3D POSE AND SHAPE ESTIMATION OF HANDS AND MANIPULATED OBJECTS FROM IMAGES AND VIDEOS

By

TZE HO ELDEN TSE

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Human-Centred Visual Learning Group
School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
April 2024

To my beloved family.

ABSTRACT

3D shape and pose estimation of hands and manipulated object is an important and long-standing problem in computer vision. This problem can be particularly challenging due to extreme variations in object shape and texture. In addition, heavy occlusions can be introduced by other objects in the scene or humans during interaction. Nevertheless, modeling hand-object manipulations is essential for understanding how humans interact with the physical world.

There are many challenges in modeling hand-object interactions. In this thesis, we focus on three outstanding challenges which unifies geometry-driven and data-driven methods: (1) estimating 3D pose and shape from 2D images is an extremely ill-posed problem due to the loss of depth information during 3D projection to 2D, (2) inexpressive and physically implausible 3D hand reconstructions, 3) inability to recognise seen actions on unseen objects.

In this thesis, we propose three main contributions to overcome these challenges. First, we present a new collaborative learning strategy in which two branches of deep neural network mutually exchange information for 3D hand-object reconstruction from single RGB image. Second, we present a new Transformer-based method that estimates the absolute root pose and shape of two-hands with extended forearm at high resolution from egocentric RGB images. Third, we present a new method for compositional action recognition by leveraging 3D geometric information from egocentric RGB videos. Specifically, we exploit superquadrics for both template-free object reconstruction and interaction recognition.

This thesis pushes the state-the-art for understanding hand and object from RGB images

and videos. First, we show that a collaborative learning framework which allows sharing of 3D geometric information across two branches of networks iteratively can tackle the problem of mutual occlusions. Through this novel network architecture design, we achieve state-of-the-art performance on several common public benchmarks. Second, we present the first method that reconstruct high fidelity two-hand meshes with extended forearms from multi-view RGB images. We demonstrate that by leveraging the properties of graph Laplacian from spectral graph theory can effectively aggregate multi-view features as well as producing smooth meshes. Third, we explore superquadrics as an alternative 3D object representation to bounding boxes and demonstrate that it is beneficial to recognising seen actions on unseen objects.

ACKNOWLEDGMENTS

I would like to thank everyone who has supported me in my PhD journey. First and foremost, I am most grateful to my supervisors Hyung Jin Chang and Aleš Leonardis for their generous advice, encouragement, faith and support in me to become a computer vision researcher. I am also immensely grateful to Kwang In Kim for his passion in pursuing challenging problems and rigor in technical details. All of my achievements during my PhD would not have been possible without their constant source of inspiration, guidance and support.

I thank all my labmates at Birmingham, especially Linfang Zheng, Zhongqun Zhang, Hengfei Wang, Jonathan Freer, Esha Dasgupta and Yuqi Hou, with whom I had many discussions spanning both research-related and unrelated academic pursuits.

I am also fortunate to have the opportunity to experience an internship at Google. I thank Danhang Tang for introducing me the hands team and Bardia Doosti for being my intern host. I also thank my collaborators, Franziska Mueller, Thabo Beeler, Zhengyang Shen, Mingsong Dou, Yinda Zhang, Sasa Petrovic and Jonathan Taylor, for sharing their valuable insights, technical support and encouragement.

I am deeply grateful to my family for their constant support and love. I thank my parents and sister for teaching me the value of hard work and perseverance. Their unwavering dedication has shaped my character and fueled my aspirations. Lastly, I would like to thank my wife, Idow. Her boundless love, constant support and encouraging words have been my guiding light throughout this PhD journey. I am forever indebted to my family.

Publications

This thesis includes work published in the following:

- **Tze Ho Elden Tse**, Kwang In Kim, Aleš Leonardis, Hyung Jin Chang. (2022, June). **Collaborative Learning for Hand and Object Reconstruction with Attention-guided Graph Convolution.** In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1664-1674).
- **Tze Ho Elden Tse**, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, Bardia Doosti. (2023, October). **Spectral Graphormer: Spectral Graph-based Transformer for Egocentric Two-Hand Reconstruction using Multi-View Color Images.** In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 14666-14677).
- **Tze Ho Elden Tse**, Runyang Feng, Jiho Park, Linfang Zheng, Yixing Gao, Jihie Kim, Aleš Leonardis, Hyung Jin Chang. (2023, November). **Collaborative Learning for 3D Hand-Object Reconstruction and Compositional Action Recognition from Egocentric RGB Videos using Superquadrics.** *Under review.*

Other publications from work not included in this thesis:

- Linfang Zheng, Aleš Leonardis, **Tze Ho Elden Tse**, Nora Horanyi, Hua Chen, Wei Zhang, Hyung Jin Chang. (2022, May). **TP-AE: Temporally Primed 6D Object Pose Tracking with Auto-Encoders.** In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (pp. 10616-10623).

– **Tze Ho Elden Tse***, Zhongqun Zhang*, Kwang In Kim, Aleš Leonardis, Feng Zheng, Hyung Jin Chang. (2022, October). **S^{2}Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning.** In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 568-584). *The first two authors contributed equally.

– Runyang Feng, Yixing Gao, Xueqing Ma, **Tze Ho Elden Tse**, Hyung Jin Chang. (2023, June). **Mutual Information-Based Temporal Difference Learning for Human Pose Estimation in Video.** In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 17131-17141).

– Runyang Feng, Yixing Gao, **Tze Ho Elden Tse**, Xueqing Ma, Hyung Jin Chang. (2023, October). **DiffPose: SpatioTemporal Diffusion Model for Video-based Human Pose Estimation.** In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 14861-14872).

– Linfang Zheng, **Tze Ho Elden Tse**, Chen Wang, Yinghan Sun, Hua Chen, Wei Zhang, Aleš Leonardis, Hyung Jin Chang. (2024, June). **GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pose Refinement.** In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Contents

List of Figures

List of Tables

Acronyms

API Application Programming Interface. 3

AR Augmented Reality. 1, 3, 8, 9, 55, 56, 59, 67

CAD Computer Assisted Design. 17, 107

CNN Convolutional Neural Network. 11, 12, 17, 23, 60–62, 70, 76, 88

DoF Degrees of Freedom. 5, 35

HDR High-Dynamic-Range. 69

MANO hand Model with Articulated and Non-rigid defOrmations. 10–12, 19, 24, 35, 55, 76, 91

MLP Multi Layer Perceptron. 17, 71, 91, 94

MPVE Mean-Per-Vertex-Error. 71, 96, 97, 100, 102

PCA Principal Component Analysis. 11, 35, 71

PnP Perspective-n-Point. 15

RGB Red Green Blue. 2, 7, 8, 12, 17–19, 22, 24, 25, 29–32, 34, 42, 56–58, 60, 83, 85–88, 90, 105, 106

RNN Recurrent Neural Network. 27

SDF Signed Distance Function. 14, 24

SMPL Skinned Multi-Person Linear. 24, 32

SVD Singular Value Decomposition. 15

VR Virtual Reality. 1, 3, 8, 9, 55, 56, 67

Chapter One

INTRODUCTION

1.1 Problem statement

Understanding human hand-object interaction is a long-standing challenge in computer vision. It delves into how humans engage with the physical world where object manipulation plays a central role. As shown in Figure 1.1, hands serve as the primary tools for daily interactions with a variety of objects in an environment to complete a wide range of tasks. Therefore, understanding hand-object interaction represents a crucial step towards general understanding of humans in unconstrained environments which finds numerous of applications such as Augmented Reality (AR), Virtual Reality (VR), and robotics.

In the current literature, the dominant way to approach this problem is to classify the action labels, such as grasping, lifting or taking, based on visual inputs (Jhuang et al., 2013; Carreira et al., 2017; Varol et al., 2017; Kantorov et al., 2014). However, despite achieving high accuracy, these models do not capture the rich and complex dynamics of how the action is performed, such as the hand pose, the object shape, the contact areas, and the applied forces. Moreover, in order to transfer the knowledge from vision to robotics, it is essential to provide semantic or geometric information about the scene, which is usually

Figure 1.1: Examples of everday's hand-object interactions from *Epic-Kitchens* (Damen et al., 2018). Object manipulation plays a crucial role in human everyday life. (Figure taken from the *Epic-Kitchens* dataset (Damen et al., 2018).)

obtained by extracting 2D segmentation maps and some additional information at the pixel level. However, these intermediate representations do not provide enough information to reason about contacts and forces. Therefore, a complete 3D understanding of hand-object interaction is essential for advancing the field and enabling various applications.

In this thesis, we make contributions towards unified recognition of two hands manipulating objects. Specifically, we study the following two tasks within hand-object interaction: **3D pose and shape estimation** and **action recognition**. Our hope is to develop methods that accurately estimate dense 3D shape and pose of hand and manipulated objects [1], as well as recognise their interactions. We consider three types of problem scenarios along with this direction in this thesis:

1. 3D pose and shape estimations from single RGB image: We estimate the 3D pose and shape of single hand and manipulated object during interaction from single RGB image.

[1]These refer to physical items that are interacted with or altered by a person or an intelligent system. In the context of computer vision and robotics, manipulated objects can include everyday items such as tools, utensils, toys, or any other objects that are handled, moved, or transformed during a task or activity.

2. 3D pose and shape estimations from egocentric view: We estimate the absolute 3D pose and shape of two-hand from egocentric multi-view images to simulate AR/VR problem settings. This motivated by the fact human usually interacts with object using both hands.

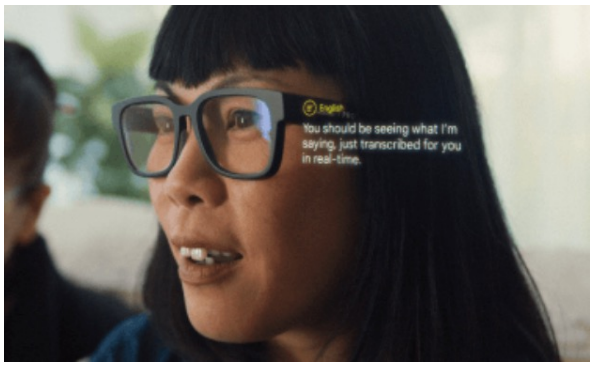
3. 3D pose and shape estimations and action recognition from egocentric videos: We jointly estimate the 3D pose and shape of hand-object and recognise interaction from egocentric videos.

1.2 Motivation

In the following, we detail applications of hand-object modeling in AR, robotics and surgical contexts.

AR. Estimating the hand and object pose accurately is useful for AR, where the view of the user is enhanced by computer-generated information. As illustrated in Figure 1.2, consumer head-mounted displays like Google Glasses or Microsoft HoloLens have a dedicated Computer Vision Application Programming Interface (API) that can assist users in performing specialised tasks that require dexterous object manipulation. These wearable devices are particularly suitable for training hand-related tasks, such as surgery or machine manipulation. Moreover, reconstructing the manipulation sequence in 3D space and time can help monitoring the interventions and detecting errors.

Robotics. This field aims to create machines capable of performing tasks that humans can do, such as manipulating objects, navigating environments, and collaborating with other agents. To achieve this goal, robotics needs to understand how humans interact with objects using their hands, which is the main tool for object manipulation. By understanding hand-



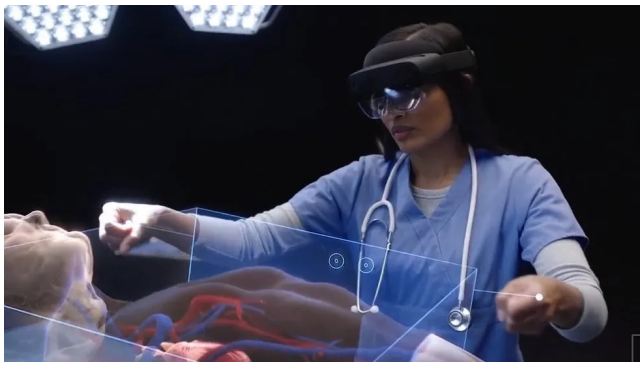


Figure 1.2: Example applications to AR devices. Left: Google Glasses provides new communication experiences by translation and transcription (Figure taken from Google VR (2024)). Right: Microsoft HoloLens offers new ways for medical surgeons to plan operations (Figure taken from Kelion (2019)).

object interaction, robotics can learn from human behavior and intention, and improve its own performance and adaptability. For example, understanding hand-object interaction can help robotics to design better grippers, plan optimal grasps, generate natural motions, and coordinate with human partners. Moreover, understanding hand-object interaction can enable robotics to transfer the knowledge from vision to action, and to simulate and evaluate different scenarios of object manipulation. Therefore, understanding hand-object interaction is essential for advancing the field and application of robotics.

Surgical contexts. Understanding hand-object interaction in surgical settings is important for several critical reasons. First, it directly impacts precision and safety during procedures as surgeons rely on their hands and instruments to execute intricate maneuvers and accurate interaction ensures optimal outcomes. Second, the skill level of a surgeon can be observed by analysing hand-object interaction and hence be improved by tailored training and identifying areas of improvement. Third, as robotic systems become more prevalent in surgery, comprehending how human operators interact with these tools is essential for seamless communication between surgeons and robots. Lastly, hand-object pose estimation aids surgical navigation systems, enhancing planning and execution.

1.3 Challenges

In this section, we discuss several the outstanding challenges that arise during 3D pose and shape estimation of hand-object, and also during action recognition.

High degrees of freedom in hands. Hand pose estimation involves optimising 51 Degrees of Freedom (DoF) (Q. Ye et al., 2016). However, due to strong correlations among the joint angles of the fingers, the effective DoF is significantly lower than 51 in practical scenarios (Pavlakos et al., 2019). Despite this, when directly predicting hand pose parameters from an RGB image using a deep neural network, these inherent correlations may not be properly captured, leading to unrealistic and implausible hand poses.

Occlusions. As hands are highly articulated, it is common that several parts of the hand tend to be self-occluded. While multi-view settings can alleviate this problem, it remains challenging in single view setting as the pose of occluded fingers are main source of ambiguity. This problem is further compounded during hand-object interactions, where accurate grasp poses are difficult to estimate. In addition, the problem of occlusions becomes more significant when under egocentric perspective, as this viewpoint frequently exhibits large degree of erratic camera motion.

Variations in shape. Human hands can exhibit large shape variations. Previous research (Garcia-Hernando et al., 2018; Baek et al., 2018; Ge et al., 2018) mostly focuses on evaluating the 3D joint locations which can be described as a function of hand shape and joint angles. Consequently, despite accurate joint angle estimation, the resulting hand pose may not be reliable if the hand shape is inaccurately estimated. Similarly, everyday objects also exhibit a wide diversity of sizes and shapes, adding further complexity to the problem.

Physically implausible shape recovery. Early works (Mueller et al., 2018; Simon et al.,

2017; Spurr et al., 2018; Zimmermann et al., 2017) primarily focused on sparse keypoint estimation which has limited reasoning ability about hand-object interactions. While dense 3D hand meshes can be estimated from images by fitting a hand mesh to detected joints or through tracking with a good initialisation (La Gorce et al., 2011), recent advancements have explored end-to-end learnable models for capturing the 3D shape or surface of a hand using depth input (Malik et al., 2018; Malik et al., 2020). However, existing methods often neglect the critical constraints that govern object interactions in the physical world. Specifically, they fail to account for physical prior knowledge that objects cannot interpenetrate each other and that, during grasping, contacts occur at the surface between the object and the hand.

Generalisation to unseen objects on seen action. While deep architectures trained on large scale datasets (Sigurdsson et al., 2016; Kay et al., 2017; Karpathy et al., 2014) exhibit strong distribution learning capabilities, mainstream action recognition models (Simonyan et al., 2014; Carreira et al., 2017; Feichtenhofer et al., 2019; L. Wang et al., 2016) primarily focus on frame appearance rather than temporal reasoning. Consequently, reversing the order of the video frame at test time will often produce the same classification result as shown in (Materzynska et al., 2020; B. Zhou et al., 2018). In particular, classical activity recognition methods like the two-stream Convolutional Neural Network (Simonyan et al., 2014) and I3D (Carreira et al., 2017) have demonstrated strong performance on various video datasets, including UCF101 (Soomro et al., 2012) and Sport1M (Karpathy et al., 2014), with only still frames and optical flow. While appearance features can be highly predictive of the action class (Santoro et al., 2017; Battaglia et al., 2018), it remains challenging for appearance-based deep networks to capture the compositionality of action and objects without temporal transformations or geometric relations (Materzynska et al., 2020).

1.4 Thesis outline and contributions

In this thesis, we propose to model hand-object interactions from colour images in both first-person and third-person perspectives. Our main objective is to jointly reconstruct the 3D geometry of hands and manipulated objects, and understand their interactions from colour images. We aim to accurately estimate the dense surfaces of the hand and object, while also capturing precise contact locations at the hand-object surfaces interface and recovering physically plausible 3D shape reconstructions.

In Chapter 3, we present an end-to-end trainable collaborative learning method for hand-object reconstruction from a single RGB frame. We design an attention-guided graph convolution to capture mesh information dynamically and tackle the problem of mutual occlusions. We introduce an unsupervised training strategy for effective feature transfer between hand-object branches and stabilise training. We achieve state-of-the-art accuracy in 3D hand-object reconstruction benchmarks at the time of publication. We also demonstrate that our model achieves highly physically plausible results without contact terms.

In Chapter 4, we combine ideas from the spectral graph theory and Transformers and present a spectral graph-based Transformer architecture that reconstructs two high fidelity from egocentric multi-view RGB images. We consider a more challenging problem setting where we directly regress the absolute root poses of two-hands with extended forearm at high resolution from egocentric view. As existing datasets are either infeasible for egocentric viewpoints or lack background variations, we create a large-scale synthetic dataset with diverse scenarios and collect a real dataset from multi-calibrated camera setup to verify our proposed multi-view image feature fusion strategy. We design an efficient soft attention-based multi-view image feature fusion in which the resulting image features are region-specific to segmented hand mesh. To make the reconstruction physically plausible, we propose two strategies: (i) a coarse-to-fine spectral graph convolution decoder to smoothen the meshes

during upsampling and (ii) an optimisation-based refinement stage at inference to prevent self-penetrations. We show that our framework is able to produce realistic two-hand reconstructions and demonstrate the generalisation of synthetic-trained models to real data, as well as real-time AR/VR applications.

In Chapter 5, we propose to leverage 3D geometric information for reasoning compositional action recognition from egocentric RGB videos. We show that using superquadrics as the intermediate 3D object representation is beneficial for 3D hand pose estimation and interaction recognition. We are the first to exploit superquadrics for both template-free object reconstruction and interaction recognition. We extend two egocentric hand-object datasets by introducing new compositional splits and investigate compositional action recognition where a subset of action verb and noun combinations do not exist during training. We achieve state-of-the-art performance on two public benchmarks in both official and our compositional settings.

In the remaining chapters, we conclude this thesis by summarising the contributions of our work, discussing its limitations and open problems, and outlining possible opportunities for future work.

Chapter Two

RELATED WORK

This chapter provides a survey of previous work related to modelling hand-object interactions. We first review the literature on independent hand and object modelling in Section 2.1 and 2.2. Then, we focus on the methods which model hands and objects simultaneously in Section 2.3 and 2.4, respectively.

2.1 Hand

Research in hand pose estimation has a rich history due to its wide range of applications, *i.e.* human-computer interaction, AR/VR devices and activity recognition. In the following, we first review the literature of hand modelling by categorising previous work into *generative* and *discriminative* methods in Section 2.1.1. Then, we focus on 3D hand pose and shape estimation from colour images in Section 2.1.2 as they are most similar in spirit to the approaches proposed in this thesis.

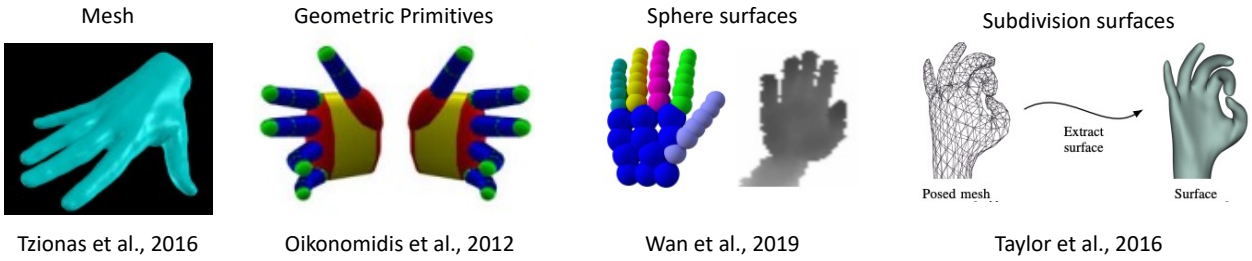


Figure 2.1: Different hand representations used in the literature. *From left to right:* modelling hand with mesh (Tzionas et al., 2016), a collection of geometric primitives (Oikonomidis et al., 2012), sphere surfaces (Wan et al., 2019) and subdivision surfaces (Taylor et al., 2016).

2.1.1 Hand modelling

Generative approaches. This class of methods assumes the availability of a generative hand model, *i.e.* 3D meshes, collections of geometric primitives and implicit representations (see Figure 2.1). These hand models are usually personalised as they are captured manually. Generative approaches typically recover hand pose estimate by optimising these explicit hand model with a set of pre-defined constraints. They have the advantages of modelling image evidence with statistical or physical likelihood of hand poses. However, they are prone to inaccurate initialisation and rely on multiple optimisation constraints to minimise uncertainty. In the following, we briefly introduce several more recent works that estimate a detailed hand shape automatically. D. J. Tan et al. (2016) present a practical method for personalising a hand shape to an individual user using depth images. Tkach et al. (2017) propose an online optimisation algorithm that jointly estimates pose and shape of the hand in each frame for real-time hand tracking. Remelli et al. (2017) present a sphere mesh tracking model for personalising user from a collection of depth measurements. Romero et al. (2017) introduce a parametric model of hand shape and pose called MANO. MANO is learned from more than 1000 high-resolution 3D scans of hands from 31 subjects across a diverse hand poses. It provides a compact mapping from hand poses to pose blend shape corrections and can be used for generative model fitting. We illustrate in Figure 2.2 the variations of hand

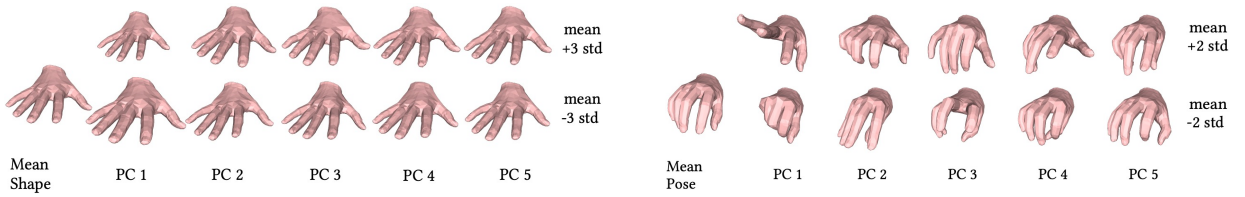


Figure 2.2: Qualitative examples of the MANO parametric hand model (Romero et al., 2017) PCA shape and pose space. As indicated, we visualise the mean shape and pose space by varying the effect of the first 5 principal components (PC). We also show the effect of each PC by adding variations on standard deviations (std).

reconstructions in the PCA shape and pose space.

Discriminative approaches. In contrast, this class of methods are data-driven which often used to perform independent per-frame pose estimation. As they are based on machine learning techniques, they usually require large-scale annotated data for model training. In earlier works, random forests have been the mainstream (P. Li et al., 2015; X. Sun et al., 2015; Tang et al., 2014; Wan et al., 2016) due to its accuracy and speed. Recently, Convolutional Neural Network (CNN) are widely used as they promise large learning capacities for this task (Baek et al., 2018; Ge et al., 2016; Ge et al., 2018; Oberweger et al., 2015; Wan et al., 2017). While these methods operate on single frame estimations and do not propagate error across frames, they are more susceptible to temporal jittering and potential data biases.

Hybrid approaches. As both generative and discriminative approaches have distinct advantages, hybrid approaches have been developed to combine the positive aspects of both. These methods typically use machine learning to initialise pose hypothesis or directly integrate the prediction into the objective function for optimisation procedure of the generative model (C. Qian et al., 2014; Sridhar et al., 2015; Taylor et al., 2016; Taylor et al., 2017; Mueller et al., 2019; Han et al., 2020; Jiayi Wang et al., 2020; Han et al., 2022).

2.1.2 Hand pose and shape estimation

Keypoint regression. With the success of deep neural networks in various computer vision tasks and hand pose datasets (Tang et al., 2014; Tompson et al., 2014; X. Sun et al., 2015; Yuan et al., 2017), many works predict 3D positions of sparse hand keypoints from depth images (Ge et al., 2018; Ge et al., 2017; Yuan et al., 2018; Wan et al., 2018). As monocular RGB cameras find wider applications, many recent works estimate 3D hand pose from RGB images. Zimmermann et al. (2017) is a pioneering work that estimates 3D hand pose by lifting detected hand keypoints in 2D. Cai et al. (2018) propose an end-to-end trainable method that leverages depth images during training as weak supervisions. Mueller et al. (2018) leverage CycleGANs (J.-Y. Zhu et al., 2017) to enhance synthetic hand image datasets by preseving hand poses during image-to-image translation. Iqbal et al. (2018) propose a CNN-based architecture that first predicts latent 2.5D heatmaps before regressing 3D hand joint locations. Spurr et al. (2018) propose to learn an unified latent space for RGB images and 3D hand keypoints, which can be used to retrieve plausible hand poses given input image as well as lifting 2D detections to 3D.

Parametric reconstructions. The introduction of the MANO hand model (Romero et al., 2017) has been influential for hand pose estimation research. This is because most earlier methods do not have access to suitable datasets that contain dense 3D ground truths. MANO offers an alternative path as it can be integrated as a differential operation in an end-to-end trainable framework. It contains prior 3D information about the hand geometry and the input model parameters can be directly regressed using CNNs. Hence, discriminative approaches which regress input MANO pose and shape parameters from single images have been a popular approach (Baek et al., 2019; Boukhayma et al., 2019; Hasson et al., 2019; S. Liu et al., 2021).

Non-parametric reconstructions. Concurrently, there is a line of work which directly regress vertex location of 3D hand meshes. As 3D hand mesh in essence are graph-structured data, graph convolution attracts much attention in hand pose estimation which can be divided into spectral- and spatial-based methods. For spectral-based methods, Ge et al. (2019) and Moon et al. (2020b) adopt the Chebyshev graph convolution (Defferrard et al., 2016) to generate 3D hand mesh at pre-defined topology. On the other hand, Kulon et al. (2020) and Xingyu Chen et al. (2021) follow a spatial-based approach which leverages spiral filters to upsample hand mesh from encoded image features.

Hand pose and shape estimation has been explored with various 3D representations. Parametric hand models, with their ability to estimate mesh surfaces, offer numerous practical advantages for explicitly reasoning about interactions and contacts with objects. On the other hand, non-parametric hand models offer more flexibility in capturing diverse hand shapes and variations, making them better suited for handling complex and unique hand poses and interactions. In Chapters 3 and 4, we explore both of the hand representations by combining the generalisation strength of neural network as well as the template hand shape prior.

2.2 Object

In this section, we first introduce representations that have been used for 3D object modelling in Section 2.2.1. Then, we review works which estimates the rigid pose and shape of objects from visual evidence in Section 2.2.2 and 2.2.3, respectively.

2.2.1 Object representation

Implicit. A family of 3D shapes can be represented using implicit functions on 3D coordinates. In the context of hand-object modelling, we discuss superquadrics (Barr, 1981) and Signed Distance Function (SDF) (Malladi et al., 1995) in the following. Superquadric is a well-studied computational primitive shape abstraction which offers a diverse range of shape representations including cuboids, ellipsoids, cylinders, octohedra and other variations. It was first proposed to model complex objects in computer graphics (Barr, 1981) and later shown to abstract simple objects using a single superquadric (Solina et al., 1990). Recently, (W. Liu et al., 2022) present a probabilistic approach to recover superquadric with improved robustness to outlier and fitting accuracy. SDFs are continuous functions which are parametrised by the 3D space coordinates and can better represent detailed object surfaces. As they can model arbitrary object geometry with adjustable resolution, the community has focused on SDFs and other derivations such as Occupancy Networks (Mescheder et al., 2019).

