1 = 1 \times 10^{-2}$ and $\eta_2 = 1 \times 10^{-1}$. The final objective is

$$\mathcal{L}_{\text{Contacts}} = \lambda_4 \cdot \mathcal{L}_{\text{rec}} + \lambda_5 \cdot \mathcal{L}_{\text{penetration}} + \lambda_6 \cdot \mathcal{L}_{\text{pose}} + \lambda_2 \cdot \mathcal{L}_{\text{shape}} \tag{5.10}$$

where $\lambda_4 = 10$, $\lambda_5 = 1000$, and $\lambda_6 = 0.5$. We optimize this loss w.r.t. the same parameters and with the same method as in the previous stage.

In the next section, we evaluate the combined $p$-CHOIR + *JointDiffusion* + TTO solution on grasp refinement and synthesis benchmarks, and show how much each component contributes to the accuracy and plausibility of our grasps.

## 5.3 Evaluation

Our solution consists of (a) a representation ($p$-CHOIR), (b) a learning method (*JointDiffusion*) for denoising and synthesizing interactions, and (c) a hand-mesh fitting algorithm (TTO). We evaluate this solution in two settings and in Section 5.3.2, we evaluate a multi-modal variant trained in both settings.

**Grasp refinement** For static grasp refinement, we employ the Perturbed ContactPose benchmark of Chapter 4, as explained in Section 4.3. We train and evaluate all methods on subject splits as well as object splits to evaluate generalizability.

**Grasp synthesis** In addition to refining noisy interactions, we demonstrate that our *JointDiffusion* can be used to synthesize novel interactions for unseen objects. To do this, we train *JointDiffusion* on ContactPose with subject splits. We encode the object mesh in BPS representation as input to the context encoder. For quantitative comparison, we retrain a recent SOTA method in grasp synthesis: GraspTTA [8]. This method also uses test-time adaptation, which makes it a good baseline to compare against. We use the same training, validation and test splits as for Section 5.3. This benchmark evaluates the capabilities of our model to learn the complex interaction between hands and objects. In grasp refinement, partial noisy information is given to the model during inference, but in grasp synthesis, the model must generate plausible grasps without prior information other than the training data. In Section 5.3.2, we evaluate our multimodal variant on the OakInk benchmark [83] against GrabNet[9].

**Keypoint baseline method** To evaluate the expressiveness and efficacy of each component of CHOIR, we design a diffusion model backbone that allows us to fit a simpler alternative to CHOIR. This simpler representation only encodes the hand pose and shape as 21 MANO joints $\boldsymbol{j}_H \in \mathbb{R}^{21\times3}$ and 32 MANO anchors $\boldsymbol{a}_H \in \mathbb{R}^{32\times3}$. The object is encoded as a vector of $K$ randomly sampled surface points $\boldsymbol{p}_O \in \mathbb{R}^{K\times3}$ where we set $K = 4096$ to match CHOIR which uses a grid of $16 \times 16 \times 16$ basis points. The final keypoint representation is defined as

$$\boldsymbol{r}_{\text{kp}} = [\boldsymbol{p}_O \in \mathbb{R}^{K\times3}, \boldsymbol{j}_H \in \mathbb{R}^{21\times3}, \boldsymbol{a}_H \in \mathbb{R}^{32\times3}]. \tag{5.11}$$

However, as in *JointDiffusion*, this model learns to predict the hand part only, defined as

$$\boldsymbol{r}^{\text{H}}_{\text{kp}} = [\boldsymbol{j}_H \in \mathbb{R}^{21\times3}, \boldsymbol{a}_H \in \mathbb{R}^{32\times3}] \tag{5.12}$$

The backbone of this diffusion model is composed only of residual blocks made of multi-layer perceptrons (MLPs). We use 4 residual blocks with a hidden dimensionality of 512.

In effect, in this baseline, we only replace the 3D U-Net component of *JointDiffusion* with a residual MLP and remove the contact prediction branch, while keeping cross-attention and the same time step conditioning scheme. The context encoder is also replaced with a residual MLP of hidden

Table 5.1: Evaluation of our approach on static grasp refinement against SOTA methods on the Perturbed ContactPose benchmark. * means reported figures. All methods are evaluated with one non-cherry-picked output per sample. *JointDiffusion* shows greater contact accuracy and outperforms all methods on most metrics, especially contact metrics (F1, Precision, Recall) and intersection volume, showing greater contact fidelity on the hand locations. Best results are in **bold** and second best are underlined.

| **Method** | **MPJPE** (mm) ↓ | **R-MPJPE** (mm) ↓ | **IV** ($cm^3$) ↓ | **F1** (%) ↑ | **Precision** (%) ↑ | **Recall** (%) ↑ |
|---|---|---|---|---|---|---|
| Perturbed data | 83.02 | 21.55 | 6.99 | 1.55 | 1.88 | 2.74 |
| ContactOpt [1] | 32.88 | 28.17 | 12.83* | 17.27 | 13.24 | **34.30** |
| S2Contact [161] | 33.65 | 26.61 | – | 17.63 | 13.98 | 32.91 |
| TOCH [27] | **26.96** | 29.24 | 10.14 | 22.23 | 21.46 | 25.09 |
| *JointDiffusion* (ours) | 27.69 | **23.54** | **6.04** | **27.20** | **25.21** | 32.80 |

Table 5.2: Quantitative evaluation of our approach on static grasp refinement against SOTA methods on the Perturbed ContactPose benchmark with object splits. * means reported figures. *JointDiffusion* is evaluated with one non-cherry-picked generated grasp per sample. *JointDiffusion* shows greater contact accuracy and outperforms other methods on most contact metrics, although ContactOpt [1] and S2Contact [161] retain greater recall scores due to their objective which maximizes the hand-object contact ratio, hence reducing false negatives. Best results are in **bold** and second best are underlined.

| **Method** | **MPJPE** (mm) ↓ | **R-MPJPE** (mm) ↓ | **IV** ($cm^3$) ↓ | **F1** (%) ↑ | **Precision** (%) ↑ | **Recall** (%) ↑ |
|---|---|---|---|---|---|---|
| Perturbed data | 83.02 | 21.55 | 6.99 | 1.55 | 1.88 | 2.74 |
| ContactOpt [1] | **35.05** | **29.13** | 12.83* | 15.39 | 12.04 | **30.36** |
| S2Contact [161] | 35.78 | 29.35 | – | 14.33 | 11.32 | 28.36 |
| TOCH [27] | 48.27 | 51.13 | 17.63 | 11.18 | 10.74 | 13.54 |
| *JointDiffusion* (ours) | 42.54 | 29.55 | **2.90** | **21.40** | **21.94** | 23.05 |

dimensionality 2048. We experimented with a PointNet++-based encoder but observed a degradation in performance.

### 5.3.1 Qualitative & quantitative results

We use several key metrics to quantify hand pose and contact error, as detailed in Section 2.8. In particular, we employ the (Root-aligned) Mean Per-Joint Pose Error (R-MPJPE and MPJPE) for either world space error (MPJPE) or object space error (R-MPJPE). However, a low pose error is not always indicative of a realistic or well-refined grasp. For denoising, recovering intended contacts is more important: the Precision score captures intended grasp locations, while a high Recall score implies fewer false negatives. The latter can be misleadingly increased by maximizing hand contact, leading to less dexterous grasps (as shown by ContactOpt [1] on Fig. 5.6). The F1 sore (harmonic mean of Precision and Recall) is the most meaningful metric for hand contact fidelity. For synthesis,

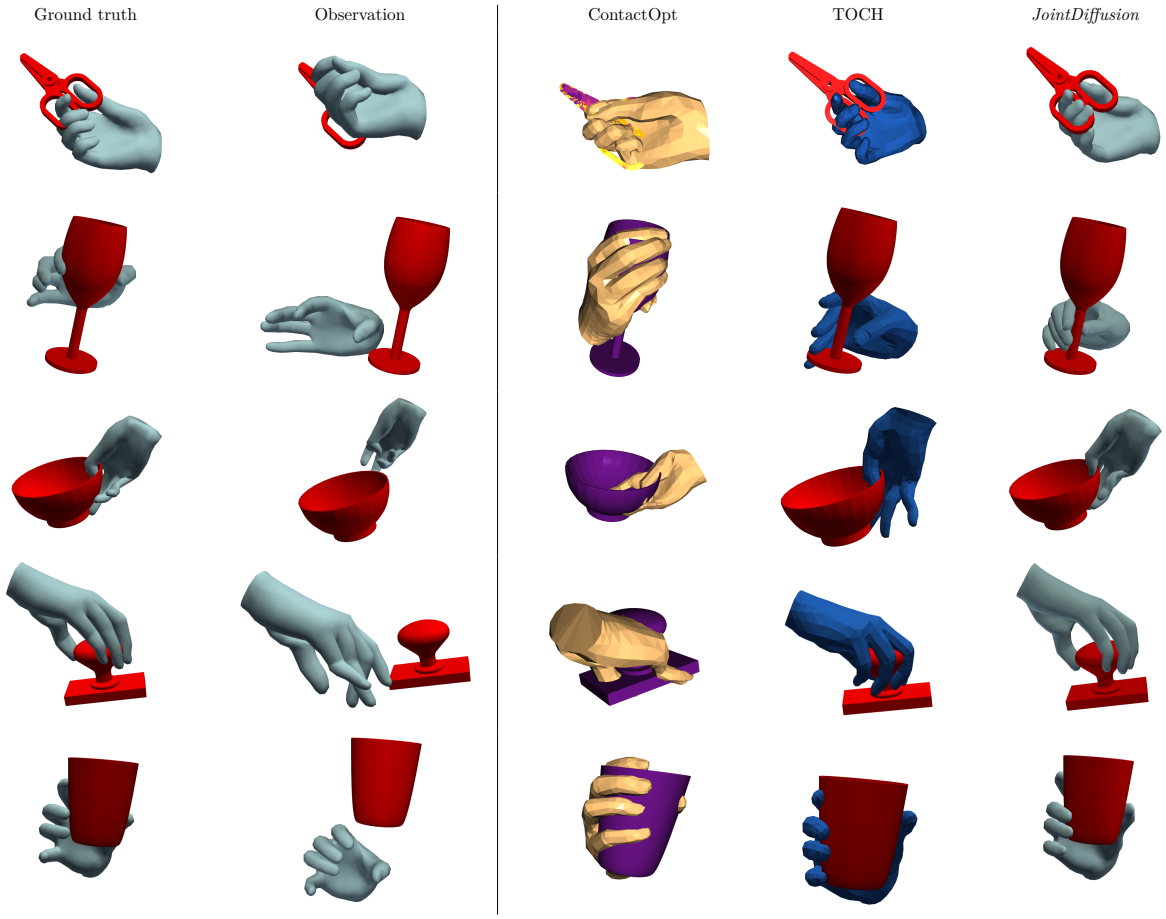



Figure 5.6: Qualitative comparison of grasp denoising on the Perturbed ContactPose benchmark. Our method produces less penetration and more natural contacts than TOCH [27], and substantially better output than ContactOpt [1] which maximizes hand-object contact.

we employ the simulation displacement (SD) metric with IV to evaluate the feasibility and stability of grasps.

For grasp refinement, Table 5.1 shows that our method *JointDiffusion* outperforms two SOTA methods on most metrics, and comes second best in the remaining metrics. In particular, our method brings a 5% improvement over TOCH [27] and 10% over ContactOpt [1] & S2Contact [161] in contact F1 score. Our method demonstrates the lowest R-MPJPE (−3mm over S2Contact) and highest contact precision (+3.8% over TOCH), indicating a higher contact and grasp fidelity than both methods. With the highest contact precision, the lowest intersection volume (−4mm over TOCH), and a high contact recall, *JointDiffusion* demonstrates the most accurate contact inference, as seen with a challenging case

Table 5.3: Evaluation of our approach against GraspTTA [8] on static grasp generation for the ContactPose benchmark. †: contact fitting enabled. Best results are in **bold**, second best are underlined.

| **Method** | **IV** ($cm^3$) ↓ | **SD** (cm) ↓ |
|---|---|---|
| GraspTTA [8] | 5.17 | 3.81 |
| *JointDiffusion* | 8.13 | 2.05 |
| *JointDiffusion* † | **4.51** | **2.07** |

on Fig. 5.6. This comes at the cost of a negligible penalty in absolute pose, where TOCH outperforms *JointDiffusion* by less than 1mm. We show more qualitative samples in Fig. 5.10, Fig. 5.11, and Fig. 5.12, where our method demonstrates fidelity in the reconstructed finger contacts, as opposed to ContactOpt [1]. ContactOpt remains 2% better in hand contact Recall since it aims to maximize hand-object contact and thus reduces false negatives, closely followed by S2Contact which builds upon the former. However, with a 12% worse contact Precision than our method, it cannot yield the intended grasp with high accuracy, as reflected by a 10% lower F1 score and a failure on a challenging case in Fig. 5.6. Fig. 5.9 shows failure cases in challenging examples. While ContactOpt [1] fails to produce a plausible grasp for each object and noisy input, our method delivers satisfying results that still closely match the contacts of the ground-truth hand pose.