Point clouds. A point cloud is a collection of points defined in a 3D metric space. Unlike structured grids, point clouds lack a regular arrangement and do not explicitly encode full 3D shape. They have become a significant data format for 3D representation due to the increased availability of acquisition devices and their applications in areas like robotics and autonomous driving (Xiaozhi Chen et al., 2017; Newman et al., 2006). Early attempts do not model local regions effectively with raw point cloud data and tend to project the original point clouds into images as point cloud data is irregular and unordered (You et al., 2018). As there exists information loss caused by the projection, PointNet (Qi et al., 2017a) is proposed to directly process unordered point sets and its extension PointNet++ (Qi et al., 2017b) can be viewed as the generic point cloud analysis framework. Recently, PointNext (G. Qian et al., 2022) shows that small modifications on PointNet++ can outperform more advanced

networks such as Point Transformer (Zhao et al., 2021).

Voxels. Computational object shape can also be described as voxels. They are 3D counterparts to pixels in 2D images and represent the object's volumetric occupancy on a discrete grid in 3D space. Each voxels are typically stored as a binary random variable and can be conveniently reason about the object's volumetric extent. However, the computational cost and memory requirement both increase cubically with the function of the grid resolution. Therefore, octree representation which adaptively subdivides voxels has been widely used to reduce the memory consumption (Laine et al., 2010; Z. Liu et al., 2019; Riegler et al., 2017; Tatarchenko et al., 2017).

Meshes. Representing object shapes using polygonal mesh are common in computer graphics. They can be defined by a set of vertices (3D points on the object surface) and faces (convex polygons that connect these vertices) which forms the surface of the 3D object. Meshes are known for their compactness in capturing the shape of an object, *i.e.* large flat regions can be represented using a sparse set of points. However, meshes do not explicitly capture the volumetric extent of an object. It typically requires computing the distance to all faces when computing the distance from a point to mesh (Möller et al., 1997).

2.2.2 Object pose estimation

Instance-level. In this problem setting, a known 3D object model is assumed to be available during training and testing for estimating the pose of the target object. They can be briefly divided into correspondence matching methods and template matching methods. Correspondence-based methods recover 2D-3D (Rad et al., 2017; Rad et al., 2018) or 3D-3D correspondences (W. Chen et al., 2020a; W. Chen et al., 2020b). Subsequently, they solve the Perspective-n-Point (PnP) and Singular Value Decomposition (SVD) problems with the

2D-3D and 3D-3D correspondences (Kabsch, 1976), respectively. While the above methods work well for textured objects, they typically fail for objects that have homogeneous textures with smooth appearance. Therefore, template matching methods has been proposed to directly match a known object template to the observed image or depth map with either hand-crafted or deep learning feature descriptors (Oberweger et al., 2018; Hinterstoisser et al., 2011; Hinterstoisser et al., 2016; Sundermeyer et al., 2020).

Category-level. Despite impressive performance has been achieved in instance-level object pose estimation, the reliance of known object models limits the generalisation ability for handling everyday objects. Therefore, the task of category-level object pose estimation has been proposed which serves as a more challenging problem setting. In this setting, the major challenge becomes the intra-class object variation in terms of shape and appearance. NOCS (H. Wang et al., 2019) is a pioneering work that tackles shape discrepancy by recovering the normalised shape of the target object and estimating the object pose via point cloud matching. This work has spurred a series of research by extending this framework with shape prior (D. Chen et al., 2020; Tian et al., 2020) and structural similarities (K. Chen et al., 2021; J. Lin et al., 2022). However, they are not applicable in practice due to the computational cost for iterative point matching. To increase the inference speed, FS-Net (W. Chen et al., 2021) and HS-Pose (Zheng et al., 2023) adopt 3D graph convolution (Z.-H. Lin et al., 2020) to enhance geometric sensitivity and feature extraction, respectively.

2.2.3 Object shape estimation

Since there is a vast amount of literature on object shape estimation, we focus on reviewing learning-based approaches that focus on recovering object shape via retrieving object model and deforming mesh templates in the following.

3D model retrieval. With the release of large-scale Computer Assisted Design (CAD) datasets such as ShapeNet (A. X. Chang et al., 2015) and ModelNet (Z. Wu et al., 2015), it established a standardised benchmark for single and multi-view shape reconstruction tasks. A common approach is to train a single CNN to extract image features and perform feature matching against 3D model rendering. Aubry et al. (2015) use a CNN pretrained on ImageNet (Russakovsky et al., 2015) as a image feature extractor and match image features against those of 3D object models rendered under multiple viewpoints to estimate both shape and viewpoint. Other than RGB images, several works perform 3D model retrieval using depth images (Z. Zhu et al., 2016; J. Feng et al., 2016). Recently, Grabner et al. (2018) demonstrate performance gain by utilising a pose prior and Tatarchenko et al. (2019) propose a retrieval baseline for single-view shape reconstruction.

Model deformation. As object model querying has limitations to a pre-defined number of candidates, continuous deformation of point clouds or mesh templates allows for the recovery of an infinite diversity of object shapes. Several early works focus on learning category-specific point cloud deformations to model intra-class variability (Kar et al., 2015; Zia et al., 2015; Prasad et al., 2010). N. Wang et al. (2018) propose to use Graph Convolution Networks to progressively deforming an ellipsoid. Their coarse-to-fine strategy is shown to be effective in producing mesh model with fine details. As spherical template limits the diversity of objects that can be modelled, Groueix et al. (2018) propose AtlasNet which learns to transform simple templates such as 2D squares into surface using Multi Layer Perceptron (MLP). They show that the learned parametric transformation maps everywhere locally to a surface and can be sampled at any desired resolution. Existing single-view reconstruction methods rely heavily on ground truth 3D label for training. However, these annotations are typically only available at scale for synthetic datasets, whereas annotating real-world datasets is prohibitively expensive and impossible to cover everyday objects.

The estimation of arbitrary object shapes enables modeling interactions with unknown items; however, recovering the shape of unseen objects still poses challenges for learning-based approaches. In Chapter 3, we delve into hand-object reconstruction using meshes as object representations, while in Chapter 5, our focus shifts to scenarios involving unseen objects, where we propose the use of superquadrics to approximate their shapes.

2.3 Joint hand-object pose and shape estimation

2.3.1 Existing annotated datasets

First-person hand benchmark (FPHA). FPHA (Garcia-Hernando et al., 2018) is the first benchmark that enables the study of egocentric hand actions with the use of 3D hand poses. This dataset contains a video collection of egocentric dynamic hand actions interacting with 3D objects captured under both RGB and depth cameras. They use marker-based capture system to obtain hand and object poses during interaction. Specifically, as shown in Fig. 2.3, six magnetic sensors are attached to the interacting hand and one for the manipulated object. However, the white markers attached to hand are visible in most sequences which can bias the training. This dataset provides 3D mesh for four objects.

Figure 2.3: Qualitative examples of the FPHA dataset (Garcia-Hernando et al., 2018). As shown, white magnetic sensors attached to hand are visible in most sequences.

ObMan. To overcome the lack of training data for learning-base methods, Hasson et al.

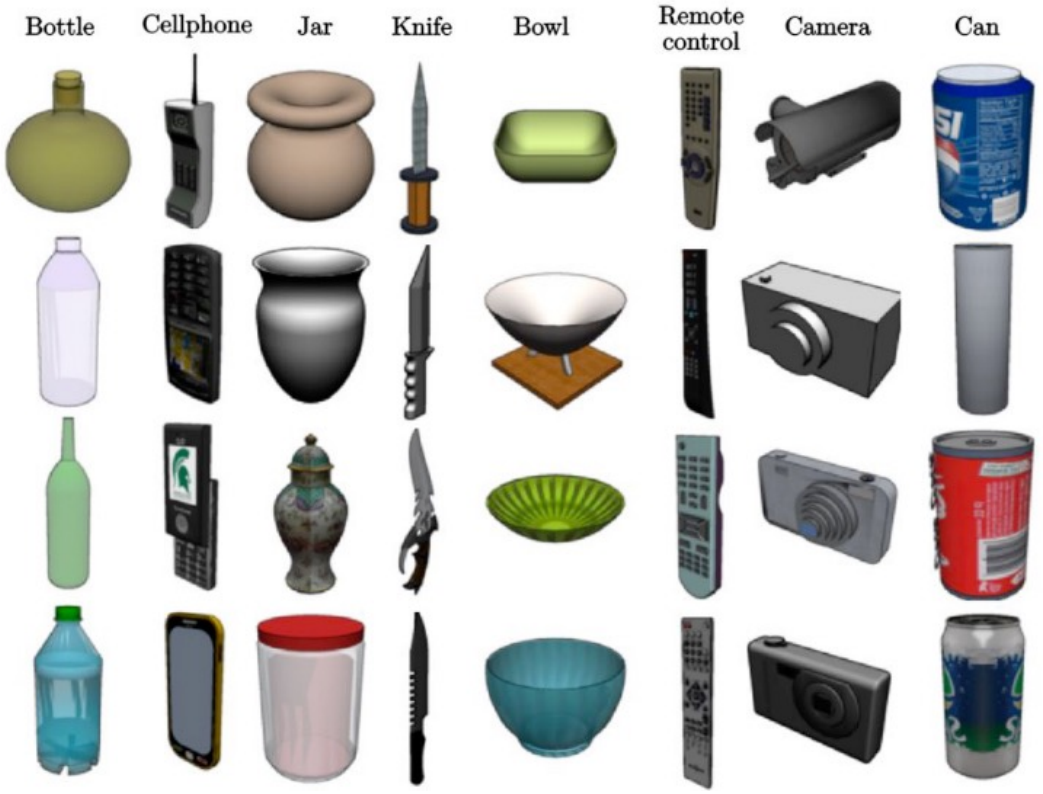


Figure 2.4: Qualitative examples of the ObMan dataset (Hasson et al., 2019). We show example graspable objects from each of the categories. (Figure taken from Hasson (2021).)

(2019) create this large-scale synthetic image dataset of hands grasping objects named ObMan. As illustrated in Figure 2.4, they first select object models for 8 object categories of everyday objects from ShapeNet A. X. Chang et al. (2015), *i.e.* bottles, bowls, cans, jars, knifes, cellphones, cameras and remote controls. Then, they render posed MANO hand model (Romero et al., 2017) to grasp a given object mesh using GraspIt (Miller et al., 2004). The resulting dataset contains in total of 2772 object meshes which are split among 154,000 RGB images.

HO-3D. Motivated by the lack of real dataset and the large domain gap between real and synthetic datasets, Hampali et al. (2020) propose an automatic method to annotate 3D hand-object poses which results in this HO-3D dataset. Their method considers known 3D object model and use the parametric hand model MANO (Romero et al., 2017) to represent

human hand. They simultaneously optimise the 6D object pose and hand pose from multi-view RGB-D images, which provide complete view of hand-object necessary to reason all the occlusions. The resulting dataset consists of 200K images of hand-object interaction at an accuracy reported around 10mm.

H2O-3D. This dataset (Hampali et al., 2022) is an extension of HO-3D (Hampali et al., 2020) with two hands. In similar fashion as in Hampali et al. (2020), the 3D poses of hand and objects automatically with a multi-view setup. This dataset captures 6 subjects interacting with 10 objects from the YCB dataset (Xiang et al., 2018a) which results in 61K training and 15K testing images.

DexYCB. Inspired by the multi-camera setup (Hampali et al., 2020), Chao et al. (2021) instrument their capture system with more cameras and a larger workspace such that it allows human subjects to interact more freely with the target objects. This is the first dataset that allows joint evaluation of the following tasks: 2D object and keypoint detection, 6D object pose estimation, and 3D hand pose estimation. The resulting dataset consists of 582K images of over 1000 sequences of 10 subjects grasping 20 different objects from 8 views.

ContactPose. To go beyond hand-object pose estimations, Samarth et al. (2020) extend ContactDB (Brahmbhatt et al., 2019) and create this dataset (ContactPose) to model hand-object contact. ContactDB (Brahmbhatt et al., 2019) introduce thermal cameras for capturing detailed ground truth contact with 3D hand and object poses. Specifically, their method observes heat transfer from hand to object through a thermal camera after the grasp as it avoids the pitfalls of modelling contact without external instrument such as hands with gloves. We present their contact capture system from ContactDB (Brahmbhatt et al., 2019) as well as qualitative examples of ContactPose dataset (Samarth et al., 2020) in Figure 2.5.

H2O. As most existing datasets consider only single-hand scenarios, H2O (Kwon et al., 2021)

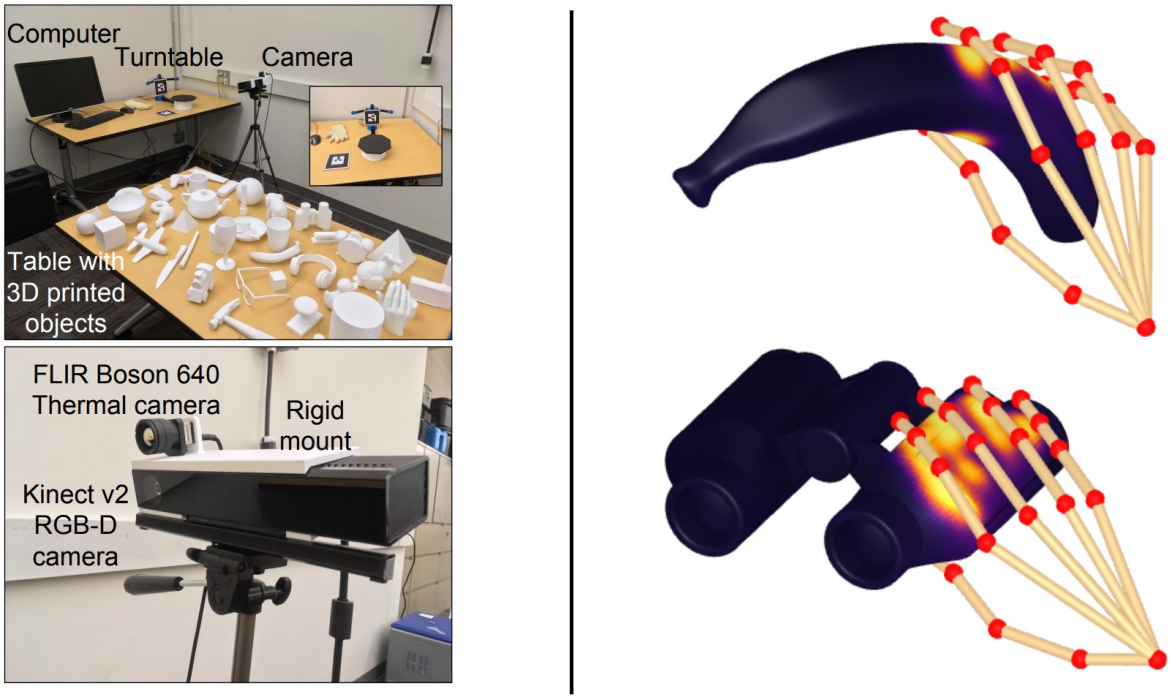


Figure 2.5: Qualitative examples of the ContactPose dataset (Samarth et al., 2020). We show the contact capture system from ContactDB (Brahmbhatt et al., 2019) (*left*) and visualise high-resolution contact maps (object meshes textured with contact) with 3D hand joints from ContactPose dataset (Samarth et al., 2020) (*right*). (Figure taken from Brahmbhatt et al. (2019).)

contains dynamic grasping sequences of two hands interacting with objects. The annotation setup is based on a multi-view RGB-D system which also contains an egocentric camera. In addition to pose annotations, the dataset also contains action labels for each sequences, resulting the first unified dataset for egocentric hand-object interaction recognition with 3D annotations of two hands and the manipulated objects.

OakInk. In order to overcome the limitations of existing real datasets, which only capture a small number of objects and lack comprehensive awareness of the affordance of objects, this dataset (L. Yang et al., 2022) is introduced to enhance visual and cognitive understanding of hand-object interactions. Specifically, this dataset is constructed with two interrelated knowledge bases, *i.e.* Object Affordance Knowledge base (Oak) and Interaction Knowledge base (Ink). The Oak base provides comprehensive descriptions of the affordances of objects within a knowledge and the Ink base captures dynamic hand interactions with objects according to its affordances. To transfer recorded affordances to another object, they also propose a learning-fitting hybrid strategy which extends the total number of distinct hand-object interactions to 50K. We provide qualitative examples of the object affordance knowledge graph and transferred affordances to other objects in Figure 2.6. One limitation of this dataset is that it does not record dynamic hand interactions for the extended objects as well as articulated objects that contain moveable parts.

ARCTIC. ARticulated objeCTs in InteraCtion (ARCTIC) (Z. Fan et al., 2023) is a recent dataset capturing 2.1M RGB images of 10 subjects interacting with 11 different articulated objects. In addition to 3D annotations of hand and objects, they retrieve full-body 3D pose estimates in SMPL-X (Pavlakos et al., 2019) as they provide more reliable global rotations and translations for each hand. This dataset enables two novel tasks of consistent motion reconstruction and interaction field estimation to study dexterous bimanual manipulation motions of hands interacting with articulated objects. As illustrated in Figure 2.7, ARCTIC

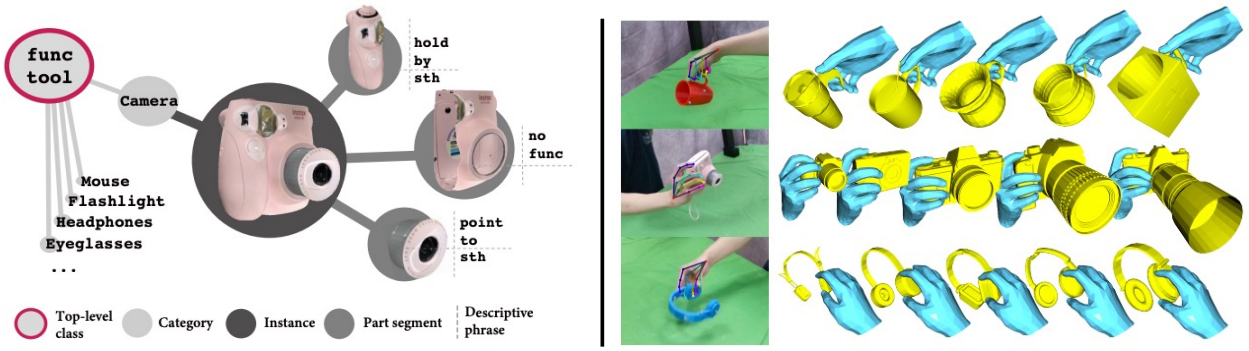


Figure 2.6: Qualitative examples of the OakInk dataset (L. Yang et al., 2022). We show an example of the object affordance knowledge graph in *left* and illustrations of transferred interactions in *right*. (Figure taken from L. Yang et al. (2022).)

contains videos from both third-person and first-person views with accurate ground truth annotations for 3D hand and object meshes.

2.3.2 Reconstruction with known object models

As discussed in Section 1.3, hands and objects inherently undergo significant mutual occlusions which makes joint 3D reconstruction extremely ill-posed during interactions. Therefore, most of the existing works (Garcia-Hernando et al., 2018; Tekin et al., 2019; Doosti et al., 2020; Karunratanakul et al., 2020; Hasson et al., 2020; S. Liu et al., 2021; Cao et al., 2021; L. Yang et al., 2021) reduce this problem to 6D pose estimation by assuming access to instance-specific object templates at inference time. Recently, there is a shift in focus from CNN-based architecture to Transformers (Vaswani et al., 2017). Hampali et al. (2022) propose an efficient Transformer-based network architecture that estimates the two hands and an object pose during various interaction scenarios such as hand-hand and hand-object interactions. Their proposed architecture address the shortcomings in heat-map-based pose estimation methods where they typically suffer from ambiguities localising 2D joint locations. They demonstrate that their Transformer-based architecture can explicitly disambiguate the identity of the keypoints and perform well even on complex configurations.

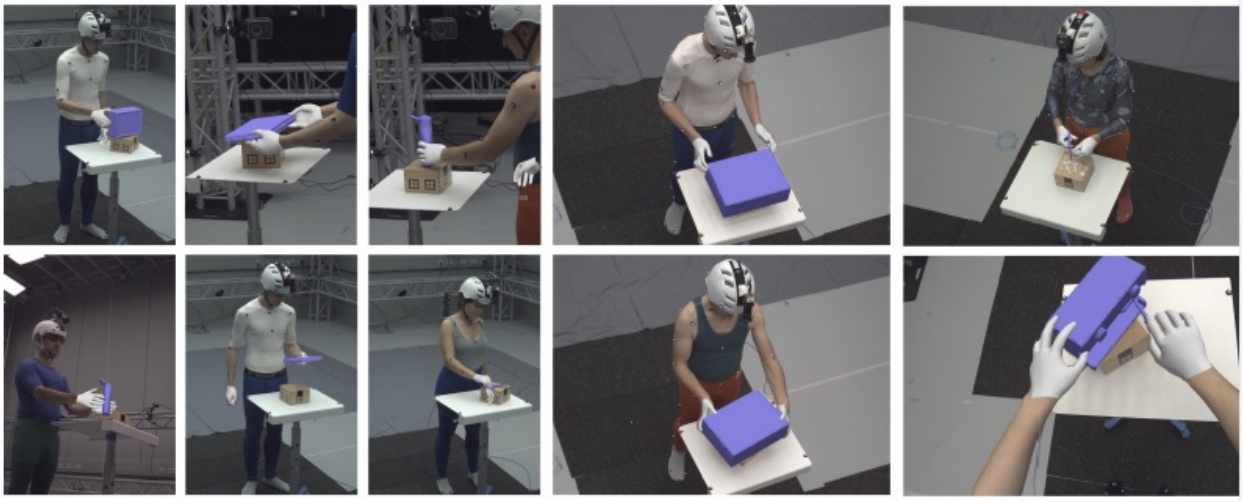


Figure 2.7: Qualitative examples of the ARCTIC dataset (Z. Fan et al., 2023). We show randomly sampled hand-object interaction examples with overlaying ground truths in the RGB images. (Figure taken from Z. Fan et al. (2023).)

2.3.3 Template-free reconstruction

While all the methods presented in Section 2.3.2 assume known object model, several recent approaches simultaneously estimate the shape of both hand and manipulated object from input RGB image. Hasson et al. (2019) differ from previous hand-object reconstruction methods by proposing an end-to-end trainable framework that takes advantages of both differentiable hand model and physical constraints on penetration and contact. Specifically, by following methods which integrate the Skinned Multi-Person Linear (SMPL) parametric body model (Loper et al., 2015) as a neural network layer (Kanazawa et al., 2018; Pavlakos et al., 2018), they also integrate the MANO hand model (Romero et al., 2017). However, the resulting meshes have limited resolution which prevents detailed modelling of contacts between hands and objects. To address this limitation and generate more detailed surfaces, several works (Karunratanakul et al., 2020; Z. Chen et al., 2022; Z. Chen et al., 2023) propose to use SDFs to reconstruct hand and object meshes. In particular, gSDF (Z. Chen et al., 2023) is an extension to AlignSDF (Z. Chen et al., 2022) which embeds strong geometric priors to SDFs by aligning the SDF shape with its underlying kinematic chains of pose

transformations. They show that by doing this can reduce ambiguities in 3D reconstruction. Recently, Y. Ye et al. (2023) propose a diffusion-guided reconstruction framework to tackle the task of hand-object reconstruction from short video clips. Their approach casts 3D inference as a per-clip optimisation and recovers a neural 3D representation of the shape of hand and object. Their key insight is to incorporate data-driven priors to guide the optimisation procedure, in which they propose to a 2D diffusion network to model the distribution over plausible geometric object renderings conditioned on estimated hand pose.

At this point, we have reviewed the literature on hands and object modelling in interaction scenarios. As discussed, most existing methods rely on known object models and constrained capture setups. In Chapter 3, we specifically address the challenging problem of scenarios where only color RGB images are available and operate within a template-free reconstruction setting. Specifically, we introduce a collaborative learning framework to address the mutual occlusion issue when reconstructing dense 3D representations of hands and manipulated objects from a single RGB frame.

2.4 Hand-object interactions

2.4.1 Existing annotated datasets

First-person hand benchmark (FPHA). In addition to hand-object pose annotations mentioned in Section 2.3.1, this dataset (Garcia-Hernando et al., 2018) also provides action classification for the sequences. Specifically, this dataset contains more than 100K frames of 45 daily hand action categories by 6 subjects interacting with 26 different objects.

H2O. While many datasets for third-person action recognition have been proposed (Gu

et al., 2018; Carreira et al., 2017; Sigurdsson et al., 2016), there is increasing interest in understanding egocentric videos (Damen et al., 2018; Goyal et al., 2017; Sigurdsson et al., 2018). However, they mostly involve 2D features which is insufficient for comprehensive understanding of the scene. Therefore, Kwon et al. (2021) create this dataset for egocentric hand-object interaction dataset with markerless 3D annotations of hand-object poses. The resulting dataset captures 571K frames involving dynamic interactions of two hands with objects, split into 36 action classes. Specifically, action labels can be represented as verb-noun pairs where there are in total of 11 verb classes and 8 noun classes.

Something-something. With the goal of enabling neural networks to develop features required for making decision that involves certain aspects of common sense information, the Something-Something dataset (Goyal et al., 2017) is created for video prediction tasks which require common sense understanding. This large-scale dataset contains more than 100K videos across 174 classes.

EPIC-KITCHENS. This dataset (Damen et al., 2018) is a large-scale egocentric dataset featuring diverse everyday tasks. In particular, the video data shows the natural multi-tasking of daily kitchen activities which has not been captured in other existing dataset before. The resulting dataset features 55 hours of video which are densely annotated for a total of 39.6K action segments and 454.3K object bounding boxes.

2.4.2 Recognising hand-object interactions

Action recognition is one of the most actively researched areas in computer vision (Jhuang et al., 2013; Carreira et al., 2017; Varol et al., 2017; Kantorov et al., 2014). In recent years, significant progress has been made with the availability of large-scale datasets (Sigurdsson et al., 2016; Kay et al., 2017; Karpathy et al., 2014; Grauman et al., 2022). In this sub-

section, we focus on methods that simultaneously estimating 3D poses of hand-object and interactions from egocentric videos. As mentioned above, FPHA (Garcia-Hernando et al., 2018) presents the first egocentric dataset and shows that 3D hand poses are beneficial for recognising actions. As the dataset contains visible magnetic sensors and does not include two-hand poses, a markerless dataset named H2O is developed to provide rich 3D annotations for egocentric 3D interaction recognition Kwon et al. (2021). This enables recent research to develop unified models for understanding hand-object interactions. Tekin et al. (2019) extend the YOLO framework (Redmon et al., 2016) to predict 3D poses for hand and object in a single forward pass. Then, the predictions with high confidence values are passed to a Recurrent Neural Network (RNN) to model temporal interactions between hands and objects. As 3D hand pose and gesture recognition are highly correlated tasks, S. Yang et al. (2020) propose a collaborative learning network which splits the two tasks into separate network branches and shares joint-aware features across both branches. They show that performances for both tasks benefit from increasing iterations, demonstrating the effectiveness of their collaborative learning framework. Recently, there are two Transformer-based architecture (Vaswani et al., 2017), *i.e.* HTT (Wen et al., 2023) and H2OTR (H. Cho et al., 2023). HTT (Wen et al., 2023) propose a hierarchical temporal Transformer with two cascaded blocks designed to leverage different time spans for 3D hand-object pose estimation and action recognition. In contrast, H2OTR (H. Cho et al., 2023) exploits contact map for robust interaction recognition.

With the availability of egocentric 3D hand-object interaction datasets, there is increasing interest in developing unified models for hand-object pose estimation and action recognition. However, existing methods still struggle to recognise seen actions on unseen objects due to the limitations in representing object shape and movement using 3D bounding boxes. Additionally, the reliance on object templates at inference time limits their generalisability to unseen objects. In Chapter 5, we explore superquadrics as an alternative 3D object

representation to bounding box and demonstrate their effectiveness on both template-free object reconstruction and action recognition tasks. Furthermore, we study the compositionality of actions by considering a more challenging task where the training combinations of verbs and nouns do not overlap with the testing split.

Chapter Three

LEARNING JOINT HAND-OBJECT RECONSTRUCTION FROM A SINGLE RGB IMAGE

This chapter presents work published at the 2022 Conference on Computer Vision and Pattern Recognition (CVPR) in New Orleans, USA (Tse et al., 2022a).

In Section 1.3, we underlined several outstanding challenges that arise during 3D pose and shape estimation of hand-object. Our goal in this chapter is to develop method which can estimate the dense 3D geometry of hands and objects from RGB images with learning-based approach. We are interested in approaches that benefit from the success of neural network and focus on a single colour image as input and predict the shape of the unknown object and the associated hand. In this chapter, we present a collaborative learning strategy where two-branches of deep networks are learning from each other. Our algorithm is agnostic to object models and it learns the physical rules governing hand-object interaction.