To evaluate the generalizability of our method in the grasp refinement setting, we retrain all methods on the Perturbed ContactPose benchmark [1] with object splits instead of subject splits. We hold 2 objects out of the validation split, and reserve 5 objects for the test split, namely: *doorknob, eyeglasses, apple, bowl, toothbrush*. This increases the difficulty of the benchmark, as all test objects were unseen during training. For a method to perform well in this setting, it must learn generalizable hand-object interaction in latent space. Table 5.2 shows that our method outperforms ContactOpt [1] & S2Contact [161] on most contact-based metrics, and TOCH [27] on all metrics. ContactOpt [1] and S2Contact [161] retain an edge on the recall score since they both maximize the hand-object contact ratio and therefore minimize false negatives, but at the cost of less contact fidelity since their precision scores are significantly lower than *JointDiffusion*. However, TOCH [27] fails to generalize to these objects, which can be explained by the lack of object representation in the TOCH field. We consider this task to be a main challenge in hand-object interaction understanding and will focus on object generalization in future work.

For grasp synthesis, Fig. 5.8 shows *JointDiffusion*'s ability to generate plausible and realistic grasps. It shows minimal penetration and consistent contact between the used fingers and the object, owing to rich contact modelling through $p$-CHOIR. These results are validated quantitatively on Table 5.3, where our solution outperforms the SOTA method, GraspTTA [8], while being more versatile in applications. *JointDiffusion* reduces the intersection volume by 13% and the simulation displacement by 46%, resulting in more stable and feasible grasps. For a qualitative evaluation, Fig. 5.7 shows samples of our generative model given an object mesh as input. The model is trained on the improved Perturbed ContactPose benchmark [1], *i.e.* all objects are seen during training. *JointDiffusion* generates visually plausible grasps with consistent finger contacts and minimal mesh penetration.

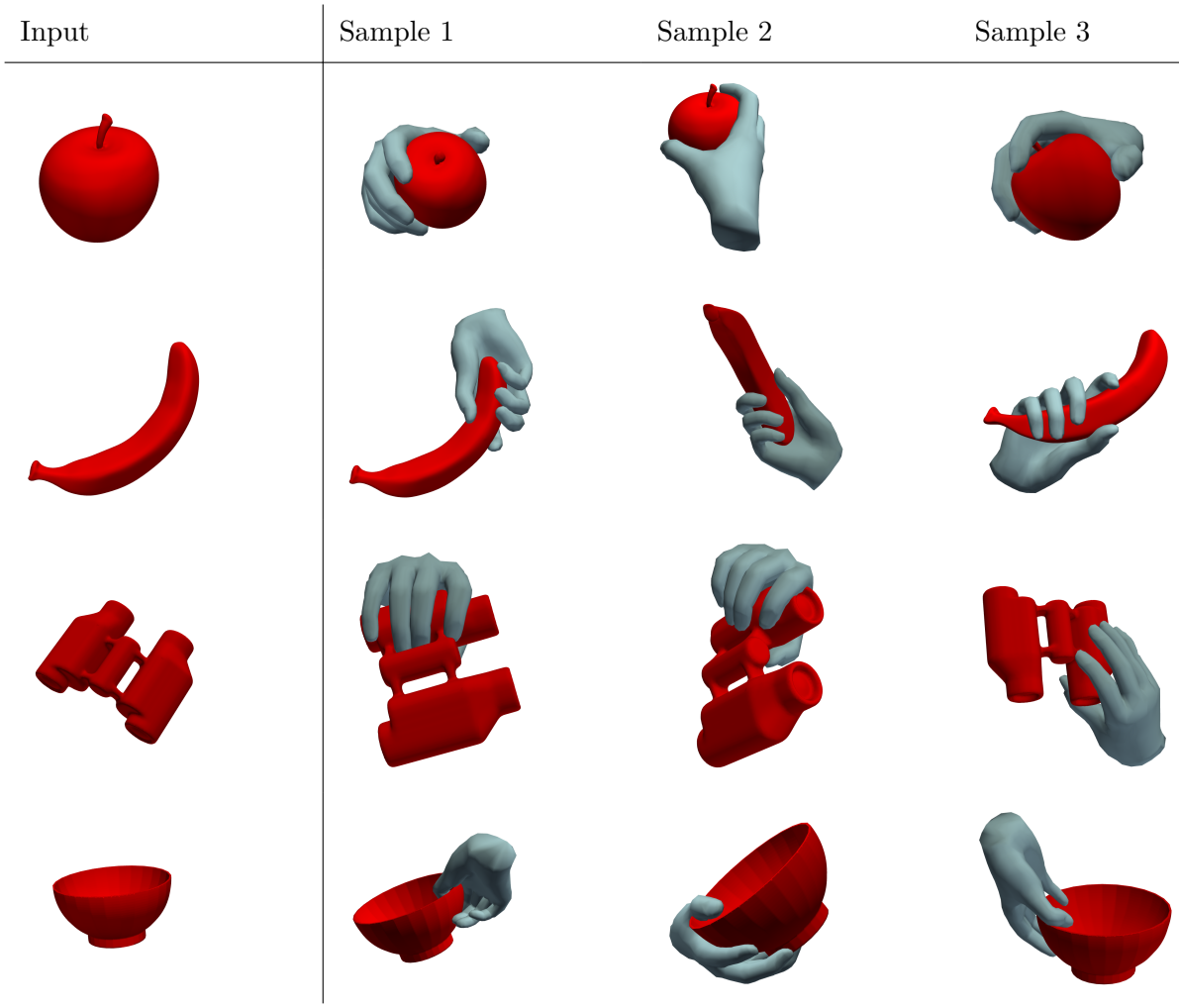



Figure 5.7: Qualitative evaluation of our method, *JointDiffusion*, trained on the object modality of input for grasp synthesis. Each sample is generated from the same input, the object mesh in canonical pose. *JointDiffusion* produces plausible grasps with minimal mesh penetration and consistent finger contacts.

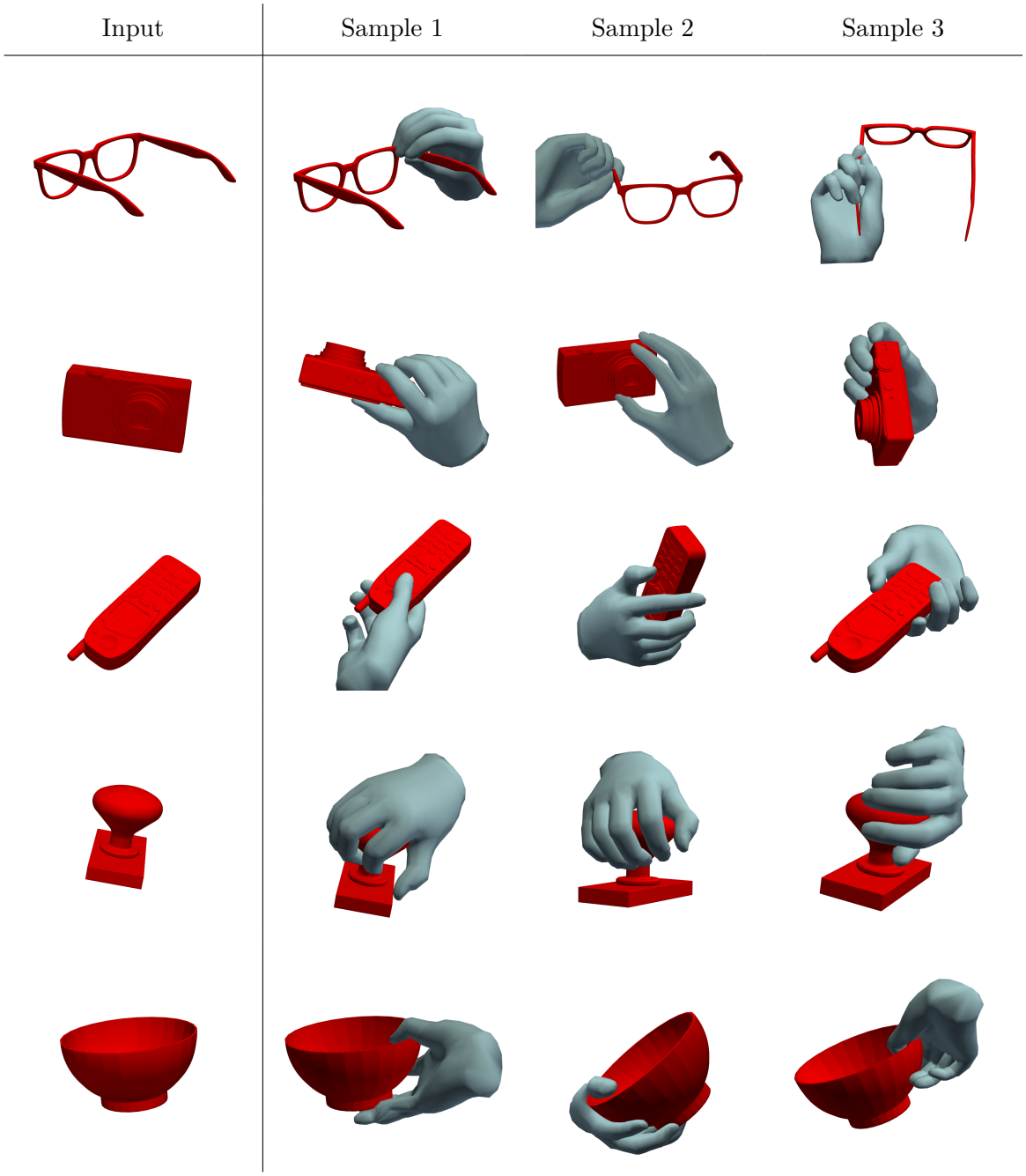



Figure 5.8: Generated grasps obtained with *JointDiffusion* on the ContactPose benchmark. The synthesized grasps show good finger-object contact and minimal penetration, showing the expressive contact modelling of CHOIR.

| Method | Ground truth | Observation | Prediction |
|---|---|---|---|
| *JointDiffusion* |  |  |  |
| ContactOpt |  |  |  |
| *JointDiffusion* |  |  |  |
| ContactOpt |  |  |  |
| *JointDiffusion* |  |  |  |
| ContactOpt |  |  |  |

Figure 5.9: Failure cases on a comparison of *JointDiffusion* and ContactOpt for the Perturbed ContactPose benchmark. While ContactOpt consistently fails at producing a plausible mesh after multiple restarts, our method results in minimal penetration and respected finger contacts with only one sample.

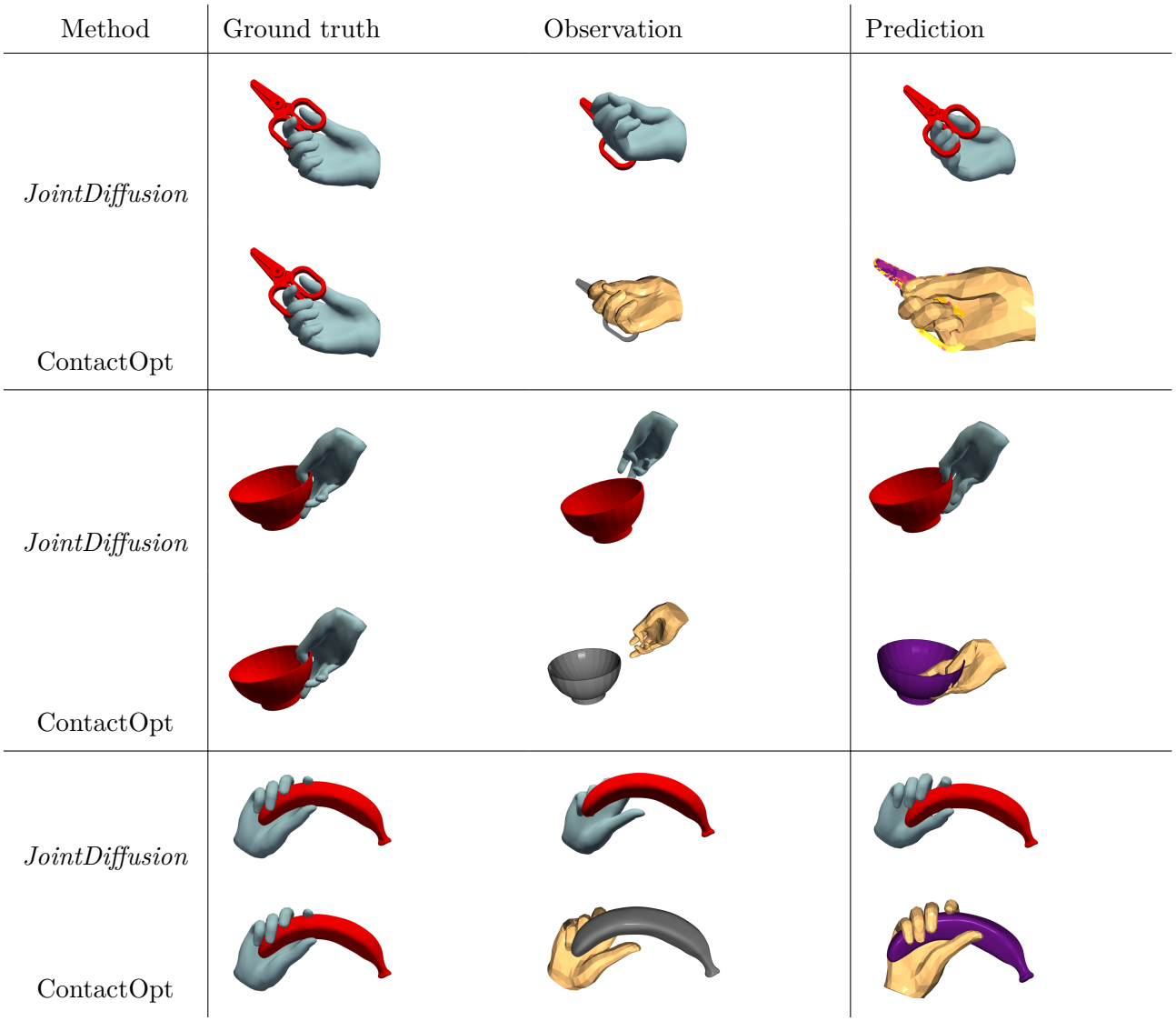



Figure 5.10: Qualitative comparison of *JointDiffusion vs.* ContactOpt on the Perturbed ContactPose benchmark. Our method, *JointDiffusion*, produces plausible grasps and maintains the fidelity of finger contacts while ContactOpt fails in challenging cases even with several random restarts. Our method only draws one sample and performs TTO without random restarts.

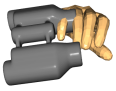
| Method | Ground truth | Observation | Prediction |
|---|---|---|---|
| *JointDiffusion* |  |  |  |
| ContactOpt |  |  |  |
| *JointDiffusion* |  |  |  |
| ContactOpt |  |  |  |
| *JointDiffusion* |  |  |  |
| ContactOpt |  |  |  |

Figure 5.11: Qualitative comparison of *JointDiffusion vs.* ContactOpt on the Perturbed ContactPose benchmark. Our method, *JointDiffusion*, produces plausible grasps and maintains the fidelity of finger contacts while ContactOpt fails in challenging cases even with several random restarts. Our method only draws one sample and performs TTO without random restarts.




Figure 5.12: Qualitative comparison of *JointDiffusion vs.* ContactOpt on the Perturbed ContactPose benchmark. Our method, *JointDiffusion*, produces plausible grasps and maintains the fidelity of finger contacts while ContactOpt fails in challenging cases even with several random restarts. Our method only draws one sample and performs TTO without random restarts.

Table 5.4: Evaluation of our multimodal model ✠*JointDiffusion* on static grasp generation against two state-of-the-art methods on two benchmarks. ✠*JointDiffusion* outperforms GraspTTA [8] on the ContactPose benchmark [23] and is on par with GrabNet [9] on the OakInk benchmark [83]. We used reported metrics for GrabNet [9] from the OakInk paper [83] and sampled one grasp per dataset sample for our method on both benchmarks. Best results are in bold.

| | ContactPose[23] | | OakInk[83] | |
|---|---|---|---|---|
| **Method** | **IV** ($cm^3$) ↓ | **SD** (cm) ↓ | **IV** ($cm^3$) ↓ | **SD** (cm) ↓ |
| GraspTTA [8] | 5.17 | **3.81** | - | - |
| GrabNet [9] | - | - | 6.60 | **1.21** |
| ✠*JointDiffusion* | **5.13** | 5.80 | **5.98** | 5.84 |

Table 5.5: Evaluation of our multimodal approach on static grasp refinement against three SOTA methods and our non-multimodal baseline on the Perturbed ContactPose benchmark. * means reported figures. Our multimodal model is marked with ✠. Our models were evaluated with one non-cherry-picked generated grasp per sample. Our full model *JointDiffusion* shows greater contact accuracy and outperforms ContactOpt [1], S2Contact [161], and TOCH [27] on almost all metrics. The multimodal version still outperforms these baselines on contact-based metrics and IV score for grasp refinement, while also being able to do grasp synthesis. Best results are in **bold**, second best are underlined.

| **Method** | **MPJPE** (mm) ↓ | **R-MPJPE** (mm) ↓ | **IV** ($cm^3$) ↓ | **F1** (%) ↑ | **Precision** (%) ↑ | **Recall** (%) ↑ |
|---|---|---|---|---|---|---|
| Perturbed data | 83.02 | 21.55 | 6.99 | 1.55 | 1.88 | 2.74 |
| ContactOpt [1] | 32.88 | 28.17 | 12.83* | 17.27 | 13.24 | **34.30** |
| S2Contact [161] | 33.65 | 26.61 | – | 17.63 | 13.98 | 32.91 |
| TOCH [27] | 31.41 | 36.67 | 13.16 | 19.18 | 18.34 | 22.08 |
| *JointDiffusion* | **27.69** | **23.54** | 6.04 | **27.20** | **25.21** | 32.80 |
| ✠*JointDiffusion* | 35.45 | 33.10 | **5.62** | 24.88 | 23.87 | 29.24 |

### 5.3.2 Multimodal model: grasp refinement & synthesis

We further explore our model expressiveness by jointly training two context encoders along with the diffusion backbone of *JointDiffusion*, as opposed to separately trained models for object conditioning and noisy hand-object pair conditioning. Fig. 5.13 shows qualitative results of the grasp synthesis from this model, while Fig. 5.14 shows qualitative results of the grasp denoising task for the same model. We trained two multimodal models: one on the ContactPose [23] dataset, and one on the OakInk [83] dataset which we only evaluate on grasp synthesis.

A quantitative evaluation on the denoising task is shown on Table 5.5, and one on the generation task is shown on Table 5.4. For the latter, the increase in simulation displacement (SD) for our method with contact fitting suggests that some hand penetration is helpful to a stable grasp. Note that the

synthetic nature of most OakInk samples results in incorrect vertex normals, adversely affecting our penetration regularization loss and performance. This could be solved with a different approach to penetration regularization, such as via the signed distance function.

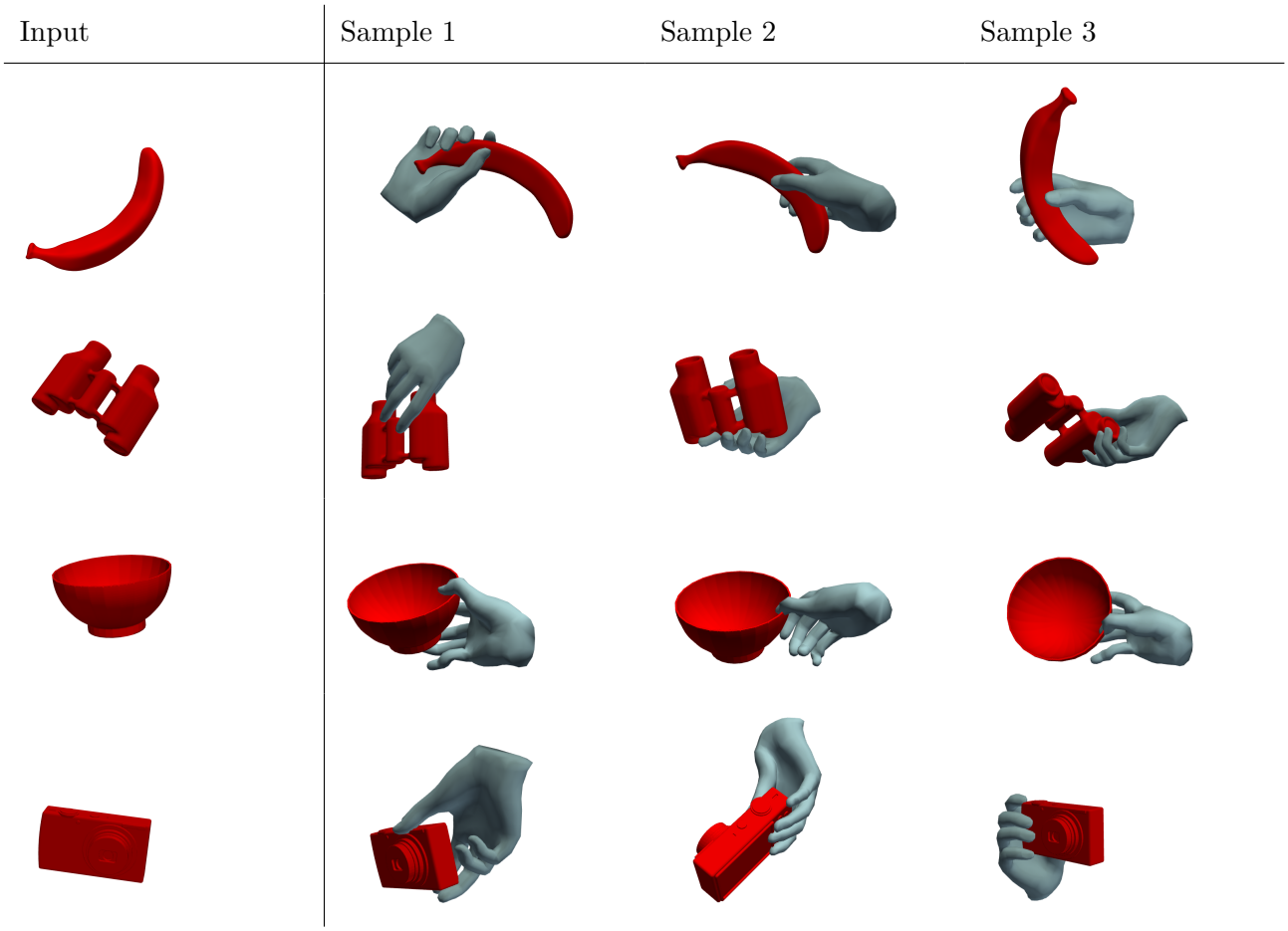



Figure 5.13: Qualitative evaluation of our multimodal *JointDiffusion*, trained on both object and noisy hand-object pair modalities, in the grasp synthesis setting. Each sample is generated from the same input, the object mesh in the canonical pose. *JointDiffusion* produces plausible grasps with minimal mesh penetration and consistent finger contacts.