Joint hand and object pose and shape estimation is a challenging problem. Firstly, self-occlusion in hand is a well-known issue, as highlighted in previous studies (Q. Ye et

al., 2018; Rangesh et al., 2016). However, when hands interact with objects, the level of occlusion between hands and objects becomes even more significant from almost any viewpoints (Nakamura et al., 2017). Secondly, in the case of first-person-view scenarios (e.g., the FPHA dataset (Garcia-Hernando et al., 2018)), there is often a considerable amount of erratic camera motion, which further complicates the estimation process. While recent works (Tekin et al., 2019; Doosti et al., 2020; S. Liu et al., 2021) have made progress in addressing some of the primary challenges in joint hand-object pose estimation using color input, they still face limitations in the absence of physical constraints and sparse keypoint detection. These limitations frequently result in erroneous pose estimation or mesh reconstructions, such as hands penetrating objects.

To fundamentally understand hand-object interactions, it is essential to fully recover dense 3D information, and accordingly, there has been significant improvement towards hand mesh estimations from single RGB image (Ge et al., 2019; Baek et al., 2019; Boukhayma et al., 2019; Zimmermann et al., 2019; Kulon et al., 2020; Y. Zhou et al., 2020; Baek et al., 2020; Choi et al., 2020; Moon et al., 2020b). To enable physically plausible reconstructions, Hasson et al., 2019 propose both attraction and repulsion loss terms during model training. Along with optimisation-based approaches (Cao et al., 2021; Hasson et al., 2021), they are limited to scenarios where hand and object are already in contact due to the contact loss terms. However, we argue that the ability to reason pre-grasp stages are equally important as it allows agents to infer human intents (Meltzoff, 1995) and learn manipulation skills from humans (Mandikal et al., 2020). Therefore, we are interested in learning a strategy that is not restricted by these contact terms and is able to learn the context of actual as well as near physical contact.

In this chapter, we present a novel collaborative learning framework which allows hand and object branches to boost each other in a progressive and iterative fashion. There are two motivations for this design: 1) estimating the pose and shape of interacting hands and objects

is a highly-correlated task and 2) mutual occlusions can be tackled by simultaneously sharing 3D geometric information. This is supported by the fact that the image encoder struggles to extract useful features under mutual occlusion. Therefore, capturing object mesh information would compensate this limitation for hand reconstruction (same for object branch). In previous approaches within this context, information sharing across branches was achieved through simple branch stacking (S. Yang et al., 2020), which introduced a communication bottleneck. However, our empirical observations revealed that this approach had limitations in terms of performance gain across network inference iterations. We explicitly address this by a new unsupervised associative loss facilitating the information transfer. Further, to address frequently occurring occlusions in hand-object interaction scenarios, we introduce an attention-guided graph convolution that demonstrates the ability to improve mesh quality as well as correct hand and object poses.

The specific contributions of this chapter are threefold. We first introduce an end-to-end trainable collaborative learning strategy for hand-object reconstruction from a single RGB image. We then design an attention-guided graph convolution to capture mesh information dynamically. We also introduce an unsupervised training strategy for effective feature transfer between hand-object branches. Lastly, we demonstrate that our approach achieves highly physically plausible results without contact terms.

In this chapter, we focus on reconstructing dense meshes of hand and manipulated objects. We first provide preliminaries on hand-object reconstruction and collaborative learning methods in Section 3.1. We then present our novel collaborative learning framework in Section 3.2. Lastly, we present experimental results which validate our approach in Section 3.3 and conclude the chapter in Section 3.4.

3.1 Preliminary

The work in this chapter tackles the problem of hand and object reconstruction from a single RGB images. In the following, we first introduce the line of work that leverages Graph Convolutional Neural Networks on hand reconstruction tasks in Section 3.1.1. Then, we provide a brief review on Collaborative Learning in Section 3.1.2.

3.1.1 Graph convolution-based methods

As both skeleton and mesh can be represented in form of a graph, graph convolution naturally attracts much attention in the field of hand pose estimation. We can split graph convolution-based methods into spectral- (Bruna et al., 2014; Defferrard et al., 2016; Kipf et al., 2017) and spatial-based methods (Gilmer et al., 2017; K. Xu et al., 2018; Monti et al., 2017). For methods that use spectral-based graph convolution, Ge et al. (2019) and Choi et al. (2020) adopt the Chebyshev spectral graph convolution (Defferrard et al., 2016) to produce 3D hand mesh. For spatial-based graph convolution, both Cai et al. (2019) and Doosti et al. (2020) leverage GCN (Kipf et al., 2017) to exploit the spatial and temporal consistencies for hand and object pose estimation. Kulon et al. (2020) exploit spiral filters to recover hand mesh directly from autoencoder. They demonstrate that spatial mesh convolutions outperform spectral methods and SMPL-based models (Loper et al., 2015; Romero et al., 2017) for hand reconstruction. In contrast, our attention-guided graph convolution is able to take dynamic graph as input and it does not assume a fixed neighbourhood for feature aggregation.

3.1.2 Collaborative learning

There has been a rich history concerning learning multiple tasks simultaneously. We can briefly divided into multi-task learning (Baxter, 1997; Baxter, 2000; Caruana, 1993), domain adaptation (Mansour et al., 2009; Mansour et al., 2008), distributed learning (Balcan et al., 2012; Dekel et al., 2011; Jialei Wang et al., 2016) and collaborative learning (Blum et al., 2017; Z. Jiang et al., 2017; G. Song et al., 2018; Nguyen et al., 2018). Collaborative learning studies to make learning more efficient through sharing of information. Blum et al. (2017) proposes a collaborative probably approximately correct (PAC) learning model which was built upon (Valiant, 1984) and Nguyen et al. (2018) and J. Chen et al. (2018) are the follow-up works. G. Song et al. (2018) introduces one form of collaborative learning framework in which multiple classifier heads of the same network are simultaneously trained on the same training data to improve generalisation and robustness without extra inference cost. There are two major mechanisms under his framework: 1) Same training datasets for multiple views from different classifiers improves generalisation and 2) Intermediate-level representation sharing. S. Yang et al. (2020) exploits joint-aware features for gesture recognition and 3D hand pose estimation. Their mechanism focuses on intermediate-level representation sharing iteratively across multiple tasks. In this work, we improve on S. Yang et al. (2020) with an attention-guided graph convolution and an unsupervised associative loss to guide the intermediate-level representation sharing process. Also, our graph convolution is based on a multi-head attention mechanism which possesses the spirit of G. Song et al. (2018) to improve generalisation with multiple views on the same dataset.

3.2 Methodology

Our goal is to develop a model which can jointly estimate the shape of hand and the manipulated object from a single colour image. As illustrated in Figure 3.1, our training pipeline takes an input RGB image $\mathbf{x} \in \mathbb{R}^{256\times256}$ and involves 4 steps for one iteration. Each branch has a ResNet-18 (He et al., 2016) encoder pre-trained on ImageNet (Russakovsky et al., 2015), *i.e.* $\text{ENC}_{hand}(\mathbf{x})$ and $\text{ENC}_{obj}(\mathbf{x})$.

The primary motivation behind our approach is to exploit the implicit hand-object relationship between hands and objects. We aim to address the challenge of mutual occlusion in hand-object interactions by simultaneously sharing 3D reconstructions within our collaborative learning framework. However, directly connecting network branches in a naïve manner often resulted in error accumulation and rendered the training process highly unstable. To address this, we design an attention-guided graph convolution to capture 3D reconstructions dynamically and introduce an unsupervised associative loss to improve the feature transfer process from hand to object, and vice versa. Our networks are trained in an end-to-end manner and Alg. 1 summarises the training process.

Algorithm 1 Collaborative learning algorithm

Require: $\mathbf{x}$: input image, P : network iteration

function OPTIMISE($\mathcal{L}_{Total}$)
  $\mathbf{r}_{hand} \leftarrow \text{ENC}_{hand}(\mathbf{x})$ ▷ Extract hand features
  $\mathbf{m}_{hand} \leftarrow \text{g}^{HME}(\mathbf{r}_{hand})$ ▷ Get hand mesh
  for $t = 1$ to P **do**
    $\phi_{hand} \leftarrow g^{conv}_{hand}(\mathbf{m}_{hand})$ ▷ Hand Graph Conv.
    $\mathbf{r}_{obj} \leftarrow \text{ENC}_{obj}(\mathbf{x}) + \phi_{hand}$ ▷ Feature update
    $\mathbf{m}_{obj} \leftarrow \text{g}^{OME}(\mathbf{r}_{obj})$ ▷ Get object mesh
    $\phi_{obj} \leftarrow g^{conv}_{obj}(\mathbf{m}_{obj})$ ▷ Object Graph Conv.
    $\mathbf{r}_{hand}{}' \leftarrow \mathbf{r}_{hand} + \phi_{obj}$ ▷ Feature update
    $\mathbf{m}_{hand} \leftarrow \text{g}^{HME}(\mathbf{r}_{hand}{}')$
  end for
end function

3.2.1 Hand mesh estimator

In Section 2.1.1, we introduced the parametric hand model MANO (Romero et al., 2017) which has the advantages of capturing the pose and shape variations in a low-dimensional parameters. Therefore, regressing these low-dimensional parameters in a learning-based framework has the potential to output statistically plausible hand reconstruction. By following (Hasson et al., 2019), we adopt the differential MANO model which maps pose ($\boldsymbol{\theta} \in \mathbb{R}^{51}$) and shape ($\boldsymbol{\beta} \in \mathbb{R}^{10}$) parameters to a mesh with $N = 778$ vertices. Pose parameters ($\boldsymbol{\theta}$) consists of 45 DoF (*i.e.* 3 DoF for each of the 15 finger joints) plus 6 DoF for rotation and translation of the wrist joint. Shape parameters ($\boldsymbol{\beta}$) are fixed for a given person. A kinematic tree is formed with the 15 joints and the wrist joint as the first parent node. Joint locations can be obtained using the kinematic tree with global rotation based on $\boldsymbol{\theta}$. In the following, we describe our hand mesh estimator g^{HME}.

Given the 512-dimensional hand feature vector $\mathbf{r}_{hand}$ from $\text{ENC}_{hand}(\mathbf{x})$, we use a fully connected layer to regress $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The original MANO model uses 6-dimensional Principal Component Analysis (PCA) subspace of $\boldsymbol{\theta}$ for computational efficiency. However, we empirically observed that full 45-dimensional pose space better captures a variety of hand poses especially over sequential datasets.

We define hand mesh as $\mathbf{m}_{hand} = (\mathbf{v}_{hand}, \mathbf{f}_{hand})$, where $\mathbf{v}_{hand} \in \mathbb{R}^{778\times3}$ refers to a set of vertices in the mesh and $\mathbf{f}_{hand} \in \mathbb{R}^{1538\times3}$ refers to a close set of edges (*i.e.* a triangle face has 3 edges). We keep the mesh faces consistent to the original MANO model (Romero et al., 2017).

Hand reconstruction loss $\mathcal{L}_{hand}$. We directly optimise root-relative 3D positions by

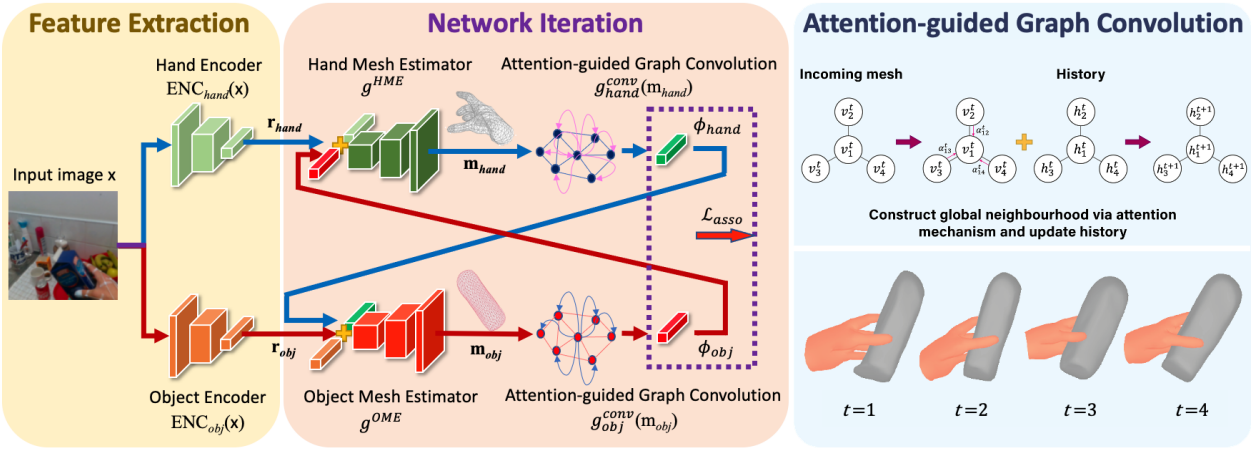


Figure 3.1: A schematic illustration of our framework. It takes an input image $\mathbf{x}$, which goes through two separate encoders, $\text{ENC}_{hand}(\mathbf{x})$ and $\text{ENC}_{obj}(\mathbf{x})$ to produce hand and object features, $\mathbf{r}_{hand}$ and $\mathbf{r}_{obj}$, respectively. Hand mesh estimator g^{HME} takes $\mathbf{r}_{hand}$ and output hand mesh $\mathbf{m}_{hand}$ which is then pass to graph convolution module g^{conv}_{hand} and output ϕ_{hand}. Object mesh estimator takes both $\mathbf{r}_{obj}$ and ϕ_{hand} to output object mesh $\mathbf{m}_{obj}$. Similarly, graph convolution module g^{conv}_{obj} takes object mesh $\mathbf{m}_{obj}$ and output ϕ_{obj} which is then combine with hand features $\mathbf{r}_{hand}$ and goes into the hand mesh estimator g^{HME}. We also introduce an unsupervised associative loss to supervise the feature transfer process under network iterations, *i.e.* ϕ_{hand} and ϕ_{obj}. We include a qualitative example on the bottom right corner which demonstrates the effect of our attention-guided graph convolution for iteration t.

minimising their L2 distance to the corresponding ground-truth vertex positions $\mathbf{v}^*_{hand}$:

$$\mathcal{L}_V(\mathbf{v}_{hand}) = \|\mathbf{v}_{hand} - \mathbf{v}^*_{hand}\|_2^2 . \tag{3.1}$$

When ground truth vertex positions are not available, we supervise on 3D joint locations $\mathbf{J} \in \mathbb{R}^{n\times 3}$ where n refers to the number of joints. The 3D joint loss is defined as:

$$\mathcal{L}_J(\mathbf{J}) = \|\mathbf{J} - \mathbf{J}^*\|_2^2 , \tag{3.2}$$

where $\mathbf{J}^*$ refers to ground truth joint positions. The resulting loss is defined as: $\mathcal{L}_{hand} = \mathcal{L}_V + \mathcal{L}_J$. We do not adopt shape regularisation as in Hasson et al. (2019) as we empirically observed that our iterative process already prevents extreme mesh deformation.

3.2.2 Object mesh estimator

In this section, we introduce our object mesh estimator g^{OME} which is adopted from Hasson et al. (2019). Given the 512-dimensional object feature vector $\mathbf{r}_{obj}$ from $\text{ENC}_{obj}(\mathbf{x})$, we use AtlasNet (Groueix et al., 2018) to estimate object mesh $\mathbf{m}_{obj} = (\mathbf{v}_{obj}, \mathbf{f}_{obj})$, *i.e.* $\mathbf{v}_{obj} \in \mathbb{R}^{642\times 3}$ refers to object vertices and $\mathbf{f}_{obj} \in \mathbb{R}^{1280\times 3}$ refers to object mesh faces.

Object reconstruction loss $\mathcal{L}_{obj}$. As object mesh is reconstructed in the camera coordinate frame, it can be directly optimised by minimising the Chamfer distance. The resulting loss is defined as:

$$\mathcal{L}_{obj}(\mathbf{v}_{obj}) = \frac{1}{2}\Big(\sum_{x\in\mathbf{v}_{obj}} d_{\mathbf{v}^*_{obj}}(x) + \sum_{y\in\mathbf{v}^*_{obj}} d_{\mathbf{v}_{obj}}(y)\Big), \tag{3.3}$$

where $\mathbf{v}^*_{obj}$ refers to the points uniformly sampled on the surface of the ground truth object, $d_{\mathbf{v}^*_{obj}}(x) = \min_{y\in\mathbf{v}^*_{obj}} \|x-y\|_2^2$, and $d_{\mathbf{v}_{obj}}(y) = \min_{x\in\mathbf{v}_{obj}} \|x-y\|_2^2$.

3.2.3 Attention-guided graph convolution

In the following, we first introduce the preliminary of message passing scheme (Gilmer et al., 2017) in graph convolution. We then provide details of our attention-guided graph convolution g^{conv}.

Preliminary. We follow the message passing scheme in graph convolution to capture mesh information and transfer to the opposite branch. By denoting vertex feature $\mathbf{v}_i^{(k)} \in \mathbb{R}^F$ of vertex i in layer k, the first step of such message passing scheme can be described as:

$$\mathbf{msg}_i^k = \text{AGGREGATE}^{(k)}\big(\{\mathbf{v}_u^{(k-1)}, u \in \mathcal{N}(i)\}\big), \tag{3.4}$$

where message $\mathbf{msg}_i^k$ is formed by aggregating neighbourhood $\mathcal{N}(i)$ around vertex i from previous layer $(k-1)$. The second step updates vertex feature with this new message:

$$\mathbf{v}_i^k = \text{UPDATE}^{(k)}\big(\mathbf{v}_i^{(k-1)}, \mathbf{msg}_i^k\big). \tag{3.5}$$

The choice for neighbourhood $\mathcal{N}(i)$, aggregating function $\text{AGGREGATE}^{(k)}$ and update function $\text{UPDATE}^{(k)}$ are crucial. While there has been a variety of functions proposed in the literature (Defferrard et al., 2016; Kipf et al., 2017; Gilmer et al., 2017; K. Xu et al., 2018), we leverage attention mechanism to construct aggregating neighbourhood and a history term for updating node features.

Objective. By defining P to be the number of iterations per forward pass, the input is a sequence of meshes $(\mathbf{m}_\theta^1, \mathbf{m}_\theta^2, ...)$ where $\mathbf{m}_\theta^t = (\mathbf{v}_\theta^t, \mathbf{f}_\theta^t)$ for $t \in [1, \ldots, P]$ is defined by vertices $\mathbf{v}_\theta^t$ and faces $\mathbf{f}_\theta^t$ for either branch $\theta \in \{hand, obj\}$. The objective is to estimate feature offset Δ_{hand}^t from the hand branch for object reconstruction, and vice versa:

$$\mathbf{r}_{obj}^{t+1} = \mathbf{r}_{obj}^t + \Delta_{hand}^t. \tag{3.6}$$

Attention-guided graph convolution. As the above sequential task involves dynamically evolving graphs, static graph convolution would not be suitable because the weights are only being updated after P iterations. Therefore, a solution should maintain the history of operations.

By assuming input mesh vertices $\mathbf{v}_\theta$ is an un-ordered set, we propose to dynamically construct neighbourhoods $\mathcal{N}(i)$ using attention mechanism (Bahdanau et al., 2015; Gehring et al., 2017). Attention coefficient $\alpha_{ij} \in [0, 1]$ is defined as the importance of vertex j's features to vertex i (Veličković et al., 2018). Node j is included in the neighbourhood $\mathcal{N}(i)$ of i when α_{ij} is larger than a threshold, *i.e.* 0.5. Finally, our graph convolution layer at

iteration t can be defined by rewriting Eqs. (3.4–3.5) as:

$$\alpha_{ij}^{t} = \frac{\exp\Big(\text{LeakyReLU}(\mathbf{a}^{\top}[\mathbf{W}\mathbf{v}_i^t \| \mathbf{W}\mathbf{v}_j^t])\Big)}{\sum_{k\in\mathbf{v}^t}\exp\Big(\text{LeakyReLU}(\mathbf{a}^{\top}[\mathbf{W}\mathbf{v}_i^t \| \mathbf{W}\mathbf{v}_k^t])\Big)} \tag{3.7}$$

where attention coefficient α_{ij}^t is computed using incoming vertices $\mathbf{v}^t = \{\mathbf{v}_1^t, ..., \mathbf{v}_N^t\}$ with N being the maximum mesh vertices and learnable weights $\mathbf{a} \in \mathbb{R}^{2F}$ and $\mathbf{W} \in \mathbb{R}^{F\times 3}$. Note that F is a hyperparameter and $\|$ is concatenation operation. We then update history $\mathbf{h}_i^t$ of vertex i:

$$\mathbf{h}_i^{t+1} = \text{LayerNorm}\bigg(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}^t(i)}\alpha_{ij}^{t,k}\mathbf{v}_j^t + \mathbf{h}_i^t\bigg), \tag{3.8}$$

where $\mathcal{N}^t(i)$ is the aggregating neighbourhood around vertex i at t, history $\mathbf{h}^t = \{\mathbf{h}_1^t, ..., \mathbf{h}_N^t\}$ and it is initialised as $\mathbf{0}$. Similar to Veličković et al. (2018) and Vaswani et al. (2017), we find multi-head attention α_{ij}^k to be beneficial and apply layer normalisation (Ba et al., 2016) to stabilise and enable faster training. We use residual connection (He et al., 2016) to track the history sequence and prevent performance drop on increasing iterations. In the final step, we use a fully connected layer to resize to the same size as image features $\mathbf{r}_\theta(\mathbf{x})$, namely ϕ_θ.

Discussions. Our graph convolution is reminiscent to GAT (Veličković et al., 2018) and any k-nearest neighbours (k-NN) based dynamic graph convolutions such as EdgeConv (Y. Wang et al., 2019). However, our approach differentiates from those because firstly, we do not assume static graph inputs. Secondly, we differentiate from GAT (Veličković et al., 2018) by how we leverage attention mechanism - they aggregate on fixed and local neighbourhood whereas we take this further by dynamically constructing global neighbourhood using attention mechanism. In addition, as the incoming mesh are 3D positions, k-NN like approaches suffer from local neighbourhood aggregation and high k-NN computational cost at each iterations. In contrast, our method is able to capture long-range dependencies from dynamic

graph in a single layer.

3.2.4 Associative supervision

Due to mutual occlusion in hand-object scenarios (Nakamura et al., 2017), it is challenging for the image encoder to capture useful information for mesh reconstruction. Instead, here we rely on the fact that hand pose changes with respect to different objects. For example, we hold cups differently depending on whether it has handle or not. We hypothesise that object branch benefits from hand mesh information (and vice versa for hand branch) and assume that good feature transfer in collaborative learning occurs when these features are highly similar within the same object class and distinctive across all other object classes. However, in practice, such object class information is not available. Hence, we introduce an unsupervised loss to facilitate effective feature transfer.

Objective. Given $\phi_\theta = \{\phi_\theta^1, ..., \phi_\theta^B\}$ with B being the input batch size, we update the image features by simple addition. In the following, we describe an unsupervised loss for ϕ_θ.

Associative loss $\mathcal{L}_{asso}$. Our approach is inspired by Haeusser et al. (2017) which was originally designed for semi-supervised learning. We imagine a walker going along $\Phi_i = [\phi_{hand}^i; \phi_{obj}^i]$ where $i \in \{1, \ldots, B\}$. As each Φ_i comes in pair with the same object class, we define a correct walk if transition is under the same object class. We define similarity between two embeddings as:

$$M_{ij} = \Phi_i^\top \Phi_j, \;\; 1 \leq i, j \leq B. \tag{3.9}$$

A single transition based on embeddings similarity is defined as:

$$P_{ij} = P(\Phi_j|\Phi_i) = \frac{\exp(M_{ij})}{\sum_{j'} \exp(M_{ij'})}. \tag{3.10}$$

The round trip probability (Markov Chain) of walking from i to j can then be defined as:

$$P_{ij}^{round} = \sum_{k \in \{1,\dots,B\}} P_{ik} P_{kj}. \tag{3.11}$$

We further extend this into an unsupervised loss by encouraging the walker to walk back to its starting batch index i. This can be achieved by leveraging the fact that batch index implicitly refers to an object class $C_{obj} \in \{1, \dots, O\}$ and $O \ll B$. An unsupervised loss $\mathcal{L}_{asso}$ can be obtained as:

$$\mathcal{L}_{asso}(\phi_\theta) = \left\| U - P^{round} \right\|_F^2, \tag{3.12}$$

where $\| \cdot \|_F$ is the Frobenius norm and U is a diagonal matrix of $\frac{1}{O}$ values: The i-th diagonal entry U_{ii} represents that the walker starts at and returns to state i. U can be adjusted if dataset is class-imbalanced.

3.2.5 Training

Our final loss $\mathcal{L}_{final}$ is defined as:

$$\mathcal{L}_{final} = \mathcal{L}_{hand} + \mathcal{L}_{obj} + \mathcal{L}_{asso}. \tag{3.13}$$

3.3 Experiment

3.3.1 Implementation details

We train all parts of the network simultaneously with Adam optimiser (Kingma et al., 2015) at a learning rate 10^{-4} for 400 epochs. We then freeze the ResNet (He et al., 2016) encoders and decrease the learning rate to 10^{-5} for another 100 epochs. We empirically fixed $K = 3$ attention heads and $P = 2$ iterations to produce the best results.

3.3.2 Datasets

First-person hand benchmark (FPHA). As discussed in Section 2.3, this is a widely-used dataset (Garcia-Hernando et al., 2018) which contains egocentric RGB-D videos on a wide range of hand-object interactions. There are 4 available objects, *i.e.* juice bottle, liquid soap, milk and salt. For fair comparisons with Tekin et al. (2019) and Hasson et al. (2020), we follow the same *action split* for evaluation where each object is present in both training and testing. We also conduct a comparison with Hasson et al. (2019), following their experimental settings, which involve the *subject split* of the dataset. Their approach includes filtering frames in which the hand is more than $1cm$ away from the manipulated object and excludes the milk object. We call this subset $FPHA^{-}$ which contains a total of 3 objects.

ObMan. Similarly, as introduced in Section 2.3, this is a synthetic dataset (Hasson et al., 2019) which was produced by rendering hand meshes with selected objects from ShapeNet (A. X. Chang et al., 2015). It captures 8 object categories and results in a total of 2,772 meshes which are split among 154,000 image frames. We pretrained the network on *ObMan* before

training on other real datasets as we observed in our preliminary experiments that their setting led to consistent improvements over training directly on real data.

DexYCB. This is a recent real dataset for capturing hand grasping of objects (Chao et al., 2021) which consists a total of 582,000 image frames on 20 objects from YCB-Video dataset (Xiang et al., 2018b). We present results on all 4 official dataset split settings.

HO-3D. This dataset (Hampali et al., 2020) is most similar to *DexYCB* where it consists of 78,000 images frames on 10 objects. We present results on the official dataset split (version 2). The hand mesh error is reported after procrustes alignment and in mm.

3.3.3 Evaluation metrics

Hand error. We report the mean end-point error (mm) (Hasson et al., 2019) over 21 joints and use the percentage of correct keypoints (PCK) score to evaluate at different error thresholds.

Object error. We measure the accuracy of object reconstruction by computing the Chamfer distance (mm) (Hasson et al., 2019) between points sampled on ground truth and predicted mesh.

Hand-object interaction. To understand hand-object interaction, we followed Hasson et al. (2019) to include penetration depth (mm) and intersection volume (cm^3). Penetration depth refers to the maximum distances from hand mesh vertices to the object's surface when in a collision. Intersection volume is obtained by voxelising the hand and object using a voxel size of $0.5cm$.

3.3.4 Results

Joint hand-object reconstruction. As discussed in Section 2.3.2 that most recent efforts on joint hand-object reconstructions assume known object models, we compare with ObMan (Hasson et al., 2019) in Table 3.1. Similar to *FPHA*, we used the default *DexYCB* split and filtered frames when hand and manipulated object are $1cm$ apart. We name this subset to be $DexYCB^{-}$ and retrain ObMan using their released code. As shown in Table 3.1, there is still a presence of interpenetration at test time and even increases the hand error by $0.7mm$ on $FPHA^{-}$ with contact loss in ObMan. This is mainly due to the fact that their model is not implicitly learning the physical rules imposed by the contact loss. In contrast, our method consistently outperforms ObMan with a higher hand-object reconstruction accuracy. In addition, we provide qualitative comparisons on *FHB* and *CORe50* (Lomonaco et al., 2017) datasets in Figure 3.2.

Table 3.1: Quantitative comparison with ObMan on *ObMan* (Hasson et al., 2019), $FPHA^{-}$ (Garcia-Hernando et al., 2018) and $DexYCB^{-}$ (Chao et al., 2021) datasets. * refers to the results with contact loss. Our collaborative learning strategy performs competitively without physical contact loss.

Datasets	*ObMan*			$FPHA^{-}$			$DexYCB^{-}$	
Method	ObMan	ObMan*	Ours	ObMan	ObMan*	Ours	ObMan*	Ours
Hand error $(mm)\downarrow$	11.6	11.6	**9.1**	28.1	28.8	**25.3**	17.6	**15.3**
Object error $(mm)\downarrow$	641.5	637.9	**385.7**	1579.2	1565.0	**1445.0**	549.4	**501.2**
Max. penetration $(mm)\downarrow$	9.5	9.2	**7.4**	18.7	**12.1**	16.1	14.6	**12.1**
Intersection vol. $(cm^3)\downarrow$	12.3	12.2	**9.3**	26.9	16.1	**14.7**	14.9	**13.4**

Hand pose estimation. We first compare with state-of-the-art methods on *HO-3D* (Hampali et al., 2020) in Table 3.2. As shown, our method performs competitively against methods that assumes known object models. Then, we benchmark on *FPHA* (both *action split* and *subject split*) and report results in Tables 3.3 and 3.4. As shown in Table 3.3, we demonstrate superior performances among all three architecturally similar networks (Hasson et al., 2019;

Figure 3.2: Qualitative comparison with ObMan (Hasson et al., 2019). Top two rows refers to models trained with *FPHA*. Bottom two rows refers to in-the-wild settings where models are only trained with synthetic dataset *ObMan*. Our method is able to refine and sharpen object mesh under the collaborative learning framework (see blue arrows) and generalise better hand pose in both settings.

Hasson et al., 2020). We attribute the performance gain in *action split* to the fact that *FPHA*$^{-}$ contains almost half of *FPHA* with incomplete object list and unseen test subjects during test time. We analyse our hand pose estimation performance using the PCK metric in Table 3.4. Note that S. Yang et al. (2020) takes sequential images as input and leverages action recognition task in their collaborative framework. We achieve state-of-the-art performance to in hand pose estimation with the advantage of object reconstruction. We also plot 3D PCK curves in Figure 3.3. Finally, we compare with a supervised version of Spurr et al.

Table 3.2: Error rates of different hand pose estimation methods on *HO-3D* (Hampali et al., 2020). Note that the reported results for S. Liu et al. (2021) output hand meshes only. We outperform two other architecturally similar networks (Hasson et al., 2019; Hasson et al., 2020) without known object models under our collaborative learning framework.

Method	Mesh error ↓	F-score @5mm ↑	F-score @15mm ↑	Known objects
Hasson et al. (2019)	11.0	46.0	93.0	✗
Hampali et al. (2020)	10.6	50.6	94.2	✓
S. Liu et al. (2021)	**9.5**	**52.6**	**95.5**	✓
Hasson et al. (2020)	11.4	42.5	93.4	✓
Ours	10.9	48.5	94.3	✗

(2020) which won the HANDS 2019 Challenge (Armagan et al., 2020) on *DexYCB* (Chao et al., 2021) and report the results in Table 3.5.

Table 3.3: Error rates of different algorithms. *FPHA* refers to *action split* and *FPHA*$^-$ refers to *subject split* of the dataset.

Method	*FPHA* Hand Error	*FPHA*$^-$ Hand Error
Tekin et al. (2019)	15.8	-
Hasson et al. (2020)	-	28.0
Hasson et al. (2019)	18.0	27.4
Cao et al. (2021)	14.2	-
Ours	**9.8**	**25.3**

3.3.5 Ablation study

To motivate our design choices, we present a quantitative comparison of our method with various components disabled. We validate that the combination of our design choices outperforms the naïve collaborative learning baseline (as illustrated in Figure 3.4), which predicts the embeddings directly and perform 3D reconstruction last.

Impact of the number of network iterations (P). Table 3.6 shows the results of varying

Table 3.4: PCK performance over respective error threshold on *FPHA*. Compared to another collaborative learning framework (S. Yang et al., 2020) and graph-based method (Doosti et al., 2020), our method performs better and is able to reconstruct both hand-object meshes.

Method	PCK@20mm	PCK@25mm
Tekin et al. (2019)	69.17%	81.25%
Garcia-Hernando et al. (2018)	74.73 %	82.10%
S. Yang et al. (2020)	81.03%	86.61%
Doosti et al. (2020)	92.17%	92.63%
Ours	**93.14%**	**95.65%**

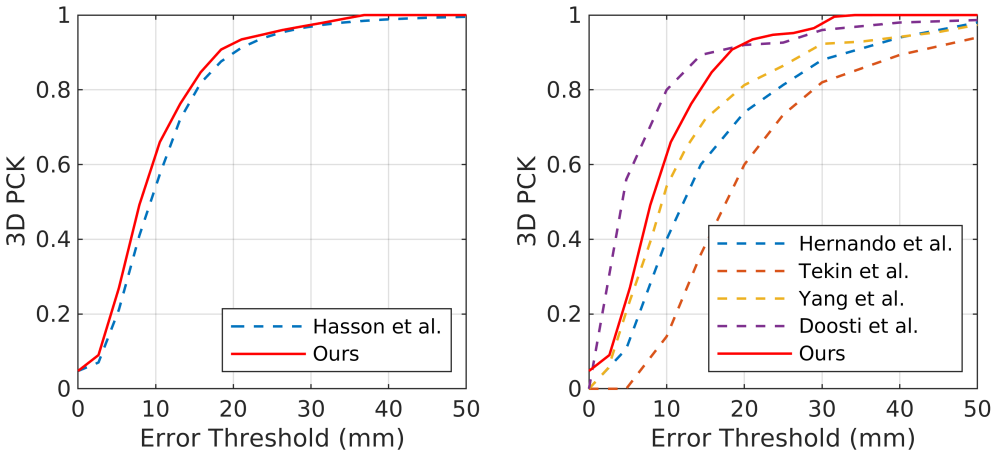


Figure 3.3: 3D PCK for *ObMan* (*left*) and *FPHA* (*right*) datasets. Note that Doosti et al., 2020 is a hand-object pose estimation method where known object is given.

P with associative loss and demonstrate that associative loss contributes to improving hand and object error. This can be expected since hand-object reconstruction are highly correlated such that learning in a collaborative manner enables performance boost to each other. The effectiveness of our proposed dynamic graph convolution can be demonstrated by the fast performance saturation at $P = 2$. Note that we took ObMan as our baseline and graph convolution is enabled from $P = 1$.

Comparison with static graph convolution. To motivate our dynamic graph convolution, we experiment with two commonly used graph convolution in Table 3.6, *i.e.* GCN (Kipf et al., 2017) and spiral mesh convolution (Gong et al., 2019). As the graph convolutions

Table 3.5: Error rates on *DexYCB* and Spurr et al. (2020) is the winner of HANDS 2019 Challenge (Armagan et al., 2020). Table indicates hand error (mm) with AUC values in parentheses. S0-S3 are the official dataset splits.

	S0	S1	S2	S3
Spurr et al. (2020)	17.34(0.698)	22.26(0.615)	**25.49(0.530)**	18.44(0.686)
Ours	**16.05(0.722)**	**21.22(0.620)**	27.01(0.521)	**17.93(0.698)**

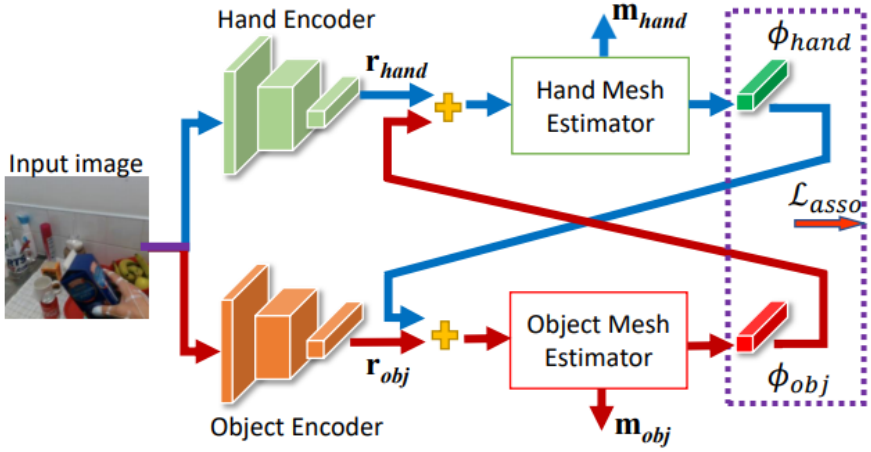


Figure 3.4: Simple collaborative learning framework design. Note that the yellow cross sign refers to addition.

weights are only updated after P iterations, increasing network iterations will have zero effects. It can be seen that static graph convolution does not benefit from increasing network iterations. We also observed that our unsupervised associative loss ($\mathcal{L}_{asso}$) consistently improves hand-object error across Table 3.6.

Effectiveness of associative loss ($\mathcal{L}_{asso}$). To further study the effect of our unsupervised associative loss $\mathcal{L}_{asso}$, we plot the training loss for the collaborative framework, with and without associative loss in Figure 3.5. Unsurprisingly, we find that increasing network iterations P contributes to a higher convergence rate (right of Figure 3.5). We also observe that $\mathcal{L}_{asso}$ is able to stabilise the training across all iterations (left of Figure 3.5). This shows that training with $\mathcal{L}_{asso}$ is crucial for this framework.

Table 3.6: Performances of different network design choices on *FPHA*$^-$. We experiment on network iterations P, associative loss $\mathcal{L}_{asso}$ and two different convolution operators (*i.e.* GCN (Kipf et al., 2017) and spiral (Gong et al., 2019). Note that the baseline on the first row is same as ObMan.

	w $\mathcal{L}_{asso}$		w/o $\mathcal{L}_{asso}$	
Method	Hand Error	Object Error	Hand Error	Object Error
Baseline	-	-	28.4	1655.2
Baseline ($P = 1$)	26.9	1600.3	27.4	1625.9
Baseline ($P = 2$)	**25.3**	**1445.0**	26.3	1618.4
Baseline ($P = 3$)	25.4	1448.2	26.4	1620.5
Baseline ($P = 4$)	25.3	1447.9	26.3	1612.9
Baseline ($P = 5$)	25.3	1445.6	26.2	1618.8
GCN ($P = 1$)	27.1	1587.6	27.8	1629.8
GCN ($P = 2$)	27.0	1590.8	28.2	1635.1
Spiral ($P = 1$)	26.8	1581.8	27.6	1630.1
Spiral ($P = 2$)	26.9	1600.2	27.6	1629.5

Mesh generation within iterations. We target the problem of mutual occlusion of interacting hand and object by sharing 3D information at each iteration via graph convolution. To validate this design choice, we construct a simpler collaborative learning framework which directly predicts embeddings ϕ_θ and reconstruct meshes m_θ at the final stage (see Figure 3.4). As *FPHA* has limited backgrounds and visible magnetic sensors, we compare the two design on *FPHA* and *DexYCB*. Table 3.7 shows that our final design consistently outperforms the naïve composition baseline across both datasets. We observe that sharing 3D mesh information across hand and object branches improves both reconstruction performance. At the bottom right of Figure 3.1, we provide a qualitative example of how reconstruction changes with graph convolution. It can be confirmed that our attention-guided graph convolution combined with collaborative learning enables better mesh quality as well as more accurate pose estimation. We provide additional qualitative results in Figure 3.6.

MANO pose representation. As described in Section 3.2.1, our hand branch outputs a 45-dimensional vector to represent the hand. We experiment with different dimensionality for the latent hand representation and summarise our findings in Table 3.8. We observe low-

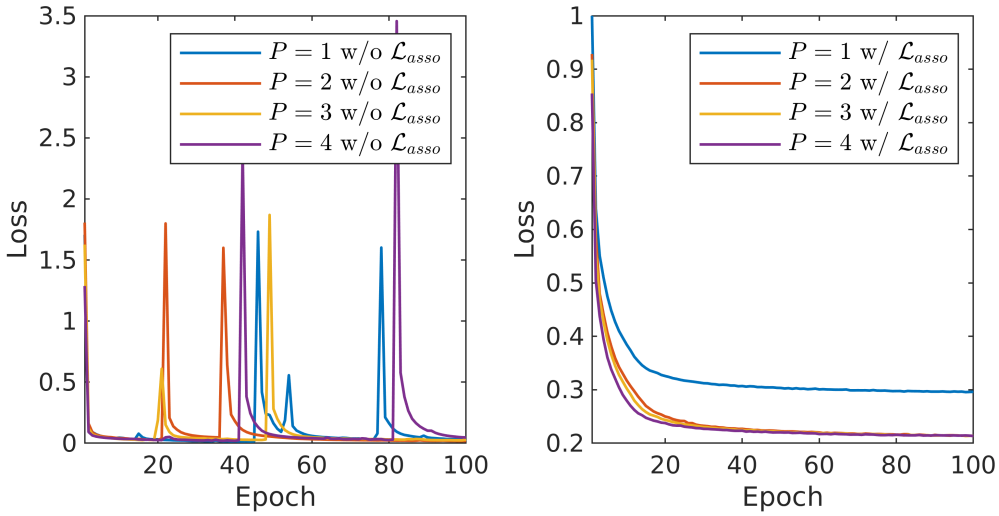


Figure 3.5: Progression of training losses for iterations $P = \{1, \ldots, 4\}$, without (*left*) and with (*right*) associative loss $\mathcal{L}_{asso}$.

Table 3.7: Ablation studies on collaborative learning framework design. We experiment on both *FPHA*$^-$ and the default *DexYCB* (S0) dataset split. * refers to the naïve collaborative learning baseline.

Method		*FPHA*$^-$		*DexYCB* (S0)	
		Hand Error	Object Error	Hand Error	Object Error
$P = 1$	Ours*	28.0	1759.4	17.9	563.4
	Ours	**26.9**	**1600.3**	**17.6**	**529.3**
$P = 2$	Ours*	27.6	1726.8	17.5	554.6
	Ours	**25.3**	**1445.0**	**16.1**	**461.1**
$P = 3$	Ours*	27.1	1678.1	17.3	542.1
	Ours	**25.4**	**1448.2**	**16.0**	**464.2**

dimensionality fails to capture some poses present in the datasets and full 45-dimensional vector is required to produce the best result.

MANO shape regularisation. Similarly, we observe that hand reconstruction performance increases with a larger saturated hand shape value than when it is trained with hand shape regularisation. We experiment with the loss on 3D joints ($\mathcal{L}_J$) and shape regularisation ($\mathcal{L}_\beta$). Table 3.9 shows that the hand reconstruction performance increases without shape regularisation ($\mathcal{L}_\beta$). As dense vertex supervision is not available in the real dataset *FPHA*$^-$,

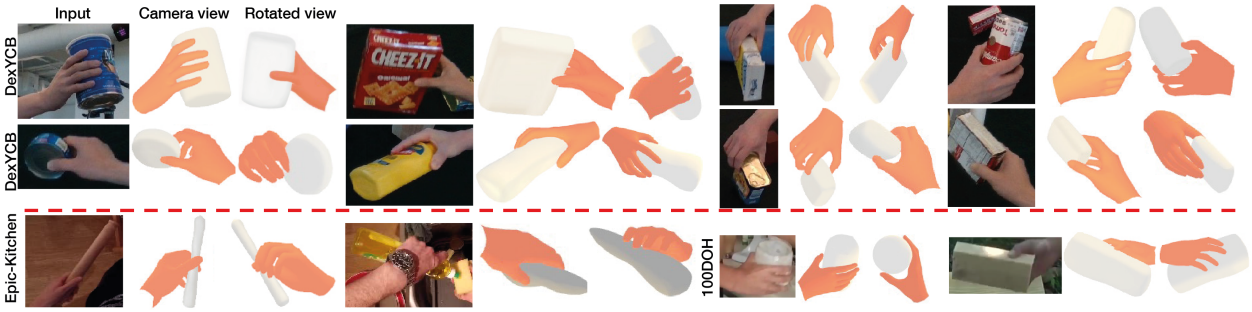


Figure 3.6: Qualitative results on *DexYCB* (top two rows), *EPIC-Kitchens* (Damen et al., 2018) (left of bottom row) and *100 Days of Hands* (100DOH) (Shan et al., 2020) (right of bottom row). The bottom row refers to in-the-wild settings. Our model, trained only on *DexYCB*, shows robustness to various hand poses, objects and scenes.

Table 3.8: We report the mean end-point error (mm) on *FPHA*$^-$ and *ObMan* to study the effect of the number of PCA hand pose components for the latent MANO representation.

PCA components	15	30	45
ObMan	11.7	9.6	**9.2**
FPHA$^-$	28.2	26.1	**25.3**

we omit experimenting on vertex loss $\mathcal{L}_V$.

Table 3.9: The mean end-point errors (mm) of two versions of our system that use 1) only the 3D joint loss ($\mathcal{L}_J$) and 2) a combination of the joint loss and shape regularisation ($\mathcal{L}_J + \mathcal{L}_\beta$). For both *ObMan* and *FPHA*$^-$, low errors are measured when shape regularisation is disabled.

	ObMan	*FPHA*$^-$
$\mathcal{L}_J$	**9.2**	**25.3**
$\mathcal{L}_J + \mathcal{L}_\beta$	10.3	27.5

Multi-head attention. As described in Section 3.2.3, we found multi-head attention to be beneficial. We experiment with different number of heads and summarise our findings in Table 3.10. We used multi-head attention $K = 3$ in all experiments as it provides the best performance.

Table 3.10: We report the mean end-point error (mm) on *FPHA*$^-$ and *ObMan* to study the effect of the number of multi-head attention mechanism.

#heads	1	2	3	4	5
ObMan	12.4	11.4	**9.2**	9.3	9.6
FPHA$^-$	28.7	26.8	**25.3**	25.5	25.4

3.4 Summary

In this chapter, we introduced a novel collaborative learning framework that addresses the challenge of mutual occlusion in hand-object interactions. Our approach enables the iterative sharing of mesh information between the hand and object branches, leveraging the highly correlated nature of estimating the pose of interacting hands and objects. To capture long-range dependencies from the dynamic graph, we proposed an attention-guided graph convolution, which effectively incorporates mesh information in a single layer.

One of the key contributions of our work is the introduction of an unsupervised associative loss that stabilizes the training process and enhances the feature transferring between the branches. By incorporating this loss, we observed improved performance compared to existing approaches on several widely-used datasets, highlighting the effectiveness of our collaborative learning framework.

However, our work had a few limitations that should be addressed in future research. Firstly, our reliance on AtlasNet for object reconstruction resulted in varying reconstruction quality based on the size of the training data. Exploring alternative or complementary techniques for object reconstruction could potentially mitigate this limitation and improve the overall quality of object reconstructions.

In addition, our current approach focused on static objects, and we did not consider the interaction between hands and articulated objects. Future works should aim to extend

our framework to handle the complexities introduced by articulated objects, as these scenarios are common in real-world hand-object interactions. By incorporating articulated objects, we can further enhance the realism and applicability of our collaborative learning framework.

Addressing these limitations will contribute to the advancement of our collaborative learning approach and enable more robust and accurate hand-object interaction analysis in various real-world scenarios.

Chapter Four

LEARNING TWO-HAND RECONSTRUCTION FROM EGOCENTRIC MULTI-VIEW RGB IMAGES

This chapter presents work published at the 2023 International Conference on Computer Vision (ICCV) in Paris, France (Tse et al., 2023).

In Chapter 3, we addressed a challenging single-view problem setting, but it was not without limitations. Firstly, our approach was limited to scenarios involving a single hand and object, which restricted its applicability to more complex interactions. Additionally, although our learning-based approach performed well on benchmarks even without the use of contact loss terms, there were instances of physically implausible reconstructions during test time. This suggests that our model did not fully learn the physical constraints of the real world.

In this chapter, our focus shifts towards tackling an even more challenging problem setting: the reconstruction of two hands in an egocentric view setting, with an emphasis on absolute root pose recovery for the applications on AR/VR. This is motivated by the fact that understanding two-hand interactions from an egocentric view is crucial for AR/VR applications as it enhances realism and user engagement by aligning virtual objects with the user's real-world perspective. Our objective is to achieve high-fidelity reconstruction of both hands, including the extended forearms. By doing so, we aim to provide a more realistic and accurate representation of hand movements and gestures compared to the parametric hand model MANO. This expansion of our reconstruction scope to encompass two hands and absolute root pose recovery presents new complexities and opportunities for improving the realism and accuracy of our reconstructions.

Several recent methods have been developed in response to the availability of the InterHand2.6M dataset (Moon et al., 2020c). These methods aim to address the self-similarity issue between interacting hands. Some of the more recent approaches tackle this problem by leveraging hand part segmentation probability (Z. Fan et al., 2021), joint visibility (D. U. Kim et al., 2021), cascaded refinement modules (Baowen Zhang et al., 2021), or keypoints using Transformer (Hampali et al., 2022). While these methods demonstrate good performance on complex hand configurations, they rely on root joint alignment and do not provide absolute root pose recovery in multi-view scenarios. This limitation becomes particularly critical for interactions in VR settings where precise root pose estimation is essential.

In addition, accurate estimation of two hands, including the extended forearms, is crucial for many immersive AR/VR applications. This comprehensive representation enables a more realistic and precise portrayal of hand movements and gestures within the virtual environment. By incorporating the forearm, the orientation and movement of the hand relative to the arm can provide valuable contextual information for the user's actions in the

virtual space. Moreover, including the forearm in the tracking system can help reduce errors and enhance the stability of the overall tracking process, which is essential for maintaining a high level of immersion and preventing disorientation in AR/VR applications (Han et al., 2020). However, similar to other pose estimation areas, there is currently a lack of suitable datasets for supervised deep learning approaches in this domain. Existing datasets either do not support egocentric views due to feasibility constraints (Garcia-Hernando et al., 2018) or lack the variations in background and lighting conditions (Moon et al., 2020c) as they were captured under constrained laboratory settings. Also, accurately estimating the pose of two closely interacting hands with extended forearms remains challenging. In such scenarios, one hand often occludes the other, making it difficult to achieve precise hand tracking and motion estimation (D. U. Kim et al., 2021).

In this paper, we make the following contributions: 1) We present a novel end-to-end trainable spectral graph-based transformer for high fidelity two-hand reconstruction from multi-view RGB image. 2) We design an efficient soft attention-based multi-view image feature fusion in which the resulting image features are region-specific to segmented hand mesh. We further demonstrate a minimal reduction of 35% in the model size with this approach. 3) We introduce an optimisation-based method to refine physically-implausible meshes at inference. 4) We create a large-scale synthetic multi-view dataset with high resolution 3D hand meshes and collect real dataset to verify our proposed method.

We first provide preliminaries in hand pose estimation and existing hand datasets in Section 4.1. We then describe our proposed model and method for leveraging the properties of graph Laplacian from spectral graph theory in Section 4.2. Finally, we present empirical evidence to show the strength of our proposed method in Section 4.3 and conclude the chapter in Section 4.4.

4.1 Preliminary

Our work tackles the problem of two-hand reconstruction from multi-view colour images. We first introduce the literature on two-hand pose estimation in Section 4.1.1. Then, we focus on the line that leverages Transformer on human body/hand modelling in Section 4.1.2. Finally, we provide a brief on existing hand datasets which motivates us to create a large-scale egocentric dataset in Section 4.1.3.

4.1.1 Two-hand pose estimation

The task of pose estimation for interacting hands can be broadly categorised into discriminative and generative (or hybrid) approaches. Initially, hybrid approaches were commonly employed, where visual cues detected by discriminative methods were utilised, followed by model fitting. For example, both Ballan et al. (2012) and Tzionas et al. (2016) incorporated collision optimisation terms and physical modeling based on detected fingertips. Other approaches (Han et al., 2020; Mueller et al., 2019; Jiayi Wang et al., 2020; Han et al., 2022) extracted image features or keypoints from RGB or depth images and fitted hand models with physical constraints. With the introduction of recent large-scale datasets focusing on interacting hands (Moon et al., 2020c), fully discriminative methods have emerged. These methods, including D. U. Kim et al. (2021), M. Li et al. (2022), Hampali et al. (2022), Rong et al. (2021), and Baowen Zhang et al. (2021), jointly estimate the 3D joint locations or hand model parameters directly from a single RGB image.

4.1.2 Transformer in 3D vision

Transformer-based architectures have gained significant popularity in the field of computer vision. In this context, we specifically focus on methods that reconstruct the human body or hands from RGB images, and we recommend referring to Khan et al. (2022) for a comprehensive survey on this topic. In a closely related work, K. Lin et al. (2021a) employs cascaded Transformer encoders to reconstruct the human body and hands from a single RGB image, achieving state-of-the-art performance. Building upon this work, K. Lin et al. (2021b) extends the approach by incorporating graph convolutions into the Transformer encoder. Furthermore, J. Cho et al. (2022) enhances the efficiency of K. Lin et al. (2021a) by disentangling image encoding and mesh estimation through an encoder-decoder architecture. Hampali et al. (2022) extends "Detection Transformer" (Carion et al., 2020) with hand-object pose estimations. While these existing works primarily focus on single-image reconstruction, extending them to the multi-view setting is challenging due to the large number of learnable parameters involved. In contrast, our proposed architecture is specifically designed to efficiently reconstruct two hands at a high resolution, thereby addressing the unique requirements of this particular scenario.

4.1.3 Hand pose datasets

The effectiveness of discriminative methods heavily relies on the availability and diversity of datasets capturing hands interacting with each other. While the InterHand2.6M dataset (Moon et al., 2020c) offers a large-scale collection of closely interacting hand motions, it lacks significant variation in background and lighting conditions. On the other hand, FreiHAND (Zimmermann et al., 2019) provides more extensive background variation, but it is limited to single-hand scenarios and third-person views. Consequently, these datasets do not adequately address the requirements of the egocentric two-hand reconstruction task.

Alternatively, as discussed in Section 2.3.1, the available egocentric datasets are either constrained by visible markers (Garcia-Hernando et al., 2018) or limited to controlled lab environments (Kwon et al., 2021). This motivates us to create a large-scale egocentric synthetic dataset with improved environment and lighting variations. To validate our proposed multi-view fusion strategy, we further collect a real dataset with more challenging camera viewing angles. Altogether, both synthetic and real datasets contain diverse egocentric multi-view and -frame data points with multiple subjects.

4.2 Methodology

We build on the state-of-the-art Transformaer-based model for 3D two-hand reconstruction from egocentric multi-view colour images. Differently from Chapter 3, we target at multi-view problem setting with the focus on egocentric views to simulate the practical scenario for immersive AR applications. The key idea of our approach is to first aggregate multi-view image features and upsample features to target mesh resolution.

In the following, we first present the overview of our proposed method and mathematical notations in Sections 4.2.1 and 4.2.2. We then detail the multi-view image feature encoder in Section 4.2.3, the spectral graph convolution decoder in Section 4.2.4 and the loss used for training in Section 4.2.5. We also present an optimisation-based refinement procedure at inference time to produce physically plausible reconstructions in Section 4.2.6.

4.2.1 Overview

As shown in Figure 4.1 , our architecture first passes N number of multi-view RGB input images $\mathbf{x} \in \mathbb{R}^{N\times 224\times 224\times 3}$ to a shared CNN backbone to extract volumetric features $\mathbf{f} \in$

$\mathbb{R}^{N\times7\times7\times2048}$, *i.e.* features before the global average pooling layer for ResNet (He et al., 2016). The volumetric features are then fed into a soft attention-based multi-view feature encoder and output K region-specific features $\mathbf{f}_r \in \mathbb{R}^{K\times C}$ where C refers to the feature channel size. The Transformer encoder takes $\mathbf{f}_r$ together with template hand meshes $\mathbf{m}' \in \mathbb{R}^{V'\times3}$ and outputs a coarse mesh representation $\mathbf{f}_c \in \mathbb{R}^{V'\times F}$. Finally, the spectral graph decoder generates hand meshes $\mathbf{m} \in \mathbb{R}^{V\times3}$ by upsampling on $\mathbf{f}_c$ where $V \gg V'$. With slight abuse of notation, $\mathbf{m}$ can either be two-hand or single-hand depending on the application.