### 5.3.3 Ablation study

To validate the design choices of $p$-CHOIR, we conduct an ablation study of its components by starting from a keypoint baseline and adding $p$-CHOIR components. The baseline uses a model similar to

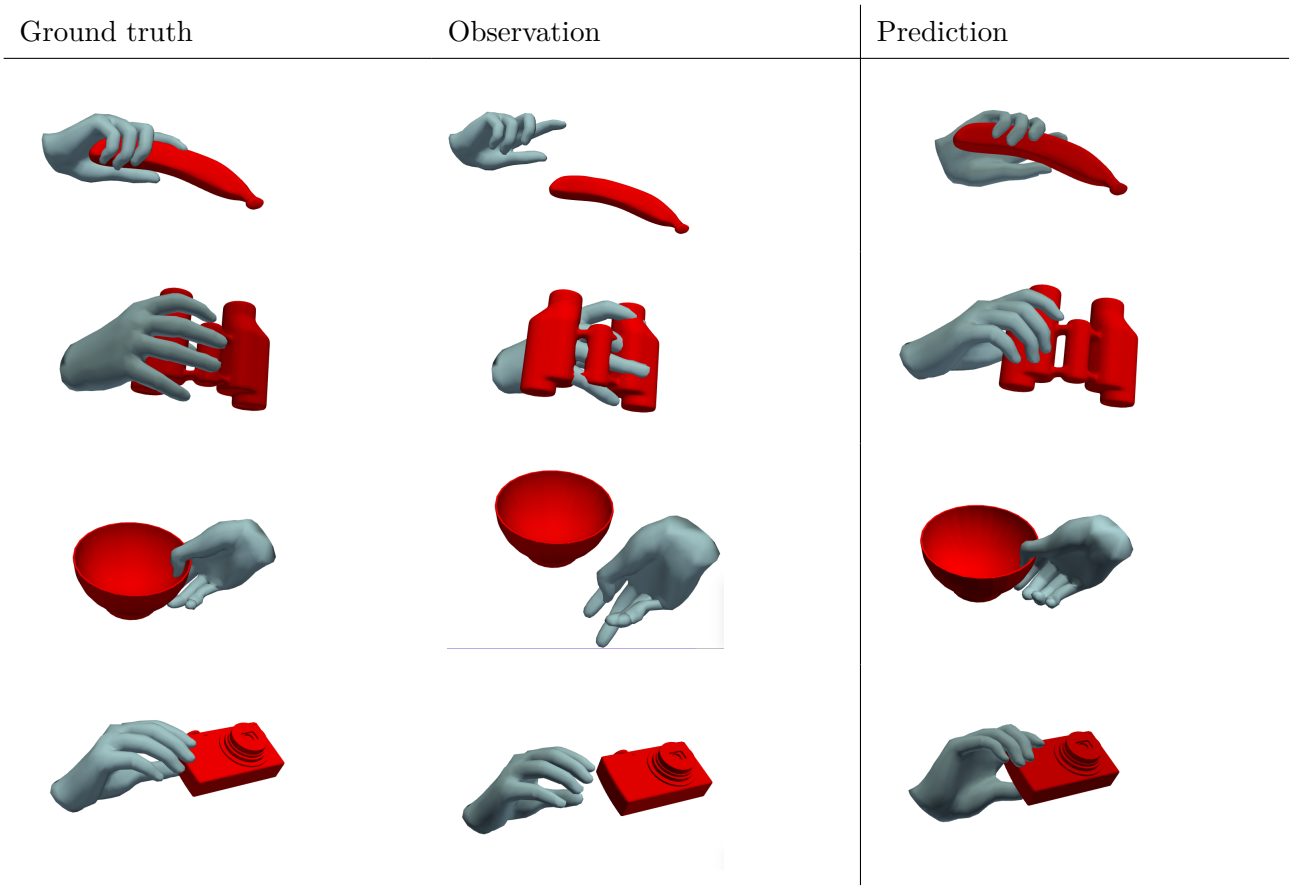



Figure 5.14: Qualitative evaluation of our multimodal *JointDiffusion*, trained on both object and noisy hand-object pair modalities, in the grasp refinement setting. *JointDiffusion* produces plausible grasps with minimal mesh penetration and respects finger contacts from the ground-truth mesh.

*JointDiffusion* to learn hand joint keypoints. In Table 5.6, *+BPS* corresponds to *JointDiffusion* without contact representation, and *+BPS +Contacts* matches *JointDiffusion* with the full $p$-CHOIR. We evaluate each row on the Perturbed ContactPose benchmark. Table 5.6 shows that while the baseline yields a lower MPJPE, it also gives the highest IV and lowest contact scores. The first two rows yield better pose accuracy by ignoring physical plausibility and contact fidelity, simplifying the learning objective. The last row improves on all contact and penetration metrics. Hence, $p$-CHOIR offers the best compromise for pose accuracy and contact fidelity in the denoising setting, while yielding plausible and stable grasps in the synthesis setting. These findings corroborate the previous experimental results.

Table 5.6: Ablation study of *p*-CHOIR on grasp refinement (best seen in colour). Best metrics are in **bold** and second best are underlined, with improvement or degradation w.r.t. the previous row in parentheses. The keypoint diffusion model is used as a baseline. The BPS representation for shape and pose encoding improves all contact metrics at the cost of a slightly higher pose error. Adding the probabilistic contacts substantially improves the contact metrics with a small cost in pose accuracy.

| **Method** | **MPJPE** (mm) ↓ | **IV** ($cm^3$) ↓ | **F1** (%) ↑ | **Precision** (%) ↑ | **Recall** (%) ↑ |
|---|---|---|---|---|---|
| KP. Baseline | **22.11** (-0.00) | 9.18 (-0.00) | 19.45 (+0.00) | 20.22 (+0.00) | 21.18 (+0.00) |
| *+BPS* | 24.69 (+2.58) | 9.12 (-0.06) | 20.77 (+1.32) | 20.86 (+0.64) | 23.38 (+2.20) |
| *+BPS +Contacts* | 27.69 (+3.00) | **6.04** (-3.14) | **27.20** (+6.43) | **25.21** (+4.35) | **32.80** (+9.42) |

### 5.3.4 Performance costs

To evaluate the computational costs of *p*-CHOIR, we timed its computation and that of TOCH for 50 grasps on an RTX 2080Ti and Intel i9-7900X. On average, TOCH takes ~ 8.89s (±3.99) while *p*-CHOIR takes ~ 0.13s (±0.015), a 68× reduction. When looking at the total inference time, including the model forward pass and TTO, ours converges in ~ 49s (± 16) and TOCH in ~ 14s (± 0.3). Our diffusion model accounts for ~ 13s of the total, hence is a major runtime bottleneck. Diffusion Models are inherently slow, but they are becoming faster, and new alternatives with similar properties can be easily integrated since our representation is agnostic to the learning method.

### 5.3.5 Limitations

The addition of a dense contact representation to *p*-CHOIR showed clear improvements in contact fidelity in the denoising setting, as well as grasp realism and physical plausibility in both denoising and generation. The net improvement brings overall more accurate hand-object interaction modelling, enabling richer priors that can be conditioned on other downstream tasks.

Despite its advancements, our method is not without limitations. The reliance on the BPS representation may limit the ability to capture detailed interactions. A side effect of this UDF-based representation is the highlights of edges and the diminishing of surfaces that are closer to the object's centre of mass. While increasing the number of basis points helps to mitigate this effect, it is not a viable solution as this entails higher compute cost and memory complexity. Furthermore, the model's focus on static interactions might restrict its application in real-world scenarios. While dynamic manipulation is outside the scope of this thesis, there are mainly two avenues to addressing

this gap. The first one is through the representation itself, by representing acceleration vectors for instance. The second one is through the learning method, by employing temporal models such as the LSTM. Since the latter is a more conventional approach which was shown to work well, even on "static features" as in TOCH [27], we leave this for future work. Moreover, by introducing a static model that demonstrates greater fidelity and realism, we raise the lower bound for dynamic models.

# Chapter 6

# Reflections, limitations and conclusions

This thesis explores new ideas for the design of hand-object interaction representations that improve the expressivity of grasp priors. This improvement is measured in terms of better quantitative and qualitative measures in the output grasps, and in terms of applicability to various HOI tasks. Since the scope of this thesis limits applications to grasp refinement from noisy inputs and generation from object inputs, the quantitative metrics used cover contact and pose accuracy (*i.e.* contact F1, pose MPJPE) as well as physical plausibility (*i.e.* penetration volume, simulation displacement). The main contribution is a novel representation, named *p*-CHOIR, which improves the SOTA in static grasp refinement and synthesis on the ContactPose [23] benchmark. *p*-CHOIR is the result of two core novelties in HOI representation: shape & pose representation with the Basis Point Set distance field [64] (Chapter 4), and contact representation with local multivariate Gaussian distributions (Chapter 5). These novelties themselves arise from the literature gaps in recent HOI representations, and follow the guidelines on HOI modelling introduced in this thesis.

To investigate the challenges in HOI modelling, in particular when dealing with distribution shifts associated with object categories, we propose a new approach to evaluating generalization in end-to-end methods. We start with the hypothesis that meta-learning can alleviate group distribution shifts that arise in the light of unseen object grasps. While the meta-learner seems to be more robust to distribution shifts, its accuracy is worse than the baseline. We investigate this with an analysis of the test-time adaptation and give detailed explanations for the under-performance of the meta-leaner. This lead to the discovery of factors that limit the learning of specialized parameters in the meta-

learner. From these factors, we introduce a set of guidelines for better HOI modelling, which we use to design *p*-CHOIR.

## 6.1 Reflections on *p*-CHOIR

The extensive quantitative and qualitative experiments on *p*-CHOIR show the importance and advantages of the two main components: the shape & pose representation, and the contact representation. These results along with the literature survey also point at the limitations of each component on their own. In this section, we adopt a retrospective view on these limitations and advantages, with additional support from the latest works in the literature.

### 6.1.1 Representing the shape and pose is not enough

Initially, CHOIR represents a static grasp as the shape and pose of a hand-object pair in an object-centric coordinate system, using an unsigned distance field in a fixed basis [64]. Representing the hand through fixed anchor distances has advantages over similar methods (*i.e.* TOCH [27]), mainly in terms of learning and fitting a mesh at test time, as no classification or search is needed to match the distances to certain vertices. However, the lack of contact representation results a grasp representation that is not explicit enough to capture the physical constraints of the hand-object interaction. This goes against the guidelines we propose in this thesis. It is clear, from quantitative and qualitative evaluation, that the produced grasps lack in physical realism and do not recover the ground truth contacts well. By contrast, when adding contact representation to *p*-CHOIR, there is a clear improvement in the fidelity of ground-truth contacts, as well as in the physical realism of generated grasps. This proves the importance of contact modelling in a grasp prior.

From an HOI perspective, recovering contacts accurately (or learning their prior distribution) is crucial. In fact, methods that only modelling contacts are able to produce plausible grasps [1, 31, 161]. On the other hand, methods that only model the shape and pose are lacking in physical realism, as shown in our experiments. However, the former suffer from certain issues such as the inability to model approaching hands or free fingers, hence contact representation alone is not enough.

### 6.1.2 Contact representation alone is not enough

Early trends in grasp priors focused on modelling contacts only (singular methods), which has the advantage of producing valid grasps with minimal mesh penetration and maximal contact. Essentially, a straightforward test-time optimization algorithm can be used to fit a hand mesh to a predicted contact map. However, this thesis argues that such methods are not able to represent full hand poses, since there are several degrees of freedom for each finger depending on the contact area. As such, these methods do not provide any information on a finger that is not in contact. Without a hand pose prior, the result can be unnatural finger poses or grasps that are devoid of free fingers, which is not realistic either.

Since *p*-CHOIR models the hand pose in addition to the hand contacts, we are able to learn a prior model that achieves the lowest pose error in a grasp denoising benchmark, as well as the highest contact F1 score, compared to several SOTA methods. These results show that holistic methods such as *p*-CHOIR are promising for richer grasp priors. In retrospect, the latest works on grasp modelling bring new support to these claims. Indeed, recent works seem to converge towards holistic methods for grasp representation [177, 180, 178].

### 6.1.3 A holistic representation is the way forward

Our experiments with *p*-CHOIR argue for the superiority of holistic HOI representations by integrating pose and contact information. Subsequently, novel HOI representations that adopt a similar approach have seen the light. In this section, we look at three notable works that tackle the problem of grasp synthesis given hand and object motion trajectories. The goal is here to generate a hand pose at each frame, given a canonical hand pose, such that the final sequence results in a smooth and realistic manipulation.

In the same vein as TOCH [27], CAMS [177] proposes to encode the hand pose as finger embeddings in a contact-centric frame derived on the object surface. The finger embeddings consist of canonical fingertip coordinates and normal vectors for the direction of each joint with respect to the fingertip. Those embeddings, combined with object contact representations made of binary flags and normal vectors, result in a holistic representation of the hand pose as a direct function of the object contact points. While *p*-CHOIR embeds the hand pose as distances from basis points to hand anchor points,

TOCH uses binary contact flags and the signed distance to the nearest hand point, as well as its coordinates in canonical pose. CAMS thus sits between *p*-CHOIR and TOCH.

GRIP [178] represents hand-object interaction in a holistic manner akin to *p*-CHOIR, in two levels of granularity. On a first level, GRIP exploits the BPS representation [64] to encode the object shape. Unlike *p*-CHOIR however, GRIP samples the basis points from a hemisphere centred around the base middle finger joint and facing outwards from the palm. This has the advantage of providing a denser object representation near the hand palm, which puts emphasis on the object grasp affordances. On a second level, GRIP encodes hand-object correspondences similarly to TOCH [27] as distances from 99 anchor points to the nearest object points. This effectively represents both the hand pose and the hand-object correspondences in an efficient and scalable form with a fixed basis point set that implicitly represents the hand pose. While GRIP demonstrates realistic grasp synthesis, the SOTA accuracy in hand pose error raises concerns as to the variety of the produced grasps, and if the model serves as a good HOI prior or essentially overfits the training data. Hence, purely distance-based HOI representations are promising, but there lacks evidence that they can entirely replace dense contact maps.