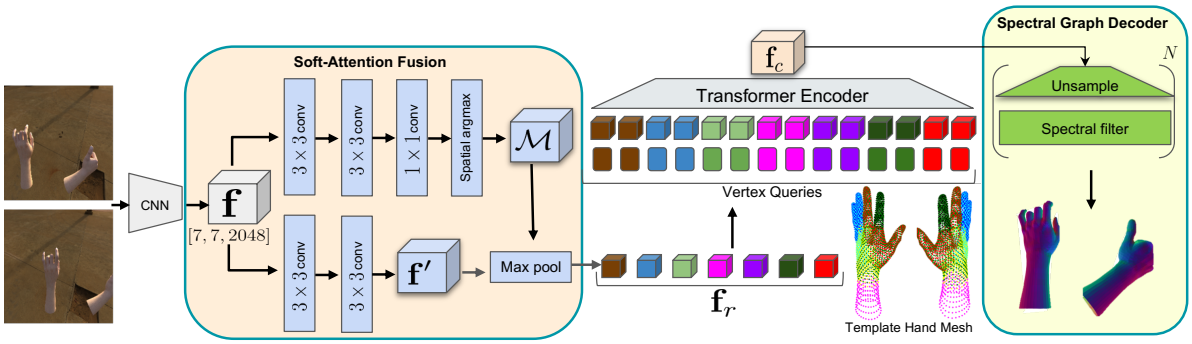


Figure 4.1: A schematic illustration of our framework. Given multi-view RGB images, we extract volumetric features $\mathbf{f}$ with a shared CNN backbone. The soft-attention fusion block generates the attention mask $\mathcal{M}$ and finer image features $\mathbf{f}'$ through multiple upsampling and convolution blocks. Region-specific features $\mathbf{f}_r$ are computed by first aggregating $\mathbf{f}'$ along the feature channel dimension via the attention mask $\mathcal{M}$, followed by a max-pooling operation across multi-view images to focus on useful features. Then, we apply mesh segmentation via spectral clustering on template hand meshes and uniformly subsample them to obtain coarse meshes. We perform position encoding by concatenating coarse template meshes to the corresponding region-specific features $\mathbf{f}_r$, *i.e.* matching colored features to mesh segments. Finally, our multi-layer transformer encoder takes the resulting features as input and outputs a coarse mesh representation $\mathbf{f}_c$ which is then decoded by a spectral graph decoder to produce the final two-hand meshes at target resolution. Here, each hand contains 4023 vertices.

4.2.2 Graph Laplacian

A 3D mesh can be represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ is a node set and $\mathcal{E}$ is an edge set. An adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ encodes information of pairwise relations between nodes. A degree matrix $\mathbf{D} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is a diagonal matrix whose diagonal

element $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$ refers to the degree value of each node. An essential operator in spectral graph theory (Chung, 1997) is the graph Laplacian $\mathbf{L}$, whose definition is $\mathbf{L} = \mathbf{D} - \mathbf{A}$, and $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where the graph Laplacian can be diagonalised by the Fourier basis $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{|\mathcal{V}|}] \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{\Lambda} = diag([\lambda_1, \ldots, \lambda_{|\mathcal{V}|}]) \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ where $\{\mathbf{u}_i\}_{i=1}^{|\mathcal{V}|}$ are the eigenvectors and $\{\lambda_i\}_{i=1}^{|\mathcal{V}|}$ are the non-negative eigenvalues of graph Laplacian ($0 = \lambda_1 \leq \ldots \leq \lambda_{|\mathcal{V}|}$).

4.2.3 Multi-view image feature encoder

In order to adapt the Transformer-based architecture to multi-view settings and prevent concatenating global image features in an overly-duplicated way, as observed in K. Lin et al., 2021a, we introduce a simple soft-attention fusion strategy. This approach is designed to enhance the aggregation of features across multiple views and enable selective attention to different hand parts through mesh segmentation. The output of this process is a set of K region-specific features denoted as $\mathbf{f}r$, which are identified using spectral clustering. These region-specific features are then fed into our Transformer encoder, allowing us to obtain a coarse mesh representation denoted as $\mathbf{f}c$.

Soft-attention fusion. Given volumetric features $\mathbf{f}$ from the shared CNN backbone, we do not apply any pooling operations to avoid losing spatial information. Instead, we aggregate multi-view features with a soft-attention mask. We first obtain a finer representation of $\mathbf{f}$, denoted as $\mathbf{f}' \in \mathbb{R}^{N\times(H\times W)\times C}$, by feeding it through two blocks each comprised of 2D upsampling with bilinear interpolation, 3×3 convolution layers, batch-normalisation (Ioffe et al., 2015) and ReLU. The soft-attention mask $\mathcal{M} \in \mathbb{R}^{N\times(H\times W)\times K}$ is obtained by applying 1) K 1×1 convolutional filters to reduce the feature channel of $\mathbf{f}'$ to K and 2) spatial soft arg-max which determines the image-space point of maximal activation in each C. At this stage, we compute per-frame features $\mathbf{f}'' \in \mathbb{R}^{N\times K\times C}$ by $\mathbf{f}'' = \mathcal{M}^{\mathrm{T}}\mathbf{f}'$. We finally obtain

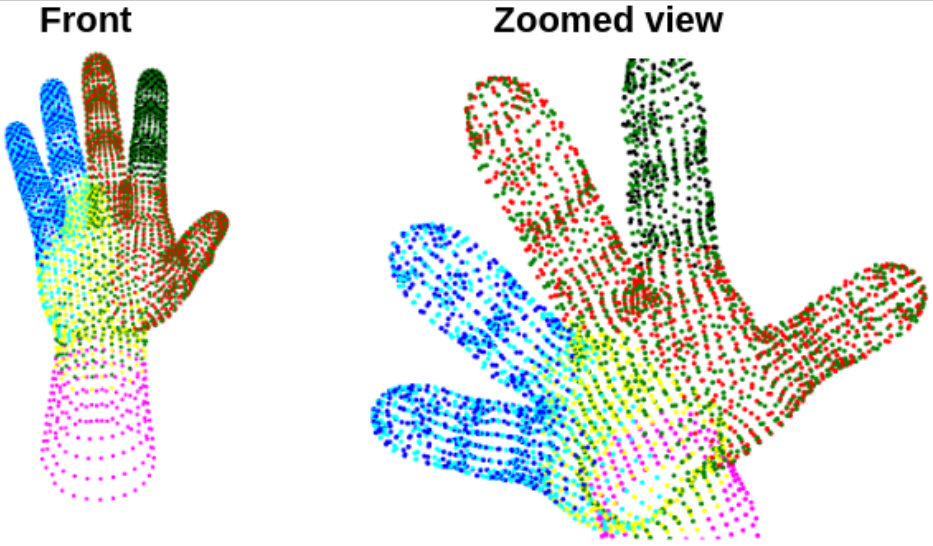


Figure 4.2: Illustration of mesh segmentation via spectral clustering. Here, we have chosen $K = 7$ clusters and each has its own color for a single right-handed template mesh. As shown, there are no rigid boundaries across the mesh. To prepare inputs $\mathbf{f}_t$ to transformer, we subsample this mesh (same for left hand) uniformly by a factor of 10, *i.e.* $V' = V/10$ and concatenate with the corresponding region-specific features $\mathbf{f}_r$.

region-specific features $\mathbf{f}_r \in \mathbb{R}^{K \times C}$ by max-pooling along the second dimension of $\mathbf{f}''$. There are two intuitions to this design: 1) the corresponding feature with higher attention weight contributes more to the final feature representation spatially and 2) max-pooling allows efficient feature selection across multiple views and it does not overfit to any multi-camera configurations as CNN filters are shared across all views.

Mesh segmentation via spectral clustering. We now perform spectral clustering to obtain a 3D mesh segmentation by applying the eigen-decomposition of the graph Laplacian $\mathbf{L}$ instead of the affinity matrix (which encodes pairwise point affinities with exponential kernel), followed by k-means clustering into K clusters. With this approach, we can efficiently segment any template model mesh without manual effort. In addition, as shown in Figure 4.2, the segmented mesh does not exhibit clear-cut boundaries, and certain clusters are scattered throughout the entire surface. This is in spirit similar to mask vertex modeling occlusions in K. Lin et al., 2021a by encouraging the transformer to consider other relevant vertex queries.

Transformer encoder. We then concatenate the C-dimensional region-specific image features to the corresponding K clusters of the segmented template hand mesh. The resulting features $\mathbf{f}_t \in \mathbb{R}^{V' \times (C+3)}$ [1] are fed into a multi-layer transformer encoder with progressive dimensionality reduction. The output is a coarse mesh representation $\mathbf{f}_c \in \mathbb{R}^{V' \times F}$, where $F = (C+3)/2^n$ with n layers of transformer encoder.

Theoretical motivations. Since there do not exist explicit coordinate systems as in grid graphs, aligning nodes of highly irregular graphs in nature is a non-trivial problem. Recent studies (Dwivedi et al., 2022; Dwivedi et al., 2021; Kreuzer et al., 2021) attempted to encode positional information by leveraging the spectral domain in a way that nearby nodes have similar values and distant nodes have different values. This can be achieved as the eigenvectors of the graph Laplacian can be interpreted as the generalised concepts of sinusoidal functions of positional encoding in Transformers (Vaswani et al., 2017). For instance with any given graph, nodes close to each other can be assigned values of similar positional features as the smaller the eigenvalue of the graph Laplacian λ_i (closer to 0) the smoother the coordinates of the corresponding eigenvector $\mathbf{u}_i$.

4.2.4 Spectral graph decoder

As illustrated in Figure 4.3, we find that simply relying on fully-connected layers to upsample meshes to target resolution is insufficient as this process introduces instability and disruption to the mesh. Therefore, we couple the fully-connected layers for upsampling with spectral filtering as meshes can be treated as graph signals $\mathbf{f}_c = (f_1, \cdots, f_{V'}) \in \mathbb{R}^{V' \times F}$, *i.e.* V' vertices with F-dimensional features for batch size of 1.

Spectral filtering. As spectral graph convolution can be defined via point-wise products

[1] +3 refers to the 3D positions of mesh template.

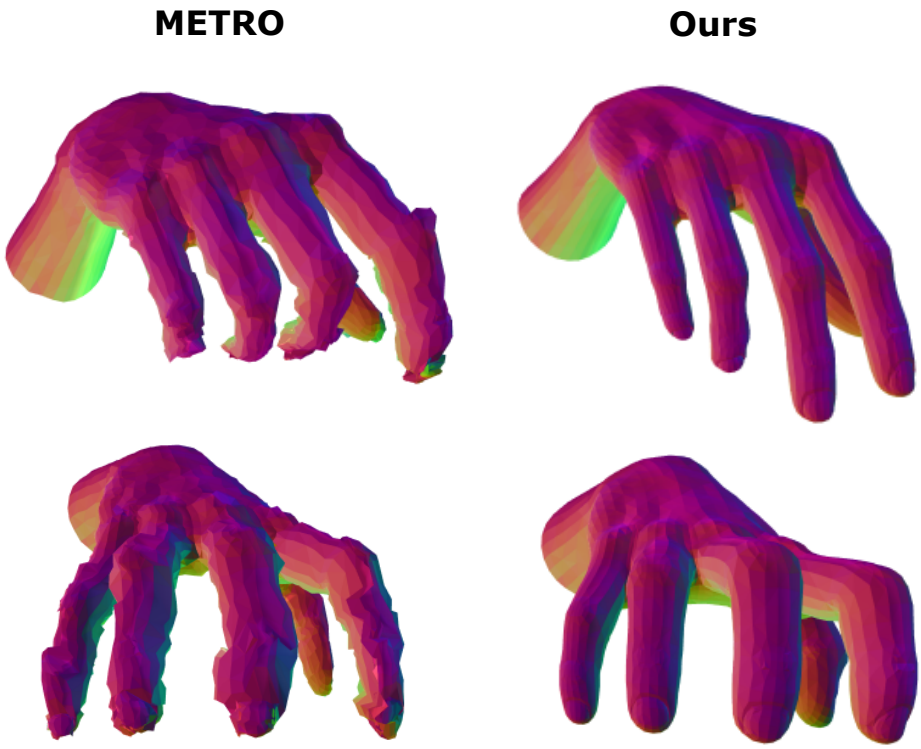


Figure 4.3: Qualitative comparison with METRO (K. Lin et al., 2021a). We show that relying on fully-connected layers to upsample meshes is inadequate for high-resolution mesh reconstruction. In contrast, our spectral graph decoder can accurately capture intricate surface features like nails.

in the transformed Fourier space, graph signals $\mathbf{f}_c$ filtered by g_θ can then be expressed as $\mathbf{U}g_\theta(\mathbf{\Lambda})\mathbf{U}^T\mathbf{f}_c$ where $g_\theta(\mathbf{\Lambda})$ is a diagonal matrix with Fourier coefficients. To ensure that the spectral filter corresponds to a meaningful convolution on the graph, a natural solution is to parameterise based on the eigenvalues of the Laplacian. In this work, we use the Chebyshev polynomial parametrisation of $g_\theta(\mathbf{L})$ for fast computation and readers are referred to Defferrard et al., 2016 for more details.

Architecture. Our decoder is based on a hierarchical architecture where the mesh is recovered by using fully-connected and graph convolution layers for upsampling. We followed Ge et al. (2019) and Defferrard et al. (2016) to pre-computed coarse graphs and used the third-order polynomial in the Laplacian. Given the coarse mesh representation $\mathbf{f}_c \in \mathbb{R}^{V' \times F}$, our key idea here is to smooth out the F-dimensional values after upsampling on V'. The fundamental difference between our approach and the previous ones (Ge et al., 2019; Choi et al., 2020) is that they have a hierarchical architecture on both V' and F dimensions, we apply only spectral filtering and keep $F = 3$ constant which massively reduces the model size while maintaining performance.

Discussions. Recall that in Section 4.2.3, we leverage the properties of Laplacian eigenvectors to perform spectral clustering. Here we can interpret from a signal processing perspective where the Laplacian eigenvectors define signals that vary smoothly across the graph, with the smoothest signals indicating the coarse community structure of the mesh.

4.2.5 Training

Our model can be trained end-to-end with L1 losses on 3D mesh vertices and 2D re-projection using the predicted camera parameters to improve image-mesh alignment. In addition, we

apply an edge length regularisation $\mathcal{L}_{edge}$ to encourage smoothness of the mesh:

$$\mathcal{L}_{edge}(\mathbf{m}) = \frac{1}{|\mathcal{E}_L|} \sum_{l \in \mathcal{E}_L} |l^2 - \mu(\mathcal{E}_L^2)|, \tag{4.1}$$

where $\mathcal{E}_L$ refers to the set of edge lengths, defined as the L2 norms of all edges and $\mu(\mathcal{E}_L^2)$ is the average of the squared edge lengths.

4.2.6 Mesh refinement at inference

There are two main approaches in the literature for producing realistic mesh reconstruction: learning-based and optimisation-based methods. Learning-based methods (Hasson et al., 2019; Moon et al., 2020b) approach the problem by proposing various repulsive losses that penalise penetration during training. However, their generalisation ability to other meshes is limited due to the need for mesh-specific pre-computation. In particular, they require manual selection of areas of interest (*i.e.* fingertips and palm) which does not prevent other forms of self-penetrations, such as finger-finger. More importantly, the presence of interpenetration at test time shows that the model is unable to learn the physical rule implicitly. On the other hand, recent optimisation-based methods (Grady et al., 2021; Tse et al., 2022b) leverage contact maps to refine meshes at inference. However, they rely heavily on accurate contact map estimations and are sensitive to initialisation as contact optimisation is local.

To overcome these limitations, we extend the repulsion loss from Hasson et al. (2019) into an optimisation-based strategy which does not require any form of pre-computation and is more generalisable to other meshes. To identify hand vertices that contribute to collision, we cast rays from each vertex and count the number of surface intersections. If the number is odd, it indicates penetration. After obtaining the collision mask $\mathcal{M}_C$, we compute the nearest point in the source mesh $\mathbf{m}$ with respect to the set of collision points. If the point

and its nearest corresponding point have a different normal, we compute their distance as loss. We minimise this collision loss $\mathcal{L}_{collision}$ with as-rigid-as-possible (ARAP) Sorkine et al. (2007) regularisation on the mesh shape:

$$\mathcal{L}_{collision}(\mathbf{m}, \mathcal{M}_C) = \sum_{v \in \mathcal{V}} \mathcal{M}_C \cdot d(v, \mathcal{V}), \tag{4.2}$$

where $\mathcal{M}_C = \mathbb{1}_{v \in \text{Int}(\mathcal{V})}$ indicates which vertices belong to the interior of the mesh and $d(v, \mathcal{V}) = \min_{v' \in \mathcal{V}} \|v - v'\|_2$ denotes distances from point v' to set $\mathcal{V}$. We show that our method is able to remove self-penetration in Figure 4.4.

4.2.7 Datasets

As discussed in Section 4.1.3, there is a lack of high fidelity two-hand dataset for training. In the following, we describe how we create a large-scale synthetic multi-view dataset as well as collecting a real-world data to validate our approach.

Synthetic dataset creation. A large-scale multi-view egocentric dataset with challenging interacting hand motion is required to train our pipeline. However, existing datasets are either infeasible for egocentric views or lack variations in terms of background and lighting conditions, which are crucial for AR/VR applications. In addition, the ground-truth mesh in InterHand2.6M dataset (Moon et al., 2020c) contains $5mm$ fitting error which is sufficient to cause inaccuracy in hand tracking or misalignment when interacting with virtual objects. Therefore, it is not suitable to validate our approach. To this end, we create a new large-scale synthetic egocentric dataset with two high-fidelity 3D hand meshes. In particular, we purchase commercial 3D hand models for both left and right hands. Each of them contains 4023 vertices and 4008 quad faces. By dividing each quad into two triangles, they can be decomposed into 8016 triangular faces. The size and textures of the hand models can vary,

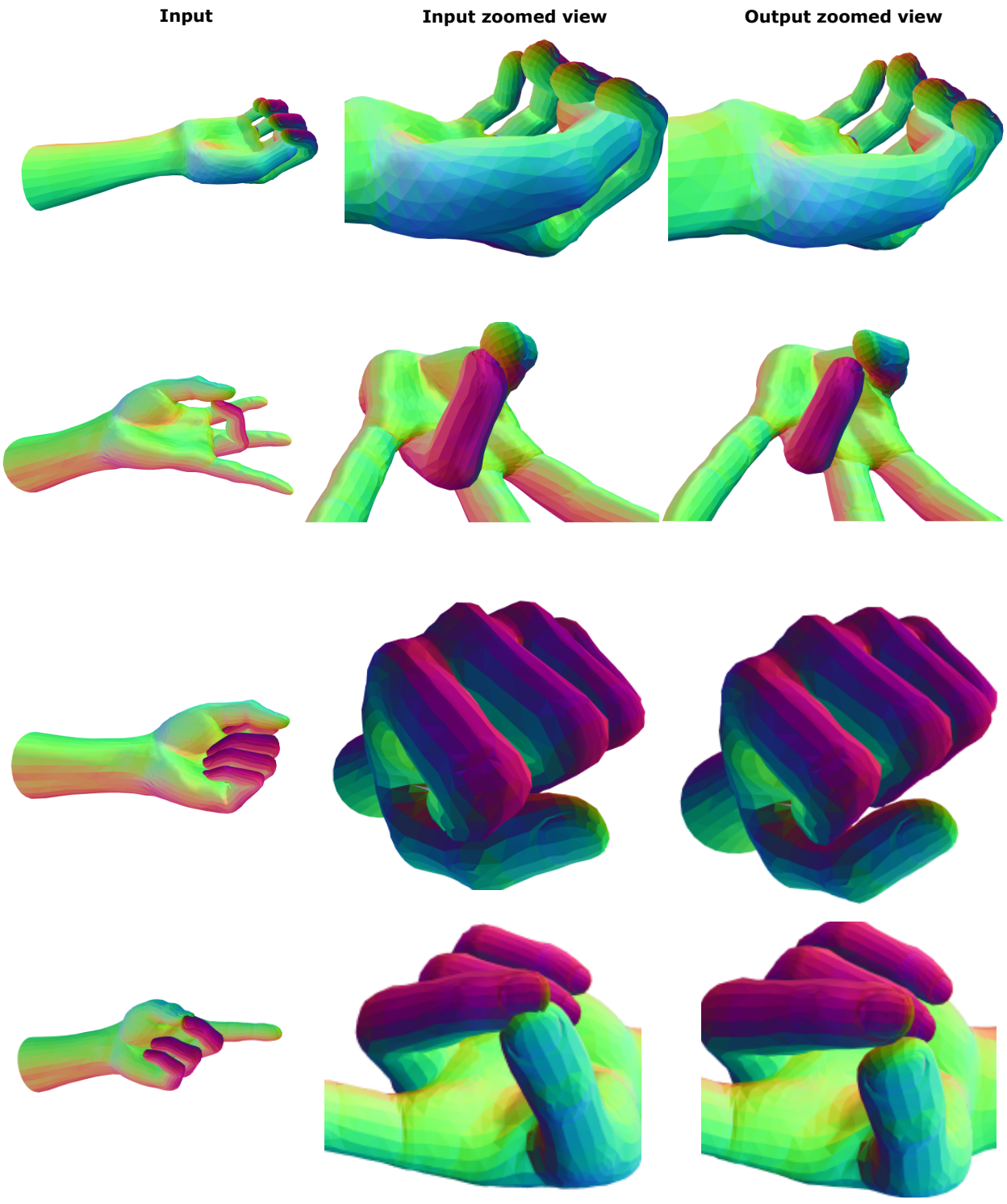


Figure 4.4: Qualitative examples of mesh refinement at inference. Our optimisation-based strategy shows robustness to various hand poses.

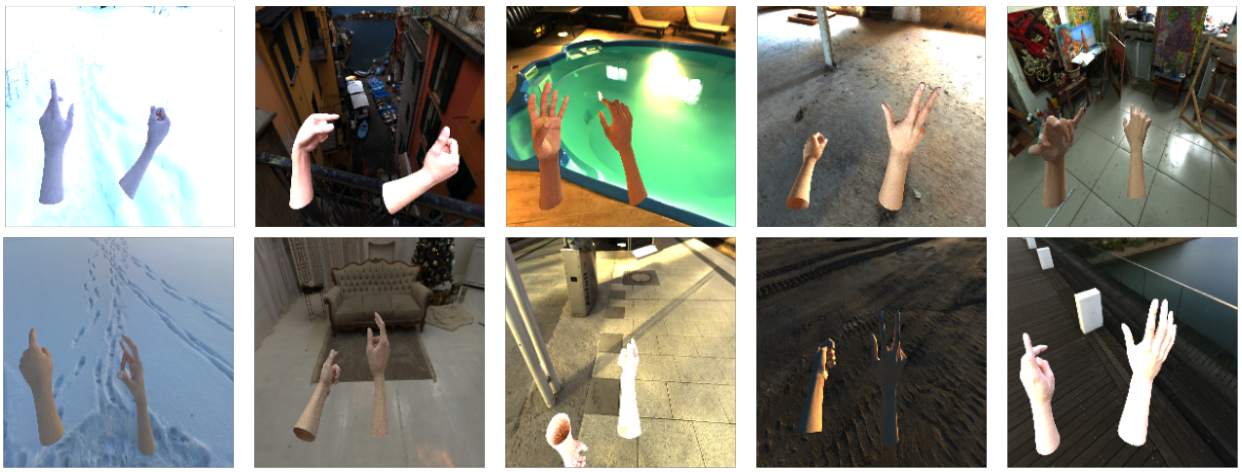


Figure 4.5: Qualitative examples of our synthetic dataset. The scenes are rendered with large variations in lighting, texture and background.

and it can also be rigged with 3D joint locations. We apply photorealistic textures as well as natural lighting using High-Dynamic-Range (HDR) images. To create realistic hand motions, we generate 100 poses and interpolate between pairs of these poses randomly over 1000 frame sequences, *i.e.* pose pair is swapped every 10 frames. Then, we render hands onto 480 4K backgrounds using the Cycles renderer (Blender, 2023). Our synthetic dataset comprises 1M data points. Each data point consists of 3D annotations for two hands, rendered in two egocentric views to simulate a customised camera headset scenario. We divide the entire dataset into 4 groups and use 25% randomly for testing and provide image examples in Figure 4.5.

Real dataset collection. Our real hand data is captured from a multi-view stereo system with 18 synchronised Z-Cams. We build a NeRF-based (Barron et al., 2022; Müller et al., 2022) reconstruction pipeline that simultaneously reconstructs and disentangles the foreground and the background, from which a meshing module extracts hand meshes as ground-truth target mesh. We then apply a mesh and tetrahedral registration approach (Smith et al., 2020) that registers a template hand model (H. Xu et al., 2020) to get a well-registered mesh for each reconstruction. We provide image examples in Figure 4.8.

4.3 Experiment

In this section, we first describe our implementation details in Section 4.3.1 and the datasets in Section 4.2.7. We then describe the corresponding evaluation protocols in Sections 4.3.2 and 4.3.3. Finaly, we validate our approach numerically and qualitatively in Sections 4.3.4 and 4.3.5.

4.3.1 Implementation details

We train all parts of the network simultaneously with the Adam optimiser (Kingma et al., 2015) using a learning rate of 10^{-4} when training on the synthetic dataset and 10^{-5} when fine-tuning on the real dataset. We use ResNet (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) for our CNN backbone. For computing region-specific features $\mathbf{f}_r$, we set $K = 7$ and $C = 256$.

4.3.2 Baselines

We compare with the state-of-the-art Transformer-based architecture METRO (K. Lin et al., 2021a) that is explicitly designed for human body mesh reconstruction. We follow their official implementation and use 3 transformer encoder layers. We attempted to include other strong baselines to compare with. However, given our challenging multi-view settings on reconstructing high resolution two-hand meshes, existing methods are constrained by either 1) relying on intermediate 3D pose supervision (Moon et al., 2020a; Choi et al., 2020) or 2) directly regressing a low resolution parametric hand model (Hasson et al., 2019; Zimmermann et al., 2019; Baek et al., 2019; Boukhayma et al., 2019). In addition, the direct extension of METRO, Mesh Graphormer (K. Lin et al., 2021b), is infeasible to extend to multi-view

settings as they tokenised volumetric features which increases the computational complexity of each transformer layer quadratically (J. Cho et al., 2022). Therefore, by considering that the performance gain is minor, we pick METRO as a strong baseline and study in-depth.

4.3.3 Evaluation metric

We report the Mean-Per-Vertex-Error (MPVE) (K. Lin et al., 2021a) in mm to evaluate the hand reconstruction error. MPVE measures the mean Euclidean distances between the ground-truth vertices and the predicted vertices.

4.3.4 Results

We perform a quantitative comparison for the two-hand reconstruction task on our synthetic dataset. As previously mentioned in Section 4.3.2, comparing with other methods is not straightforward since the majority of existing methods focus on single-view settings. We extend METRO (K. Lin et al., 2021a) to the multi-view setting by applying max-pool to image features and concatenating them with vertex queries. For fair comparisons, we share the same hyperparameter setting for the multi-layer transformer encoder and report the results in Table 4.1. In this experiment, we downsample the template hand mesh by 10 times. To provide more context regarding the difficulty of our synthetic dataset, we also experiment on a parametric baseline. We create a parametric hand model which uses the 200-dimensional PCA subspace from 25% of the training data. We use ResNet-50 as the backbone for all models in Table 4.1 and the parametric baseline has an MLP head to predict the input parameters to recover the hand mesh. Our method significantly outperforms both baseline methods with less than half the model size of METRO. We show that our soft-attention feature fusion strategy coupled with mesh segmentation using spectral clustering can effectively

reduce the feature channel size from 2048 to 256 without performance drop. In addition, the performance of the parametric baseline is in-line with existing single-hand benchmarks. Furthermore, we report results on single-view benchmark FreiHAND (Zimmermann et al., 2019) in Table 4.2. Lastly, we provide additional qualitative examples in Figures 4.6 and 4.7.

Table 4.1: Error rates on our synthetic dataset. Parametric baseline is trained and tested on right hand only. Our proposed method achieves strong performance with less than half the METRO model size.

	Hand error	# Params
Parametric baseline	24.5	40.9M
METRO	7.09	116.1M
Ours (w/o graph decoder)	3.72	75.2M
Ours	**1.38**	58.3M

Table 4.2: Error rates (in *mm*) on *FreiHAND* dataset.

	METRO	Graphormer	Ours
PA-MPVPE ↓	6.7	5.9	**5.5**
PA-MPJPE ↓	6.8	6.0	**5.6**

4.3.5 Ablation study

To motivate our design choices, we present a quantitative evaluation of our method with various components disabled. We validate that each of our proposed technical component contributes meaningfully.

Effects of multi-view feature fusion. Table 4.3 shows the results of varying number of spectral clusters K (full results in supp.). The combination of soft-attention fusion and mesh segmentation consistently improves the performance. As our synthetic dataset contains only two views, we further verify our multi-view fusion strategy on more challenging viewing

Figure 4.6: Additional qualitative examples on our real dataset.

Figure 4.7: Additional qualitative examples on our synthetic dataset.

angles captured in our real dataset. We also demonstrate that our method does not overfit to camera setup when tested on unseen camera views in Table 4.4. In these experiments, we divide the 18 camera views into 2 groups, *i.e.* the first group contains the first 15 camera views and the second group contains the last 3 unseen camera views for evaluation. Note that the evaluation group contains the only egocentric view. We consider 3 experimental settings by varying number of camera views present in the training data: (a) 15 views, (b) 6 views and (c) 3 views in Table 4.4.

Table 4.3: Performance of different multi-view fusion strategies. We report hand error for both settings. K refers to the number of clusters for template hand mesh. Note that we do not include spectral filtering in the graph decoder here.

	Single-view	Multi-view
METRO	10.87	-
METRO + avg. pool	-	8.71
METRO + max pool	-	7.09
Ours ($K = 1$)	-	6.59
Ours ($K = 2$)	-	5.71
Ours ($K = 3$)	-	5.19
Ours ($K = 4$)	-	4.79
Ours ($K = 5$)	-	4.59
Ours ($K = 6$)	-	5.28
Ours ($K = 7$)	-	**3.72**
Ours ($K = 8$)	-	3.79

Table 4.4: Ablations on multi-view feature fusion. We report hand error on 3 experimental settings ((a)-(c)) with different feature fusion strategies. We keep the same transformer and spectral graph decoder for all fusion strategies.

	(a)	(b)	(c)
Direct concatenation	11.3	13.9	14.7
Max pool	9.3	11.2	12.5
Soft-attention + max pool	**6.7**	**7.2**	**7.4**

Effects of spectral filters. We experiment with three commonly-used spectral filters: Gaussian g_{gau}, Laplacian g_{lap} and Chebyshev filters g_{cheb} with varying order of polynomials

Table 4.5: Ablations of different spectral filters.

	g_{gau}	g_{lap}	g_{cheb_3}	g_{cheb_4}	g_{cheb_5}	g_{cheb_6}
Hand error	1.56	1.45	1.38	1.47	1.38	1.39

in Table 4.5. We choose g_{cheb} for efficiency and do not find increasing order of polynomials improves performance further. In the following, we detail Gaussian and Laplacian filters.

Given eigenvalue λ, the Gaussian filter function f_{gau} can be described as:

$$f_{gau}(\lambda) = e^{\frac{-\lambda^2}{2\sigma^2}}, \tag{4.3}$$

where σ is the standard deviation of the Gaussian distribution. We set $\sigma = 0.5$ in our experiments. Similarly, the Laplacian filter function f_{lap} can be described as:

$$f_{lap}(\lambda) = \lambda^{-\frac{1}{2}}. \tag{4.4}$$

As eigenvalues can have zero and negative values, the inverse square root computation can result in NaN (not a number) values. Therefore, we put a tolerance value that is close to zero to avoid this.

Synthetic to real transfer. Large-scale synthetic dataset can be used to pre-train models in the absence of suitable real datasets. In the following, we demonstrate a use case of deploying our pre-trained model which was trained on synthetic data to general real images. First, we train on a subset of our real dataset and freeze all part of the pre-trained model except CNN backbone. As the registered meshes for our real dataset have a different mesh topology, we drop the pink cluster of our original template mesh (shown in Figure 4.2) before training. This approach is analogous to MANO-based 3D hand mesh estimation methods (Baek et al., 2019; Boukhayma et al., 2019) and has been demonstrated in Moon

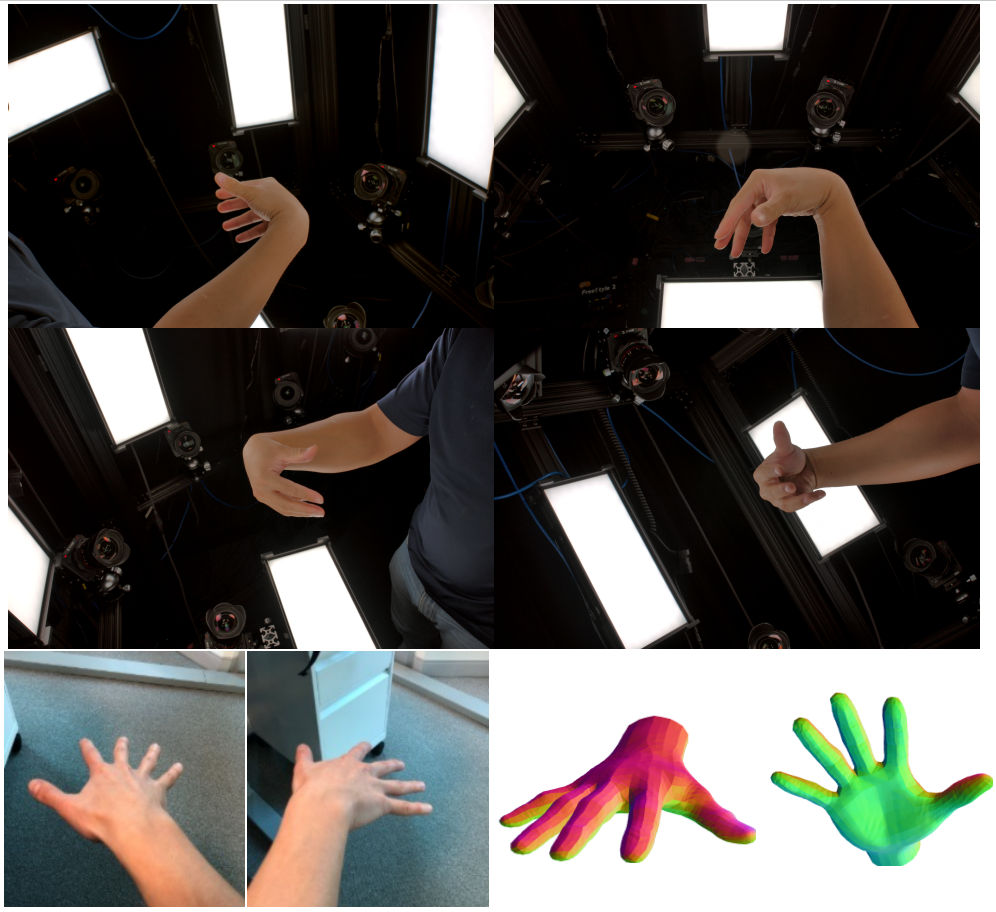


Figure 4.8: Training examples of our real dataset (*top*) and 3D hand mesh estimation results on in-the-wild image (*bottom*). The resulting mesh contains 3745 vertices.

et al., 2020b. We show that our model can generalise to in-the-wild-images in Figure 4.8.

Model compression. We are interested in finding the minimal size to which model can be compressed while maintaining the METRO baseline performance in Table 4.1. First, we perform computational complexity analysis on the self-attention layer in the Transformer encoder. There are two key steps to compute self-attention: 1) linear projection to C-dimensional query, key and value matrices from V' vertex queries requires $O(V'C^2)$ and softmax operation for layer output requires another $O(V'^2C)$. The total computational complexity of each transformer layer is therefore quadratic no matter whether V' or C dominates. Recall that as we uniformly subsample template hand meshes by 10 times to

Table 4.6: Ablations of different backbones and hyperparameters. We denote P_{cnn} and P_{total} to be the number of parameters for CNN backbone and total model, respectively.

Backbone	P_{cnn}	V'	C	Error	P_{total}
ResNet-50	23.5M	804	256	1.38	58.3M
EfficientNet-B3	12.9M	804	256	2.53	47.7M
EfficientNet-B2	9.2M	804	256	2.55	44M
EfficientNet-B1	7.8M	804	256	2.74	42.6M
EfficientNet-B0	5.3M	804	256	2.85	40.1M
EfficientNet-B3	12.9M	804	128	3.00	40.6M
EfficientNet-B2	9.2M	804	128	3.20	36.9M
EfficientNet-B1	7.8M	804	128	3.20	35.5M
EfficientNet-B0	5.3M	804	128	2.91	33M
EfficientNet-B3	12.9M	804	64	4.00	38.4M
EfficientNet-B2	9.2M	804	64	3.87	34.7M
EfficientNet-B1	7.8M	804	64	4.04	33.3M
EfficientNet-B0	5.3M	804	64	4.31	30.8M
EfficientNet-B3	12.9M	804	32	89.7	37.3M
EfficientNet-B0	9.2M	804	32	90	29.7M
EfficientNet-B0	5.3M	160	256	4.12	42.8M
EfficientNet-B0	5.3M	160	128	4.96	37.8M
EfficientNet-B0	5.3M	80	64	6.89	34.2M

obtain $V' = 804$ and C is empirically set to 256 in earlier sections, V' contributes more to overall complexity as it is dominantly larger. As shown in Table 4.6, we gradually decrease V' before C while using a variant of EfficientNet (M. Tan et al., 2019).

Influence of different loss terms. In Table 4.7, we analyse the influence of different loss terms. In these experiments, we consider L1 losses on 3D mesh vertices and 2D re-projection loss, *i.e.* $\mathcal{L}_{mesh}$ and $\mathcal{L}_{2D}$, respectively. In addition, we experiment with mean squared euclidean distance loss $\mathcal{L}_{MSE}$ on hand mesh. For mesh regularisation, we apply edge length regularisation $\mathcal{L}_{edge}$ and minimise the Chamfer distances $\mathcal{L}_{cham}$. We find that the combination of $\mathcal{L}_{mesh}$, $\mathcal{L}_{2D}$ and $\mathcal{L}_{edge}$ delivers the optimal results.

Table 4.7: Impact of different loss terms on our synthetic dataset. Hand errors are given in millimeters (mm).

$\mathcal{L}_{mesh}$	$\mathcal{L}_{2D}$	$\mathcal{L}_{MSE}$	$\mathcal{L}_{edge}$	$\mathcal{L}_{cham}$	Error
✓	✓				1.55
✓	✓	✓			1.48
✓	✓		✓		1.38
✓	✓	✓	✓		1.38
✓	✓	✓	✓	✓	1.49

Analysis on mesh refinement at inference. For quantitative evaluation, we use penetration depth (mm) and intersection volume (cm^3). Penetration depth refers to the maximum distances from hand mesh vertices to the other/self hand's surface when in a collision. Intersection volume is obtained by voxelising the meshes using a voxel size of $0.5cm$. We report the results in Table 4.8. These results demonstrate the robustness of our optimisation-based mesh refinement strategy.

Table 4.8: Quantitative evaluation on the impact of mesh refinement at inference.

	Before refinement	After refinement
Max. penetration (mm)	5.3	0.16
Intersection vol. (cm^3)	2.0	0.09

4.4 Summary

In this chapter, we have introduced a novel spectral graph-based transformer framework designed for the reconstruction of high fidelity two-hand meshes from egocentric views. The main idea behind this work was to explore how the fundamental properties of the graph Laplacian, derived from spectral graph theory, can be effectively incorporated into a Transformer architecture. Through our extensive experiments, we have demonstrated that our

proposed multi-view feature fusion strategy is most effective when combined with mesh segmentation based on spectral clustering. This combination allows for improved aggregation of features across multiple views while enabling selective attention to different hand parts. To enhance the quality and physical plausibility of the reconstructions, we have also introduced spectral filtering and an optimisation-based refinement step. These additional components contribute to the generation of more accurate and realistic meshes.

Furthermore, our framework is not limited to two-hand reconstructions alone but can be extended to other multi-view reconstruction tasks as well. The versatility and generalisability of our approach provide opportunities for its application in various domains.

However, it is important to note that our method may encounter challenges when dealing with highly complex self-penetrations, as shown in Figure 4.9. While our current optimisation-based mesh refinement step addresses self-penetrations, future work can focus on tackling interpenetrations during hand-hand interactions. We believe that incorporating temporal information and more advanced physical modeling techniques into our framework will be instrumental in addressing this issue.

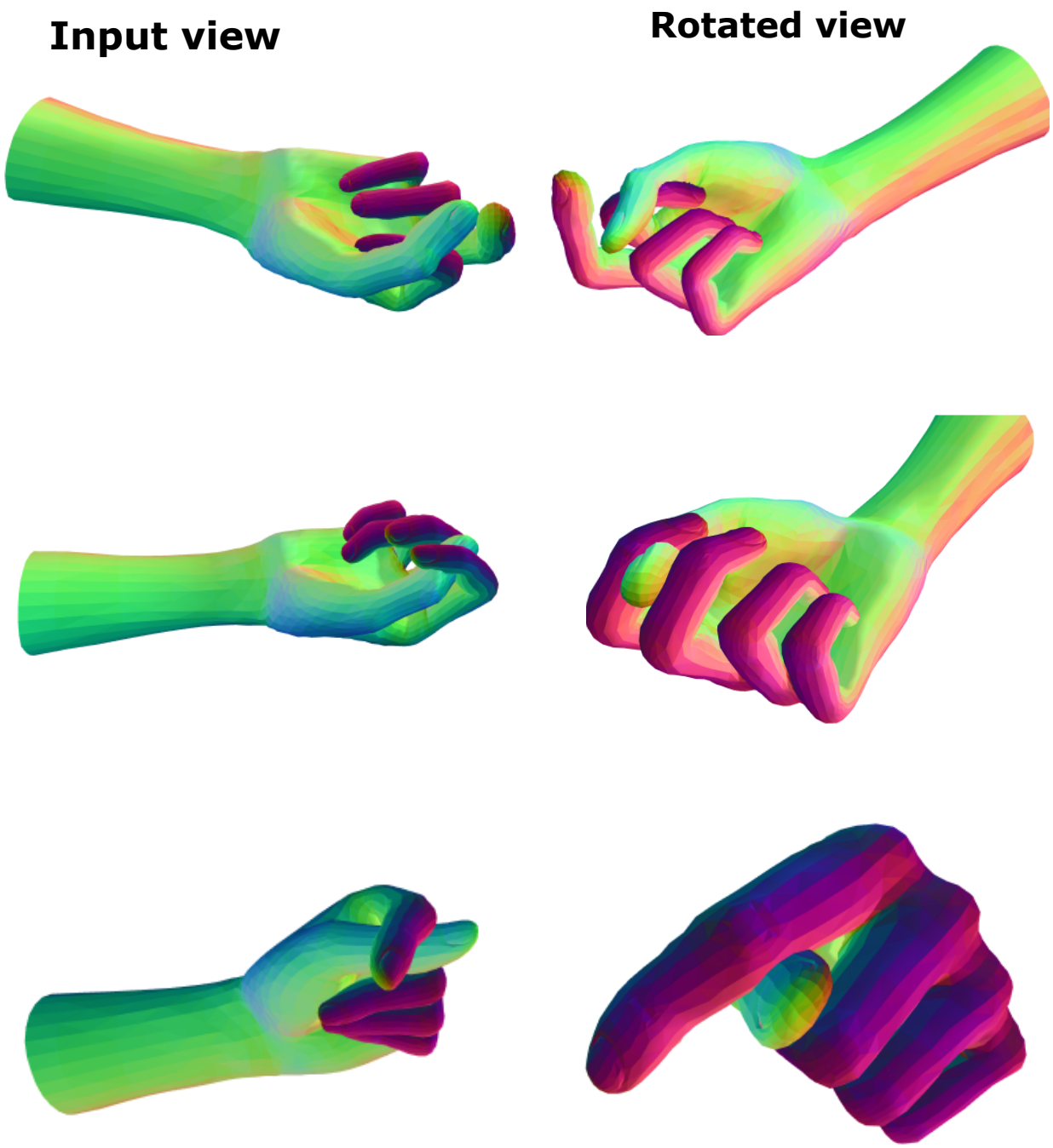


Figure 4.9: Failure examples for mesh refinement.

Chapter Five

LEARNING HAND-OBJECT RECONSTRUCTION AND COMPOSITIONAL ACTION RECOGNITION FROM EGOCENTRIC RGB VIDEOS

In the previous chapters, we developed learning-based method for reconstructing hands and objects where accurate 3D ground truth annotations is required for every objects. However, in practice, existing method does not scale well to the number of objects in the world. This limitation hinders the ability of learning-based approaches to generalise beyond their specific training domains. As a result, these methods often face challenges when encounter objects or scenarios that differ from their training data.

In this chapter, we are interested in developing a learning-based approach that can demonstrate generalisability to unseen objects. In particular, we study jointly the problem

of hand-object reconstruction and interaction recognition from egocentric RGB views. Our goal is to enhance understanding of hand-object interactions by developing approaches that can recognise seen actions on unseen objects. By achieving this, we aim to improve the versatility and applicability of action recognition systems in real-world scenarios.

As discussed in Section 2.4, recent egocentric hand-object interaction datasets (Kwon et al., 2021; Garcia-Hernando et al., 2018) with 3D annotations enable the development of a unified framework for estimating hand-object poses and interaction classes (Tekin et al., 2019; S. Yang et al., 2020; Kwon et al., 2021; Wen et al., 2023; H. Cho et al., 2023). These methods couple 3D geometric cues (*i.e.* hand-object poses and/or contact maps) and appearance features to predict interaction class. Despite a unified understanding of the hand and manipulated object dynamics being crucial for recognising egocentric interactions, we find that pure appearance-based methods (*i.e.* Multiscale Vision Transformers (H. Fan et al., 2021; Yanghao Li et al., 2022)) can achieve comparable performance to the state-of-the-arts. This raises immediate questions on when or how 3D geometric features can benefit interaction recognition.

In contrast, while deep architectures trained on large-scale datasets (Sigurdsson et al., 2016; Kay et al., 2017; Karpathy et al., 2014) exhibit strong distribution learning capabilities, mainstream action recognition models (Simonyan et al., 2014; Carreira et al., 2017; Feichtenhofer et al., 2019; L. Wang et al., 2016) primarily focus on frame appearance rather than temporal reasoning. Consequently, reversing the order of the video frame at test time will often produce the same classification result (Materzynska et al., 2020; B. Zhou et al., 2018). In particular, classical activity recognition methods like the two-stream Convolutional Neural Network (Simonyan et al., 2014) and I3D (Carreira et al., 2017) have demonstrated strong performance on various video datasets, including UCF101 (Soomro et al., 2012) and Sport1M (Karpathy et al., 2014), with only still frames and optical flow. While appearance features can be highly predictive of the action class (Santoro et al., 2017;

Battaglia et al., 2018), it remains challenging for appearance-based deep networks to capture the *compositionality* of action and objects without temporal transformations or geometric relations (Materzynska et al., 2020).

To address the aforementioned problems, Materzynska et al. (2020) extends the Something-Something dataset (Goyal et al., 2017) and introduces the Something-Else task with a new compositional split. This presents a novel task known as compositional action recognition, in which methods are required to recognise an action with unseen objects. Under this problem setting, the combinations of actions and object instances do not overlap in the training and testing split. Therefore, models are encouraged to learn the compositionality of action *verb* and *noun*, and not overfit to the correlation between appearance features and action classes. Nonetheless, the current research in compositional action recognition is primarily generic approaches using 2D geometric cues such as 2D instance bounding boxes. The potential benefits offered by 3D geometric information remain an open problem.

Therefore, in this chapter, we take an alternative approach which exploits the compositionality of actions using 3D geometric information. To achieve that, we first extend the two existing 3D annotated egocentric hand-object datasets, H2O (Kwon et al., 2021) and FPHA (Garcia-Hernando et al., 2018), by introducing new compositional splits. We show that the existing approaches (S. Yang et al., 2020; Wen et al., 2023; H. Fan et al., 2021; Yanghao Li et al., 2022) still face significant challenges in recognising a seen action when facing new objects. This is because the current methods (either single or dual branches) are unable to tackle the problem of appearance bias in objects, as they have to take the combination of appearance and geometric information as a whole. In addition, these approaches focus on extracting features for the whole scene and do not explicitly recognise objects as individual entities. Hence, they cannot fully capture the compositionality of the action [1].

[1]We follow the same definition of action compositionality as defined in Materzynska et al. (2020).

In this chapter, we present a new collaborative learning framework that allows an action verb and object to interact and complement each other. The key motivation for this strategy is that the tasks of estimating hand-object poses and recognising interactions are naturally closely-correlated. Existing collaborative learning methods in understanding hand-object interactions typically follow an iterative approach where the multiple target learning tasks (*i.e.* hand pose estimation, object reconstruction or action recognition) boost each other mutually and progressively. However, connecting branches iteratively can lead to highly unstable training (demonstrated in Chapter 3). This is because gradients from one branch can propagate through the connections to affect the other branches which causes unstable gradients. We explicitly address this by a new transformer-based architectural design to exploit the compositionality of actions and avoid branch stacking. In addition, we exploit to use superquadrics (Barr, 1981) as the intermediate 3D object representation. This is motivated by the fact that existing action recognition methods have limitations in accurately representing objects' shape and movement with only 2D or 3D bounding boxes. But at the same time, accurately reconstructing complete 3D objects without an object template remains highly challenging, especially in scenarios involving unseen objects. Therefore, superquadrics offer a compact representation with their ability to represent a wide range of shapes with few parameters. In addition and more importantly, it allows models to interpret objects with basic geometric primitives.

Our contributions in this chapter are fourfold: 1) We present an end-to-end trainable collaborative learning framework to leverage 3D geometric information for compositional action recognition from egocentric RGB videos. 2) We then show that using superquadrics as the intermediate 3D object representation is beneficial for 3D hand pose estimation and interaction recognition. 3) Further, we extend two egocentric hand-object datasets by introducing new compositional splits and investigate compositional action recognition where a subset of action verb and noun combinations do not exist during training. 4) Lastly, we

demonstrate state-of-the-art performance on two common datasets, H2O (Kwon et al., 2021) and FPHA (Garcia-Hernando et al., 2018), in both official and compositional settings.

5.1 Preliminary

Our work tackles the joint problem of 3D hand-object reconstruction and action recognition from egocentric RGB videos. While a comprehensive review of the existing literature on understanding hand-object interactions is covered in Sections 2.3 and 2.4, we introduce compositional action recognition and superquadrics recovery in Sections 5.1.1 and 5.1.2, respectively.

5.1.1 Compositional action recognition

This task is designed to alleviate the problem of appearance bias by disjointing the combination of actions and objects between training and testing. STIN (Materzynska et al., 2020) models actions as transformation of geometric relations in both spatial and temporal domains using 2D instance bounding boxes. This approach generalises well to most actions but fails when there are intrinsic state changes of objects. T. S. Kim et al. (2020) proposes to fuse RGB information with instance bounding boxes to capture more complex actions. P. Sun et al. (2021) removes the appearance effect by counterfactual debiasing inference. While these methods have proven effective, they are primarily designed to leverage 2D geometric information and do not fully explore the potential of 3D geometric cues. As a result, their performance remains comparable to I3D (Carreira et al., 2017). In this work, we focus on leveraging 3D geometric cues for compositional action recognition.

5.1.2 Superquadrics recovery

Superquadric is a well-studied computational primitive shape abstraction, offering a diverse range of shape representations including cuboids, ellipsoids, cylinders, octohedra, and other variations. It was first proposed to model complex objects in computer graphics (Barr, 1981). Solina et al. (1990) presents a method for abstracting simple objects from range images using a single superquadric. Subsequently, Leonardis et al. (1997) and Chevalier et al. (2003) extend to recover more complex objects with multiple superquadrics. Recently, W. Liu et al. (2022) proposes a probabilistic approach to improve robustness to outlier and fitting accuracy.

5.2 Methodology

Our training pipeline, as shown in Figure 5.1, takes a sequence of T RGB frames $\mathbf{I} \in \mathbb{R}^{T\times 256\times 256\times 3}$ of dynamic hands manipulating objects as input. We first obtain spatial features $\mathbf{x} \in \mathbb{R}^{T\times d}$ by passing each frame into a ResNet-18 (He et al., 2016) encoder where d refers to feature dimensions. To enhance the interaction between visual and geometric cues, we introduce a simple collaborative learning framework by leveraging the Transformer encoder and decoder (Vaswani et al., 2017) as basic building blocks. Specifically, we design a two-branch network where the appearance branch extracts video features from the entire sequence in Section 5.2.1 and the geometric branch aims to recover 3D hand-object geometric information in Section 5.2.2. In addition, we explicitly model the compositionality of the interaction by decomposing the action class into a verb-and-noun pair in Section 5.2.3. Finally, we combine video appearance and geometric representations for recognising egocentric hand-object interactions in Section 5.2.4.

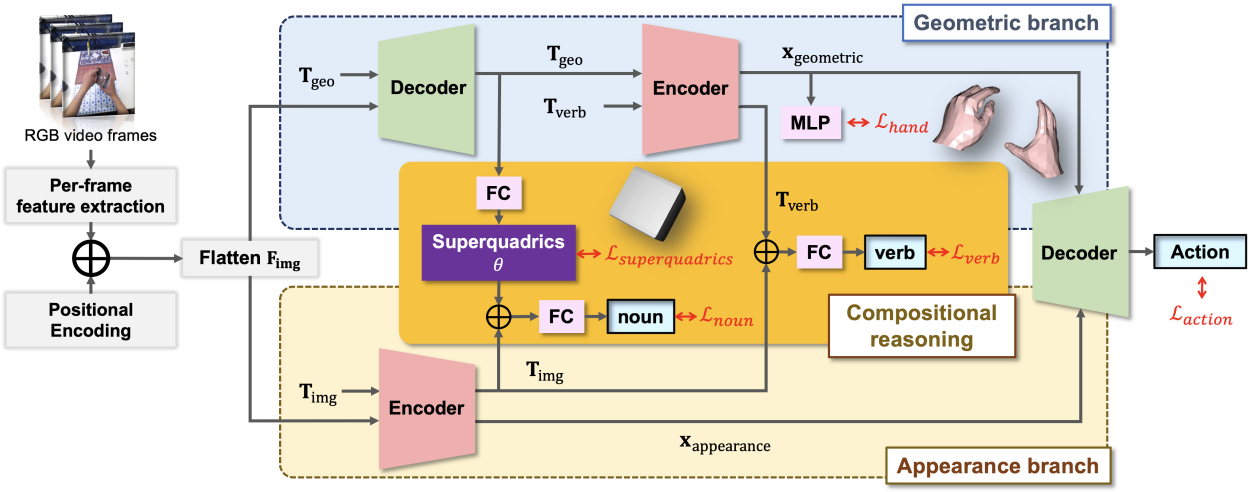


Figure 5.1: Overview of our approach. Our framework takes RGB videos as input, which are processed by a CNN backbone to produce per-frame spatial features $\mathbf{x}$. The appearance branch (bottom) applies positional encoding to $\mathbf{x}$ and combines it with a learnable token $\mathbf{T}_{\text{img}}$ before feeding into a Transformer encoder. Similarly, the geometric branch (top) extracts geometric features $\mathbf{T}_{\text{geo}}$ from flatten spatial features $\mathbf{F}_{\text{img}}$ using a Transformer decoder. The features from both branches are combined to predict superquadrics and object category. In addition, the geometric features are aggregated through self-attentions in another Transformer encoder to create global context-aware features between object shape and hand poses. The outputs of this Transformer encoder are aggregated geometric features $\mathbf{x}_{\text{geometric}}$ and verb token features $\mathbf{T}_{\text{verb}}$. These features are used to predict hand pose and action verb. Finally, the action class is predicted by feeding $\mathbf{x}_{\text{geometric}}$ into a cross-attention mechanism with the aggregated spatial representation $\mathbf{x}_{\text{appearance}}$ through a Transformer decoder.

5.2.1 Appearance branch

Video feature encoder. Given image features $\mathbf{x}$, we concatenate with a learnable token $\mathbf{T}_{\text{img}} \in \mathbb{R}^d$ and apply positional encoding before feeding them into a Transformer encoder. The encoder models the relationships of different spatial regions and $\mathbf{T}_{\text{img}}$ through self-attentions. The learnable token $\mathbf{T}_{\text{img}}$ captures essential global contexts from backbone representation which are used for compositional reasoning in Section 5.2.3. In addition, the encoder outputs aggregated spatial representation $\mathbf{x}_{\text{appearance}} \in \mathbb{R}^{T \times d}$ which is later used for interaction recognition in Section 5.2.4.

Learnable attention mask. As the input sequence $\mathbf{I}$ is densely sampled, successive video

frames can introduce redundancy to spatial features $\mathbf{F}_{\text{img}}$. This directly limits the length of the input sequence as the computational requirement of the transformer encoder is proportional to the number of input frames T. To address this problem and encourage the model to attend across multiple video segments, we introduce a learnable attention mask $\mathcal{M} \in \mathbb{R}^{T \times T}$ for the video feature encoder. We first apply the sigmoid function $\sigma(\cdot)$ on $\mathcal{M}$ to obtain continuous activation of range 0 to 1. We then regularise model training for long-range video sequences by minimising:

$$\mathcal{L}_{\text{mask}}(\mathcal{M}) = \sum_{i=1}^{T} \sum_{j=1}^{T} \sigma(\mathcal{M}_{i,j}). \tag{5.1}$$

5.2.2 Geometric branch

As estimating the poses of hand and object requires more local or nearby frames, we divide the video sequence into N consecutive segments. Specifically, we follow Wen et al. (2023) and use a shifting window strategy with window size t, *i.e.* $N = T/t$. The frames beyond sequence length T are padded but masked out from attention computation.

Instead of aiming to reconstruct the hand and the manipulated object simultaneously, we first estimate the shape and poses of the manipulated object and leverage this geometric information to predict hand poses. The reason behind this approach is that joint estimation poses a significantly harder problem. As discussed in Section 1.3, self-occlusion and self-similarity between the joints of two hands are unique problems in interacting hands. Moreover, when interacting with objects, hands and objects often exhibit even greater occlusions. This problem is further amplified under egocentric view setting due to large degree of erratic camera motions.