Then, GEARS [180] takes a similar approach to our proposed contact Gaussians (Chapter 5) for local, geometry-aware, hand-object interaction synthesis. To generate smooth HOI sequences, GEARS first predicts coarse hand poses for all frames, then computes local finger joint features, and finally predicts joint displacement vectors to drive the deformation of a MANO [120] hand mesh in its initial coarse pose. Then, to predict a coarse hand pose, GEARS encodes the object as a cropped point cloud in a canonical frame set around the wrist joint of the canonical MANO mesh. Finally, to refine this pose, GEARS adopts a similar approach to *p*-CHOIR and defines spheres around each joint, in which object surface points are sampled. These joint-local object points are then transformed from the global frame (where the hand is in its initial pose) to the MANO canonical frame. The result is a set of $K$ joint-centred point clouds, for $K$ MANO hand joints, which are fed to a PointNet [59] encoder to obtain feature vectors for all joints. *p*-CHOIR represents contacts as local probabilistic contact maps (in the form of 3D Gaussian distributions) in canonical anchor frames. Therefore, GEARS employs a similar idea to *p*-CHOIR, where contacts are represented as local 3D patches around fixed hand points. However, our compact parametric formulation allows to efficiently learn a distribution of

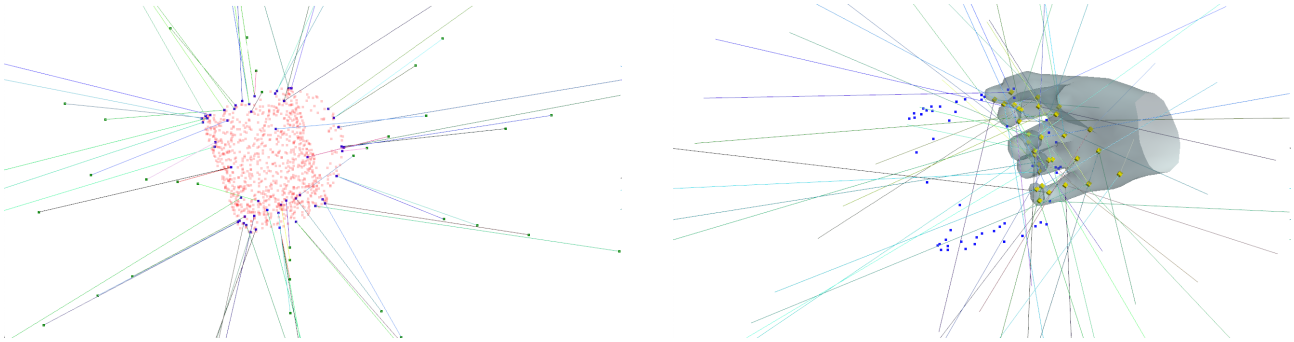
contact maps, which GEARS is not able to do without resorting to point cloud generation, an albeit harder task. Thus, GEARS brings new evidence that hand-centric, local contact representations such as $p$-CHOIR are a promising way to model hand-object interactions.

Finally, the recently published G-HOP [30] proposes an interaction prior based on a diffusion model with an SDF representation on a regular fixed gridR. Unlike $p$-CHOIR, G-HOP represents the hand-object pair in a canonical hand-centric and normalized frame, which has the advantage of capturing more nuance in common grasps. The hand is represented as a pose-conditioned semantic distance field with 15 channels for the distance to each joint. In contrast, $p$-CHOIR encodes 32 anchors on the hand surface with a fixed assignment for each basis point, thus removing redundancy and spreading the information. It is unclear whether G-HOP's representation is more expressive than $p$-CHOIR due to this fact, as this would require substantial ablation studies on this recently released work. As for the object, it is represented as a latent SDF embedding using a VQ-VAE [40] to lower the dimensionality and memory requirements. For the learning method, G-HOP adopts a 3D U-Net similar to *JointDiffusion*. One advantage of G-HOP over *JointDiffusion* is its use of score distillation sampling (SDS) [241] which allows using the trained diffusion model as a flexible prior. Effectively, SDS provides log-probability gradients which point towards the data manifold, with a single forward pass of a sample. The authors demonstrate the use of SDS with image-based reconstruction, without conditioning the diffusion model. G-HOP presents a major advancement in HOI modelling, with a truly flexible and rich prior for several downstream tasks.

In summary, these recent works bring evidence that confirm the advantage of holistic representations against singular pose and contact representations. This evidence is complementary to our experimental results, which bring support to our hypotheses on HOI representations, and further validate our guidelines on *canonical frames of reference* and *explicit HOI formulation*.

## 6.2 Limitations of $p$-CHOIR

The main limitation of $p$-CHOIR resides in the object representation. We recall that the object mesh is encoded as the distance to the nearest surface point from each basis point. As a side effect,



(a) Out of all object points (in red), only the blue ones are represented.

(b) The only points representing the object are on its edges and corners.

Figure 6.1: Illustration of the limitations of the BPS representation with a few dozen points; the rays represent distances from the basis points to their nearest object point. In Fig. 6.1a, only the blue points are encoded with the BPS representation [64], which tend to be on the edges and corners of the object. In Fig. 6.1b, the hand-object pair representation is very coarse with very few points on the object surface near the fingers. As the BPS resolution increases, the sparse areas become denser, but the disparities between outer edges and inner surfaces still remains and the memory cost increases quadratically.

this representation emphasizes the object's edges and protrusions, while smooth areas are sparsely represented (as seen in Fig. 6.1). This results in a loss of information in the object shape representation, which may degrade the modelling capabilities of the grasp prior. This is especially true for unknown objects, where the object shape is not well represented by the training set. While the effect can be dampened by increasing the number of basis points, it does not fix the problem but increases the memory footprint. This is a design flaw in the BPS [64] representation itself and would need to be addressed in future work.

While $p$-CHOIR is a scalable representation that takes advantage of modern GPUs, the learning method we employ hinders this efficiency. $p$-CHOIR is in fact $\sim 70\times$ faster than the closest equivalent, TOCH [27]. However, for $p$-CHOIR (Chapter 5) we employ a diffusion model with a 3D U-Net backbone and self-attention blocks. While this architecture enables the learning of a rich and versatile grasp prior, it incurs high inference costs. By design, DDPMs have a slow inference since they rely on traversing a Markov chain to generate samples. This effectively means running the forward pass of the backbone model (3D U-Net in *JointDiffusion*) multiple times, typically 100 times. However, we have used the original DDPM formulation [45] while there has been recent works that focus on reducing

the number of steps via knowledge distillation [242], using ordinary differential equations [243, 244] or even switching to a different paradigm for distribution learning and one-step generation [245, 246]. Thus, the inference cost of *JointDiffusion* can be significantly reduced by using advanced DDPM formulations, or by adopting another learning method altogether, as demonstrated with AggVED (Chapter 4).

On a functional level, $p$-CHOIR is not designed to represent dynamic manipulation. As this is a design choice since this thesis focuses on improving static grasp modelling, we consider dynamic manipulation to be outside the scope of this work. Nevertheless, using a static representation does not preclude the modelling of dynamic manipulation sequences. In fact, TOCH [27] and GRIP [178] are recent examples that show how a static representation combined with a temporal model (*i.e.* LSTM autoencoder) is a valid approach to model manipulation sequences. This is also true for image-based HOI reconstruction methods, where temporal sequences are processed on a per-frame basis [247, 33, 142], with some works using temporal smoothing [178, 248, 7]. There are countless examples where the learning method is entirely modelling the temporal dynamics from static inputs. However, there exist counter examples where the inputs to the model are comprised of temporal features (*i.e.* velocity vectors), such as HuMoR [249] for human body reconstruction from monocular videos. A more relevant case is CAMS [177], which proposes a novel HOI representation that specifies grasps as static finger embeddings in an object-centric and contact-centric frame. To model motion, they train continuous-time decoders which convert a sequence of static representations to continuous-time sequences. To achieve this, the decoders are trained as implicit models taking as input a latent vector for the static fields and a time embedding, akin to the Transformer [240] and DDPMs [45]. This is an interesting approach that could be explored in future work.

## 6.3 Future work

In retrospect, the past and recent literature show that $p$-CHOIR fills literature gaps that are relevant to the progress on HOI modelling. In particular, recent works following $p$-CHOIR bring forth new support to our claims that holistic representations are a worthwhile research direction for HOI modelling. However, the recent trends show a focus on dynamic manipulation, full-body interaction or bi-manual

manipulation. While these works show that static representations are still relevant for these advanced tasks, there seems to be an unexplored area for dynamic manipulation fields.

**Dynamic representation** Representing dynamic hand-object interaction is distinct from grasp modelling, as it requires the modelling of motion or forces and necessarily includes transitions from one functional grasp to another. While TOCH [27] achieves this with a fully static representation, the proposed learning method (an LSTM autoencoder) effectively models these motions and transitions. Subsequently, GEARS [180] and GRIP [178] adopted a similar strategy with more elaborate learning mechanisms, such as latent temporal consistency or spatio-temporal transformers. However, CAMS [177] distinguishes itself by proposing to learn an implicit function for time-continuous grasp embeddings. This is a step forward, although static representations and data augmentation are used to learn this implicit model. The main disadvantage of static representations lies in the inference fitting algorithm often used to obtain a mesh or animation sequence, although some works opt for a feed-forward model instead [178]. These algorithms typically use heuristics and rely on hyperparameters to adjust the grasp penetration-stability trade-off. Instead, a dynamic representation may allow removing the inference algorithm or feed-forward network altogether [249]. Therefore, we argue that there is motivation for a fully dynamic representation of hand-object manipulation, in the same vein as recent works on animation generation [52, 250].

**Richer object representation** While investigating dynamic representations represent a consequent leap from the current state of *p*-CHOIR, there is a direct limitation of *p*-CHOIR to be addressed. As previously mentioned, the BPS representation [64] severely hinders the expressivity of the object shape. One straightforward way to alleviate this is proposed in GRIP [178], where the basis points are centred around the middle finger and sampled from a hemisphere. This increases the resolution of object surface points that are encoded near the hand, but it does not remove the inherent limitation. Future work should focus on enhancing the granularity of the object geometry local to the hand contact regions, as in GEARS [180]. One interesting approach would be to exploit rich shape priors (*i.e.* in the form of a distance field) and infer expressive posteriors given an arbitrary amount of measurements at test time [42, 73].

**Representation learning** A review of the literature shows an increasing effort in designing better HOI representations. In contrast, and to our knowledge, very few works explore the area of representation learning [248]. Engineering HOI representations that are richer and address the current limitations is a challenge. For this reason, there is an unmet need in learning HOI representations that conform with our current needs. Through representation learning methods such as deep metric learning [251] and contrastive learning [252], we may learn to extract richer features than the ones we can design. This may also be used in conjunction with the SOTA representations, such as to embed them in a semantically meaningful and well partitioned space. Deep metric learning for instance, may enable the mapping of a dense HOI latent space where grasps and object shapes are embedded by similarity. With the recent rise of contrastive learning methods [253], the field of HOI modelling is equipped with new tools that remain unused.

## 6.4 Closing statement

In this thesis, we found that encoding the hand-object shape and contacts together, as a holistic representation, addresses limitations found in methods that use singular representations for either pose or contacts. We also found that using a compact and parametric form for contact representation allows learning a good prior distribution. This is an important and useful tool to have, as a rich prior can be conditioned on diverse observations, such that we can infer a good posterior in different tasks.

By focusing on learning a good grasping prior, and hence improving HOI modelling as a whole, researchers may enable downstream tasks across many domains to be improved. From the fields of robotics, graphics or mixed reality, learning a rich hand-object interaction prior has a direct positive impact on any related task. Thus, we hope that this thesis brings convincing evidence that this research direction is worth pursuing.