In addition, existing work which relies on 2D/3D bounding boxes has limitations in

accurately representing the shape and movement of objects. However, at the same time, it remains challenging to accurately reconstruct unseen objects from RGB images. Therefore, we exploit superquadrics as a new object representation for improving action recognition. In the following, we present the *preliminaries* of superquadrics and detail our two-stage approach consisting *superquadrics decoder* and *hand pose estimator*.

Preliminaries. As shown in Figure 5.2, superquadrics are a family of geometric primitives, *i.e.* cuboids, cylinders, ellipsoids, octahedra and their intermediates, which can be defined by an implicit function $f(\cdot)$ (Barr, 1981):

$$f(\mathbf{p}) = \left(\left(\frac{x}{a_x} \right)^{\frac{2}{\epsilon_2}} + \left(\frac{y}{a_y} \right)^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left(\frac{z}{a_z} \right)^{\frac{2}{\epsilon_1}} = 1, \tag{5.2}$$

where points $\mathbf{p} = [x, y, z] \in \mathbb{R}^3$ satisfying Equation 5.2 form the surface of a superquadric. It can be encoded using 5 parameters: shape parameters $\epsilon_1, \epsilon_2 \in [0, 2] \subset \mathbb{R}$ and scale parameters $a_x, a_y, a_z \in \mathbb{R}_{>0}$. While the shape parameters can exceed 2 and result in non-convex shapes, we limit them within the convex region in this paper. We can now fully parameterise a superquadric by including the Euclidean transformation $g \in SE(3)$, *i.e.* $g = [\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3]$.

Superquadrics decoder. We use superquadrics $\boldsymbol{\theta}$ to be the intermediate 3D object representation, as it offers a compact way to represent a wide range of geometric primitives, *i.e.* $\boldsymbol{\theta} = \{\epsilon_1, \epsilon_2, a_x, a_y, a_z, g\} \in \mathbb{R}^{11}$. As shown in Figure 5.3, superquadrics can provide sufficient expressiveness to reasonably model a diverse range of everyday objects. To this end, we extract geometric features $\mathbf{T}_{\text{geo}} \in \mathbb{R}^{N \times d}$ from the segmented spatial features using a Transformer decoder. We train a fully-connected layer to predict $\boldsymbol{\theta}$ from the flatten $\mathbf{T}_{\text{geo}}$ by minimising the L1 loss $\mathcal{L}_{\text{superquadrics}}$.

Hand pose estimator. Given geometric features $\mathbf{T}_{\text{geo}}$, we concatenate with a learnable

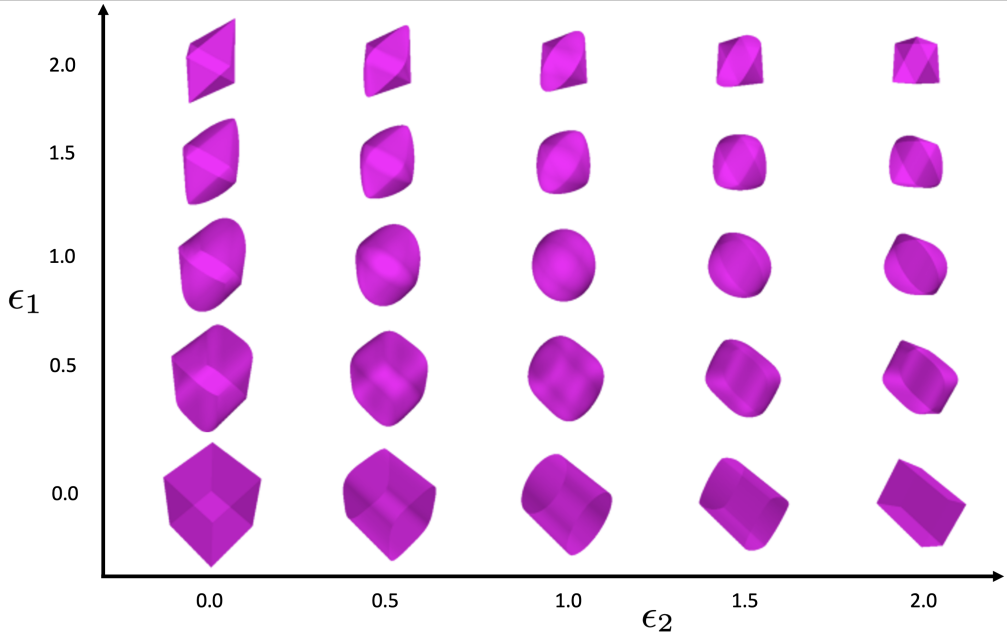


Figure 5.2: Qualitative examples of convex superquadrics. We show that superquadrics can model diverse shapes by varying the shape parameters, ϵ_1 (y-axis) and ϵ_2 (x-axis).

verb token $\mathbf{T}_{\text{verb}}$ and use self-attention in a Transformer encoder to create global context-aware features between object shape and hand poses. The Transformer encoder outputs aggregated geometric features $\mathbf{x}_{\text{geometric}} \in \mathbb{R}^{N \times d}$ and verb token features $\mathbf{T}_{\text{verb}} \in \mathbb{R}^{d}$.

We map the aggregated geometric features $\mathbf{x}_{\text{geometric}}$ to hand pose space by a 3-layer MLP and use MANO joint angles (Romero et al., 2017) for hand pose representation. Specifically, we estimate 16 3D joint angles under the hand kinematic tree and MANO hand shape parameter per hand. Then, we can compute the 21 root-relative 3D joint locations of each hand by using the predicted joint angles and hand shape parameters. They are learned by minimising L1 loss $\mathcal{L}_{\text{hand}}$.

5.2.3 Compositional reasoning

In the following, we describe how we leverage the compositional nature of actions by exploiting the action class as verb-noun pair with 3D geometric cues, *i.e.* superquadrics $\boldsymbol{\theta}$ and

Figure 5.3: Qualitative examples of superquadrics. We extract superquadrics from everyday objects obtained from *YCB* (Calli et al., 2015), *ShapeNet* (A. X. Chang et al., 2015), *FPHA* (Garcia-Hernando et al., 2018) and *H2O* (Kwon et al., 2021) datasets. We show that superquadrics have sufficient expressiveness to represent everyday objects. We also present an example failure case in the red box.

geometric-aware verb token $\mathbf{T}_{\text{verb}}$.

Object category predictor. We first predict the category of the manipulated object as it corresponds to the noun of an action. To achieve that, we leverage the basic primitive geometric information from superquadrics $\boldsymbol{\theta}$ as object shape provides strong signals to estimating object category. More specifically, we predict the classification probability vector for object category by linearly projecting the concatenation of superquadrics $\boldsymbol{\theta}$ and $\mathbf{T}_{\text{img}}$. We supervise this linear layer by minimising the cross-entropy loss $\mathcal{L}_{\text{noun}}$.

Verb predictor. Similarly, we predict action verb by feeding the concatenation of $\mathbf{T}_{\text{verb}}$ and $\mathbf{T}_{\text{img}}$ to a linear layer. It is also trained by minimising the cross-entropy loss $\mathcal{L}_{\text{verb}}$.

Discussion. The key idea for concatenating with $\mathbf{T}_{\text{img}}$ is to allow verb and noun of an action to interact with the appearance branch. Also, it generates loss gradients for both branches to develop a collaborative learning relationship. In addition, the motivation for estimating superquadrics first in the geometric branch is based on the fact that the human visual system flavours abstracting scenes into canonical parts for better perceptual understanding (W. Liu et al., 2022). This enables robust action recognition using basic geometric primitives instead of relying on accurate point-wise estimation. In summary, our design targets the problem of recognising a seen action when facing new objects by enabling the network to capture the compositionality of an action explicitly.

5.2.4 Interaction recognition

Besides explicitly modelling the compositionality of actions, our proposed framework can easily combine with any video-level appearance representation. The impact of appearance features can be two-fold: 1) The presence of appearance features can be particularly beneficial

for action classes that lack prominent inter-object dynamics (Materzynska et al., 2020). 2) Conversely, appearance bias can inhibit the model learning ability by making strong correlations on spatial appearance rather than temporal or geometric transformations (P. Sun et al., 2021). To overcome the limitations of existing methods that can only accept or reject appearance information as a whole, we use a Transformer decoder for recognising interaction. This decoder takes the aggregated geometric and spatial features, *i.e.* $\mathbf{x}_{\text{geometric}}$ and $\mathbf{x}_{\text{appearance}}$ as input and extracts relevant image features through cross-attention between geometric features. The vector output of this decoder is fed to a 3-layer MLP classifier of width and is supervised with cross-entropy loss $\mathcal{L}_{\text{action}}$. We investigate and analyse our design choices in Section 5.3.

5.2.5 Training

Our final loss $\mathcal{L}_{\text{final}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{final}} = \; & \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{superquadrics}} + \mathcal{L}_{\text{hand}} \\ & + \mathcal{L}_{\text{noun}} + \mathcal{L}_{\text{verb}} + \mathcal{L}_{\text{action}}.\end{aligned} \tag{5.3}$$

5.3 Experiment

In this section, we first describe our implementation details in Section 5.3.1 and the datasets in Section 5.3.2. We then describe the corresponding evaluation protocols in Sections 5.3.3 and 5.3.4. Finaly, we validate our approach numerically and qualitatively in Sections 5.3.5 and 5.3.6.

5.3.1 Implementation details

We train all parts of our model simultaneously with the Adam optimiser (Kingma et al., 2015) using an initial learning rate of 3×10^{-5} and halve the learning rate in every 15 epochs. We keep the relative weights between different losses and normalise them such that the sum of all the weights equals to 1 for all experiments. We use ResNet (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) for our backbone. For all Transformer encoders and decoders, we use 2 encoding/decoding layers where each layer has 8 attention heads. We use the fixed sine/cosine functions for positional encoding and add layer normalisation before the attention and feed-forward computations (Vaswani et al., 2017). We follow Wen et al. (2023) by setting $T = 128$, $t = 16$, $d = 512$ and training for 45 epochs with batch size of 2.

5.3.2 Datasets

We conduct experiments on 3 interacting hand-object datasets and detail below.

ObMan (Hasson et al., 2019). We precomputed superquadrics for all object meshes using the EMS algorithm (W. Liu et al., 2022) and pretrained the geometric branch on *ObMan* before training on other real datasets. We observed consistent improvements over training directly on real data as the number of objects in hand-object interaction dataset is very limited.

First-person hand benchmark (FPHA) (Garcia-Hernando et al., 2018). We evaluate on the *action split* where all subjects and actions are present in both training and testing. This split consists of 600 and 575 videos for training and testing, respectively.

H2O (Kwon et al., 2021). We use the sequences of egocentric view for training and

testing. Specifically, the training split consists of 569 videos from the first 3 subjects, while the testing split includes 242 videos from the remaining unseen subjects.

5.3.3 Baselines

We compare our method against MViTv2 (Yanghao Li et al., 2022) and HTT (Wen et al., 2023). MViTv2 is widely adopted for video recognition tasks and can serve as a strong appearance-based baseline. We train the base variant of MViTv2 with weights pretrained on Kinetics-400 dataset (Kay et al., 2017) using the PySlowFast library (H. Fan et al., 2020). For the pose-based baseline, we consider HTT as it is a recent method that achieves state-off-the-art performance on both *FPHA* (Garcia-Hernando et al., 2018) and *H2O* (Kwon et al., 2021) datasets. We also consider two Transformer-based baselines ('Ours w/o CR' and 'Our w/o SQ') which do not contain compositional reasoning (CR) and superquadrics (SQ), respectively. They are useful for understanding the importance of superquadrics and compositional reasoning for recognising interactions with unseen objects.

5.3.4 Evaluation metrics

We report the MPVE in *mm* to evaluate *pose estimation*. MPVE measures the mean Euclidean distances between predictions and ground-truths. We also report the top-1 classification accuracy for *action recognition*.

5.3.5 Results

Hand and object pose estimations. We report quantitative comparisons with the state-of-the-art methods on *H2O* and *FPHA* datasets in Table 5.1. H+O (Tekin et al., 2019)

Table 5.1: Error rates of pose estimation on *H2O* (Kwon et al., 2021) and *FPHA* (Garcia-Hernando et al., 2018). We report MPVE in *mm* for hand (left/right) and object error. Our proposed method performs competitively without known object templates at inference.

Method	*H2O*		*FPHA*	
	Hand (left/right) ↓	Object ↓	Hand ↓	Object ↓
H+O (Tekin et al., 2019)	41.4/38.9	50.4	15.8	24.9
H2O (Kwon et al., 2021)	41.5/37.2	47.9	-	-
HTT (Wen et al., 2023)	35.0/36.1	-	15.8	-
H2OTR (H. Cho et al., 2023)	**24.4/25.8**	45.2	15.0	21.0
Ours	28.9/30.2	**43.5**	**13.6**	**20.1**

Figure 5.4: Qualitative examples on *FPHA* (Garcia-Hernando et al., 2018).

is a single hand method so the results are reported separately. H2OTR (H. Cho et al., 2023) is a Transformer-based framework which achieves state-of-the-art accuracy for hand pose estimation. All of the compared methods require manual selection of object models at test time. In contrast, our method performs competitively without known object templates and outperforms all methods on object pose estimation. We attribute this to the fact that superquadrics can provide dense 3D geometric information about the manipulated object, whereas 2D or 3D bounding boxes have limitations on representing object shape and movement. We show qualitative results on *FPHA* and *H2O* datasets, in Figure 5.4 and 5.5 respectively.

Action recognition. We also report the top-1 classification accuracy for recognising egocentric hand-object interactions in Table 5.2. We split the table into two sections: appearance-based (X. Wang et al., 2018; Carreira et al., 2017; Feichtenhofer et al., 2019; Yanghao Li

Figure 5.5: Qualitative example on *H2O*. We show that our model can recover plausible interaction across different object categories and hand-object configurations without object templates.

et al., 2022) and geometric (Tekin et al., 2019; S. Yang et al., 2020; Kwon et al., 2021; Wen et al., 2023; H. Cho et al., 2023) methods for clear comparison. S. Yang et al. (2020) is the closest method to ours as it is a collaborative learning framework which allows appearance and geometric cues to interact under a two-branch architecture. H2OTR (H. Cho et al., 2023)

Table 5.2: Classification accuracy of action recognition on *H2O* (Kwon et al., 2021) and *FPHA* (Garcia-Hernando et al., 2018). We show that adding compositional reasoning with superquadrics allow us to outperform the current state-of-the-arts.

Method	*H2O* Top-1 accuracy (%) ↑	*FPHA* Top-1 accuracy (%) ↑
C2D (X. Wang et al., 2018)	70.66	-
I3D (Carreira et al., 2017)	75.21	-
SlowFast (Feichtenhofer et al., 2019)	77.69	-
MViTv2 (Yanghao Li et al., 2022)	90.08	98.45
H+O (Tekin et al., 2019)	68.88	82.43
Collab. (S. Yang et al., 2020)	-	85.22
H2O (Kwon et al., 2021)	79.25	-
HTT (Wen et al., 2023)	86.36	94.09
H2OTR (H. Cho et al., 2023)	90.90	98.4
Ours w/o compositional reasoning	86.01	94.87
Ours w/o superquadric prediction	88.46	96.28
Ours	**92.25**	**98.74**

leverages the estimated contact map to guide interaction recognition. As shown in Table 5.2, the appearance-based baseline MViTv2 performs competitively with the state-of-the-art geometric method H2OTR (H. Cho et al., 2023). It raises an immediate question as to when 3D geometric cues be beneficial to recognising interaction as collecting ground-truth contact maps or other 3D annotations are non-trivial (Brahmbhatt et al., 2019; Tse et al., 2022b; Kwon et al., 2021). We will address this question in the following paragraph. Nonetheless, we demonstrate the effectiveness of explicit compositional reasoning with superquadrics by outperforming all methods.

Compositional action recognition. We further evaluate our model on the compositional recognition task in Tables 5.3 and 5.4. Following Materzynska et al. (2020), we first create new splits for the task of compositional action recognition by extending existing egocentric hand-object datasets. Specifically, we remove the sequences that contain the predefined

object category from the train split, such that the combinations of a verb (action) and nouns do not overlap in the testing set. To gain a deeper understanding of the model generalisation ability on unseen objects, we evaluate the model using $N_{\text{obj}}-$fold cross validation where N_{obj} refers to the number of total objects presented in the dataset. We keep the original testing splits (named $\mathcal{S}_0$) to illustrate the difficulty of this compositional task. We further experiment on a more challenging split where two object categories are randomly removed in the *H2O* dataset. We name the base splits by $\mathcal{S}_1$ and the more difficult splits where additional verb-nouns combinations are removed from training by $\mathcal{S}_2$. We report the mean and the standard deviation of top-1 classification accuracy for all experiments.

As shown in Tables 5.3 and 5.4, our collaborative learning framework consistently outperforms both the appearance and geometric baselines, *i.e.* MViTv2 (Yanghao Li et al., 2022) and HTT (Wen et al., 2023). We find that the performance of MViTv2 drastically drops by 29.44% and 37.76%, in $\mathcal{S}_1$ and $\mathcal{S}_2$ of *H2O* respectively. These results are in line with previous studies (Materzynska et al., 2020; B. Zhou et al., 2018; P. Sun et al., 2021) where deep architectures tend to overfit the object appearance. By adding geometric cues in HTT, we observe a small performance gain by an average of 5.82% on both compositional splits. We further evaluate hand pose estimation accuracy under this compositional setting. Similarly, we report the mean and the standard deviation of MPVE in *mm* for all experiments. By comparing with HTT, we achieve state-of-the-art performance in hand pose estimation with the advantage of object reconstruction using superquadrics. Our strong performances across all settings demonstrate the importance of explicit reasoning about interactions with 3D geometric information.

Table 5.3: Error rates of compositional action recognition *H2O* (Kwon et al., 2021). We report classification accuracy in % and hand error in *mm* for the official split $\mathcal{S}_0$ and two additional compositional split settings, $\mathcal{S}_1$ and $\mathcal{S}_2$. The results are reported as the mean and standard deviation of the performance metric of interest, providing a comprehensive understanding of the model's performance across different split settings.

	H2O					
Method	$\mathcal{S}_0$(%) ↑	Hand ↓	$\mathcal{S}_1$(%) ↑	Hand ↓	$\mathcal{S}_2$(%) ↑	Hand ↓
MViTv2	90.08	-	60.64±2.3	-	52.32±5.7	-
HTT	86.36	35.6	71.13±2.7	37.4±2.3	59.88±2.6	41.5±1.7
Ours	**92.25**	**29.6**	**80.59±1.6**	**31.8±2.1**	**69.93±2.5**	**33.4±1.5**

Table 5.4: Error rates of compositional action recognition *FPHA* (Garcia-Hernando et al., 2018). We report classification accuracy in % and hand error in *mm* for the official split $\mathcal{S}_0$ and two additional compositional split settings, $\mathcal{S}_1$ and $\mathcal{S}_2$. The results are reported as the mean and standard deviation of the performance metric of interest, providing a comprehensive understanding of the model's performance across different split settings.

	FPHA			
Method	$\mathcal{S}_0$(%) ↑	Hand ↓	$\mathcal{S}_1$(%) ↑	Hand ↓
MViTv2	98.45	-	69.02±1.8	-
HTT	94.09	15.8	74.21±1.8	18.8±1.2
Ours	**98.74**	**13.6**	**85.80±1.4**	**13.9±0.9**

5.3.6 Ablation study

To motivate our design choices, we perform additional qualitative evaluation of our method with various key components disabled in Table 5.5. We evaluate the effectiveness of our collaborative learning framework by experimenting on a single appearance branch (row 1) and a two-branch network without gradient flow (row 2), which yields the lowest performance in *H2O*. Then, we observe significant performance drops by removing either one of the inter-branch classifiers (row 3, 4). These results demonstrate the effectiveness of our collaborative learning framework which encourages information sharing between two-branches. Further,

we are interested in finding out whether incorporating the interaction decoder is important as action class can be obtained by combining verb and noun predictions. We show that the interaction decoder can bring performance gain for verb and noun classifiers by additional supervision in row 5. Finally, we show that superquadric predictions can push the limit to the new state-of-the-art in all metrics.

Table 5.5: Ablation study of model architecture design on *H2O*. We report top-1 classification accuracy (%) and MPVE for hand error in *mm*. In addition to final action prediction, we include classification predictions for verb and noun from compositional reasoning.

Method	Verb(%) ↑	Noun(%) ↑	Top-1(%) ↑	Hand(*mm*) ↓
w/o geometric branch	-	-	78.91	-
w/o compositional reasoning	-	-	81.82	38.91
w/o verb classifier	-	85.61	83.45	37.15
w/o noun classifier	86.24	-	83.96	37.62
w/o interaction decoder	88.07	90.04	-	32.67
w/o superquadrics	90.18	91.56	89.85	31.85
Ours	**92.23**	**96.89**	**90.0**	**29.6**

5.4 Summary

In this chapter, we showed that we could recognise actions performed on unseen objects much more accurately than existing state-of-the-arts by explicitly leveraging 3D geometric information. We also demonstrated that superquadrics as a new object representation for action recognition to be effective. We validated our approach by extending existing datasets with compositional splits and achieved state-of-the-art performance.

Our approach relies on the expressiveness of superquadrics. First, we found that multi-superquadrics recovery is necessary to model more complex shapes, where preliminary point cloud segmentation can be helpful. Second, it remains challenging to capture non-

convex everyday objects such as cups, papers and clothes. Third, we found superquadrics recovery relies on the quality of the object template. In addition, we present an example failure case where a broken object model can heavily degrade the accuracy of superquadrics recovery in Figure 5.6. As illustrated, we show that the problem can be resolved by uniformly resampling the input object mesh.

Figure 5.6: Example failure case for superquadrics extraction. **First row:** We show input target object and the rotated views with wireframe. The visualisation of wireframe shows that the top and bottom of the input mesh is broken. **Second row:** We visualise the extracted superquadrics in grey and sampled point clouds in red. As the sampled point clouds are unable to capture the enclosed surface of the top and bottom, the estimated superquadric fails to represent the input object accurately. **Third row:** To fix this problem, we uniformly resample the input mesh such that it is watertight. We show that the resulting superquadric estimation can well-represent the target object.

Chapter Six

CONCLUSION

In this chapter, we summarise the contributions of this thesis in Section 6.1 and outline several limitations and potential directions for future works in Sections 6.2 and 6.3.

6.1 Contributions

This thesis has focused on modelling hand-object interactions from RGB images. The main contributions of our work are as follows:

- In Chapter 3, we presented a collaborative learning strategy to reconstruct hand and object from singe view RGB images. We demonstrated that sharing geometric information across hand and object learning branches can tackle the problem of mutual occlusion. We also designed a novel attention-guided graph convolution which can capture long-range dependencies from dynamic graph. As training such iterative learning framework is highly unstable, we introduced an unsupervised associative loss to stabilise the training and show that it can improve the feature transferring process. We found that our model is able to achieve highly physically plausible results without

contact loss terms.

- In Chapter 4, we proposed a novel Transformer-based framework that reconstructs two high fidelity hands from multi-view RGB images. We showed that the fundamental properties of the graph Laplacian from the spectral graph theory can be applied to a Transformer architecture. We also introduced a multi-view feature fusion strategy by leveraging soft-attention. To produce physically plausible hand reconstructions, we proposed an optimisation-based strategy which does not require any form of pre-computation and is more generalisable to other meshes as it is fully automatic.

- In Chapter 5, we extended our scope from reconstruction to action recognition. We studied the problem of compositional action recognition and showed that our approach is able to recognise actions performed on unseen objects much more accurately than existing state-of-the-arts by explicitly leveraging 3D geometric information. We also demonstrated that superquadric is an effective object representation for interaction recognition.

6.2 Limitations

We outline three major limitations of our approaches in the following.

Object topology. Our object reconstruction quality presented in Chapter 3 heavily relies on AtlasNet (Groueix et al., 2018). However, this architecture has a limitation: it is restricted to genus 0 topology, meaning that it can only handle sphere-like objects. Consequently, it can only reconstruct water-tight objects and unable to tackle objects with more complex topologies. Everyday objects like cups with handles or plates with significant curvature fall outside its reconstruction capabilities.

Reliance on training datasets. Our work presented in Chapters 3 and 4 was enabled by densely annotated datasets. Although these datasets allow different approaches to benchmark against, model trained on these dataset do not show generalisation ability to unseen problem settings or domains. This is because they are trained to regress instance-specific objects and prone to overfit to training data. As a result, our models can not explicitly tackle unseen object stances and do not exhibit cross-domain generalisation.

Superquadrics recovery. Our approach in Chapter 5 relies on the expressiveness of superquadrics. While single superquadric can well-represent objects within existing datasets, it remains challenging to capture non-convex everyday objects such as papers and clothings. Therefore, multi-superquadrics is necessary to model objects with more complexities. However, superquadrics recovery at this point still requires a significant amount of manual effort and it requires high quality object CAD models which is non-trivial in practice.

6.3 Future work

Despite the limitations described above, we believe our work has demonstrated promising results, and opened avenues for future research.

From modeling to understanding and reasoning. This thesis has focused on reconstruction of hands and objects without reasoning about the semantic meaning of human motions. We believe future works could investigate the problem of human (hand) pose estimation by embedding contextual information. This can be achieved by leveraging the power of large, pre-trained and multi-modal language models (LLM). Some progress has already been demonstrated by fine-tuning LLM to produce 3D poses or descriptions of poses from language queries (Y. Feng et al., 2023). The key difference between existing approaches is that

LLM embeds the world knowledge about the visual scenes and it can reason about human poses and motions. This allows for the development of more robust models to tackle conventional challenges, such as occlusion, which traditionally required large training datasets for effective solutions. In addition, it opens up research on human behaviour understanding beyond the traditional action recognition settings by investigating contextualised human activity. By the integration of visual information, language and 3D poses, we believe by developing multi-model models that are capable of leveraging highly diverse data can establish strong connection between human pose and the surrounding physical environment which offers a new path towards computers/agents that can perceive and understand humans.

Towards in-the-wild generalisation. Future research could focus on increasing the diversity and quantity of annotated hand-object interaction datasets to enhance generalisation ability. This can be achieved by improving the ease and speed of annotation for outside of laboratory settings, such as web images. Additionally, developing models that can learn from noisy annotations would alleviate the need for highly accurate dataset annotation and increase robustness to challenging viewing conditions and interactions. By exploring these avenues, we believe the field of hand-object interaction modeling can be advanced, allowing models to better perceive and understand real-world scenarios.

Hand-object contact modelling. Another promising future work direction would be to accurately estimate contact points during dynamic grasps and incorporate them as additional constraints within the pose annotation framework. Annotating hand-object contact points is crucial for robot interactions, especially for object-pick-up tasks and human-robot object handover tasks. Furthermore, an interesting application related to contact modeling is the modeling of object affordances. Learning affordances from diverse demonstrations has the potential to improve object manipulation planning. By representing affordances as 3D motions or trajectories, one can directly apply them to simulated environments. This ap-

proach enables efficient retargeting of affordances, enhancing the capabilities of the system in handling various objects and tasks.

References

Armagan, Anil, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, MingXiu Chen, Boshen Zhang, Fu Xiong, et al. (2020). "Measuring Generalisation to Unseen Viewpoints, Articulations, Shapes and Objects for 3D Hand Pose Estimation under Hand-Object Interaction". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 85–101.

Aubry, Mathieu and Bryan C Russell (2015). "Understanding Deep Features with Computer-generated Imagery". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 2875–2883.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer Normalization". In: *arXiv preprint arXiv:1607.06450*.

Baek, Seungryul, Kwang In Kim, and Tae-Kyun Kim (2018). "Augmented Skeleton Space Transfer for Depth-based Hand Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 8330–8339.

— (2019). "Pushing the Envelope for RGB-based Dense 3D Hand Pose Estimation via Neural Rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1067–1076.

— (2020). "Weakly-supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6121–6131.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Balcan, Maria Florina, Avrim Blum, Shai Fine, and Yishay Mansour (2012). "Distributed Learning, Communication Complexity and Privacy". In: *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, pp. 26–1.

Ballan, Luca, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys (2012). "Motion Capture of Hands in Action using Discriminative Salient Points". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 640–653.

Barr, Alan H (1981). "Superquadrics and Angle-Preserving Transformations". In: *IEEE Computer Graphics and Applications*, pp. 11–23.

Barron, Jonathan T, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman (2022). "Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5470–5479.

Battaglia, Peter W, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. (2018). "Relational inductive biases, deep learning, and graph networks". In: *arXiv preprint arXiv:1806.01261*.

Baxter, Jonathan (1997). "A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling". In: *Machine Learning*, pp. 7–39.

— (2000). "A Model of Inductive Bias Learning". In: *Journal of Artificial Intelligence Research*, pp. 149–198.

Blender (2023). *Cycles*. URL: https://www.cycles-renderer.org/.

Blum, Avrim, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao (2017). "Collaborative PAC learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Boukhayma, Adnane, Rodrigo de Bem, and Philip HS Torr (2019). "3D Hand Shape and Pose from Images in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10843–10852.

Brahmbhatt, Samarth, Cusuh Ham, Charles C Kemp, and James Hays (2019). "ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 8709–8719.

Bruna, Joan, Wojciech Zaremba, Arthur Szlam, and Yann LeCun (2014). "Spectral Networks and Locally Connected Networks on Graphs". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Cai, Yujun, Liuhao Ge, Jianfei Cai, and Junsong Yuan (2018). "Weakly-supervised 3D Hand Pose Estimation from Monocular RGB Images". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 666–682.

Cai, Yujun, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann (2019). "Exploiting Spatial-temporal Relationships for 3D Pose Estimation via Graph convolutional Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 2272–2281.

Calli, Berk, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar (2015). "The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research". In: *International Conference on Advanced Robotics (ICAR)*. IEEE, pp. 510–517.

Cao, Zhe, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik (2021). "Reconstructing Hand-Object Interactions in the Wild". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 12417–12426.

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (2020). "End-to-End Object Detection with Transformers". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 213–229.

Carreira, Joao and Andrew Zisserman (2017). "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6299–6308.

Caruana, Rich (1993). "Multitask Learning: A Knowledge-Based Source of Inductive Bias". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 41–48.

Chang, Angel X, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. (2015). "ShapeNet: An Information-Rich 3D Model Repository". In: *arXiv preprint arXiv:1512.03012*.

Chao, Yu-Wei, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. (2021). "DexYCB: A Benchmark for Capturing Hand Grasping of Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 9044–9053.

Chen, Dengsheng, Jun Li, Zheng Wang, and Kai Xu (2020). "Learning Canonical Shape Space for Category-Level 6D Object Pose and Size Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11973–11982.

Chen, Jiecao, Qin Zhang, and Yuan Zhou (2018). "Tight Bounds for Collaborative PAC Learning via Multiplicative Weights". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, Kai and Qi Dou (2021). "SGPA: Structure-Guided Prior Adaptation for Category-Level 6D Object Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 2773–2782.

Chen, Wei, Jinming Duan, Hector Basevi, Hyung Jin Chang, and Aleš Leonardis (2020a). "PointPoseNet: Point Pose Network for Robust 6D Object Pose Estimation". In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 2824–2833.

Chen, Wei, Xi Jia, Hyung Jin Chang, Jinming Duan, and Aleš Leonardis (2020b). "G2L-Net: Global to Local Network for Real-Time 6D Pose Estimation with Embedding Vector Features". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4233–4242.

Chen, Wei, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Aleš Leonardis (2021). "FS-Net: Fast Shape-based Network for Category-Level 6D Object Pose Estimation with Decoupled Rotation Mechanism". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1581–1590.

Chen, Xiaozhi, Huimin Ma, Ji Wan, Bo Li, and Tian Xia (2017). "Multi-View 3D Object Detection Network for Autonomous Driving". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1907–1915.

Chen, Xingyu, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng (2021). "Camera-Space Hand Mesh Recovery via Semantic Aggregation and Adaptive 2D-1D Registration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 13274–13283.

Chen, Zerui, Shizhe Chen, Cordelia Schmid, and Ivan Laptev (2023). "gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 12890–12900.

Chen, Zerui, Yana Hasson, Cordelia Schmid, and Ivan Laptev (2022). "AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 231–248.

Chevalier, Laurent, Fabrice Jaillet, and Atilla Baskurt (2003). "Segmentation and Superquadric Modeling of 3D Objects". In: *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*.

Cho, Hoseong, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek (2023). "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.

Cho, Junhyeong, Kim Youwang, and Tae-Hyun Oh (2022). "Cross-Attention of Disentangled Modalities for 3D Human Mesh Recovery with Transformers". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 342–359.

Choi, Hongsuk, Gyeongsik Moon, and Kyoung Mu Lee (2020). "Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 769–787.

Chung, Fan RK (1997). *Spectral Graph Theory*. American Mathematical Society.

Damen, Dima, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray (2018). "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 720–736.

Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst (2016). "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Dekel, Ofer, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao (2011). "Optimal Distributed Online Prediction". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 713–720.

Doosti, Bardia, Shujon Naha, Majid Mirbagheri, and David J Crandall (2020). "HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6608–6617.

Dwivedi, Vijay Prakash and Xavier Bresson (2021). "A Generalization of Transformer Networks to Graphs". In: *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*.

Dwivedi, Vijay Prakash, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson (2022). "Benchmarking Graph Neural Networks". In: *Journal of Machine Learning Research*, pp. 1–48.

Fan, Haoqi, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer (2020). *PySlowFast*. https://github.com/facebookresearch/slowfast.

Fan, Haoqi, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer (2021). "Multiscale Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 6824–6835.

Fan, Zicong, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges (2021). "Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-Pixel Part Segmentation". In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 1–10.

Fan, Zicong, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges (2023). "ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 12943–12954.

Feichtenhofer, Christoph, Haoqi Fan, Jitendra Malik, and Kaiming He (2019). "SlowFast Networks for Video Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 6202–6211.

Feng, Jie, Yan Wang, and Shih-Fu Chang (2016). "3D Shape Retrieval using a Single Depth Image from Low-Cost Sensors". In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1–9.

Feng, Yao, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black (2023). "PoseGPT: Chatting about 3D Human Pose". In: *arXiv preprint arXiv:2311.18836*.

Garcia-Hernando, Guillermo, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim (2018). "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 409–419.

Ge, Liuhao, Yujun Cai, Junwu Weng, and Junsong Yuan (2018). "Hand PointNet: 3D Hand Pose Estimation using Point Sets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 8417–8426.

Ge, Liuhao, Hui Liang, Junsong Yuan, and Daniel Thalmann (2016). "Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3593–3601.

— (2017). "3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1991–2000.

Ge, Liuhao, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan (2019). "3D Hand Shape and Pose Estimation from a Single RGB Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10833–10842.

Gehring, Jonas, Michael Auli, David Grangier, and Yann N Dauphin (2017). “A Convolutional Encoder Model for Neural Machine Translation”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 123–135.

Gilmer, Justin, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl (2017). “Neural Message Passing for Quantum Chemistry”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1263–1272.

Gong, Shunwang, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou (2019). “SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE.

Goyal, Raghav, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. (2017). “The "Something Something" Video Database for Learning and Evaluating Visual Common Sense”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 5842–5850.

Grabner, Alexander, Peter M Roth, and Vincent Lepetit (2018). “3D Pose Estimation and 3D Model Retrieval for Objects in the Wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3022–3031.

Grady, Patrick, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp (2021). “ContactOpt: Optimizing Contact to Improve Grasps”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1471–1481.

Grauman, Kristen, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. (2022). “Ego4D: Around the World in 3,000 Hours of Egocentric Video”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012.

Groueix, Thibault, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry (2018). "A Papier-Mâché Approach to Learning 3D Surface Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 216–224.

Gu, Chunhui, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. (2018). "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6047–6056.

Haeusser, Philip, Alexander Mordvintsev, and Daniel Cremers (2017). "Learning by Association–A Versatile Semi-Supervised Training Method for Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 89–98.

Hampali, Shreyas, Mahdi Rad, Markus Oberweger, and Vincent Lepetit (2020). "HOnnotate: A Method for 3D Annotation of Hand and Object Poses". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3196–3206.

Hampali, Shreyas, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit (2022). "Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11090–11100.

Han, Shangchen, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. (2020). "MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality". In: *ACM Transactions on Graphics (TOG)*, pp. 87–1.

Han, Shangchen, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. (2022). "UmeTrack: Uni-

fied multi-view end-to-end hand tracking for VR". In: *ACM Transactions on Graphics (TOG)*.

Hasson, Yana (2021). "Reconstructing Hands and Manipulated Objects from Images and Videos". PhD thesis. Inria.

Hasson, Yana, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid (2020). "Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 571–580.

Hasson, Yana, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid (2019). "Learning Joint Reconstruction of Hands and Manipulated Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11807–11816.

Hasson, Yana, Gül Varol, Ivan Laptev, and Cordelia Schmid (2021). "Towards Unconstrained Joint Hand-Object Reconstruction from RGB Videos". In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 659–668.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 770–778.

Hinterstoisser, Stefan, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit (2011). "Gradient Response Maps for Real-Time Detection of Textureless Objects". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 876–888.

Hinterstoisser, Stefan, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige (2016). "Going Further with Point Pair Features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 834–848.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 448–456.

Iqbal, Umar, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz (2018). "Hand Pose Estimation via Latent 2.5D Heatmap Regression". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 118–134.

Jhuang, Hueihan, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black (2013). "Towards Understanding Action Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 3192–3199.

Jiang, Zhanhong, Aditya Balu, Chinmay Hegde, and Soumik Sarkar (2017). "Collaborative Deep Learning in Fixed Topology Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Kabsch, Wolfgang (1976). "A Solution for the Best Rotation to Relate Two Sets of Vectors". In: *Acta Crystallographica*, pp. 922–923.

Kanazawa, Angjoo, Michael J Black, David W Jacobs, and Jitendra Malik (2018). "End-to-end Recovery of Human Shape and Pose". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 7122–7131.

Kantorov, Vadim and Ivan Laptev (2014). "Efficient Feature Extraction, Encoding and Classification for Action Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2593–2600.

Kar, Abhishek, Shubham Tulsiani, Joao Carreira, and Jitendra Malik (2015). "Category-Specific Object Reconstruction from a Single Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1966–1974.

Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei (2014). "Large-scale Video Classification with Convolutional Neural

Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1725–1732.

Karunratanakul, Korrawe, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang (2020). "Grasping Field: Learning Implicit Representations for Human Grasps". In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 333–344.

Kay, Will, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. (2017). "The Kinetics Human Action Video Dataset". In: *arXiv preprint arXiv:1705.06950*.

Kelion, Leo (2019). *Microsoft HoloLens 2 augmented reality headset unveiled*. URL: https://www.bbc.com/news/technology-47350884.

Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah (2022). "Transformers in Vision: A Survey". In: *ACM computing surveys (CSUR)*.

Kim, Dong Uk, Kwang In Kim, and Seungryul Baek (2021). "End-to-End Detection and Pose Estimation of Two Interacting Hands". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 11189–11198.

Kim, Tae Soo and Gregory D Hager (2020). "SAFCAR: Structured Attention Fusion for Compositional Action Recognition". In: *arXiv preprint arXiv:2012.02109*.

Kingma, Diederik P and Jimmy Ba (2015). "ADAM: A Method for Stochastic Pptimization". In: *Proceedings of the International Conference on Learning Representations (Proceedings of the International Conference on Learning Representations (ICLR))*.

Kipf, Thomas N and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kreuzer, Devin, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou (2021). "Rethinking Graph Transformers with Spectral Attention". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21618–21629.

Kulon, Dominik, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou (2020). "Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4990–5000.

Kwon, Taein, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys (2021). "H2O: Two Hands Manipulating Objects for First Person Interaction Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 10138–10148.

La Gorce, Martin de, David J Fleet, and Nikos Paragios (2011). "Model-based 3D Hand Pose Estimation from Monocular Video". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1793–1805.

Laine, Samuli and Tero Karras (2010). "Efficient Sparse Voxel Octrees". In: *Proceedings of the Interactive 3D Graphics and Games*. ACM, pp. 55–63.

Leonardis, Aleš, Ales Jaklic, and Franc Solina (1997). "Superquadrics for Segmenting and Modeling Range Data". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1289–1295.

Li, Mengcheng, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu (2022). "Interacting Attention Graph for Single Image Two-Hand Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2761–2770.

Li, Peiyi, Haibin Ling, Xi Li, and Chunyuan Liao (2015). "3D hand pose estimation using randomized decision forest with segmentation index points". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 819–827.

Li, Yanghao, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer (2022). "MViVv2: Improved Multiscale Vision Transformers for Classification and Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4804–4814.

Lin, Jiehong, Zewei Wei, Changxing Ding, and Kui Jia (2022). "Category-Level 6D Object Pose and Size Estimation using Self-Supervised Deep Prior Deformation Networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 19–34.

Lin, Kevin, Lijuan Wang, and Zicheng Liu (2021a). "End-to-End Human Pose and Mesh Reconstruction with Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1954–1963.

— (2021b). "Mesh Graphormer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 12939–12948.

Lin, Zhi-Hao, Sheng-Yu Huang, and Yu-Chiang Frank Wang (2020). "Convolution in the Cloud: Learning Deformable Kernels in 3D Graph Convolution Networks for Point Cloud Analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1800–1809.

Liu, Shaowei, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang (2021). "Semi-Supervised 3D Hand-Object Poses Estimation with Interactions in Time". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 14687–14697.

Liu, Weixiao, Yuwei Wu, Sipu Ruan, and Gregory S Chirikjian (2022). "Robust and Accurate Superquadric Recovery: A Probabilistic Approach". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2676–2685.

Liu, Zhijian, Haotian Tang, Yujun Lin, and Song Han (2019). "Point-Voxel CNN for Efficient 3D Deep Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Lomonaco, Vincenzo and Davide Maltoni (2017). "CORe50: A New Dataset and Benchmark for Continuous Object Recognition". In: *Conference on Robot Learning*. PMLR, pp. 17–26.

Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black (2015). "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Transactions on Graphics (TOG)*, pp. 851–866.

Malik, Jameel, Ibrahim Abdelaziz, Ahmed Elhayek, Soshi Shimada, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, and Didier Stricker (2020). "HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation from a Single Depth Map". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 7113–7122.

Malik, Jameel, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier Stricker (2018). "DeepHPS: End-to-end Estimation of 3D Hand Pose and Shape by Learning from Synthetic Depth". In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 110–119.

Malladi, Ravi, James A Sethian, and Baba C Vemuri (1995). "Shape Modeling with Front Propagation: A Level Set Approach". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 158–175.

Mandikal, Priyanka and Kristen Grauman (2020). "Dexterous Robotic Grasping with Object-Centric Visual Affordances". In: *arXiv preprint arXiv:2009.01439*.

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh (2008). "Domain Adaptation with Multiple Sources". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

— (2009). "Domain Adaptation: Learning Bounds and Algorithms". In: *Conference on Learning Theory*.

Materzynska, Joanna, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell (2020). "Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1049–1059.

Meltzoff, Andrew N (1995). "Understanding the Intentions of Others: Re-enactment of Intended Acts by 18-Month-Old Children." In: *Developmental Psychology*, p. 838.

Mescheder, Lars, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger (2019). "Occupancy Networks: Learning 3D Reconstruction in Function Space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4460–4470.

Miller, A.T. and P.K. Allen (2004). "GraspIt! A Versatile Simulator for Robotic Grasping". In: *IEEE Robotics & Automation Magazine*, pp. 110–122.

Möller, Tomas and Ben Trumbore (1997). "Fast, Minimum Storage Ray/Triangle Intersection". In: *Journal of Graphics Tools*.

Monti, Federico, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein (2017). "Geometric Deep Learning on Graphs and Manifolds using Mixture Model CNNs". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5115–5124.

Moon, Gyeongsik and Kyoung Mu Lee (2020a). "I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 752–768.

Moon, Gyeongsik, Takaaki Shiratori, and Kyoung Mu Lee (2020b). "DeepHandMesh: A Weakly-supervised Deep Encoder-Decoder Framework for High-fidelity Hand Mesh Modeling". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 440–455.

Moon, Gyeongsik, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee (2020c). "InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 548–564.

Mueller, Franziska, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt (2018). "Ganerated Hands for Real-Time 3D Hand Tracking from Monocular RGB". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 49–59.

Mueller, Franziska, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt (2019). "Real-Time Pose and Shape Reconstruction of Two Interacting Hands with a Single Depth Camera". In: *ACM Transactions on Graphics (TOG)*, pp. 1–13.

Müller, Thomas, Alex Evans, Christoph Schied, and Alexander Keller (2022). "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". In: *ACM Transactions on Graphics (TOG)*, pp. 1–15.

Nakamura, Yuzuko C, Daniel M Troniak, Alberto Rodriguez, Matthew T Mason, and Nancy S Pollard (2017). "The Complexities of Grasping in the Wild". In: *International Conference on Humanoid Robotics*. IEEE, pp. 233–240.

Newman, Paul, David Cole, and Kin Ho (2006). "Outdoor SLAM using Visual Appearance and Laser Ranging". In: *Proceedings of the International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1180–1187.

Nguyen, Huy and Lydia Zakynthinou (2018). "Improved Algorithms for Collaborative PAC Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Oberweger, Markus, Mahdi Rad, and Vincent Lepetit (2018). "Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 119–134.

Oberweger, Markus, Paul Wohlhart, and Vincent Lepetit (2015). "Training a Feedback Loop for Hand Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 3316–3324.

Oikonomidis, Iasonas, Nikolaos Kyriazis, and Antonis A Argyros (2012). "Tracking the Articulated Motion of Two Strongly Interacting Hands". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1862–1869.

Pavlakos, Georgios, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black (2019). "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10975–10985.

Pavlakos, Georgios, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis (2018). "Learning to Estimate 3D Human Pose and Shape from a Single Color Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 459–468.

Prasad, Mukta, Andrew Fitzgibbon, Andrew Zisserman, and Luc Van Gool (2010). "Finding Nemo: Deformable Object Class Modelling using Curve Matching". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1720–1727.

Qi, Charles R, Hao Su, Kaichun Mo, and Leonidas J Guibas (2017a). "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 652–660.

Qi, Charles R, Li Yi, Hao Su, and Leonidas J Guibas (2017b). "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Qian, Chen, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun (2014). "Realtime and Robust Hand Tracking from Depth". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1106–1113.

Qian, Guocheng, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem (2022). "PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 23192–23204.

Rad, Mahdi and Vincent Lepetit (2017). "BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 3828–3836.

Rad, Mahdi, Markus Oberweger, and Vincent Lepetit (2018). "Feature Mapping for Learning Fast and Accurate 3D Pose Inference from Synthetic Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4663–4672.

Rangesh, Akshay, Eshed Ohn-Bar, and Mohan M Trivedi (2016). "Hidden Hands: Tracking Hands with an Occlusion Aware Tracker". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 19–26.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi (2016). "You Only Look Once: Unified, Real-Time Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 779–788.

Remelli, Edoardo, Anastasia Tkach, Andrea Tagliasacchi, and Mark Pauly (2017). "Low-Dimensionality Calibration through Local Anisotropic Scaling for Robust Hand Model Personalization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 2535–2543.

Riegler, Gernot, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger (2017). "OctNetFusion: Learning Depth Fusion from Data". In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 57–66.

Romero, Javier, Dimitrios Tzionas, and Michael J. Black (2017). "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Transactions on Graphics (TOG)*.

Rong, Yu, Jingbo Wang, Ziwei Liu, and Chen Change Loy (2021). "Monocular 3D Reconstruction of Interacting Hands via Collision-Aware Factorized Refinements". In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 432–441.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)*, pp. 211–252.

Samarth, Brahmbhatt, Tang Chengcheng, D Twigg Christopher, C Kemp Charles, and Hays James (2020). "ContactPose: A Dataset of Grasps with Object Contact and Hand Pose". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 361–378.

Santoro, Adam, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap (2017). "A Simple Neural Network Module for Relational Reasoning". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Shan, Dandan, Jiaqi Geng, Michelle Shu, and David Fouhey (2020). "Understanding Human Hands in Contact at Internet Scale". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 9869–9878.

Sigurdsson, Gunnar A, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari (2018). "Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos". In: *arXiv preprint arXiv:1804.09626*.

Sigurdsson, Gunnar A, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta (2016). "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 510–526.

Simon, Tomas, Hanbyul Joo, Iain Matthews, and Yaser Sheikh (2017). "Hand Keypoint Detection in Single Images using Multiview Bootstrapping". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1145–1153.

Simonyan, Karen and Andrew Zisserman (2014). "Two-Stream Convolutional Networks for Action Recognition in Videos". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Smith, Breannan, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori (2020). "Constraining Dense Hand Surface Tracking with Elasticity". In: *ACM Transactions on Graphics (TOG)*, pp. 1–14.

Solina, Franc and Ruzena Bajcsy (1990). "Recovery of Parametric Models from Range Images: The Case for Superquadrics with Global Deformations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 131–147.

Song, Guocong and Wei Chai (2018). "Collaborative Learning for Deep Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah (2012). "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild". In: *arXiv preprint arXiv:1212.0402*.

Sorkine, Olga and Marc Alexa (2007). "As-Rigid-As-Possible Surface Modeling". In: *Symposium on Geometry processing*. Citeseer, pp. 109–116.

Spurr, Adrian, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz (2020). "Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 211–228.

Spurr, Adrian, Jie Song, Seonwook Park, and Otmar Hilliges (2018). "Cross-modal Deep Variational Hand Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 89–98.

Sridhar, Srinath, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt (2015). "Fast and Robust Hand Tracking using Detection-guided Optimization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3213–3221.

Sun, Pengzhan, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, and Chuang Gan (2021). "Counterfactual Debiasing Inference for Compositional Cction Recognition". In: *Proceedings of the International Conference on Multimedia*. ACM, pp. 3220–3228.

Sun, Xiao, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun (2015). "Cascaded Hand Pose Regression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 824–832.

Sundermeyer, Martin, Zoltan-Csaba Marton, Maximilian Durner, and Rudolph Triebel (2020). "Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection". In: *International Journal of Computer Vision (IJCV)*, pp. 714–729.

Tan, David Joseph, Thomas Cashman, Jonathan Taylor, Andrew Fitzgibbon, Daniel Tarlow, Sameh Khamis, Shahram Izadi, and Jamie Shotton (2016). "Fits Like a Glove: Rapid and Reliable Hand Shape Personalization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5610–5619.

Tan, Mingxing and Quoc Le (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR.

Tang, Danhang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim (2014). "Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3786–3793.

Tatarchenko, Maxim, Alexey Dosovitskiy, and Thomas Brox (2017). "Octree Generating Networks: Efficient Convolutional Architectures for High-Resolution 3D Outputs". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 2088–2096.

Tatarchenko, Maxim, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox (2019). "What Do Single-view 3D Reconstruction Networks Learn?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3405–3414.

Taylor, Jonathan, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. (2016). "Efficient and Precise Interactive Hand Tracking through Joint, Continuous Optimization of Pose and Correspondences". In: *ACM Transactions on Graphics (TOG)*, pp. 1–12.

Taylor, Jonathan, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi (2017). "Articulated Distance Fields for Ultra-Fast Tracking of Hands Interacting". In: *ACM Transactions on Graphics (TOG)*, pp. 1–12.

Tekin, Bugra, Federica Bogo, and Marc Pollefeys (2019). "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4511–4520.

Tian, Meng, Marcelo H Ang, and Gim Hee Lee (2020). "Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 530–546.

Tkach, Anastasia, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew Fitzgibbon (2017). "Online Generative Model Personalization for Hand Tracking". In: *ACM Transactions on Graphics (TOG)*, pp. 1–11.

Tompson, Jonathan, Murphy Stein, Yann Lecun, and Ken Perlin (2014). "Real-Time Continuous Pose Recovery of Human Hands using Convolutional Networks". In: *ACM Transactions on Graphics (TOG)*, pp. 1–10.

Tse, Tze Ho Elden, Kwang In Kim, Aleš Leonardis, and Hyung Jin Chang (2022a). "Collaborative Learning for Hand and Object Reconstruction with Attention-guided Graph Convolution". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1664–1674.

Tse, Tze Ho Elden, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, and Bardia Doosti (2023). "Spectral Graphormer: Spectral Graph-based Transformer for Egocentric Two-Hand Reconstruction using Multi-View Color Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 14666–14677.

Tse, Tze Ho Elden, Zhongqun Zhang, Kwang In Kim, Aleš Leonardis, Feng Zheng, and Hyung Jin Chang (2022b). "S^{2}Contact: Graph-based Network for 3D Hand-Object Contact Estimation with Semi-Supervised Learning". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 568–584.

Tzionas, Dimitrios, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall (2016). "Capturing Hands in Action using Discriminative Salient Points and Physics Simulation". In: *International Journal of Computer Vision (IJCV)*.

Valiant, Leslie G (1984). "A Theory of the Learnable". In: *Communications of the ACM*, pp. 1134–1142.

Varol, Gül, Ivan Laptev, and Cordelia Schmid (2017). "Long-term Temporal Convolutions for Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1510–1517.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio (2018). "Graph Attention Networks". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

VR, Google AR & (2024). *AR Glasses Experiences*. URL: https://arvr.google.com/.

Wan, Chengde, Thomas Probst, Luc Van Gool, and Angela Yao (2019). "Self-supervised 3D Hand Pose Estimation through Training by Fitting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10853–10862.

Wan, Chengde, Thomas Probst, Luc Van Gool, and Angela Yao (2017). "Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 680–689.

— (2018). "Dense 3D Regression for Hand Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5147–5156.

Wan, Chengde, Angela Yao, and Luc Van Gool (2016). "Hand Pose Estimation from Local Surface Normals". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 554–569.

Wang, He, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas (2019). "Normalized Object Coordinate Space for Category-Level 6D Object

Pose and Size Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2642–2651.

Wang, Jialei, Mladen Kolar, and Nathan Srerbo (2016). "Distributed Multi-Task Learning". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR.

Wang, Jiayi, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt (2020). "RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video". In: *ACM Transactions on Graphics (TOG)*, pp. 1–16.

Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool (2016). "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 20–36.

Wang, Nanyang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang (2018). "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 52–67.

Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He (2018). "Non-local Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 7794–7803.

Wang, Yue, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon (2019). "Dynamic Graph CNN for Learning on Point Clouds". In: *ACM Transactions on Graphics (TOG)*, pp. 1–12.

Wen, Yilin, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang (2023). "Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21243–21253.

Wu, Zhirong, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao (2015). "3D ShapeNets: A Deep Representation for Volumetric Shapes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1912–1920.

Xiang, Yu, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox (2018a). "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes". In.

— (2018b). "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes". In: *Robotics: Science and Systems (RSS)*.

Xu, Hongyi, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu (2020). "GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6184–6193.

Xu, Keyulu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka (2018). "How Powerful are Graph Neural Networks?" In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yang, Lixin, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu (2022). "OakInk: A :arge-scale Knowledge Repository for Understanding Hand-Object Interaction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 20953–20962.

Yang, Lixin, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu (2021). "CPF: Learning a Contact Potential Field to Model the Hand-Object Interaction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 11097–11106.

Yang, Siyuan, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot (2020). "Collaborative Learning of Gesture Recognition and 3D Hand Pose Estimation with Multi-Order

Feature Analysis". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 769–786.

Ye, Qi and Tae-Kyun Kim (2018). "Occlusion-aware Hand Pose Estimation Using Hierarchical Mixture Density Network". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 801–817.

Ye, Qi, Shanxin Yuan, and Tae-Kyun Kim (2016). "Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 346–361.

Ye, Yufei, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani (2023). "Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 19717–19728.

You, Haoxuan, Yifan Feng, Rongrong Ji, and Yue Gao (2018). "PVNet: A Joint Convolutional Network of Point Cloud and Multi-View for 3D Shape Recognition". In: *Proceedings of the International Conference on Multimedia*. ACM, pp. 1310–1318.

Yuan, Shanxin, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. (2018). "Depth-based 3D Hand Pose Estimation: From Current Achievements to Future Goals". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2636–2645.

Yuan, Shanxin, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim (2017). "BigHand2.2M Benchmark: Hand Pose Dataset and State of the Art Analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4866–4874.

Zhang, Baowen, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang (2021). "Interacting Two-Hand 3D Pose and Shape Reconstruction from

Single Color Image". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 11354–11363.

Zhao, Hengshuang, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun (2021). "Point Transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 16259–16268.

Zheng, Linfang, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang (2023). "HS-Pose: Hybrid Scope Feature Extraction for Category-level Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 17163–17173.

Zhou, Bolei, Alex Andonian, Aude Oliva, and Antonio Torralba (2018). "Temporal Relational Reasoning in Videos". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 803–818.

Zhou, Yuxiao, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu (2020). "Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5346–5355.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 2223–2232.

Zhu, Zhuotun, Xinggang Wang, Song Bai, Cong Yao, and Xiang Bai (2016). "Deep Learning Representation using Autoencoder for 3D Shape Retrieval". In: *Neurocomputing*, pp. 41–50.

Zia, M Zeeshan, Michael Stark, and Konrad Schindler (2015). "Towards Scene Understanding with Detailed 3D Object Representations". In: *International Journal of Computer Vision (IJCV)*, pp. 188–203.

Zimmermann, Christian and Thomas Brox (2017). "Learning to Estimate 3D Hand Pose from Single RGB Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 4903–4911.

Zimmermann, Christian, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox (2019). "FreiHand: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 813–822.