# Bibliography


[1] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021. xiv, 3, 7, 33, 34, 35, 39, 41, 67, 73, 74, 76, 77, 80, 83, 84, 85, 88, 92, 93, 94, 96, 97, 98, 99, 106, 112

[2] P. Zuccotti, *Everything we touch: A 24-hour inventory of our lives*. National Geographic Books, 2015. 1

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 2018. 3

[4] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11799–11808, 2019. 3

[5] Y. Hasson, G. Varol, I. Laptev, and C. Schmid, "Towards unconstrained joint hand-object reconstruction from rgb videos," *ArXiv*, vol. abs/2108.07044, 2021. 3, 31

[6] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall, "Hope-net: A graph-based model for hand-object pose estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6607–6616, 2020. 3, 29, 54

[7] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," *2020 IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 568–577, 2020. 3, 39, 117

[8] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the International Conference on Computer Vision*, 2021. 3, 33, 35, 39, 83, 84, 85, 93, 95, 98, 99, 106

[9] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision (ECCV)*, vol. 12349, pp. 581–600, 2020. 3, 28, 33, 39, 40, 74, 95, 106

[10] Z. Fan, M. Parelli, M. E. Kadoglou, M. Kocabas, X. Chen, M. J. Black, and O. Hilliges, "Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video," *ArXiv*, vol. abs/2311.18448, 2023. 3, 39

[11] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," *ArXiv*, vol. abs/1804.01654, 2018. 3

[12] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1749–1759, 2021. 3

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021. 3, 19, 90, 92

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *ArXiv*, vol. abs/2302.13971, 2023. 3

[15] R. Balestriero, J. Pesenti, and Y. LeCun, "Learning in high dimension always amounts to extrapolation," *ArXiv*, vol. abs/2110.09485, 2021. 3

[16] C. Bakker, G. Parks, and J. Jarrett, "On the application of differential geometry to mdo," in *12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, AIAA, 2012. 5

[17] A. Armagan, G. Garcia-Hernando, S. Baek, S. Hampali, M. Rad, Z. Zhang, S. Xie, M. Chen, B. Zhang, F. Xiong, Y. Xiao, Z. Cao, J. Yuan, P. Ren, W. Huang, H. Sun, M. Hrúz, J. Kanis, Z. Krňoul, Q. Wan, S. Li, L. Yang, D. Lee, A. Yao, W. Zhou, S. Mei, Y. Liu, A. Spurr, U. Iqbal, P. Molchanov, P. Weinzaepfel, R. Brégier, G. Rogez, V. Lepetit, and T.-K. Kim, "Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction," in *ECCV*, 2020. 6

[18] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *ArXiv*, vol. abs/1811.12231, 2018. 6

[19] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *ArXiv*, vol. abs/1911.08731, 2019. 6

[20] L. Tu, G. Lalwani, S. Gella, and H. He, "An empirical study on robustness to spurious correlations using pre-trained language models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 621–633, 2020. 6

[21] P. Izmailov, P. Kirichenko, N. Gruver, and A. Wilson, "On feature learning in the presence of spurious correlations," *ArXiv*, vol. abs/2210.11369, 2022. 6

[22] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018. 6

[23] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "Contactpose: A dataset of grasps with object contact and hand pose," *ArXiv*, vol. abs/2007.09545, 2020. 6, 36, 69, 74, 106, 111

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022. 6

[25] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "Humor: 3d human motion model for robust pose estimation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11468–11479, 2021. 6, 40, 70, 71

[26] I. M. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003. 6

[27] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll, "Toch: Spatio-temporal object-to-hand correspondence for motion refinement," in *European Conference on Computer Vision*, 2022. 6, 7, 34, 35, 36, 39, 61, 63, 64, 65, 67, 69, 73, 74, 76, 78, 79, 83, 86, 92, 96, 97, 98, 106, 110, 112, 113, 114, 116, 117, 118

[28] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang, "Grasping field: Learning implicit representations for human grasps," *2020 International Conference on 3D Vision (3DV)*, pp. 333–344, 2020. 6, 7, 31, 35, 37, 61, 62, 63, 65, 66

[29] H. Zhang, Y. Ye, T. Shiratori, and T. Komura, "ManipNet: Neural manipulation synthesis with a hand-object spatial representation," *Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 121:1–121:14, 2021. 6, 34, 36, 83

[30] Y. Ye, A. Gupta, K. Kitani, and S. Tulsiani, "G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis," in *CVPR*, 2024. 6, 32, 35, 37, 115

[31] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11087–11096, 2021. 7, 112

[32] R. Wang, W. Mao, and H. Li, "Interacting hand-object pose estimation via dense mutual attention," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5724–5734, 2022. 7

[33] A. T. Aboukhadra, J. N. Malik, A. Elhayek, N. Robertini, and D. Stricker, "Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1001–1010, 2022. 7, 30, 117

[34] Z. Yu, L. Yang, Y. Xie, P. Chen, and A. Yao, "Uv-based 3d hand-object reconstruction with grasp optimization," in *British Machine Vision Conference*, 2022. 7, 83

[35] J. Hu, H. Zhang, Z. Chen, M. Li, Y. Wang, Y. Liu, and Z. Sun, "Learning explicit contact for implicit reconstruction of hand-held objects from monocular images," *ArXiv*, vol. abs/2305.20089, 2023. 7, 35

[36] H. Cho, C. Kim, J. Kim, S. Lee, E. Ismayilzada, and S. Baek, "Transformer-based unified recognition of two hands manipulating objects," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4769–4778, 2023. 7

[37] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, and C. Lu, "Cpf: Learning a contact potential field to model the hand-object interaction," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11077–11086, 2020. 7, 35, 36, 39, 60, 61, 62, 63, 64, 65, 68, 82, 83, 86

[38] H. T. H. Piaggio, "Calculus of variations," *Nature*, vol. 173, pp. 1156–1157, 1954. 16

[39] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013. 18, 71

[40] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. 18, 115

[41] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017. 18

[42] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. A. Eslami, and Y. W. Teh, "Neural processes," *ArXiv*, vol. abs/1807.01622, 2018. 18, 44, 70, 71, 72, 118

[43] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1530–1538, PMLR, 07–09 Jul 2015. 18

[44] J. N. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *ArXiv*, vol. abs/1503.03585, 2015. 19

[45] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020. 19, 20, 116, 117

[46] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. 19

[47] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, 2023. 19

[48] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *ArXiv*, vol. abs/2204.03458, 2022. 19

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015. 19, 21, 90

[50] S. J. Prince, *Understanding Deep Learning*. MIT Press, 2023. 20, 21

[51] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014. 21

[52] W. Zhou, Z. Dou, Z. Cao, Z. Liao, J. Wang, W. Wang, Y. Liu, T. Komura, W. Wang, and L. Liu, "Emdm: Efficient motion diffusion model for fast and high-quality motion generation," 2024. 21, 118

[53] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "Meshcnn: a network with an edge," *ACM Transactions on Graphics (TOG)*, vol. 38, pp. 1 – 12, 2019. 22

[54] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, "Deep mesh reconstruction from single rgb images via topology modification networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9963–9972, 2019. 22

[55] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," 2024. 22

[56] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5806–5815, 2021. 23

[57] K. Schwarz, A. Sauer, M. Niemeyer, Y. Liao, and A. Geiger, "Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids," *ArXiv*, vol. abs/2206.07695, 2022. 23

[58] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *ArXiv*, vol. abs/2206.00630, 2022. 23

[59] C. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2016. 23, 62, 63, 65, 66, 114

[60] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Neural Information Processing Systems*, 2017. 23, 84

[61] M. A. Uy, Q.-H. Pham, B.-S. Hua, D. T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1588–1597, 2019. 23

[62] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19291–19300, 2021. 23

[63] C. Huang, F. Jiang, Q. Huang, X.-Z. Wang, Z. Han, and W.-Y. Huang, "Dual-graph attention convolution network for 3-d point cloud classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 4813–4825, 2022. 23

[64] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4331–4340, 2019. 23, 25, 33, 34, 36, 66, 67, 68, 87, 111, 112, 114, 116, 118

[65] F. Bernardini, J. Mittleman, H. E. Rushmeier, C. T. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 5, pp. 349–359, 1999. 23

[66] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4455–4465, 2018. 24, 25

[67] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *CVPR*, 2020. 24

[68] Wikipedia, "Signed distance function — Wikipedia, the free encyclopedia." http://en.wikipedia.org/w/index.php?title=Signed%20distance%20function&oldid=1234617806, 2024. [Online; accessed 19-August-2024]. 24

[69] J. Chibane, A. Mir, and G. Pons-Moll, "Neural unsigned distance fields for implicit function learning," *ArXiv*, vol. abs/2010.13938, 2020. 25, 66

[70] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *European Conference on Computer Vision (ECCV)*, 2020. 25

[71] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 25

[72] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 25

[73] V. Sitzmann, E. Chan, R. Tucker, N. Snavely, and G. Wetzstein, "Metasdf: Meta-learning signed distance functions," *ArXiv*, vol. abs/2006.09662, 2020. 25, 44, 118

[74] B. Y. Feng, Y. Zhang, D. Tang, R. Du, and A. Varshney, "Prif: Primary ray-based implicit function," in *European Conference on Computer Vision*, 2022. 25

[75] J. Shue, E. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein, "3d neural field generation using triplane diffusion," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20875–20886, 2022. 25

[76] S. Puhlmann, F. Heinemann, O. Brock, and M. Maertens, "A compact representation of human single-object grasping," in *2016 IEEE International Conference on Intelligent Robots and Systems (IROS)*, p. 1954–1959, IEEE, 2016. 27

[77] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proceedings of International Conference on Computer Vision (ICCV)*, October 2017. 27

[78] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 11807–11816, 2019. 27

[79] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "ContactDB: Analyzing and predicting grasp contact via thermal imaging," in *Computer Vision and Pattern Recognition (CVPR)*, 2019. 27, 33

[80] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "ContactPose: A dataset of grasps with object contact and hand pose," in *European Conference on Computer Vision (ECCV)*, vol. 12358, pp. 361–378, 2020. 27, 33, 61, 85

[81] S. Hampali, S. D. Sarkar, and V. Lepetit, "Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset," *ArXiv*, vol. abs/2107.00887, 2021. 27, 48

[82] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *CVPR*, 2020. 27, 46, 48

[83] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20921–20930, 2022. 27, 40, 95, 106

[84] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox, "DexYCB: A benchmark for capturing hand grasping of objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 27, 30, 40, 46, 47, 48

[85] J. Jian, X. Liu, M. Li, R. Hu, and J. Liu, "Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose," 2023. 27, 40

[86] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 28, 46

[87] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *IEEE International Conference on Computer Vision (ICCV)*, 2019. 28

[88] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2O: Two Hands Manipulating Objects for First Person Interaction Recognition," in *International Conference on Computer Vision (ICCV)*, pp. 10138–10148, 2021. 28

[89] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21013–21022, June 2022. 28

[90] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 28, 40

[91] W. Xie, Z. Yu, Z. Zhao, B. Zuo, and Y. Wang, "Hmdo: Markerless multi-view hand manipulation capture with deformable objects," 2023. 28

[92] T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin, "AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12999–13008, 2023. 28

[93] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. Lu, "Oakink2: A dataset of bimanual hands-object manipulation in complex task completion," 2024. 27, 28

[94] Y. Liu, H. Yang, X. Si, L. Liu, Z. Li, Y. Zhang, Y. Liu, and L. Yi, "Taco: Benchmarking generalizable bimanual tool-action-object understanding," *arXiv preprint arXiv:2401.08399*, 2024. 27, 28

[95] Z. Zhu, J. Wang, Y. Qin, D. Sun, V. Jampani, and X. Wang, "Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation," *2024 International Conference on 3D Vision (3DV)*, pp. 201–212, 2023. 28

[96] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys, "Motion Capture of Hands in Action Using Discriminative Salient Points," in *European Conference on Computer Vision (ECCV)*, pp. 640–653, 2012. 28

[97] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *International Conference on Computer Vision (ICCV)*, 2011. 28

[98] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from RGB-D input," in *European Conference on Computer Vision (ECCV)*, 2016. 28

[99] C. Zimmermann and T. Brox, "Learning to Estimate 3D Hand Pose from Single RGB Images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice), pp. 4913–4921, IEEE, Oct. 2017. 29

[100] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB," *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. Publisher: IEEE. 29

[101] F. Lin, C. Wilhelm, and T. Martinez, "Two-hand global 3d pose estimation using monocular rgb," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2373–2381, January 2021. 29

[102] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, "Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints," *arXiv:2003.09282 [cs]*, Aug. 2020. arXiv: 2003.09282. 29

[103] I. Kourbane and Y. Genc, "A hybrid classification-regression approach for 3D hand pose estimation using graph convolutional networks," *arXiv:2105.10902 [cs]*, May 2021. arXiv: 2105.10902. 29

[104] J.-H. Song and S.-J. Kang, "3D Hand Pose Estimation via Graph-Based Reasoning," *IEEE Access*, vol. 9, pp. 35824–35833, 2021. Conference Name: IEEE Access. 29

[105] W. Zhao, W. Wang, and Y. Tian, "Graformer: Graph-oriented transformer for 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20438–20447, June 2022. 29

[106] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker, "HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation from a Single Depth Map," *arXiv:2004.01588 [cs]*, Apr. 2020. arXiv: 2004.01588. 29, 31

[107] S. Li, H. Wang, and D. Lee, "Hand pose estimation for hand-object interaction cases using augmented autoencoder," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 993–999, 2020. 29

[108] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu, "Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data," *arXiv:2003.09572 [cs]*, Apr. 2020. arXiv: 2003.09572. 29

[109] D. Kulon, R. A. Güler, I. Kokkinos, M. Bronstein, and S. Zafeiriou, "Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild," *arXiv:2004.01946 [cs]*, Apr. 2020. arXiv: 2004.01946. 29

[110] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," *arXiv:2006.10214 [cs]*, June 2020. arXiv: 2006.10214. 29

[111] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: Fast Monocular 3D Hand and Body Motion Capture by Regression and Integration," *arXiv:2008.08324 [cs]*, Aug. 2020. arXiv: 2008.08324. 29

[112] Z. Fan, J. Liu, and Y. Wang, "Adaptive computationally efficient network for monocular 3d hand pose estimation," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, (Berlin, Heidelberg), p. 127–144, Springer-Verlag, 2020. 29

[113] Y. Wu, S. Ma, D. Zhang, and J. Sun, "3D Capsule Hand Pose Estimation Network Based on Structural Relationship Information," *Symmetry*, vol. 12, p. 1636, Oct. 2020. 29

[114] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, (Berlin, Heidelberg), p. 769–786, Springer-Verlag, 2020. 29

[115] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Dual Grid Net: hand mesh vertex regression from single depth maps," *arXiv:1907.10695 [cs]*, July 2019. arXiv: 1907.10695. 29

[116] L. Fang, X. Liu, L. Liu, H. Xu, and W. Kang, "JGR-P2O: Joint Graph Reasoning based

Pixel-to-Offset Prediction Network for 3D Hand Pose Estimation from a Single Depth Image," *arXiv:2007.04646 [cs]*, July 2020. arXiv: 2007.04646. 29

[117] C. Zhu, B. Hu, J. Chen, X. Ai, and S. K. Agrawal, "SARN: Shifted Attention Regression Network for 3D Hand Pose Estimation," *Bioengineering*, vol. 10, p. 126, Jan. 2023. 29

[118] Z. Lin, C. Ding, H. Yao, Z. Kuang, and S. Huang, "Harmonious feature learning for interactive hand-object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12989–12998, June 2023. 29, 30

[119] Z. Yu, S. Huang, C. Fang, T. P. Breckon, and J. Wang, "ACR: Attention Collaboration-based Regressor for Arbitrary Two-Hand Reconstruction," Mar. 2023. arXiv:2303.05938 [cs]. 29, 30

[120] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, Nov. 2017. 29, 31, 33, 39, 62, 63, 68, 82, 114

[121] K. Karunratanakul, A. Spurr, Z. Fan, O. Hilliges, and S. Tang, "A Skeleton-Driven Neural Occupancy Representation for Articulated Hands," Sept. 2021. arXiv:2109.11399 [cs]. 30

[122] V. Wirth, A.-M. Liphardt, B. Coppers, J. Bräunig, S. Heinrich, A. Kleyer, G. Schett, M. Vossiek, B. Egger, and M. Stamminger, "ShaRPy: Shape Reconstruction and Hand Pose Estimation from RGB-D with Uncertainty," Mar. 2023. 0 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv:2303.10042 [cs]. 30

[123] H. Qi, C. Zhao, M. Salzmann, and A. Mathis, "HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields," Feb. 2024. arXiv:2402.17062 [cs]. 30

[124] L. Huang, J. Tan, J. Meng, J. Liu, and J. Yuan, "HOT-Net: Non-Autoregressive Transformer for 3D Hand-Object Pose Estimation," in *Proceedings of the 28th ACM International Conference on Multimedia*, (Seattle WA USA), pp. 3136–3145, ACM, Oct. 2020. 30, 54

[125] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "HandsFormer: Keypoint Transformer for Monocular 3D Pose Estimation ofHands and Object in Interaction," *arXiv:2104.14639 [cs]*, Apr. 2021. arXiv: 2104.14639. 30

[126] H. Cho, C. Kim, J. Kim, S. Lee, E. Ismayilzada, and S. Baek, "Transformer-based unified recognition of two hands manipulating objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4769–4778, June 2023. 30

[127] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing Hands in 3D with Transformers," Dec. 2023. arXiv:2312.05251 [cs]. 30, 31

[128] W. Cheng, H. Tang, L. Van Gool, and J. H. Ko, "HandDiff: 3D Hand Pose Estimation with Diffusion on Image-Point Cloud," Apr. 2024. arXiv:2404.03159 [cs]. 30

[129] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward More Reliable 3D Pose Estimation," Apr. 2023. arXiv:2211.16940 [cs]. 30

[130] A. Armagan, G. Garcia-Hernando, S. Baek, S. Hampali, M. Rad, Z. Zhang, S. Xie, M. Chen, B. Zhang, F. Xiong, Y. Xiao, Z. Cao, J. Yuan, P. Ren, W. Huang, H. Sun, M. Hrúz, J. Kanis, Z. Krňoul, Q. Wan, S. Li, L. Yang, D. Lee, A. Yao, W. Zhou, S. Mei, Y. Liu, A. Spurr, U. Iqbal, P. Molchanov, P. Weinzaepfel, R. Brégier, G. Rogez, V. Lepetit, and T.-K. Kim, "Measuring Generalisation to Unseen Viewpoints, Articulations, Shapes and Objects for 3D Hand Pose Estimation under Hand-Object Interaction," *arXiv:2003.13764 [cs]*, Mar. 2020. arXiv: 2003.13764. 30

[131] Z. Cheng, S. Chen, and Y. Zhang, "Semi-supervised 3d hand-object pose estimation via pose dictionary learning," *ArXiv*, vol. abs/2107.07676, 2021. 30

[132] K. Li, L. Yang, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu, "Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis," *ArXiv*, vol. abs/2109.05488, 2021. 30, 54

[133] G. Garcia-Hernando, E. Johns, and T.-K. Kim, "Physics-Based Dexterous Manipulations with Estimated Hand Poses and Residual Reinforcement Learning," *arXiv:2008.03285 [cs]*, Aug. 2020. arXiv: 2008.03285. 30

[134] A. Boukhayma, R. de Bem, and P. H. S. Torr, "3D Hand Shape and Pose from Images in the Wild," *arXiv:1902.03451 [cs]*, Feb. 2019. arXiv: 1902.03451. 31

[135] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt, "RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video," *ACM Transactions on Graphics*, vol. 39, pp. 1–16, Nov. 2020. 31

[136] P. Chen, Y. Chen, D. Yang, F. Wu, Q. Li, Q. Xia, and Y. Tan, "I2UV-HandNet: Image-to-UV Prediction Network for Accurate and High-fidelity 3D Hand Mesh Modeling," July 2021. arXiv:2102.03725 [cs]. 31, 39

[137] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik, "Reconstructing hand-object interactions in the wild," *ArXiv*, vol. abs/2012.09856, 2020. 31

[138] Y. Chen, Z. Tu, D. Kang, R. Chen, L. Bao, Z. Zhang, and J. Yuan, "Joint Hand-object 3D Reconstruction from a Single Image with Cross-branch Feature Fusion," *arXiv:2006.15561 [cs]*, June 2020. arXiv: 2006.15561. 31

[139] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf," *Communications of the ACM*, vol. 65, pp. 99 – 106, 2020. 31, 32

[140] A. K. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. M. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhofer, "State of the art on neural rendering," *Computer Graphics Forum*, vol. 39, 2020. 31

[141] Y. Ye, A. Gupta, and S. Tulsiani, "What's in your hands? 3D Reconstruction of Generic Objects in Hands," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3885–3895, June 2022. Conference Name: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781665469463 Place: New Orleans, LA, USA Publisher: IEEE. 31, 39

[142] Z. Chen, S. Chen, C. Schmid, and I. Laptev, "gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction," *ArXiv*, vol. abs/2304.11970, 2023. 31, 117

[143] C. Zhang, Y. Di, R. Zhang, G. Zhai, F. Manhardt, F. Tombari, and X. Ji, "DDF-HO: Hand-Held

Object Reconstruction via Conditional Directed Distance Field," Aug. 2023. arXiv:2308.08231 [cs]. 31, 39

[144] J. Hu, H. Zhang, Z. Chen, M. Li, Y. Wang, Y. Liu, and Z. Sun, "Learning Explicit Contact for Implicit Reconstruction of Hand-held Objects from Monocular Images," May 2023. arXiv:2305.20089 [cs]. 31, 39

[145] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging Photometric Consistency Over Time for Sparsely Supervised Hand-Object Reconstruction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 568–577, June 2020. ISSN: 2575-7075. 32

[146] D. Huang, X. Ji, X. He, J. Sun, T. He, Q. Shuai, W. Ouyang, and X. Zhou, "Reconstructing Hand-Held Objects from Monocular Video," Nov. 2022. arXiv:2211.16835 [cs]. 32

[147] H. Choi, N. Chavan-Dafle, J. Yuan, V. Isler, and H. Park, "HandNeRF: Learning to Reconstruct Hand-Object Interaction Scene from a Single RGB Image," Sept. 2023. arXiv:2309.07891 [cs]. 32

[148] Z. Fan, M. Parelli, M. E. Kadoglou, M. Kocabas, X. Chen, M. J. Black, and O. Hilliges, "HOLD: Category-agnostic 3D Reconstruction of Interacting Hands and Objects from Video," Nov. 2023. arXiv:2311.18448 [cs]. 32

[149] Z. Leng, S.-C. Wu, M. Saleh, A. Montanaro, H. Yu, Y. Wang, N. Navab, X. Liang, and F. Tombari, "Dynamic Hyperbolic Attention Network for Fine Hand-object Reconstruction," Sept. 2023. arXiv:2309.02965 [cs]. 32

[150] A. T. Aboukhadra, J. Malik, A. Elhayek, N. Robertini, and D. Stricker, "THOR-Net: End-to-end Graformer-based Realistic Two Hands and Object Reconstruction with Self-supervision," Oct. 2022. arXiv:2210.13853 [cs]. 32

[151] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani, "Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips," Sept. 2023. arXiv:2309.05663 [cs]. 32, 39

[152] N. S. Pollard and V. B. Zordan, "Physically based grasping control from example," in *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 311–318, 2005. 33, 82

[153] S. El-Khoury, A. Sahbani, and P. Bidaud, "3D objects grasps synthesis: A survey," in *IFToMM World Congress on Mechanism and Machine Science*, 2011. 33

[154] P. G. Kry and D. K. Pai, "Interaction capture and synthesis," *Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 872–880, 2006. 33

[155] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on Robotics*, vol. 30, pp. 289–309, 2014. 33

[156] Y. Li, J. L. Fu, and N. S. Pollard, "Data-driven grasp synthesis using shape matching and task-based pruning," *Transactions on Visualization and Computer Graphics (TVCG)*, vol. 13, no. 4, pp. 732–747, 2007. 33

[157] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez, "GanHand: Predicting human grasp affordances in multi-object scenes," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 5030–5040, 2020. 33

[158] K. Karunratanakul, A. Spurr, Z. Fan, O. Hilliges, and S. Tang, "A skeleton-driven neural occupancy representation for articulated hands," in *2021 International Conference on 3D Vision (3DV)*, pp. 11–21, 2021. 33, 39

[159] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang, "Grasping Field: Learning implicit representations for human grasps," in *International Conference on 3D Vision (3DV)*, pp. 333–344, 2020. 33

[160] T. Zhu, R. Wu, X. Lin, and Y. Sun, "Toward human-like grasp: Dexterous grasping via semantic representation of object-hand," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15721–15731, 2021. 33

[161] T. H. E. Tse, Z. Zhang, K. I. Kim, A. Leonardis, F. Zheng, and H. J. Chang, "S2contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning," *ArXiv*, vol. abs/2208.00874, 2022. 33, 34, 35, 74, 76, 83, 85, 86, 96, 97, 98, 106, 112

[162] S. Liu, Y. Zhou, J. Yang, S. Gupta, and S. Wang, "Contactgen: Generative contact modeling for grasp generation," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20552–20563, 2023. 33, 35, 39

[163] H. Li, X. Lin, Y. Zhou, X. Li, Y. Huo, J. Chen, and Q. Ye, "Contact2grasp: 3d grasp synthesis via hand-object contact constraint," in *International Joint Conference on Artificial Intelligence*, 2022. 34, 35, 39

[164] K. C. Liu, "Dextrous manipulation from a grasping pose," *Transactions on Graphics (TOG)*, vol. 28, no. 3, p. 59, 2009. 34

[165] I. Mordatch, Z. Popovic, and E. Todorov, "Contact-invariant optimization for hand manipulation," in *Symposium on Computer Animation (SCA)*, pp. 137–144, 2012. 34

[166] W. Zhao, J. Zhang, J. Min, and J. Chai, "Robust realtime physics-based motion control for human grasping," *Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 207:1–207:12, 2013. 34

[167] Y. Ye and C. K. Liu, "Synthesis of detailed hand manipulations using contact sampling," *Transactions on Graphics (TOG)*, vol. 31, no. 4, p. 41:1–41:10, 2012. 34

[168] Y. Ye and K. C. Liu, "Synthesis of detailed hand manipulations using contact sampling," *Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 41:1–41:10, 2012. 34

[169] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "GOAL: Generating 4D whole-body motion for hand-object grasping," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 34

[170] M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. W. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," *Int. J. Robotics Res.*, vol. 39, no. 1, 2020. 34

[171] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Proceedings of Robotics: Science and Systems*, (Pittsburgh, Pennsylvania), June 2018. 34

[172] K. Bergamin, S. Clavet, D. Holden, and J. R. Forbes, "Drecon: Data-driven responsive control of physics-based characters," *ACM Trans. Graph.*, vol. 38, nov 2019. 34

[173] S. Park, H. Ryu, S. Lee, S. Lee, and J. Lee, "Learning predict-and-simulate policies from unorganized human motion data," *ACM Trans. Graph.*, vol. 38, nov 2019. 34

[174] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "DeepMimic: Example-guided deep reinforcement learning of physics-based character skills," *Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 143:1–143:14, 2018. 34

[175] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, "Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning," *ACM Trans. Graph.*, vol. 36, jul 2017. 34

[176] S. Christen, M. Kocabas, E. Aksan, J. Hwangbo, J. Song, and O. Hilliges, "D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions," in *Computer Vision and Pattern Recognition (CVPR)*, 2022. 34

[177] J. Zheng, Q. Zheng, L. Fang, Y. Liu, and L. Yi, "Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis," 2023. 34, 36, 39, 113, 117, 118

[178] O. Taheri, Y. Zhou, D. Tzionas, Y. Zhou, D. Ceylan, S. Pirk, and M. J. Black, "Grip: Generating interaction poses using spatial cues and latent consistency," in *2024 International Conference on 3D Vision (3DV)*, pp. 933–943, 2024. 34, 35, 36, 39, 113, 114, 117, 118

[179] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll, "Gears: Local geometry-aware hand-object interaction synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20634–20643, June 2024. 34

[180] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-moll, "Gears: Local geometry-aware hand-object interaction synthesis," 2024. 34, 36, 39, 113, 114, 118

[181] C. Diller and A. Dai, "Cg-hoi: Contact-guided 3d human-object interaction generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19888–19901, June 2024. 35

[182] L. Gui, Y.-X. Wang, D. Ramanan, and J. M. F. Moura, "Few-shot human motion prediction via meta-learning," in *ECCV*, 2018. 37

[183] E. Park and A. C. Berg, "Meta-tracker: Fast and robust online adaptation for visual object trackers," in *ECCV*, 2018. 37

[184] H. Xiao, B. Kang, Y. Liu, M. Zhang, and J. Feng, "Online meta adaptation for fast video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 1205–1217, 2020. 37, 46

[185] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9367–9376, 2019. 37, 46, 47

[186] Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *ICML*, 2020. 37

[187] M. Hao, Y. Li, Z. Di, N. B. Gundavarapu, and X. Wang, "Test-time personalization with a transformer for human pose estimation," *ArXiv*, vol. abs/2107.02133, 2021. 37, 46

[188] X. Li, S. Liu, S. D. Mello, K. Kim, X. Wang, M.-H. Yang, and J. Kautz, "Online adaptation for consistent mesh reconstruction in the wild," *ArXiv*, vol. abs/2012.03196, 2020. 37, 46

[189] M. Michalkiewicz, S. Parisot, S. Tsogkas, M. Baktash, A. P. Eriksson, and E. Belilovsky, "Few-shot single-view 3-d object reconstruction with compositional priors," in *ECCV*, 2020. 37

[190] S. Yang, M. Xu, and H. Yao, "Single-view 3d object reconstruction from shape priors in memory," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3160, 2021. 37

[191] B. Wallace and B. Hariharan, "Few-shot generalization for single-image 3d reconstruction via priors," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3817–3826, 2019. 37

[192] Y. Hasson, G. Varol, I. Laptev, and C. Schmid, "Towards unconstrained joint hand-object reconstruction from RGB videos," *arXiv:2108.07044 [cs]*, Aug. 2021. arXiv: 2108.07044. 39

[193] J. Chen, M. Yan, J. Zhang, Y. Xu, X. Li, Y. Weng, L. Yi, S. Song, and H. Wang, "Tracking and reconstructing hand object interactions from point cloud sequences in the wild," 2022. 39

[194] S. Ravi, P. Climent-Perez, T. Morales, C. Huesca-Spairani, K. Hashemifard, and F. Flórez-Revuelta, "Odin: An omnidirectional indoor dataset capturing activities of daily living from multiple synchronized modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6487–6496, June 2023. 40

[195] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. Cambridge University Press, 2023. https://D2L.ai. 43

[196] M. Juarez, S. Yeom, and M. Fredrikson, "Black-box audits for group distribution shifts," 2022. 43

[197] J. Liang, R. He, and T.-P. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *ArXiv*, vol. abs/2303.15361, 2023. 43

[198] M. Andrychowicz, M. Denil, S. G. Colmenarejo, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. de Freitas, "Learning to learn by gradient descent by gradient descent," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 3988–3996, 2016. 43, 45, 46

[199] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *arXiv:1703.03400 [cs]*, July 2017. arXiv: 1703.03400. 43, 44

[200] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," in *International Conference on Learning Representations*, 2018. 44, 46, 51

[201] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse? towards understanding the effectiveness of maml," in *International Conference on Learning Representations*, 2019. 44, 51

[202] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *ArXiv*, vol. abs/1807.05960, 2019. 44

[203] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv: Learning*, 2018. 44

[204] A. Santoro, S. Bartunov, M. M. Botvinick, D. Wierstra, and T. P. Lillicrap, "One-shot learning with memory-augmented neural networks," *ArXiv*, vol. abs/1605.06065, 2016. 44

[205] A. Santoro, S. Bartunov, M. M. Botvinick, D. Wierstra, and T. P. Lillicrap, "Meta-learning with memory-augmented neural networks," in *ICML*, 2016. 44

[206] M. Garnelo, D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. J. Rezende, and S. M. A. Eslami, "Conditional neural processes," in *ICML*, 2018. 44

[207] H. Kim, A. Mnih, J. Schwarz, M. Garnelo, S. M. A. Eslami, D. Rosenbaum, O. Vinyals, and Y. W. Teh, "Attentive neural processes," *ArXiv*, vol. abs/1901.05761, 2019. 44, 72

[208] H. Edwards and A. J. Storkey, "Towards a neural statistician," *ArXiv*, vol. abs/1606.02185, 2017. 44

[209] P. Bachman, R. Islam, A. Sordoni, and Z. Ahmed, "Vfunc: a deep generative model for functions," *ArXiv*, vol. abs/1807.04106, 2018. 44

[210] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021. 44

[211] M. Huisman, J. N. van Rijn, and A. Plaat, "A survey of deep meta-learning," *Artif. Intell. Rev.*, vol. 54, pp. 4483–4541, 2021. 44, 55

[212] L. Huang, B. Zhang, Z. Guo, Y. Xiao, Z. Cao, and J. Yuan, "Survey on depth and rgb image-based 3d hand shape and pose estimation," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 3, pp. 207–234, 2021. 46, 48

[213] S. Thrun and L. Y. Pratt, "Learning to learn," *arXiv: Learning*, pp. 354–354, 1998. 46

[214] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2016. 46

[215] Dawson-Haggerty et al., "trimesh," 2019. 48

[216] J. Rajendran, A. Irpan, and E. Jang, "Meta-Learning Requires Meta-Augmentation," *arXiv:2007.05549 [cs, stat]*, Nov. 2020. arXiv: 2007.05549. 50

[217] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn, "Meta-learning without memorization," *ArXiv*, vol. abs/1912.03820, 2020. 51, 52, 56

[218] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 51, 52, 62

[219] S. M. R. Arnold, P. Mahajan, D. Datta, I. Bunner, and K. S. Zarkias, "learn2learn: A library for Meta-Learning research," *ArXiv*, Aug. 2020. 52

[220] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015. 52

[221] L. Biewald, "Experiment tracking with weights and biases," 2020. Software available from wandb.com. 53

[222] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, p. 861, 2018. 56, 59

[223] M. J. Anderson, *Permutational Multivariate Analysis of Variance (PERMANOVA)*, pp. 1–15. John Wiley & Sons, Ltd, 2017. 55

[224] R. Miller, C. Acton, D. A. Fullerton, J. Maltby, and J. Campling, "Analysis of variance (anova)," *The SAGE Encyclopedia of Research Design*, 2022. 55

[225] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. 56

[226] I. L. Dryden and K. V. Mardia, "Statistical shape analysis," in *Statistical Shape Analysis*, 1998. 56

[227] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987. 62

[228] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 5745–5753, 2019. 65

[229] G. V. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989. 65

[230] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 2004. 65

[231] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Neural Information Processing Systems*, 2017. 65

[232] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015. 66, 91

[233] A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei, "Avoiding latent variable collapse with generative skip models," *ArXiv*, vol. abs/1807.04863, 2018. 71

[234] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017. 72

[235] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. 73, 93

[236] T. Morales and G. Lacey, "A new benchmark for group distribution shifts in hand grasp regression for object manipulation. can meta-learning raise the bar?," in *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 77

[237] A. T. Miller and P. K. Allen, "Graspit! A versatile simulator for robotic grasping," *IEEE Robotics Automation Magazine (RAM)*, vol. 11, no. 4, pp. 110–122, 2004. 82

[238] K. Ehsani, S. Tulsiani, S. Gupta, A. Farhadi, and A. K. Gupta, "Use the force, luke! learning to predict physical forces by simulating effects," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 221–230, 2020. 82

[239] B. Smith, C. Wu, H. Wen, P. Peluse, Y. Sheikh, J. K. Hodgins, and T. Shiratori, "Constraining dense hand surface tracking with elasticity," *ACM Transactions on Graphics (TOG)*, vol. 39, pp. 1 – 14, 2020. 82

[240] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017. 92, 117

[241] T. Alldieck, N. Kolotouros, and C. Sminchisescu, "Score distillation sampling with learned manifold corrective," in *European Conference on Computer Vision (ECCV)*, 2024. 115

[242] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *CVPR*, 2024. 117

[243] C. Lu, Y. Zhou, F. Bao, J. Chen, C. LI, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 5775–5787, Curran Associates, Inc., 2022. 117

[244] B. Zheng and T. Yang, "Diffusion models are innate one-step generators," *ArXiv*, vol. abs/2405.20750, 2024. 117

[245] X. Liu, C. Gong, and qiang liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *The Eleventh International Conference on Learning Representations*, 2023. 117

[246] X. Liu, X. Zhang, J. Ma, J. Peng, and Q. Liu, "Instaflow: One step is enough for high-quality diffusion-based text-to-image generation," in *International Conference on Learning Representations*, 2024. 117

[247] K. Li, L. Yang, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu, "ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis," *arXiv:2109.05488 [cs]*, Sept. 2021. arXiv: 2109.05488. 117

[248] A. Ziani, Z. Fan, M. Kocabas, S. Christen, and O. Hilliges, "Tempclr: Reconstructing hands via time-coherent contrastive learning," in *International Conference on 3D Vision (3DV)*, 2022. 117, 119

[249] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "HuMoR: 3D human motion model for robust pose estimation," in *International Conference on Computer Vision (ICCV)*, pp. 11488–11499, 2021. 117, 118

[250] W. Wan, Z. Dou, T. Komura, W. Wang, D. Jayaraman, and L. Liu, "Tlcontrol: Trajectory and language control for human motion synthesis," *arXiv preprint arXiv:2311.17135*, 2023. 118

[251] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Spectral, probabilistic, and deep metric learning: Tutorial and survey," 2022. 119

[252] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020. 119

[253] Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. LeCun, "On the duality between contrastive and non-contrastive self-supervised learning," in *The Eleventh International Conference on Learning Representations*, 2023. 119